Analog IC design - the obsolete book

Ricardo Erckert

December 2022

Contents

1 Preface

'Analog IC design – the obsolete book' – So why is it obsolete and I write it nevertheless? Probably because I am a nerd writing useless stuff for other nerds. No, not quite. The truth is analog ICs are almost extinct like relays, valves, discrete transistors..... All these things now are part of system integration.

Analog design does not scale with the advances of modern technologies. So manufacturers tend to push more and more functionality into the digital design. Today's so called analog chips consist of some 10K analog transistors together with some millions of digital gates on one chip. But those 10K analog transistors may well consume the same area as the 2 million gate logic sitting on the same dice.

So why am I writing about analog design is stead of mixed signal? Because my digital know how is ridiculously low.

How did I learn analog design? To tell you the truth: I never did! In the old days we learned about telegraphy equations, power plants, acoustics, electrical networks, how to build digital filters with the legendary PDP11..... And I had some data books holding transistor level circuits of pioneers like Bob Widlar or Bruno Murari. Later I even met some of them (So now you can guess the color of the hair I am loosing) and learned some more circuits.

In the beginning this wasn't a book. It was a collection of drawings I used to train younger engineers starting on the field of analog chip design. By the time I figured out I have to add comments to the drawings. Slowly the collection of drawings turned into an unorganized book...

Today I'm often confronted with the question: W.T.F. ! I know I have solved that some years ago, but I forgot the most important considerations and equations!

So today I'm writing this book for myself to document how I solved all the stuff that keeps repeating again and again.

This isn't really a scientific book. I don't see a point in poking out the last percent while I exactly know production spread is magnitudes higher than what I optimize. For this reason it doesn't make sense to develop equations that describe 3rd order effects. This book is more something like a cook book. If you need more details look at the literature index.

1.1 Using the book

The book holds some equations I regard as essential. In the recent years I found out that the mathematical education of students today seems to have fallen way behind my own mathematical education. Some of this degradation seems to be caused by the PISA comparison mechanism and not by the students themselves. (PISA tests the capability of reproducing rather than the ability of finding new solutions to compare education levels in Europe. I think this is a fatal approach because it doesn't support engineering creativity.) To give some basic mathematical help for younger engineers I added the mathematical appendix without giving any mathematical proofs (so again it is a cook book. For further details I suggest having a look at[16]).

1.2 If you find mistakes

The book was written without any help of editors or proof readers. I would be surprised if there are no mistakes. In case you find severe mistakes please let me know so I can correct them. The book is a living document. I will keep writing and updating. Updates will be made available time after time on my WEB page.

1.3 Copying

The contents of this book is publicly known from earlier publications (Some circuits in fact date back to the 1920s!). Nothing new under the sun. Since I learned that licenses owned by publishing companies have become a major obstacle for public education at schools and universities today this book intentionally is free.

- No publishing company and no other editors are involved in this book project.
- All figures were drawn by myself (in most cases using public domain software to avoid any conflicts) and are free.
- You may make copies
- You may distribute copies provided you name the source (Title, author, year of the revision)
- You may include pages of this book in your own presentations provided you name the source (Title, author, year of the revision)
- You may convey a work based on this book, provided that you also meet all of these conditions:
  a) The work must carry prominent notices stating that you modified it, and giving a relevant date.
  b) The work must carry prominent notices stating that it is released free for public use including copying and conveying it.
  c) You must license the entire work, as a whole, under the same conditions as this book to anyone who comes into possession of a copy.

1.4 Warranty and Liability

The book holds simplified (conceptual) schematics, drawings and some code snippets. These simplified examples can't be expected to be fool proof for a chip integration. In most cases you will need to enhance them (adding power down, test modes etc.) to really make them fit to your application.

- Disclaimer of Warranty:
  THERE IS NO WARRANTY FOR THE PROGRAM OR SCHEMATIC, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM OR SCHEMATIC "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM OR SCHEMATIC IS WITH YOU. SHOULD THE PROGRAM OR SCHEMATIC PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

- Limitation of Liability.
  IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MODIFIES AND/OR CONVEYS THE PROGRAM OR SCHEMATIC AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM OR SCHEMATIC (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM OR SCHEMATIC TO OPERATE WITH ANY OTHER PROGRAMS OR CIRCUITS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Hm, looks a bit like the GNU public license. I fact it is my intention to make this book public available similar to the way the GNU public license does it for software.

2 The top level of a chip

2.1 How to get the correct data of a chip

In the old days the data of a chip was kept in a UNIX directory. To produce the masks you needed the GDS2 data. GDS2 mainly deals with geometries and some hierarchical information. Besides that it is a file similar to a tar file. To produce the GDS2 file you need a layout. The layout is usually in a design library. To get going you just needed the following informations:

1. The UNIX path to the library
2. The name of the top cell (usually the view layout)
3. The path where to place the GDS2 file

To check consistency of schematic and layout you just had to run an extract run and run an (This creates a netlist of the layout) LVS (layout versus schematic). The LVS compares if the netlist of the schematic and the netlist of the layout are equal. If both are equal you get a pass, if they differ you get an error listing. If the error listing holds more than only parameter errors you better forget about the GDS2 and start fixing.....

If the LVS is error free you could do a stream out and send the GDS2 file to the mask shop.

Typical naming convention was in the top level cell. It consisted of the chip name and a trailing version description such as A1 for first release, first metal or B2 for second release (all layer change) second metal version. To make things even more fool proof it was common to have a chip name written in metal on the layout and to name the GDS2 file in exactly the same way. (Something like U412B2.gds)

Note: Usually the stream out procedure uses view layout only! views such as layout1, layout2 etc for different versions of the layout are ignored! Never create different layouts of the same schematic with names differing from layout. GDS stream out will ignore them even if your design system shows them correctly!

2.1.1 Versioning tools

These simple days are over! Now we have versioning tools that can create a lot of more confusion. Within one database you may have dozens of versions of the schematic and the layout (and whatever view else you might have). Which schematic version is consistent with which layout version is only mapped by the tag! The person issuing the tag is fully responsible for the correctness! It may well be that there is a schematic version 1.52 that maps to a layout version 1.98 and you have almost no chance finding this out without a correct tag. It is no more the name of the top cell that is giving you the information. Even worse, now there may be a U412B2 layout version 1.3 that

is completely different from U412B2 layout version 1.32 although both are still named U412B2! It is the tag only telling you what you work on.

Tags are used in an inflationary way not only describing tape out versions but also all kinds of intermediate versions used for simulation etc.

To order that mess it is mandatory to have some way of assigning which version of a block belongs to which top level. Unfortunately this usually is done in the population process. Population is the process of linking the data sitting in something called vault (Any relation Vault <-> Voldemort? The dark magician in Harry Potter? Who knows?). he vault is a big chunk of data that everybody hopes he will not have to debug. In this vault there are dozens of versions of each item (no matter if it is a schematic, symbol, layout, netlist, software, grandma's cookie recipe.....whatever). Populating means some software creates links between these files representing the items and your local directory.

In most cases default is simply linking the latest version found in the vault - no matter if it is consistent or not! (So on the top level schematic you might find buses with a different width than the module pins require. The linking software has no intelligence to see that.)

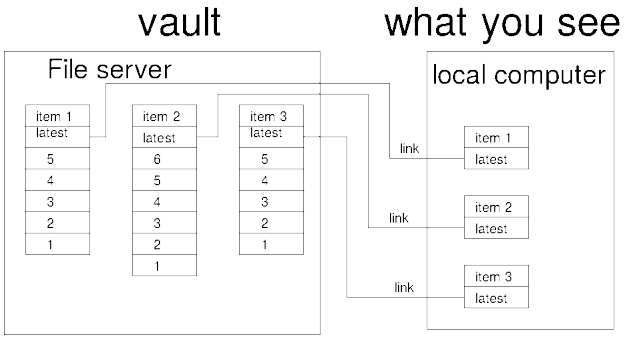


Figure 2.1: Population using the default "latest"

Alternatively the links can point to older versions. This makes sense if the latest version still is work in progress. For instance the layout corresponding to the latest schematic is not yet finished. Using such a work in progress version would lead to LVS (layout versus schematic test) and DRC (design rule check) errors in the top level data base.

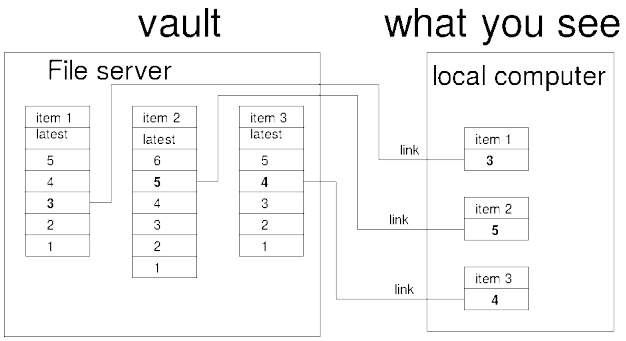


Figure 2.2: Population of a tagged vault represents the versions item 1 ver. 3, item 2 ver. 5 and item 3 ver. 4 to the chip designer.

The configuration of the correct links is one of the most risky activities in the whole design process of a chip. One single wrong link can lead to a useless mask set. Therefore it is good engineering practice to create the tagged version, close the data base and repopulate using the automatic procedure that creates the links. Once this is done run DRC and LVS to verifify that at least there is no mayor disaster such as shorts or opens in the design.

This still does not guarantee full functionality! There could be versions of modules that have identical pins but still have different behavior. So these will not even be caught by DRC and LVS. To find out such issues the whole top level should also be resimulated before ordering the masks. For this reason project planning should provide some weeks between tape out and mask making.

Getting started: One of the first questions entering a project is how to start the population. Typically there is a script (usually a shell script) to be called. Something like

```
enter_project 'devicename'
```

This command should list all tagged versions and 'default'. 'default means populate the latest version. Every tagged version if selected should search a database describing which versions are to be linked. In very simple cases populating a tagged version could simply call some shell code holding many lines such as:

```
ln -s /vault.../module/view/version ./module/view
```

Not very elegant but working. More elegant would be a perl script (or whatever else is good in handling regular expressions) reading a data base.

Once this is done the chip designer should see a library at his local computer that corresponds the version selected.

Unfortunately one of the best kept secrets in many design departments is where to find the list of module versions belonging to a specific database. So desperate chip designer frequently fuzz around simulating and redesigning the wrong module version. (I wish I could have 1% of the money burnt by working on the wrong version. I wouldn't need to write a book and would go climbing instead!)

Check out: To edit data the data first has to be checked out. Checking out creates a copy of the version checked out (typically 'latest' will have a number only and there now is a new 'latest'.)

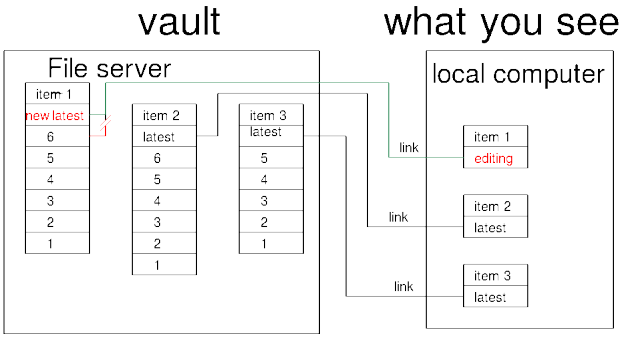


Figure 2.3: Editing under design sync. The green link is created new after checkout.

During editing revision 6 is kept but revision 7 is not yet assigned a version number. It can still be revoked by canceling the check out. The 'new latest version becomes a valid revision (that can no more be revoked) as soon as it is checked in.

The DSS crash problem Design sync stores who checked out data and from which server the checkout was done. If the session crashes while something was checked out and the same user starts a new session but gets a different server editing is no more possible. (Miller can't check out because the design is already checked out by Miller....). There are two possible solutions to fix the problem:

1. An administrator must fix the problem hacking the files directly inside the vault
2. Copy the design to a new name and edit the new design. In the design level above now the new design must be instantiated. The trace-ability of changes now is lost.

The same happens if you close the server connection while the CAD software is still running.

Team blockage by design sync: If a design (possibly by accident without intention) was checked out by one user nobody else can edit anymore. This happens frequently using for instance the Cadence mixed mode simulation environment. Before going into vacation check if there are any checkouts that you may have done unintentionally. If you block the team they can do nothing but let an administrator hack the vault or rename everything.

As a consequence working in a team it is required to check in much more often than there really are completed design steps. So most libraries hold about three times more checked in intermediate versions of a design than consistent design states. Version numbers of schematic, symbol and layout will run out of sync rapidly. Therefore it is necessary to tag versions. Tagging assigns the version numbers of different views to each other. (For instance a symbol ver. 3 could be associated with schematic ver. 20 and layout ver. 8 and be used in top level ver. 4.....) A loss of these assignments makes a design almost useless even if the complete libraries still exist. (This is similar to the configuration of a big software project such as a complete operating system.)

Risks of versioning: The highest risk of versioning tools is to work on an outdated copy of a data base without being aware of it. Using DesSync GUI to check out parts of a data base can't be recommended at all!

Splitting a project into several databases will sooner or later lead to a crash because one database is checked out as current and in edit mode while you see outdated parts of the other data base (in read mode, not repopulated recently).

Imagine you see obsolete symbols coming from one data base while you are editing the schematics instantiating these symbols in the other data base. You will end up with broken connections at tape out! If you are lucky an LVS check will intercept this. If the LVS uses the same partial check out even the LVS won't find it and it depends on your luck which data eventually is used for GDS stream out. There is no automatic verification tool checking the correctness of your population!

Now since you are aware of the main risks, let's get going!

2.2 System on a chip

SOC or system on a chip - the newest buzz word! Modern chips try to include more and more functions. Logic is cheap because it scales nicely with every new generation of technologies. So IC manufacturers try to add value plugging more and more logic into the chips. These chips require big teams to design them. Depending on the role of the team member everybody has a different perception of what a certain block does and what are the important interfaces.

An analog designer will regard a pin as something providing voltage and current with a certain impedance and signal swing. Some of them regard ground and supply as superconductive. The more experienced ones might even already have heard of pin inductance.

The system designer being more used to tools like Mathlab or compilers may already forget these unimportant details. He is more of a generalist. A pin is something providing a floating point number - and that's it!

The digital designer works with bits and bytes. Floating point MUST be avoided because it leads to very expensive arithmetic units. If any possible it all gets truncated to integer. (You can simulate a complete delta sigma converter in verilog. For the analog signal you just create a 32 bit bus and choose the LSB as $10\mu V$. Why fuzz with such strange concepts as floating point or voltages that just consume computation power?). Furthermore digital often has no power supply. This is SOP (somebody other's problem).

The project manager is in charge of dates and cost. Why should he worry about volts, amperes or data types?

Since cost matters the project manager may have hired architects ,WEB designers or unemployed professors to do the layout. Expertise I found in the layout teams can vary extremely from zero to Einstein's level!

CAD software specialists neither care for physical properties (like voltage and current) nor for such primitive things like integers. They rather work with software objects. Usually these guys are in love with inherited connections.

Technology developers want to make the transistors smaller and don't understand why these stupid analog guys want to work with such horribly high voltages like 3.3V. If analog designers were more clever and would learn to use transistors with 100mV V_{dsmax} the design could get much smaller! (at some loss of SNR, why worry?)

So there is plenty of chances for a desasterous misunderstanding in the team. Thus the most important question is: With whom do I discuss and which parameter does he have a concept of. So if you run into a problem and you have to discuss it with the project manager express it in cost and delay. If you have to discuss the same problem with the digital designer try to express it in bits or integer numbers. And if Einstein is doing layout for you you can discuss how much heavier you transistor gets due to the impact of an ESD pulse....(relative to a matched transistor that is not getting an ESD pulse).

2.2.1 Where exactly is "ground"?

Since we are discussing mixed signal ICs and analog ICs this is the most decisive question of all! We don't use superconductors and every conducting structure is surrounded by a magnetic field. Chip ground has not much to do with board ground. There is a bond wire between chip ground and board ground plane. At 1 GHz RF currents of 10mA can easily produce ground bounce of 400mVpp due to pin inductance. If you add some parasitic capacities you even can get a resonance boosting this to the volt-level.

Power transistors can produce pulse currents of several amperes. In case of rush in shorts these currents can have current edges of some hundred mA/ns. This can lead to ground or supply transients of several volts for some ns. So power ground must be separated from the ground of low voltage CMOS circuits (a 3nm gate oxide logic will not survive if ground jumps up 3V because there is a short of a power transistor).

In power ICs substrate often must be tied to the power ground because substrate acts as a free wheeling diode in case the power stage has to drive inductive loads. As a consequence substrate must be regarded as noisy from the point of view of precision analog circuits.

Substrate is tightly coupled to the dice pad. In case of using an exposed dice pad the substrate may be connected to board ground via a parasitic inductance of only some 10pH while all other grounds have inductance to board ground in the range of some nH. Low resistive substrate will rather follow the dice pad voltage than the voltage of any of the on chip ground traces.

At frequencies higher than about 200MHz the capacitive coupling between different ground nets becomes dominant (capacities of ESD protections etc.). So ground nets expected to be independent of each other at high frequencies have a lot of cross talk. In the supply concept shown the power ground (PGND) is expected to have RF disturbances of some hundred mV rms plus transients of +-5V (relative to board GND). The digital ground and the I/O ground is expected to have not more than some 10mV of noise.

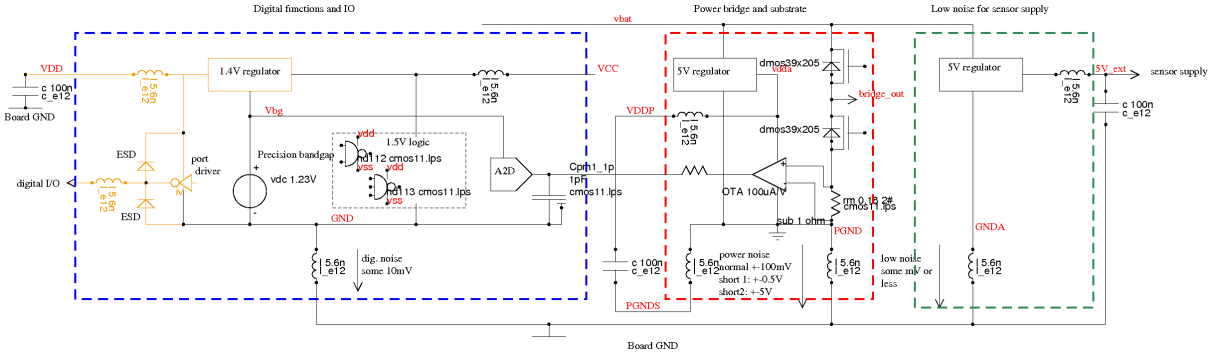


Figure 2.4: Example of a supply concept with separate grounds for power stages, digital functions and low noise analog functions

Both noise levels still are not acceptable for some high performance analog functions (sensor supply on the right side). So these analog functions have a third ground pin called GNDA.

All circuits belonging to the power stage refer to PGND. Even the regulator blocking of the driver supply must follow the bounce of the power ground (100nF capacitor connected to PGNDS). In this specific case of a power bridge the substrate diode may have to carry several amperes. Therefore here the substrate must be connected to PGND.

All circuits belonging to the high precision analog part refer to GNDA. Since substrate is connected to PGND every component of the precision analog part must be shielded from substrate noise!

Logic and the delta sigma A2D are in the digital domain. The input signal of the A2D is low pass filtered with reference to the digital ground (GND). The input of the A2D must be shielded from substrate noise.

All application measurements refer to board ground, which is neither identical with PGND nor with AGND, nor with GND!

2.2.2 Signals: If anything can go wrong it will!

The most difficult part of a system on chip design is the team communication. To solve this problem software designers have invented bug tracking tools and versioning tools. I learned in a bug tracking tool you can create a lot of useless noise if many people without understanding of the technical details add their comments. And since every comment has to be answered - no matter how meaningless it may be - this creates a loss of efficiency.

Nevertheless at a certain level of complexity bug tracking must be used to not overlook something. But it must be used with discipline.

- Work with consequence. Either you use a bug tracking tool or you dont. There is nothing in between.
- Before adding a new thread check if it has not yet been started by somebody else. Avoid duplicate threads.
- Don't add low brainer comments. They only create work without benefit.

Names of cells are not worth making a big discussion. You will see it if you confuse cells. Pins are different. Any pin can be connected to any net. If a block as N pins there are at least N!-1 possibilities to connect them wrong. So the first measure is:

- Keep the number of pins low.

Many software tools do not distinguish between upper case and lower case. So two pins named P1 and p1 will lead to a short.

- Don't mix upper case and lower case in names of pins or circuits.

Names that may look meaningful to one person often are not meaningfull to an other. Once upon a time I called a 3.3V logic pin pd (for power down). In the top level someone connected it to a 60V pad with inductive loads assuming pd means power dmos.

- Use signal names that indicate the supply domain and signal type

The indication of supply and type can be done as post fix such as lowside_p or prefix like p_lowside. Prefix offers the advantage that it can be sorted in a more easy way. Typical examples could be:

Digital 1.2V of domain vdd: d_vdd1v2_enable

Power 60V, 2A: p_60v_2a_lowside

Analog 3.3V: a_vdda3v3_op1in

Long names but once this convention is clear it is less likely someone connects p_60v_10a_pd to input d_vdd1v8_power_down.

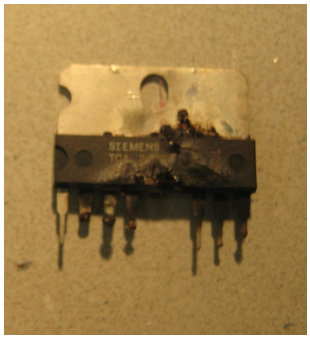


Figure 2.5: something went wrong with the signals connected

- avoid names that might be reserved words such as “input” or “output”.

Reserved names may lead to unexpected surprises with the CAD software!

- Check bus interfaces for correct width and position of the MSB and LSB

Clever digital designers may have written tools that automatically connect buses even if the number of bits differ by automatically using the first bits and deleting the rest in the code. What if you need trim<9:2> but the clever tool has automatically connected trim<0:7> and removed trim<9:8> to correct the bus width? Since digital simulation has no concept of analog trimming this is likely not to be detected in the verification!

- Can the logic turn off the clock oscillator?

Using synchronous logic this can lead to a dead lock. You need a wake up pin that is not synchronized to get out of the dead lock again. Even better: The logic must be prevented from doing suicide turning off the oscillator.

- Do you have low power modes?

Changing from low power mode to an active mode means a sudden increase of the current consumed by the logic. Without special precaution your regulators will respond too slowly and you end up in a reset or a brown out.

- Check that coming out of a low power mode the memories on the chip can not write against each other

If you have several memories (for instance a RAM and a ROM) it must be guaranteed that they never can write on the bus simultaneously. This is not just an X in the digital simulation. It means excessive current consumption pulling down the regulators. This can lead to a power on reset.

- Isolate the logic from the analog part during scan test

Analog functions are magnitudes slower then scan test patterns. I doesn't make sense to have access to the logic via analog functions during scan test. The testers don't support a mix of scan and analog anyway. Keeping analog and digital connected during scan test only is one more source of unexpected events.

- Analog test buses are extremely dangerous if they can manipulate reference voltages.

Increasing the reference of a voltage regulator during testing can destroy the chip or lead to damages that lead to field returns after some hundred operating hours. Add test buffers between the inside references and the test output to protect the regulators from reference changes.

- Where is the ground reference?
- Where will the edge seal be connected?
- What is the possible bounce of the substrate relative to ground inside the chip?

This sounds trivial, but it is not! Ground inside the chip has little in common with the system ground outside. There are bond wires and pins between the system ground and the ground pad of the chip. This means a complex impedance. Furthermore ground traces on the chip have resistance. Depending on current flowing in the ground system ground inside the chip can significantly deviate from system ground. The same applies to supply pins.

In case of using an exposed dice pad substrate is tightly connected to board ground while the circuit ground is detached from board ground by the inductance of the bond wire.

- global signals (vdd!, vss!) and hierarchical connectors are convenient for digital design. For analog design they are killing because you wont recognize supply drops and ground drops anymore.

Avoid using globals and hierarchical connectors in the analog part for better trace ability.

2.2.3 Timings

Digital design expects events to take place synchronously with a clock. Delays of analog signals have a spread that usually is in the range of many clock cycles. During system design this usually is not considered at all. During implementation most digital models only assume a typical delay. It is a good idea to verify the complete chip with corner models of the analog functions having different (worst case) timings.

Is the system able to tolerate that a state exactly changes at the clock edge of the digital part? (This is a setup time violation and leads to an undefined state of the logic for at least one clock cycle!) Is every logic input double buffered? In some million events there surely is a case where the state change exactly hits the clock edge!

Classical example: The SPI is running with a different clock than the logic. Data coming from the SPI shift register must be synchronized to the internal clock in an extremely reliable way. Loosing one out of a million bits means loosing 4 telegrams per second!

2.3 Top level

Puh, what a question. What is so difficult about the top level. The view! Which kind of presentation do you want? Modern CAD systems allow you an unbelievable variety of views. And depending on what you want to do with it you may want to see completely different things.

- The board layouter will want a footprint showing him where to place the soldering pads.
- The designer of the system wants a drawing of mechanical outlines.
- The circuit designer wants a symbol, a schematic, an netlist (which language?).
- The software engineer wants a header file and a C-model.
- The chip layouter wants a layout to plug in his cells.
- The EMC engineer wants RF-models.
- They all want a specification..

So a top level can easily look like this:

- Library name
  - chip name
    * ams (for mixed mode simulation)
    * config (to tell the AMS-netlister what view to use)
    * doc (documentation view – whatever is inside)
    * extracted (extracted layout for LVS)
    * footprint (for the board layouter)
    * gds (graphical design station file)
    * gerber (for board layout systems)
    * ghdl (gnu hardware description language)
    * gnucap (netlist for gnucap simulator)
    * layout (for the layout engineer)
    * maestro (setup file for the maestro simulation environment)
    * mechanical (package outline for the system designer)
    * netlist (for simulation)
    * saber (saber is a simulator used in automotive)
    * schematic (for the analog chip designer)
    * schematic_ac (for EMC simulation)
    * schematic_noise (for EMC simulation)
    * spectre (netlist for spectre simulator)
    * spice (netlist for spice simulator)
    * state (setup for the spectre simulation environment)
    * symbol (to plug the design into a test bench)
    * verilog (for digital simulation)
    * verilog_a (for mixed signal simulation)
    * vhdl (for digital simulation)
    * vhdl_a (for mixed mode simulation)

Well, fortunately chip designers are lazy. The never fill all the views unless they are forced to do it. So in a typical sparse library you only find:

- Library name
  - chip name config (if you are lucky)
    * layout (usually there should be one)
    * schematic (sometimes you don't even find this!)
    * symbol (if you are lucky)
    * verilog (sometimes)
    * verilog_a (rare) vhdl (Europe only)
    * vhdl_a (very rare)

Of course the lists are not complete. There are hundreds of further tools you can create a view for. Usually under the view you will find a directory holding the associated files.

(Now I'm sorry we hit the revision management again. check in and check out procedures don't handle single files well. Usually the routines are fine for often used views. If you have something exotic they fail! If you have trouble checking in and out search for a file called data.dm. This must be checked in and out as well to mke design sync work)

2.4 Hierarchy of the chip

For the analog circuit designer usually the views layout, schematic and symbol are the most interesting ones.

The schematic holds the circuit including all subcircuits. Every subcircuit is a symbol sitting in the top level circuit. If you dive into a symbol there should be an other schematic again. At least in most of the cases. (In RF design it is common practice to only have a layout and simulate using a netlist derived from some kind of layout extraction! So there may be cases you don't find schematics below the symbols.)

2.4.1 Analog on top

Classically on the top level of the chip you should find a pad ring holding everything that is in direct contact with pads (ESD protections, power transistors, I/O structures). The analog top level holding all functions that can be simulated analog (Amplifiers, bandgaps, ADCs, DACs ...) and the digital top level holding all the logic that can be synthesized.

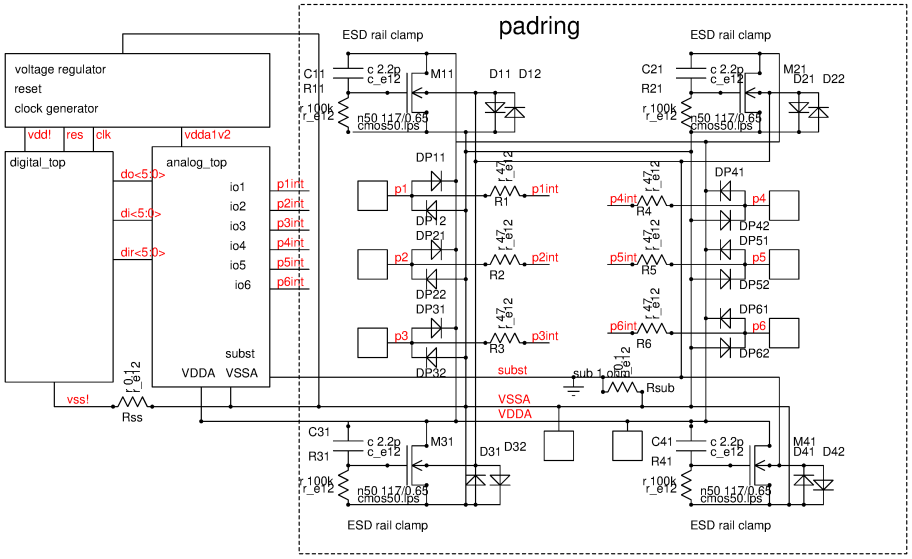


Figure 2.6: a simple example of a mainly digital chip

In the example shown above we have a digital part called digital_top. The digital design usually assumes supplies of logic gates to be connected to global signals vdd! and vss!.The cell digital_top drives the I/O cells sitting inside analog_top.

In most cases the supply voltage of the logic is in the range of 700mV to 2.5V. The block sitting on top is the bias block. It contains a voltage regulator, reset generator and the clock generator.

The I/O cells must be considered as analog because the most important properties of the I/O cells are not logical functions but analog parameters such as voltage swing, current limitation, output impedance... The cell analog_top is connected to the pad ring via signals p1int to p6int. The resistors R1 to R6 act as secondary ESD protection.

In the pad ring every signal is limited in voltage by diodes to VDDA and VSSA. Since ESD can occur in situations when the supply VDDA is not connected there must be a rail clamp between VDDA and VSSA. In our case these are 4 grounded gate NMOS transistors sitting in the corners of the chip. These transistors turn on whenever there is a rapid increase of the voltage between VDDA and VSSA. Very often these ESD rail clamps are placed directly in the substrate (acting as the bulk of the transistors). To prevent excessive voltage between the substrate and VSSA there usually are local protection diodes close to each of the rail clamp transistors.

Depending on application substrate can have a dedicated pad or can be connected to VSSA by a metal resistor (Rsub).

Real chips of course can be significantly more complex. This simple example only is intended to serve as a demonstration of the concept of analog on top design style separating protections (inside the pad ring), analog functions (inside analog_top) and digital functions (inside digital_top).

2.5 Down at the bottom

If you dive deep enough into the hierarchy you will find basic components like transistors or resistors. These basic components may still have a layout (probably scaleable because these usually are p-cells) but have no more schematic. In stead the lowest simulatable level you will find are views like netlist, au_sch, spectre, spice... For the netlister this means we are down to the bottom. Take the textual description found and don't try to further descend the hierarchy. Most netlisters are controlled by so called stop-lists. If the netlister finds a view that is listed in the stop list it will not try to descend deeper into the hierarchy.

2.5.1 Electric

Electric [1] expects an ASCI description of the cells. (symbols are called artwork in electric). Here comes an example:

```
# Cell r_0r01;1{ic}
Cr_0r01;1{ic}||artwork|1357499386323|1357763805695|E|ATTR_SPICE_template(D5G0.5;NTX7;Y3;)S
(node_name) $(plus) $(minus) 0.01
Ngeneric:Facet-Center|art@0||0|0||||AV
NOpened-Polygon|art@1||0|0|4|10|||SCHEM_function(D5G2;)Sr_10m r_e12|trace()V[-2/-5,-2/5,2/
Nschematic:Bus_Pin|pin@0||0|-5||||
Nschematic:Wire_Pin|pin@1||0|-4||||
Nschematic:Bus_Pin|pin@2||0|5||||
Nschematic:Wire_Pin|pin@3||0|4||||
NPin|pin@4||0|5|1|1||
NPin|pin@5||0|3|1|1||
NPin|pin@11||0|-3|1|1||
NPin|pin@12||0|-5|1|1||
NPin|pin@13||-1|3|1|1||
NPin|pin@14||-1|-3|1|1||
NPin|pin@15||1|-3|1|1||
NPin|pin@16||1|3|1|1||
Aschematic:wire|net@0|||900|pin@1||0|-4|pin@0||0|-5
Aschematic:wire|net@1|||2700|pin@3||0|4|pin@2||0|5
ASolid|net@2|||FS900|pin@4||0|5|pin@5||0|3
ASolid|net@9|||FS900|pin@11||0|-3|pin@12||0|-5
ASolid|net@10|||FS0|pin@5||0|3|pin@13||-1|3
ASolid|net@11|||FS900|pin@13||-1|3|pin@14||-1|-3
ASolid|net@12|||FS1800|pin@14||-1|-3|pin@15||1|-3
ASolid|net@13|||FS2700|pin@15||1|-3|pin@16||1|3
ASolid|net@14|||FS0|pin@16||1|3|pin@5||0|3
Eminus||D5G2;|pin@0||B
Eplus||D5G2;|pin@2||B
X
```

The netlister is called in the second line of the symbol code. The pin definition is in the last three lines of the code. (Well, don't ask me what all these other lines do. But I am not a software engineer. May be they know)

Electric displays the call of the spice netlister in the symbol.

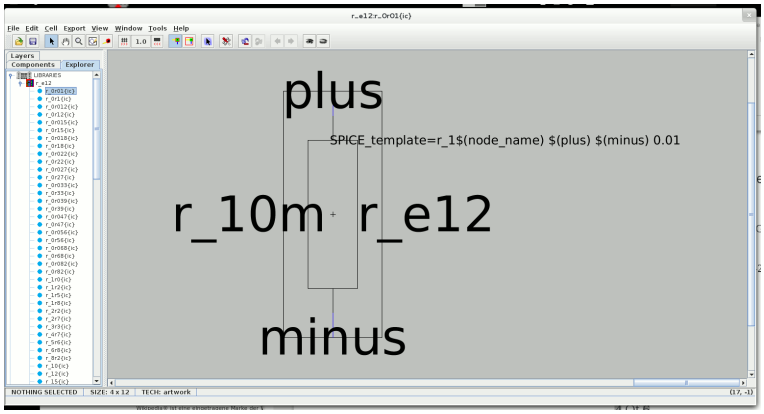


Figure 2.7: Example of a resistor in the electric CAD system

2.5.2 Xcircuit

Xcircuit [2]looks for spice code inside the symbol. Here comes an example:

```
/c10p {
begingate
1 1.00 -96 -16 0 -16 2 polygon
1 1.00 -48 -16 -48 -64 2 polygon
1 1.00 -96 16 0 16 2 polygon
1 1.00 -48 16 -48 64 2 polygon
1.00 0 -48 -64 dot 1.00 0 -48 64 dot
1.000 0.000 0.000 scb
(top) {/Helvetica cf} 2 16 0 1.00 -48 64 pinlabel
(bottom) {/Helvetica cf} 2 16 0 1.00 -48 -64 pinlabel
sce
(c 10p) {/Helvetica cf} 2 16 0 1.00 16 0 label
(c_e12) {/Helvetica cf} 2 16 0 1.00 16 -32 label
0.180 0.545 0.341 scb
(spice:C%i %ptop %pbottom 10p) {/Times-Roman cf} 2 4 0 1.00 -244 -139 infolabel
(sim:n %ptop %pbottom) {/Times-Roman cf} 2 4 0 1.00 -244 -187 infolabel endgate
} def
```

In stead of following a stop list xcircuit simply pastes the line after the marker '(spice:' into the spice netlist. It does some interpretation: %i: I is a counter variable. It is incremented each time xcircuit finds a new instance. %p: this is a pin in the spice netlist xcircuit is less powerful than the electric design system, but much easier to understand and to maintain the libraries.

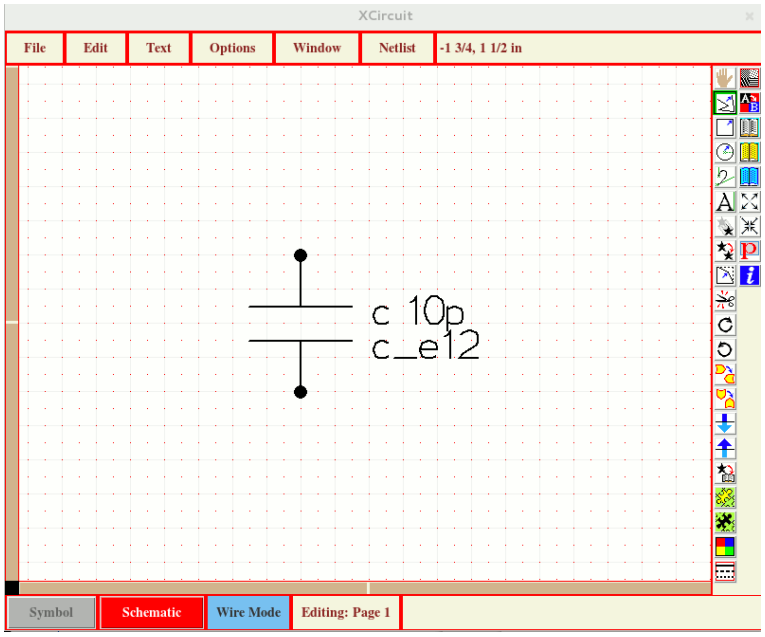


Figure 2.8: Capacitor symbol in xcircuit

Xcircuit does not show the call of the netlister in the symbol. You have to look at the code to see it.

3 From wafer to chip

Chip production always starts from a plain wafer. A wafer is a disk of mono crystalline silicon. In the beginning it does not hold any structures. Usually it comes with an initial doping. It can be p-doped on n-doped.

The level of doping can be very low leading to a resistivity of the material in the range of 4 Ωcm to 20 Ωcm. Usually these low doped wafers are used for technologies using implantation, but no epitaxy.

High resistive substrate requires restrictive design rules for substrate contacts to prevent latch up. It offers the advantage that you can create a very cheap process plugging the transistors right into the substrate.

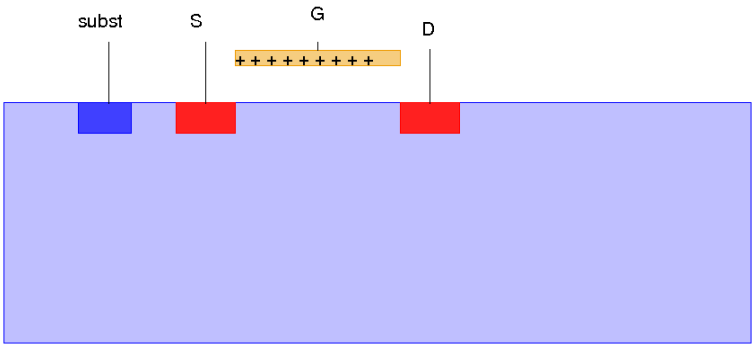


Figure 3.1: A very simple process with NMOS transistor sitting in the substrate

For technologies using epitaxy highly doped wafers are frequently used. The high doped wafers offer a lower substrate resistivity in the range of 5mΩcm to 20mΩcm. Low resistive substrate requires less restrictive latch up rules. In extreme cases having a backside contact and an edge seal only if fully sufficient (Using high doped substrate you may find technologies that do not require any substrate contacts in the middle of the chip)

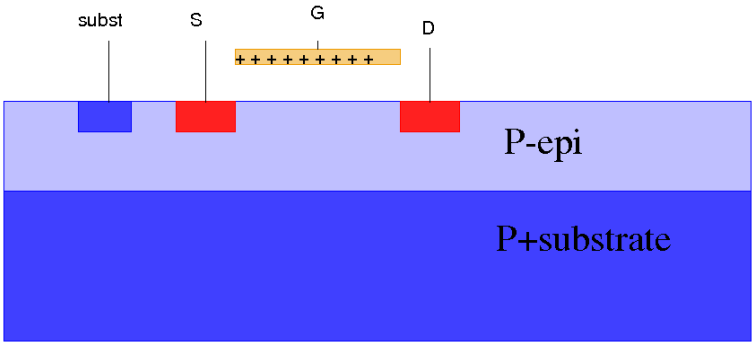


Figure 3.2: NMOS transistor in a technology using P+ substrate and P- epitaxy

In some cases the epitaxy is doped inverse to the substrate. This is mainly used for bipolar and for power technologies.

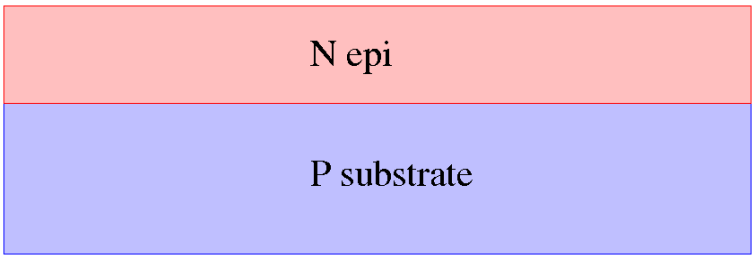


Figure 3.3: First step of creating a N-epitaxy

Here the transistors must be separated after epitaxy adding a deep diffusion of P-doping material after epitaxy .This step is called isolation.

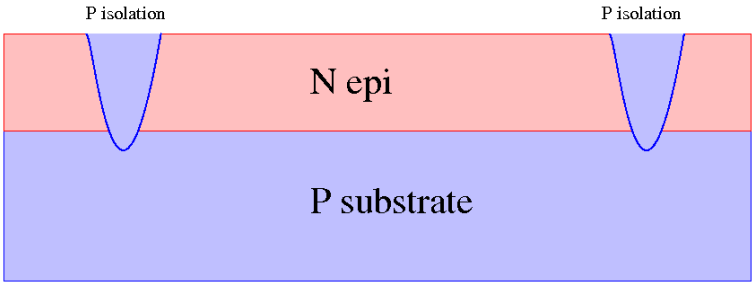


Figure 3.4: separating the epi pockets by P-isolation

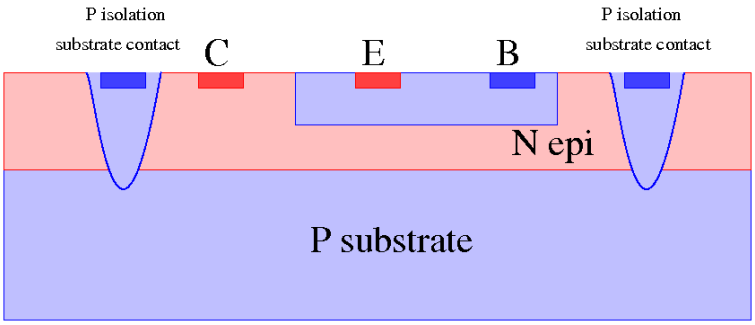


Figure 3.5: Creation of a component inside the N-epi pocket using N-epi as one node of the component.

The polarity of the wafer doping depends on the application. In most cases the back side of the chips is connected to ground and P-doping is preferred (about 90% of the market). In some cases the back side of the chip is intended as an active node (usually drain of a power transistor or collector of a power transistor). In these cases N-doping is preferred. (frequently used for high side switches). Here is an example of a darlington transistor.

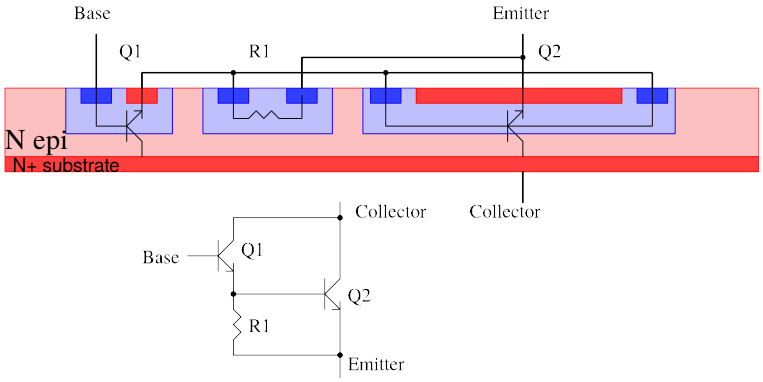


Figure 3.6: NPN darlington transistor

Et voila, we have designed the most simple IC. It holds two transistors sharing the collector diffusion (the N-epi and the N+ bottom contact) and a resistor using the base doping with two contacts. There is one little trick in it: At the end of the process the N+ substrate layer usually is lapped as thin as possible. This reduces the thermal resistance from Q2 to the bottom, that usually is soldered to a massive metal plate acting as a heat sink.

Besides using a junction isolation between substrate and epi it is possible to create a fully isolated technology with an oxide layer between the substrate (now usually called handler wafer) and the circuit components (SOI, silicon on isolation). This kind of technology is more expensive than using junction isolation but it opens additional possibilities to protect the circuit against high voltage and reverse polarity situations (Junctions used as an isolation act as diodes. In normal operation they are in blocked state. At reverse polarity the diodes become conducting. This may lead to a destruction of the chip)

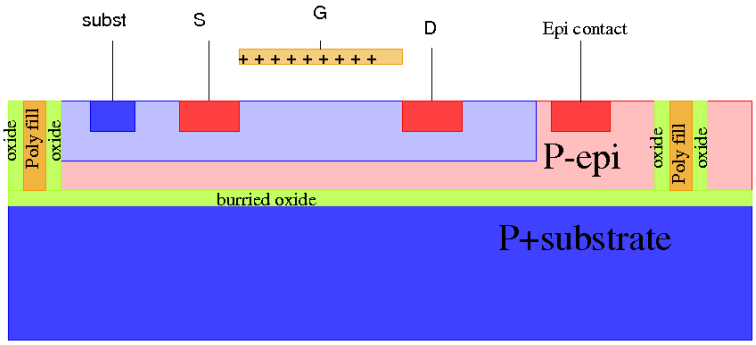


Figure 3.7: Oxide isolated process (SOI)

3.1 Choice of the right substrate

The choice of the right substrate for a product is one of the most important decisions of a project. This decision has an impact on:

- Process cost
- Thermal performance
- ESD performance
- EMI performance
- Choice of circuit topologies
- Electrical design parameters like PSRR

3.1.1 Junction isolated P- substrate

Junction isolated P- substrate is a reasonable choice for standard CMOS logic ICs. The NMOS transistors can directly be embedded in the substrate saving masks. The NMOS transistors are sitting in an nwell.

In this low cost variant the bulks of all NMOS transistors are connected to substrate.

- Process cost is very low due to the low number of masks required.
- Thermal performance is like most other junction isolated process variants.

- EMI performance usually is poor because substrate noise modulates all signals coming from NMOS transistors. (Back gate is directly connected to the substrate and noise present in the substrate. The high resistivity of the substrate makes these designs prone to latch up problems. Latch up [7] can be triggered by transients, ESD during operation or RF injected into the pins.
- The choice of circuit topologies is very limited. Building NMOS differential amplifiers with the bulk connected to the source is not possible. Offset of NMOS amplifier stages is affected by the backgate matching.
- PSRR of analog amplifiers can be designed moderately well if the design as well as the layout is done in an clean way. Asymmetries of substrate contacts can significantly influence the PSRR (power supply rejection ratio) and the CMRR (common mode rejection ratio) of analog amplifier stages.

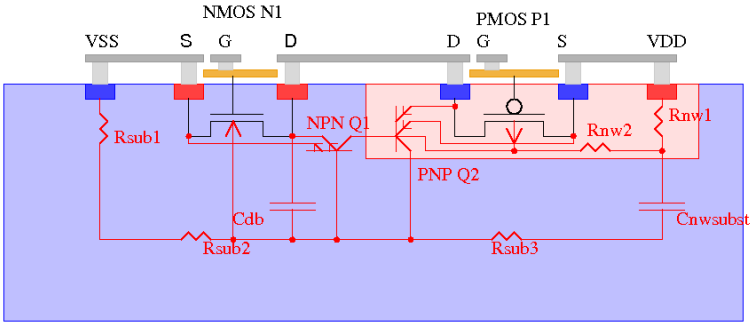


Figure 3.8: A simple CMOS process using P- substrate

This most simple CMOS process uses only the following 7 masks: active (defines areas with gate oxide), P+ (for PMOS and NMOS bulk contact), N+ (for NMOS and PMOS bulk contact), NWELL (for PMOS bulk), Poly silicon (for gates), contact, metal. Probably this is the cheapest possible CMOS process we can build, but also the least performant one.

The most important parasitic components are drawn in red. In later chapters we will see these parasitic components in more detail. For comparing different process variants it is important to consider that Rsubs2 and Rnw2 can be quite high using P- substrate and an nwell without buried layer.

3.1.2 Junction Isolated P+ substrate

This process uses a P+ substrate to reduce the risk of latch up. Since you can't implant a low doped well into a P+ substrate there is an additional production step called epitaxy (most chip designers simply abbreviate this as epi).

epitaxy is a fairly time consuming production step. So P+ substrate with P- epitaxy is more expensive than using P- substrate. This kind of process typically is used for medium price products as well as high speed logic.

- Process cost still is low
- Thermal performance like other standard CMOS processes
- EMI performance can be improved significantly if exposed die pads are used. Latch up risk is much lower than using P- substrate
- Choice of circuit topologies still is limited because NMOS bulks are tied to a common ground node (the substrate)
- PSRR and CMRR can be very good if clever design is used
- CMRR and PSRR is less sensitive to asymmetries of substrate contacts than using P- substrate

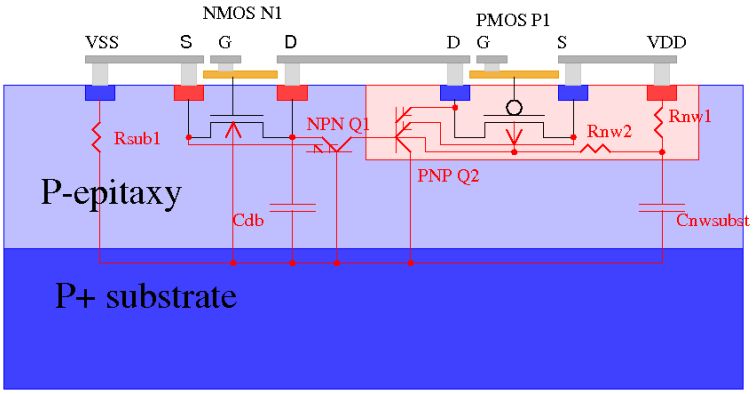


Figure 3.9: Epi technology on P+ substrate

The P+ material is typically 3 magnitudes less resistive than P- material. Compared to the P- substrate technology the resistors Rsub2 and Rsub3 are eliminated. The remaining vertical resistor through the epitaxy layer can be made low resistive making the epitaxy as thin as possible. Typical thickness of the epitaxy layer is in the range of 2um to 15um depending on the voltage requirements. The P+ wafer usually is about 200um to 400um thick. It can be regarded like a metal plate. Typical resistivity of the substrate is in the range of 3..15 mili Ohm cm. The epitaxy usually has 3..15 Ohm cm.

3.1.3 Junction isolated N- substrate

Using N- substrate in stead of a P- substrate is unusual because it connects most of the parasitic components to the positive supply voltage (while P substrate connects the parasitic components to the negative supply node). In applications using the negative supply as a reference ground this makes the product more supply noise sensitive.

Now the PMOS transistor bulks are automatically connected to the positive supply while the NMOS bulks can be wired freely . This may offer slight advantages for analog RF amplifiers (usually NMOS transistors are slightly faster than PMOS transistors).

The latch up properties can be slightly improved compared to a P- substrate process because the electron mobility inside the N- substrate is about factor 2 to 3 better than the hole mobility inside P- substrate.

- Very low process cost
- Thermal performance like most other processes
- EMI sensitive unless the positive node is the ground reference
- Choice of circuit topologies is limited because all PMOS transistors share the same bulk potential
- PSRR is very poor
- PSRR and CMRR are sensitive to asymmetries of bulk contacts
- Latch up performance slightly better than using P- substrate

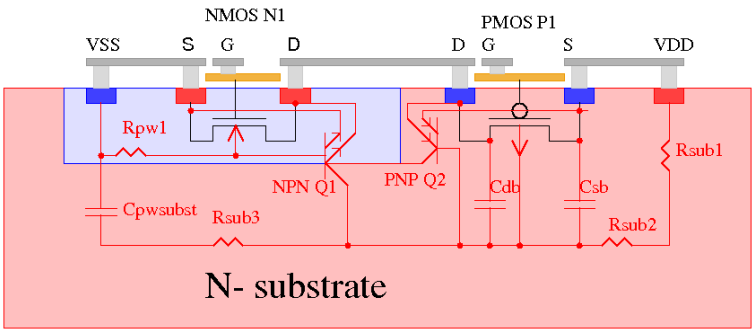


Figure 3.10: CMOS transistors inside an N- substrate

3.1.4 Junction isolated N+ substrate

Using an N+ substrate and an N- epitaxy offers a very area efficient way of building powerful high side drivers. We have already seen something similar discussing the NPN darlington transistor. In the following technology we want

to combine a vertical DMOS (diffused MOS) transistor with some CMOS circuits. The crux is to produce a voltage higher than the N+ substrate needed to drive the gate of the power transistor.

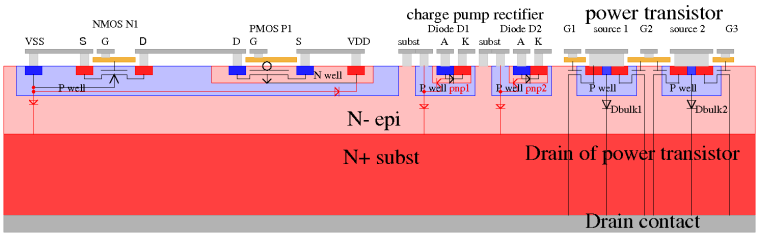


Figure 3.11: CMOS DMOS technology using N+ substrate

In this technology we added an isolated nwell that can be placed inside the pwell. This additional mask is needed to create diodes that can operate above the substrate voltage. These diodes are required for the charge pump rectifier.

As a side effect we now have NMOS transistors with isolated bulk and PMOS transistors with isolated bulk. We are no more forced to connect the bulk of the PMOS transistors to the N substrate.

We also can use the nwell as a drain extension to create a high voltage nmos. This transistor however will not have a high performance.

Furthermore we can build the following bipolar transistors:

- Vertical PNP using P+ as an emitter, the nwell as the base and the pwell as the collector.
- Vertical NPN with the collector tied to the substrate (drain of the power transistor), pwell used as base and N+ used as it's emitter.

So we can summarize the properties of an N+ substrate process (with additional nwell inside pwell) as follows:

- Ideal for building high side power drivers.
- Low process cost (only one mask more than N- substrate processes. Theoretically 9 masks are sufficient. In practice further masks to improve the performance of the power transistors are added in most cases).
- excellent thermal performance (the drain region of the power transistor is not at the surface).
- EMI, PSRR can be a problem in systems using the negative supply as ground (most stray capacities go to N-epi).
- CMRR depends on symmetrical layout of well ties.
- Latch up may be critical because nwell and pwell have high resistance.
- low voltage NMOS and PMOS have freely wire-able bulks.
- Additional bipolar components become available (diode, NPN, PNP).

3.1.5 Oxide isolated technology with N-tubs

This kind of process uses a burred oxide as a bottom isolation and trenches with oxide as side wall isolation of the components. Parasitic diodes present in junction isolation processes can be avoided. (Well, this is not quite true because the number of permissible trenches is limited. So CMOS parts still use wells and junction isolation. Only the well regions can be separated from other wells using oxide isolation).

Although all circuit activity is taking place inside a layer of only some micrometers of thickness a carrier with the same thermal expansion as the active silicon is required. This carrier is called the handler wafer.

- More expensive due to a complex manufacturing process
- Poor thermal performance because burried oxide is a heat barrier
- EMI performance typically is 1 magnitude better then junction isolated process
- PSRR usually is good if the handler wafer is connected to system ground
- good CMRR (all transistors can have freely wire-able bulks)
- Very good latch up performance due to oxide isolation (parasitic bipolars can be avoided)

There are several ways of producing a burred oxide.

The classic way splicing the wafer: The most classic way to create a burried oxide is to splice the wafer. In a monocrystalline structure cracks will easily propagate following certain directions. This allows splicing the wafer. If one of the splices is very thin (a few μm) the thin layer becomes flexible like an aluminum foil. This is a pure mechanical process. If there are defects in the crystall the thin slice will get damaged!

Figure 3.12: The first step to create a thin layer of mono crystalline silicon

After creating a thin layer of about 5..6 μm the thin slice is attached to a chuck. One side will be oxidized. The best oxide quality would be simple thermal oxide. But this process is too slow. Chemical vapor deposition (CVD) works faster. The same process is done with the thick part that from now on is called the handler wafer.

After oxidization both sides have to be planarized to a roughness of only one atom layer. Both halfs must have the same oxide lattice. At the surface of the oxide there are dangling bonds. If both oxides are planar enough the oxides will immediately attach to each other when brought in contact. (This process originally was used to attach the Brewster mirrors of He-Ne lasers. So attaching glass to glass is in use at least since the 1960s)

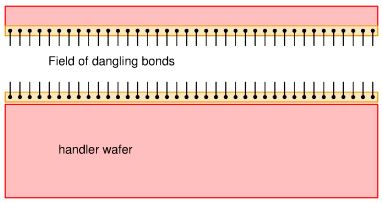


Figure 3.13: wafers to be attached have the same pattern of surface bonds of the oxide

After attachment (ideally) each open bond of one wafer connects to an atom of the other wafer.

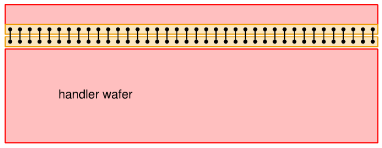


Figure 3.14: After attachment each surface atom of one oxide finds a partner in the other oxide

The achieved attachment can be as good as if the two layers are one monolithic block! Taking apart the two halfs isn't possible anymore without destroying them. Now we have a buried oxide and etching trenches allows the creation of completely isolated components.

Process for partial buried oxide: A more sophisticated method that allows building chips with some regions with buried oxide as well as regions without buried oxide is shown on the following pages. The advantage is that in regions with power transistors the buried oxide is not present. This provides a lower thermal resistance in the power areas while we still can benefit from oxide isolation in the low power areas. The process is described in literature but I have no clue if it is really in use.

In the following the most important process differences compared to a junction isolation are explained [12]. As a first step Si-Ge (silicon germanium) is grown on a silicon wafer. Afterwards silicon is grown on top of the Si-Ge. The surface of the silicon gets oxidized (dry oxide, some nm thick). At the end the oxide is covered with nitride. These steps can be done without any masks.

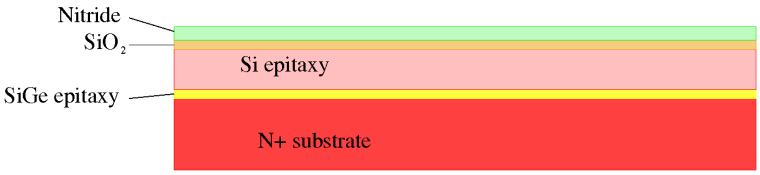


Figure 3.15: Preparation of the wafer

The next step is to separate regions of silicon epitaxy to form the isolated components. This is done etching trenches through the silicon and the Si-Ge.

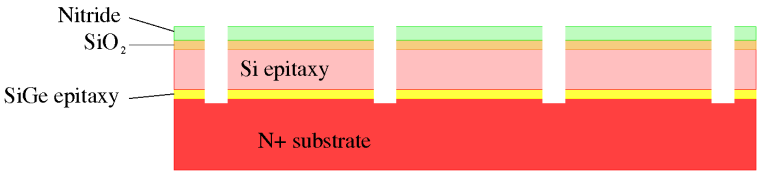


Figure 3.16: Definition of isolated regions

The trenches are filled with silicon oxide. Part of these oxide fills are covered with nitride used as a mask in later process steps.

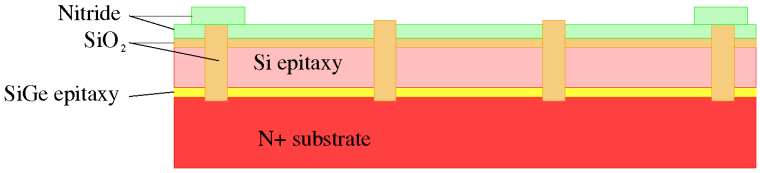


Figure 3.17: Trench fill and definition which area will be etched

Using an etch for silicon oxide the exposed trenches are reopened. The trench regions covered with nitride will not be opened again.

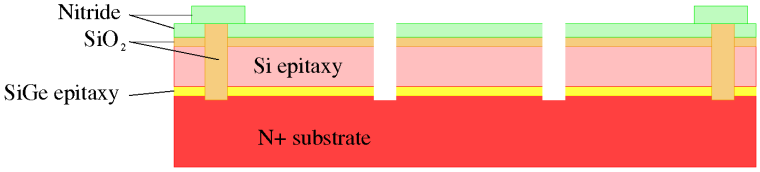


Figure 3.18: Opening the exposed trenches

Through the open trenches an etch optimized for Si-Ge can be used to horizontally remove the Si-Ge adjacent to the open trenches. The trenches that still are filled act as a mechanical support so the epitaxy will not fall down. Every epitaxy region must be held by several pillars of covered oxide (not shown in the middle)

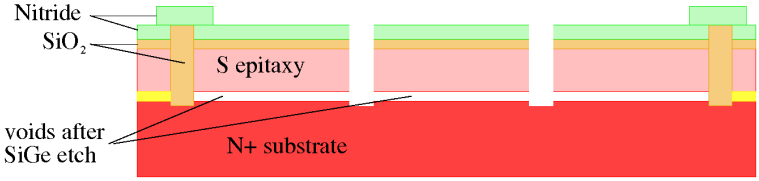


Figure 3.19: Voids after Si-Ge etch and Si epi held by pillars of oxide

The voids are filled by oxidation (The oxide has more volume than the silicon it has consumed during oxidation. So the substrate as well as the bottom side of the epitaxy will grow an oxide into the voids).

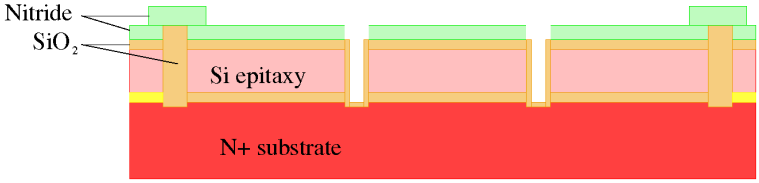


Figure 3.20: Voids closed again after oxidation

At last the trenches that are not yet fully closed by oxide must be filled again. The fill can be oxide just as well as some other material that has an appropriate thermal expansion coefficient. In the following a fill with oxide deposition is shown.

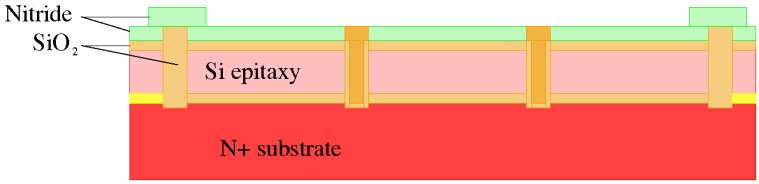


Figure 3.21: Trenches are filled again

Bottom line the oxide isolation process requires one more mask than using junction isolation (the nitride cover for the oxide pillars holding the structure in place while the voids for the buried oxide are produced). The process steps used in between are similar to the DTI (deep trench isolation) processes used in other technologies to reduce the size of high voltage components.

The process description given is simplified. Usually the Si-Ge layer is made some nm thick only to guarantee the lattice does not change from the silicon wafer to the silicon epitaxy. To create thicker buried oxide layer there usually are additional etching steps to widen the void further into the silicon. This can be enhanced doping the epitaxy right above the Si-Ge layer and using selective etches that preferably remove the doped silicon. But these are process details that go too far in this conceptual description. (For further details see [12]).

Now the covering nitride and oxide can be removed and standard process steps can be done to create all the components needed.

3.1.6 Oxide isolated technology with P-tubs

This kind of process is closer to common CMOS technologies. Therefore growing Si-Ge and P- epitaxy is a more common approach than using N-epitaxy. The process of creating the buried oxide is basically the same as described before except that the doping polarity of the silicon is changed now. The following figure shows an example of the resulting cross section after implementing the usual CMOS devices. Details like n-extensions an p-extensions or halo implants are not shown here to limit the complexity of the figure. These details will be discussed in the device section.

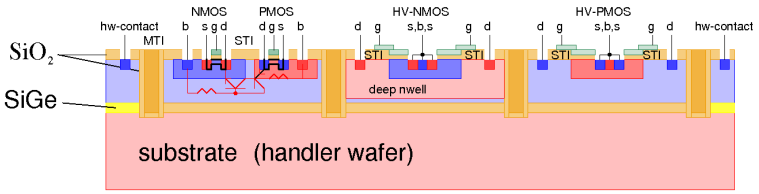


Figure 3.22: Example of a complete oxide isolated process featuring CMOS components together with high voltage MOS transistors

This kind of process offers standard CMOS logic with the same density as a junction isolated process (and the same parasitic components shown in red) together with oxide isolated high voltage components. Even if the thickness of the oxide isolation of the high voltage transistors is similar to the junction with of their counter parts in standard technologies the substrate capacities are significantly lower. The reason is that silicon oxide has a dielectric constant of 3.9 while depleted silicon has 12.

A significant draw back of oxide isolation is the thermal resistance of the silicon oxide. Silicon oxide has about 0.9W/m K to 1.2W/m K thermal conductivity while silicon has about 150W/m K (at room temperature) to about 110W/m K at 450° C. Especially for minimum high voltage transistors a thermal resistance of up to 5K/mW must be expected.

3.2 Examples of more complex processes

This chapter is intended to give an overview which features typical processes that are more rich than classical CMOS offer.

3.2.1 A bipolar process

In the beginning of IC design bipolar processes we used for analog application. Typically this kind of process offers bipolar transistors and in some cases I2L logic.

Table 1: typical parameters of bipolar transistors

component	Vce	B	Ikf	ft
NPN	12..60V	100-300	1..3mA	200MHz..1GHz
lat. PNP	12..60V	20..100	0.1mA	1..15MHz
I2L transistor	3..6V	0.3..10	0.05mA	1..30MHz
vert. PNP	12..60V	40..300	0.1..1mA	3..50MHz

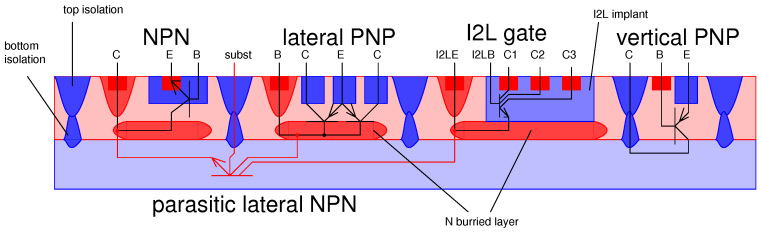


Figure 3.23: Bipolar process including I2L logic

The process offers the following components and typical range of parameters for minimum transistors:
Typical component parameters of bipolar processes

Usually the transistors tend to become faster the lower the voltage the process is optimized for.

I2L usually does not work properly anymore if the maximum process voltage becomes higher than 25V. Reason is that the higher the process voltage the higher the distance between buried layer (acting as I2L emitter) and emitter doping (acting as I2L collector). This means with increasing process voltage the base width of the I2L gates increases and the performance of the I2L logic decreases.

Bipolar process still is an option for products with very low logic contents and operating voltages higher than 5V. Below 5V there are many standard CMOS process lines that offer smaller feature size at the same or even cheaper wafer cost.

Note that there is a parasitic lateral NPN between the NPN collectors and the base regions of the PNP transistors. This parasitic transistor becomes active if one of the N-tubs is pulled below ground. All N-tubs adjacent to the N-tub pulled below ground become collectors of this parasitic transistor in this case. The lateral NPN having a big base width acts as a low pass filter. Transients in the ns range are integrated and only the total energy of the pulse matters. Longer transients in the us range and ms range propagate to adjacent tubs without being integrated.

EMC performance of bipolar ICs at high frequency usually is poor because the components are big and the capacities to substrate usually are in the range of some tenth of a pF to some pF. RF injected into an N-tub excites the substrate and capacitively couples from the bouncing substrate to other N-tubs. This effect is worsened by the fact that usually low doped substrate is used to allow high buried layer doping (low resistive buried layer). Since the buried layer is high doped the depletion area must be in the substrate forcing manufacturers to use low doped high resistive substrate.

3.2.2 A BCD (bipolar, CMOS, DMOS) power process

In the 1990s many manufacturers introduced process lines offering bipolar components, CMOS logic and DMOS transistors (diffused MOS) and poly silicon resistors. In the beginning these process lines were based on a bipolar base process using N-epitaxy. (ST Microelectronic BCD1 to BCD4, Siemens SPT1 to SPT5). When CMOS became the main stream manufacturers started to use P-epitaxy to take advantage of cost scaling of preproduced standard wafers. (ST Microelectronics BCD6 and up, Infineon SPT7 and higher, Texas instruments LBC lines and Freescale's smartmos lines.) As a consequence of this change bipolar transistors more and more became a secondary component. This makes the design of classical bipolar blocks like band gaps more and more difficult. Part of this analog performance loss can be compensated by better trimming means (all kind of memory cells coming with CMOS technologies.) The drawback of compensating the loss of bipolar performance by trimming is a significantly more complex analog testing strategy. (During test trimming must be emulated before writing the best trim values into non volatile memories.) As a pro carrying more and more know how into test procedures makes copying chips more difficult. (Reverse engineering of the silicon alone does not yet allow manufacturing of the hijacked product.)

Changing to more CMOS based process flows P+ substrate is becoming more common to take benefit of the better latch up ruggedness.

The introduction of P+ substrate makes the use of N-buried layers more complex and expensive. (You can not place the buried layer right on top of a P+substrate. This would lead to low break down voltages.) In stead the epitaxy must either be grown in two steps and the buried layer must be produced in between these two epitaxy steps or the buried layer must be produced by a deep implant with very high doping doses. Alternatively circuits using lateral components only without any buried layer have to be used. (For instance the rectifier diodes of power stages

driving inductive loads become very lossy and must be replaced by active rectifiers. The same applies to rectifiers in charge pumps.) Mixing components with buried layer and without buried layer is not a good idea because the atoms doping the buried layer add to the volume of the material. If the buried layer can be masked this means the surface of the chip will no longer be planar. Violations of planarity will move some of the structures out of the optical focus in the following production steps.

Modern BCD process technologies often use trenches filled with oxide (DTI: deep trench isolation) for lateral isolation of the components. The oxide can withstand higher electric fields. So the trenches can be made more narrow than P-isolation regions between the components.

For lithographic reasons (better optical focus needed for narrow metals and contacts) a planarized surfaces is preferred. This STI process step mainly helps to reduce the wiring pitch in the CMOS logic and reduces the junction capacity of small CMOS transistors. So the logic becomes smaller and faster.

For the power transistors this means: In stead of the classical field oxide the drift regions of the high voltage components will use STI (shallow trench isolation) that was planarized by a CMP (chemical mechanical polishing). (The oxide grows from the surface down into the silicon as well as upward because every atom of silicon is accompanied by two atoms of oxygen. Oxidizing means an increase of material. During CMP the oxide having created a hill gets removed.) CMP is a very critical process step because particles in the polish can create scratches in the STI. Therefore modern high voltage MOS designs require special drain leakage test modes.

The limitations accepted to make the process more planar drive circuit complexity to higher levels and increases the design effort significantly. As a consequence the number of chips to be produced to reach economical break even is increasing dramatically using such a modern process. For products not requiring a high performance of the logic a more old fashioned process could be a more economical choice.

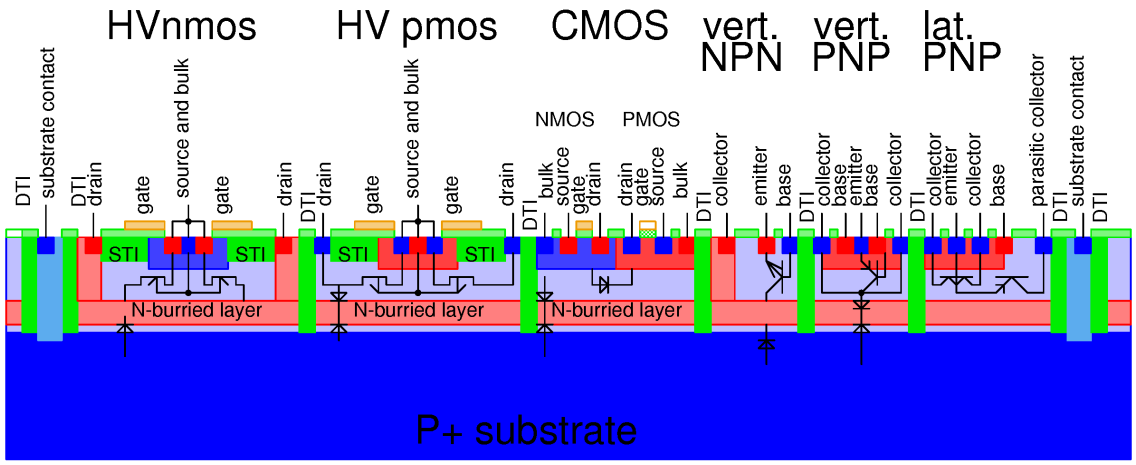


Figure 3.24: A full blown BCD process

The process shown in the figure above is based on a P+ substrate with P-epitaxy. The buried layer usually is done with N-doping before the P- is grown. In some technologies it is allowed to leave the buried layer floating. In others it is mandatory to connect the buried layer to a voltage higher than the voltage of the P-epitaxy to prevent multiplication of leakage current with the gain of the vertical PNP transistor (collector=substrate, base=buried layer, emitter=P-epitaxy)

3.2.3 An isolated (ABCD - advanced BCD) power process

This kind of process typically offers low voltage CMOS components together with high voltage power MOS transistors and bipolar transistors. Poly silicon can be manufactured with salicide (metal ions doping the poly silicon leading to about $1\Omega/\#$) and unsalicided doped poly silicon (Typically in this case we get $100\Omega/\#$to about $2000\Omega/\#$depending on the doping.)

Different from junction isolated process diffused resistors can be built fully isolated.

Since every component can be built fully isolated designing circuits operating beyond the supply rails is fairly easy (remember, in a junction isolated technology the parasitic diodes are limiting the use of most components beyond the supply rail voltages). This offers interesting opportunities to build ESD protection with the same behavior for both pulse polarities.

State of the art ABCD process (ABCD9, smartmos10) have to deal with the same limitations caused by planarization requirements (STI testing, certain diffusions may not be masked anymore) as modern standard BCD process families (bcd9, spt9).

3.2.4 A non volatile memory process

Non volatile memories use hot carriers or Fowler Nordheim tunneling to program the memory cells. This requires extremely tight control of the oxides between the two gates involved and the oxide at the edge of the drain (where the hot carriers during programming get injected into the floating gate). The complete process is optimized for the memory. All other process features are of secondary importance. This has a severe impact on analog design: If the memory doesn't perform, the process will be tweaked until the memory does work. The analog design simply has to follow the process change. If such a tweak happens the analog part of the design will very likely require a redesign. For this reason analog reuse cells often have to be reworked implementing them in an NVM process.

3.2.5 State of the art (2020) digital processes

Modern processes optimized for digital designs mainly focus on the transistor size. The current flow inside the transistor is extremely inhomogenous. As a consequence such a process is almost unusable for any analog application.

Usually the most important parameter is the minimum channel length of the transistors. However the smaller the transistors get the more area is used for contacts, wires, protections and guard rings while the percentage of the active area decreases. In 2020 numbers published about TSMC and Intel processes show that the component density is not only dependant on the pure channel length:

Table 2: Typical logic densities in 2020

manufacturer	node	transistors per mm^2
TSMC	7nm	$96 * 10^9$
Intel	10nm	$100 * 10^9$

Thes numbers show that the Intel 10nm process is better optimized regarding interconnects than the TSMC 7nm process.

One more important parameter is the average switching loss per gate. This depends on the capacity of the transistors plus the capacity of the interconnects. A process using bigger transistors can almost reach the same performance as a process using smaller transistors if the inteconnects have less stray capacity (thicker oxides, low ε dielectrics for the interconnects). Many publications only state power dissipation at a certain frequency per CPU core. In the following tables columns 3 and 4 show published data. The statndardized value for 1GHz is calculated in column no. 5:

Table 3: power dissipation per area if logic technologies in 2020

manufacturer	node	Power/core	fclk	calculated for 1GHz
Intel	14nm	8.75W	3.1GHz	2.8226 W/GHz
Intel	10nm	7W	2.3GHz	3.0435 W/GHz
TSMC	7nm	6.5W	3.5GHz	1.8571 W/GHz

The result isn't too surprising. The Intel 14nm process is optimized for servers. The 10nm process is optimized for low cost laptop applications. Here transistor density is more important (for marketing reasons) than power efficiency. The TSMC process is optimized for cellular phones, where efficiency is the key parameter.

3.2.6 Programmable chips

Chips can be programmed in various ways. This does not necessarily mean you have to use an NVM process. Using an NVM (non volatile memory) is only one of the options. The way of programming is part of the technology qualification. This makes programability an important part of the technology choice.

Laser adjustable resistors: Cutting resistors with a laser during wafer sort is a nice way of tuning analog parameters directly. This works nicely with SiCr resistors that have enough silicon oxide underneath because the cut may not unintentionally open a path for contamination down to the silicon-silicon oxide interface (If the cut unintentionally reaches the silicon sodium contamination can creep into the chip!)

Cutting resistors has a severe draw back. At the location we want to cut the passivation must be opened to let the vaporized material escape. This leaves the adjuster resistor vulnerable to corrosive chemicals. This is a clear reliability issue.

Laser cutting and fusing: Instead of cutting a trench into a resistor the connection can be cut open completely. This means we have two states: closed or cut open. This approach is often used in analog designs as long as only a few bit to maximum 128bits are to be programmed. Cutting poly silicon bridges is the most frequently used method.

Since the poly silicon usually only has a shallow layer of oxide underneath the chance of opening the oxide until down to the bare silicon is high. To prevent contamination the structures to be cut MUST be surrounded with an edge seal like structure. This protection structure is significantly bigger than the trim structure itself.

A second issue is the growth of whiskers. If the structure is cut open reading the structure creates an electric field over the cut open gap. At the same time some of the vaporized material has settled close to the gap (just outside of the focus of the laser). The electrical field will attract this condensated material and whiskers will grow and after a certain time close the gap again. The programming gets lost. The time needed to close the gap at a certain electrical field is a matter of time (some hundred to some thousand hours) and temperature. To minimize the time the gap is exposed to an electric field each laser fuse may only be read with a short pulse and copied into a shadow register. This can either be done once at reset or periodic (for instance read pulse $10\mu s$,repetition rate 0.1Hz). The periodic read out is preferred because if the register loses data (for instance due to EMC) the system will recover after one read period.

Summarizing a LASER cut structure always requires additional infrastructure:

- pulse generator
- read amplifier
- shadow register
- edge seal

The sweet spot for LASER cutting is about 10..100 bit.

Memresistor: Heating up an poly silicon resistor to the melting temperature of the grains changes it's resistivity by merging adjacent grains (so some boundaries acting as barriers for the current flow are getting removed) . Exposing poly silicon resistors to high power pulses can be used to reduce the resistance. This is called a Memresistor. This kind of device was proposed around 2003 but I never have seen it in real production.

Reliability concern are, that the grain boundaries could get reestablished from small seeds left over after a certain operating time of the chip.

Metal fuses: Metal fuses are thin wires that can be fused with a current pulse. The metal vaporizes. To let the vapor escape the passivation must be opened on top of the metal fuse. This leads to the same requirements as for LASER fuses. (edge seal). Like a LASER fuse a metal fuse can close again due to whisker creation. For this reason a metal fuse must be read in pulsed mode too.

The additional infrastructure for metal fuses is:

- pulse generator
- power switch (for fusing)
- read amplifier
- shadow register
- edge seal

The sweet spot for metal fuses is about 10bit to 100bit.

Zener Zap: A zener zap structure is a zener diode or an NPN transistor (operated as a zener diode) covered with excessive metal over the N+ contact. As long as the voltage applied is lower than the zener voltage it can be regarded as an open.

To short the structure a high current pulse is required that melts the metal over the N+ region. The positive charged aluminum gets soaked into the junction. This creates a metal filament connecting the two nodes (in case of the preferred NPN transistor from the N+ emitter through the base to the N-collector).

Currents needed to melt the aluminum are in the range of 0.5A to 1A at 20V for up to 2ms (DOPL structure of 1986).

Zener zap structure won't work in technologies using tungsten plugs as contacts because the tungsten doesn't melt!

Zener zapping is mainly used in pure bipolar technologies. The zap structure is fairly big (about $5000\mu m^2$per bit). In addition we need a pad to apply the pulse (about $10000\mu m^2$per bit). So we end up at about $20000\mu m^2$per bit.

The sweet spot for zener zapping is 1..10 bit.

Non volatile memories: Non volatile memories use MOS transistors with a floating gate acting as a storage element. Usually as soon as a technology is offering a NVM the whole technology is optimized for this memory and analog parameters have to follow the memory requirements. This may mean we have to sacrifice analog performance only to get the NVM feature.

Non volatile memories have a certain data retention time. Approaching the (temperature dependent!) data retention time we have to expect data loss. This is more or less a statistical effect. The more bits we have the higher the chance of encountering a bit flip. Data loss increases exponentially with temperature. For this reason non volatile memories sometimes only allow a reduced temperature range. (Check the technology specification for disclaimers as soon as you use the NVM.)

The memory itself usually comes as part of the technology library. Usually these are fixed cells including all the additional infrastructure (memory array, write map, read amp, reference generators, charge pump, error detection and possibly error correction, shadow register, BIST such as marginal read circuits etc.)

For memories without error correction the sweet spot is about 1000bit to 10000bit.

For memories with error correction the sweet spot is above 10000bit

Conclusion regarding trimming: Trimming doesn't come for free. Each trimming method has it's specific risks. As long as you can reach your target without trimming do it that way. Use trimming only as a last resort. Limit your trim bits to the lowest possible number.

Today some customers want to have traceability of each single chip in the field. This means each chip has an individual serial number written into trim bits. Compared to parameter trimming this is a low hanging fruit because usually only very few of these chips are read back (usually only the field returns).

Having a chip code stored in some bits does not necessarily mean you have trim bits reliable enough for parameters. Better do your design right as far as possible rather than fixing it in production with trim bits.

4 Components

4.1 Wires

Wires are a component? Yes, they are! They have physical properties such as resistance, inductance, capacity and current capability. So we have to regard them as components.

Until about 2005 aluminum wires and wires made of aluminum alloys (AlSi, AlSiCu with up to 5% copper) were the standard solution found in the chips. Increasing current densities mainly in the supply rails lead to the introduction of copper wires.

Aluminum is well compatible with silicon technologies. Aluminum diffusing into the silicon surface at the contacts is not causing severe process problems. Etching aluminum or aluminum alloys is simple using photo-resist masks and phosphorous acid.

Copper is much more sophisticated. Copper may not get directly in touch with the silicon underneath. So the silicon must be connected to the copper traces with an interfacing material preventing the copper from diffusing into the silicon (copper dramatically changes the recombination of minority carriers. This changes the parameters of almost every component. Copper contamination can lead to complete failure of all components of the chip!).

Typical thickness of metal traces ranges from about 300nm (usually metal 1, aluminum) to $3\mu m$ used for aluminum power routing. Copper power routing can be as thick as $15\mu m$.

The minimum width of wires is defined by the lithography and the etching process. If wet etching is used the minimum width is about 3 times the thickness of the metal. The same applies to the spacing between two wires.

Dry etching (plasma etching) creates vertical edges of the metal. This allows a reduction of the minimum width to about the thickness of the metal. Using dry etch the spacing between the wires can be reduced too.

4.1.1 Electromigration

The second limiting factor for the wire width is electromigration. Electromigration depends on the current flowing (mainly the DC component), the temperature and the desired operating life time. The most basic equation applicable to direct current was proposed in [20].

$$\frac{1}{MTF} = A * J^2 * exp(-\frac{\phi}{k * T}) \tag{4.1}$$

In Black's equation J is the current density in A/m^2, A is a constant factor including the cross section of the conducting film, ϕ is the activation energy of the material in eV, k is the Boltzman constant and T is the absolute temperature. MTF is the time at which 50% of the devices under test have failed.

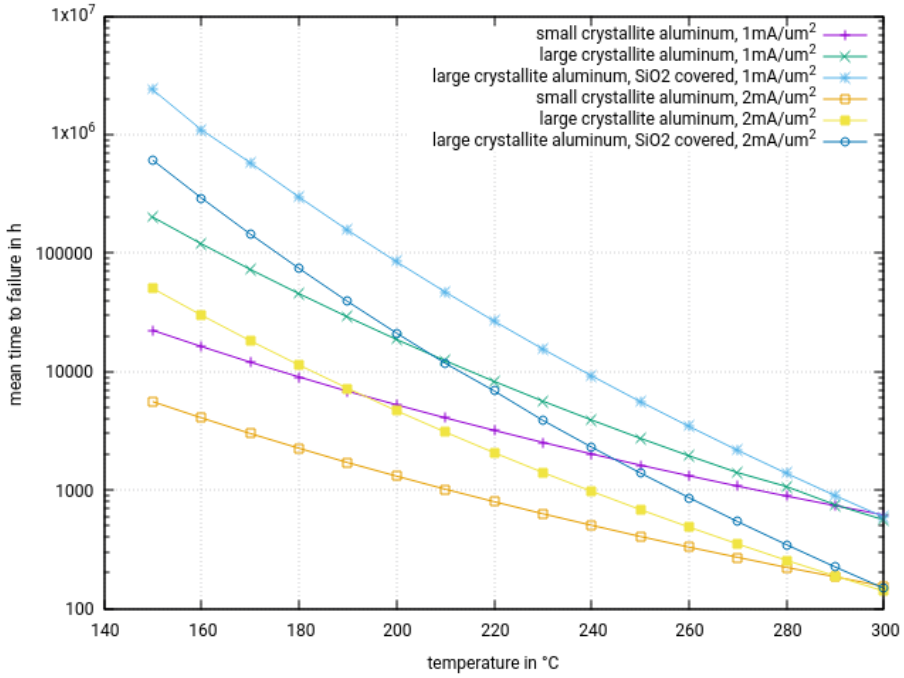


Figure 4.1: Mean time to failure (MTF) versus temperature for Aluminum

Usually we are more interested in the permissible current density to achieve a certain mean time to failure (MTF). So the equation has to be rearranged accordingly.

$$J = \sqrt{\frac{1}{A * MTF} * exp(\frac{\phi}{k * T})} \tag{4.2}$$

Parameters A and ϕ depend on the material used for the conducting film. (In some cases I didn't find all parameters in the publication. So some parts of the following table still are not filled with numbers.). The numbers of the following table were mainly extracted from the graphs shown in [20]. Besides the material the size of the crystallites and the coverage has an impact. The larger the crystallites the more rugged the conductive layer gets.

Table 4: Electromigration coefficients of various materials

material	A	unit	ϕ	unit
Al (small crystallites)	0.404[20]	$\frac{m^4}{A^2 * h}$	0.5 [20]	eV
Al (large crystallites)	290[20]	$\frac{m^4}{A^2 * h}$	0.82 [20]	eV
Al covered with SiO2 (large crystallites)	268060[20]	$\frac{m^4}{A^2 * h}$	1.16 [20]	eV
AlSi (covered with SiO2)		$\frac{m^4}{A^2 * h}$	1.2 [20]	eV
AlSiCu (5% Cu, covered)		$\frac{m^4}{A^2 * h}$	1.2 [20]	eV
Cu (SiO2 covered)		$\frac{m^4}{A^2 * h}$	1.0 [23]	eV
Cu (Ta/TaN covered)			1.4 [23]	eV
Cu (CoWP covered)			1.9-2.4 [23]	eV

The numbers of factor A look a lot different for different metalizations. But this is partially compensated by the different exponents (ϕ). Improving the metalization (larger crystallites) and covering the film with oxide mainly helps at low temperatures. Approaching 290°C all curves merge and the electromigration accelerates due to additional mechanisms becoming more significant. At about 450°C aluminum conductors start to melt and life time drops to 0.

The following plot shows the MTF of different metal films described in [20] for current densities of $1mA/\mu m^2$and $1mA/\mu m^2$. These are common current densities permitted in automotive and industrial chips.

4.1.2 Rules of thumb

For practical design it is handy to have some rules of thumb:

$$MTF \sim J^{-2}$$

$$MTF \sim exp(-\Delta T/20K).....exp(-\Delta T/10K)$$

Since modern technologies try to exploit the advantages of large crytallites and good passivation covering the metal with oxides or nitrides the roll off at high temperatures gets more significant the more modern a metalization process is. The benefit of more modern process technologies mainly applies to low temperatures below 200°C. Below 150°C the chemistry of the metalization makes a big difference. Above 200°C this advantage more and more disapears.

Restricting the temperature range (for instante to $T_{jmax} = 100°C$ as often observed in consumer products offers the oportunity to make the wiring of the chip magnitudes more dense than at higher temperatures.

Important note: The table applies to thin film conductors (some μm). If the conductor is thicker we get Joule-heating and lower current densities must be used!

If the current is pulsed [22]the MTF follows the MTF of a direct current over the squared duty cycle m:

$$MTF_{pulsed} = MTF_{DC} * m^{-2} \tag{4.3}$$

The metal is taken out of the lattice by the impact of the electron flowing. As a consequence the metal follows the movement of the electrons similar to material transported by a river. Metal gets eroded at the location with the highest current density and settles downstream (towards the positive node) when the current density decreases again. Where the metal is deposited we will find hillocks (sometimes called whiskers if found on board level).

Gradients of the current density as well as changes of the passivation covering the metal trace lead to an increase of electromigration [23].

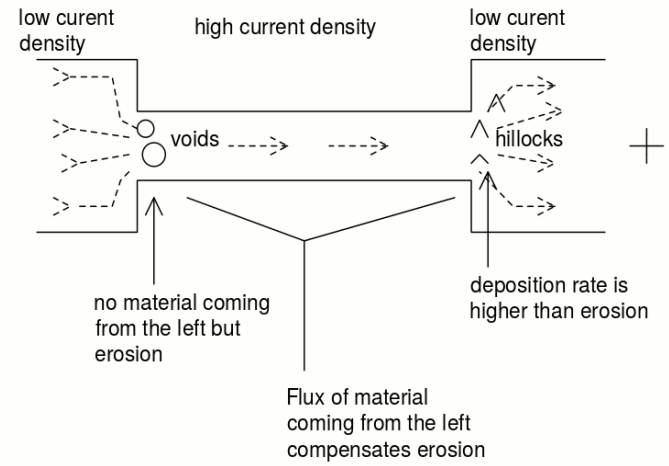


Figure 4.2: Current density gradients leading to increased electromigration

Furthermore later publications [21]suggested refinements of Black's equation to better represent multiple mechanisms contributing. These models are fairly complex and in some cases depending on the geometry of the conductor (e.g. if the current density changes locally or if additional mechanical stress is involved!). So Black's equation can be regarded as a first guess. However it is strongly suggested to perform reliability (High Temperature Operating Life tests, HTOL) tests before going into production with a new device.

4.1.3 Resistance of a wire

Usually wires on a chip are thin films. The resistance is convenient to calculate if we have the resistance per square of this film. Typical values found in many technologies at 300K are about:

Table 5: Wire resistivities of typical metalizations of a chip

Material	Thickness/μm	Resistivity / $m\Omega/\#$	usage
Aluminum	0.3	94	signal wires
Aluminum	0.7	40	signal wires
Aluminim	1.0	28	supply routing
Aluminum	3.0	9.4	supply routing
Copper	3.0	5.6	power routing
Copper	15.0	1.12	power routing

Most metal resistors have temperature coefficients of about 0.003/K to 0.005/K!

4.1.4 The via how to

Vias are connecting different layers of metal. Usually vias and contacts are the smallest features of every process. Therefore vias and contacts belong to the most expensive masks and to the most critical process steps!

Since these process steps are especially critical everything is optimized for a certain via or contact size (with the exception of the edge seal). One via or contact only can transport a limited amount of current. Usually this current is defined by the circumference of the via and the thickness of the metal trace attached. (in case the metal layers connected by the via differ in thickness the thinner one becomes the limiting factor.)

$$I_{via} = 2 * (w + l) * d * J \tag{4.4}$$

In this equation w and l are the lengths of the two sides of the (normally more or less rectangular) via. d is the thickness of the metal attached. i is the maximum permissible current density of the metal trace.

Example: w=0.3μm, l=0.3μm,d=0.25μm,J=1mA/μm^2leads to

$$I_{via} = 2 * (0.3 + 0.3) * 0.25\mu m * 1mA/\mu m = 0.3mA$$

Power traces that are designed will need much more than just one via. Power transistor often are connected with thousands of vias. Since the current through the vias should be distributed homogeneously the vias can not be placed one behind the other but must be placed in parallel!

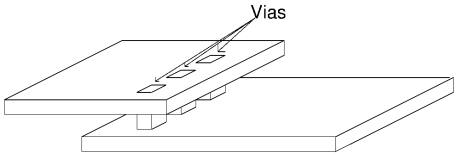


Figure 4.3: good placement of vias

If the vias are placed as shown below the current is very inhomogeneous. The current follows mainly the red path. Only the first via (V1) and the last via (V6) will carry current. In between the voltage drop on the top side (V16top) and on the bottom side (V16bottom) is equal. The voltage drop across vias V2 to V5 is zero.

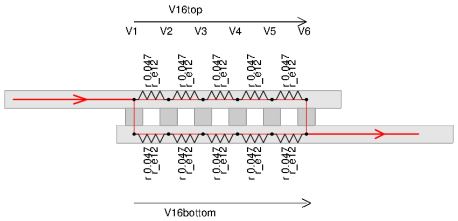


Figure 4.4: Poor placement of vias

Fig.4.1.3.2: Poor placement of vias

The situation improves if the vias have a certain resistance. Tungsten is about 3 times more resistive than aluminum. Taking the via resistance into account we may find a network similar to the following:

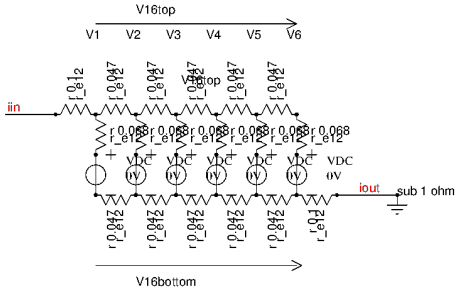


Figure 4.5: Test circuit representing resistive vias

The 0V sources are needed to measure the currents ib the simulation. A simple SPICE simulation of the above circuit using a 1mA current source shows that even in this less extreme case the current distribution is unbalanced.

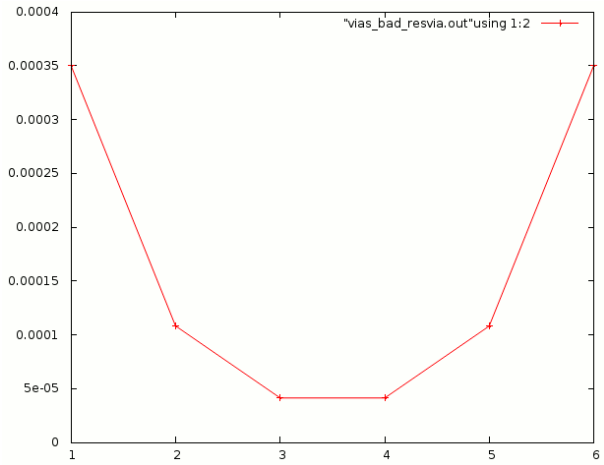


Figure 4.6: Current carried by the different vias. X-axis: via number, Y-axis current in A

The conclusion of this current distribution is that multiple rows of vias or vias aligned one behind the other in stead of parallel will not significantly increase the current capability anymore when the number of rows exceeds about 4 to 5. The vias in the middle will not contribute anymore!

There is a possibility of improving the situation decreasing the width of the top metal from left to right and at the same time increasing the width of the bottom metal. This forces more current into the inner vias. Intentionally thinning wires in a defined way to homogenize the current flow in serial aligned vias leads to a very sophisticated layout and should only be considered in exceptional cases.

4.1.5 Wires acting as antennas

Short compared to the wave length: Wires on a chip (in most cases even on a board) are short compared to radio waves (unless you are working in the range of tens of GHz). Therefore in most cases short wires can be regarded as capacitive antennas to infinite distance. The effectiveness of such short antennas is poor but in some cases chasing the micro volts of radiated emission these capacitive antennas still matter. To estimate the stray capacity acting as an antenna the far field of a wire can be regarded as a spherical capacitor.

$$C_{sphere} = 4 * \pi * \varepsilon_0 * \varepsilon_r * \frac{R_1 * R_2}{R_1 + R_2} \tag{4.5}$$

If the wire is above a ground plane only one half of the sphere is to be considered as a capacity to infinite distance.

$$C_{halfsphere} = 2 * \pi * \varepsilon_0 * \varepsilon_r * \frac{R_1 * R_2}{R_1 + R_2} \tag{4.6}$$

If $R_2 \to \infty$ the capacity of the half sphere simplifies to

$$C_{halfsphere\infty} = 2 * \pi * \varepsilon_0 * \varepsilon_r * R_1 \tag{4.7}$$

In the near field (R<length/2) the wire can be regarded as a cylinder capacity (Ccyl). The ends of the wire can be regarded as half spheres again (Cend).

The total capacity to infinity can be approximated by the series capacity.

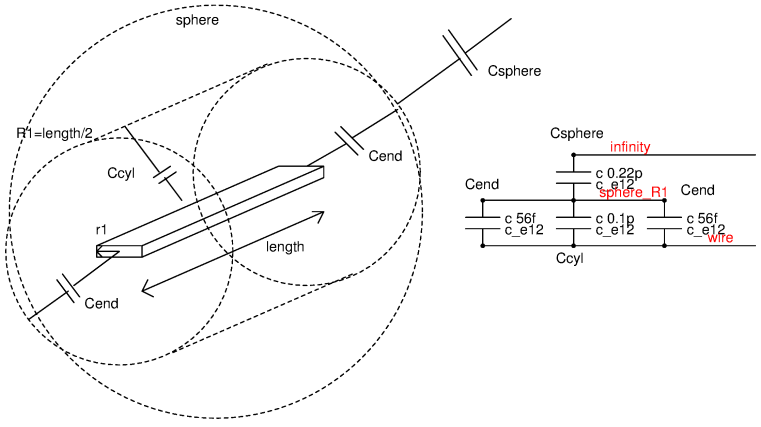


Figure 4.7: Estimation of a capacitive antenna (short compared to the wave length)

So a piece of wire in free air can be approximated by a series of capacitors connected to infinity.

$$C_{sphere} = 4 * \pi * \varepsilon_0 * \varepsilon_r * R_1 \tag{4.8}$$

with $R_1 = length/2$

$$C_{cyl} = length * \frac{\varepsilon_0 * \varepsilon_r * 2 * \pi}{ln(R_1/r_1)} \tag{4.9}$$

$$C_{end} = 2 * \pi * \varepsilon_0 * \varepsilon_r * \frac{r_1 * R_1}{r_1 + R_1} \tag{4.10}$$

$$C_{antenna} \approx \frac{C_{sphere} * (C_{cyl} + 2 * C_{end})}{C_{sphere} + C_{cyl} + 2 * C_{end}} \tag{4.11}$$

In case of a wire over a ground plane only half of the field lines are going to infinity and the capacity halfs.

$$C_{antennaplane} \approx \frac{1}{2} * \frac{C_{sphere} * (C_{cyl} + 2 * C_{end})}{C_{sphere} + C_{cyl} + 2 * C_{end}} \tag{4.12}$$

Example: A wire 1mm wide, 1cm long on a board with ground plane couples to a reference plane in far distance from the board (for instance the shielded wall of an EMC test facility) with about

$$C_{sphere} = 4 * \pi * \varepsilon_0 * 0.005m = 0.55pF$$

$$C_{cyl} = 0.01m * \frac{2 * \pi * \varepsilon_0}{ln(5mm/0.5mm)} = 0.24pF$$

$$C_{end} = 4 * \pi * \varepsilon_0 * \frac{0.0005m * 0.005m}{0.0055m} = 50.3fF$$

$$C_{antennaplane} = 0.5 * \frac{0.55pF * 0.346pF}{0.55pF + 0.346pF} = 0.106pF$$

If the antenna is excited with an RF signal there will be a current flow through this capacity to the wall of the EMC test facility. The return current will flow through the ground wire of the system under test. The RF voltage measured between the ground of the device under test and the reference ground connected to the wall depends on the current flowing and the inductance of the ground wire. Since the ground wire in most EMC test setups is in the range of 20cm (200nH!) the voltage drop can be significant! This leads to the impression that a signal is noisy although it has nothing to do with the actual RF emission source that is radiating via a short piece of wire producing a ground bounce with it's return current.

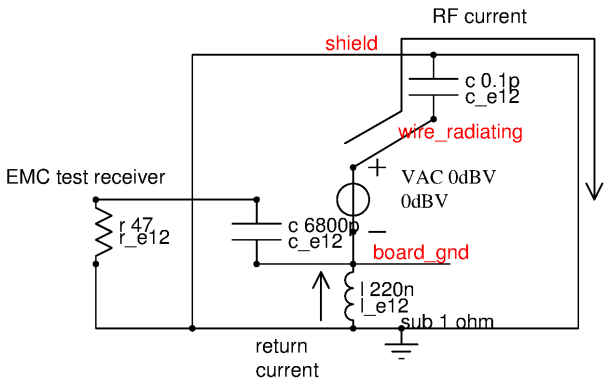


Figure 4.8: current flow of radiated power in an EMC test facility

Long compared to the wave length: As soon as the wire becomes longer than about $\lambda/10$ we have to regard the wire as an efficient antenna. This means it has two field components:

- the electric field
- the magnetic field

This is totally different from a short antenna described above. We will observe a real electromagnetic wave starting to travel away from the antenna!

The H-field starts to create rings of opposit directivity. The E-field starts to create toruses that are getting bigger and bigger (in Y direction) the further we move away from the antenna. The rings and toruses travel away from the antenna.

The drawing shows a dipole. A classical $\lambda/2$ antenna can be regarded as a half of a dipole beginning at the symmetry line (horizontal dotted line). The other half is simply replaced by a ground plane

The real antenna has no field directly above the antenna and directly below the antenna. The wave only travels in horizontal direction. This is why a dipole has a certain directivity leading to an antenna gain.

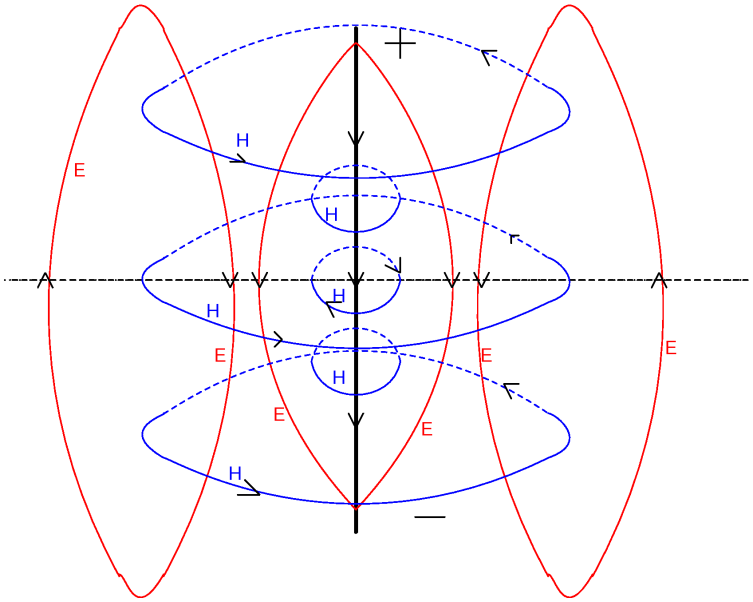


Figure 4.9: A long wire acts as an antenna that really creates a travelling wave.

Since the wave consists two components, the magnetic (H) field and the electrical (E) field the wave must be represented as a power flowing away from the antenna instead of a single field component. The power density observed at a distance R from the antenna can be described as the total power divided by the surface of the sphere at radius R:

$$p = \frac{P_{tx} * G_{tx}}{4 * \pi * R^2} \tag{4.13}$$

(Note: This assumes an isotropic antenna without any directivity!)

$P_{tx}$is the power of the transmitter. $G_{tx}$is the antenna gain, that depends on the directivity of the antenna design. R is the distance at which the power density p (in W/m^2) is measured.

The power a receiving antenna can pick from the field depends on the equivalent receiving area.

$$P = p * \frac{\lambda^2 * G_{rx}}{4 * \pi} \tag{4.14}$$

Combining both equations provides the ideal received power:

$$P = \frac{P_{tx} * G_{tx} * G_{rx}}{(4 * \pi * R)^2} \tag{4.15}$$

This equation applies to simple wire antennas up to one wavelength. Antennas with reflectors that are bigger than the wave length have a bigger receiving area because all the waves are reflected in phase to the antenna in the focus of the reflector. There are two possibilities of calculating the received power:

1. Using the area $A_{rx}$of the reflector:

$$P = \frac{P_{tx} * G_{tx}}{4 * \pi * R^2} * A_{rx} \tag{4.16}$$

2. Using the antenna gain of the receiving antenna

In this case we are using the standard equation with wave length λ again. The ratio between the wave length and the reflector area simply coded into the receiving antenna gain G_{rx}. For a round reflector with radius r the antenna gain becomes:

$$G_{rx} = \frac{4 * \pi^2 * r^2}{\lambda^2} \tag{4.17}$$

In real applications reflections and adsorption will take place. Furthermore peridic antenna structures (for example multiple antennas excited in phase but with $n * \lambda$ spacing) can create a much higher antenna gain in one or two directions (Yagi antenna). The special antenna properties usually are expressed by the antenna gain. The real signals observed at the receiver can be higher or lower depending on these properties of the environment. Nevertheless this simple equation is a good starting point to estimate the power transfer of a long antenna.

4.2 Contacts

Usually a contact connects a more or less resistive (doped) silicon to a metal path. The metal connecting to a low doped silicon would create a schottky diode. Therefore the silicon at the contact must be high doped (Well, this still would lead to a schottky barrier, but the high doping makes the depletion area so short that the electrons can cross it and we get a contact in stead of a diode)

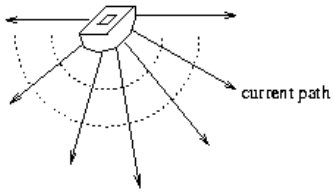


Figure 4.10: Small single contact

The current close to the contact flows vertically through a half sphere. The surface of this half sphere grows with the distance from the contact. The surface A(x) of this half sphere calculates as:

$$A(x) = 2 * \Pi * x^2 \tag{4.18}$$

The differential resistance dr over a way dx at the half sphere calculates as:

$$dR = \frac{r * dx}{A(x)} \tag{4.19}$$

In this equation r is the resistivity of the material. To get the contact spread resistance dr has to be integrated from the end of the high doped implant under the contact to the location we are interested in.

$$R = \frac{r}{2 * \Pi} * \int \frac{1}{x^2} dx \tag{4.20}$$

Solving the integration we get:

$$R = \frac{r}{2 * \Pi} * (\frac{1}{x_1} - \frac{1}{x_2}) \tag{4.21}$$

Example: We have a contact with 2μm diameter (leading to x1=1μm) that is 10μm away (leading to x2=10μm) from the point of interest. As a resistivity r we assume 10Ωcm. Using the above equation we calculate a contact resistance of:

$$R = \frac{10\Omega cm}{2 * \Pi} * (\frac{1}{0.0001cm} - \frac{1}{0.001cm}) = 14324\Omega$$

So a single minimum contact is quite high resistive!

Very often in a chip design lines of contacts or a strip shaped contact is used. These long contacts can be approximated assuming the current is flowing vertical to the surface of a half torus.

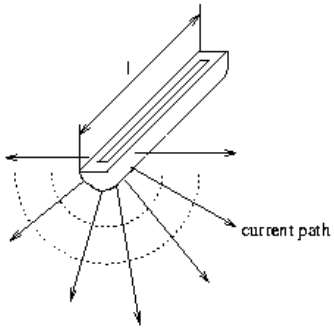


Figure 4.11: Strip shaped contact

The area the current has to flow through calculates:

$$A(x) = x * l * \Pi \tag{4.22}$$

So the differential resistance becomes:

$$dR = \frac{r}{\Pi * l} * \int \frac{1}{x} dx \tag{4.23}$$

Solving the integral we get:

$$R = \frac{r}{\Pi * l} * ln(\frac{x_2}{x_1}) \tag{4.24}$$

Example: A contact strip of 2μm width (leading to $x_1 = 1\mu m$) and 100μm length. As before the point of interest is 10μm away. (Well, it is more something like a conductive surface collecting the current).

$$R = \frac{10\Omega cm}{0.01cm * \Pi} * ln(10) = 733\Omega$$

Still some resistance. So there is only one replacements for contacts: More contacts!

4.2.1 Local interconnect

Some processes offer a highly conductive salicided surface of the silicon. This increases the effective size of the contact. The contact itself can be made very small allowing very small wires to contact the devices. Below the contact (inside the silicon or the poly silicon layer) there is a local interconnect layer (usual layer name "LI") that distributes the current over a wider area. This way the contact spread resistance can be reduced significantly.

Some logic technologies even permit using the local interconnect as a signal layer or to connect the bulk! Use this feature with caution. A metal trace for sure is better.

4.2.2 Device matching

Contacts have different mechanical properties than silicon oxide (Tungsten is harder, Aluminum is softer). Mechanical tension can change the bandgap energy by several meV. For matching devices always use the same kind of contacts and place them at the same position.

4.2.3 Seebeck effect

Contacts always are interfaces between materials with different affinity to electrons [32, 67]. As a consequence each of these interfaces has a built in voltage. As long as all contacts connecting a piece of silicon have the same temperature the built in voltage cancels. As soon as the temperature of contacts differs we can measure a voltage between the contacts a different locations (and temperatures). This effect is used for peltier elements and thermocouplers. In resistor ladders (e.g. if the voltage drop per resistor segment is low!) temperature gradients can produce significant offsets. Usually these contact voltages connecting semiconductors are in the range of 0.2mV/K to 1.4mV/K.

Table 6: Seebeck coefficients

material	usage/position	thermo voltage
Si	gate contact, source contact	$\approx 440\mu V/K$
Ge	emitter contacts	$\approx 300\mu V/K$
NiCr	Nichrome	$25\mu V/K$
Fe	Pin of ICs	$19\mu V/K$
W	Tungsten vias	$7.5\mu V/K$
Au	bond wires	$6.5\mu V/K$
Ag	RF inuctors	$6.5\mu V/K$
Cu	wires, bond wires	$6.5\mu V/K$
Pb	lead solder	$4.0\mu V/K$
Al	pad, wires on chip	$3.5\mu V/K$
C	Carbon resistors	$3.0\mu V/K$
Pt	reference	0
Ni	Nickel	$-15\mu V/K$
	Constantan resistors	$-35\mu V/K$
Bi	Bismuth solder	$-72\mu V/K$

4.3 Resistors

Modern technologies offer a big variety of resistors. Usually almost any layer can be used. In most cases not all possible layers are made available to keep the CAD support effort low. Typical resistor ranges are:

Metal resistors using copper usually are in the range of 1..5 mOhm/#. These resistors are used to separate nets (to control the way the wiring is done by LVS) or to measure very high currents in the range of Amperes.

Metal resistor using aluminum layers are in the range of 10 mOhm/# to 50mOhm/# depending on the thickness of the metal.

Poly silicon resistors can be salicided or unsalicided. Salicided poly silicon resistors are in the range of 1..10 Ohm/#. Unsalicided poly silicon resistors are in the range of 50 Ohm/# to 2000 Ohm/#. Usually poly silicon resistors of about 100 Ohm/# to 200 Ohm/# are the best controlled ones. Poly silicon resistors with higher resistivity tend to have more spread and usually have a higher temperature coefficient.

Silicium-Chrome (SiCr) resistors are used for high precision resistors. SiCr usually has a better matching than poly silicon resistors. SiCr resistors usually require additional masks.

4.3.1 Poly silicon resistors

Poly silicon resistors (and SiCr resistors) are oxide isolated. For this reason in most modern processes the poly resistor is the standard device. The only parasitic components remaining are capacities. **These capacities however should not be underestimated. If resistors are placed over a bouncing substrate (relative to the ground of the block they belong to) the substrate noise is coupled into the circuit**.

One of the disadvantages of poly silicon resistors is the poor thermal conductivity of the oxide surrounding the poly silicon. Self heating becomes a problem if good matching is required and parts of the resistor network use parallel devices while other parts of the network do not. (The parallel devices carry less current per device and will not heat up in the same way.)

ESD and electrical over stress (EOS) may be a second limitation if the resistors heat up to more than 300°C. (The thermal capacity of poly silicon resistors is very low!) Usually electrical over stress partially melts the poly silicon. This changes the grain boundaries making the resistors less resistive.

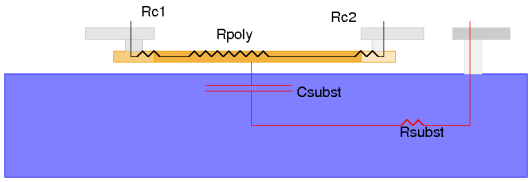


Figure 4.12: Poly silicon resistor over substrate

Most important parameters of these resistors are:

Table 7: Parameters of poly silicon resistors

parameter	symbol	unit	comment
resistivity	r#	Ohm per square	resistivity of Rpoly is needed to calculate the resistor value
current density	i_{max}	mA per μm	maximum permissible current density
contact resistance Rc1 and Rc2	Rc	Ohm	The contact resistance usually has a lot of spread. Rpoly>>Rc is desirable
parasitic substrate capacity	Csubst	fF per μm^2	needed to calculate parasitic capacitive couplung to substrate
Oxide break down voltage	Vox	V	permissible voltage between poly silicon and substrate

If capacitive coupling to substrate has to be avoided poly silicon resistors over nwell must be used. The nwell then can be tied to a silent signal (relative to the ground of the block the resistor belongs to).

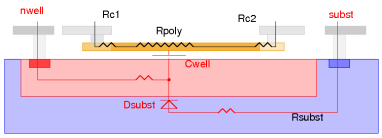


Figure 4.13: Poly silicon resistor over nwell

Most important parameters of poly resistors over nwell are:

Table 8: additional parameters of poly silicon resistors placed over wells

parameter	symbol	unit	comment
resistivity	r#	Ohm per square	resistivity of Rpoly is needed to calculate the resistor value
current density	i_{max}	mA per μm	maximum permissible current density
contact resistance Rc1 and Rc2	Rc	Ohm	The contact resistance usually has a lot of spread. Rpoly>>Rc is desirable
parasitic substrate capacity	Cwell	fF per μm^2	needed to calculate parasitic capacitive couplung to substrate
Oxide break down voltage	Vox	V	permissible voltage between poly silicon and substrate
substrate diode break down voltage	Vbrd	V	Needed to define which well voltage can be used

Some technologies offer triple wells that can be used to isolate resistors from substrate noise. Using multiple well leads to the best possible decoupling from substrate noise.

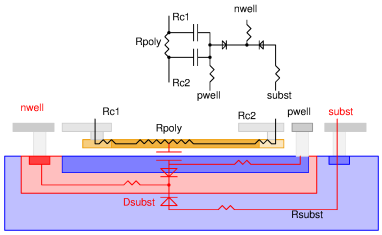


Figure 4.14: Double well noise isolation of a poly silicon resistor

4.3.2 Thermal Noise of a resistors

Resistors are a source of thermal noise. The power of the noise is a function of the absolute temperature. $v_n$is the density of noise in $V/\sqrt{Hz}$.

$$v_n = 2 * \sqrt{R * k * T} \tag{4.25}$$

For room temperature the noise voltage of a 1K resistor is about $4nV/\sqrt{Hz}$. A resistor with 100K has $40nV/\sqrt{Hz}$. To obtain the total noise this density must be multiplied with the square root of the bandwidth.

$$V_n = v_n * \sqrt{BW} = 2 * \sqrt{R * k * T * BW} \tag{4.26}$$

4.3.3 Poly silicon resistor aging:

The resistance of a poly silicon resistor depends on carrier mobility and the grain boundaries. Carrier mobility and carrier density can be affected by electrical fields at the surface and by mechanical stress as well as thermal stress. For low aging the following general rules can be given:

- Minimize self heating (don't allow more than 10K temperature difference versus the rest of the chip)
- Minimize mechanical stress (caused by the package). Place critical resistors in the middle of the chip.
- Avoid electrical fields close to the resistors (no high voltage wires crossing).

- Place resistors over wells with similar voltage to minimize fields between resistors and the well underneath.
- resistor heads with inhomogenous doping are more sensitive to aging. Keep ratio L/W $>$ 6 to 10.
- Don't allow ESD events (ESD may alter grain boundaries and this way change resistance).

Design rule manuals usually state a very wide range of aging even for similar process nodes. This range stretches from less than 0.3%/10 years (following the rules listed above) to about 2%/10 years (No special indications about layout given).

4.3.4 Resistor matching

Resistor matching depends on doping and mask accuracy (focus of the photolithography!). Processes with good planarization usually have better photolithography (everything is in the focus plane and resistor edges are well defined). But even assuming an excellent process at the end the statistics of doping determine the achievable matching. The smaller the structure the lower the number of dopants and the higher the spread. At the end it boils down to the same as the current matching of MOS transistors.

Resistors in a planarized process have a matching of about 1..2 $\%\mu m$.

Resistors in a non planarized process can have between 1 and 10 $\%\mu m$.

Gradients over the wafer (assuming a mature process) are typically 1%/cm or 1ppm/um. Small resistor arrays (smaller than about 30μm * 30μm) don't need interdigited design. If the array gets bigger interdigited layout is worth considering.

Good layout practice Resistors that are intended to match should be close to each other, have identical width and identical length. If ratios different than 1 are needed the resistors should be composed of unity modules.

To prevent offsets caused by the Seebeck effect multi module resistors should be placed on isotherms in a way that Seebeck voltages either are avoided or are canceled. The current direction should be intermittent forward and backward. Consequently the number of modules should be even numbers to make thermovoltages cancel.

Current flow in contacts is prone to tiny differences of the contact geometry. Since the contacts are very small compared to the resistors the matching of contacts is poor. To keep the contact spread far from becoming a dominant effect it is good practice to design resistor modules at least factor 6 longer than they are wide. Furthermore the heads of the resistors can have additional dopings to make the contact spread resistance lower.

Building RF dividers even the lateral coupling to the adjacent resistor must be taken into account. Building dividers with dozens of kilo ohm a parasitic coupling in the range of a fF matters at 20MHz and higher! Therefor RF dividers should not use interdigited layout.

4.3.5 Diffused resistors

Diffused resistors are common for applications with high power density because heat transport isn't hindered by silicon oxide. Often diffused resistors are used for ESD protections.

Diffused resistors are junction isolated. Every diffused resistor has one or more parasitic diodes to the associated well. The break down voltage of the parasitic diode depends mainly on the well (usually the well doping is lower than the resistor doping). In case of ESD protections the well diode is part of the protection and the well contacts are intentionally designed to carry ESD pulses.

Since diffused resistors are junction isolated the effective thickness and the effective width depends on the depleation zone between the well and the resistor. The resistor becomes voltage dependent! High resistive material can even behave like a J-FET.

In some processes an anisotropy was observed. The resistor slightly depends on the polarity of the current. (Observed change of resistance of p-body resistors in BCD2 technology: 0.3% using identical layout - same cell - but flipped placement)

4.4 Capacitors

There are dozens of possible implementations of capacitors on a chip. Basic idea however is always to create conductive planes with a non conductive layer in between. This non conductive layer can be an isolation such as silicon oxide or nitride or a depletion zone using an blocked junction. Combinations are possible as well.

To achieve high area capacities the layer in between is made as thin as possible. The higher the break down voltage the thicker the isolation layer and the lower the specific capacity. (low specific capacity=high silicon cost per pF).

All capacitors available on a chip also have parasitic capacities to the layers below (often this is substrate) and above (metal traces crossing them or passing in close proximity). These parasitic capacities must be considered in circuit design. It is strongly suggested to use symbols that also have a pin for the parasitic capacitor(s) to make everybody aware that he is dealing with components that are far from ideal. Sometimes it is possible to operate the parasitic capacitor like a driven shield to minimize the impact on circuit performance.

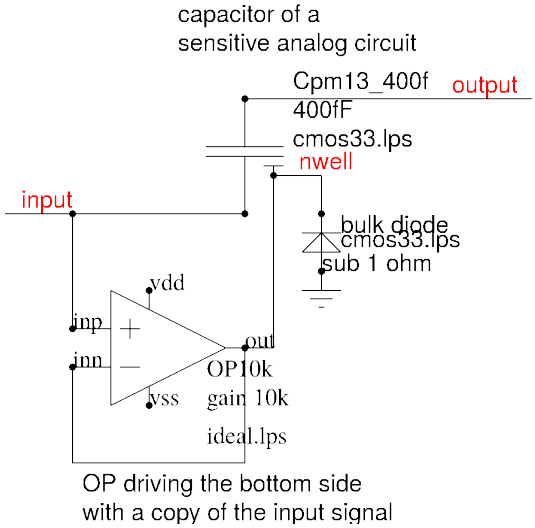


Figure 4.15: driven shield to minimize charge loss into the bottom layer under the capacitor

4.4.1 Junction capacitors

Junction capacitors were the first capacitors used on integrate circuits. Basically a junction capacitor is a large area diode which is reverse biased. The distance between the plates is the length of the depletion zone. Usually one side of the junction is higher doped than the other side. So most of the electrical field is accommodated by the low doped material.

Typically there is no specific capacitor diffusion available. In stead designers use the junction that satisfies their break down voltage requirements. The width of the depletion zone calculates as:

$$d = \sqrt{\frac{2 * \varepsilon_0 * \varepsilon_r * (V_b - V_{bi})}{e * N_D}} \tag{4.27}$$

In this equation $V_b$is the voltage applied. $V_{bi}$is the built in voltage - well bottom line it is the forward voltage if the capacitor is operated as a diode, about -0.6V for silicon (we count against the forward direction). N_D is the doping of the low doped side of the junction. ε_r is the relative permittivity. In case of silicon this is about 12.

Knowing the length of the depletion zone the capacity calculates as:

$$C = \varepsilon_0 * \epsilon_r * \frac{w * l}{d} \tag{4.28}$$

Combining both equations we get

$$C = w * l * \sqrt{\frac{e * N_D * \varepsilon_0 * \epsilon_r}{2 * (V_b - V_{bi})}} \tag{4.29}$$

The voltage dependence of the junction capacitor immediately becomes visible.

Reading design rule manuals there often is no number for N_D. In stead we have to estimate it from the break down voltages given. In literature [28, 26] we find numbers of break down field strengths of $30V/\mu m$ to $60V/\mu m$ for silicon. The higher number is more a theoretical value for ideal plane junctions. Since this will never found in practice using the lower number of $30V/\mu m$ is the more appropriate choice. Assuming a triangular shape of the electrical field in the depletion zone we can calculate the doping.

$$N_D = \frac{E_{br}^2 * \varepsilon_0 * \varepsilon_r}{2 * e * (V_{br} - V_{bi})} \tag{4.30}$$

As a rule of thumb we can expect typical junctions used for depletion capacitors:

Table 9: typical bipolar device properties

junction	V_{br}	V_{usage}	typical N_D	capacity at V_{usage}	shield
base-emitter	5V - 9V	up to 5V	$5 * 10^{22}/m^3$	290pF/mm^2	possible
base-collector	15V - 100V	up to Vce	$5 * 10^{21}/m^3$	29pF/mm^2at 15V	no
epi-substrate	50V - 150V	up to Vcb	$2 * 10^{21}/m^3$	6pF/mm^2at 60V	no
drain/source-bulk	3V-10V	3V	$5 * 10^{22}/m^3$	400pF/mm^2at 2.5V	possible

Resistance of the diffusion: The diffusions or implanted layers used for the capacitor of course have a significant resistance. So a junction capacitor is not ideal. There always is a serial resistance in the device. This serial resistance depends on the shape of the capacitor. For large capacitors it is recommended to connect the diffusions involved with interdigited metal fingers to keep the resistive paths in the range of only a few squares.

4.4.2 Poly silicon capacitors

Most processes offering poly silicon also feature thin (gate) oxide. Thin oxide usually is a good choice for capacitors because it usually is well controlled (thickness only varies in a very little range). Typical 6 sigma spread is in the range +-15% or even better.

Unfortunately the gatepoly silicon often is used as a mask for N+ and P+ (drain and source of NMOS and PMOS transistors. As a consequence the capacity is between the bulk doping and the gate poly silicon rather than P+ or N+. Care must be taken about the polarity of the capacitor. If the poly is over P-well the poly plate must be negative versus the pwell (pwell and p+ contact). If the poly plate is over nwell the poly plate must be positive versus the nwell. This is required to prevent inversion of the well close to the oxide. Inversion of the well would change the thickness of the dielectric layer. This would reduce the capacity.

There is one exception to the rule: If the capacitor is designed as a MOS transistor (pwell but n+ contact) the gate of the nmos transistor must be positive to create the channel acting as the opposit side of the capacitor. If the capacitor is designed as a PMOS transistor the gate must be negative versus the channel.

Some processes offer a very low resistive sinker diffusion that can be used as the bottom plate of the capacitor. The advantage of the sinker is that the high doping at the silicon surface prevents inversion. Poly silicon capacitors using sinker (or an other very high doped material) are polarity insensitive.

4.4.3 Isolated poly silicon capacitors

If a process offers multiple levels of poly silicon fully isolated capacitors can be built. One of the poly silicon layers becomes the bottom plate, the other poly silicon layer becomes the top plate. The oxide between the capacitor plates will be an CVD (chemical vapor deposition) oxide. CVD oxide is less homogeneous than an oxide grown from a noncrystalline silicon. Typical break down field strength of a CVD oxide is in the range of 0.1V/nm to 0.3V/nm depending on the details of the CVD process (speed, temperature..). Pure CVD oxide is too inefficient for creating poly-poly capacitors.

Typically in stead of a simple oxide a sandwich of CVD oxide and CVD nitride is used. This so called ONO benefits from the much higher break down field strength of the nitride. The oxide mainly serves as an interface to the poly silicon layers. ONO capacitors offer break down field strengths of 0.5 to 1V/nm inside the nitride layer. So a 15nm to 20nm nitride is typically used for 5V (long term operation) capacitors.

ONO between two poly silicon layers also can be used to build EEPROM cells.

The bottom poly silicon always has a certain parasitic capacity to the silicon layer underneath. It is strongly suggested to represent poly-poly capacitors as 3 terminal devices (Nodes poly1, poly2, subst).

4.4.4 Metal capacitors

Technologies having multi layer metalization offer the possibility to implement metal capacitors. There are various flavors of such capacitors.

High voltage capacitor: The most classical approach building a metal capacitor is simply having two metal rectangles on top of each other. Between these rectangles there is a layer of CVD (chemical vapor deposition) oxide. The thickness of this oxide ranges from some hundred nm (modern logic process) to about 1um (classical 40V bipolar process of the 1990s).

Between the bottom plate and the silicon substrate there usually is a CVD oxide of similar thickness plus some nm of thermally grown silicon oxide (similar to the gate oxide. It comes for free during the thermal process steps). So this kind of capacitor always has a parasitic capacity between the bottom plate and the silicon underneath in the same magnitude as the intended capacity between the plates.

Since CVD oxide is amorphous the break down field strengths reported in literature are in the range of 0.1V/nm to 0.3V/nm.

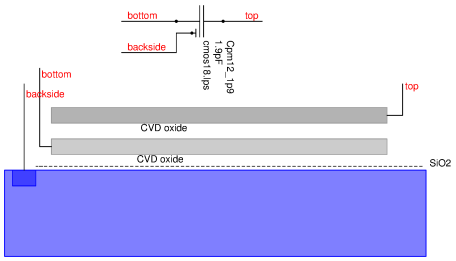


Figure 4.16: High voltage metal capacitor

The size of such capacitors is limited. In most semiconductor processes the maximum metal size without slicing (slicing is an automatic step done at mask generation. All pieces of metal larger than a certain size will be modified such that there are holes inside the metal. These holes provide a mechanical coupling of the oxide below the metal and above the metal by silicon oxide, which is much harder than aluminum.) So there usually is a maximum capacitor size. All capacitors larger than this size must be composed of capacitor arrays. Typical area capacities are in the range of $30pF/mm^2$ to $100pF/mm^2$.

Low voltage metal capacitor (MIM-cap): The thick oxide between the metal layers leads to a very low capacity per area. To achieve higher capacities the oxide is thinned. This is done by creating via holes between the capacitor metals. But in stead of connecting the metal layers the holes are left open and will get oxidized in the next process steps. So at the location of the holes we will find thin oxide. Above the thin oxide the holes are filled with conducting material connecting to the top plate (So there are two kinds of vias in the process: vias without oxide and vias with a thin oxide or nitride breaking the connection between the via and the metal layer below). These capacitors are called MIM-cap (Metal-Isolation-Metal capacitor). 5V MIM caps can have area capacities as high as $2nF/mm^2$. MIM caps are ideal for RF because they have almost no parasitic resistance (ESR).

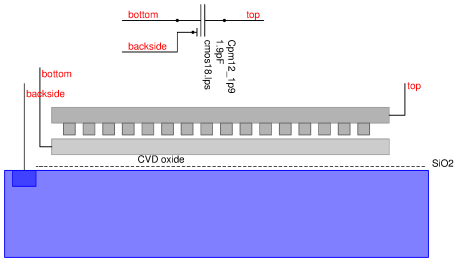


Figure 4.17: Low voltage MIM capacitor

Fringe capacitors: An other way of creating higher capacities on a chip is using the lateral capacity between metal fingers. This becomes attractive if the spacing between the metals and the minimum trace width get less than the thickness of the CVD oxide. Different from a MIM capacitor a fringe capacitor doesn't require an extra mask marking the area where the vias are not created immediately after opening the holes. The parasitic capacity to the silicon underneath is symmetrical on both sides of the intentional capacitor.

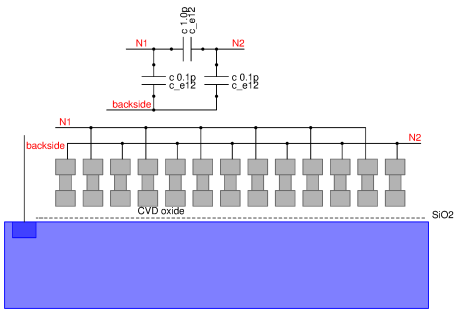


Figure 4.18: Fringe capacitor

The voltage of a fringe capacitor can be scaled changing the lateral spacing of the metal traces.

One big problem of fringe capacitors is their sensitivity to disalignement of the vias and the metal traces. This becomes especially critical if the metal overlap over via is reduced (specifically for the capacitor) to achieve a higher area capacity.

The second big problem of fringe capacitors are etching residuals in the gaps between the metals. Therefore fringe capacitors require screening in production. Screening voltage typically is twice the voltage level used in application.

This requires additional circuitry (even worse: high voltage circuits for screening!) and leads to non ideal effects (capacity to substrate due to the screening circuits and parasitic capacity to the test bus used for screening!)

Slicing rules are not applicable anymore because there is oxide all few micrometers. So mechanically fringe capacities are superior to metal plate capacities.

4.4.5 Noise of a capacitor (KTC-noise)

Capacitors don't produce noise but in combination with resistors they limit the bandwidth and thus the total noise. On the other hand the higher the resistance the more noise voltage (at lower bandwidth) is available. Thus the resistor cancels. The noise density of an RC network becomes:

$$v_n = \sqrt{\frac{k * T}{C}} \tag{4.31}$$

In case of a switch v_n can be regarded as the statistical mean value of the offset left on the capacitor opening an ideal switch.

Example:

A 1pF capacitor at 300K has a noise voltage of:

$$v_n = \sqrt{\frac{8.51733 * 10^{-5}\frac{V}{K} * 1.607 * 10^{-19}As * 300K}{10^{-12}\frac{As}{V}}} = 64.08\mu V$$

4.4.6 Parasitic resistance

Capacitors on integrated circuits in most cases use layers with a significant resistance. So the ESR of a capacitor must be taken into account. It is good layout practice to connect capacitors on the wide side and rather create fingers or small modules to minimize the path length the current has to flow in high resistive layers.

4.4.7 Capacitor matching

Plate capacitor matching mainly depends on the uniformity of the oxide layer. The bigger a capacitor the better non uniformities average out. As a rule of thumb the matching of plate capacitors is in the range of $0.3..2\%\sqrt{fF}$. This is more or less independent of the oxide thickness because the thicker the oxide the more area is needed. So the increase of thickness variations of the oxide gets compensated by the increase of the area needed for the capacitor.

Fringe capacitors are more dependent on the photolithography than plate capacitors. So here additional mismatch contributors are in the game. The achievable matching of fringe capacitors in a mature process is in the range of $0.3..10\%\sqrt{fF}$.

4.4.8 Maximum voltage of a capacitor

Today most capacitors are designed using poly over nwell or poly-poly capacitors. The dielectric is silicon oxide. Silicon oxide time to break down was well investigated in the 1980s [19, 73]. Main parameters to determine the life time of a capacitor or a gate oxide are:

- Oxide thickness (at the weakest spot)
- Temperature
- applied voltage

The equation stated in [19] is:

$$t_{BD} = t_0 * e^{\frac{G * X_{ox}}{V_{ox}}} \tag{4.32}$$

In this equation $X_{ox}$is the effective thickness of the oxide and $V_{ox}$is the voltage over the oxid. In his first paper Chenming Hu simply stated that for silicon oxide G is in the range of 350MV/cm or 35V/nm. The time of $t_0$is given as $1 * 10^{-11}$s. These numbers simply seem to come from the measurement results.

In a later paper [73] Chenming Hu additionally provides information, that the parameter G is temperature dependent.

$$G(T) = G * (1 + \frac{\delta}{k} * (\frac{1}{T} - \frac{1}{300K})) \tag{4.33}$$

In this equation K is the Boltzmann constant expressed in eV/K and δ is derived from measurements at different conditions. The temperature is in Kelvin. Chenming Hu states $\delta = 0.0167eV$. The factor δ/k becomes:

$$M = \frac{\delta}{K} = 194K \tag{4.34}$$

In most technologies life time of a gate oxide refers to a certain - usually high - temperature. (For instance 150°C).

Example: For a 7nm gate oxide the maximum gate voltage for 10000h operating time is give as 3.5V at 150°C. To calculate the highest permissible voltage at -40C we can calculate:

$$t_{BD150} = t_{BD-40}$$

Since $t_0$is a constant (10ps) we simply have to find equal exponents:

$$\frac{G(150C) * X_{ox}}{V_{ox150}} = \frac{G(-40C) * X_{ox}}{V_{ox-40}}$$

$$V_{ox-40} = V_{ox150} * \frac{G(-40C)}{G(150C)}$$

Now we just have to plug in the equation of G(T) and our factor M. To make the result more general insteat of using -40°C and 150°C we just call the two temperatures $T_{cold}$and T_{hot}.

$$V_{oxcold} = V_{oxhot} * \frac{1 + M(\frac{1}{T_{cold}} - \frac{1}{300K})}{1 + M * (\frac{1}{T_{hot}} - \frac{1}{300K})} \tag{4.35}$$

With this equation we can scale the maximum operating conditions at one temperature to a different operating temperature. Using the temperature -40°C and 150°C as an example we find:

$$V_{ox-40C} = V_{ox150C} * 1.46$$

Regard this as a rule of thumb. Other researchers found slightly different values for δ . To be on the save side better calculate with 1.3 instead of 1.46.

Using other dielectrics than silicon oxide different values for $M = \delta/k$ must be used.

4.4.9 External capacitors

On chip capacitors are limited to very low values. This is especially true for high voltage capacitors where dielectric layers of some hundred nm have to be used. Therefore for some applications using external capacitors can't be avoided. There external capacitors aren't ideal either.

1. External capacitors (even SMD devices) have a parasitic inductance (ESL, equivalent series L)
2. ESR (equivalent series resistance) is usually in the range of some 10 mili Ohm.
3. X7R dielectrics have a voltage dependent ϵ_r . This means they are non linear (up to some 10%!)
4. X7R dielectrics have a hysteresis in the range of 1%. This leads to harmonic distortions!
5. High permittivity dielectrics like X7R have a so called battery effect. There is energy stored in the polarization of the dielectric material that has a time constant of some ms to seconds.
6. The dielectric material responds to mechanic tension. It can act as a piezzo microphone.
7. COG capacitors are much more linear but usually more expensive and bigger.

Models provided by manufacturers like Murata [49] can be quite complex. Usually these models are provided as spice netlists. These models hold inductances and even battery effects but hysteresis usually is missing.

The model of course is from metal cap to metal cap of the capacitor. For RF applications the trace and the inductance even of the solder pad is in the same magnitude as most of the inductances inside the capacitor model! The board does definitely play an important role simulating chip and application.

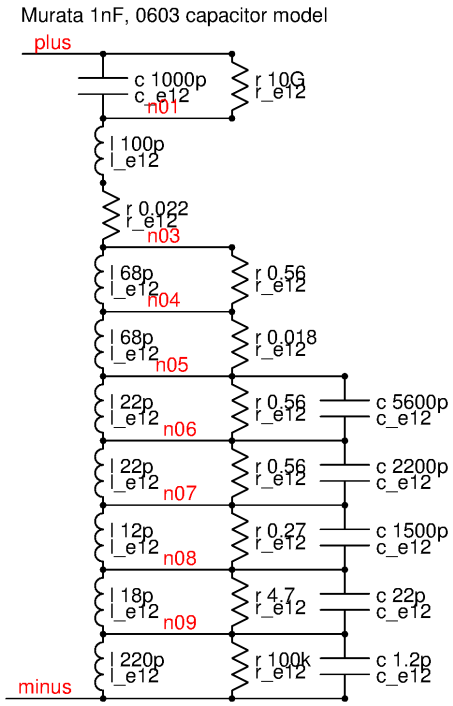


Figure 4.19: Example of a 1nF X7R capacitor model

In most cases the precision of a 9th order model is not needed and more simple models can be used. However above the serial resonant frequency no capacitor acts like a theoretical capacitor anymore. So at least the ESR and the ESL of a first order model should be included for all external components.

4.5 Inductors

Inductors on chip usually are limited to the range of pH to very few nH. Therefore on chip inductors are used in the range of hundreds of MHz. Furthermore on chip inductors usually suffer from eddy currents making them very lossy. The voltage induced in a single winding [51]calculates:

$$V_{ind} = \frac{d\Phi}{dt} \tag{4.36}$$

with

$$\Phi = \int BdA \tag{4.37}$$

A is the area enclosed by the inductor. B is the magnetic field. The magnetic flux Φ can change by either changing the magnetic field or by changing the area. The voltage induced can be increased adding more windings to the inductor.

$$V_{indN} = n * \frac{d\Phi}{dt} \tag{4.38}$$

To give a rough idea of magnetic fields encountered in the environment:

Table 10: Typical magnetical fields of technical systems

source	typical field	remark
earth magnetic field	$47\mu T$	varies depending on location on earth
field in the air gap of a transformer	0.5T	maximum value
field inside an NMR	1..20T	depends on size of the chamber
inductive cooking	5mT	at 5 kHz

If the area is kept constant the calculation simplifies to

$$V_{indn} = n * A * \frac{dB}{dt} \tag{4.39}$$

Units: $1T = 1Vs/m^2$

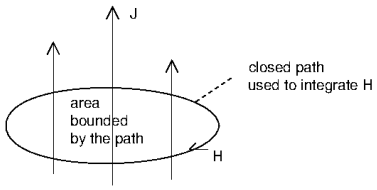


Figure 4.20: Faraday's law

Example: An area of $1cm^2$is exposed to a filed of 0.1T that changes polarity within 5us.

$$V_{ind} = 1 * 10^{-4}m^2 * 0.2\frac{Vs}{m^2} * \frac{1}{5 * 10^{-6}s} = 4V$$

So having a loop with an area of only $1cm^2$ in a field that changes periodically from 0.1T to -0.1T with 100kHz leads to quite some voltage!

4.5.1 External inductors

External inductors typically are found in in application circuits such as switch-mode power supplies or class D amplifiers. Normally they are not matter of concern of the chip designer. Nevertheless it is good practice of circuit design to verify reasonable scaling of the application circuit. For this reason a brief description of these inductors is given here.

Theoretical books on magnetic fields usually present Ampere's law using the following equation [59]:

$$\oint Hdl = \int_{area} Jds \tag{4.40}$$

H is the electromagnetic field intensity. Integrating the electromagnetic field intensity along a closed path equals the total current flowing through the area bounded by this path (J is a current density).

Well, the equation looks a bit abstract. It becomes a bit handy if we consider that the integration over the current density is nothing else than the total current flowing through the area. The total current can of course consist if several (n) wires with a current I flowing in each wire. In addition in reasonable technical inductors H is more or less piece wise constant. (there are cores with intentional air gaps. So the field H is constant in the core material and in the air gap. At the interface between the core material and the air gap the field changes. Inductors with air gap will be discussed a bit later.) Using these assumptions the equation becomes a simple sum [51]:

$$n * I = H_1 * l_{mag_1} + H_2 * l_{mag_2} ... = \sum_{i=1}^{n} H_i * l_{mag_i} \tag{4.41}$$

The following figure shows a simple transformer with a (typically iron) core without air gap.

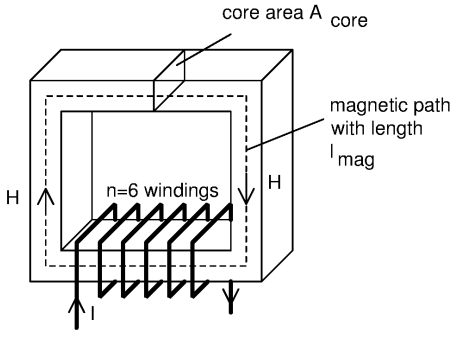


Figure 4.21: A simple inductor with an iron core or a ferrite core.

The core can be magnetized up to a certain limit of the flux density calledB_{sat} . This saturation flux density determines the maximum current the inductor can be used for. (Well, it can be used for higher currents too, but then it doesn't act as a linear inductor anymore.) B_{sat} depends on the core material. Common materials range from 200mT (ferrite) to 1.5T (iron cores).

$$B = \mu_0 * \mu_r * H \tag{4.42}$$

Combining these two equations yields:

$$I_{sat} = \frac{B_{sat} * l_{mag}}{\mu_0 * \mu_r * n} \tag{4.43}$$

The second important parameter is the voltage to be applied. It is a function of the speed of change of the flux Φ, the number of windings (n) and the area bounded by the windings (A_{core}) .

$$\Phi = \int B dA_{core}$$

and

$$V_{indn} = n * A_{core} * \frac{dB}{dt}$$

in this equation the flux density B must be expressed by the current flowing through the windings

$$B = \frac{I * n * \mu_0 * \mu_r}{l_{mag}}$$

This replacement yields:

$$V_{indn} = \frac{n^2 * \mu_0 * \mu_r * A_{core}}{l_{mag}} * \frac{dI}{dt} \tag{4.44}$$

The expression

$$L = \frac{n^2 * \mu_0 * \mu_r * A_{core}}{l_{mag}} \tag{4.45}$$

is called the inductance of the inductor.

Inductor with an air gap: Inductors with a closed core as shown above saturate easily. To increase the saturation current inductors with an air gap are used.

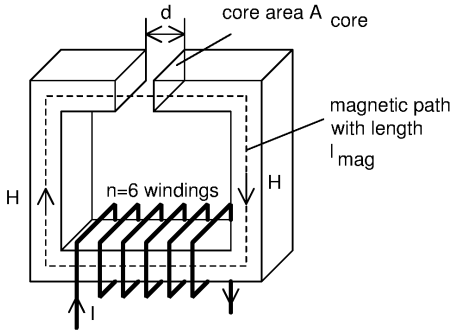


Figure 4.22: Same inductor as before but now with an air gap of length d

This air gap changes the equations a little bit. We have to consider two different fields H_1 with a magnetic path length of $l_{mag} - d$ in the metal core and a different field $H_2$with a magnetic path length of d in the air gap. The flux density remains constant along the whole closed path. This leads to:

$$n * I = H_1 * (l_{mag} - d) + H_2 * d$$

Replacing H by the flux density in the path by the corresponding permeabilities we can again calculate the saturation current.

$$n * I_{sat} = \frac{B_{sat}}{\mu_0 \mu_r} * (l_{mag} - d) + \frac{B_{sat}}{\mu_0} * d$$

(This equation assumes that the relative permitivity of the air gap is 1. If the gap is filled with a different material the corresponding relative permitivity must be added for the gap.)

Reordering the equation leads to:

$$I_{sat} = \frac{B_{sat}}{n} * \frac{(l_{mag} - d) + \mu_r d}{\mu_0 \mu_r} \tag{4.46}$$

Especially for cores with high permitivity $\mu_r d$ will get significantly bigger than $l_{mag} - d$. This means the saturation current increases significantly compared to a core without air gap. The air gap of course also changes the flux density.

$$B = \frac{I * n * \mu_0 * \mu_r}{(l_{mag} - d) + \mu_r * d} \tag{4.47}$$

So this changes the inductance:

$$L = \frac{n^2 * \mu_0 * \mu_r * A_{core}}{(l_{mag} - d) + \mu_r * d} \tag{4.48}$$

If the distance d approaches zero this equation approaches (45) again.

Example: $l_{mag} = 6cm$, $A_{core} = 1cm^2$, $\mu_r = 3000$, $n = 6$, $d = 0$ and $d = 2mm$, $B_{sat} = 0.5T$ (these are typical numbers for a nice little ferrite core inductor used in switch-mode power supplies).

Calculation without air gap: $I_{sat} = 1.32A$, $L = 226\mu H$, Energy storage $0.5 * I_{sat}^2 * L = 199\mu J$

Calculation with air gap: $I_{sat} = 803A$, $L = 2.24\mu H$, Energy storage $0.5 * I_{sat}^2 * L = 722mJ$

So without an air gap the coil would fit the design of a 1A switch-mode power supply while the version with 2mm air gap fits the design of a 500A switch-mode power supply!

The energy of the magnetic field stored in the air gap is much higher than the energy stored in the ferrite core!

4.5.2 Transformers

The most simple way of describing a transformer is the use of the primary inductance (output windings are open), the stray inductance (output windings are shorted), coupling between the windings K and resistance of the windings. The coupling K can be calculated from the inductances measured in open and in shorted operation [11].

$$K_{ind} = \sqrt{1 - \frac{L_{short}}{L_{open}}} \tag{4.49}$$

With data obtained by open and short measurements the equivalen circuit can be drawn.

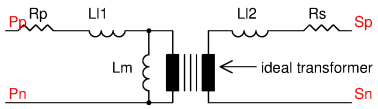


Figure 4.23: equivalent circuit of a transformer

In the equivalent circuit the ideal transformer is assumed to work down to 0Hz and have an coupling K of 1. It has a ratio of N=Np/Ns=Vp/Vs (Np is the number of windings of the primary side, Ns is the number of windings of the secondary side). The minimum operating frequency is determined by inductor Lm. The parameters Ll1, Ll2 and Lm can be calculated:

$$L_{l1} = (1 - K_{ind}) * L_{open} \tag{4.50}$$

$$L_{l2} = (1 - K_{ind}) * \frac{L_{open}}{N^2} \tag{4.51}$$

$$L_m = K_{ind} * L_{open} \tag{4.52}$$

The resistors Rp and Rs must be measured with an ohmmeter.

For circuit simulation in addition the winding capacities and the stray capacity between the primary side and the secondary side may be needed. These capacities are important to determine resonances and RF feed through. Typically the effects of parasitic capacities become dominant at frequencies about one or two magnitudes higher than the operating frequency of the transformer.

Example: $L_{open} = 2.32mH$, $L_{short} = 6.7\mu H$, N=1: $K_{ind} = 0.9985$, $L_m = 2.3166mH$, since N=1 $L_{l1} = L_{l2} = 3.36\mu H$.

4.5.3 Transducers

Today the use of a transducer is somewhat out of date. A transducer is an inductor with a core that is operated in partial saturation. It has a few windings that carry a large AC current. To regulate the current there is a second winding with one or two magnitudes more windings. This second winding carries a DC path current used to drive the core into saturation to lower the AC impedance or out of saturation to increase the AC impedance of the high current path.

The result of this design is the somewhat nonlinear regulation of the AC current. This kind of current regulation was used for power supplies before the advent of silicon power transistors in the 1950s and beginning 1960s. The following figure shows the simplified concept of the power supply of an old Siemens 2002 computer.

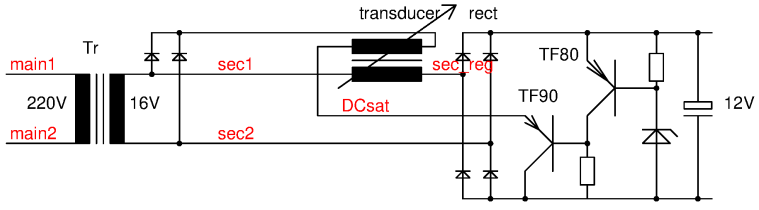


Figure 4.24: Concept of a power supply using a transducer for regulation

The loop is simple: When the output voltage increases TF80 will pull up the base of TF90. This reduces the DC current flowing in the upper windings of the transducer. The core leaves saturation and the inductance of the transducer increases. This kind of regulation requires a certain minimum load at the 12V DC output because the inductance of the transducer only can be changed between 1 and 2 magnitudes. It can't be increased to infinite.

The concept of regulating high power with variable inductances became obsolete with the introduction of switch-mode power supplies in the mid 1960s.

I have no idea if similar concepts still are in use to compensate or regulate complex impedances in a power grid. The transducer offers the advantage that the DC current path can be supplied from a low voltage rail and the regulating transistors don't need to be high voltage devices. But today this may be not required anymore because stacking of thyristors up to 800kV is possible.

4.5.4 Piezzo transducers used for energy transfer

Piezzo transducers normally are used for applications like SONAR or to analyse metal construction elements for cracks. The efficiency of piezzo transducers operated at the resonant frequency can be very high in the range of 50%. Using one transducer as a transmitter and a second one as a receiver efficiencies of 20% are reported [85]. If stray capacities of transformers must be avoided piezzo transducers can be an option to circumvent the use of transformers.

4.6 MOS transistors

MOS transistors are a big family of components. The most simple transistors are used for low voltage applications. In most cases the logic inside a chip is designed using the transistors of the lowest voltage class. The lower the voltage class the smaller the transistors, the lower the capacities and the faster the technology (at least the logic part). Modern technologies offer transistors from about 7nm minimum gate length (This is a drawn gate length. Since the gate is a trench or sometimes called a T-gate the effective channel length is longer! Some companies are already discussing 5nm) to 250nm minimum gate length. The permissible operating voltages for these transistor reach from 500mV to about 2.5V.

To interface the logic inside the chip with the outside world higher voltage classes are needed. Standard voltage swings at the I/O cells are 3.3V and 5V. This requires a second voltage class with minimum gate lengths of 350nm to 650nm. This second class of transistors is fairly simple as well. Of course the gate oxides have to be scaled for the different voltage. [19] states the following equation for the life time of a gate oxide:

$$t_{BD} = \tau_0 * exp(\frac{G * t_{ox}}{V_{ox}}) \tag{4.53}$$

with $\tau_0 = 1 * 10^{-11} s$ and G=350MV/cm = 35V/nm. This equation can be rearranged:

$$t_{ox} = \frac{V_{ox}}{G} * ln(\frac{t_{BD}}{\tau_0}) \tag{4.54}$$

If we want to achieve a life time of 10 years operating at 5V gate voltage this leads to 6.4nm. Since the production of oxides has certain tolerances (production spread and roughness of silicon, planarity errors..) the manufacturers usually design with at least:

$$t_{ox} = V_{gs} * 1nm/0.5V$$

Here for simplicity the factor G was regarded as a temperature independent constant. This is not exactly true. G depends on temperature. Details about the temperature dependence can be found at the discussion of oxide capacitors some pages before and in literature [73] . At lower temperature slightly higher gate voltages are permissible for the same life time.

Above 5V supply voltage this straight forward scaling becomes uneconomical. Here additional tricks are used such as different oxide thickness close to the gate and close to the drain. This leads to special high voltage transistor designs that are discussed in an extra subsection.

As long as we are discussing standard MOS transistors we assume NMOS and PMOS to be following the same process steps. Of course hole mobility and electron mobility differ. But the basic equations as well as the basic techniques of patterning the devices is assumed to be similar. Therefore most of the things common to both polarities is explained at the example of the NMOS transistor.

4.6.1 NMOS transistors

The cheapest way of implementing NMOS transistors is simply placing them in the substrate or a pwell tied to substrate.

The more flexible approach is to place the NMOS transistors inside pwell regions that are isolated from substrate. These pwell regions usually are sitting inside nwell regions (or n-epi regions). This way of isolating the NMOS

transistors from substrate allows building floating circuits.It even is possible to build NMOS circuits operating below substrate potential.

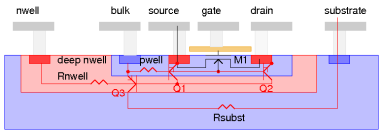


Figure 4.25: Isolated NMOS transistor

Electrical properties of the NMOS transistor To use NMOS transistors in a circuit we need an understanding of what happens physically inside the transistor.

When a positive voltage is applied at the gate the gate is charged. On the opposite side inside the silicon the opposite charges (electrons) get accumulated. This is called the influence charge. Inside the P-doped bulk a thin layer with electrons is created. This layer is called the channel. The channel is connected to the N-doped region at the drain and the source of the transistor. So the channel becomes a conductive path from the drain to the source. As long as the voltage at the drain is close to the voltage at the source the channel is more or less equally thick on both sides of the transistor. So at low Vds the transistor behaves like a resistor that is controlled by the gate charge.

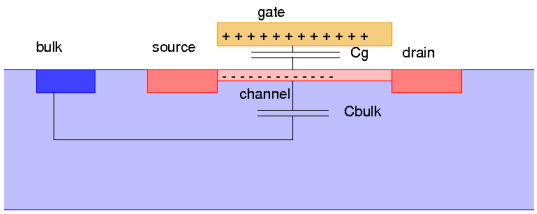


Figure 4.26: Development of a channel inside an NMOS transistor

The voltage available at the top of the silicon depends on the ratio of the capacitors Cg and Cbulk. These capacitors can be regarded as a voltage divider reducing the gate voltage.

$$V_{surface} = V_g * \frac{C_g}{C_g + C_{bulk}} = V_g * k = V_g/n \tag{4.55}$$

k=1/n is called the coupling factor. It depends on the gate oxide thickness and the depletion zone below the channel. (Here comes the relationship to bipolar transistors. In a bipolar transistor Cg is bypassed by the base contact. In stead of capacitively producing a channel the base inside a bipolar transistor is pulled up galvanically until electrons coming from the emitter start to flood the base. So a bipolar NPN transistor can be considered as a degenerated NMOS transistor with k=1and a very leaky gate producing the base current.)

The gate to channel capacity is simply calculated using the oxide thickness and the gate area A.

$$C_g = \frac{A * \varepsilon_{sio_2}}{t_{ox}} \tag{4.56}$$

To calculate $C_{bulk}$the depletion width t_{si}below the channel is required. This can be calculated regarding it as a pn junction.

$$t_{si} = \sqrt{\frac{2 * \varepsilon_{si} * (\phi - V_{bs})}{q * N_b}} \tag{4.57}$$

In this equation ϕ is the junction built in voltage of about 0.6V. q is the charge of an electron. $N_b$is the bulk doping concentration. The channel to bulk capacity becomes:

$$C_{bulk} = \frac{A * \varepsilon_{si}}{t_{si}} \tag{4.58}$$

Combining the equations yields the inverse of the coupling factor n:

$$n = 1 + \frac{C_{bulk}}{C_g} = 1 + \frac{\varepsilon_{si} * t_{ox}}{\varepsilon_{sio_2} * \sqrt{\frac{2 * \varepsilon_{si} * (\phi - V_b)}{q * N_b}}} \tag{4.59}$$

Example:

$t_{ox} = 7nm$, $\varepsilon_{si} = 1pF/cm$,$\varepsilon_{sio_2} = 0.34pF/cm$,$\phi = 0.6V$,$q = 1.6 * 10^{-19}As$,$N_b = 4 * 10^{17}cm^{-3}W = 10\mu m$, $L = 0.35\mu m$.bulk connected to drain and source.

$$C_g = \frac{3.5 * 10^{-8} cm^2 * 0.34 pF/cm}{7 * 10^{-7} cm} = 17 fF$$

$$t_{si} = \sqrt{\frac{2 * 1 * 10^{-12} F/cm * 0.6V}{1.6 * 10^{-19} As * 4 * 10^{17} cm^{-3}}} = \sqrt{\frac{1.2 * 10^{-12} cm^2}{6.4 * 10^{-2}}} = 43nm$$

$$C_{bulk} = \frac{3.5 * 10^{-8} cm^2 * 0.34 pF/cm}{43 * 10^{-7} cm} = 2.77 fF$$

$$k = \frac{17 fF}{19.77 fF} = 0.86$$

Operating the NMOS transistor as a resistor For $V_{ds} < V_{gs} - V_{th}$ ($V_{th}$is called the threshold voltage of the transistor) the patch from the drain to the source acts more or less like a resistor.

$$R_{on} = \frac{L}{W} * \frac{t_{ox}}{\mu * \varepsilon_{sio_2}} * \frac{1}{V_g - V_{th}} \tag{4.60}$$

μ is the mobility of the carriers in the channel. In case of an NMOS transistor these are electrons with a mobility of $\mu_n = 600 cm^2/Vs$.

For more convenience let us abbreviate the expression

$$V_g - V_{th} = V_{gseff} \tag{4.61}$$

Example:
$V_{gseff} = 2V$, $L = 0.35\mu m$, $W = 1\mu m$, $t_{ox} = 7nm$

$$R_{on} = \frac{0.35}{1} * \frac{7 * 10^{-7} cm}{600 cm^2/Vs * 0.34 * 10^{-12} As/(V * cm)} * \frac{1}{2V} = 600\Omega$$

Operation in strong Inversion In strong inversion the transistor usually is operated at $V_{ds} > V_{gseff} = V_g - V_T$. This is called saturated operation because the drain current saturates. It is (more or less) independent of the drain voltage. The drain current follows the square of the effective gate voltage.

$$I_d = k' * \frac{W}{L} * V_{gseff}^2 \tag{4.62}$$

with

$$k' = \frac{\mu * \varepsilon_{sio_2}}{2 * n * t_{ox}} \tag{4.63}$$

n of course depends on the capacitive coupling between the gate and the channel and between the channel and the bulk. Thus it depends on the bulk doping and the bulk biasing. Typical values of n are about 1.2..1.5.

Example:
$V_{gseff} = 1V$, $L = 0.35\mu m$, $W = 1\mu m$, $t_{ox} = 7nm$

$$I_d = \frac{600 cm^2/Vs * 0.34 * 10^{-12} As/Vcm}{2 * 1.2 * 7 * 10^{-7} cm} * \frac{1}{0.35} * 1V^2 = 347\mu A$$

Velocity saturation The speed of the electrons in the channel can not be increased in an unlimited way. There are more and more collisions between the electrons and the atoms. The speed of electrons in the silicon is limited to about $v_{sat} = 10^7 cm/s$. Reaching velocity saturation the current increases linearly with the gate voltage. Transconductance gm does not increase anymore - but current consumption does! So operating a transistor in velocity saturation does not provide any more benefit.

$$I_{dvs} = W * c_{sio} * v_{sat} * V_{gseff} \tag{4.64}$$

The gate voltage at which velocity saturation is found calculates as:

$$V_{gseff_{vs}} = 2 * n * L * \frac{v_{sat}}{\mu} \tag{4.65}$$

Details how to find these equations can be found in [10] in chapter 1.

Operation in weak inversion At low gate voltage the channel vanishes. Using the square model of strong inversion at $V_g = V_T$ the channel disappears and the drain current drops to 0. In practice this is not quite true. The energy of electrons is statistically distributed. The energy barrier in the weak biased channel is very low and allows diffusion of electrons through the channel similar to the current flow found in bipolar devices.

$$I_{dwi} = I_{do} * \frac{W}{L} * exp(\frac{V_{gseff}}{n * k * T/q}) \tag{4.66}$$

The transconductance in weak inversion becomes

$$gm_{wi} = \frac{I_{dwi}}{n * k * T/q} \tag{4.67}$$

The transition between strong inversion and weak inversion is at the point of the characteristic where the gm of the weak inversion equation is equal to the gm of the strong inversion. Note that the weak inversion transconductance is almost independent of any technology parameters. The only parameter remaining is the ratio of the gate capacity and the bulk capacity. Since the bulk doping usually is controlled quite well the gm (and thus the gain of a amplifier) in weak inversion is a very stable parameter.

$$\frac{I_{dwi}}{n * k * T/q} = 2 * k' * \frac{W}{L} * V_{gseff} \tag{4.68}$$

As an additional condition the currents must be equal no matter if we are coming from weak inversion using the exponential characteristic or if we are coming from strong inversion using the quadratic behavior. (Otherwise it would create a discontinuity!)

$$\frac{k' * \frac{W}{L} * V_{gseff}^2}{n * k * T/q} = 2 * k' * \frac{W}{L} * V_{gseff} \tag{4.69}$$

$$V_{gseff_{ws}} = 2 * n * k * T/q \tag{4.70}$$

So the transistion between weak inversion and strong inversion takes place when the effective gate voltage reaches about 70mV. Knowing this transition point we can even determine I_{d0} (although I have no clue if we ever will need it).

$$I_{d0} * \frac{W}{L} * exp(2) = k' * \frac{W}{L} * (2 * n * k * T/q)^2 \tag{4.71}$$

$$I_{d0} = \frac{2 * \mu * \varepsilon_{sio_2} * n * (k * T/q)^2}{t_{ox} * exp(2)} \tag{4.72}$$

Well, this is not really a physical explanation of I_{d0} . We are rather fitting I_{d0} in such a way that we get a smooth transition between weak inversion and strong inversion.

The following figure shows the simulated characteristic of an NMOS transistor. The current is scaled logarithmic. The weak inversion (with an exponential increase of the current with Vgs) can clearly be recognized because log(Id) becomes a straight line. In strong inversion the current increases more slowly (following the square of Vgs. In the plot this is the region where the current seems to increases slower and slower due to the logarithmic scale.

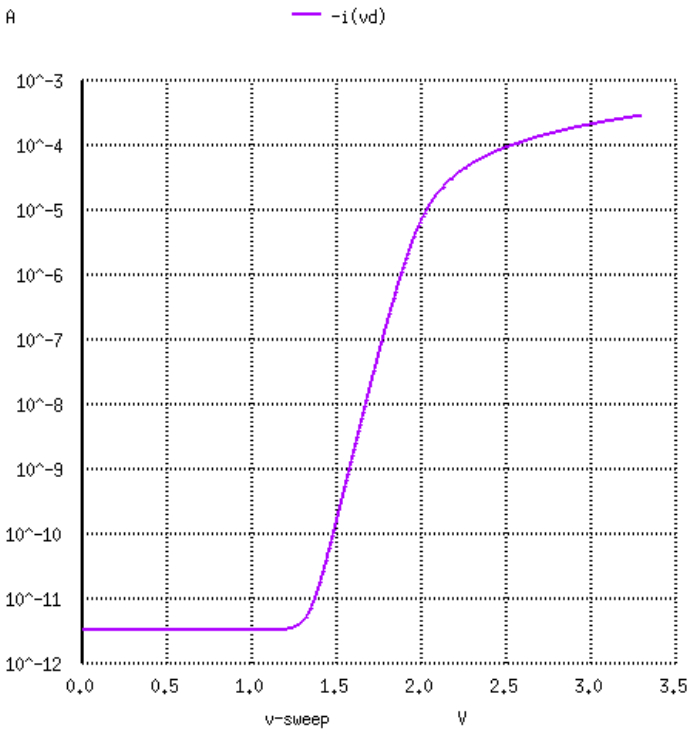


Figure 4.27: NMOS transistor transition from weak inversion to strong inversion.

Technological differences

Classical junction isolated process In standard technologies the transistors are isolated from each other by junctions. The drain and the source contacts are in heavily doped n regions. These N-regions are embedded in a P-bulk region. Ideally this P-bulk is connected to the source. under the gate oxide there is no N-doping because it was masked with the gate poly silicon. The gate oxide between the gate and the channel is thin. Outside of the active area the oxide between silicon and poly or metal is thick.

The following figures illustrate this classical manufacturing process.

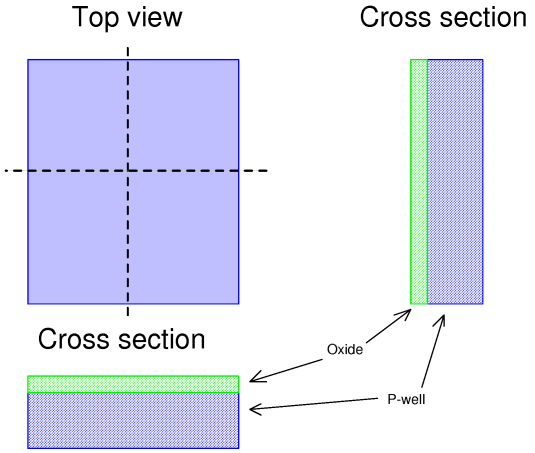


Figure 4.28: Definition of the pwell

Next the oxide is getting removed in the active area of the transistor. Then a thin oxide is grown again on the exposed silicon. To achieve a high quality this oxide is grown with dry hot air. (humid air is faster but not usable for gate oxides because too much hydrogen would be left inside the oxide providing unwanted charges affecting the transistor.)

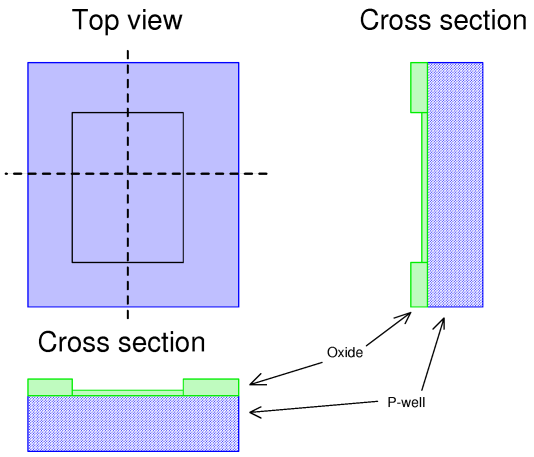


Figure 4.29: active area

Next the poly silicon gate is produced covering part of the active area. This part later will be the channel of the MOS transistor.

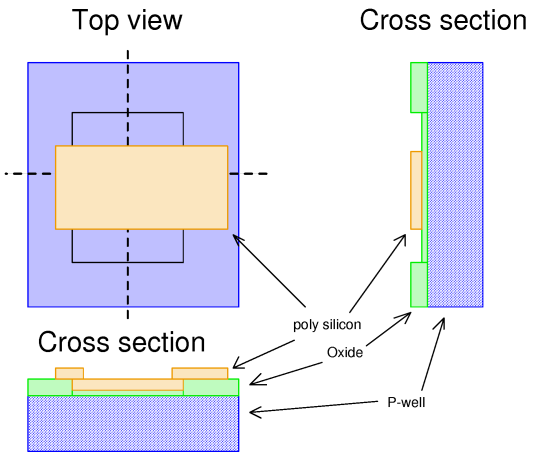


Figure 4.30: Poly silicon (gate poly)

The poly silicon and the thick oxide act as a mask. The implantation of N-doping will only succeed through the thin oxide. Wherever there is thick oxide or poly silicon the doping is masked. Using gate poly as a mask the drain and the source are self aligned to the gate. In the following figure the drain mask is intentionally drawn disalligned to show the effect of the masking by thick oxide and poly silicon. (Often the implant mask is open over the complete transistor!)

As a last step the thin oxide at the drain and source is opened to produce the contacts. Wherever there is no contact additional oxide is grown serving as an isolation.

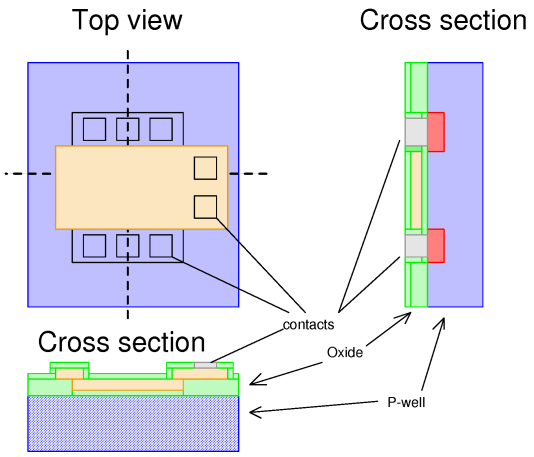


Figure 4.31: Contact of the MOS transistor

Now it is time to apply a voltage at the gate and see where we get the inversion (the channel). The channel

length is defined by the poly silicon. The width of the channel is a bit wider than the active area opening because there also is a little bit of an electrical field that is not fully vertical. In fact we have a big NMOS transistor with a well defined gate oxide plus two very narrow transistors in parallel sitting at the edges having a thicker gate oxide (because the electrical field there is not vertical but has to cross more oxide).

These 'side transistors' are the reason why precision current mirrors must be designed of unity modules. It is important to note that these side transistors having a thicker effective gate oxide and the same bulk doping as the main transistor have a **higher threshold** than the main transistor. So the non ideal effect remains low even in weak inversion and tracks the main transistor.

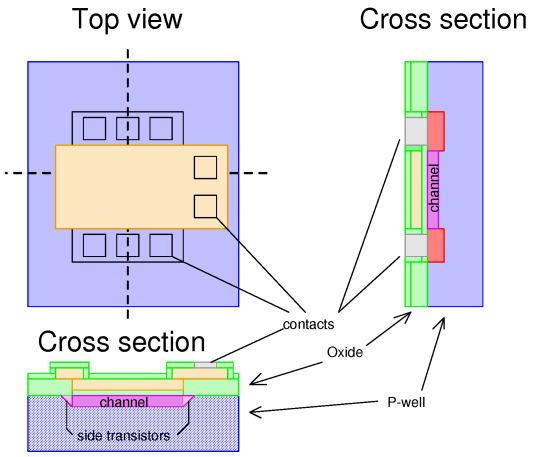


Figure 4.32: Inverted regions with a gate voltage is applied

Short channel transistors with halo implant While the transistor is off ideally the carriers (electrons in case of an NMOS transistor) are on both sides of the channel while in the bulk there are no carriers (holes but no electrons in case of an NMOS transistor). This applies as long as the channel is long compared with the uncertainty of the position of the carriers (electrons). The lower the threshold of the transistor the wider the density function of the electrons laps into the channel. The following figure shows a long channel transistor. The energy barrier the electrons have to overcome (this energy corresponds the threshold of the transistor) depends on the bulk doping. A high threshold means the electron density function decays rapidly (solid line). A low threshold means the electron density function decays slowly (dashed lines).

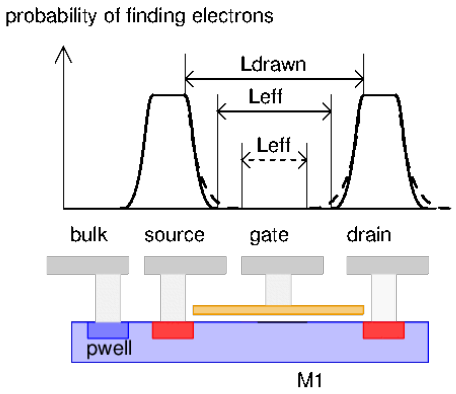


Figure 4.33: Electron density function and effective length of a transistor

The effective length of the transistor is shorter than the drawn channel length. Furthermore the difference between drawn length and effective length depends on the threshold of the transistor.

If additionally the channel becomes very short the density function reaches the other end of the channel. This leads to an increase of the leakage current (short channel subthreshold leakage). The lower the threshold of the transistor the more critical the situation gets. On the other hand using low threshold voltages can not be avoided when the operating voltage of the logic is decreased more and more to allow shorter transistors. This leads to short channel leakage.

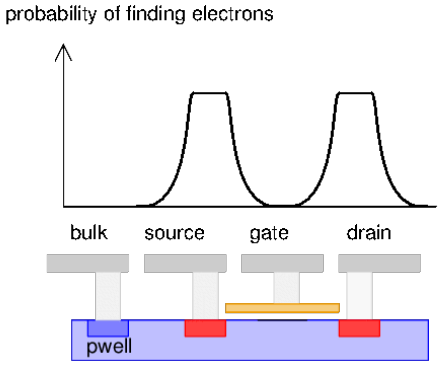


Figure 4.34: Electron density functions of a short channel transistor

With electron density functions touching we will find electrons in the whole channel although the gate is a 0V. The transistor will start to leak. To produce a faster decay of the electron density function the following options are available:

- Increase the bulk doping and the threshold (leads to a slower technology)
- Apply a negative voltage (versus source) to the bulk (reduces speed just like higher bulk doping)
- Increase the bulk doping locally at the drain and source edges.

Today the last option of locally modifying the bulk doping is the most frequent approach for fast logic technologies. This process step is called halo implant. The halo implant uses a P-implant through the same opening as the n implant used for drain and source. So we need no additional mask. halo implant must be scattered a bit more than the n implant to achieve an overlap over the drain and source region.

Often the halo is implanted with a tilt. The tilt can even change rotating the waver. This leads to a halo that depends on the orientation of the transistor. Transistors orientated differently will suffer from poor matching.

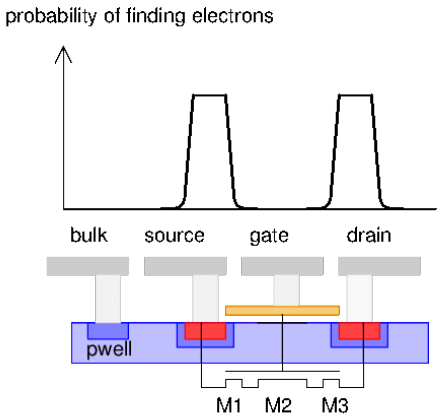


Figure 4.35: short channel transistor with halo implant

The resulting transistor can be regarded as a series of three transistors. M1 and M3 have an extremely short channel but a higher threshold than M2. Typically the halo implant is overlapping the drain and the source by some 10nm. Operating close to the threshold of M1 and M3 almost all of the voltage drop is across M1 and M3. (Operating in saturation M3 being closer to the drain takes most of the voltage drop.) Due to the short channels of M1 and M3 the matching of halo transistors is poor and the early voltage is very low (some V only). 1σ mismatch of the currents can be in the range of +-20% and offset voltages of several 10mV have been reported.

At higher currents the drop over M2 increases thus improving analog behavior at high current densities.

Practical use of halo transistors is restricted to pure switching operation. For analog circuits they are more or less unusable. To illustrate the approximate behavior of such a halo transistor a series of 3 transistors is simulated. M1 and M3 re assumed to be about 50nm long (each of them) and M2 being an analog transistor is assumed to be 1.9μm long. The width is 4μm. Halo transistors typically are used in technologies having 120nm (minimum length) or less with about 3nm or thinner gate oxides.

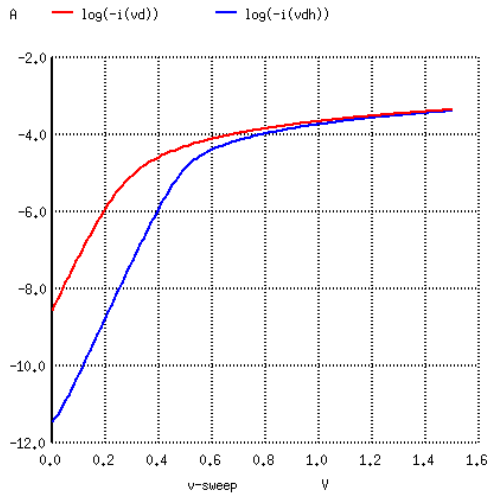


Figure 4.36: Comparison of the drain currents with (blue) and without (red) halo implant

The halo implant shifted the threshold from about 350mV to 450mV without having a significant impact on the drain current at Vgs=1.5V.

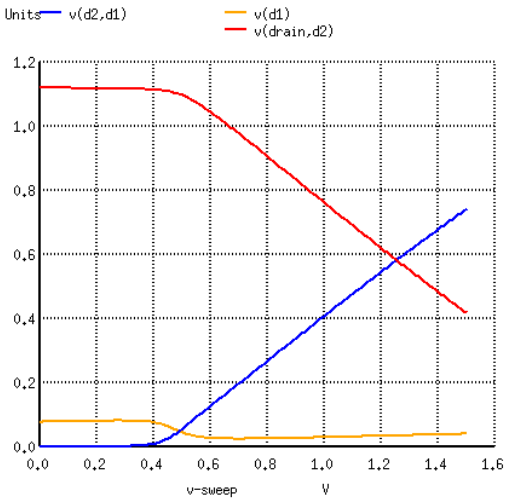


Figure 4.37: Vds of the different parts of the halo implant transistor

At low gatevoltage the inner transistor acts as a resistor. since the current is low the drop is very low too. Most of the voltage drop can be found in the halo implanted region close to the drain (red curve). Above 0.4V the drop over the very short channel affected by the halo decreases while the drop over the inner transistor (blue curve) increases. Since the inner transistor is much longer here the current matching improves significantly at higher current densities. The device can be used for analog applications above about Vgs=0.7V (when the inner transistor reaches a voltage drop of about 200mV).

The source sided halo region has almost no effect (brown curve).

Shallow trench isolation (STI) transistors To achieve higher transistor densities shallow trench isolation was introduced. Shallow trench isolation offers the following advantages:

- Higher logic density (because the spacing between the transistors is reduced)
- Lower capacity at the side walls of the transistor (silicon oxide has a lower ε_r than depleted silicon)
- faster switching speed (e.g. when using transistors with minimum width)

To achieve a better focus of the (still optical) photolithography several planarization steps and oxide back etch steps are needed. This can lead to non ideal effects at the edge of the transistors the analog designer should be aware of. As an explanation a very simplified shallow trench process (STI) is shown.

In step 1 the active area is masked to keep oxygen away from the active are.

In step 2 the wafer is oxidized. Since we are adding atoms the oxidized areas will bulge.

$$Si + O_2 \leftrightarrow SiO_2$$

In step 3 the mask is getting removed. Usually this is a pure chemical process.

Since CMP (chemical mechanical polish) involves mechanical interaction the wafer must be protected adding additional oxide. This happens in step 4.

In step 5 the wafer is getting planarized by a combination of etching and mechanical polishing. Ideally this polishing process should stop when the wafer is planar but the protective oxide is still present.

After polishing the oxide over active is not uniform enough to be used as a gate oxide. It must be removed by a selective etch. Usually this process is slightly overetching thus producing a slightly embossed active area. For good analog performance of the transistors this overetch should be kept as low as possible. In NVM (non volatile memory) technologies a deeper overetch sometimes is desirable (reason will be explained later)

In step 7 the gate oxide is grown. The best process for a well defined gate oxide is dry oxidation. It requires some silicon to react with the oxygen. So it will not fill the overetched regions over the trenches.

In step 8 the gate poly silicon is deposited. It follows the shape of the slightly empossed active area.

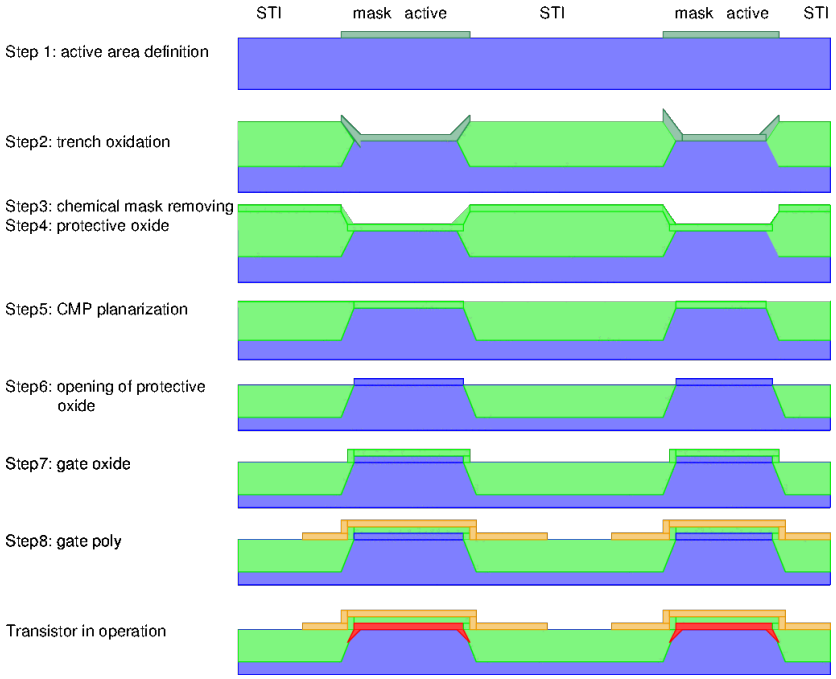


Figure 4.38: Simplified process flow of an STI process

Operating the transistors the inversion will be found in the red areas. The edges of the active area (side walls of the transistors) are exposed to a gate on the top and on the side of the transistor. Furthermore the back etching may have diluted the bulk doping at the edge of the transistors. (Some authors claim that oxidation (of the STI) pushing charges into the silicon has more effect than the overetch itself. [24]) As a consequence the inverted layer forming the channel is not fully homogeneous. This leads to several effects:

1. The threshold of the transistor is lower at the edges. So the real transistor consists of 3 transistors operating in parallel: a wide transistor in the middle and two narrow transistors with reduced threshold at the edges.

2. Besides differences of the gate coupling mechanical stress produced by the STI (shallow trench isolation) changes the bandgap energy locally at the edge of the channel.

3. The electrical field is more concentrated on the edges. Operating the transistor in current saturation this leads to increased tunneling of hot electrons. For NVM (no volatile memories) this effect can intentionally be used.

4. The silicon oxide defining the width of the transistor has a lower capacity than silicon. The technology will become faster and switching losses will be reduced

If a process has to satisfy analog requirements overetching of the protective oxide should be kept as low as possible. If a process is used for NVMs overetching is advantageous but leads to a process that has poor analog performance especially at low currents. A nice way of detecting such effects is shown in [24, 25] measuring the subthreshold slope of Ids. If the side transistor has a lower threshold than the inner transistor this will be visible as a deviation from the ideal slope.

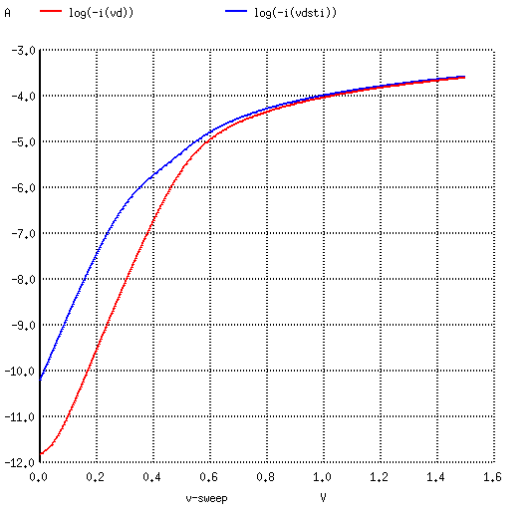


Figure 4.39: Comparison of the subthreshold slope of an ideal NMOS (red) and an NMOS with reduced threshold close to the STI (blue)

The result seen in the simulation above is not quite as significant as in the lecture [24] but still clearly showing that for Vgs<0.4V the side transistors close to the STI are dominating the behavior of the device in weak inversion.

One possibility to avoid the subthreshold hump is to modify the threshold close to the edges by intentionally p-doping the gate [25] (The second proposal of extending the active area beyond the gate area might have reliability issues!). A second solution is to create ring gates for the analog transistors while the logic and NVM is designed with standard rectangular shaped transistors. If halo implant is used such round transistors suffer if the halo depends on the orientation of the transistor.

If we are only interested only in the speed but not at all in the analog properties of the process it is tempting to intentionally maximize the overetch to create U-shaped gates! This maximizes the effective width of the transistor without making the drawn transistor wider. This leads to the concept of a fin-FET. Long term stability of such an extreme process optimization for pure digital applications however may suffer from hot carriers.

Technology considerations: The technology used can have a significant impact on matching and drift of transistors.

Nitride enhanced gates: Basic idea of using a stack of silicon oxide and silicon nitride is to reduce gate leakage and to make the gates more robust. The draw back of having a layer of silicon nitride (Si_3N_4) in the gate is that between the silicon oxide and the silicon nitride the lattice doesn't match. As a consequence there are a lot of open bonds that can collect charges. This leads to an increased drift of the transistor threshold. MOS transistors with an ONO stack (oxide-nitride-oxide) in the gate are barely usable for analog circuits!

4.6.2 PMOS transistors

Usually PMOS transistors are sitting inside an nwell that is tied to a supply voltage. The break down voltage between drain or source and the nwell is low (in the range of some volt) but the break down voltage between the nwell and substrate can be significantly higher than the Vds of the PMOS transistors.

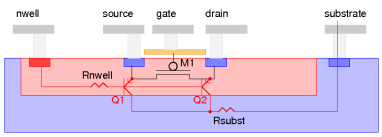


Figure 4.40: PMOS transistor and the most important bipolar parasitic transistors

MOS transistor matching: A lot of work regarding transistor matching was done by Marcel Pelgrom [24] . As long as the channel region is homogeneous the transistor matching depends on oxide thickness, mask accuracy and dopants in the channel. The bigger a MOS transistor the better statistics of errors level out. In addition thin oxides usually can be manufactured with better accuracy. As a rule of thumb we can expect a matching constant of:

$$A = 1mV * \mu m/nm$$

The mismatch of a transistor pair becomes:

$$V_{os} = \frac{1mV * \mu m}{nm} * \frac{t_{ox}}{\sqrt{W * L}} \tag{4.73}$$

This rule of thumb applies to transistors without halo implant and with $t_{oc} > 2nm$.

For transistors with halo implant the matching strongly depends on the current density (see above). Matching in strong inversion can be acceptable for transistors with halo implant while weak inversion matching is poor. If transistors with halo implant become very short the halo implant regions abutt. In this case we can regard them as similar to a homogeneous transistor again. The matching constant of a 100nm halo implant transistor may in fact be OK again. For deep sub-micron technologies it may make sense to compose well matching transistors of many short channel devices in series to exploit this effect of abutting halos.

Rule of thumb current matching of a mature process is always in the range of $1..2\%\mu m$.

4.7 Bipolar transistors and diodes

4.7.1 Bipolar Diodes

Wherever there is a junction the diode comes for free - whether the designer likes it or not! In many cases the diode is a parasitic device rather than an intended device. In designs with several supply domains the circuit designer MUST be aware of the diodes. There is one (partial) exception to the rule: If you can afford silicon on insulation (usually a process similar to the one shown before featuring a burried oxide, but there have been other ones like SOS - Silicon on Sapphire - as well.)

Basic equations As long as there is no external voltage applied we have to consider the charges inside the lattice of the semiconductor and the generation (and recombination) of carriers. Silicon (as well as Germanium) has 4 valence electrons. Inside the lattice every silicon atom is attached to 4 neighbor atom in a tetraeder like structure. Replacing one of the Silicon atoms by Phosphor or Arsen introduces an additional electron that can be mobilized easily (by thermal energy). Once the electron is mobilized we can regard the remainder of the Phosphor or Arsen as a positive ion sitting in the lattice while the compensating charge of the electron is hoovering around in proximity of this charge (some hundred lattice periods around the ion). In other words N-doped material means we have positive charges in the lattice with the corresponding cloud of electrons moving around. Vice versa P-doping (usually Boron, Indium or Gallium) means we an electron too many sitting in the lattice (completing the bond) and a missing electron in the conducting band. This leads to a negative charge of the lattice and a so called defect electron or hole in the conduction band.

At the interface between P-doped material and N-doped material we have a region where the electrons are pulled into the region with the positive lattice charges and the holes are pulled to the region with the negative lattice charges. In the middle we have the depletion zone that is neither holding a significant number of holes nor a significant number of electrons.

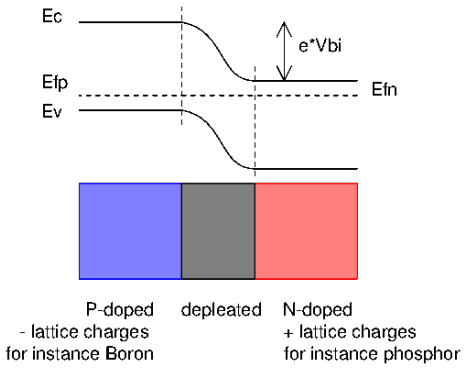


Figure 4.41: PN diode in equilibrium

As long as no external voltage is applied the energy of the valence electrons depends on the charges sitting in the lattice. On the left side we have negative charges in the lattice that will reject the electrons. This corresponds an increase of the energy of the electrons in the conducting band.

On the right side the lattice has positive charges attracting the electrons. This corresponds a lower energy. The electrons will tend to move into the region with the lowest energy of the conducting band.

The difference of energy can be expressed by a built in voltage Vbi multiplied with the electron charge e. At room temperature the built in voltage of a typical silicon diode is in the range of 0.7V. (A typical germanium diode has about 0.2V). The built in voltage depends on the doping levels on both sides of the junction (N_a and N_d) and on the number of intrinsic carriers n_i. Note that $n_i$has a strong dependence on the temperature of the semiconductor!

$$V_{bi} = \frac{k * T}{e} * ln(\frac{N_a * N_d}{n_i^2}) \tag{4.74}$$

To achieve a significant current flow the energy of the electrons on the right side must be increased and the energy on the left side must be lowered. This can be done connecting the negative node of a voltage source to the N-doped material (called the cathode of the diode) and the positive node to the P-doped side (called the anode of the diode).

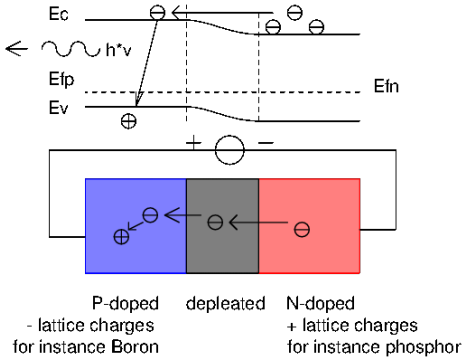


Figure 4.42: PN diode getting forward biased

Applying an external voltage cancels the energy difference between both sides of the diode and the electrons can diffuse from the right side to the P-doped left side. On the other side of the depletion (in the P-doped area) zone the electrons become the minority carriers and will recombine with the holes. This can also be described as falling from the conductive band to the lower energy of the valence band. The energy difference can be emitted as optical photons. The efficiency of producing optical photons depends on the semiconductor material. Silicon has a very poor efficiency because the momentum of the holes and the momentum of the electrons differ. This leads to an indirect recombination. The momentum difference must be converted into phonons.

Direct recombination takes place when the momentums of the holes and electrons are equal. In case of direct recombination the intermediate step of converting momentum differences into phonons is not required. Direct recombination produces magnitudes more photons than indirect recombination. Usually LEDs (light emitting diodes) are built using III-V semiconductors such as GaAs.

The energy of the electrons in the conducting band has a more or less Gaussian distribution. Therefore the current does not start abruptly increasing the voltage applied. In stead following the Gaussian distribution far from its maximum we will observe an exponential increase of the current with increasing forward voltage. The number of electrons reaching the left side of the depletion zone becomes

$$n_p = n_{n0} * exp(\frac{-e * (V_{bi} - V_a)}{k * T}) = n_{n0} * exp(\frac{-e * V_{bi}}{k * T}) * exp(\frac{e * V_a}{k * T}) \tag{4.75}$$

In this equation $V_a$is the voltage applied with the external voltage source. n_{n0} is the density of electrons on the right side of the depletion zone. $n_p$is the density of electrons on the right edge of the P-doped zone (before recombination starts). If we replaces the term

$$n_{p0} = n_{n0} * exp(\frac{-e * V_{bi}}{k * T}) \tag{4.76}$$

we can simplify the equation.

$$n_p = n_{p0} * exp(\frac{e * V_a}{k * T}) \tag{4.77}$$

with n_{p0}being the electron density on the P-doped side without applying an external voltage.

Similarly for holes on the left edge of the N-doped material we find:

$$p_n = p_{no} * exp(\frac{e * V_a}{k * T}) \tag{4.78}$$

Assuming the thermal energy is sufficient to mobilize the doping (This usually is the case at temperatures higher than about 200K) the number of majority carriers corresponds the concentration of the doping.

$$n_{n0} = N_d \tag{4.79}$$

$$p_{p0} = N_a \tag{4.80}$$

And the product of holes and electrons is equal the square of the intrinsic carrier density.

$$n_{n0} * p_{n0} = n_i^2 \tag{4.81}$$

$$n_{p0} * p_{p0} = n_i^2 \tag{4.82}$$

The total current flowing consists of holes and electrons.

$$J = (\frac{e * D_p * p_{n0}}{L_p} + \frac{e * D_n * n_{p0}}{L_n}) * (exp(\frac{e * V_a}{k * T}) - 1) \tag{4.83}$$

To make the equation a bit user friendly let us replace $p_{n0}$by $n_i^2/N_d$and $n_{p0}$by n_i^2/N_a. Thus we get

$$J = e * n_i^2 * (\frac{D_p}{L_p * N_d} + \frac{D_n}{L_n * N_a}) * (exp(\frac{e * V_a}{k * T}) - 1) \tag{4.84}$$

Or even better usable:

$$J = J_s * (exp(\frac{e * V_a}{k * T}) - 1) \tag{4.85}$$

with

$$J_s = e * n_i^2 * (\frac{D_p}{L_p * N_d} + \frac{D_n}{L_n * N_a}) \tag{4.86}$$

$D_n$and $D_p$are the diffusion constants of electrons and holes. $L_n$and $L_p$are the diffusion widths before recombination takes place.

These equations apply to weak injection where the diffusion of carriers defines the current. If the injection increases recombination inside space charge region will reduce the number of carriers crossing the space charge region. This leads to a non ideality factor η.

$$J = J_s * exp(\frac{e * V_a}{\eta * k * T}) \tag{4.87}$$

η ranges from 1 (weak injection) to 2 (recombination inside the space charge region dominates).

The diffusion capacity of a diode calculates as:

$$C_d = \frac{e}{2 * k * T} * (I_{p0} * \tau_b + I_{n0} * \tau_n) \tag{4.88}$$

If the voltage applied to the diode is getting reversed minority carriers have to be pulled back to empty the space charge region and the adjacent silicon volume that was flooded with minority carriers. This leads to a turn off delay consisting of a storage time and a current decay time. During the storage time the reverse current is constant.

$$t_s = \tau_{p0} * ln(\frac{I_F}{I_R}) \tag{4.89}$$

In reverse polarity the junction capacity depends on the width of the depletion zone. The depletion zone width is a function of the doping and the blocking voltage. Usually diodes must be designed for a target break down voltage. In silicon we can approximately allow a peak of the field strength of $60V/\mu m$ [26]. Other authors state lower break down voltages [28] in the range of $30V/\mu m$. Worst case the diode has a high doping on one side and a low doping on the other side. This leads to a triangular field strength distribution. Thus neglecting the built in voltage the width of the depletion zone for a given break down target must be designed as

$$d = 2 * |V_{br}|/E_{br}$$

Example: A 100V Silicon diode must have a depletion zone width at break down of $d = 2 * 100V * \mu m/30V = 6.6\mu m$.

In practical devices the field is not homogeneous. The field in corners can have a significantly higher field strength. Therefore it is good practice to design with a factor 2 lower values than stated by [26]. Thus classical 100V diode usually have depletion zones of about $6\mu m$ to $7\mu m$.

This distance calculation defines the minimum size of high voltage components. Semiconductor devices with a snap back behavior (transistors) should be designed even more conservative!

Knowing the maximum field strength and the target break down voltage the doping level on the low doped side can be defined [28].

$$N_D = \frac{E_{br}^2 * \varepsilon_0 * \varepsilon_r}{2 * e * (V_{br} - V_{bi})} \tag{4.90}$$

In silicon $\varepsilon_r = 12$. Note that e has a negative sign and V_{br} is negative as well in reverse operation. The width of the depletion zone becomes

$$l = \sqrt{\frac{2 * \varepsilon_0 * \varepsilon_r * (V_b - V_{bi})}{e * N_D}} \tag{4.91}$$

If we set the operating voltage equal the break down voltage ($V_b = V_{br}$) we get the break down depletion width of the diode. Replacing $N_D$and using the actual reverse operating voltage we can calculate the width of the depletion zone without needing the doping level.

$$l = \frac{2}{E_{br}} * \sqrt{(V_b - V_{bi}) * (V_{br} - V_{bi})} \tag{4.92}$$

Example: We use a 100V silicon diode ($V_{bi} = 0.6V$) designed for a maximum field strength of $30V/\mu m$ at a blocking voltage of 20V. This leads to

$$l = \frac{2}{30V/\mu m} * \sqrt{(-20V - 0.6V) * (-100V - 0.6V)} = 3.035\mu m$$

The width of the depletion zone follows the square root of the voltage V_b applied.

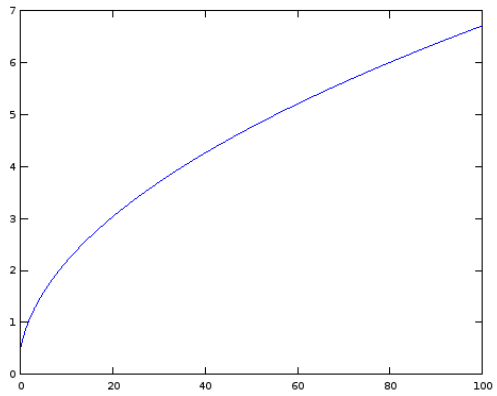


Figure 4.43: Width of the depletion zone of a 100V diode as a function of the blocking voltage. (width in μm,$V_b$in V)

This can be used to build a voltage dependent capacitor capacitor. The capacity calculates as

$$C = A * \varepsilon_0 * \varepsilon_r * 1/l(V_b) \tag{4.93}$$

Choosing $A = 1\mu m^2$we get between 0.2fF at 0V and 0.02fF at 100V

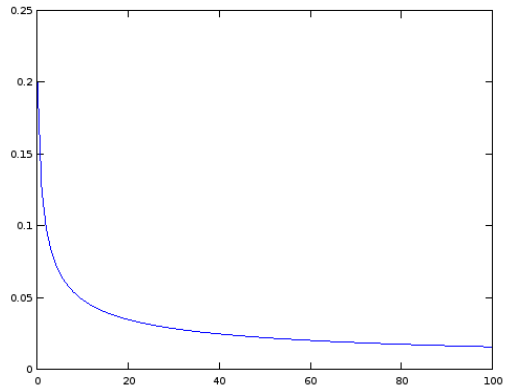


Figure 4.44: Junction capacity of a 100V diode in $ff/\mu m^2$.

Zener diode: The zener diode is operated in reverse polarity with a limited current. At a certain field strength the diode breaks down and limits the voltage over the junction. This break down is determined by a combination of 2 effects:

- Tunneling
- Avalanche break down

Standard discrete zener diodes usually have specifications of $\pm 3\%$ for about 5V break down voltage. Typical temperature coefficients are in the range of $\pm 0.02\%/K$. As soon as the required zener voltage deviates from about 5V the performance suffers. The following plot shows temperature coefficients (in %/K) of a standard precision zener diode versus the zener voltage.

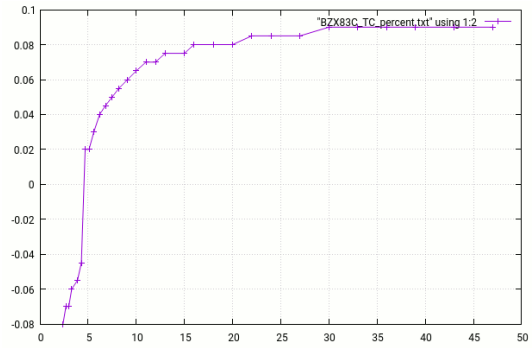


Figure 4.45: Temperature coefficient in %/K versus zener voltage [68]

This relationship is very similar for almost all zener diodes because it relates to basic physical properties of silicon.

There are some integrated "zener diodes" that consist of a zener diode plus a standard diode (in forward operation) to tweak the temperature coefficients. The negative temperature coefficient of the diode operating in forward direction is intended to compensate the positive temperature coefficient of a zener diode of about 7..8V.

Integrated zener diodes usually are not the components the process is optimized for. In most integrated circuits processes the zener diodes more or less perform like standard zener diodes. Even worse in many processes the zener diodes are components of minor interest. Typical examples are the base-emitter surface zener diodes that come for free in most bipolar processes. The junction is optimized for the performance of the NPN transistors but not for the characteristic of the zener diodes! Such BE-diodes used as zener diodes offer typical break down voltages in the range of 7V $\pm 0.5V$ and temperature gradients in the range of 0.04%/K. In addition surface zener diodes are susceptible for contamination and show long term drift of the break down voltage and the small signal resistance.

Making things worse a zener diode is more or less a combination of two different physical effects:

- Tunneling through a very narrow depleation zone usually has a negative temperature coefficient
- avalanche effects have a positive temperature coefficient
- avalanching produces noise that in extreme cases can reach up to $1V_{pp}$for a 7V surface zener diode

The zener diode achieves the best temperature stability if the temperature coefficients of the tunneling and the avalanching exactly cancel - which is rarely the case!

In some processes the junction acting as a zener diode can be moved away from the surface. This is called a burried zener diode. Burried zener diodes typically operate close to 5V and have about one magnitude less avalanching noise than 7V zener diodes. Long term drift (1000h operating time) can be expected between +100mV and +500mV depending on the current the diode is operated with. Due to lattice defects caused by the avalanching part of the current the small signal resistance can change by about one magnitude. (I have seen cases with an increase of the small signal resistance changing from 50Ω end of production to 800Ω after 1000h operation at 80°C.)

On chip zener diodes can't be recommended as a voltage reference unless the process is especially tweaked for the zener diode. (Tweaking the process for the zener often has negative side effects on the performance of other components on the chip. For this reason this usually isn't done in standard semiconductor processes. Some NVM processes do offer zener diodes that are exactly tweaked for the requirements of the memory.)

On ICs zener diodes mainly are used for ESD protection and for gate protection. Using them for voltage references normally is avoided because a bandgap offers better performance.

Substrate diode: This diode is present in all junction isolated processes. Assuming a P-substrate the substrate diode is present wherever there is an N-region embedded in the substrate.

4.7.2 Vertical NPN transistors

Before the advent of CMOS technologies the NPN transistor was the workhorse of integrated circuit design. Usually NPN transistors are designed as veritcal transistors. This means the current is flowing vertically from the emitter through the base into the collector. To carry the current back to the surface the collector holds a highly n-doped burried layer. In most cases the burried layer is connected to the metal contact by the higly n-doped sinker.

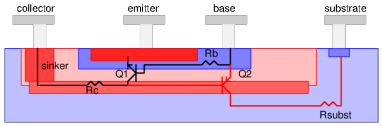


Figure 4.46: Vertical bipolar transistor

Q1 is the vertical bipolar transistor. As long as the collector voltage is higher than the base voltage the parasitic substrate PNP transistor Q2 is inactive. For a minimum size device the base resistance Rb usually is in the range of 50 to 300 Ohm. The collector resistor Rc of a a minimum NPN usually is in the range of 30 to 300 Ohm. In some cases the sinker is omitted to reduce the size of the transistor. In this case the collector resistance can increase up to some kΩ.

Design considerations of an NPN transistor Main targets of the design of an NPN transistor are:

1. current gain of the transistor (B=Ic/Ib)
2. Maximum collector-emitter voltage Vce
3. Minimum silicon real estate

The gain of the transistor mainly is a function of the ratio of the emitter doping and the base doping. Usually the emitter is at least 2 decades higher doped than the base. Additionally the base should be kept as thin as possible to minimize recombination of electrons in the base layer (because this recombination would reduce the gain). As a consequence the base-emitter blocking voltage V_{BEr} usually is in the range of 5..8V. Driving the base emitter junction into zener break down produces hot electrons and immediately (within fractions of ms) reduces the gain of the transistor. This can be regarded as a permanent damage. In some cases the damage can patially be annealed by hot storage. (Often observed after light ESD damage).

The collector-base break down depends on the area available between the base layer and the burried layer to accommodate the collector depletion zone. So the thickness of the epitaxy (together with the doping used) minus the base thickness defines the vertical break down voltage V_{CB}. Once the break down current starts to flow it will pull up the base. The transistor starts to open partially. The break down voltage will snap back some V to some 10V. Typically the spacing between base and burried layer is chosen about 100nm/V.

The same spacing (or even more as a security margin) is needed for the lateral distance from the base to the sinker. Furthermore the sinker width roughly follows 1.5 times the depth. The depletion zone between sinker and substrate again requires about 100nm/V. In modern technologies the size of the sinker (together with the required spacings needed for the depletion zones) is dominant. Shrinking NPN transistors to sizes less than about 1000$\mu m^2$is barely possible (5V transistors to 10V transistors). Higher voltages need even bigger sizes (40V transistors require 3000 to 10000μm^2)

The limitation of scaling now restrict bipolar transistors to special applications that still justify this size (bandgaps, RF input stages, low offset differential amplifiers that can not be chopped).

Forward active operation In forward active operation the base voltage is about 600mV higher than the emitter voltage and the (inner) collector voltage is higher than the base voltage. The base-emitter junction is conducting.

Since the emitter is higher doped than the base more electrons will diffuse from the emitter into the base than holes will be traveling from the base to the emitter. In a well designed bipolar transistor most of the electrons will not recombine but travel to the collector-base junction. There the electric field caused by the higher voltage of the collector will soak the electrons into the collector. Assuming a thin base the current gain

$$B = \frac{I_C}{I_B} = \frac{\mu_n * n_e}{\mu_p * n_p} \tag{4.94}$$

This equation is extremely simplified because it neglects the following contributions to the gain:

- recombination of electrons in the base is neglected
- The emitter usually is significantly smaller than the base region. So the electrical field in the base-emitter junction is far from homogeneous

So the equation above can be regarded as a rough guess, not more!

The emitter current calculates as:

$$I_E = I_{ES} * (e^{\frac{V_{BE}*e}{k*T}} - 1) \tag{4.95}$$

The emitter current is the sum of the base current and the collector current.

$$I_E = I_C + I_B = (B+1) * I_B = \frac{B+1}{B} * I_C \tag{4.96}$$

Building amplifiers the current of highest interest is the collector current. Defining

$$I_{CS} = \frac{B}{B+1} * I_{ES} \tag{4.97}$$

the collector current becomes:

$$I_C = I_{CS} * (e^{\frac{V_{BE}*e}{k*T}} - 1) \tag{4.98}$$

To calculate the transconductance we are interested in the derivative of the collector current.

$$\frac{dI_C}{dV_{BE}} = I_{CS} * e^{\frac{e}{k*T}} * \frac{e}{k*T} \tag{4.99}$$

During operation the base emitter voltage $V_{BE}$is much higher (typically 600mV) than the thermal voltage $\frac{k*T}{e}$ (26mV at room temperature). So the above equation can be approximated with less than a percent of error using:

$$\frac{dI_C}{dV_{BE}} \approx \frac{I_C}{V_T} = \frac{I_C * e}{k*T} \tag{4.100}$$

Very similar to the equation of a MOS transistor in weak inversion! Only the coupling factor representing the ratio of the gate and the bulk capacity is missing (replaced by 1). Well, we can regard a bipolar transistor as a MOS transistor in weak inversion with perfect coupling (at the cost of having a base current).

Saturated Operation Different from MOS transistors the expression saturated operation of a bipolar transistor refers to the saturation of the collector-emitter voltage at about 200mV. In saturated operation the base-collector diode starts to get conducting. As a consequence there are holes injected into the collector. The gain of the transistor decreases because we are loosing holes into the collector. For this effect the voltage inside the transistor - not outside at the pin - is decisive.

Turning off a saturated bipolar transistor leads to a significant turn off delay. The delay is caused by the holes in the collector region that either have to recombine or must be soaked back out of the base. Holes not soaked back and not recombining will at turn off start to travel to the base emitter junction. The presences of holes (coming from the collector when the collector voltage starts to increase) act like a base current keeping the transistor on. Typical turn off delays for AF (audio frequency) transistors can be in the range of hundreds of ns to some μs! In the 1960s and 1970s it was a common practice to dope fast switching transistors with gold. The gold accelerates recombination of minority carriers because it offers an energy level approximately in the middle of the bandgap. The disadvantage of gold doping is a lower gain and higher noise due to more frequent generation and recombination of minority carriers. Factories contaminated with gold can not be used for high performance analog chip production. For the factory the introduction of gold doping is a one way street! Once a production line is contaminated with heavy metal (gold etc.) there is no more way to clean it!

Modeling of turn off delay after saturation of a bipolar transistor is very difficult because the life time of minority carriers plays a dominant role. Turn off delay (SPICE parameter tr - ideal reverse transit time) and turn on delay (SPICE parameter tf - ideal forward transit time) often are determined experimentally for one operating point only. In most cases tr is about 10 times higher than tf. Often the models dramatically underestimate the turn off delay tr for low current densities! [18]

Reverse operation In principal a bipolar transistor can be operated in reverse mode as well. In reverse operation the collector (usually low doped) acts as an emitter and the emitter (the high doped side) acts as a collector. Reverse current gain B_R usually is below 1 (Typically 0.3). For I^2L logic the collector doping can be increased to reach reverse gains in the range of 3 to 8. Since this optimization increases the collector doping I^2L logic usually can only be combined with low voltage bipolar technologies (Up to about $V_{CE}$of 5..15V).

Reverse operation may be extremely interesting if a very low saturation voltage is required. The bipolar transistor has two junctions. In most technologies the base is higher doped than the collector while the emitter is higher doped than the base. As a consequence the forward voltages of the resulting diodes differ. Approaching forward saturation the base-collector diode opens before the collector reaches 0V. Operating the same transistor in reverse saturation mode the base-emitter diode (that now has taken the place of the former base-collector diode) remains off even if the emitter voltage (that now acts as a collector) reaches 0V. This behavior can be used for bipolar chopper amplifiers [36].

4.7.3 High Voltage NPN transistors

The simple transistor shown before has a low doped collector. Where the emitter trace or the base trace crosses the collector region the metal will act as a parasitic gate. This gate can create a P-channel from the base to the isolation when the voltage between collector and emitter gets too high. Typically this threshold (depending on collector doping and oxide thickness) is in the range of 10V to 30V.

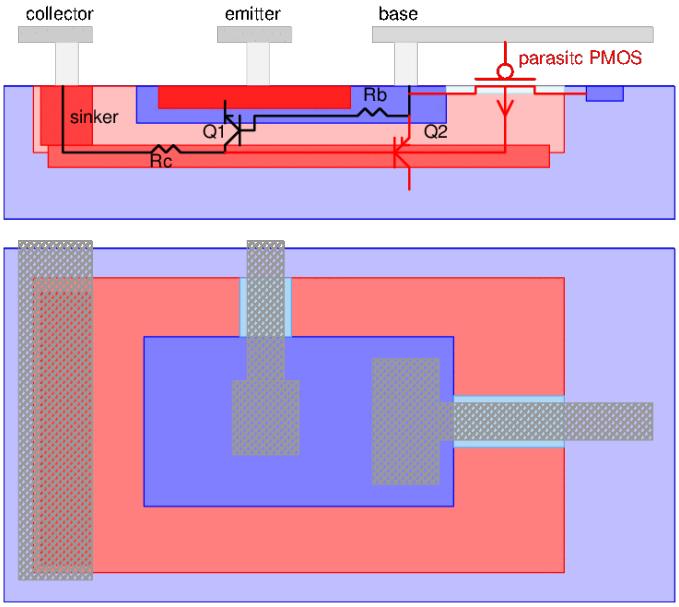


Figure 4.47: Parasitic P-channel MOS in a low voltage transistor

Fig.4.7.3.1: Parasitic P-channel MOS in a low voltage transistor

To shift the threshold of the parasitic P-channel metal gate MOS transistor to higher levels the surface of the transistor can be implanted with a light N-doping (VT-implant). As long as there is no mix with transistors without this threshold shift implant needed this implant can be done without a mask. So the additional process cost is low.

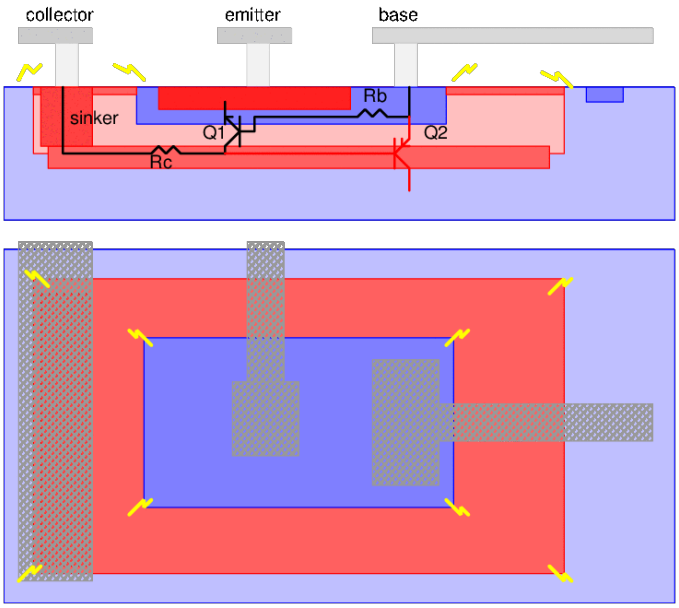


Figure 4.48: Threshold shift (VT) implant

Fig.4.7.3.2: Threshold shift (VT) implant

The disadvantage of the VT-implant is an increase of the electrical field strength at the edge of the base and at the edge of the P-isolation surrounding the transistor. So this method is limited to about 40V to 50V. To reach higher voltages the VT-implant has to be removed again. In stead of implanting without a mask the base must be surrounded by a N-doping that does not touch the base or the P-isolation. Typically the emitter mask is used to create a channel stop ring.

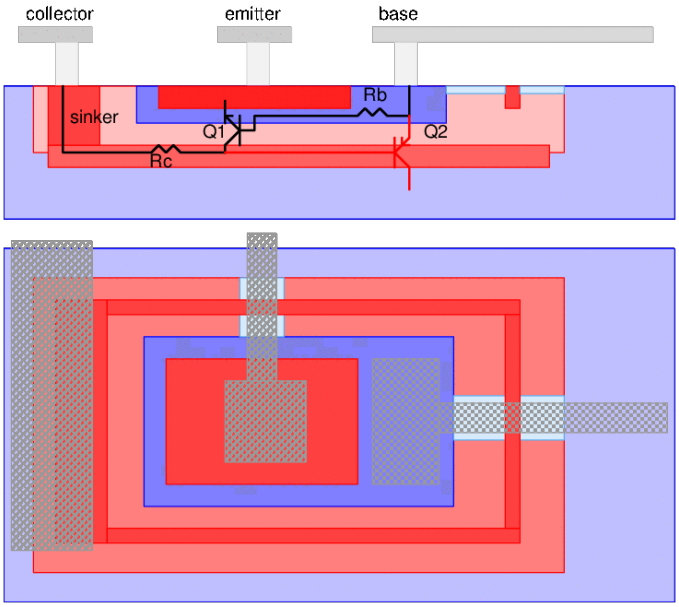


Figure 4.49: Transistor with channel stop ring and without VT-implant

This approach replacing VT-implant by a channel stop interrupts the parasitic PMOS transistor without the penalty of an early break down due to higher surface doping. The limiting factor now becomes the electrical field close to the surface of the silicon. These devices usually can handle a $V_{CES}$of 60V to 70V. To further increase the voltage capability of the transistor field plates are needed to smoothen the shape of the electrical field close to the silicon surface. These field plates are needed at the edge of the base and at the edge of the P-isolation. In some of the older technologies these field plates were built using metal 1 and metal 2 had to be used for the wires connecting the pins of the high voltage transistors. Modern technologies typically use poly silicon field plates (Poly over field oxide).

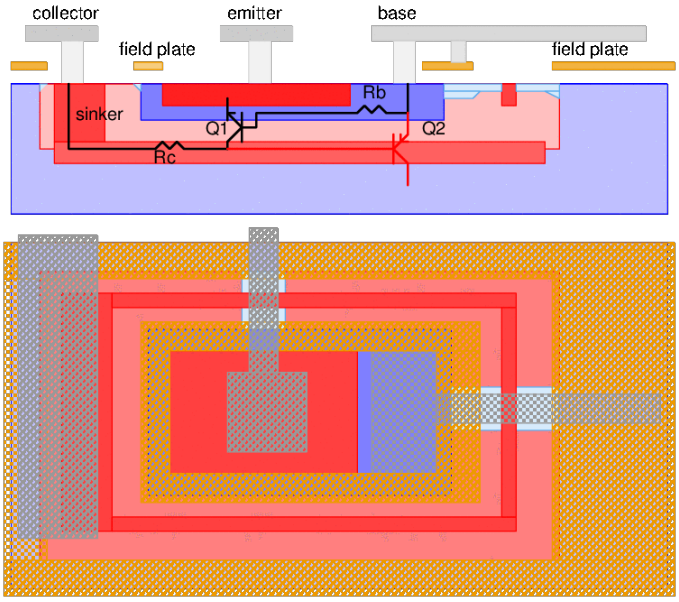


Figure 4.50: High voltage transistor with channel stop and field plate

Adding the field plate break down voltages up to $V_{CES} = 150V$ have been reported. Further pushing the limit rounding the corners in stead of using rectangular base and collector regions is common. Adding P- rings around the base and the P-isolation still is possible too but this adds masks to the process that are specific to the high voltage components only (additional cost).

Besides taking care about the parasitic MOS transistors and surface break down the vertical distance between the base and the burried layer must be adjusted accordingly to prevent a vertical break down. This increases the

thickness of the epitaxy (and the process time tor this production step).

For high voltage minimum NPN transistors the area needed for field plates and channel stop rings is dominant.

4.7.4 NPN power transistors

The saturation currents $I_{CS}$and I_{ES} strongly depend on the thermal generation of minority carriers and on the statistical distribution of the energy of the electrons on both sides of the base-emitter junction. As a consequence the base-emitter voltage required for a constant collector current decreases with about -2mV/K. As a consequence the hottest center of a power transistor tends to drain more current than the colder edge of the transistor. Transistors designed to dissipate a high power must have emitter resistors to balance the current and prevent thermal run away of the center of the transistor. Bipolar power transistors found on integrated circuits either are designed for pure switching (So they either carry current but have no voltage drop or have voltage drop but no current is flowing) or are composed for many transistors in parallel with emitter degradation resistors to balance the current.

Transistors designed for pure switching application do not tolerate a current limiting short circuit protection. The short circuit protection MUST turn off the transistor.

Transistors designed for class A operation (current flow and voltage drop found simultaneously) typically have a voltage drop of 100mV to 200mV over the emitter degradation resistor at their nominal operating point. In extreme cases the emitter resistors can even be designed higher resistive at the anticipated hot spot and lower resistive at the anticipated colder edge of the transistor.

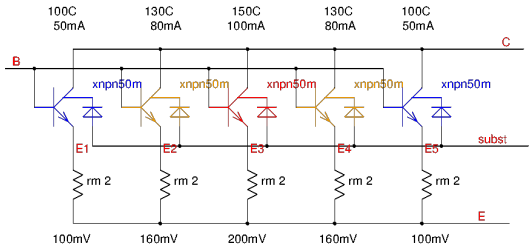


Figure 4.51: Example of a power transistor with already significant current concentration at the hot spot

The resistors in the emitter can be implemented in various ways. A very elegant method is to use the emitter diffusion itself.

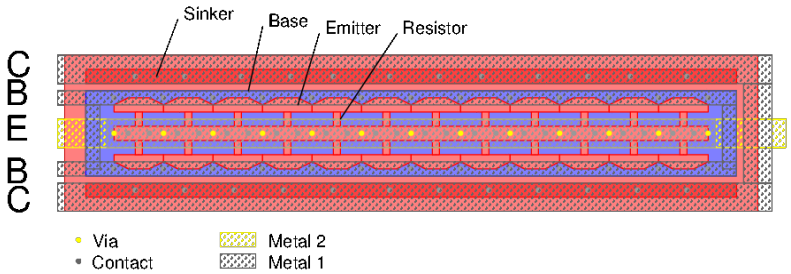


Figure 4.52: Layout example of a power transistor using the emitter diffusion for the emitter resistor.

A base resistor network can additionally improve thermal stability. But usually the current gain B increases with temperature making base resistor symmetrization much less effective than emitter degradation.

Power transistors designed for high voltage have channel stop rings and guard rings as discussed before. This is not shown here to focus on the essentials.

4.7.5 Lateral NPN transistors

In most cases lateral NPN transistors are activated unintentionally. Only very few technologies use lateral NPN transistors as intentional components. The reason is that most technologies are based on a P-substrate. To build an NPN transistor there must be a second P-region that is isolated from substrate. So it must be embedded in an N-well. As soon as the process offers the N-well it doesn't make sense anymore to accept the poor performance of a lateral transistor because the vertical NPN is already available.

Unintentional lateral NPN transistors can be wherever an n-region is pulled below the potential of the surrounding P-region. Then the negative polarized N-region becomes the emitter and surrounding N-regions with higher potential become parasitic collectors. A typical case is an output stage (NMOS or NPN) driving an inductive load that is pulled below ground by a flyback pulse.

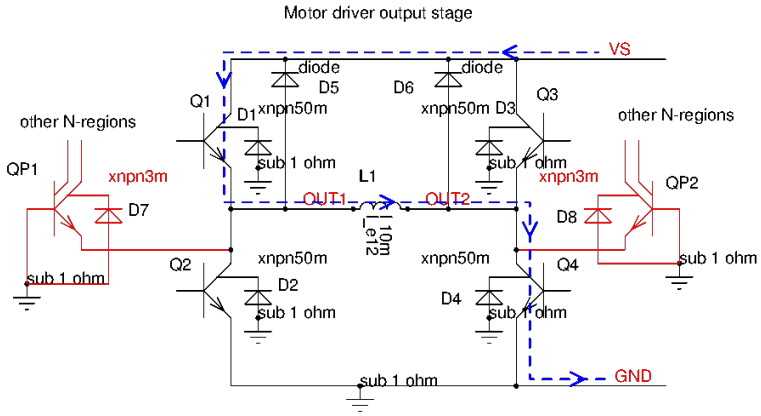


Figure 4.53: Motor bridge building up current

As long as the current flows through the transistors in forward direction the parasitic transistors are not activated. As soon as Q1 turns off because the current exceeds the target value (chopping regulation) or if we turn off Q1 and Q4 because we want to turn on Q2 and Q3 (polarity change) the inductive load forces the current to continue to flow against the polarity of the transistors. The diodes D2 and D6 will take over most of the current. Since D2 is the substrate diode of Q2 we open the junction between the collector of Q2 and the substrate. So the collector of Q2 becomes a parasitic emitter. The lateral transistor QP1 turns on and pulls all surrounding N-regions (that now become the parasitic collectors) down.

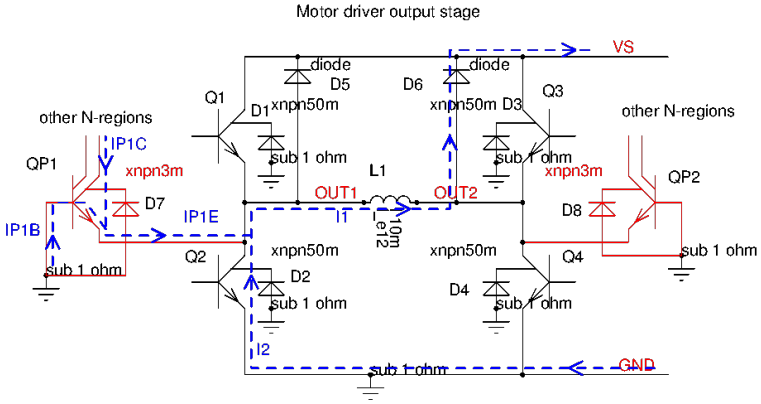


Figure 4.54: Motor bridge in flyback mode at current direction change

The parasitic lateral NPN QP1 (and QP2 as well) usually has a fairly low gain because the base is wide (some hundred um to some mm). But operating at motor current of up to some A even a gain B=0.01 is a serious issue! Therefore the gain of the parasitic lateral transistor must be made as low as possible.

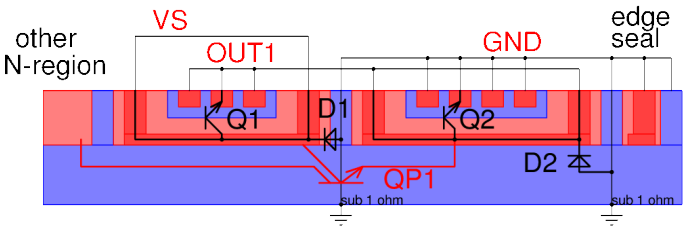


Figure 4.55: Cross section of Q1 and Q2 of the power bridge shown above.

To reduce the gain of QP1 (and QP2 respectively) the following option are available:

- Increase the base width (Distance from Q2 to the other N-regions)
- Avoid touching of burried layer and substrate (puh, needs a lot of epitaxy. expensive!)
- increase substrate doping as far as possible to reduce emitter efficiency (Limited by collector to substrate break down requirements)
- Reduce minority carrier life time in the substrate (Warning: doping substrate with heavy metals may spoil a lot of other parameters of the desired components too!)

- Make the base of parasitic QP1 high resistive my removing substrate contacts between Q1 and Q2. Risk: latch up!

4.7.6 PNP transistors

Almost every technology offers a substrate PNP transistor. The collector of the substrate PNP transistor is tied to the substrate node.

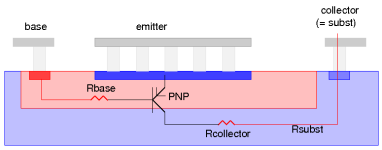


Figure 4.56: substrate PNP transistor

Matching of bipolar transistors: Matching of bipolar transistors mainly depends on the emitter area. In a mature process the matching constant of bipolar transistors is in the range of

$$V_{os} = \frac{0.3..1mV\mu m}{\sqrt{A_{emitter}}} \tag{4.101}$$

The matching slightly can be improved avoiding current crowding at the corners of the emitter. For this reason some design kits offer transistors with octagon shaped emitter instead of pure rectangular transistors (for instance ST Microelectronic HDS2P2) or even round emitters (Some old versions of DOPL of the 1980s). Round emitter however cause a lot of trouble at mask making. (I personally rather use octagon emitters drawn on a layer that is not oversized. Oversizing to compensate outdiffusion might lead to off grit corners. Generating masks while there are off grit corners may lead to unexpected shape variation due to snapping to the next grit point!)

Current matching of bipolar transistors is in the range of $0.3..2\%\mu m$.

Layout techniques for matching In a mature process gradients on the wafer are in the range of $1\%/cm$ or $1ppm/\mu m$. With these gradients common centroid layout only makes sense for structure bigger than about $30\mu m * 30\mu m$. As long as the matching structures are smaller it is sufficient to use same device orientation and same environment (dummies, no heavy doped structure close to the matching transistors, dummy wires to have the same mechanical stress at all matching transistors, keep away from pads and asymmetric trenches, check for temperature gradients)

4.7.7 Lateral PNP transistor

In many applications making the collector of the PNP transistor available as an individual node is required. In technologies offering an n-buried layer this is possible. The buried layer prevents the diffusion of the holes from the emitter to the substrate. Holes reaching the buried layer will recombine thus increasing the base current. But only about 1% or less of them will reach the substrate. The PNP transistor becomes a surface device with a weak parasitic substrate PNP in parallel.

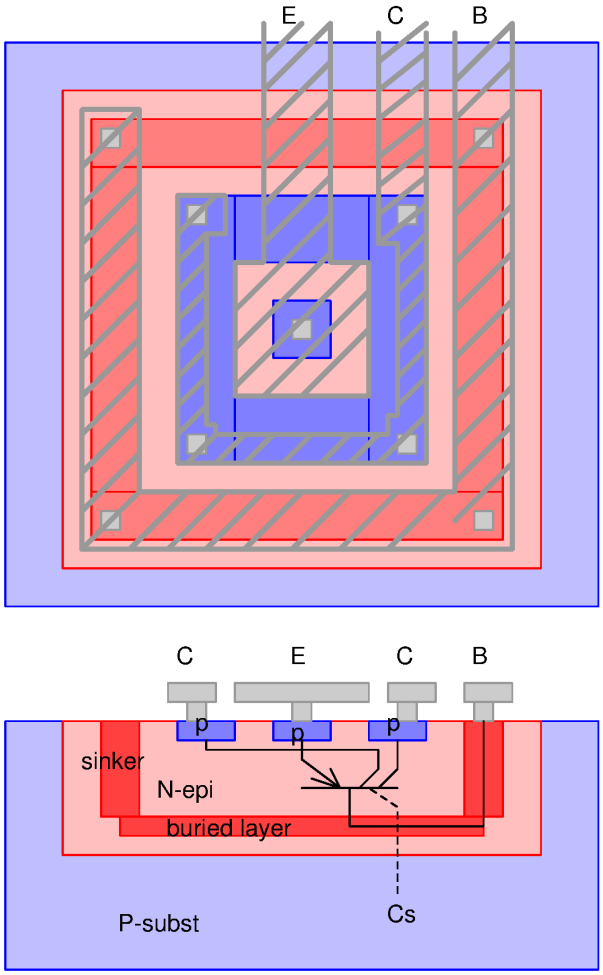


Figure 4.57: Lateral PNP transistor

The P-region in the middle is the emitter. The emitter is surrounded by a collector ring. The emitter and the collector typically use the same mask. This way the transistor's base width is self adjusted and the spread remains reasonably low.

The emitter metal intentionally overlaps the base for two reasons:

1. The metal acts as a shield against possible parasitic PMOS transistors created by traces at low voltage crossing the transistor in metal 2.

2. The metal cover prevents surface contamination and aging of the transistor.

Ideally the lateral PNP transistor has the same break down voltage between base and collector and between base and emitter. Due to the metal shield connected to emitter operating the transistor in reverse way (emitter negative versus the base) we will create a P-channel under the shield. This metal gate PMOS limits the reverse operation to a lower voltage than the junction break down theoretically allows.

Some technologies offer lateral PNP transistors with the shield intentionally connected to the base instead of the emitter to make the PNP transistor work in reverse mode until the junction break down limit is reached. These transistors can be recognized by the additional metal or poly ring surrounding the emitter.

4.8 Special devices

Standard processes used for integrated circuits just hold some low voltage devices and some bipolar devices. In most cases even the number of bipolar components is reduced to the absolute limit. Since modern processes are optimized for the CMOS transistors only the bipolar components are regarded as some kind of step children only permitted to build a bandgap or some very basic NTAT current generators.

High voltage components usually are restricted to few process variants especially designed for high voltage applications such as the ST's BCD lines, Infineon's SPT lines, NXP's ABCD process lines and TI's LBC lines. Most of these lines have a numbering scheme referring to the feature size of the CMOS process they are based on. The numbers of the technology nodes are fairly similar!

Table 11: Special processes used for power ICs

IFX	NXP	Freescale	ST	TI	CMOS channel length
			BCD1		3um
SPT1			BCD2		2um
SPT4	ABCD3	IDR-HV	BCD3		1.2um
SPT5, SPT6		Hyper80	BCD5		0.8um
SPT7		Smartmos7	BCD6	LBC7	0.35um
	ABCD9			LBC8	0.18um
SPT9, C11-HV		Smartmos9	BCD9	LBC9	0.11um

In recent years some of these process lines were licensed to silicon foundries such as TSMC or UMC making them more generally available for fab-less companies.

The components more or less unique to these process lines are high voltage NMOS and high voltage PMOS transistors. Usually these transistors are designed using the outdiffusion of the bulk defining the channel length of the high voltage transistors. Matching of the high voltage transistors usually is poor due to the short channels that in most cases can not be influenced by the circuit designer. If good matching is required the standard design approach is to build low voltage circuits with HV transistors on top of them acting as cascodes.

As a consequence of these similarities at least the CMOS parts of these process lines in fact are comparable!

NXP's ABCD lines offer dielectric isolation in stead of junction isolation permitting a fairly different high voltage design style than all other lines.

In the following the junction isolated high voltage components are shown because these flavors are more common.

Substrate transistors are feasible in all junction isolated process flavors. Nevertheless they usually are not used except in very specific cases because the substrate node often is the ground terminal (or supply terminal) for the rest of the circuit.

In packages having an exposed die pads (usually for thermal reasons) the substrate voltage may differ significantly from circuit ground. This is especially true for RF. (The inductive voltage drop over the ground bond wire is seen as a substrate noise by all circuits referring to circuit ground.) Voltage difference between substrate and circuit ground capacitively couples into some of the circuit components. (Usually bipolar transistors and high voltage MOS transistors are affected most.)

4.8.1 Vertical DMOS

In a vertical DMOS most of the electrical field of the drain is vertical. The drain is connected by the buried layer.

Since the electrical field is vertically down into the silicon (not close to the surface) usually vertical DMOS transistors suffer less from hot carriers than lateral DMOS transistors. Vertical DMOS transistors start to become more area efficient than lateral transistors above about 60V. Usually vertical transistors have several rows of sources that share one common sinker serving as the drain contact.

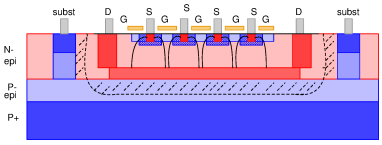


Figure 4.58: Vertical DMOS transistor over P+ substrate

The current flow follows the solid lines. The electrical field between the drain and the substrate is in the dashed areas. The buried layer must be doped very high to achieve a sufficiently low ON-resistance. The low resistive (high doped) N buried layer enforces one of the following process flows:

1. Use P+ substrate, grow P- epi, deposit the N+ buried layer, grow N- epi
2. Use P+ substrate, grow N- epi, deposit N+ buried layer, continue growing N- epi
3. Use P- substrate, deposit N+ buried layer, grow N- epi

Variant 1 offers the smallest minimum transistor that can also be used in the signal processing part of the design. If the chip size is dominated by big power transistors this benefit becomes negligible.

Variant 2 is less complex because you just use one epitaxy process. The bottom side depletion zone moves into the N- between the substrate and the N+ buried layer. Since the transistors reach deeper down the isolation spacings become more difficult to control and the small signal transistors must be designed with more margin and become bigger.

Variant 3 is the cheapest approach but the high resistive substrate increases the risk of having a latch up. This kind of process leads to design rules with certain maximum spacings between the P- iso contacts. Big power transistors

must be broken into several modules to find space for the substrate contacts between them. This way the process gets less effective for high power applications.

Which variant eventually is chosen is a financial trade off and a design risk trade off in case of variant 3.

4.8.2 Lateral DMOS

In a lateral DMOS transistor the current flows at the surface. Usually this means a power concentration and thus a heat concentration at the drain extension close to the poly silicon over field oxide. Normally lateral DMOS transistors are less rugged at ESD (electrostatic discharge) than vertical DMOS transistors.

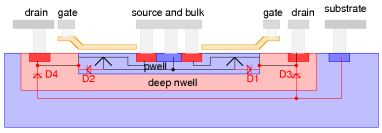


Figure 4.59: cross section of a lateral DMOS transistor including parasitic diodes

Using junction isolated technologies the drain is at the edge of the transistor for minimum size. If the source and the bulk were at the edge the spacing between source would have to accommodate two depletion areas (Junction deep nwell to substrate and junction deep nwell - may be even n+ as channel stop - to bulk). Placing the drains at the edge we only need space for the junction deep nwell to substrate.

Using deep trench isolation (DTI) instead of junction isolation it becomes more beneficial to have the sources at the edge of the transistor.

Since the lateral DMOS is a surface element there are less constraints regarding substrate doping and placement of buried layers than in a vertical DMOS (In fact a lateral DMOS can be built without a buried layer at all. However if the lateral DMOS is built without buried layer the substrate PNP gets a much higher current gain. This high PNP current gain increases losses if the LDMOS is operated in power bridges with flyback currents.)

4.8.3 High voltage PMOS

High voltage PMOS transistors usually are built very similar to low voltage PMOS transistors. They just have an additional P- drain extension and similar to the high voltage NMOS the gate is extended over field oxide acting as a field plate reducing the peak of the electrical field at the edge of the drain.

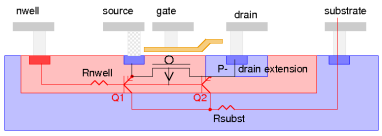


Figure 4.60: High voltage PMOS transistor

4.8.4 Substrate power PNP transistor

Compared to lateral PNP transistors vertical PNP transistors offer higher knee-currents (the gain rolls off at higher currents) and higher current gains. If the collector can be connected to substrate vertical power PNPs can be quite a performant device.

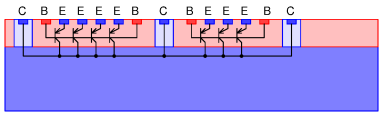


Figure 4.61: Vertical power PNP transistor

Substrate PNPs designed for high power require a very good substrate grounding. If the substrate locally gets pulled higher than adjacent nwells this may lead to latch up. So power PNPs typically are built as strips or arrays with substrate contacts in between.

Very often power PNPs are used as ESD protections. This is less critical regarding latch up because in most applications ESD is expected to take place only while the chip is unsupplied.

If substrate power PNPs are used as devices operating while the chip is supplied the following layout recommendations can be made:

1. Add enough ground contacts to keep substrate low even if the PNP is shorted to high voltage
2. Avoid nwells tied to low supply voltages in close proximity (logic wells tied to 1V etc can lead to enormous latch up currents)

4.8.5 Substrate power DMOS

This is the standard approach used to build power MOS transistors with Rdson values in the range of few milli-Ohms. Basic idea is to move the drain extension holding the drift zone vertically under the channel. The bottom metalization becomes the drain of the transistor. In most cases this is a discrete transistor. If it is to be combined with an IC technology the drain becomes coincident with the substrate of the process. There are semiconductor IC processes especially designed to built high side switches with power NMOS transistors. These processes have an N-substrate in stead of the more common P-susbstrate.

The total resistance of the component consists of three main contributors:

Rch is the channel resistance. It can be estimated knowing the effective channel length and width and the gate voltage.

Rdrain is the resistance of the drain region. It depends on the thickness of the N-material used for the drain and the doping.

Rmet1 and Rmet2 are the metal resistances of the package material (dice pad) usually their contribution is very low.

Rbond is the bond wire resistance. It depends on the length and width of the bond wire.

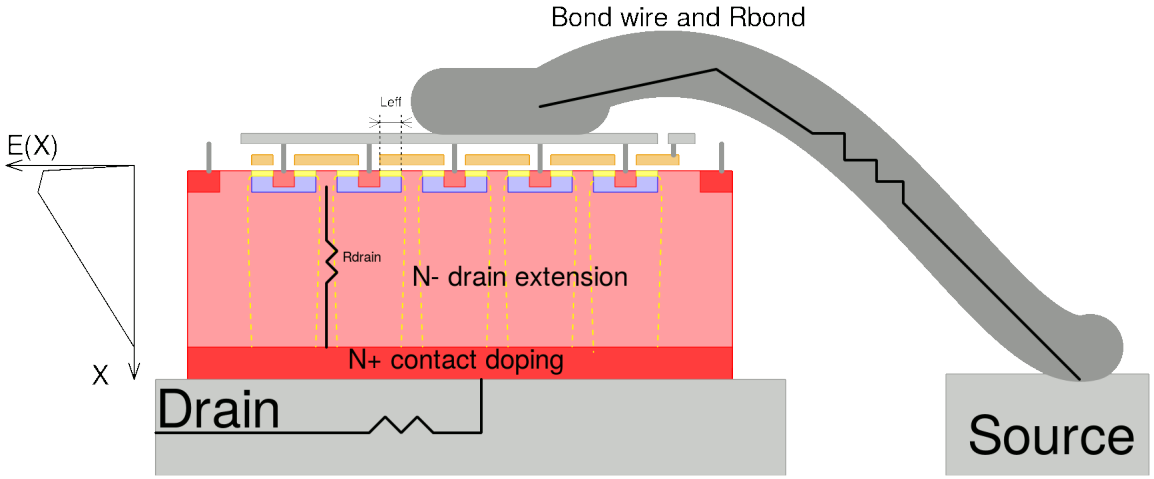


Figure 4.62: Vertical power DMOS transistor including bod wires and dice pad

Example calculation of the total Ron of a power transistor fitting in a TO220 package:

tox=40nm, W=3m, L=1μm, Vgs=10V, Vth=4V, dice size=12mm^2(active area only), thickness of the N- drain extension layer=0.005mm, N- resistivity=10 Ohm cm. Wafer thickness is 0.3mm at a resisitivity of 1mΩcm Aluminum bond wire diameter 0.2mm, bond wire length=3mm. Dice pad material is copper, Thickness of the dice pad is 0.2mm, width is 3mm.

$$R_{ch} = \frac{L}{W} * \frac{t_{ox}}{\mu * \epsilon_{SiO2}} * \frac{1}{V_{gs} - V_{th}} = 0.27m\Omega$$

Well, may be the ideal calculation is a bit too optimistic and we have to expect a factor 2 more. Nevertheless this means the pure channel resistance can be neglected! The more interesting part is the resistance of the vertical path down to the drain.

$$R_{drain} = \frac{0.005mm}{12mm^2} * 10\Omega * 10mm + \frac{0.3mm}{12mm^2} * 0.01\Omega mm = 41m\Omega + 2.5m\Omega = 43,5m\Omega$$

The bond wire contributes

$$\frac{3mm * 28.2n\Omega * m}{(0.1mm)^2 * 3.1415} = 2.7m\Omega$$

The frame (dice pad at the drain) and the source pin usually contribute an other 300 to 600$\mu\Omega$ (depending on pin length mainly at the source).

So the complete transistor reaches about 47mΩ at room temperature.

Since the 1990s semiconductor manufacturers learned to lap the wafers as thin as 50μm. Additionally multiple bond wires and thicker source metal layers (up to 60μm) became available. Today discrete 40V power transistors with an Ron of less than 1mΩ are available.

Today package (bond wire and pin) are the limiting factors! Therefore these extremely low resistive transistors don't come in TO220 like packages anymore. To reduce the pin resistance leadless packages are used.

The standard DMOS design shown above has one big limitation. If it is scaled for higher voltage the break down field strength may not be exceeded. Since this is a material constant the E-field triangle can only be widened in the ground side. The transistor must be made thicker. The minimum thickness can be calculated integrating the field.

$$V_{dsmax} = \int E(x)dx \tag{4.102}$$

Looks more complicated than it is. Since E(x) is more or less a triangle the break down voltage simply becomes:

$$V_{dsmax} = E_{br} * d/2 \tag{4.103}$$

d is the thickness of the N- layer. So the thickness we need becomes:

$$d = 2 * V_{dsmax}/E_{br} \tag{4.104}$$

To be more clear d is the length of the drain extension RESISTOR!

Furthermore the doping of the drain extension must be adjusted such that the peak of the field strength just reaches the break down field strength and that the depleted area operating at V_{dsmax}exactly fills the range from the drain contact to the bulk. The equation can be stolen from the diode calculations :-) .

$$N_D = \frac{E_{br}^2 * \varepsilon_0 * \varepsilon_r}{2 * e * V_{dsmax}}$$

That sucks! The resistivity of our drain resistor is:

$$r = \frac{1}{N_D * e * \mu_n} = \frac{2 * V_{dsmax}}{E_{br}^2 * \varepsilon_0 * \varepsilon_r * \mu_n} \tag{4.105}$$

And the resistance of the drain extension of a transistor with area A becomes:

$$R_{drainext} = \frac{r * d}{A} = \frac{4 * V_{dsmax}^2}{A * E_{br}^3 * \varepsilon_0 * \varepsilon_r * \mu_n} \tag{4.106}$$

The more convenient description of the silicon limit is the equation of the area resistance that can best case be achieved using silicon DMOS transistors with this standard cross section. The result is a number normally expressed with the unit Ohm*mm^2.

$$r_{area} = \frac{4 * V_{dsmax}^2}{E_{br}^3 * \varepsilon_0 * \varepsilon_r * \mu_n} \tag{4.107}$$

So far the resistance scales with the square of V_{dsmax}. Practical values are worse because additional area gets lost at the edge of the transistors. Practical results found in experimental transistors show:

$$R_{dson} \sim V_{dsmax}^{2.5}$$

This more than quadratic proportionality is called the silicon limit of a conventional DMOS transistor. (It applies to any MOS transistor that uses a normal drain extension as well.)

Example: We do not want to have corner break down etc. So we only want to exploit the capabilities of silicon to a maximum field strength of 20V/μm. for a 50V transistor the silicon limit calculates as

$$r_{area50V} = \frac{4 * (50V)^2}{(20V/\mu m)^3 * 12 * 8.8pF/m * 600cm^2/Vs} = 0.2\Omega mm^2$$

Super Junction Transistor: Super junction DMOS transistors (Infineon sales name cool MOS) are a trick to get around the silicon limit [37]. The following figure shows the cross section.

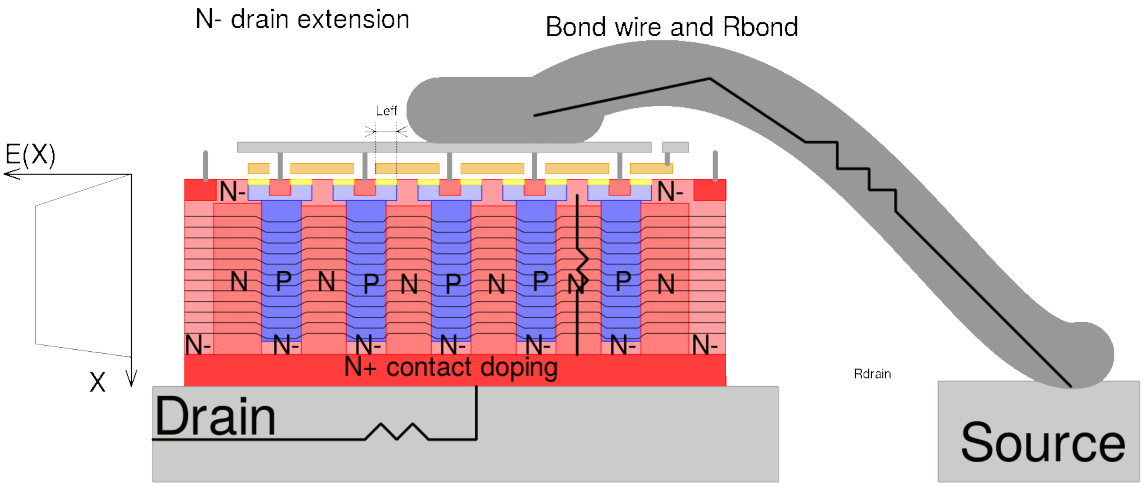


Figure 4.63: Cross section of a super junction transistor (cool MOS transistor)

The P and N doped pillars have exactly the same doping densities. As long as the drain voltage is low (some V) the N-pillar acts as a low resistive path from the drain contact to the end of the channel (yellow).

When the pillars deplete the capacity between drain and source (Cds) and the capacity between drain and gate (Cdg) changes in an extremely abrupt way! The change typically can be observed at about Vds~20V and can be more than one magnitude within a few volt!

Main field of application of super junction transistor is the voltage range from 200V to 800V.

The dopings of the P-pillars and the N-pillars must be exactly balanced to make both pillars deplete simultaneously.

The cell pitch defines the doping level needed in the pillars. Cell pitch variations lead to a degradation of the break down voltage too.

The smaller the pitch the higher the doping of the pillars can be chosen. High doping of the pillars leads to low R_{dson}. So the achievable R_{dson}depends on the technological capabilities of the manufacturer (pitch, control of the doping matching of P and N) and the voltage (height of the pillars needed). Kondekar [37] states the following relationship:

$$R_{on} * area = 2.6 * 10^{-5} * C_p * BV[\Omega * mm^2] \tag{4.108}$$

In this equation BV is the break down voltage in Volts and C_p is the pitch of the pillars μm.

One of the drawbacks of the super junction transistor is that in ON-state the depletion area between the pillars disappears. So electrons from the N-pillar can diffuse into the P-pillar and holes from the P-pillar can diffuse into the N-pillar. In other words, we have minority carriers! Turning off the transistor these minority carriers have to recombine. This makes the super junction transistor significantly slower (at least at turn off) than a normal DMOS transistor. This effect limits the switching speed of super junction transistors in power inverters to some 10kHz.

A second limitation is the matching of the doping of the P-pillars and the N-pillars. Since this matching never is perfect current (2018) super junction transistors are limited to about 650V.

Alternative materials for power transistors: The silicon limit shown above has the break down field strength at the power of 3 in the denominator. Finding materials with a higher break down field strength is extremely interesting for high voltage power transistors. Very interesting candidates offering break down field strength values up to 300V/μm are silicon carbide (SiC) and gallium nitride (GaN) [26]. These materials allow such thin drain extension layers that in theory they beat super junction transistors (using silicon) by several magnitudes!.

Furthermore the concept of a super junction transistor could just as well be used for SiC. This would enable power transistors of some 10mΩ at break down voltages of several 10kV!

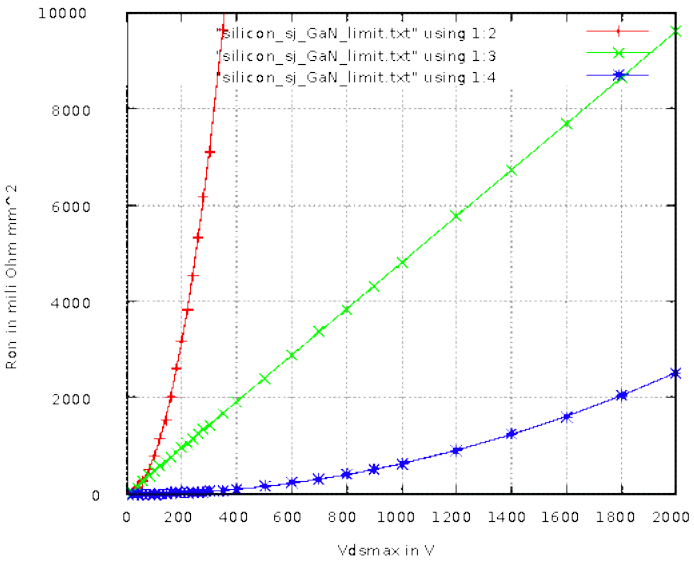


Figure 4.64: Comparison of the silicon limit of a standard cross section, SJ cross section (using silicon) and wide bandgap semiconductors (GaN) limit using a standard cross section transistor

The above figure shows that a SiC transistors and GaN transistors theoretically beat a SJ silicon transistor by 2 magnitudes! The red curve is the silicon limit of a standard cross section. The green curve is the limit of a SJ transistor using silicon. The blue curve is the theoretical calculation of a SiC transistor (SiC limit) or a GaN transistor (Both materials have similar break down voltages).

Practical designs of SiC transistors don't reach these values yet because for handling reasons the wafers can't be lapped down to just a few micrometers. The path resistance through the (typically 50um to 150um thick) material can easily be significantly higher than the theoretical ON-resistance of the pure channel. Today (2020) SiC transitors offer ON-resistance in the range of $2000m\Omega * mm^2$ at Vbr=1200V. (Estimated from the data sheet of STC50N120). This is about factor 3 away from the theoretical limit.

Gallium nitride is a special case because in Gallium Nitride (GaN) the electrons can only move well at the surface (two dimensional electron gas) but not vertically in the bulk.

The reality the benefits of SiC and GaN can't be fully exploited. Limiting factors are contact resistance and the resistivity of the bulk material. Silicon can be doped very high and the path from the 'inner drain' to the die pad can be made as low resistive as $3m\Omega cm$. SiC doesn't permit such high doping concentrations. The lowest reported values for the path from the 'inner drain' to the die pad are in the range of $20m\Omega cm$. For this reason the drain resistance of SiC is significantly higher than anticipated simply calculating the SiC limit.

GaN being a 2-dimensional surface element mainly suffers from the contact spread resistance of the shallow conductive layer.

An other material that theoretically would allow even better performing transistors is diamond. Currently (2018) is is possible to produce amorphous diamond using CVD (chemical wafer deposition). For commercial use this way of producing diamonds for semiconductors is too expensive. In addition there is not yet a reasonably good way found to dope diamonds and to produce a gate oxide. But research is running and the high break down fields, high carrier mobility and good thermal properties could make diamond an interresting material in the future.

4.8.6 SiC transistors

Until about 2015 the main manufacturing problem of SiC MOS transistors is the gate oxide [45]. Therefor cascodes using SiC J-FET transistors in combination with a low voltage silicon MOS transistor were a common design choice in the beginning of SiC power transistor design. In this combination the SiC J-FET is used as a cascode to limit the voltage while the Si MOS transistor acts as a switch.

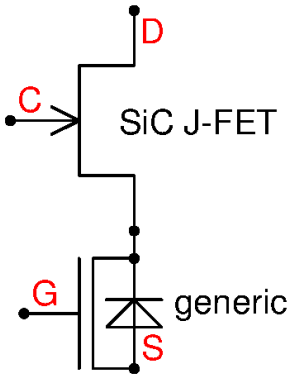


Figure 4.65: SiC and a low voltage MOS transistor acting as a cascode

The device acts as a switch with the gate of the low voltage transistor (G) as the control node. Pin C is the cascode gate voltage. It usually is connected to system ground or the source (S) of the power transistor.

Slew rate control of such a device is difficult because the miller capacity of the MOS transistor is not exposed to a voltage swing. So the classical approach of limiting the slew rate by driving the device with a controlled gate current doesn't work anymore.

Since about 2017 producing real SiC MOS-FETs is getting common. Cascodes are getting more and more obsolete in recent time. Instead several chip producers offer SiC MOSFETs. Using SiC MOSFETs the classical driver approach limiting the gate current to define the slew rate is applicable again.

Typical applications of SiC MOSFETs are power bridges for several hundred Volts and some hundred Amperes.

SiC MOSFET transistors Today SiC MOSFET transistors are more common than the JFET cascodes used in about 2015. To estimate the capabilities of this material it is recommended to compare the most important parameters of silicon and SiC. The most frequently used variant of SiC used for transistors is 4H-SiC.

Table 12: comparison of silicon and SiC

parameter	Si	4H-SiC	unit
break down field E_{max}	25	250	$V/\mu m$
bandgap voltage V_{bg}	1.23	3.23	V
carrier mobility	600..1100	800..1000	cm^2/Vs
saturation speed	10^5	$2*10^5$	m/s
dielectric constant ε_r	12	9.7	

Due to the higher break down field strength the channel of a SiC transistor can be made one magnitude shorter than the channel of a silicon MOS transistor. For the comparison let us have a look at a typical 1200V device used for classical power converters.

Table 13: achievable specific resistance using different power transistor technologies for 1200V transistors

	Si	Si super junction	SiC	remark
V_{ds}	1200V	1200V	1200V	design target
channel length	$96\mu m$	$48\mu m$	$9.6\mu m$	triangular field for non s.j.
Ron*area	$58.2\Omega * mm^2$	$10\Omega * mm^2$?	$43.6 m\Omega mm^2$	calculated with worst case mobility

The extremely low specific resistance of a SiC MOSFET of course is a theoretical number because it would require wafers as thin as the depletion zone. For handling reasons wafers are significantly thicker than $10\mu m$. For this reason transistors practically available (from volume production) *are about 3 to 50 times higher resistive*. (wafer thickness $50\mu m$ instead of the ideal $10\mu m$,wafer resistivity of SiC in the range of $20 m\Omega cm$ compared to high doped silicon having about $3 m\Omega cm$).

Turn off behavior: A SiC MOS transistor in a majority carrier device. So the switching speed of SiC MOSFETs (e.g. at turn off) is significantly faster than the speed of super junction transistors of the same voltage.

behavior at high gate voltage: If the gate voltage of a SiC transistor approaches the velocity saturation limit the current increase becomes less than quadratic. In velocity saturation the drain current only increases linear with the effective gate voltage. The effective gate voltage at which velocity saturation occurs can be calculated:

$$V_{gseffvs} = 2 * n * L * \frac{v_{sat}}{\mu}$$

For a 1200V SiC transistor we need a Chanel length of about $5\mu m$(assuming a graded junction that makes the electrical field distribution more rectangular). Let's use this length to calculate the gate overdrive to reach velocity saturation. (assumption n=1.4)

$$V_{gseffvs} = 2 * 1.4 * 5 * 10^{-6} m * \frac{2 * 10^5 m/s}{0.1 m^2/Vs} = 28V$$

This number is not far away from the static maximum gate voltage of most SiC power transistor (typical limits are in the range of 18V to 25V.) Between the region with strong inversion with it's quadratic characteristic and the deep velocity saturation there is a transition zone in which the current already increases at a lower rate.

Table 14: ranges of operation of a MOS transistor

Vgs	$< V_{th}$	$V_{th} to 0.5 * V_{gsvs}$	$0.5 * V_{gsvs} to V_{gsvs}$	$> V_{gsvs}$
range	WI	strong inversion	transition	vel. sat.
behavior	$I_d \sim e^{V_{gs}}$	$I_d \sim V_{gseff}^2$	$I_d \sim V_{gseff}^m, 1 < m < 2$	$I_d \sim V_{gseff}$

A standard silicon MOSFET has the same behavior. But since the channel length is much higher (for the same maximum V_{ds}) this roll off of the transconductance comes much later.

4.8.7 GaN transistors

First generation GaN (Gallium Nitride) transistors use an N-doped gallium nitride material with an P-doped "gate". In GaN the electrons have high mobility along the surface of the layers but very low mobility inside the bulk. The shallow layer in which the electrons can move is extremely sensitive to electrical fields. A positive charge close to this layer will attract electrons that can carry the current. A negative charge will deplete the layer the electrons cam move in. This characteristic is used to create a field effect transistor using the P-doped "gate" to control the shallow layer in which the electrons can move. The "gate" is without gate oxide. It simply is a PN junction. The voltage needed to turn on the transistor is positive but lower than the forward voltage of the diode. If a driver attempts to apply a higher voltage than the diode forward voltage (about 1.5V) the diode clamps the gate voltage.

In the GaN transistor the path of the electrons is not interrupted by a P-doped region like in a bipolar transistor. It is a majority carrier device in which the electrons simply can't bypass the depleted surface by flowing through the bulk (like they would do in silicon if there is no P-bulk around the source)

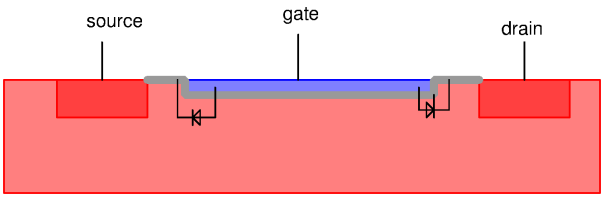


Figure 4.66: Cross section of a GaN transistor. The electrons can only move in the gray interface layer (majority carrier). Since the gate is a P-doped region without an oxide the two diodes represent the PN junction

Drivers for GaN transistors limit the steady state current applied to the diode acting as a gate. For fast turn on and turn off the gate is driven with up to some amperes. Once the transistor is on the gate current limits reduced to some mA.

Typical applications for GaN transistors are fast switching power supplies (above 1MHz, up to 300V) and RF amplifiers up to the GHz range. (Cellular phone base stations).

Currently (2018) GaN MOS FETs are used commercially for a voltage range of 100V to 650V.

Second generation GaN transistors use real MOS gates instead of a junction. This eliminates the need of limiting the gate current and the driver stages are by far less complex.

Dynamic Ron shift of GaN transistors: While the transistor is in off state electrons can leave the "two dimensional electron gas". When the transistor is turned on again these electrons are missing and it takes some ns to replace them and to build a fully conducting channel again. The on resistance is higher than anticipated for a certain time.

This dynamic Ron shift depends on the field applied during off state (because the electrical field pulls the electrons away from the surface.

One counter measure against dynamic Ron shift is to create a P-doped ring around the active area of the transistor. The P-doped ring prevents electrons from leaving the 2 dimensional electron gas.

4.8.8 IGBT (Isolated gate bipolar transistor)

Most IGBT (Insulated gate bipolar transistor) transistors are designed similar to the standard DMOS cross section but the N+ layer connecting the drain is replaced by a P+ layer becoming the emitter of a PNP transistor [38]

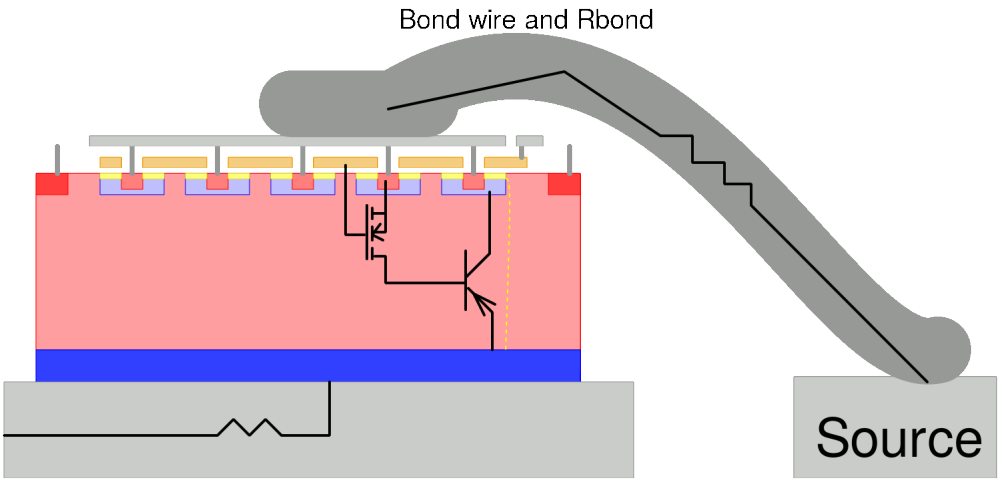


Figure 4.67: Cross section of an IGBT

The advantage of an IGBT is that the P+ acting as an emitter will flood the N-epi layer when the transistor is turned on. This makes the transistor as low resistive as a bipolar transistor. Typically silicon IGBTs are used for voltages higher than about 800V.

Typical applications are switch-mode power supplies and motor bridges with switching frequencies in the range of some KHz to some 10kHz. Higher frequency applications suffer from the slow turn off (current tail) of the bipolar part of the IGBT. Switching off at zero current crossing is helping to reduce the turn off losses (resonant mode power supplies).

4.8.9 Thyristors

Thyristors are feared as parasitic components by chip designers. They can be regarded as a combination of a PNP transistor and an NPN transistor. The most frequent equivalent circuit is shown below.

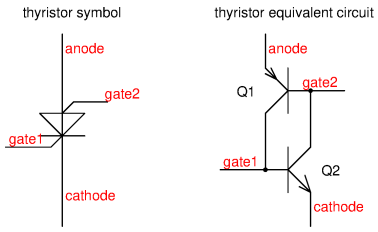


Figure 4.68: thyristor equivalent circuit

The behavior is simple. If there is a voltage applied between nodes anode and cathode first nothing happens. As soon as a current is injected into gate1 Q2 turns on. The collector current of Q2 flows into gate2 and immediately turns on Q1. Now Q1 injects into gate1. The structure latches.

There are three ways of turning it off again:

1. disconnect the power supply from nodes anode and cathode.
2. pull gate1 negative. The required current is higher than what the collector of Q1 can deliver!
3. pull gate2 positive. The required current is higher than what the collector of Q2 can deliver!

The most classical way is to disconnect the power supply. Pulling gate1 down or pulling gate2 up often doesn't work due to the contact spread resistance of the gate nodes.

Thyristors are characterized by the following parameters:

1. The voltage drop between anode and cathode at a specified current (V_{on})
2. The holding current required to flow into the anode to keep it conducting. (I_{hold})
3. The forward and the reverse break down voltage.(V_{bdf} and V_{bdr})
4. The critical slew rate at which the thyristor turns on due to injection via the collector-base capacities of the transistors (dV_{ac}/dt)
5. The turn off delay required to recombine minority carriers while there is no anode current flowing. (t_{doff})

Unintentional parasitic thyristors often are found in CMOS logic.

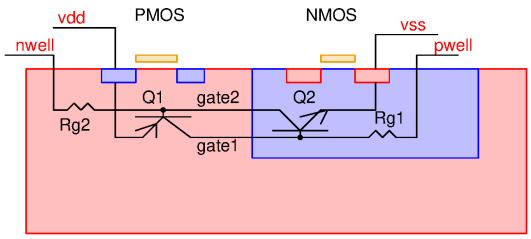


Figure 4.69: parasitic thyristor in CMOS logic

Under normal circumstances the nodes nwell and vdd are shorted and nodes vss and pwell are shorted to each other. Now we can calculate the hold current of the parasitic thyristor.

$$I_{hold} = max(V_f/R_{g1}, V_f/R_{g2}) \tag{4.109}$$

If the miller capacities of the parasitic bipolar transistors are known the critical slew rate can be calculated as well:

$$dV/dt = \frac{V_f}{R_{gx} * C_{cb}} \tag{4.110}$$

These equations shows the importance of making the path resistance from the well contact to the parasitic bipolar transistor under the logic transistors as low resistive as possible.

Power Thyristors: Power thyristors are implemented as vertical components. A typical cross section looks like this:

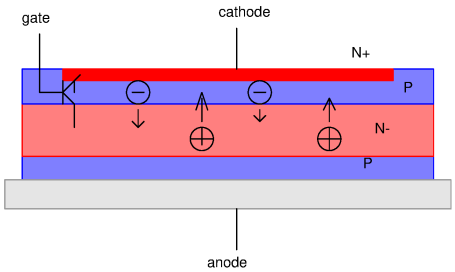


Figure 4.70: cross section of a power thyristor

The cathode side is high doped similar to the emitter of an NPN transistor. The gate is P-doped similar to the base of an NPN transistor. The N- zone accommodates the electric field while the thyristor is off. The break down voltage mainly is determined by the thickness of the N- layer. Thyristors can have break down voltages up to several kV. The anode is connected to the P doped are under the N- drift zone.

To turn on the thyristor the little NPN transistor in the left corner is activated. As soon as the gate injects current into the cathode the electrons of the cathode diffuse through the gate into the N- zone. Now the PNP transistor turns on and holes start to drift through the N- drift zone from the cathode into the P doped gate. What started as a local activation in one corner leads to a global turning on of the device in the whole device area.

For power electronics this is a desired behavior because the current flow becomes very homogeneous. Power thyristors can handle thousands of Amperes without severe current crowding problems. The size of a thyristor can be made as big as a complete wafer. This makes thyristors a preferred component used for AC power grid control.

Figure 4.71: Examples of thyristors for some 10A to some 100A

Usually the anode is connected to the metal case of the thyristor. The cathode is the big pin and the gate is the small pin.

4.8.10 Laser diodes

Usually laser diodes aren't present on the chip itself. They are described here because some of the properties of laser diodes have to be considered designing chips driving them. [57] gives a brief introduction to laser diodes.

The laser diode can be regarded as a constant voltage load in series with a little resistance. Usually laser diodes are designed using III-V semiconductors such as GaAs or GaAsInP etc. to achieve a direct band gap (carriers in the conductive band have the same momentum as carriers in the valence band. So no interaction with phonons is needed). Typical forward voltages are in the range of 1.4V per junction. (You might have to drive a stack of several diodes.)

To achieve stimulated emission there must be more carriers in the conductive band than in the valence band. Then the probability of triggering a carrier's descending from conductive band to valence band becomes higher than the probability of a carrier's ascending from valence band to conductive band. This is called inverted population.

To achieve an inverted population a diode junction is needed. This junction has to be operated at a high current density to invert the population locally at the junction.

A LASER diode has three main areas of operation:

- Blocked: The diode is reverse biased. There is no current flow. The only load component visible to the driver is the junction capacity and possibly the inductance of the bond wires.
- Spontaneous emission: The diode operates in forward mode but the current is too low to pump the diode into inversion. The laser diode operates as a classical light emitting diode (LED). The spectrum of the light resembles to small bandwidth noise.
- Laser emission: The diode is operating with sufficient current to pump it into inversion. There is light amplification and the spectrum of the laser has discrete lines.

Pumping the LASER diodes into inversion takes a certain time [58].

$$t_d = t_s * ln(\frac{I_{on} - I_{off}}{I_{on} - I_{th}}) \tag{4.111}$$

$t_s$is the carrier life time due to spontaneous emission. Usually it is in the range of some ns (3..4ns are common values for AlGaAs laser diodes [58]). $I_{th}$is the threshold current of the laser diode.

Light emission of laser diodes: In the non inverted operating range there only is spontaneous emission, but no amplification. Increasing the number of electron is the conductive band the laser diode reaches inversion. As soon as inversion is reached the laser diode has an optical gain. There are more photons making an electron jump from the conductive band to the valence band than lifting an electron from the valence band to the conductive band. The optical output power increases rapidly.

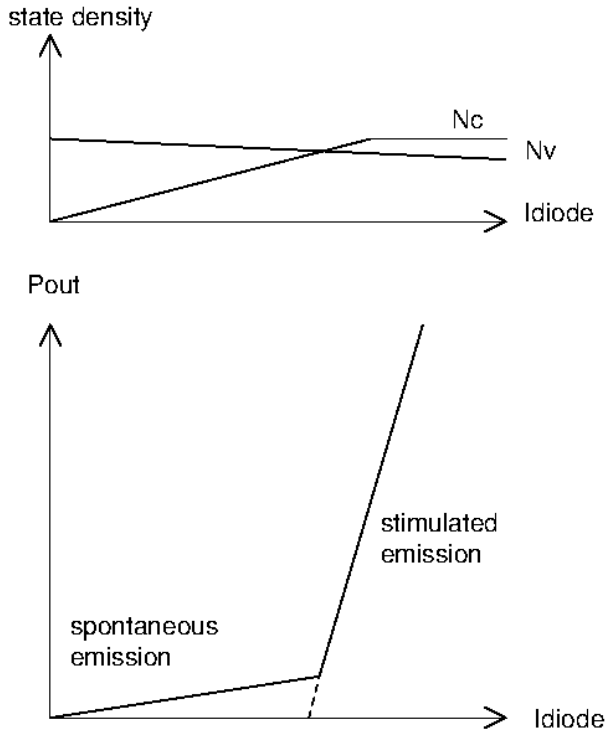


Figure 4.72: operating ranges of a LASER diode

The intersection of the tangent (dashed line) with the 0-emission line is called the threshold of the laser diode.

If the LASER diode is pumped with current pulses first the conductive band has to be filled with electrons. As soon as inversion is reached the loop gain of the LASER beam moving between the mirrors exceeds 1. In the beginning the number of photons increases exponentially. Each photon generated by stimulated emission of course means an electron falling from the conductive band back to the valence band. This empties the valence band and the gain collapses. As a consequence the optical output of the LASER collapses as well until the inversion gets reestablished by the pumping current.

This leads to a relaxation oscillation of the optical output power of the LASER.

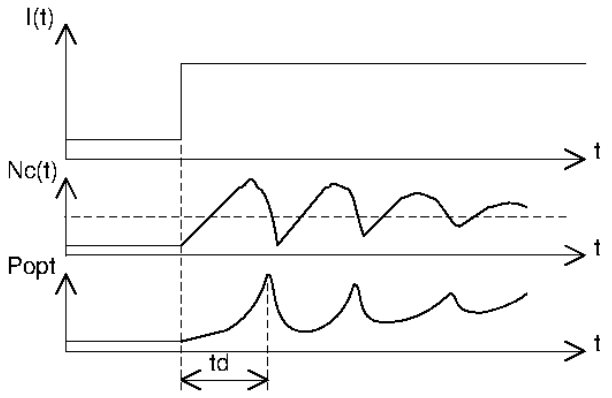


Figure 4.73: Relaxation oscillation of a LASER diode

Pulse shortening: The relaxation behavior of the LASER diode can intentionally be used to create LASER pulses in the pico second range. The rectangular drive pulse simply is turned off again at td. The LASER diode will simply produce only one singel pulse because pumping is stopped after this first pulse. This way it is possible to create single light pulses of only a few ps.

4.8.11 Photo diodes

A photo diode basically is a diode operated in blocked mode. The current flowing through the photo diode is proportional to the number of light quantas absorbed in the junction.

4.8.12 CCD image sensors

A CCD (charged coupled device) photo sensor consists of a P- silicon with gates on top to collect electrons and to shift electrons. There is no simple circuit representation of a CCD using simple capacitors and MOS transistors. Instead a cross section of the CCD is needed to understand the function. Modern cameras of course use two dimensional CCDs. For understanding the concept it is sufficient to draw a one dimensional CCD.

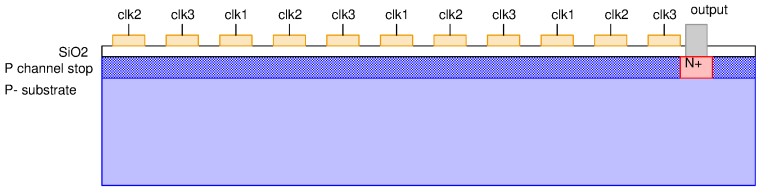


Figure 4.74: The CCD in it's initial dark state

While no voltage is applied at the gates and the CCD isn't exposed to light there is no photo generation. The P-doped material holds mainly holes. Electrons generated thermally will sooner or later recombine with the holes.

The poly silicon gates are thin enough to be transparent. The clock lines for clk1 to clk3 are thin enough to be neglected.

As soon as the CCD is exposed to light this will lead to the generation of hole-electron pairs. The number of holes exceeds the number of dopants by the number of electrons generated. Still assuming there is no gate voltage these pairs still maintain neutrality and will recombine again after some time. In the following figure only the excess holes (those exceeding the number of dopants) is shown.

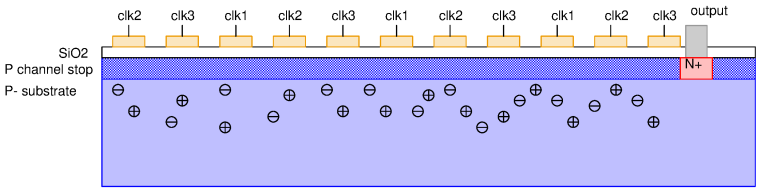


Figure 4.75: The CCD in exposed (illuminated) state but without charges at the gates.

The situation changes as soon as some of the gates are charged with a positive gate voltage. The field will separate the electrons from the holes before they can recombine. This leads to 'buckets' (represented by the dashed red lines in the following figure) filled with electrons generated by the light. The holes are either pushed down to the substrate or to the areas with a negative charged gate on top. So each 'bucket' can be regarded as an N-conducting area created by the positive charges of the gate above.

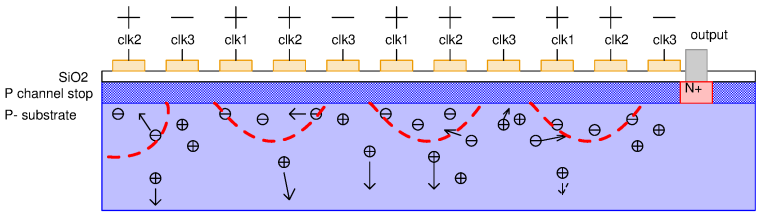


Figure 4.76: The CCD with 2 out of three electrons charged positive to collect the electrons.

Once the electrons are concentrated in the 'buckets' the shutter gets closed and the light exposure ends. The photo now is stored by the electrons in the buckets.

To read out the sensor the electrons now have to be transported to the electrical contact on the right side of the sensor. To accomplish this movement of the 'buckets' the following clock scheme is required:

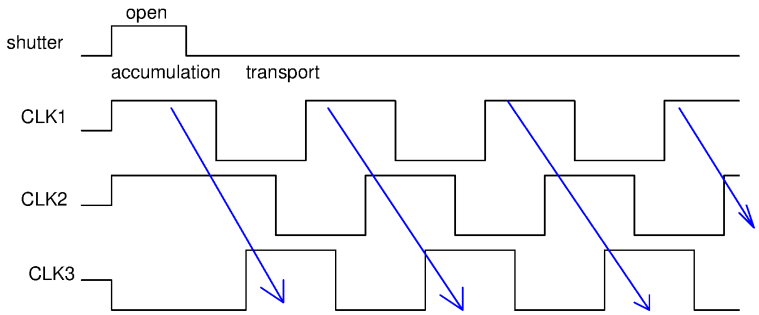


Figure 4.77: clocks required for accumulation and transport of the charges. The blue arrow represents the movement of the negative charges enclosed in the 'bucket'

The three clocks must be overlapping in a way that there always is a positive overlap maintaining the 'bucket' and a negative overlap maintaining the 'outside of the bucket'. The area of the 'bucket' may change during transportation. During the time only one gate is charged positive the charge of this gate must be high enough to overcompensate the negative charge of the electrons inside the 'bucket'. The following figure shows the transportation of the negative charges in the silicon.

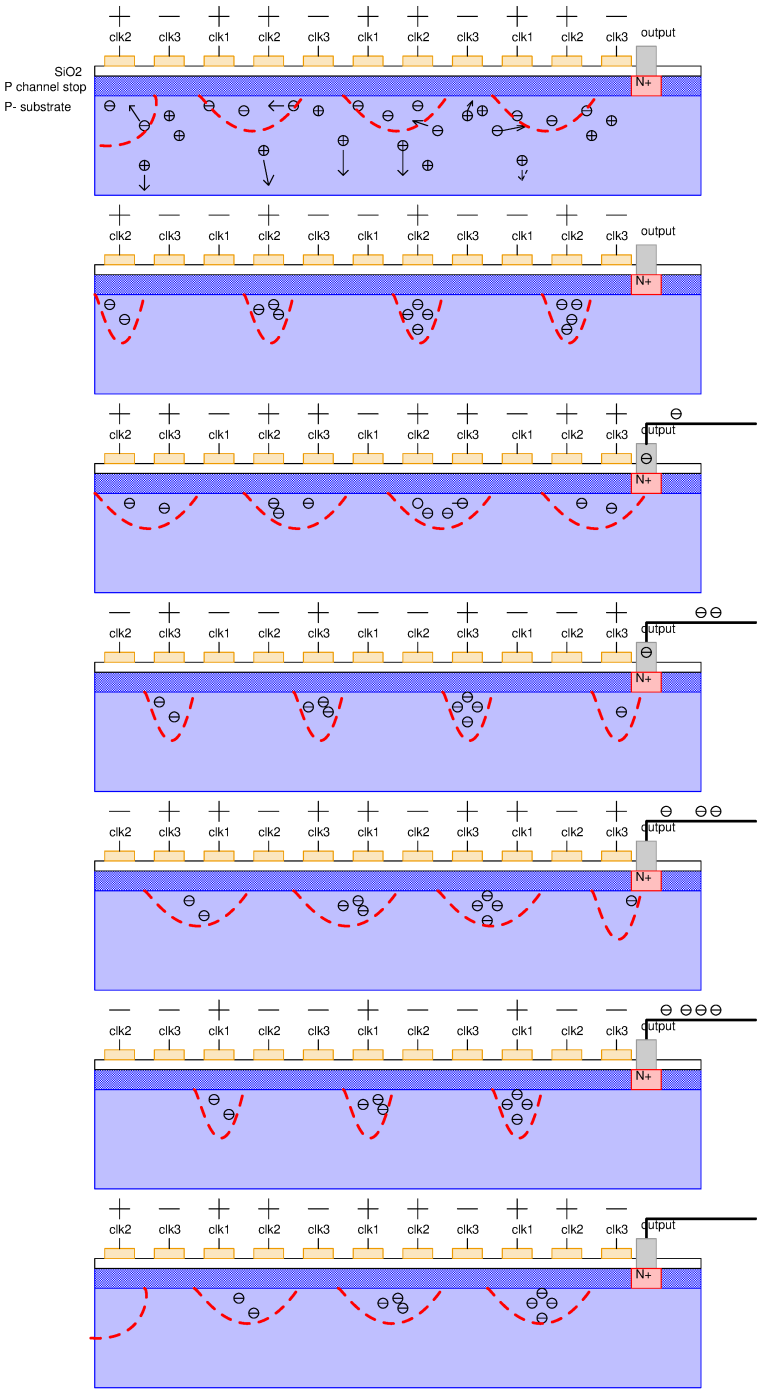


Figure 4.78: Transportation of the electrons to the cathode on the right side of the CCD

Two dimensional CCDs: For a full picture we need many parallel shift lines each having a length of N cell (3N gates). Each output of the horizontal shifts ends in a vertical collection CCD. This vertical CCD is not exposed to the light. It only is used to transport the data to the read amplifier. The vertical line must of course be clocked N-times faster than the horizontal lines

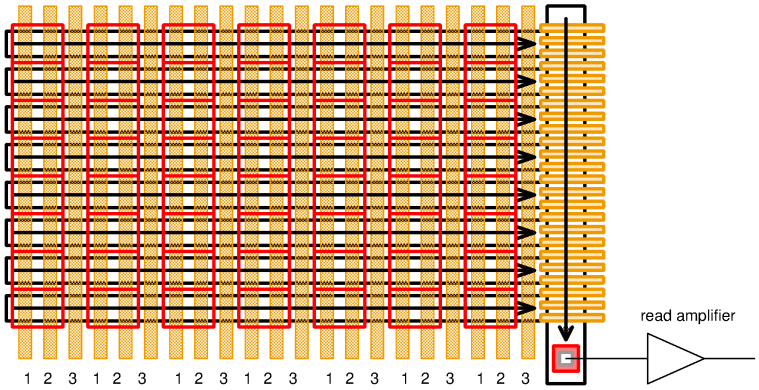


Figure 4.79: 2 dimensional CCD with a 1 bit wide bit stream output.

The red rectangles are the sensitive areas. The data is shifted to the right side to fill the vertical analog shift register. This topology works nicely for black and white photos.

Modern cameras have 3 colors: Red, green, blue. At minimum one pixel must consist of 3 sensitive spots sensitive to different colors. Well, not really compatible to a rectangular pixel. Usually the most important color is green. For this reason the green sensitive fields are doubled. (green is exactly in the middle of the spectrum humans can see). Classical color sensors have a filter on top of each sub-pixel. This filter is shown by the fill color of each of the red sensitive area rectangles. Thus one color pixel (black rectangle) consists of 4 color pixels (2 green ones + 1 blue +1 red).

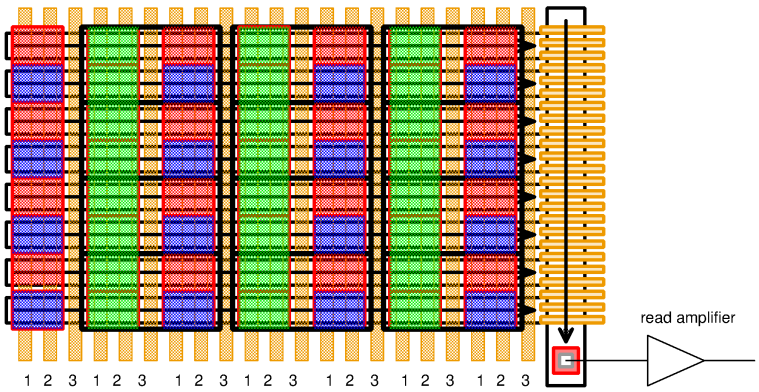


Figure 4.80: 2 dimensional classical color sensor

The information of one color pixel is 2 components green + 1 component red + 1 component blue.

$$data = 2 * G + R + B$$

Each color reaches the amplifier in a pure form. The exaggeration of the green color can simply be corrected by the digital post processing of the signal. This is needed anyway because the quanta efficiency of silicon changes with the wave length of the light. So in the post processing each color is multiplied with a constant factor that depends on the sensor used.

The periodic color pattern has some undesired side effects:

- The green picture has a horizontal offset of half a pixel width to the red and the blue picture.
- The red and the blue picture have half a vertical pixel offset

Under certain circumstances this leads to visible artifacts in the picture (e.g. if very small periodic patterns are in the photo - the classical stripe and checker board aliasing sometimes seen in TV if someone wears plait.)

Dual Pixel CCD Image Sensor: Reducing the pixel size reduces the sensitivity of the sensor. (Less light quantas per pixel and time at low light) One proposal to reduce the offset between the different colors of the sub-pictures and to keep the sensor size untouched is the dual pixel CMOS CCD image sensor [62] presented at the IS&T's 1999 PICS conference. The following picture compares the pixel of a classical sensor and the array of the dual pixel image sensor.

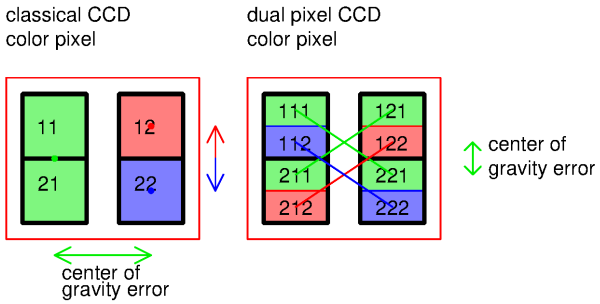


Figure 4.81: comparison of the classical sensor pixel and the dual pixel sensor

The classical sensor has a shift of the center of gravity of the green sub-pixels versus the positions of the blue and the red sub-pixels. The dual pixel CCD image sensor has exactly the same center of gravity for the red and the blue sub-pixels. The center of gravity of the 4 green sub-pixels is shifted vertically versus the center of gravity of the blue and the red ones by only a quarter of a pixel. ([62] proposed this placement of the colors for a more simple CCD structure. Swapping green 211 and red 212 and swapping green 221 and blue 222 would have been even better! But since it is described in the literature this way I stick to this less than perfect example)

The colors can be calculated from the charges in the areas 111 to 222:

green=0.5*(q111+q121+q211+q221)

red=q122+q212

blue=q112+q222

Assuming the areas 11 and 111+121, 12 and 122+212, 21 and 211+221, 22 and 112+222 are identical this should lead to the same signal to noise ratio.

The simple looking adding procedure has some hidden risks:

- Adding directly on the CCD simply merging for instance 'bucket' 112 and 'bucket' 222 to get the blue sub-pixel in an analog way avoids quantisation errors. This however makes the CCD sensor much more complex building an analog charge accumulator for sub-pixels of the same color. This increase of sensor complexity leads to a long time to market!
- Simply streaming all 'buckets' to a signal processor and first converting it to digital values and afterwards adding them in a digital way adds up systematic quantisation errors. The simple CCD sensor is either paid with higher requirements of the ADC or with a loss of performance at low light. Assuming we keep using the same ADC resolution the loss would be a factor 4 because we have twice the quanttisation error at half of the signal per 'bucket'. The advantage of processing digital instead of adding the 'buckets' directly on the sensor is a shorter time to market. (Note: the data rate of the dual pixel sensor is twice of the classical one. If the ADC is already at the cutoff frequency for the classical one the loss of resolution using the same ADC for a dual pixel sensor can even be a factor 8!)

Cameras available on the market in 2017 (for instance a d6-markII falls back behind the old d6 significantly at low light) seem to have chosen the digital adding probably due to lack of development time of something cleaner.

Analog adding of charges in a CCD: The concept is simple. just let two charges merge by controlling the gates accordingly. To do so stop the movement of the first charge before it reaches the contact. The following charge (of the next sensor of the same color) will catch up until both charges merge. The both charges are transferred together to the contact.

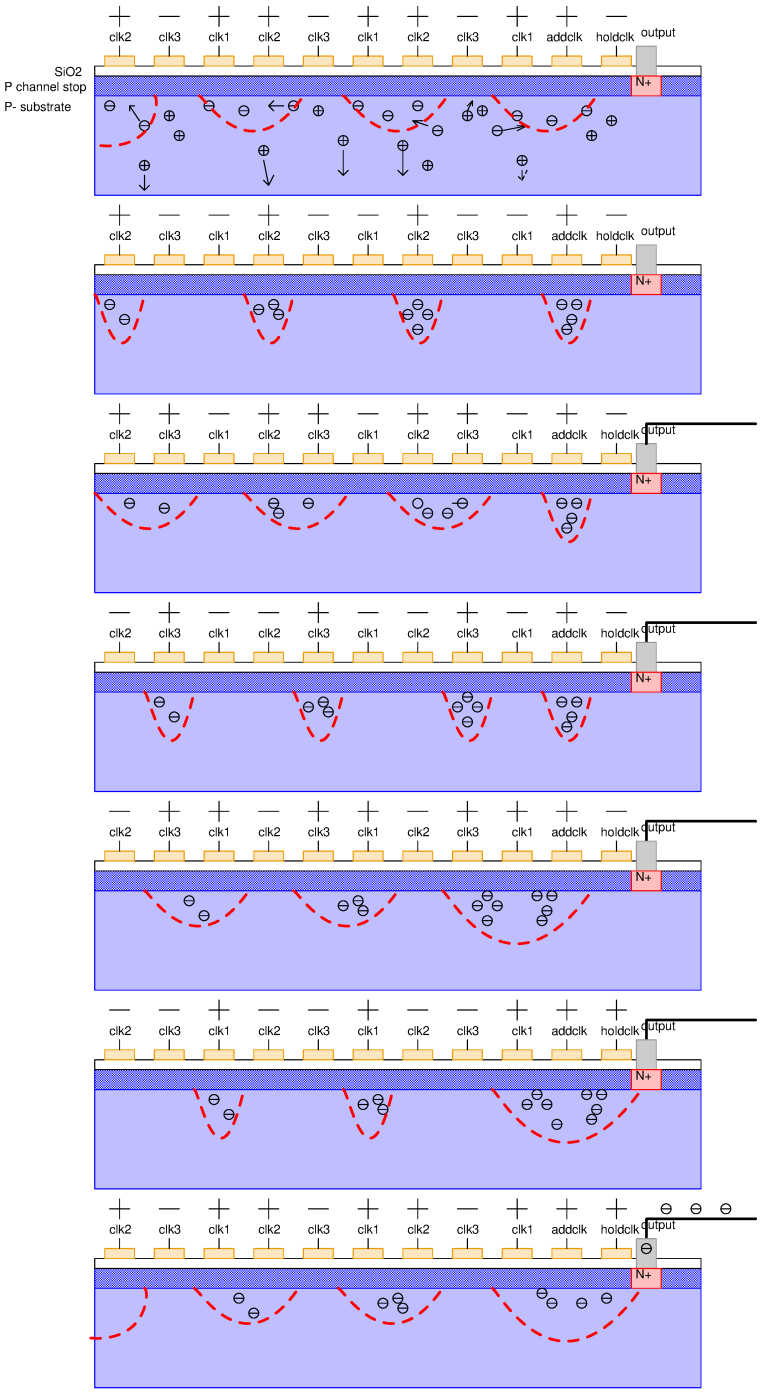


Figure 4.82: charge adding by letting two 'buckets' merge

If a two dimensional structure of gates is available it even is possible to let one 'bucket' pass an other and to add charges in an arbitrary sequence. Well, practical implementation however is somewhat complex. The paths of the charges could finally look a bit like a railway station.

4.8.13 Single Photon Avalanche Device

A single photon avalanche device (SPAD) is a photo diode that is operated at avalanche break down. These devices are used to detect very weak optical signals with a very short response time. If a photon is detected this triggers an avalanche. As soon as the signal is detected the avalanche has to be stopped again by reducing the voltage applied. This can be done passive using a resistor or by an active circuit. After triggering the avalanche the time required to stop the avalance and then return to the regular operation voltage again (so the device is ready to detect the next photon) is in the range of some ns to some 10ns.

Typical applications for SPADs are time of flight LIDAR systems.

4.9 HAL sensors

Hal sensors are used to measure magnetic fields. The HAL effect [28] is based on the Lorentz force. Every electrom moving in a magnetic field is exposed to a force perpendicular to the direction of movement and the magnetic field. This is described by the vector product.

$$F_B = e * (v \times B) \tag{4.112}$$

As soon as the electrons are deviating from the original direction of the movement an electrical field builds up. Reaching equilibrium this electrical field compensates the deflection of the electrons by the magnetic field.

$$F_E = E_H * e \tag{4.113}$$

$$F_B = F_E \tag{4.114}$$

Solving this equation leads to

$$E_H = v \times B$$

The speed of the electrons is determined by the electrical field in parallel with the movement direction of the electrons.

$$v = E * \mu \tag{4.115}$$

The following figure shows a typical HAL element. To avoid short circuiting the output voltage V_H by the electrodes supplying the element the HAL element usually is designed with small contacts far away from the measurement contacts. This leads to a diamond like shape of most HAL elements built on chips.

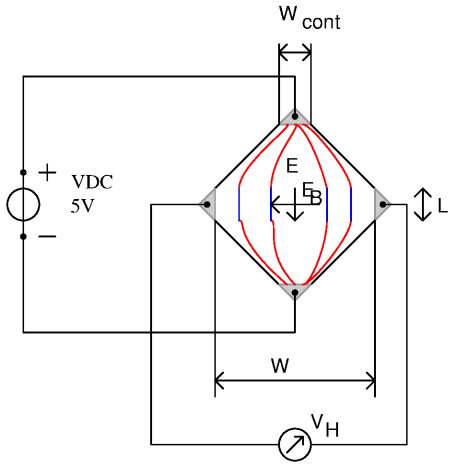


Figure 4.83: HAL element

Close to the supply electrodes the current flow is inhomogenous and the electrons may even move close to the saturation velocity close to the contacts (red areas). To achieve reasonable control characteristics the region where the electron reach saturation velocitiy must be kept short (directly at the contacts only) . In the middle of the HAL element the movement is fairly homogenous (blue area). Assuming we want to leave some margin to get a reasonably linear control characteristic to adapt the supply to changing resistivity the usual choice for the spped of the electrons at the contacts is about 50% of the saturation velocity. In the blue area the achieved speed becomes

$$V = m * v_{sat} * \frac{W_{cont}}{W} \tag{4.116}$$

m is the 50% margin factor. The HAL voltage measured is:

$$V_H = E_H * W \tag{4.117}$$

Typical ratios between the contact width and the width of the HAL sensor are in the range of 5..10. achievable HAL voltages depend mainly on the semiconductor material used and on the width W of the HAL element.

Table 15: Important parameters for the design of HAL elements

Material	v_{sat}	v_{sat}	typical sensitivity	scaled for mT and μm
Si	$10^5 m/s$	$1 * 10^3 m/s$ to $3 * 10^3 m/s$	$2 * 10^3 \frac{V}{m*T}$	$2 \frac{\mu V}{mT * \mu m}$
GaAs				

Example: To build a compass with HAL sensors we need a resolution of about $1\mu T$ (preferably even less). With a size W of $100\mu m$ the amplifier reading the HAL sensor must have an offset of less than 200nV.

4.10 General problems of high voltage components

Manufacturing high voltage components - no matter whether these are simple diodes, bipolar transistors of MOS transistors - require special considerations to produce reliable chips. In the following some, but for sure not all effects are described.

4.10.1 Single event burnout

Single event burn out is caused by an avalanche break down triggered by particle radiation. The particles are coming from radiation hitting the atmosphere (solar wind). These particles get scattered in the ionosphere and create secondary particles. These particles may hit the depletion area of a high voltage component. Typically heavy particles such a neutrons are the most dangerous ones to trigger an avalanche. To trigger a single event burnout (SEB) the field strength in the depletion zone must be high enough to produce an avalanche and an energy reservoir must be connected to the device that delivers enough energy to destroy it (charged capacitor at the supply rail!).

The probability of a SEB depends on:

1. Operating voltage of the device (rule of thumb: if the device is operated at less than 60% of it's maximum rating the avalanche will not run and there is no SEB)
2. Area of the device (The bigger the area the higher the chance of catching a particle)
3. activation energy (wide bandgap materials such as SiC are less sensitive)
4. altitude (At sea level there are less high energetic particles than at high altitude or in space)
5. solar activity
6. location in the magnetic field of earth (polar lights are caused by particles)
7. artificial radiation (for instance operation inside a reactor or a particle accelerator)

Recently strong lightning has been reported to produce roentgen and gamma rays as well. The correlation between SEB and thunderstorm activity has not yet been investigated, but I wouldn't exclude that there is a correlation.

4.11 Figure of Merit of a technology

To benchmark technologies many people rely on some kind of a figure of merit. Depending on the design the figure of merit however can be completely different from one project to an other!

4.11.1 Digital figure of merit

For digital designs the logic density is of highest interest. Usually this is the number of gates per area. Other parameters such as analog properties of a design usually don't matter for pure digital designs. In some cases in addition to the gates per area the memory bits per area may be important for logic designs as well.

For the gates per area it is reasonable to use some kind of a reference gate. A triple NAND is a reasonable choice. (It is worth while asking the technology provider to which gates the figure of merit refers!). In the following a rough estimation of the area of a NAND gate is shown.

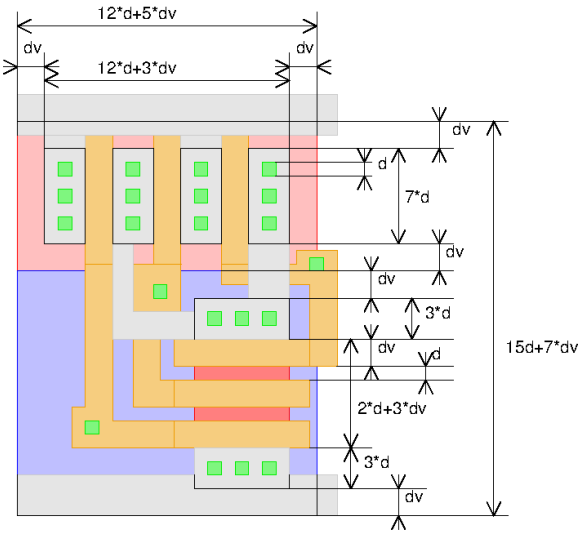


Figure 4.84: Example layout of a triple NAND gate

In the simplified layout (bulk contacts are omitted) shown above there are two important spacings. Distance d is the minimum feature size defined by the lithography. Usually the minimum contacts are defined by distance d.

All spacings of metal or poly and all overlaps over the contacts follow the lithography limit d as well. Typical values of d range from $1\mu m$ (typical 5V CMOS technology of the 1990s) to about $0.1\mu m$ (typical 1V CMOS technology of 2015).

All channel lengths and overlaps of wells over the transistors are defined by the break down capability of the semiconductor material and the operating voltage of the cell. Using a dv of:

$$dv = 0.1\frac{\mu m}{V} * Vdd$$

is a typical choice.

Since we want to have a drive strength of the NAND gate that is close to the drive strength of a minimum inverter the NMOS transistors should have about 3 time the minimum width. The PMOS transistor should have about the same strength as the tree NMOS transistors in series. So the PMOS transistors are 3 times wider than a minimum transistor as well. These considerations lead to a approximate cell size of:

$$A_{cell} = (15d + 7 * 0.1\frac{\mu m}{V} * Vdd) * (12d + 5 * 0.1\frac{\mu m}{V} * Vdd)$$

$$A_{cell} = (15d + 0.7\frac{\mu m}{V} * Vdd) * (12d + 0.5\frac{\mu m}{V} * Vdd)$$

This equation is a rule of thumb but it is already giving a reasonable estimation.

$$FOM_{logic} = 1/A_{cell}(d, Vdd)$$

Some Examples:

5V logic with minimum feature size $d = 0.8\mu m$

$$A_{cell5V} = 187.55\mu m^2$$

$$FOM_{logic5V} = 5332 gates/mm^2$$

3.3V logic with minimum feature size $d = 0.5\mu m$

$$A_{cell3V3} = 75\mu m^2$$

$$FOM_{logic3V} = 13333 gates/mm^2$$

1.8V logic with feature size $d = 0.2\mu m$

$$A_{cell1V8} = 14.1\mu m^2$$

$$FOM_{logic1V8} = 70922 gates/mm^2$$

1V logic with feature size $d = 0.1\mu m$

$$A_{cell1V} = 3.74\mu m^2$$

$$FOM_{logic1V} = 267380 gates/mm^2$$

These examples show the dramatic effect of decreasing the supply voltage and of using deep UV lithography and phase shift masks for logic technologies. From one logic generation to the next the number of gates increases by about factor 4!

$$FOM_{logic} \sim \frac{1}{vdd^2} \tag{4.118}$$

Limitations of this rule: This rule works well down to a supply voltage of about 1V. For channel leakage reasons the thresholds can't be made lower than about 0.5V. Therefore the minimum supply voltage is about 0.7V. This means here we come to a limit of standard CMOS technologies.

Further decrease of the cell size requires creating a vertical channel instead of a horizontal channel. This is already done using T-gates or vertical gates. A so called 35 nanometer technology has a gate that in fact is a poly trench of 35nm width but about 40nm deep. This leads to about 100nm real channel length.

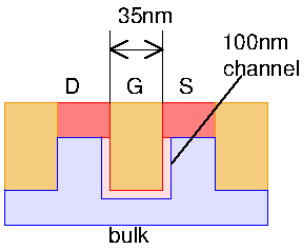


Figure 4.85: T-gate NMOS transistor

To decrease the overlap of the well over the drain and the source isolating the logic transistor with trenches (instead of junctions) helps. Using these two tricks Moore's law could be kept alive for some more years. The price semiconductor manufacturers had to accept was an explosion of equipment cost for these demanding processes.

Processes creating vertical channels don't work well for analog design anymore because most of the transistor is not defined by a drawn width and length anymore but by process parameters. Scaling current by modifying transistor sized doesn't work anymore.

Using trench isolation changes the calculation of the size of a standard triple NAND. The width of the trench is limited by lithography. So all the well overlaps now become d (opposed to dv using junction isolation). The same applies to the gate length. The voltage capability of a technology now is in the vertical direction.

$$A_{cell-trenchiso} = 22d * 17d = 374d^2$$

Example: A technology uses trench isolation and T-gates. d=35nm: $A_{cell} = 0.72\mu m^2$, $FOM_{trench35nm} = 1.38 * 10^6 gates/mm^2$.

The shrink path of a technology now is independent of the logic voltage. It only depends on the lithography and the capability of etching well controlled vertical trenches.

Using trench isolation and vertical gates or T-gates becomes beneficial at about d=35nm. Today (2018) technologies with feature sizes down to 14nm have been reported.

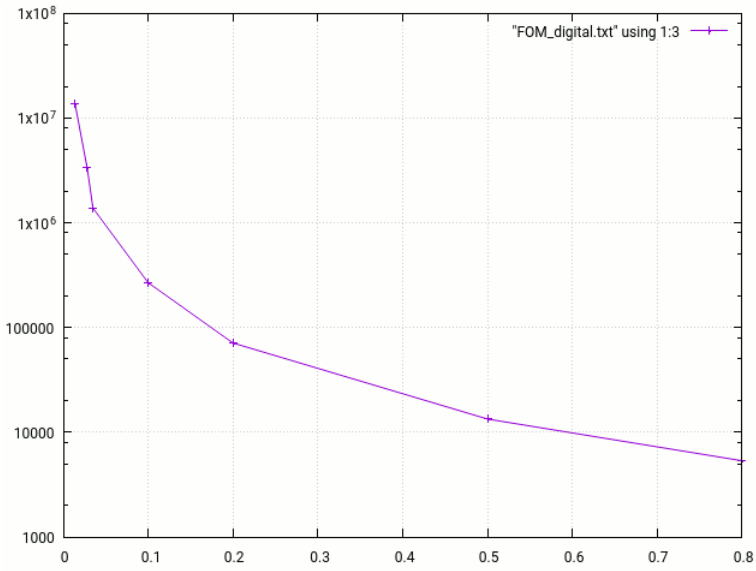


Figure 4.86: number of gate (NAND3) per square mm as a function of feature size (in μm)

Trench width usable for isolating transistors reaches it's physical limit when the electrons start tunneling through the gate oxide. This is about 2nm. So the minimum trench width needed for a gate is expected to be in the range of 5..6nm (we need a gate oxide on both sides of the T-gate plus some conducting fill acting acting as a gate in the middle).

5 Parasitic Components

Parasitic components are components designed unintentionally. Some of them can be found by layout back annotation. For simulation speed reasons in most cases back annotation only checks for linear components (resistors and capacities) and first order bipolar elements (diodes). In most cases stacks of diodes are not converted to the corresponding transistors by extraction tools!

In some design libraries special layout checks for parasitic MOS transistors are available. The use of parasitic MOS checkers usually requires labeling all high voltage nets!

5.1 Passive Parasitics

Wherever there are two nets there is a capacity between the nets. Wherever there is current flow there is a resistance and an inductance. Usually it depends on the impedances of the circuit and the frequency range which of these linear parasitc component matters most.

Low resistive circuits and power switches: Parasitic resistance and inductance is the highest risk. Typical metal trace resistance ranges from $40m\Omega/\#$ (Aluminium trace, 300nm thick) to $3m\Omega/\#$ (copper trace $15\mu m$thick). (See 4.1.2 ff).

High speed design and high impedance circuits: Parasitic capacities and inductances are the highest risk. Often calculating simple plate capacitors gives a first idea. Most dielectrics used on integrated circuit are in the reange $\epsilon_r = 4..12$. For pins and traces about 1nH/mm is a good first guess.

5.1.1 Parasitic Capacities

The highest area capacities found on most chips are the junction capacities betwen high doped layers of opposit doping. NMOS transistors with an isolated bulk always have a parasitic capacity from the pbulk to the nwell they are placed in. A classical example is the tail capacity of a differential amplifier.

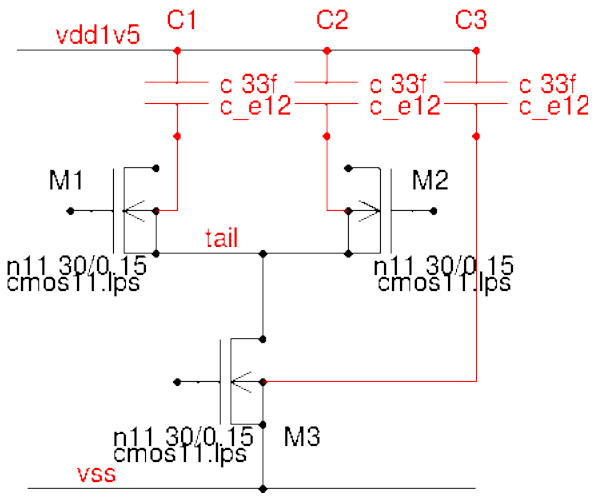


Figure 5.1: Example of parasitic bulk capacities in an NMOS differential amplifier

In the differential stage shown above the bulks of M1 and M2 are connected to the tail node. This frequently is done to avoid back gate effect (reduction of the gain by typically 30%) and to obtain best possible DC matching. C1 and C2 can carry supply noise into the amplifier and will lead to a gain boost at high frequencies because the tail node impedance is reduced for RF. Furthermore fast common mode signals at the inputs of the amplifier will lead modulate the differential mode gain because the current in both transistors changes. In extreme cases M1 and M2 will completely turn off if the common mode signal has a falling edge faster than M3 can discharge C1 and C2.

Most models used on schematic level only include the bulk of the MOS transistors but not the capacity between the bulk and the nwell. In some technologies C1 and C2 will get recognized running layout extraction. But that is a bit late to find out a circuit has an essential problem!

The same parasitic capacity exists in PMOS transistors. Usually there the bulk is an nwell sitting inside substrate. So there the parasitic bulk capacity goes to substrate. Keep in mind substrate often is not coincident with circuit ground! It strongly depends on where the substrate contacts are and how low resistive they are.

High voltage components: High voltage NMOS transistors use the nwell or an epitaxy region as their drain. The same applies to the collector of NPN transistors. The high voltage drain and the collector have a high capacity to substrate in ther range of 0.1pF to 1pF. High voltage NMOS transistors and NPN transistors are sensitive to substrate noise coupling into the drain or collector.

High voltage PMOS transistors usually are placed inside an nwell. This nwell often is coincident with the source of the HV-pmos. Substrate noise often gets coupled into the nwell of the HV-pmos transistor. HV-pmos transistors operating in common gate configuration can amplify the substrate noise!

PNP transistors have the base connected to the nwell. Substrate noise gets coupled into the base of PNP transistors.

5.1.2 Parasitic Inductances

Parasitic inductance starts to play a significant role on the chip above about 1GHz.

Usually the first parasitic inductance causing trouble is the ground bond wire. Depending on the number of pins it is in the range of 1nH (one pin) down to about 200pH (about 10 ground pins that are mutually coupled).

Traces on chip running over substrate or a ground plane can be regarded as something similar to a microstrip line. As long as the wave length is significantly longer than the chip size it can be regarded as a lossy inductor (losses caused by eddy currents in the substrate). When the wave length approaches the lenth of the wire it behaves like a strip line.

Above about 10GHz the on chip wires must be checked for inductance too.

5.2 Surface Parasitics

Surface parasitics are caused by the physics taking place at the surface of a chip. Usually MOS components are much more affected by surface parasitis than bipolar components that are depending on junctions burried somewhere inside the silicon.

5.2.1 Parasitic metal gate MOS transistors

Parasitic metal gate MOS transistors are possible wherever a metal trace or poly silicon (resistors!) crosses low doped silicon that is having a significantly different voltage than the metal. Metal gate MOS can have both polarities:

PMOS as well as NMOS. Metal traces crossing nwell or N-epi can create unintentional PMOS transistors. Metal traces crossing pwell can create a parasitic NMOS transistor. The following picture shows some typical examples:

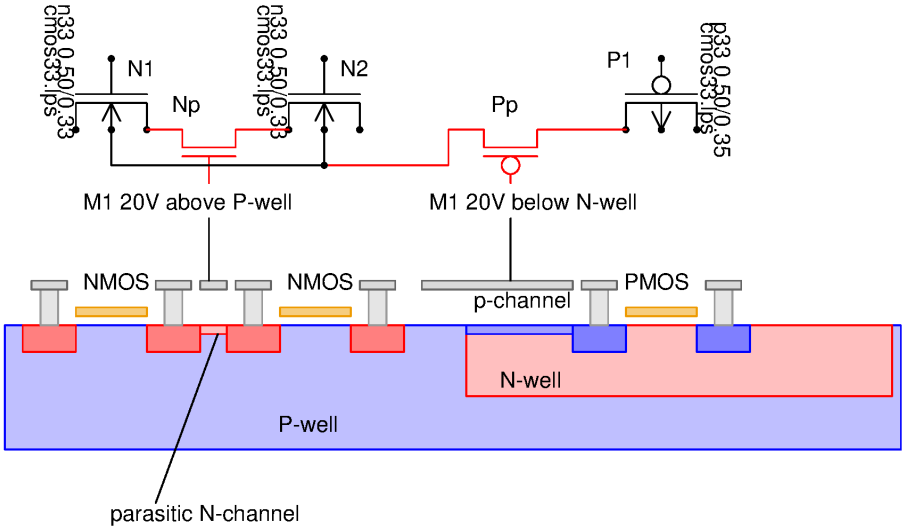


Figure 5.2: Examples of parasitic metal gate MOS transistors

The parasitic transistors created by the two metal traces are drawn in red in the schematic above the cross section. Np creates an unintentional connection between the designed transistors N1 and N2. Pp connects one node of P1 with the adjacent P-well.

To prevent parasitic metal gate MOS transistors several measures are possible:

1. Don't allow metal traces with high voltage over logic and low doped regions.
2. Interrupt parasitic channels with high doped guard rings. (drawback: outdiffusion can degrade matching of transistors)
3. shield the wells from the field of the high voltage traces by poly silicon field plates.
4. Use trenches to isolate the transistors instead of pure junction isolation (careful! trenches produce mechanical stress and degrade device matching.)
5. Take high voltage signals into higher metal levels. This increases the threshold of the parasitic MOS transistor.

Typical thresholds found in most processes are around:

Table 16: Parasitic MOS thresholds of poly silicon and metal gate MOS transistors

gate	oxide	bulk	threshold
poly	CVD 300nm	nwell, pwell, epi	10..20V
M1	CVD 300nm	nwell, pwell, epi	10..20V
M2	CVD 600nm	nwell, pwell, epi	15..30V
M3	CVD 900nm	nwell, pwell, epi	20..50V
M4	CVD 900nm	nwell, pwell, epi	25..60V
M5	CVD 1200nm	nwell, pwell, ep	30..70V

The table already shows that moving signals into higher layers only shifts the threshold of the parasitic MOS by a view Volt. Most of the parasitic MOS transistors are long channel devices following the shape of the metal trace. Therefore parasitic metal gate MOS transistors often remain undetected in low resistive circuits. In low current consumption designs however the situation changes and even metal gate MOS transistors with some nA can change the behavior of the circuit.

Characterization of parasitic MOS transistors often is poor. In low consumption designs I found parasitic MOS transistors influencing my parameters with thresholds of 8V while the process manual stated it has more than 15V! You can't be too careful!

5.2.2 Accumulation of ionic contamination (BTI and NBTI)

The surface of silicon has dangling bonds. These dangling bonds are attractive for ionic contamination. In case of classical MOS transistors the ions (in most cases sodium, H2) creap between the gates of the transistors and the channels. This changes the threshold voltage of the transistors. NMOS transistors with sodium ions between the gate and the channel have a lowered threshold. PMOS transistors with sodium ions between gate and channel have an increased threshold.

Sodium moves along the interface between silicon and silicon oxide. To protect against sodium the chip usually has an edge seal using a strip contact sou rounding the complete chip. The metal of this strip contact is connected to a low potential (usually substrate).

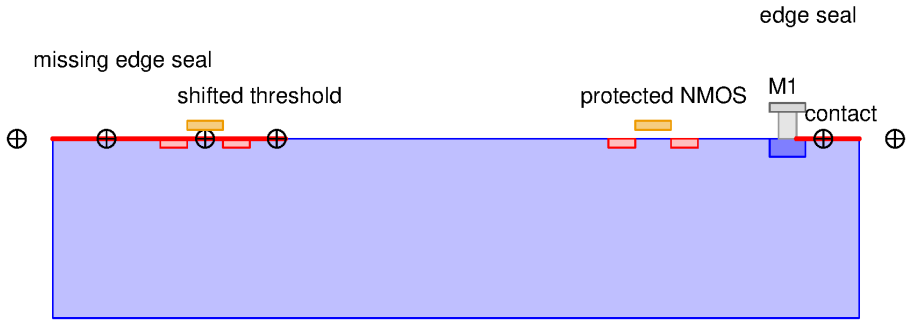


Figure 5.3: Edge seal

In the figure above the left side doesn't have an edge seal. The sodium can move into the transistor following the red colored path. On the right side sodium will be stopped by the contact interrupting the path to the transistor.

Pads: Pads lead to passivation openings. As long as the metal is wider than the passivation opening there is no path for sodium to reach the silicon-SiO2 interface. Well designed pads inside the edge seal aren't a problem.

Laser trim and zener zap: Laser trim cells require a passivation opening. After cutting the trim cell the silicon surface may be exposed. For this reason laser trim cells must be surrounded by an edge seal. The edge seal surrounding the laser trim cell often defines the final size of the trim array.

Zener zaps often break open the oxide due to vaporization of part of the aluminum. In many technologies the passivation has an opening over zener zaps to avoid propagation of cracks. This exposes the silicon surface. Zener zap cells must be surrounded by edge seals. In pure bipolar technologies violating this rule didn't cause much damage because the junctions defining the functionality are below the surface. As soon as the design includes MOS components zener zap openings must be fenced in by an edge seal.

Qualification: To verify correctness of the edge seal most chips during qualification are stressed in a hot, humid atmosphere operating at the highest possible voltage (no current flow is needed for this test). Usually the chips are data logged before stress and data logged after the stress. The test searches changes of transistor parameters (BTI and NBTI).

Production test: In recent years it became common to test that there is no crack in the edge seal of automotive chips. A correct edge seal is expected to behave like a metal resistor. The test basically measures the resistance of the edge seal from one corner of the chip to the opposite corner of the chip. If there is a crack in the edge seal the resistance of this path deviates from the expectation. The test has some issues because even if the edge seal has a crack the crack has a bypass resistance through the substrate. So the change of resistance can be very low.

N-substrate: Some technologies (mainly dedicated to pure high side driver applications) use an N-substrate. During operation this substrate is connected to the highest voltage. The positive polarization of the N-substrate pushes sodium out of the chip. For this reasons some N-substrate technologies don't require an edge seal.

5.3 Bulk Parasitics

Bulk parasitics mainly are caused by carriers flowing inside the silicon. Most of the bulk parasitics are related to junctions and their properties.

5.3.1 Parasitic substrate PNP

Vertical PNPs can be found in almost every technology. They usually consist of a P+ region acting as the emitter, a N well or an N-epi region surrounding the P+ region acting as the base and the P-substrate acting as the collector of the substrate PNP.

Especially if a high resistive substrate is used substrate PNPs in combination with an NPN structure often create a thyristor that can lead to a latch up.

Substrate PNP included in an NPN transistor: Every vertical NPN transistor of a classical bipolar technology has a parasitic PNP. The emitter of this PNP transistor is the base of the NPN transistor. The collector becomes the base of the parasitic PNP. The substrate is the collector of the parasitic PNP transistor.

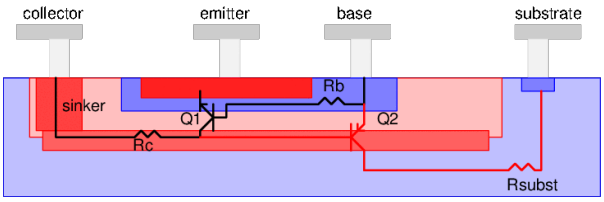


Figure 5.4: parasitic substrate PNP included in an NPN transistor

Q1 is the designed transistor. Q2 is the parasitic vertical PNP transistor. Q2 gets activated as soon as the NPN transistor reaches saturation. As soon as there is injection of holes from the base into the collector (Vcb<-0.6V) Q2 starts to pull the base current down into the substrate. The properties of the parasitic PNP transistor strongly depend on the thickness and the doping of the buried layer. The current gain of such a PNP transistor can range from about 0.1 (thick, high doped buried layer) to 300 (no buried layer).

Due to the path resistance Rsubst the activation of Q2 can locally pull up the substrate under the transistor. Rsubst plays an important role at latch up events.

A reasonable model of an NPN transistor on a chip is always a subcircuit holding both, Q1 and Q2.

Substrate PNP in CMOS logic: In CMOS logic the most important substrate PNP transistor consists of the drain or source region of the PMOS transistors acting as an emitter, the nwell acting as the base and the substrate acting as the collector.

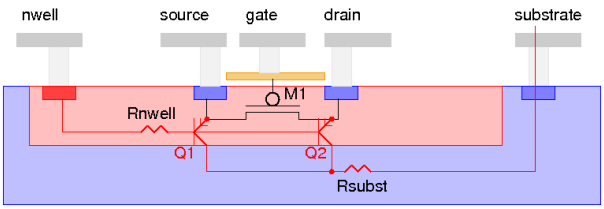


Figure 5.5: parasitic PNP included in CMOS logic

The parasitic transistor Q1 or Q2 gets activated when the drain or the source of the PMOS transistor is pulled above the supply of the nwell. Usually CMOS technologies don't have a buried layer inside the nwell. Furthermore the nwell often is only a few μm thick. Therefore the transistors Q1 and Q2 often have a fairly high current gain. Values reaching from B=3..100 are common. The worst case I have seen (at hot) was B=900!

Q2 is the reason why CMOS outputs may not be pulled above the supply of the logic.

If the drain or the source are pulled higher than Vnwell with a resistive source Q1 and Q2 will limit the voltage typically at Vnwell+0.6V

To reduce simulation time Q1 and Q2 often aren't modeled! This is a high risk designing charge pumps using CMOS components as rectifiers.

Substrate PNP in a high voltage NMOS transistor In high voltage power NMOS transistors usually the source and the bulk are connected by a shared contact. This leads to a P-emitter at the source of the device, The drain (usually N-epi) acts as the base of the parasitc PNP transistor. The substrate becomes the collector of the parasitic PNP. The following figure shows a classical vertical DMOS power transistor.

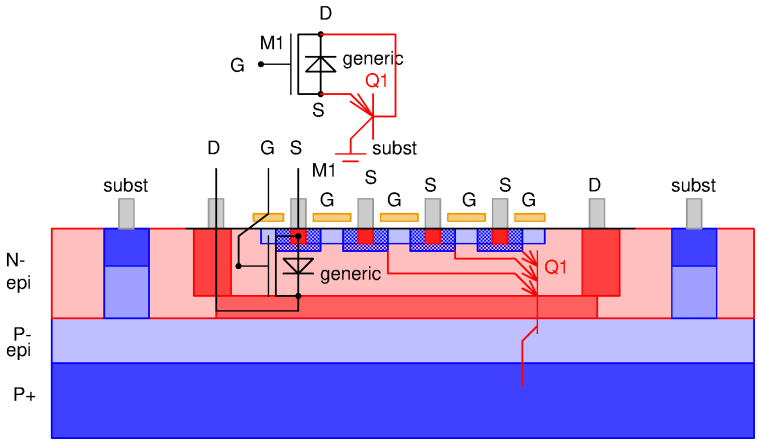


Figure 5.6: parasitic PNP included in a DMOS transistor

The parasitic PNP transistor gets activated as soon as the source (bulk) is pulled one diode forward voltage above the drain. Classical vertical DMOS transistors have a buried layer. So the current gain of Q1 remains in a range of B=0.03..0.3 depending on the properties of the buried layer.

In lateral DMOS transistor (LDMOS) the buried layer often is omitted for cost reasons. In this case the current gain of Q1increases significantly to B=0.3..30.

Designing power bridges that are driving inductive loads the power dissipation of the parasitic transistor Q1 can be magnitudes higher than the power dissipation of the designed power transistor M1.

Most models of power transistors include the bulk diode. Often the parasitic PNP Q1 is omitted or forgotten! Sometimes Q1 is part of the bulk diode model.

5.3.2 Parasitic lateral PNP transistors

Lateral PNP transistors can be found were ever there are P-doped regions close to each other sitting an an nwell or N-epi. If one of these P-doped region is pulled one forward voltage above the nwell potential the lateral transistor gets activated. The gain of such a lateral PNP depends on the spacing of the involved P-regions. The most frequently found lateral transistor is in analog multiplexers using PMOS transistors as switches. Here the base width often is in the range of only a μm or vene less (minimum lenth of the PMOS transistor). The short distance between the drain and the source of an analog switch makes the associated lateral PNP transistor quite efficient. Of course the vertical PNP down to substrate gets activated as well. The current distribution between the lateral PNP and the vertical PNP depends on the thickness of the nwell and the lateral distance of the collector(s) from the emitter.

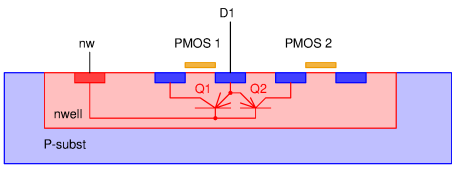


Figure 5.7: Lateral PNP bypassing two PMOS transistors

In the above figure it is assumed that the node D1 is pulled one forward voltage above node nw. This activates the parasitic lateral PNP transistors Q1 bypassing PMOS1 and Q2 creating a connection between PMOS1 and PMOS2. For simplicity the vertical PNP from D1 (acting as the emitter) to P-subst (acting as a third collector) is not shown.

5.3.3 Parasitic lateral NPN

Lateral NPN transistors can be found wherever there are N-regions embedded in a pwell or P-substrate. If one of these N-regions is pulled negative versus the isolating P-well of P-substrate a lateral NPN transistor gets activated. Usually these parasitic lateral transistors have a current gain below 1. Nevertheless these parasitic transistors can have a severe impact on circuit functionality if the node pulled negative (emitter of the parasitic transistor) is driven from a powerful source (compared to the impedance at the collector of the parasitic transistor).

Parasitic NPN in a transmission gate: Transmission gates use MOS transistors as switches. Often transmission gate are used for analog multiplexers. The output of the transmission gate usually is connected to a high impedance input of a measurement system (could be a buffer amplifier or an ADC input). In some cases the driving side ha a high impedance as well (for instance internal reference generators such as bandgaps). Even currents in the $\mu A-$range at the "collector" can already lead to fatal results ranging from wrong measurements to complete failure of the chip.

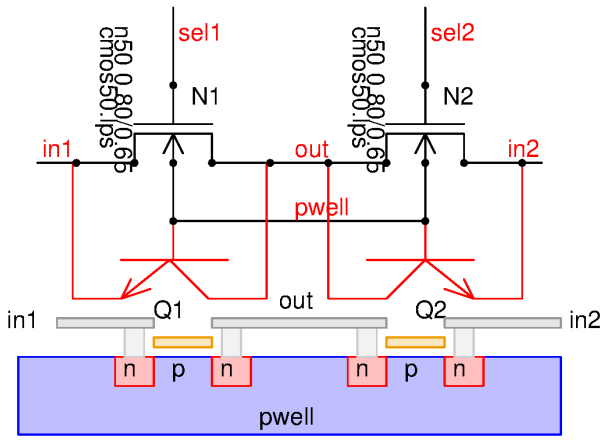


Figure 5.8: Lateral NPNs inside an NMOS analog multiplexer

Usually the pwell acting as the base of the lateral NPN transistor is connected to ground in a low resistive way. If one of the nodes in1 or in2 is pulled negative the parasitic bipolar transistors Q1 or Q2 can easily override the designed NMOS transistors M1 or M2. The signal at node out will go low even if the gain B of the NPN transistors is less than 1 (It only is a question of the base impedance respective pwell resistance and the emitter spread resistance. Usually these signals are low resistive compared to the channel resistance of the MOS transistors). The analog multiplexer will fail if one of the parasitic bipolar transistors gets activated. For proper operation the n-active regions may never drop below the voltage of the p-bulk.

Lateral NPN transistors inside power bridges: Power bridges have a very similar problem regarding parasitic NPN transistors. The currents flowing in the power bridge can be in the range of several Amperes. The driver stages operate in the range of some 10 to some hundred μA. Even lateral transistors with a gain of B=0.00001 to B=0.0001 can already affect the driver stages (of the own bridge or an other bridge). The only difference is, that the base of the parasitic NPN transistor now is the substrate instead of an pwell used as the bulk of an NMOS.

The following figure shows the typical case of a power IC with a bridge output stage.

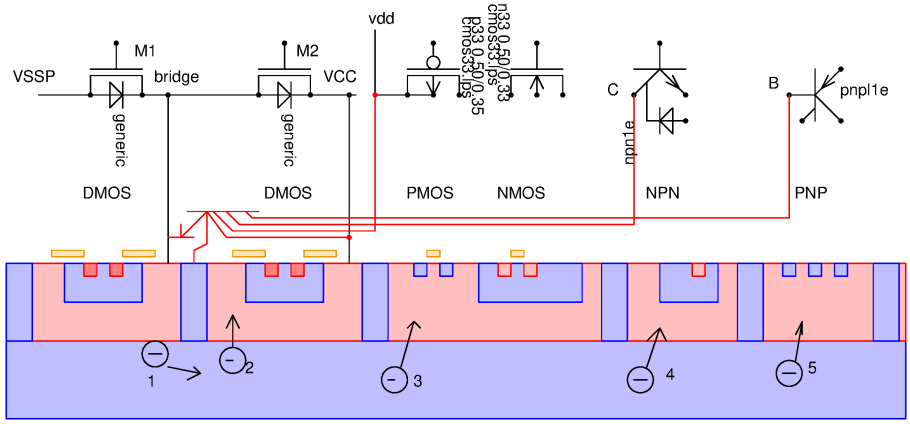


Figure 5.9: Lateral NPN with the substrate acting as the base

If the drain of M1 is pulled below the substrate voltage the N-epi of M1 acts as an emitter. Electrons will be injected into the substrate (1). Most of these electrons only will travel a short distance and move into the drain of M2 (2). This causes an increased current consumption of the power bridge during flyback events.

Electrons traveling a bit further will reach the bulk of the PMOS transistor (3). This increases the current consumption at the logic supply vdd.

Electrons reaching the collector of an NPN (4) will create an additional collector current. This can take sensitive circuits like bandgaps out of operation.

Electrons reaching the base of a lateral PNP transistor (5) will turn on the transistor in an undesired way.

The gain B of the parasitic lateral NPN transistor depends on the distance the electrons have to travel. Short paths just to the adjacent device may have gains in the range of B=0.1 to 1, $\alpha = 0.05..0.5$. Longer distances in the mm-range typically have a gain of B=1/1000. This sounds harmless, but it isn't! Imagine the power bridge has to carry a flyback current of 1A activating the parasitic lateral NPN and the collector of the NPN transistor has to operate at $10\mu A$. The function of the circuit the NPN transistor is a part off will at least degrade if not even fail completely.

To reduce the impact of the lateral NPN several measures are possible:

1. Increase recombination of the electrons in the substrate using a P+ substrate material.
2. create a drift field in the substrate by grounding substrate only at the edge of the chip. (Drawback: floating substrate in the middle increases the risk of RF coupling via substrate and latch up)
3. Use a triple well process to shield the NPN and PNP transistors from electrons in the substrate
4. Use a process with burred oxide to isolate (ABCD process, smartmos 10 or similar. More costly!)
5. Use circuits that avoid high resistive NPN collector nodes and PNP base nodes (Drawback: Often these circuits compromise accuracy)
6. Move logic far away from the power stages to minimize current consumption at vdd

More details and empirical results from BCD technologies can be found in [74].

5.3.4 Parasitic vertical NPN

A parasitic NPN transistor can be created in all twin well processes that have a shallow pwell sitting inside a deep nwell. The source and the drain of the NMOS transistor acts as an emitter. The shallow pwell becomes the base and the deep nwell becomes the collector of the parasitic NPN transistor.

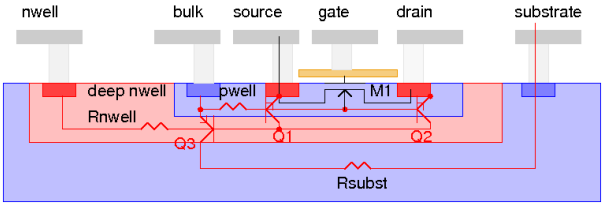


Figure 5.10: NMOS transistor (M1) and associated vertical NPN transistors (Q1, Q2)

The parasitic NPN transistors get activated as soon as the drain or the source of the NMOS transistor are pulled one forward voltage below the pwell. Since the N-active regions usually are much higher doped than the pwell the transistors Q1 and Q2 usually have a very high current gain (B=10..300 are common parameters). In many processes the parasitic NPN in fact is identical with the NPN transistor belonging to the design library (Well, the shape of the emitter usually is optimized, but the dopings often are identical)

Vertical NPN transistors have already been used intentionally to design bipolar bandgaps if no NPN transistor was available in the design library.

Warning about simulation models: To speed up simulation the junction between the N-active and the pwell often only is modeled as a simple diode. The collector of the vertical NPN transistor often is omitted to reduce netlist complexity. If you design charge pumps with CMOS transistors acting as synchronous rectifiers it is recommended to add an NPN transistor from the design library to the simulation schematic to be on the safe side simulating the charge pump start up. Once the charge pump is proven to start correctly the NPN transistor can be taken out again (If you miss taking out the NPN transistor after finishing the simulations the layouter will place it!)

5.3.5 Substrate resistance

Under normal circumstances most chip designer assume substrate to be a single node. In reality substrate is a resistive network! Regarding substrate a single node is a comfortable simplification that only is valid as long as the currents injected into substrate are only in the range of some μA. Substrate resistivity depends on the doping. It ranges from about $10m\Omega * cm$ to some $\Omega * cm$. There are two classic approaches to choose the substrate resistance:

1. small signal RF designers often prefer high resistive substrate to be able to use local guard rings for noise decoupling. ($1\Omega cm$ to $10\Omega cm$)
2. High voltage designs try to have the depletion zone in the substrate to allow a thinner epitaxy. This makes the process cheaper. ($3\Omega cm$ to $30\Omega cm$)
3. Logic designes ofter prefer a low resistive substrate because this reduces the effort to protect against latch up. (about $10m\Omega cm$)
4. Low voltage high current designs often use low resistive substrate to take benefit of the fast recombination of minority carriers (about $10m\Omega cm$. This way parasitic lateral bipolar transistors have lower gain.)

The connection of the substrate to the dice pad plays an important role too. The interface to the dice pad depends on the mounting technique and the doping of the bottom side of the chip.

High doped substrate in combination with back side metallization and conducting glue or soft solder provides an ohmic contact.

Low doped substrate together with conducting glue leads to a very poor Schottky contact to the dice pad. Usually most of the area is covered with oxide and only a few spots (where the oxide got scratched during the handling at the packaging site) really act as a Schottky diode. For high frequencies the oxide capacity is the main current path from substrate to the dice pad. The thickness of the oxide depends on storage conditions of the wafers before packaging. Usually the wafers are simply stored at room temperature in dry environment. This leads to just a few nm to some 10nm of oxide between the substrate and the dice pad. The resulting capacity can be in the range of several nF.

As long as we don't need to consider skin effect a resistor and capacitor network is a reasonable approach for majority carrier injection (capacitive noise injection or resistive coupling, but no activation of parasitic transistors).

5.3.6 Well resistance

The well resistance consists of two parts:

1. The contact spread resistance
2. The resistance of the path from the contact to the place of interest

Since most wells are only some micrometers deep the integration to calculate the contact spread resistance only runs from the contact radius (in case of a single, more or less round or rectangular contact) to the depth of the well. Everything further away than the depth of the well can simply be calculated similar to a film resistor.

If there are double wells (for instance a shallow pwell inside a deep nwell) the remaining thickness of the deep well minus the thickness of the pwell sitting in it must be used.

If the voltage difference between the wells is high high the depletion zone must be subtracted additionally.

The same applies in there is a big voltage difference between the nwell and the substrate.

The following figure shows an example how to approximately calculate the resistance of an nwell.

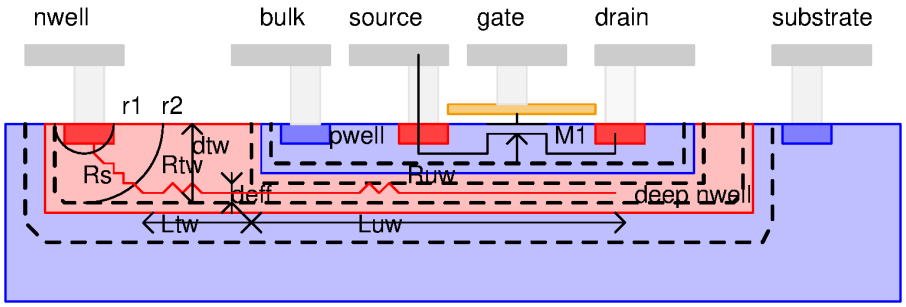


Figure 5.11: Example of a path calculation

The well resistance consists of 3 parts:

1. The contact spread resistance Rs
2. The path from radius r2 to the beginning of the depletion zone of the pwell represented by Rtw
3. The path under the pwell to the position of interest Ruw

The contact spread resistance Rs is calculated integrating the differential resistance from r1 to r2. Whether we have to use equation (21) or (24) depends on the shape of the contact.

From r2 to the edge of the depletion zone we have to use the thickness dtwell of the resistor Rtw and it's width W. (r is the resistivity of the deep nwell).

$$Rtw = \frac{r * Ltw}{W * dtw}$$

Under the pwell the effective thickness of the deep nwell is pinched to deff. The resistance under the well Ruw calculates as:

$$Ruw = \frac{r * Luw}{W * deff}$$

The effective thickness deff can be significantly less than the depth of the deep nwell. The thickness of the depletion zones (marked by the dashed lines) depends on the dopings of the wells and the voltages of the deep nwell and the pwell.

Rs and Ruw can be surprisingly high!

5.3.7 Thyristors

Thyristors can be created wherever there is a sequence such as PNPN or NPNP. The most feared thyristor is found inside CMOS logic. If this thyristor triggers it shorts the logic supply.

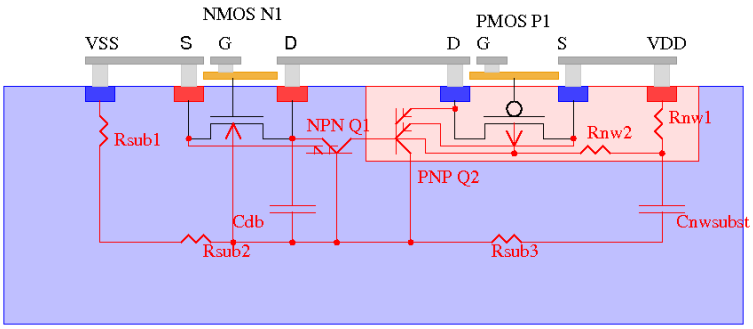


Figure 5.12: Cross section of a classical CMOS process

In the figure above there are two bipolar parasitic transistors Q1 and Q2. In classical CMOS designs the biggest concern is latch up triggered by the activation of Q1 and Q2. There are two possibilities to activate the thyristor:

1. Pulling node D above VDD first turns on Q2. The collector of Q2 pulls up the base of Q1. As soon as the drop over (Rsub1+Rsub2) reaches Vbe the transistor Q1 will turn on too and the thyristor switches ON.
2. Pulling node D below VSS first turns on Q1. The collector of Q1 pulls down the base of Q2. As soon as the drop over (Rnw1+Rnw2) reaches Vbe of Q2 the thyristor switches ON.

To keep the thyristor conducting the following two conditions must be satisfied:

$$Ic_{Q2} > \frac{Vbe_{Q1}}{Rsub1 + Rsub2} \tag{5.1}$$

$$Ic_{Q1} > \frac{Vbe_{Q2}}{Rnw1 + Rnw2} \tag{5.2}$$

The higher one of these two currents determines the trip point.

$$I_{trip} = max(Ic_{Q1}, Ic_{Q2}) \tag{5.3}$$

Both transistors involved have two emitters. One of them going to node D, the other ones going to nodes VSS and VDD. When the overvoltage event at node D ends the alternative emitters will take over. The thyristor remains conducting as long as the two conditions stated above are satisfied. In CMOS terminology this is called a latch up.

Classical CMOS logic (for example the 74Cxx or the 4xxx series) has a latch up specification. This latch up specification states the current needed at node D to trigger the thyristor and states a hold current flowing at nodes VDD and VSS to keep the thyristor in ON state.

The trigger current and the hold current usually differ because the geometries of the emitters (and therefore the gain B of the different partial transistors Q1 and Q2) differ.

Since the base emitter voltage decreases with temperature the trigger current and the hold current usually are specified at the highest operating junction temperature of the chip.

To make the trigger and hold currents as high as possible the following measures can be taken:

1. Use a P+ substrate to lower Rsub2.
2. Use a sinker to reduce Rnw1 (provided the technology offers this option).
3. Maximize the size of the well ties.
4. Make the contact of the substrate as big as possible to reduce Rsub1. Add as much P+ doping as possible to the substrate contacts
5. Place all contacts as close to the CMOS transistors as possible.
6. If possible move the substrate contact and the nwell contact in the middle between the involved CMOS transistors to eliminate Rsub2 and Rnw2.
7. Use a buried layer to reduce the gain of Q2 (provided the technology offers this option).
8. Place a deep trench of a deep P+ iso between the NMOS transistor and the nwell to reduce the gain of Q1 (provided the technology offers this option).

9. Maximize the distance between the NMOS transistors and the PMOS transistors to increase the base width of Q1.

10. Limit the current flow in node D by adding protection resistors to your circuit.

Some of these measures however will have an impact on transistor matching! (mechanical stress, outdiffusion)

I addition you can try to limit the supply current at VDD to a value below the expected hold current of the thyristor. Using this measure the thyristor still can be triggered, but once the supply has collapsed the thyristor will turn off again. The drawback of this trick is that the supply might collapse due to a load current transient of the logic in operation as well making your logic fail in application. Use this trick with care!

Futher trigger paths: Besides triggering the thyristor by pulling node D out of the supply range there are further possible trigger paths.

- The nwell (base of Q2) can be pulled down by a lateral NPN transistor such as the drain of a power DMOS transistor that gets pulled below VSS (substrate) by an inductive load.
- The substrate (base of Q1) can locally be pulled up by a vertical PNP transistor of an adjacent power stage.

To prevent these trigger paths the following layout strategies will help:

1. Keep power stages as far away from the CMOS logic as possible (increase the base width of lateral NPNs of low side drivers)

2. Use solid substrate grounding where ever you expect a vertical PNP transistor (high side drivers of power stages)

3. Use buried oxides if the technology provides this option to eliminate parasitic transistors that can trigger the thyristor.

5.4 Package Parasitics

In most cases the chip is mounted inside a package. The path from the chip to the solder joint outside of the package consists of the pin and the bond wire. Usually the pin is much wider than the bond wire. So the pin is the main contributor to the pin capacity while the bond wire mainly contributes to the inductance and the resistance of this path.

5.4.1 Bond wire resistance

Well, in reality we should have a look at the resistance of the pin and the bond wire. Usually the pin is much wider (typically we find a width of about 0.1mm, a thickness of 0.1mm and a length of 0.3mm leading to a resistance of about $1\mu\Omega$ to $3\mu\Omega$.)

The bond wire resistance depends on the diameter D of the bond wire and the length l. The resistance calculates as:

$$R_{wire} = r * \frac{4 * l}{D^2 * \Pi} \tag{5.4}$$

The resistivity of the most important bond wire materials at room temperature are:

Table 17: Bond wire specific resistances

Material	resistivity	unit
copper (Cu)	16.78	$\mu\Omega * mm$
gold (Au)	24.4	$\mu\Omega * mm$
Aluminum (Al)	28.2	$\mu\Omega * mm$

The diameter used depends on the current flowing though the bond wire. For 1.5A a typical choice is D=$30\mu m$ for copper bond wires and for gold bond wires.

Example: A 1mm bond wire made of copper with a diameter of $30\mu m$ has a resistance of $23.7m\Omega$.

Skin Effect: For high frequency this calculation is not fully correct anymore. The magnetic fields forces the current to flow at the surface of the conductor. The skin effect increases the resistance of the conductor. The skin depth approximately calculates as [52]:

$$\delta = \sqrt{\frac{r}{f * \pi * \mu}} \tag{5.5}$$

For convenience this equation in literature often is normalized for copper wires and the frequency in MHz.

$$\delta = \frac{66\mu m}{\sqrt{f(MHz) * k * \mu_r}}$$

with k being the ration of the conductivity of the metal used and the conductivity of copper. k is the ratio of the resistivity of the material used and the resistivity of copper.

$$k = r_{copper}/r_{used}$$

μ_r is the relative permeability of the conductor. In most cases μ_r is close to 1. However if iron wires are used μ_r deviates from 1 quite a lot!

Of course for copper k becomes 1 (The equation is normalized for copper) and μ_r is close to 1 as well and the equation simplifies a lot more..

Example: Skin depth of a copper wire at 100MHz:

$$\delta_{co} = \frac{66\mu m}{\sqrt{100 * 1 * 1}} = 6.6\mu m$$

For round wires the relative change of the rsistance can be neglected as long as the diameter of the wire is less than 3 times the skin depth.

$$\frac{R_{AC}}{R_{DC}} \approx 1$$

If the ratio between the wire diameter and the skin depth exceeds 3 $\frac{d}{\delta} > 3$ the change of resistance becomes about:

$$\frac{R_{AC}}{R_{DC}} \approx \frac{d}{4 * \delta} + 0.25$$

For typical bond wires the skin effect starts to significantly contribute to the real part of the impedance above 100MHz. At about 300MHz the resistance of a 25μm copper bond wire doubles (from 35mΩ/mm to about 70mΩ/mm) due to skin effect.

5.4.2 Bond wire inductance

Classically an inductance always is related to a magnetic flux through a closed current loop. A bond wire is a conducting piece of wire and the loop gets closed by the other pins. So the normal calculation doesn't work anymore. In stead the inductance can be approximated using field solver such as fasthenry or regarding the bond wire as a piece of strip line over the ground plane of the board. A bit dirty but the best we can do.

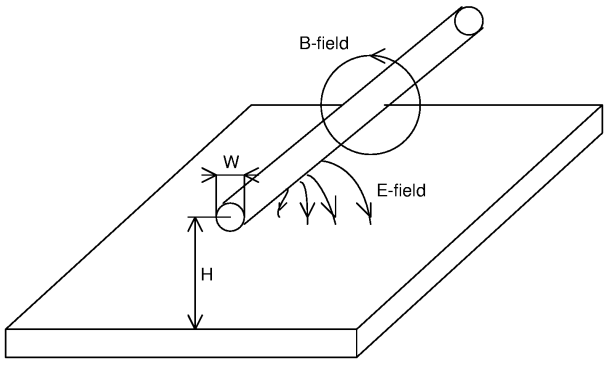


Figure 5.13: Bond wire aproximated as a microstrip line

The width W of the microstrip line representing the bond wire is the same as the thickness Tmet. A simple strip line calculator such as wcalc delivers the following parameters regarding the bond wire as a strip line 1mm above the ground plane:

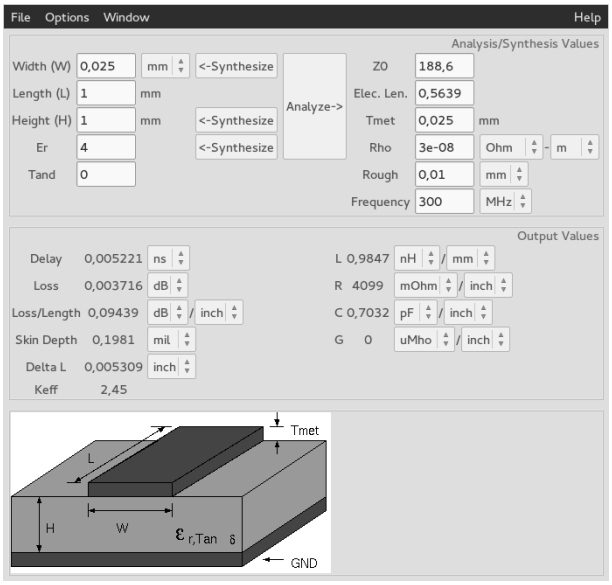


Figure 5.14: Calculation of a bond wire using a strip line approximation

The most simple model of such a bond wire looks like this:

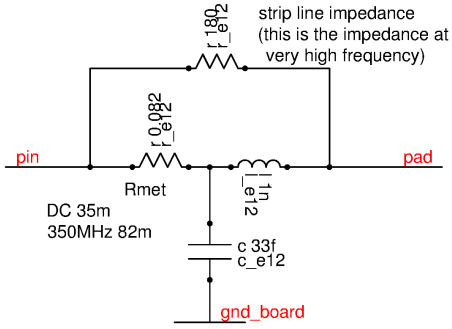


Figure 5.15: A simple bond wire model derived from the strip line approximation, fitted for 300MHz

This model is not a correct equivalent circuit. It is an approximation.

5.4.3 Pin capacities

Most of the stray capacity of the pin of an IC is connected to the two adjacent pins. The pin capacity mainly depends on the frame. Smaller packages usually have lower pin capacities. Corner pins in most cases have a lower capacity.

The most simple way of estimating pin capacities is to regard the two facing sides of the pins as two plates. This approach approximately underestimates the stray capacity by about 30%. Nevertheless it is a first guess.

Since a pin is a 3-dimensional shape more precise calculation can be done using field solvers.

The most simple way is to simply look up the package data published by various sources.

Table 18: Typical pin capacities of various packages

package	center pin	corner pin	source
QFP	1pF	0.6pF	[75]
LLP	0.45pF	0.2pF	[75]
Mini SOIC	0.03pF	0.03pF	[75]
SC-70		0.06pF	[75]
PLCC	0.6pF	0.45pF	[75]
SSOP	0.27pF	0.1pF	[75]
MDIP	1.1pF	0.4pF	[75]
Micro SMD		0.012pF	[75]
CSP	0.03..0.08pF	0.03..0.08pF	[75]
PBGA-208	0.1..0.15pF	0.15..0.35pF	[75]

Usually the pin inductance and the capacity of the ESD protection limits the bandwidth of a system long before the pin capacity starts to act as a low pass filter.

5.4.4 Pin Inductance

Pin inductance can be treated similar to the bond wire inductance. of course with W, thickness Tmet an possibly the hight over the ground plane migth be different. A typical SSOP pin has something like W=0.3mm, Tmet=0.2mm. The spacing between the pins can be less than the thickness of the metal. Therefore it makes sense to take into consideration the adjacent pin.

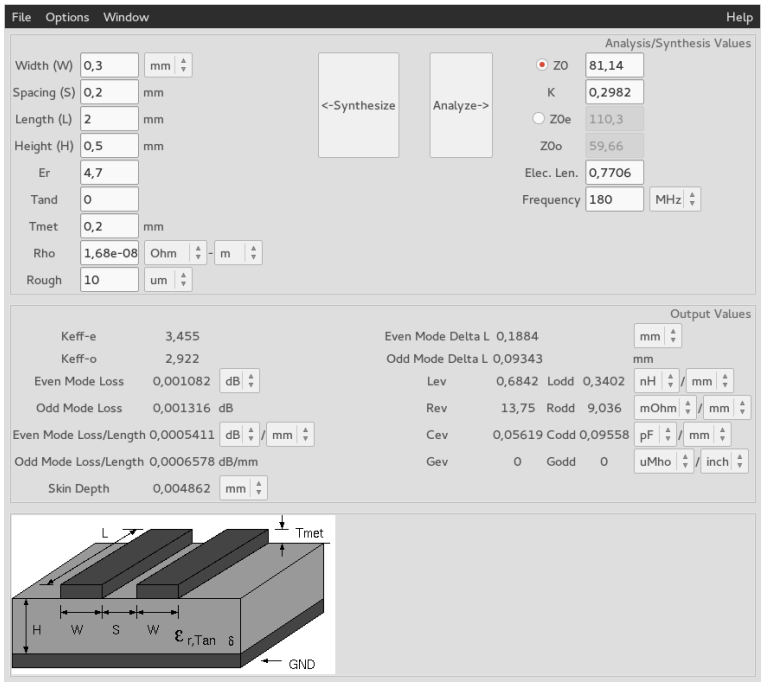


Figure 5.16: calculation of the inductance of a pin.

Due to the close proximity the result depends on the signal in the neighbouring trace. If the signal in the neighbouring trace has opposit current direction the inductance of our example is 0.34nH/mm. (odd mode). If the current in the neighbouring pin has the same direction the inductance of the example is 0.684nH/mm. This is caused by the magnetic coupling of both pins.

Since pins are significantly thicker than bond wires skin effect and proximity effect start to become significant at far lower frequencies.

Well, for those who need fast first guesses here are some typical numbers found in literature:

Table 19: Typical pin inductance of various packages

package	center pin	corner pin	source
QFP (12x12)	2.9nH	2.4nH	[75]
QFP (20x14)	4.5nH	2.4nH	[75]
QFP (28x28)	12nH	8nH	[75]
LLP	0.008nH	0.008nH	[75]
Mini SOIC	0.45nH	0.45nH	[75]
SC-70	0.45nH		[75]
PLCC 28	4.4nH	3.3nH	[75]
SSOP 28	2.9nH	1.3nH	[75]
MDIP 14	7.0nH	3.0nH	[75]
Micro SMD		0.011nH	[75]
CSP	0.4..0.6nH	0.4..1.0nH	[75]
PBGA-208	6..9nH	2.5..5.5nH	[75]
PDIP14		5.6nH	[76]
CDIP14		6.8nH	[76]
SO 14		2.2nH	[76]
PDIP20		8.0nH	[76]
CDIP20		25.6nH	[76]
SO20		4.3nH	[76]
PDIP40	3.5nH	19.2nH	[76]
CDIP40	10.5nH	48.7nH	[76]
PCC 52		6.0nH	[76]
PCC 68		7.2nH	[76]

Of course these are approximate numbers because the length of the bond wire plays an important role. And this bond wire length changes with the size of the chip and the size of the package. As a general trend it can be observed the larger the package the higher the parasitic inductance.

Ceramic packages tend to have higher parasitic inductance because usually in ceramic packages longer bond wires are permitted.

5.4.5 Die pad capacity

If the die pad is isolated the stray capacity (usually to board ground) should be considered. The capacity strongly depends on the thickness of the mold and on the spacing between the package and the ground plane (air gap between package and ground plane has $\varepsilon_r = 1$ while the mold has typically $\varepsilon_r = 4.7$.)

Example: SO package directly sitting on the ground plane. Area 4mm*4mm, Mold thickness 1mm, $\varepsilon_r = 4.7$.

$$C_{diepad} = \frac{A * \epsilon_0 * \epsilon_r}{d} = 0.66pF$$

5.4.6 Exposed die pad inductance and substrate inductance

Exposed die pads are mainly used to transport the heat produced by the power dissipation of the chip to the ground plane of the board. As a side effect an exposed die pad additionally acts as a very efficient substrate grounding. There are several cases to be distinguished:

chip soldered to the die pad: If the chip is soft soldered to the die pad the conditions are well defined. The soft solder galvanically connect the bottom side of the chip to the die pad. This has the following advantages:

1. Well defined coupling between die pad and substrate of the chip (it simply is a short).
2. Best possible cooling through metal connection.

Disadvantages of soft solder are:

1. More expensive than glue because back side metal is an additional process step.
2. Soft solder is more brittle than glue and tends to delaminate at thermal cycles.

chip glued to the die pad: This is the case found more frequently in IC manufacturing. Gluing the chip to the die pad offers advantages for production:

1. glue doesn't require a back side metalization of the chip (one step less in production, cheaper)
2. glue remains a little bit elastic. Delamination can be be controlled better than at soft solder.

The price to be paid using glue is:

1. Electrical interface between die pad and substrate is not well defined.
2. If glue with metal fill is used the electrical interface between the substrate and the die pad is a shottky diode with some resistive bypasses caused by random handling scratches.
3. In parallel with the shottky diode there is a oxide capacity. Oxide thickness depends on exposure of the waver to air during manufacturing. So tox can vary from a few nm to about 100nm!
4. If non conductive glue is used the capacity between the substrate and the handler wafer depends on the thickness of the glue.

Depending on the situation the equivalent circuit of an exposed die pad and the substrate above either is an inductor to ground or some RLC circuit. Assuming an ideal ground plane the die pad and the chip can be regarded as a vertical stump of conductive material. The substrate current flow through this stump and creates a magnetic field around the stump. The substrate current of course must be closed to a loop again. This closure is provided by the sum of the currents flowing through the pins of the chip.

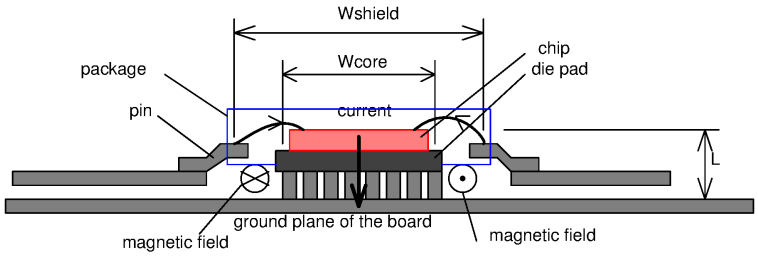


Figure 5.17: IC with exposed die pad placed over thermal vias and a ground plane and magnetic field surrounding it.

In a very simplified way this kind of structure can be regarded as a short piece of coax cable with length L, shield diameter Wshield and core diameter Wcore. This approximation allows an estimation of the inductance of the stack consisting of the thermal vias, the die pad and the chip itself.

Example: Wcore=4mm, Wshield=6mm, L=0.6mm

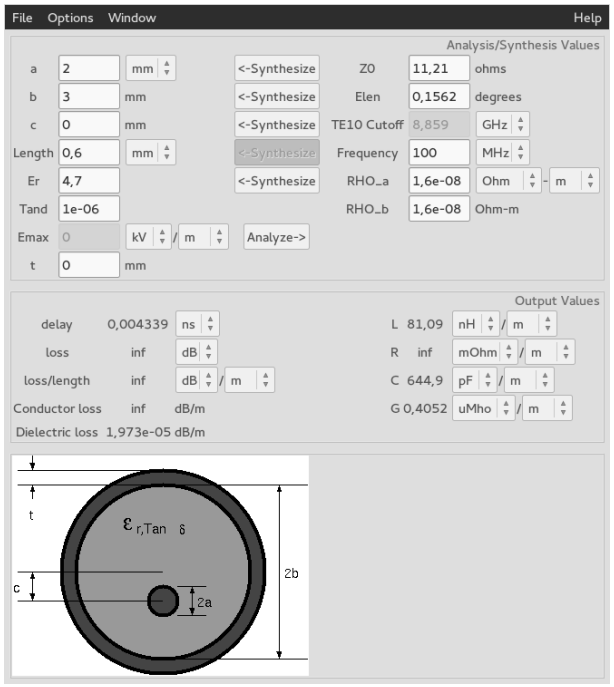


Figure 5.18: Estimation of the vertical inductance is 81.09nH/mm

Since the stack of thermal vias, die pad and chip only has a length L of 0.6mm the resulting inductance becomes 48pH. Other tools such as the field solver fasthenry find slightly lower inductances because they take into account that part of the magnetic field gets shorted by the eddy currents flowing in the ground plane. Deviations of fasthenry and the more simple to use wcalc are in the range of factor 2 (fasthenry results are a bit lower than wcalc results).

Depending on the height of the contributors a simplified equivalent circuit can be created. The values of the resistors depend on the properties of the conductors used. With increasing frequency the skin effect affects the resistor values.

The capacitor Coxide is only present if the chip is glued. In case the chip is mounted using soft solder the nets epad_top and chip_bottom are to be shorted.

Resistor Rbypass depends on oxide damages of the (unintentional but unavoidable) oxide of a grinded backside if the chip is glued.

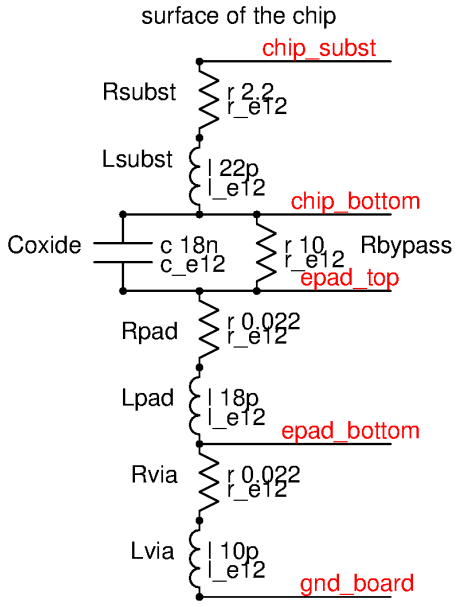


Figure 5.19: Equivalent circuit of a chip with exposed die pad.

The values of the capacitor Coxide are debatable. As soon as a grinded wafer is exposed to the atmosphere the grinded back side starts to oxidize. The thickness of this unintentional oxide depends on the time the wafer is exposed to air before it gets packaged. Thus numbers given for the oxide range from 3nm (mass production, chips get tested and molded immediately) to 100nm (engineering samples. wafers have been stored up to months before packaging). This leads to a wide range for Coxide ranging from about $350pFmm^{-2}$ to values as high as $11nFmm^{-2}$.

5.4.7 Exposed die pad and substrate resistivity change at high frequency due to skin effect

The die pad and the substrate of the chip can both be regarded as thick conductor carrying the substrate current to the board ground. The skin depth depends on the frequency and on the resistivity of the conductor.

$$\delta = \sqrt{\frac{r}{f * \pi * \mu}}$$

With increasing frequency the current flows more and more at the surface. The effect first gets visible in the die pad because the die pad usually has the lowest resistivity. (copper: $16.78 * 10^{-9}\Omega m$). The substrate of most chips has between $1 * 10^{-5}\Omega m$ and $0.1\Omega m$. The following table lists the skin depth for different materials.

Table 20: Skin depth depending on frequency

f in MHz	copper	$1m\Omega cm$	$10\Omega cm$
0.01	$666\mu m$	$1.59cm$	$1.59m$
0.1	$208\mu m$	$5.03mm$	$50.3cm$
1	$66\mu m$	$1.59mm$	$15.9cm$
10	$20.8\mu m$	$503\mu m$	$5.3cm$
100	$6.6\mu m$	$159\mu m$	$1.59cm$
1000	$2.08\mu m$	$50.3\mu m$	$5.03mm$

This means, high resistive substrate material barely is affected by skin effect. Low resistive substrate material starts to show a significant skin effect at about 10MHz. Since the skin depth follows the square root of the resistivity

low resistive material nevertheless always has a lower real part of the impedance than high resistive material. The following figure shows how the current with increasing frequency first starts to flow at the surface of the die pad and then also starts to flow at the surface of a $1m\Omega cm$ substrate.

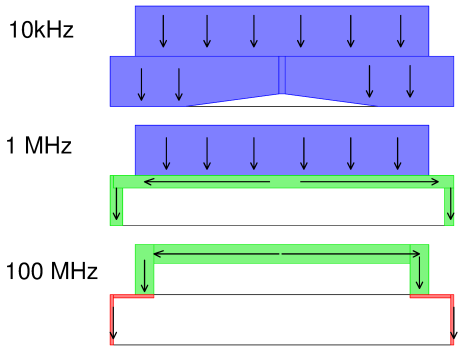


Figure 5.20: Current flow in the die pad (copper) and the substrate ($1m\Omega cm$) with increasing frequency.

For low resistive substrate the resistor network describing the substrate coupling changes with the frequency. This effect has to be taken into account above 10MHz for $1m\Omega cm$ material.

Modeling high resistive substrate ($10\Omega cm$) this effect starts to change the network above 100MHz.

5.4.8 Charge transport in mold material

Mold used for packaging electronic components starts to dissolve at an electric field of about $200V/\mu m$. This is not an immediate break down. It rather is a slow drift of ions and electrons in the mold. The transport of ions in the package continues until the charged material settles where the field strength drops below the critical value. The impact on the chip embeded in the mold depends on the protection of the surface against parasitic MOS transistors. Since the voltages involved are in the range of several 10V to hundreds of volts a charged package can activate parasitic surface transistors even through an oxide of $2\mu m$.

6 Simulation

Designing integrated circuits you will have to hop between dozens of design environments and tools. The most difficult thing usually is to get the most basic simulation running the first time. This chapter intends to just give you a kick start into your first simulation by just some lines to read. For details of your simulation tools consider using the much more complete tool manual. These manuals covering much more details usually hold hundreds of pages for each tool.

There are some tools I experienced to be a bit tricky and unstable. I added some remarks where I ran into severe issues with some of the tools.

6.1 Device simulation

Device simulation is the lowest simulation level. It directly uses physical models. There are various simulators used for specific applications.

6.1.1 wcalc

wcalc [41] is a simple calculator calculating the properties of transmission lines and coils. It uses equations found in literature. Since it is an open source tool a lot of information can be found in the source code.

wcalc offers the following calculations:

1. air core inductor
2. parallel rectangular bars
3. coaxial transmission line
4. coplanar wave guide
5. coupled micro strip line
6. coupled strip line
7. IC micro strip line

8. micro strip line
9. series / parallel RC
10. series / parallel RL
11. strip line

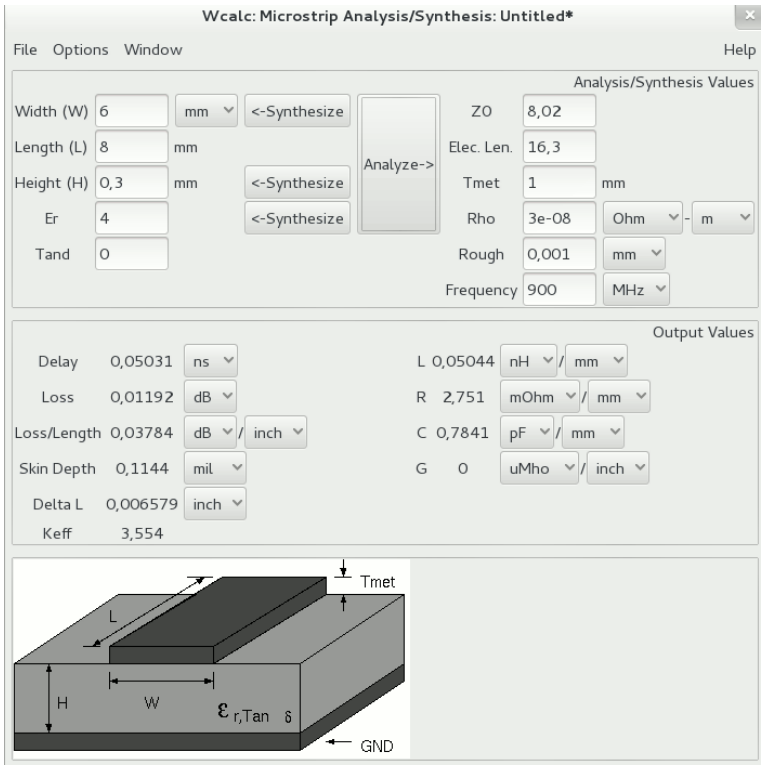


Figure 6.1: Screen shot of wcalc

wcalc is not able to calculate arbitrary shapes. The strength of wcalc is the calculation of long strip line structures with a simple form to be filled. Running wcalc with very short structures (strip lines that are wider than long etc.) leads to significant deviations from field solvers.

6.1.2 Fasthenry

Fasthenry [42] is used to determine the inductance depending on the shape of a conductor. It can handle arbitrary shapes. It is based on numeric field solving techniques.

To describe the shape a control file is used. The language mainly defines nodes and geometries connecting these nodes. Here comes a simple example used to calculate the impedance of the path to board ground of a chip in a package with exposed dice pad. The control file is named fasthenry_in.

```
** fast henry control file to calculate the inductance of dice pad and dice
.units mm .default z=0 sigma=5.8e4
* default conductivity of copper in 1/(mm*Ohms) because unit is mm
* dice pad is assumed to be 4mm*4mm, 0.3mm high
* center position of the dice pad
n0 x=0 y=0 z=0
* top of the dice pad
np x=0 y=0 z=0.3
* dice pad described as a very thick segment described as 5*5 filaments
ep n0 np w=4 h=4 nhinc=5 nwinc=5
* top of the dice (280um thick)
nsub x=0 y=0 z=0.58
echip np nsub w=2.25 h=2.4 sigma=0.133 nhinc=5 nwinc=5
* port to be calculated
.external n0 nsub
* frequency range 100kHz to 10GHz
.freq fmin=1e5 fmax=1e10 ndec=1
.end
```

This piece of code can be run using the command

```
fasthenry fasthenry_in
```

As a default fasthenry write the result to Zc.mat .

```
Row 1: n0 to nsub
Impedance matrix for frequency = 100000 1 x 1
      0.389865  +1.49684e-05j
Impedance matrix for frequency = 1e+06 1 x 1
      0.389865  +0.000144378j
Impedance matrix for frequency = 1e+07 1 x 1
      0.389865   +0.00144296j
Impedance matrix for frequency = 1e+08 1 x 1
      0.389865    +0.0144295j
Impedance matrix for frequency = 1e+09 1 x 1
       0.38988     +0.144295j
Impedance matrix for frequency = 1e+10 1 x 1
      0.391297      +1.44272j
```

The imaginary parts describe the inductance which can be calculated by:

$$L = \frac{X_L}{j * 2\Pi * f} \tag{6.1}$$

Example: at 1 GHz we get L=0.144295/(2*3.1415*1e9) = 22.9pH.

Display mode: Fasthenry in combination with zbuf can visualize the input netlist s a geometry. This requires several steps.

```
fasthenry -f OPTION inputfile
```

In this command OPTION is a string:
'simple' leads to a simple drawing without shading
'refined' provides a refined graphics.
The visibility of the ground plane is determined by the option -g on or -g off. Here comes an example:

```
fasthenry -f simple -g on test1
```

The result is a file with the default name zbuffile. To convert this file into a postscript the program zbuf is required.

```
zbuf zbuffile
```

This command produces a file with the default name zbuffile.ps. The file can be displayed using for instance gv.

```
gv zbuffile.ps &
```

There are further options available. To find out more options use the command

```
fasthenry -h
```

or check the fasthenry manual. (It is part of the archive holding the fasthenry code).

Some remarks about the code:

1. Fasthenry only accepts one single ground plane. Assigning several planes to different layers of a board is not possible. In stead all further layers must be represented by a mesh of paths.
2. Paths only connect at nodes. Letting one path end at the edge of an other path doesn´t make a connection!
3. placing orthogonal segments is better for convergence.

For further details please see [27].

6.2 Transistor level analog simulation

Transistor level analog simulation usually employs simulators such as SPICE, SPECTRE, ELDO and in some cases SABER. To view the simulation results multiple wave form viewers are available (gaw, gnuplot, wv, xelga). Additionally some more or less automatic verification tools are available (ADE-XL, Avenue, Maestro). Usability and productivity of those automatic tools is debatable.

In the following the tools are discussed in alphabetic order.

Modeling Hints: The following holds some general hints to be considered running analog transistor level simulation. Most of these remarks applies to different simulators and design environment in the same way.

Monte Carlo Simulation: Model files usually offer corner models as well as spread parameters for Monte Carlo simulation. Process corners are defined assuming large structures that don't have much statistical spread. Especially for small transistors deviations due to mismatch can exceed the process corners. Real devices found on a chip may deviate from nominal because the production run was a corner run and additionally deviate from each other due to random effects. Therefore is is suggested to run corner simulations and add Monte Carlo on top of the corners.

Some design kits do this automatically setting selection "MC with mismatch and technology spread". Some design kits require explicitly running corner and MC on top. (The implementation of the models depends on the philosophy of the model provider. Check this!)

6.2.1 ADE L

ADE L is a graphical user interface provided by Cadence. It can be run in combination with SPECTRE as well as with other simulators such as SPICE or TITAN. ADE L is optimized for running simulations interactively during the design of a circuit.

The simulation can be run on the local machine, remote or distributed (distributed automatically searches for available CPUs in a server farm). To set up the machine use the window 'setup', 'Simulator/Directory/Host ...'. Ideally the design environment is set up to automatically propose the correct command line (Example for distributed: 'qsub -l -lic_mmsim=1 -N cellname'). The results are stored below a directory having the job name. Different jobs can be accessed by the result browser using the 'File' menu. The status of jobs in a distributed simulation can be checked using menu 'tools', 'job monitor'.

In addition to simple simulation ADE L can run measurements. The following figure shows an example.

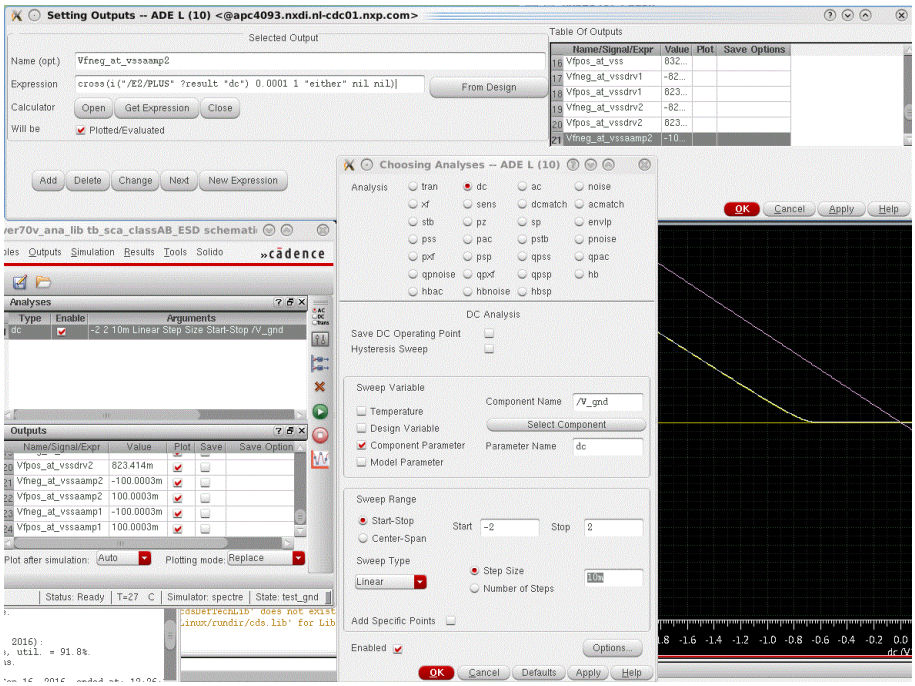


Figure 6.2: Example of setting up a measurement inside ADE L

Typically first a simulation is run without measurements. With the first simulation result the wave form calculator is used to define a measurement (here the "cross" mathematical function was used). This expression is taken from the calculator and inserted into the Expression field of the output setup pop up window. (Syntax explanation: cross(i("/E2/PLUS" ?result "dc") 0.0001 1 "either" nil nil) reads the current flowing into the PLUS node of voltage controlled voltage source E2. When the current crosses 0.0001A in either direction the measurement gets a trigger

and stores the sweeped "dc" parameter value when the crossing takes place. The result is saved into the scalar variable name "Vfneg_at_vssamp2").

After having added the expression the simulation can be rerun and the measurement is taken automatically.

ADE l doesn't check if the measured results are within specified ranges. This check requires the more complex ADE XL environment.

Once the measurements are running it is recommended to save the state to be able to repeat the measurements defined every time the circuit gets modified.

The following table lists some of the available functions:

Table 21: Some frequently used ADEL commands

function	what it does	output	typical command example
v("/node" ?result "dc")	plot of the DC voltage		
cross	determines when a certain value is crossed	scalar	cross(i("/E2/PLUS" ?result "dc") 0.0001 1 "either" nil nil)
deriv	calculates the derivative of a signal	vector	deriv(v("/node1" ?result "dc"))
differential volt.	measures differential voltage		getData("/node1" ?result "tran")-getData("/node2" ?result "tran")
dft	discrete fourrier transform	vector	
sample	pick one value at -50C	vector	sample(v("/temp" ?result "dc") -50 -50 "linear" 1)
sample	pick one value at 2.5V	vector	sample(v("/vos" ?result "dc") 2.5 2.5 "linear" 1)

Scalar results of such operations can be used by ADE-XL to produce distribution plots.

Important: if a DC sweep is done and a distribution at a certain temperature is needed the "sample" function must be used!

Note: Take care that parameter names aren't coincident with any reserved words of the netlist. Which are the reserved words depends on the environment. If verologA and VHDLA models are used the list of reserved words differs from plain vanilla spectre!

6.2.2 ADE XL

ADE XL is a graphical user interface provided by Cadence. To set up test benches and to netlist ADE XL calls the ADE L program. (But you can't start the simulation run itself from ADE L then). The main differences to ADE L are:

- enhanced corner simulation capabilities
- Monte Carlo simulation
- Parallel processing (each simulation run can be executed on a different CPU)
- Automated result documentation (similar to Avenue)
- Specification check

ADE XL uses ADE L in distributed simulation mode. It simply starts multiple distributed simulation run and stores all the run names internal to get access to the result files. (The job names are coincident with the directory names of the results of the different runs.)

Important tricks: ADE-XL only saves signals if requested to do so! Default nothing gets saved!

- check plotting options for saving signals, waveform expressions and scalar expressions. Add check marks accordingly.
- Monte Carlo simulation: check the green tooth wheel for the simulator options and add a check mark at "Save Data to Allow Family Plots"
- Distributions will only be created using scalar expressions. You typically need something like 'sample(v("/net-name" ?result "dc") 2.5 2.5 "linear" 1)' in the output setup (this example samples a dc voltage sweep from 2.5V to 2.5V at a step rate of 1V leading to exactly one sample point that can be used to calculate a distribution).
- Parameter names may not be coincident with reserved words of any of the netlist languages used (spectre, verilog, verilogA, VHDL, VHDLA, system C)

6.2.3 Avenue

Avenue is an attempt to create an automatic back end for Titan and Spectre simulation (probably it can be adjusted to work with other simulators as well). The strength of Avenue is Monte Carlo simulation with plots of distribution width versus a parameter (for instance temperature). It is possible to plot average and standard deviation versus temperature or versus other design parameters.

Working with Spectre the basic work flow is the following:

1. Create a test bench using the Virtuoso schematic editor
2. Start ADE and run a first simulation
3. Define measurements either typing in the test directly or using the Cadence waveform calculator. The measurement can be added using ADE menu "outputs", Add. Copy paste from the waveform calculator into ADE is cumbersome but possible.
4. Rerun the simulation with the measurement.
5. Save the state after successfully finishing the simulation
6. Generate Avenue plan (Menu "Session"). The menu item "Generate Avenue plan" sometimes is missing"Generate Avenue plan" sometimes is missing (e.g. if ADE-XL has been run on the same test bench or in the same Virtuoso session). Usually it helps to remove the ADE-XL setup and to restart Virtuoso (Provided the resource code is set up correctly in the project).
7. After having pressed the button Avenue will start in a new window using a Java GUI.

In the GUI there are 5 main tasks to be done.

Setup: In this tap parameters to be varied are selected.
"PARAMs" holds parameters taken over from ADE.
"Temperature" sets different temperatures
"corner0" and "corner1" holds simulation corners.
When the setup is done (The selected parameters to vary are in the text field) click "Apply".

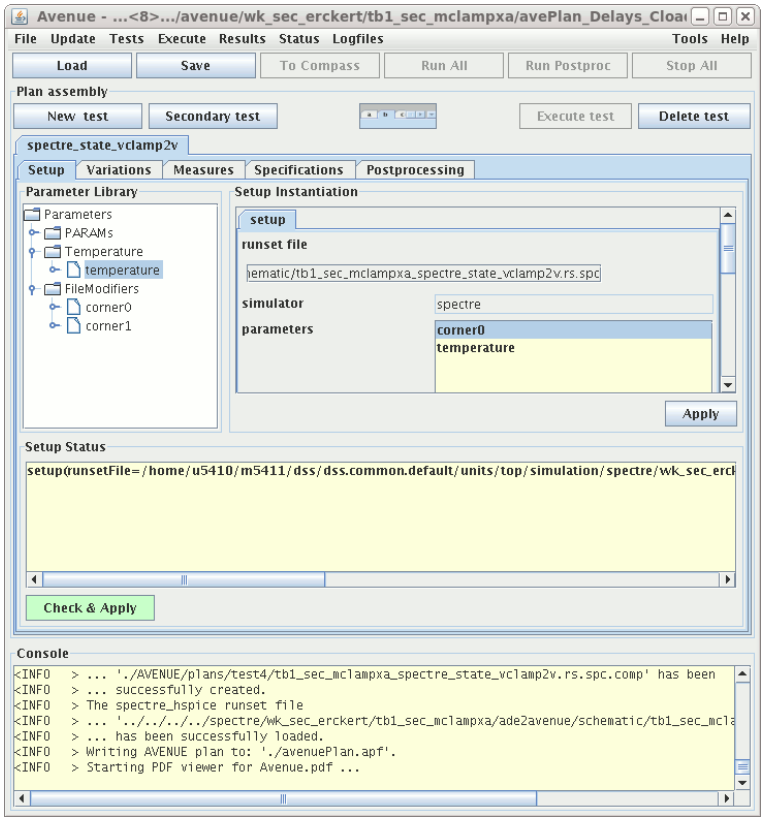


Figure 6.3: Example setup for a 2dimcorner

If you want to combine all choices of one parameter with all choices of a second parameter rather use "listSweep". Using this setup you don't need to explicitly write down all combinations.

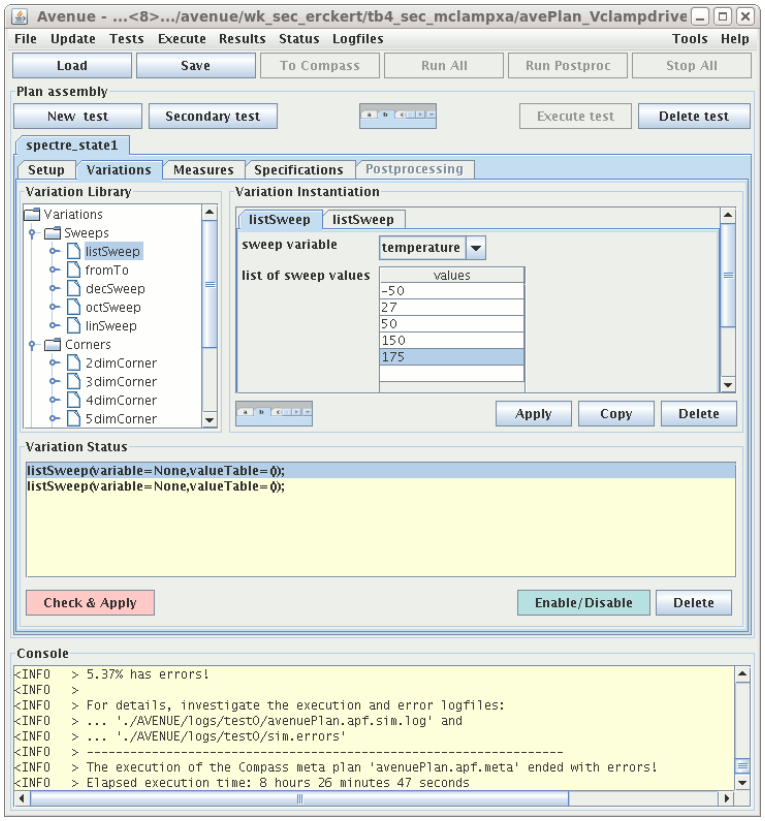


Figure 6.4: Using listSweep

Variations: First choose the number of parameters to be modified. For 2 parameters choose 2dimCorner, for 3 parameters choose 3dimCorner...

In the table visible right choose which parameter is to be placed in which column. Then the setup can be edited. This is a lot of typing work because each combination has to be typed in individually and copy paste of the tables only works in a very restricted way.

Table 22: Simple example to describing corners to be simulated using anvenue

temperature	corner0
-50	nom
50	slow
150	fast

Means only 3 simulations (-50, nom; 50, slow; 150, fast) will be performed in stead of 3*3 simulations as one would expect! To really simulate all 9 cases the table must look like this:

Table 23: A more complex table simulating more corners in avenue

temperature	corner0
-50	nom
-50	slow
-50	fast
50	nom
50	slow
50	fast
150	nom
150	slow
150	fast

So we end up in hundreds of lines of typing!

Be very careful pressing the delete button. There are two of them (one only gets visible scrolling the inner window!). One of them deletes a single line (the inner one). The other deletes the whole table. THERE IS NO UNDO FUNCTION!

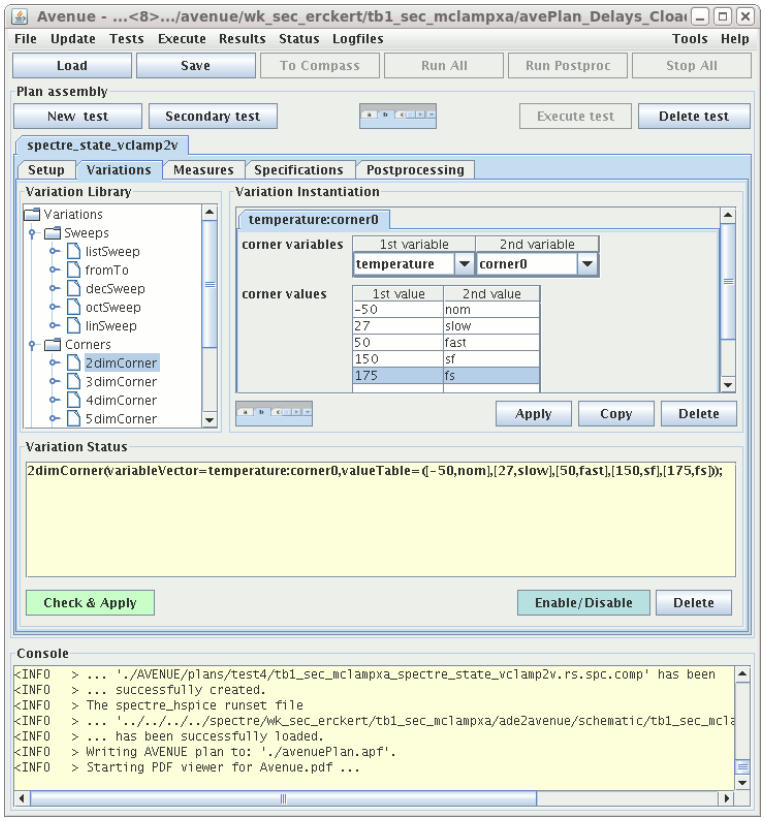


Figure 6.5: setting the variations with corners

MC simulation using Avenue: This is a tricky and undocumented feature. First get a simulation running using ADE-L. Once it is running (for instance using nom models) switch the models to mc and save state. Create avenue plan from the state with the mc models. (If the state is created from other models the simulator will complain about missing statistic information and simulation fails in all runs!)

Found now way to fix the model from within the Avenue GUI.

Measures: Here the measurements must be defined. n the beginning the table is empty and if nothing is added avenue will do nothing at all.

Specifications: The specifications column defines pass/fail criterias. If left empty no pass/fail information will be provided.

Postprocessing: Only tests listed here will be added to the output report. If this setup is left empty avenue will launch the simulations but have no output.

"SummaryTable" only provides the min and max values found. Bugs of the pdf function might not harm if the table is short enough. better use ASCII.

"measureReport" lists the detailed corners. The eps and pdf output is buggy. Better use the ASCII option.

"RunAll" This button starts all simulations. When the simulations are done open a shell using "tools", "Open Shell in Plan Directory". In the shell the results can be found at ./AVENUE/results/test..

Note: the data will not be written before the last job has finished or has been killed (using bjobs and bkill).

"Results" Behind this button the Aviator result viewer is hidden. In Aviator the pdf export only is possible after the save session button was pushed. Rest is more or less understandable.

"Tools - Open shell in Plan Directory" Warning, the shell in the avenue plan directory caches commands. If an "ls" is done and then files are changed or added the next "ls" still shows the status of before! Close the shell and restart it every time a new result is expected.

6.2.4 ELDO

ELDO is very similar to SPICE. It offers some speed up options such as "latency" to allow fast simulation of big netlists. Simulating fast comparators with slow input signals option "latency" is stongly deprecieted! (May lead to totally wrong results!)

6.2.5 gaw

gaw is the gnome analog waveform viewer. It works with gnome 3. To view the wave form SPICE must be informed to save the wave form using .write filename. (ngspice: use interactive shell write filename.)

gaw can read SPICE binary dumps (so the .spiceinit file must outcomment the ASCII setting placing an asterix in the first column '* set filetype=ascii').

The only ASCII format gaw is able to read must have a X-coordinate in the first column and the Y-coordinates in the following columns. There the first line is a description of the columns. Here is an example of an ASCII file gaw can read.

```
f          vdb(out)
1.000000e+00      5.841637e+01
1.258925e+00      5.841637e+01
1.584893e+00      5.841637e+01
1.995262e+00      5.841637e+01
```

Well, if the format is such a reduced one gnuplot will do the job as well...But gnuplot is more powerful once you get used to the gnuplot command line.

6.2.6 GNUCAP

GNUCAP uses the same netlist style as SPICE. In addition GNUCAP offers very simple switch models for standard logic gates. The strength of GNUCAP is the simulation of circuits consisting of a mix of logic and analog transistor level.

Unfortunately the development of GNUCAP stopped some years ago. More recent model levels for analog transistors have not been integrated into GNUCAP. So GNUCAP can not be recommended anymore as soon as weak inversion operating range is required.

6.2.7 Mica and Descover

Mica is the the SPICE version used at Freescale. The name originally referred to Motorola Integrated Circuit Analog simulator. The graphical front end for Mica is Descover.

6.2.8 Maestro

Maestro is just an other graphical user interface (GUI) for the cadence simulators. It is very similar to ADE-XL. The data is stored in smaller chunks (distributed over more files). Sometimes maestro causes inode problems in the file system because it creates so many little files! In this case the simulation crashes with a file system I/O error (this gets displayed in the CIW).

A second problem of maestro is the poor stability of the GUI. If you click something unexpected it crashes and takes the whole cadence design system (virtuoso) with it. The last message you will see in the Xterm is "segmentation fault". If a drop down window is hanging try to escape with a save and then exit maestro before the crash happens instead of trying to continue the work!

Copy & Paste of tests is fairly poor because you only can copy single fields (I never worked with a tool requiring more clicks for so little editing. My average: 4 clicks plus 1 scroll to change one number of a test + 4 more clicks to set the limits! Creating a test bench for an 8 bit DAC easily takes you about 2000 clicks which keeps you busy for a whole day before you can start your simulation.). Copying whole lines like in a classical spread sheet doesn't work. Copying of test limits doesn't work at all. This makes the design of complex test benches very unproductive using the Maestro GUI.

The setup is stored in some kind of an XML file. Usually these setup files are in the cadence data base under the directory representing the cell view maestro. The structure there is too complex and too unregular for just reworking the files with an editor. If you have a lot of time (several days to weeks) consider writing perl scripts to filter and sort the XML-code if you want to bypass the poor copy & paste of Maestro by hacking the XML-code in multiple setup files directly. The following screen shots show an example.

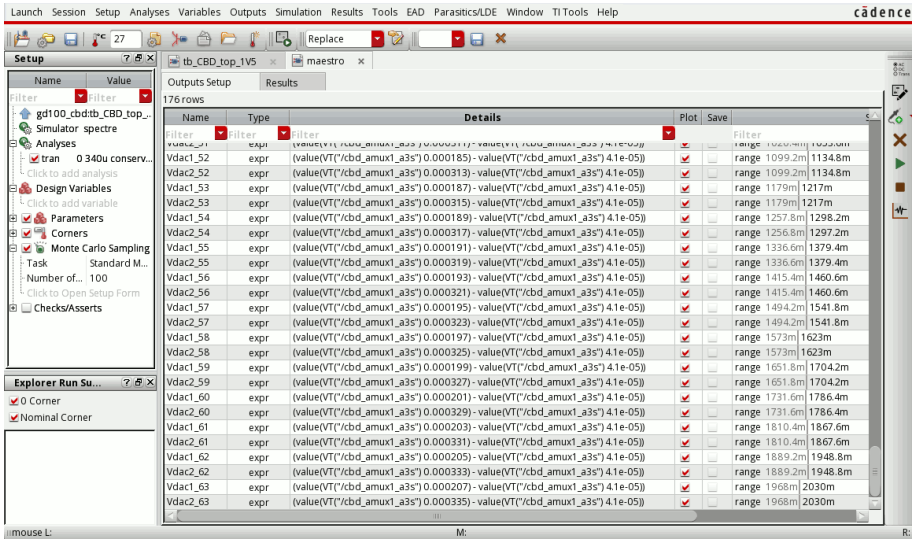


Figure 6.6: The test setup in the Maestro GUI

In the following we want to focus on the last line. This is a test of a DAC output voltage. The column "Details" gets converted into an ocean script (file oceanScript_gd100_cdb:tb_CDB_top_1V5:1.state). The file name is composed of the segment "oceanScript_" indicating what kind of file it is, the segment "gd100_cdb" indicating the cadence library and behind the ":" acting as a separator follows the cell name "tb_CBD_top_1V5". Behind the next separator follows the description if the state "1.state".

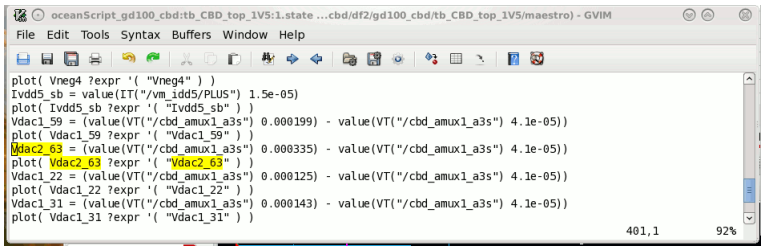


Figure 6.7: The sampling command of Vdac2_63 in the ocean script file

The test limits are hidden in the file maestro.sdb. Here we have a look at the corresponding lines:

Figure 6.8: maestro.sdb holding the limits for the DAC test shown above

A script bypassing the poor GUI must manipulate both, the ocean script and the maestro.sdb file.

Important tricks and issues: There are some unexpected behaviors that can help you speed up a simulation or will slow it down unexpectedly:

- If the number of CPUs is changed this change affects all simulations running simultaneously!

Example 1: You start a first test bench with 4 CPUs. Then you start a second test bench with 3 CPUs. You expect a total of 7 CPUs to be running. Wrong! The reduction in the second test bench also reduces the number of CPUs of the first test bench! There are only 6 CPUs working.

Example 2: The first test bench has finished. Now you have more CPUs available. => Open the "setup", "job setup" menu and increase the number of CPUs. This will add CPUs to a running simulation on the fly! The increase of CPUs however is limited to a number less or equall to the setup you had in the beginning.

Never trust the maestro graphical user interface (GUI): The maestro GUI can't be trusted. I caught maestro showing different parameters in the GUI than what is actually in the netlist!

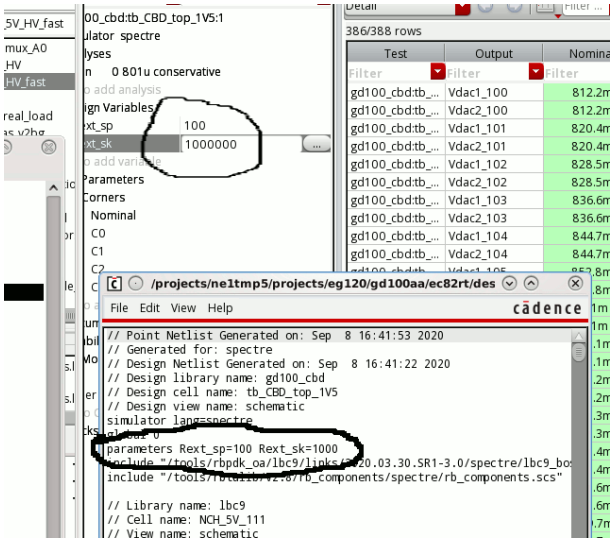


Figure 6.9: Maestro netlists different than what it shows in the GUI

The netlist shown was created with the menu sequence "Simulation-Netlist-Recreate" because I already observed somewhat strange behavior letting the hierarchical netlister decide on it's own what to netlist and what to keep. But even forcing a full netlist obviously fails. A user's nightmare!

Further tests showed that after changing the parameter you MUST click on an other entry field to make maestro accept the change. If you miss doing that last click maestro creates a netlist that differs from what you expect.

Maestro looses probes: If the list of tests and probes becomes too long I observed a loss of probes. This becomes visible at the end of the simulation when all of a sudden completely unexplainable results pop up. The pass fail test takes place although maestro lost the data. Trying to remove the probe and set it again fails with a message in the CIW like:

```
WARNING (EXPLORER-2807): The signal '/to_adc_as' already exists in the list of outputs.
```

But if you try to find the signal (using vi or similar) in the ocean scripts and the maestro.db file is doesn't exist! I have no idea where the software bug really is. The only solution I found is to rename the signal and set the probe again. (Of course you have to correct all the tests reading this signal as well!)

Maestro sometimes forgets the job setup: Sometimes maestro forgets the number of CPUs to be used. In this case it falls back to a default of 5. I observed this effect in about 1 out of 5 runs. It seems to be arbitrary. Check setup before starting a Monte Carlo simulation!

6.2.9 Powermill, Timemill

Powermill and Timemill use table models. Before running the simulation these tools simulate each component found individually. With the simulation results Powermill and Timemill set up tables. The simulation of the circuit fetches the calculation results from the tables in stead of recalculating every step to reduce the computation effort of the simulation.

Powermill and Timemill are intended to simulate big CMOS circuits using transistor level. Different to a digital simulator these tools have o concept of current and charge. So load capacities and resulting delays and energy consumption can be determined.

Setup before running the circuit simulation requires some experience (The user must have an idea which voltage and current range of the transistors will be needed for the table models.). As far as I have seen Powermill and Timemill mainly supported MOS transistor. Bipolar components were not supported well (This may have changed since I used it in the 1990s).

6.2.10 SABER

SABER has its own model language called MAST. MAST is not only able to describe electrical systems. The strength of SABER and MAST is the simulation of mixed systems than can even consist of combinations of electrical, mechanical, thermal, fluidic systems.

Since the MAST code must be interpreted by the simulator the speed of simulation of a pure electrical system is significantly slower than SPICE or SPECTRE.

For SABER a spice import function exists. This import function however is limited due to the following details of the SABER simulator concept:

- SABER does not use global nets except the reference node 0. Signals such as subst! or vdd! will not be connected correctly through the design hierarchy. The signals must be added to the netlist as subcircuit pins.
- SABER does not understand parameters of subcircuits. Subcircuit invocations such as

  ```
  x763 vlogic net0292 logic_model_tripple_well_noise
  + logicon=rlogicon logicoff=rlogicoff
  ```

  will be misinterpreted. SABER will complain about a wrong number of pins and about syntax errors. The parameters must be defined inside the subcircuit.

These limitations of the import function of SABER must be adjusted upfront in the netlist to be imported.

6.2.11 SPECTRE

SPECTRE is a simulator offered by Cadence. It has its own netlist syntax. Different from SPICE the models are not part of the simulator code. The model description is fairly free. Since every IC manufacturer can define his own SPECTRE model syntax an automatic conversion of a SPECTRE netlist into an other netlist type is more or less impossible.

Besides its own netlist style SPECTRE can be run in a SPICE compatibility mode to read SPICE netlists. The sections with SPICE code must be notified to spectre using the line "simulator lang=spice" to start the SPICE compatible section and "simulator lang=spectre" to end the section. Here comes a little example of inserting a SPICE model of a diode.

```
simulator lang=spice
.model dnr1 D (IS=1e-12 tt=20n rs=4 cjo=5e-13 vj=0.6)
simulator lang=spectre
```

Usually such a model is included using the ADE model setup. The setup is somewhat tricky because ADE seems to cache models. So the checkmarks must be switched each time the model is updated! Furthermore SPECTRES is more picky on syntax than SPICE. SPECTRE MUST have the brackets around the parameters of the model (Spice doesn't need them).

One more strange effect of SPECTRE is that SPECTRE seems to immediately fall back to it's internal verilogA style models as soon as the call of the model has a parameter the SPICE model doesn't support. A call similar to:

```
D0 (net1 net2) dnr1 area=1e-6
```

makes SPECTRE fall back to internal models because dnr1 doesn't support 'area=1e-6'! Including SPICE models requires thorough testing and consideration if the result really is plausible or if something unexpected was taken by the simulator following some fall back strategies!

A very limited conversion of SPECTRE netlists into SPICE netlists is possible as long as only the basic components of the analogLib are used. But even here certain parts of the SPECTRE netlist need manual interaction. (Example: switches need a model in SPICE netlist style while SPECTRE works without a switch .model line.)

Further differences between SPECTRE and SPICE: SPECTRE is more tolerant finding floating nets in a netlist. A floating net does not necessarily lead to 'no convergence found' or 'singular matrix' errors like in SPICE. If SPECTRE runs into convergence problems it reorders the netlist and tries to find a solution using the reordered netlist. (Well, sometimes the SPECTRE simulations show slight discontinueties where the netlist was reordered!)

- SPECTRE noise analysis does not require an input source to refer the input noise to it. SPECTRE allows noise analysis without input noise sources.
- SPECTRE offers a DC hysteresis simulation. This kind of DC simulation does the same sweep twice in two different directions. Warning: This kind of hysteresis simulation doesn't show oscillation due to capacitve feed through of analog switches used in comparators with hysteresis! (See figure 7.7.6.7. This oscillation doesn't get detected!)

Important options of SPECTRE: Spectre has a lot of sparsely documented options that can change convergence or speed up (or slow down) simulation. Here comes a short description of some of these options.

The lines starting with "simulatorOptions options" may hold some of the following commands:

cmin=10f: Giving the simulator a minimum capacity assigned to every node usually speeds up transient simulation. Of course cmin must be chosen low enough to not significantly modify the results.

dochecklimit=yes/no: enables or disables limitchecks. disabling limit checks speeds up simulation.

strobeperiod=1n: To reduce the amount of waveform data the strobe function can be used. (In very large simulations it may happen that the data to be displayed becomes to big and the whole virtuoso environment crashes when the waveform viewer gets invoked with the final message of death "segmentation fault"). In the GUI this option is hidden inside the Analysis window, tran, options, Output. If the value of strobeperiod is chosen lower than the sample points with automatic time step spectre is forced to run a simulation for each strobe point and simulation may get slower than without using this option.

If spectre is launched from the ADE L,or ADE XL GUI be aware that most of the cadence scripts are written in a dirty way placing many spectre options multiple times! Furthermore there are multiple menus in ADE L to modify parameters. Some are accessible double clicking on the analysis, some are hidden under the simulator options button. Don't expect options to be set correctly just because it is correct in the GUI because there may be an other menus you didn't adjust jet that will overwrite. Better verify the netlist manually to be sure it really holds the correct options you expect to be there.

Spectre offers different output format for the waveform file. If you want to have a certain freedom choosing different waveform viewers choose psf in stead of the new standard format psfxl. (This is hidden in the ADE "outputs", "save all" menu!) psfxl mainly differs from psf by an index accelerating the display of single waves.

Visualization of results using Spectre: The wave form display of Spectre offers the possibility to call a powerful wave form calculator. Select a signal (left mouse click on the signal), click tools, select calculator. The calculator has a function panel offering a big variety of functions. (You can just as well just type in the function - provided you know the syntax.) Here are some of the most versatile functions and their usage:

Table 24: some frequently used mathematical functions that can be used in spectre

function	what it does	output	typical command
v("/node" ?result "dc")	plot of the DC voltage		
deriv	calculates the rerivative of a signal	vector	deriv(v("/node1" ?result "dc"))
dft	discrete fourrier transform		
gainMargin	calculates gain margin at 180 degres	scalar (dB)	gainMargin(VF("/signal"))
phaseMargin	calculates phase margin at 0dB	scalar (deg.)	phaseMargin(VF("/signal"))
sample	pick on or more values	vector	sample(v("/temp" ?result "dc") -50 -50 "linear" 1)

6.2.12 SPICE

SPICE was one of the first simulators used for analog circuit simulation[18]. Usually the models and the algorithms are compiled into the code of SPICE. Introduction of new model equations usually requires a recompilation of SPICE.

Most of the original SPICE code is coming from Berkeley University. It is open source. The sourcecode is freely available. Throughout the years many derivatives offering additional features or additional models have been developed out of this original code. Today's dialects of SPICE support a wide variety of models. Some have a graphical backend mostly based on NUTMEG.

Some of the behavior of spice can be defined in the file .spiceinit. Usually this initialization lile is found in the home directory. The following lidle code snippet shows a very little example of .spiceinit:

MOS models implemented in most SPICE dialects: The model type usually is defined in the model description. Usually among the model parameters there is a statement like level=1. If this model level statement is missing SPICE falls back to a default level. Usually this is level=1. If the level statement is not consisting with the model parameters SPICE usually ignores inconsistent parameters. Normally it shows a warning. In some cases the warning is already issued at read in of the netlist. This is shown in the following example that combines level=8 with some

Algorithm 1 Example for .spiceinit setting colors for NUTMEG

```
* this is a default setting for ngspice
set color0=white
set color1=black
```

incompatible parameters. (Level=8 calculates the threshold vto from physical parameters and ignores vto given in the model.

```
ngspice 102 ->  source nmos_dc.cir
Circuit: * nmos dc transfer characteristic
Warning -- Version not specified on line
"tox=14.1e-9 nsub=6e+16 cgdo=3.3e-10 cgso=3.3e-10 cgbo=1e-10 js=1.107e-4 cj=1.284e-4
cjsw=4.43e-11 uo=513 kf=1e-28 vto=0.8 level=8"
Setting version to 'default'.
Model issue on line 8 : .model n33 nmos tox=14.1e-9 nsub=6e+16 cgdo=3.3e-10 cgso ...
unrecognized parameter (uo) - ignored
unrecognized parameter (vto) - ignored
ngspice 103 ->
```

Running this example using ngspice simply ignored the uo and vto parameters and additionally found some more warnings. Nevertheless the simulation was executed following the model level level=8.

```
ngspice 103 -> dc vg 0 3.3 1m
Doing analysis at TEMP = 27.000000 and TNOM = 27.000000
Warning:
Pd = 0 is less than W. Warning:
Ps = 0 is less than W.
No. of Data Rows : 3301
ngspice 104 ->
```

Typical model levels found in most SPICE implementations are listed below:

Table 25: Model levels used in spice and ngspice

level	description	features	remarks
level=1	MOS1	strong inversion only	simple, fast simulation
level=2	MOS2		
level=3	MOS3		
level=4	BSIM1		
level=5	BSIM2		
level=6	MOS6		
level=7	UFET		not supported by ngspice
level=8		strong, weak inversion	ignores vto, uo
level=44	EKV		
level=47	BSIM3v2		
level=53	BSIM3		not supported by ngspice
level=54	BSIM4	strong, weak inversion	ignores tox, vto, uo
level=55	B3SOIv1		
level=56	B3SOI		
level=57	SOI3		
level=58	UFS		
level=60			
level=61		strong, weak inversion	
level=62			

Simulation types of SPICE: Spice supports a big variety if simulations. The simulation can be described in the netlist. In some cases there is a interactive frontend (for instance ngspice). If the simulation is launched from the netlist the netlis usually holds a command such as '.dc' or '.ac' for a DC-analysis or an AC-analysis. In case of using an interactive SPICE front end the dot (.) often is omitted.

.DC The most used type of analysis. Even if the DC-analysis is not requested SPICE often performs a DC-analysis to determine the operating points. SPICE requires every node of the netlist to have a direct current path to node 0 (the global ground). Floating signals or signals only connected by ideal capacitors will lead to error messages such as 'singular matrix' or 'no convergence'. If one of these errors occurs it is required to find those nodes that have no DC-path to node 0. The gmin algorithme of SPICE is not able to solve floating nets because in the last step gmin is set to 0 again.

```
.DC V1 0 3.3 1m
```

This command sweeps source V1 from 0V to 3.3V using a step width of 1mV. In DC analysis multi parameter sweeps are possible.

```
.dc v1 1 10 0.1 r1 1k 10k 100
```

sweeps voltage source v1 from 1V to 10V in 100mV steps and sweeps r1 from 1K to 10K in 100 Ohm steps.

.AC Alternating current analysis. SPICE first performs a DC-anaysis to determine the complex small signal impedances. Then it linearizes the circuit and performs an AC-analysis. Normally there only may be one independent AC source in the circuit. This AC source is sweeped.

Parametric sweeps like in the .DC analysis are not supported in .AC analysis. There is a way to work around this using a control loop. The following code sequence shows a simple example.

```
* test of parameter usage
v1 1 0 dc 1 ac 1
r1 1 2 1k
r2 2 0 1k
c1 1 2 1p
**********************
.control
let start_r=1k
let stop_r=10k
let deltar=1k
let r_act = start_r
while r_act le stop_r
  alter r1=r_act
  alter r2=r_act
  ac dec 100 1e4 1e9
  plot v(2)
  let r_act=r_act+deltar
end
.endc
************************
.end
```

The spice while loop increases r_act until it becomes higher than stop_r. Each time the loop is executed a simulation is done and the result is plotted. Since nutmeg starts a new plot each time the loop is executed the result looks like this:

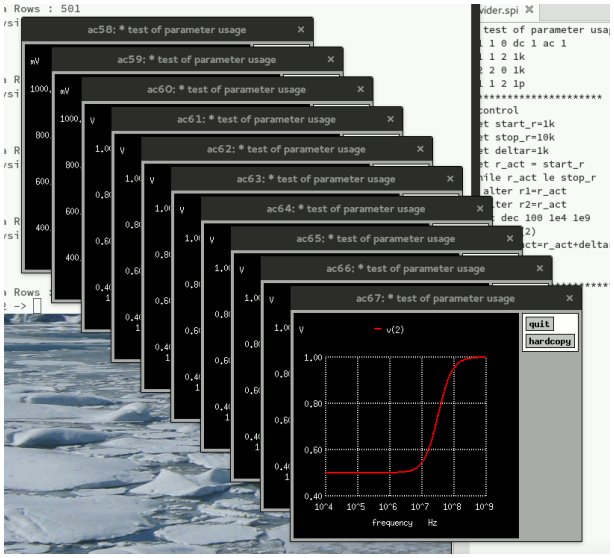


Figure 6.10: Result of a looping simulation with plot

There is one way to bypass the multiple plots limitation of nutmeg. ngspice can be made to write the results to disk after each simulation by the following code:

```
* test of parameter usage
v1 1 0 dc 1 ac 1
r1 1 2 1k
r2 2 0 1k
c1 1 2 1p
**********************
.control
let first=1
let inc=1
let file=first
let start_r=1k
let stop_r=10k
let deltar=1k
let r_act = start_r
while r_act le stop_r
  alter r1=r_act
  alter r2=r_act
  ac dec 100 1e4 1e9
  write $&file
*  plot v(2)
  let r_act=r_act+deltar
  let file=file+inc
end
.endc
************************
.end
```

The command "write $&file" converts the real number stored in file into an integer. This integer number is used as a file name to write the data. So on the disk there will be the files 1, 2, 3, 4, 5, 6, 7... These files are in spice raw format. They can be opened by gaw. gaw is able to plot results of different raw-files in one graph.

.NOISE Noise analysis. SPICE first runs a DC-analysis to determine the impedances and gains. Then it runs the noise analysis. In the noise analysis SPICE calculates the noise of every component and the propagation of the noise to the node requested. The noise sources are assumed to be independent of each other. SPICE adds the squares of the voltages without taking care of the phase. The result is stored in V^2/Hz . Often we want to know the level in dBV. To get a plot in dBV use the command shown below. Additionally SPICE calculates the input referred noise of the input source indicated in the command line.

```
.NOISE v(output) vinput dec 100 1e5 1e9
```

This example means: Calculate the output noise at node output. Calculate the input referred noise assuming vinput is the input signal of the amplifier. Use a logarithmic scale with 100 points per decade. Start frequency is 100kHz. Stop frequency is 1GHz.

```
plot 0.5*vdb(onoise_spectrum)
```

(This command refers to ngspice and the NUTMEG backend. Possibly you have to tell nutmeg which simulation result to use giving the interactive command:

```
setplot noise1
```

(This means it picks the first noise simulation found in the RAM. In case you have no clue what is available just use

```
setplot
```

and the nutmeg backend will list which results are available.

```
set color0=white
```

changes the background color to white.

```
set color1=black
```

changes the color of the grid to black.

```
set color3=blue
```

changes the color of the first trace to blue.

At least some of these setups can be stored in your home directory in the file .spiceinit . In my case .spiceinit looks like this:

```
* this is a default setting for ngspice
set color0=white
set color1=black
```

To use gaw save the wave form using the interactive command

```
write filename [netname]
```

if no netname is given all nets will be written. Example:

```
write vdb.raw vdb(net3)
```

writes the voltage of net3 in dBV.

```
write vdbn3n5.raw vdb(net3, net5)
```

works as well. This saves the differential voltage of net3 and net5 in dBV.

.TRAN The .tran command performs a transient analysis. The following command preforms a transient analysis with 1ns step width and a total duration of 1us.

```
.tran 1n 1u
```

ngspice offers a fast Fourier transformation being part of the nutmeg backend. Before running the FFT the sampling points should be placed in linear spacings. This is done by the command linearize:

```
linearize nodename
```

The linearize command sets a new current plot vector holding the linearized nodes with a equidistant time step. Now the FFT can be run:

```
fft v(nodename)
```

This again creates a new plot vector holding the FFT results. The results of the FFT can be plotted with the command:

```
plot mag(V(nodename))
```

Returning to the old transient simulation results is possible using the setplot command.

Rawfile formats: As a default the rawfile consists of an ASCII header and binary data. This behavior can be changed by creating a local file named .spiceinit inside the directory spice is running. Using this file the rawfile format can be changed to ASCII. Here is an example of the .spiceinit setting ASCII code.

```
* Standard ngspice init file modified to produce ASCII rawfile
alias exit quit
alias acct rusage all
set x11lineararcs
*set rndseed=12
** ascii rawfile **
set filetype=ascii
** frontend debug output **
*set ngdebug
** no asking after quit **
set noaskquit
** set the number of threads in openmp
** default (if compiled with --enable-openmp) is: 2
*set num_threads=4

strcmp __flag $program "ngspice"
if $__flag = 0

* For SPICE2 POLYs, edit the below line to point to the location
* of your codemodel.
 codemodel /usr/lib64/ngspice/spice2poly.cm
* The other codemodels
        codemodel /usr/lib64/ngspice/analog.cm
        codemodel /usr/lib64/ngspice/digital.cm
        codemodel /usr/lib64/ngspice/xtradev.cm
        codemodel /usr/lib64/ngspice/xtraevt.cm

end unset __flag
```

This modification of the rawfile format should be done with care. Some tools such as gaw can't handle the ASCII rawfiles. If the rawfile format is set to ASCII the result looks something like this:

```
Title: *spice circuit <opamp_open_loop_test> from xcircuit v3.7 rev 55
Date: Thu Mar  9 13:47:24  2017
Plotname: AC Analysis
Flags: complex
No. Variables: 2
No. Points: 901
Variables:
        0       frequency       frequency grid=3
        1       vdb(fb, gnda)   notype
Values:
 0      1.000000000000000e+00,0.000000000000000e+00
        5.841636710844818e+01,0.0

 1      1.023292992280754e+00,0.000000000000000e+00
          5.841636710780362e+01,0.0
.......
```

This ASCII rawfile format might be easier to use for self written tools but it requires significantly more disk space.

Spice automatic measurements: Spice can perform automatic measurements while running a simulation. The following code shows an example of such a measurement command.

```
* test of the spice measurement capability
rload 1 0 1k
vtest 1 0 pulse 0 1 0 1n 1n 1u 3u
* measurement of a period
.measure tran period trig v(1) val=0.5 rise=1 targ v(1) val=0.5 rise=2
```

```
*evaluate y-value of v(1) at t=0.5us
.measure tran yeval find v(1) at=0.5u
.end
```

The result returned by spice looks like this:

```
  Measurements for Transient Analysis
period               =  3.000000e-06 targ=  3.000500e-06 trig=  5.000000e-10
yeval                =  1.000000e+00
```

There are more complex measurements possible such as finding the value of a voltage when a certain event takes place. For details check [18].

Forbidden node names: ngspice doesn't accept node name gnd. Naming a node gnd usually leads to simulation errors or convergence problems.

6.2.13 TITAN

TITAN is a simulator derived from SPICE. It is more feature rich than most SPICE implementations. Usually it is invoked using a control file. I haven't seen any GUI (graphical user interface) yet.

Since version 6 TITAN requires a software license. To run TITAN the whole environment (for Titan and the license manager) must be set up using the script "module load dessup/titan".

TITAN usually is invoked with the command "titan -f netlistname".

Different from some SPICE implementations TITAN comes without a waveform display. In stead it offers a big variety of output formats for different waveform viewers in the save command. Here are commands producing an output file named test.out.

Table 26: Saving commands and formats of TITAN

options of the save command	format produced
.save test.out	default format p2
.save test.out format=adda	same as format=csdf
.save test.out format=csdf	ASCII working with viewer ws
.save test.out format=gnuplot	ASCII file for gnuplot including a title line
.save test.out format=gwave	ASCII file for gwave (gaw?) or gnuplot, no title line
.save test.out format=ppr	printer format
.save test.out format=psf	ASCII file similar to gnuplot
.save test.out format=psf-artist	Couldn't get it running!
.save test.out format=scope4	SABER format (old style) for SABER viewer
.save test.out format=scope5	wv and SABER new style. Data: .pl file. Info (headers): .ai_pl file

To obtain a netlist in the Infineon environment choose the CIW menu "infineon", "frontend tools", "netlisters", "spice". This will open a form to fill in which cell to be netlisted and where to write the netlist..

6.2.14 wv

wv is a waveform viewer often used in combination with SPECTRE or TITAN. In the infineon environment wv must be launched from the "units" directory to have the correct environment information.

wv can't use the compressed psfx format of SPECTRE. Set the output format to psf if you want to use wv. (This setup is hidden in the ADE "outputs", "save all" menu in one of the last lines!)

6.3 Digital simulation

Digital design usually is based on textual description of a logic. The most common languages used for this description are verilog (or open source version iverilog) and VHDL (or the open source variant ghdl). Verilog is very close to real hardware descriptions. It is limited to data types that can be stored in registers (bit, hex, integer ..). Verilog has no concept of real or complex numbers. Since the data types are predefined the code gets very compact and easy to read.

VHDL (and ghdl) also has a concept of real numbers and permits modeling continuous analog systems as well as logic. The draw back of this flexibility is that a lot of data type definitions and conversions have to be defined in the code. This makes the code much longer compared to verilog.

6.3.1 Verilog and iverilog

To start with the most easy we first have a look at verilog. (before starting check that the GNU C compiler gcc is installed because iverilog calls this compiler.) A verilog code usually consists of two sections. The initial section is executed once. All the logical functions usually are is an "always at" loop. Here comes a simple example.

```
module main;
reg a, b, c, D, E;
        initial
                begin
                $write("!           abcDE\n");
                $monitor("%10d  %b%b%b%b%b", $time,a,b,c,D,E);
                a = 0;
                b = 0;
                c = 0;
                #10 a = 1;
                #10 a = 0;
                #20 b = 1;
                #10 a = 1;
                #10 b = 0;
                #30 c = 1;
                #10 a = 0;
                #10 a = 1;
                #20 a = 0;
                #10 a = 1;
                #10 a = 0;
                #20 b = 1;
                #10 a = 1;
                #30 c = 1;
                #10 a = 0;
                #10 a = 1;
                #10 b = 0;
                #10 a = 1;
                #10 a = 0;
                #20 b = 1;
                #10 a = 1;
                #30 c = 1;
                #10 a = 0;
                #10 a = 1;
                #10 $finish;
                end
/* simple shift register with non blocking assignment.
Blocking assignment leads to shifting through the whole shift reg. in one clock */
always @ (posedge a) begin
        D <= b;
        E <= D;
        end
endmodule
```

The $write and $monitor commands are simply used to store the results into a file (similar to probes in an analog simulation). #10 means "wait 10 time steps before changing the register value. At $finish the simulation stops.

The always @ statement is getting executed in parallel with the initial each time the start condition is fulfilled (posedge a). If the open source version iverilog is used the code first has to be converted into an executable. Typically the command is

```
iverilog -o test1.out test1.v
```

test1.out is an executable. To run the simulation test1.out must be executed.

```
./test1.out
```

The simulation result will be printed into the console. Here is the screen shot.

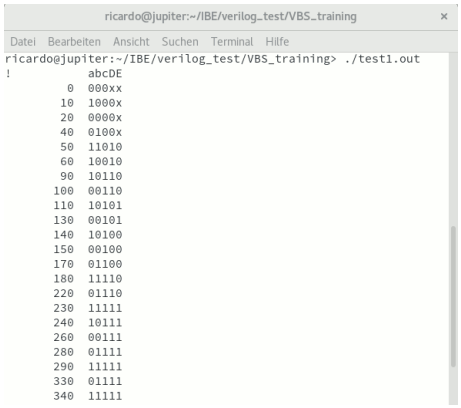


Figure 6.11: A typical console output of verilog

Instead of just throwing the result on the console it is more handy to dump the result into a file that can be read and scrolled with a wave form viewer such as gtkwave or dinotrace. In this case the initial statement must hold something like

```
$dumpfile ("test_delta_sigma.vcd");
$dumpvars (0, main);
```

These lines create a file called test_delta_sigma.vcd. The file is a vcd-dump of all variables. To read the file use

```
gtkwave test_delta_sigma.vcd
```

Alternatively the console output can be redirected into a trace file. This requires the command

```
./test1.out > test1.tra
```

The file can be opened with dinotrace reading the file type “DECSIM”. (dinotrace can also read vcd-dump files using file type “Verilog VCD”.)

6.3.2 VHDL and ghdl

VHDL and ghdl must be regarded as a modeling programming language (like Mathlab or Octave) rather than a circuit netlist. On one side it is very powerful on the other side it is more complex than writing C-code! VHDL and ghdl require more preparation (e.g. definition of data types etc.) before actually writing the code. In addition at least ghdl relies on the ada part of the GNU C compiler. (So besides gcc the package gcc-ada must be installed)

1. describe which libraries have to be used
2. define an entity
3. describe an architecture of the entity

Using ghdl there are 3 steps to be taken. These are called:

-a analysis (well, it is a kind of compilation using the gcc-ada package)

-e elaborate (well, in C-programming I would call it a linker run)

-r run the code

–xyz the double - is used for further options such as dump file or run time.

Here comes the famous hello world example using file name hello_world.vhd:

```
-- hello world
use std.textio.all;

entity hello_world is
end hello_world;

architecture behaviour of hello_world is
begin
```

```
        process
                variable l :line;
        begin
                write (l,string'("Hello world"));
                writeline (output, l);
                wait;
        end process;
end behaviour;
```

In this first example we are using the library std.textio.all. Next we have to create an empty entity called hello_world. At last we have to assign a behavior to this entity.

To make this example run the first step is the compilation or analysis of the code.

```
ghdl -a hello_world.vhd
```

The analysis produces a control file. The name of the control files is always work-obj93.cf. Inside the control file there is a hex-code part and the listings of the entity (or entities if there are more of them) and the behavior. This is what the control file looks like:

```
v 4
file . "hello_world.vhd" "986fdabd069e1d67e6172325cae350fb696f9795" "20190530153527.932":
  entity hello_world at 2( 15) + 0 on 11;
  architecture behaviour of hello_world at 7( 76) + 0 on 12;
```

Next the code must be elaborated (in C this is called a linker run).

```
ghdl -e hello_world
```

Note: in the elaboration run the name of the entity must be used!

After elaboration the code must be run using the command:

```
ghdl -r hello_world Hello world
```

Note: in the run command the name of the entity must be used!

Next is an example that really produces data:

```
-- test of the clock and writing data
  library ieee;   use ieee.std_logic_1164.all;

  entity heartbeat is     port ( clk: out std_logic);
  end heartbeat;

  architecture behaviour of heartbeat is
    constant clk_period : time := 10 ns;
  begin
    -- Clock process definition
    clk_process: process
    begin
      clk <= '0';
      wait for clk_period/2;
      clk <= '1';
      wait for clk_period/2;
    end process;
  end behaviour;
```

The analysis, elaboration and run commands are almost the same as before:

```
ghdl -a clktest.vhd
ghdl -e heartbeat
ghdl -r heartbeat --vcd=clktest.vcd --stop-time=200ns
```

The run command has two new options defining the dump file and the run time of the simulation. There are a lot of further options available running the simulation. These are documented at https://ghdl.readthedocs.io/en/latest/using/Simulation.html#export-waves

```
gtkwave clktest.vcd &
```

opens the dump file in a waveform viewer.

The next example is a test bench including a reset generator and a clock generator. The clock generator is executed periodically and all waits have a limited time of cl_period/2. The reset generator is executed once. The last wait is infinite. (Compared to verilog this code is complex and hard to read because VHDL doesn't have something like the verilog initial.) The test bench doesn't have any pins. Nevertheless it must be declared as an entity. In the architecture the components must all be declared. The connection of the components is done under the comment line "– the netlist".

```
-- a flat testbench providing clock and reset
  library ieee;
  use ieee.std_logic_1164.all;

entity tb_flat is
end tb_flat;

-- list of components used
architecture behaviour of tb_flat is
  component clock
    port(cl: out std_logic);
  end component;
  component resgen
   port(rsn: out std_logic);
  end component;

-- the netlist
signal clk, resn: std_logic;
begin
  clock50m: clock port map (cl => clk);
  resetgen: resgen port map (rsn => resn);
end behaviour;

-- clock generator ----------------------------------------
  library ieee;
  use ieee.std_logic_1164.all;
  entity clock is
    port ( cl: out std_logic);
  end clock;

  architecture behaviour of clock is
    constant cl_period : time := 10 ns;
  begin
    -- Clock process definition
    clk_process: process
    begin
      cl <= '0';
      wait for cl_period/2;
      cl <= '1';
      wait for cl_period/2;
    end process;
  end behaviour;

-- reset generator ----------------------------------------
  library ieee;
  use ieee.std_logic_1164.all;

  entity resgen is
    port ( rsn: out std_logic);
```

```
  end resgen;

  architecture behaviour of resgen is
    constant tres : time := 50 ns;
  begin
    -- reset process definition
    reset_process: process
    begin
      rsn <= '0';
      wait for tres;
      rsn <= '1';
      wait;
    end process;
  end behaviour;
```

The result of the test bench run is shown in the following figure:

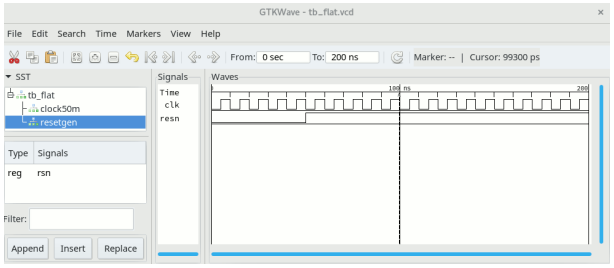


Figure 6.12: The result of the simulation of the flat test bench

The design itself usually is distributed over many different files. It isn't convenient to add the complete design to the test bench in a flat way. To create a hierarchical design test bench includes the components with a "use" statement. It looks like this:

```
-- DUT ----------------
 for counter: counter4bit use entity work.counter4bit;
```

Nevertheless the component and the pins must still be described inside the code.

```
-- DUT ----------------
  component counter4bit
    port(resn, clk: in std_logic; cnt: out std_logic_vector(3 downto 0));
  end component;
```

And the netlist including the name declarations of the wires looks like this:

```
-- the netlist
signal clk, resn: std_logic;
signal cntr: std_logic_vector(3 downto 0);
begin
  clock50m: clock port map (cl => clk);
  resetgen: resgen port map (rsn => resn);
  counter: counter4bit port map (clk => clk, resn => resn, cnt=>cntr);
end behaviour;
```

Important: Before the test bench can be analyzed, elaborated and run all the cells belonging to the design must first be analyzed. in or example the sequence of command is:

```
ghdl -a counter4bit.vhd
ghdl -a tb_with_test.vhd
ghdl -e tb_with_test
ghdl -r tb_with_test --stop-time=500ns --vcd=tb_with_test.vcd
```

If the analysis of the cells was forgotten the test bench won't run!

Functions, procedures and process: In ghdl there are two kinds of timing controls. A process consumes time. It only executes one element of a loop per triggering event.

Things that are intended to run without consumption of time must be outside of the process and inside of a function or a procedure. The following code of a 4 bit DAC shows the concept:

```
  library ieee;
  use ieee.std_logic_1164.all;
  use ieee.numeric_std.all;

entity dac4bit is
  port (resn, clk: in std_logic; d_in: in std_logic_vector(3 downto 0); refp, refm: in rea
end dac4bit;

architecture behaviour of dac4bit is
-- calculation outside the process to make it non time consuming
function convert(vector:std_logic_vector(3 downto 0)) return natural is
variable helper:natural;
begin
  for i in 0 to 3 loop
    if vector(i)='1' then
      -- std_logic_vector doesn't allow math and must be converted
      helper:=helper+2**i;
    end if;
  end loop;
  return helper;
end function convert;

begin
  process (resn, clk)
  begin
    if resn='0' then
        aout<=0.0;
      else
        if (rising_edge(clk)) then
          aout<=refm+(refp-refm)*real(convert(d_in))/15.0;
        end if;
      end if;
  end process;
end behaviour;
```

In this code the function convert converts the standard_logic_vector into a natural number without using any time. Inside the process this function is called for the calculation of the real voltage aout.

Limitations of ghdl: ghdl provides real numbers but it doesn't have a discipline "electrical" like VHDL-AMS. Probably it is possible to write a package to extend ghdl. But I haven't tried it yet.

6.4 Mixed signal simulation

Mixed signal simulation combines analog simulation (simulators SPICE, SPECTRE, ELDO) and digital simulation (simulators Verilog, VHDL). Before a mixed mode simulation can be started for every cell it must be defined, which kind of simulation is is to be used. The top level cell (often called testbench) holds a view named config.

Some languages offer the possibility to carry out a mixed mode simulation without coupling different simulators. This however requires writing models of the analog part in this specific language. Typical examples are VHDL, ghdl and verilogA.

6.4.1 Config view

The config view is is a list holding for every cell the following items:

1. Library name
2. Cell name
3. view name

The library name is required because there may be identical cell names coming from different libraries (Typical example: Multi chip designs with different technologies used for the two chips but same component names such as nmos or pmos).

The view name determines which simulator is used or whether the cell is further resolved. Typical views invoking the simulator are SPECTRE, verilog, verilogams, SPICE. These views hold code that can directly be executed by the simulator.

Views such as netlist are ambiguous. A netlist can be simulated by a digital simulator or an analog simulator. What really happens depends on the design environment.

Views such as schematic or cmos_sch usually invoke the analog (transistor level) simulator.

Views verilog or VHDL invoke a digital simulator.

After every change of a schematic it is recommended to update the config view.

6.4.2 Interface elements

Analog simulation has a concept of voltage and current. Interfacing a digital simulator and an analog simulator requires a definition which voltage is regarded as a logic 1 or a logic 0. In addition to the normal design lib a so called interface library. This library holds atd (analog to digital) and dta (digital to analog) cells. Signal coming from the analog side are converted to digital values by a atd cell. Signal coming from the digital side are converted to voltages by the dta cells. The cells are placed automatically when the config view s getting elaborated.

6.4.3 Simulator usage

The most flexible way to simulate is using the ADE GUI. The cell view to be simulated is config in stead a schematic view. The config view tells the netlister which view to use. (Different from a schematic netlist and simulation the options menu hold no more views to be used and no more stop views because these now are defined by the config views)

Netlisting usually works no matter if the views are consistent or not. The simulator however stops with an error if one of the views calls a component that is not available. (example: A digital gate holds a transistor called nmos_dig but there is no spice or spectre model for nmos_dig. In this case an other view of the cell - may be veriloga or something else used for behavioral simulations - must be used.)

6.4.4 Company specific tools

Some companies use their own setup, netlist and waveform viewer launch scripts. These little programs usually set standard paths to models, netlisting directories and run some kind of a make file to create the netlists (some parts may already exist in a precompiled form such as INCA-files. These only need some kind of a linker run called "elaborate" in the Cadence world).

Table 27: Company specific mixed signal environments

company	invocation	from	what is does
BOSCH	db.sim	shell in project directory	sets paths, netlist, run (1)
BOSCH	IFS	from maestro	sets paths, netlist, run (1)

db.sim: (1) db.sim is based on system C. So it can be used for almost anything! It is controlled by a sim file. Usually it searches at /projects/.../.../user_home/current/database_dig/.../sim_data/sim_file

db.sim calls a tcl interpreter to run the sim file. The sim file holds a lot of shell variables defining paths. Then it cleans the directories before launching a new simulation. Among others the sim file holds a section describing the lib and the top level cell to be simulated and the view to use and the state file the simulator uses to start (Parameters ams_tb_lib, ams_tb_cell, ams_tb_config, ams_tb_state).

Next thing the sim file does is to set all the elaborate and the simulator options.

The AMS simulation options should be checked. Standard simulation runs 27C, nominal, liberal. Especially "liberal" seems a bit daring.

As soon as the command "Scan for new sim-files" in the File menu found something usable the "Local Simulation" drop down menu gets available.

The sim file fetches the stimuli from an IFS command file. Usually called something like a "test.cmd". This command file contains digital stimuli written in the IFS language. IFS is a language specialized to describe test patterns executed with more or less standardized modules such as an SPA (used to measure analog values). Execution of the code is - like most digital simulators - taking place in a concurrent way. To get everything synchronized again a special SYNC command is used.

In stead of using the library analogLib it is recommended to use rb_behavioral_ifs. This library holds standard components similar to the analogLib but there the models are implemented using VHDLA. These models can be controlled by an IFS command file to create a standardized test bench for mixed signal simulations.

IFS commands: The following table list some of the most frequently used IFS commands in alphabetical order:

Table 28: List of some important IFS commands

module	command	params	what it does	
ACT	SWITCH	name on/off	opens or closes a switch	
ALL	QUIT		stop the simulation	
ALL	SYNC	ALL or module list	wait for the slowest command	
ALL	WAIT	deltime 1 ms	wait 1ms	
CLK	FREQ	n MHz	set clock to n MHz	
DIN	PWM	name f D	logic PWM with frequency f and duty cycle D	
DIN	WAIT	deltime t	stop logical PWM for time t	
JTAG				
SCT	VPWL_RAMP	source V t	ramps to new voltage within time t	
SPA	CHECK_V	net, min, max	checks the voltage of a net	
SPI	DATA1	addr, r/w, data, 0	DATA1 drives a 32 bit SPI. data is 1 byte, 15 dummy bits	
#inc	<filename>		include a command file here	
#loop				

Some of the IFS modules are coded especially for a specific test bench. In this case the module documentation typically can be found in the comment lines of the module code (Typical example: SPI. There are too many flavors of SPIs to pack them all into one piece of system C code. So every logic designer writes his own - more or less well documented - test interface).

The log files of db.sim: The db.sim simulation invokes system C code, VHDL code, verilog code, model sim code, analog simulation. Each of the tools involved dumps its log files to a different output file. Thus the logs are distributed over the files transcript, trace.log (from VHDL and verilog), trace_sc.log.

6.4.5 The checkout problem

One of the major bugs of Design Sync in combination with Cadence is the habit to check out the cells in edit mode if a simulation is running. Cadence claims the issue is solved, but it is not. If you annotated DC operating points to the schematic the design environment without asking checks out all cells in edit mode! This can create severe problems close to tape out when all the top level simulations are running.

Project management: Expect 1-2 days only for cleaning up the unintentional checkouts at the end of the project. (You will have a whole team puzzling a long time what check out was on purpose and which one was due to software stupidity.)

6.5 System simulation

System simulation is covered by a very wide range of approaches. For testing concepts often tools based on equations are used rather than transistor level or logic level simulations. Typical tools for such tasks are:

- Mathlab
- Octave
- jupyter
- Scilab

and of course we can also (miss)use electrical simulators such as SPICE and SPECTRE.

6.5.1 Mathlab and Octave

Mathlab and Octave are very similar, almost twins! Mathlab is a numeric equation solver frequently used in the Windows world. Octave is the open source counter part used in the Linux world. The developers of octave try to be as compatible to mathlab as possible. However there are some differences. Modern versions of octave come with a nice GUI.

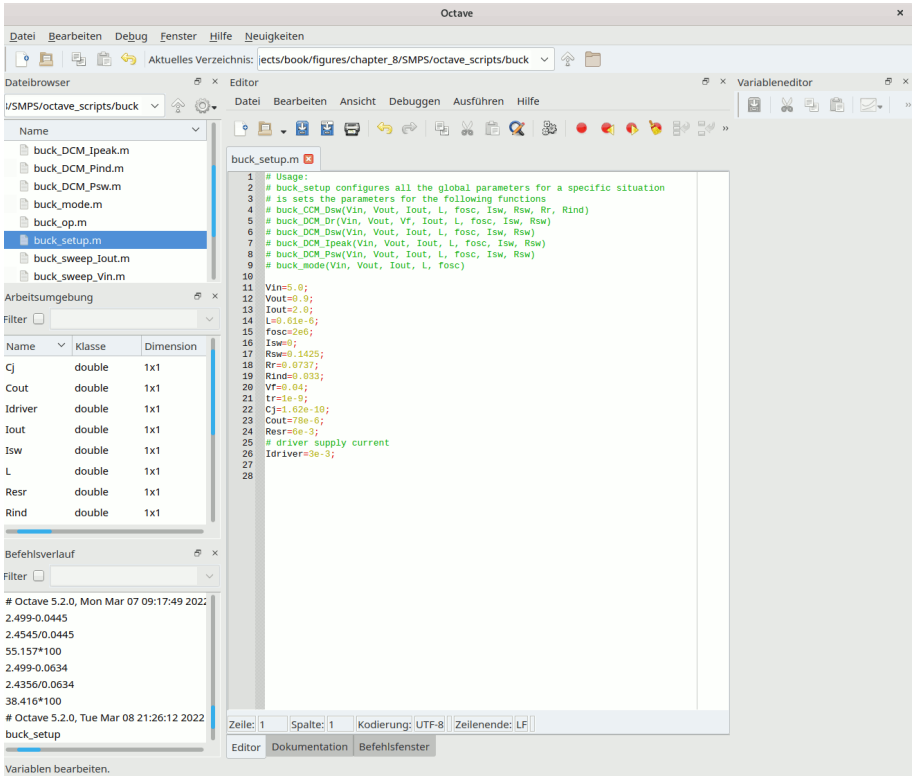


Figure 6.13: Octave GUI

6.5.2 Scilab

Scilab falls into the same class but has some more simulation features than octave using the xcos simulator. Code compatibility of scilab and mathlab is not as good as compatibility of octave and mathlab. Scilab installation is somewhat tricky. Usually the latest version (and some older ones) can be found at https://www.scilab.org.

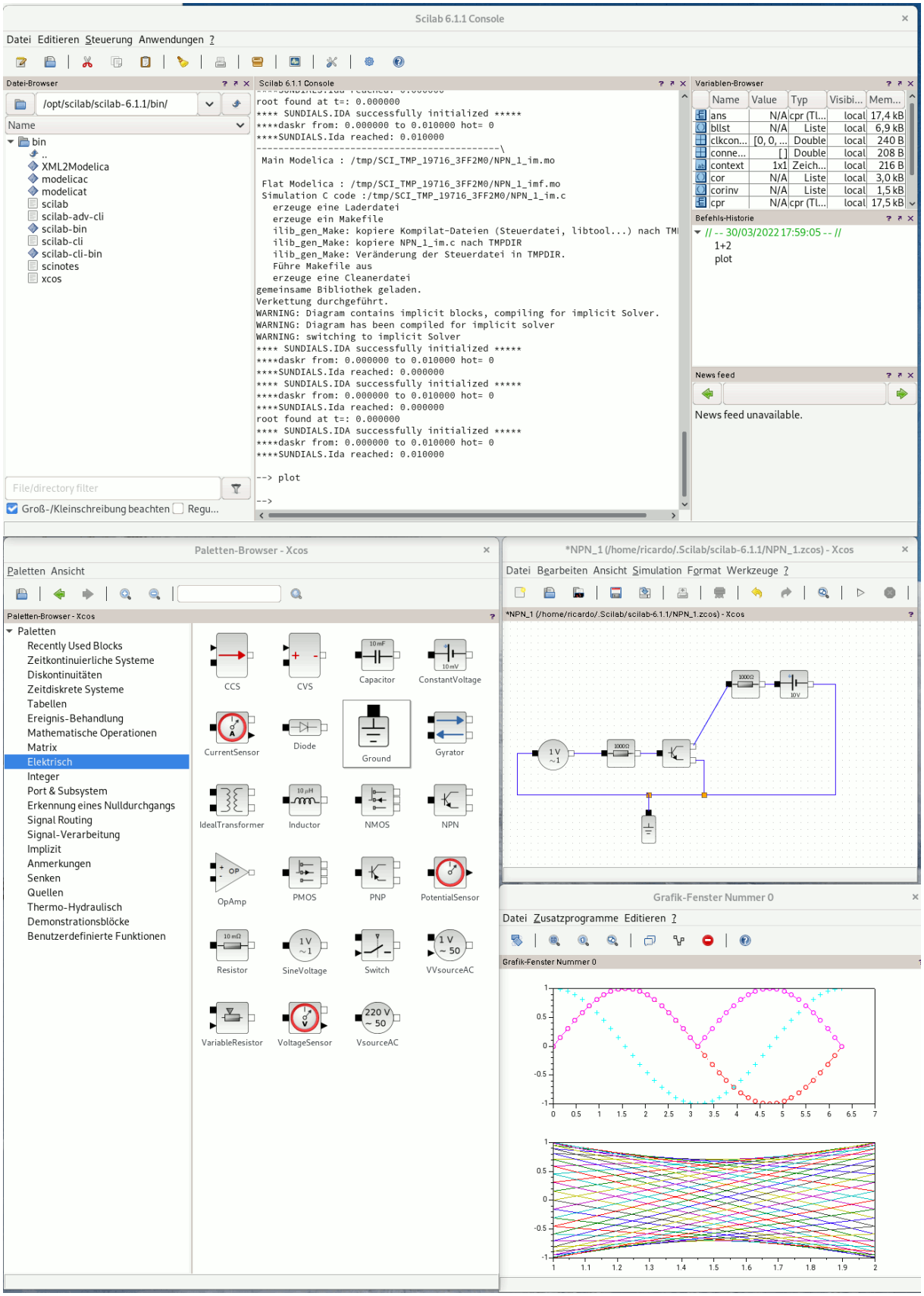


Figure 6.14: Scilab and Xcos GUI

All these programs are designed to work with vector and matrix calculations. The strength of these numerical solvers is that any equation independent of the domain (electrical, mechanical, fluidic, thermal...) can be included. Whatever can be described by mathematics can be part of the calculation and/or simulation.

6.5.3 Jupyter

Jupyter differs a little bit from the approach of mathlab, octave and scilab. In stead of emphasizing the use of vectors jupyter can be regarded as an interactive interface to python. Python itself is a powerful programming language similar to C or perl. The typical command to start jupyter is:

Algorithm 2 Invocation of jupyter using the command line

```
jupyter notebook
```

Jupyter then launches a GUI running in a browser window. To start a new session you have to create a new notebook in the 'file' menu. Inside this browser window you can enter python code snippets and run them.

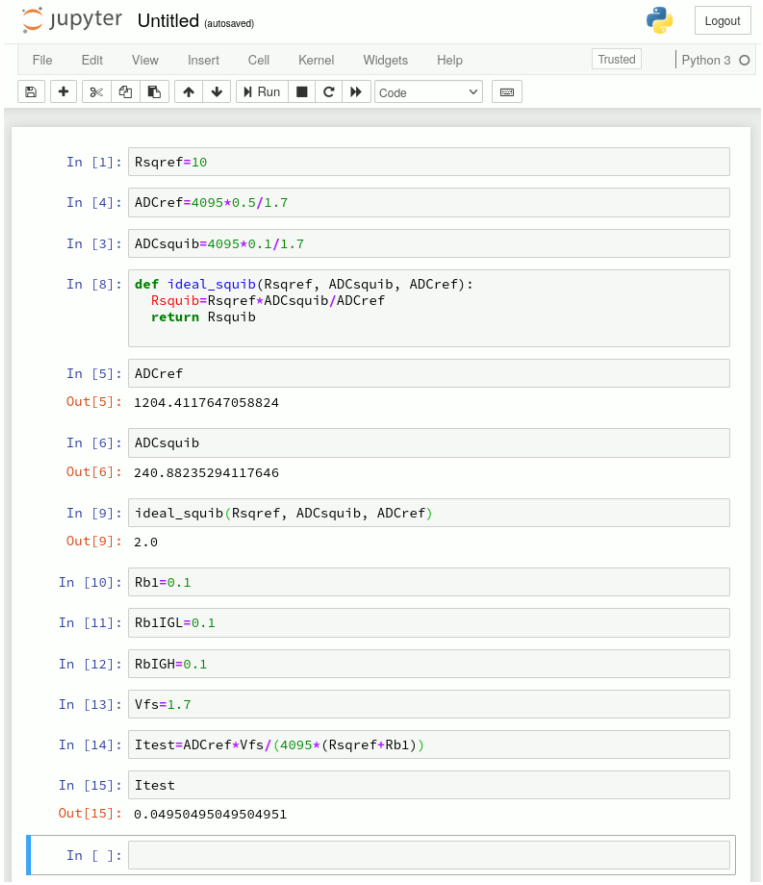


Figure 6.15: jupyter running in a firefox window

On one hand this gives access to the complete and extremely powerful syntax of python. On the other hand it is more complex than using tools like octave or mathlab. Visibility of parameters used is much better in the octave GUI.

6.5.4 Behavioral simulation using analog simulators

To speed up simulation blocks are described mathematically. Typical examples are:

- replace analog amplifiers by behavioral sources

SPICE behavioral code: In SPICE code such a replacement could look like this:

```
.subckt op10k1 out inp inn vdd vss
B10k outint 0
+ V=(v(vdd)+v(vss))/2
+ + abs(v(vdd)-v(vss))*tanh((2*10000/abs(v(vdd)-v(vss)))*(v(inp)-v(inn)))/2
ROUT outint out 100
.ends op10k1
```

The code describes an amplifier with a gain of 10000 and output clipping at vdd and vss. To achieve a nicely converging soft clipping a tanh function is used. To be sure things won't get messed up if vdd and vss are swapped the differences are taken as absolute values. In the model above the condition vdd=vss is not intercepted. This condition would lead to a divide by zero.

Since the output source is between node outint and node 0 the model has no concept of current consumption. Furthermore the bandwidth of this simple model is not limited.

SPECTRE behavioral code: The netlist language of SPECTRE is much more complex. For this reason Cadence provides a standard library called analogLib. This library holds basic elements such as voltage sources, current sources and polynomial sources. The most basic one is the pvcvc (polynomial voltage controlled voltage source). It has a gain (like a normal voltage controlled voltage source vcvc) and additionally 5 polynomial coefficients. The meaning of the coefficients is:

coeff 0: 0 order DC value of the output.
coeff 1: 1st order. The input signal is simply multiplied.
coeff 2: 2nd order. The input is squared and multiplied.
coeff 3: 3rd order. The input power of 3 multiplied with coeff 3
coeff 4: 4th order. The input power of 4 multiplied with coeff 4

Using this polynomial source non linear functions can be modeled. Here is a simple example of the resulting spectre code:

```
E1 (outp outn inp inn) pvcvs gain=1.0 coeffs=[ 1 2 3 4 5 ]
```

This line produces an output voltage following the equation:

$$V(oup, outn) = 1.0 * (1 + 2 * V(inp, inn) + 3 * V(inp, inn)^2 + 3 * V(inp, inn)^3 + 4 * V(inp, inn)^4$$

Using a two input polynomial source things get more interesting.

```
E1 (outp outn inp1 inn1 inp2 inn2) pvcvs gain=1.0 coeffs=[ 1 2 3 4 5 ]
```

This line produces an output voltage with the equation:

$$V(outp, outn) = 1.0 * (1 + 2 * V(inp1, inn1) + 3 * V(inp2, inn2) + 4 * V(inp1.inn1)^2$$

$$+5 * V(inp1, inn1) * V(inp2, inn2)$$

So the 5th coefficient can be used to perform multiplications.

6.5.5 RF emission (EMC) simulation

A matter of perspective: Before we even look at the first chip parameter it must be clear what the customer really measures! Most of the EMC discussion is coming from automotive system level considerations. The electronics may not disturb the radio inside the car. So EMC measurements either try to measure radiated RF or they refer to the ground on the receiver. In the standardized setups the ground of the spectrum analyzer is connected to the metal table (Not to the board!). The metal table is connected to the application board with a (more or less) standardized ground wire of typically 20cm length. The spacing between the board and the metal surface of the table is typically 50mm. The insulation is either air or some isolating foam such as styrodur.

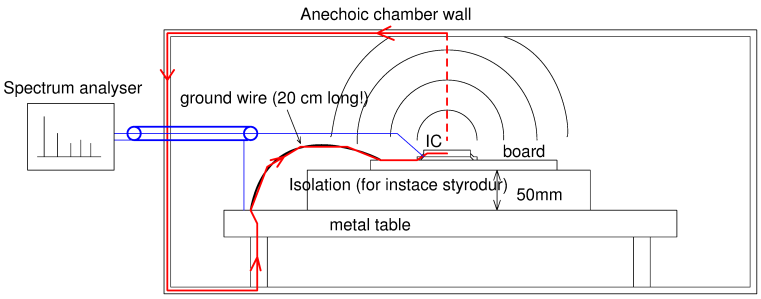


Figure 6.16: EMC measurement setup used by most customers in the automotive field.

The radiated RF current reaches the walls of the anechoic chamber, flows back to the table and through the ground wire back into the application board. The current flowing through the ground wire is the sum of the RF currents radiated by the antennas existing on the application board. As a consequence minor changes of the board layout changing the antenna characteristics has a dominant impact on the radiation and the RF current flowing back through the ground wire. The reference ground is the most important question of all! **. The setup with**

the spectrum analyzer grounded at the table measures the sum of the radiated RF differentiated by the inductance of the ground cable.

For chip evaluation the setup should be independent of board radiation. Therefore in chip characterization the spectrum analyzer is grounded at the board in stead of the table. The RF return current doesn't affect the measurement anymore.

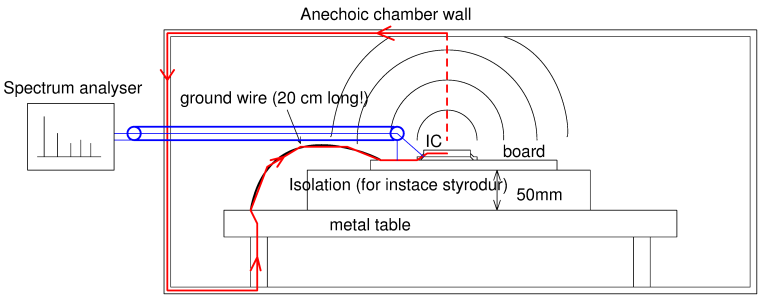


Figure 6.17: Characterization setup avoiding miss measurement due to ground wire drop

The following figure shows the equivalent schematic. Depending on the ground connection of the spectrum analyzer the RF voltage measured at the 5V DC source (for instance an on chip voltage regulator) differs dramatically!

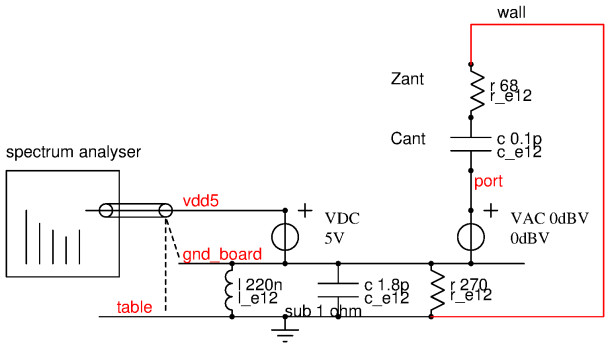


Figure 6.18: equivalent circuit of the two measurements.

The antenna is modeled with a simple serial circuit of 100fF and 68 Ohm. The radiating port is assumed to have a level of 0dBV. Thus the result of the AC simulation can be regarded as an AC transfer function.

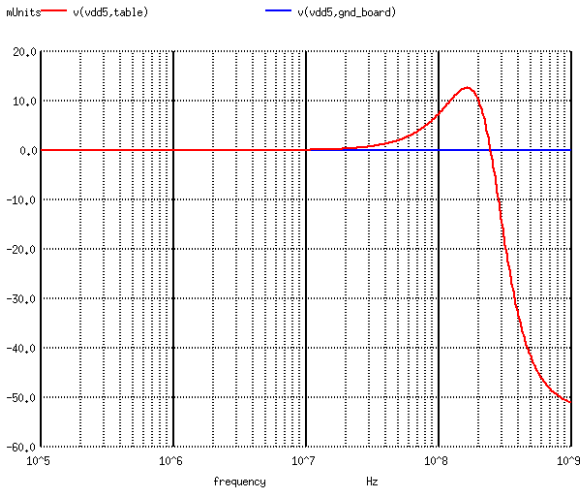


Figure 6.19: Simulation of the AC voltages between vdd5 and table and between vdd5 and gnd_board.

Measuring relative to the board ground the signal stays at 0V! The measurement relative to the table continuously increases. The zero crossing only is due to the phase (there the real part disapears). Plotting in a logarithmic scale the phase information is removed and we see the signal in dBV.

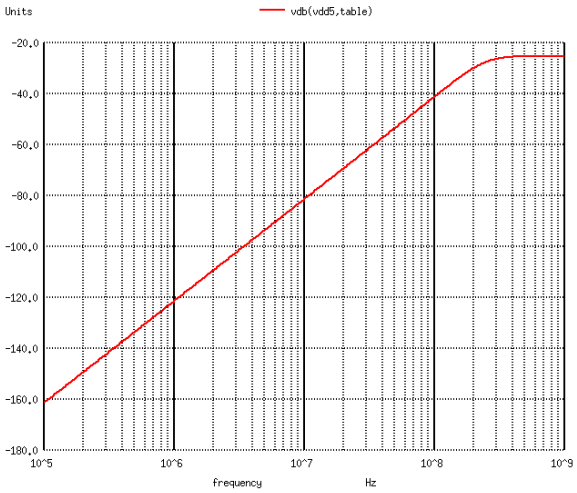


Figure 6.20: Signal between vdd5 and table plotted in dBV

In other words: Using the wrong ground can lead to totally misleading results. The distortionless 5V rail looks noisy while the real bad guy is the pin ´port´ just because we grounded the spectrum analyzer at the table in stead of the board. But this is exactly what most system level EMC test setups do!

The computation power bottle neck: "With state of the art software running on state of the art computers it will never be possible to simulate state of the art chips completely analog." This statement is already about 20 years old, but it still is true. The reason is that state of the art chip designs will be used in future computers. So the software and the computers unavoidably lags behind chip design.

On the other hand to simulate RF emission we have to look into individual junctions of the components. We even have to take into account parasitic components! The only possible solution is to create an AC model that provides an envelope of the expected emission spectrum. The AC simulation simply linearizes the problem and can run fairly fast. The main effort lies in creating reasonable AC models. For complex chips the creation of the AC models - which has to be done manually using educated guesses - can take several months.

Why do we spend this effort? Accepting a several month task of engineering must be well justified. The RF model of the chip usually is useless for the chip manufacturer. It only will be used to develop the application board of the customer. So under normal circumstances the customer that is taking the benefit of the AC model would be interested in creating the model. The crux is that the customer in most cases doesn't have enough information about the circuit and the layout with it's associated parasitic components. The key to solve this discrepancy is to either find a way the customer contributes to the development cost (common practice for application specific integrated circuits - ASICs) or to achieve an additional revenue from the RF-model.

The basic idea of emission models [40]: Every signal of the time domain can be converted into a frequency domain signal. Most signals found in digital systems as well as in switching systems (switchmode power supplies) are more or less trapezoid signals. They roll off with -20dB/decade at 3.14 times the fundamental frequency. If rise and fall times are equal the second roll off starts at 3.14*1/tr.

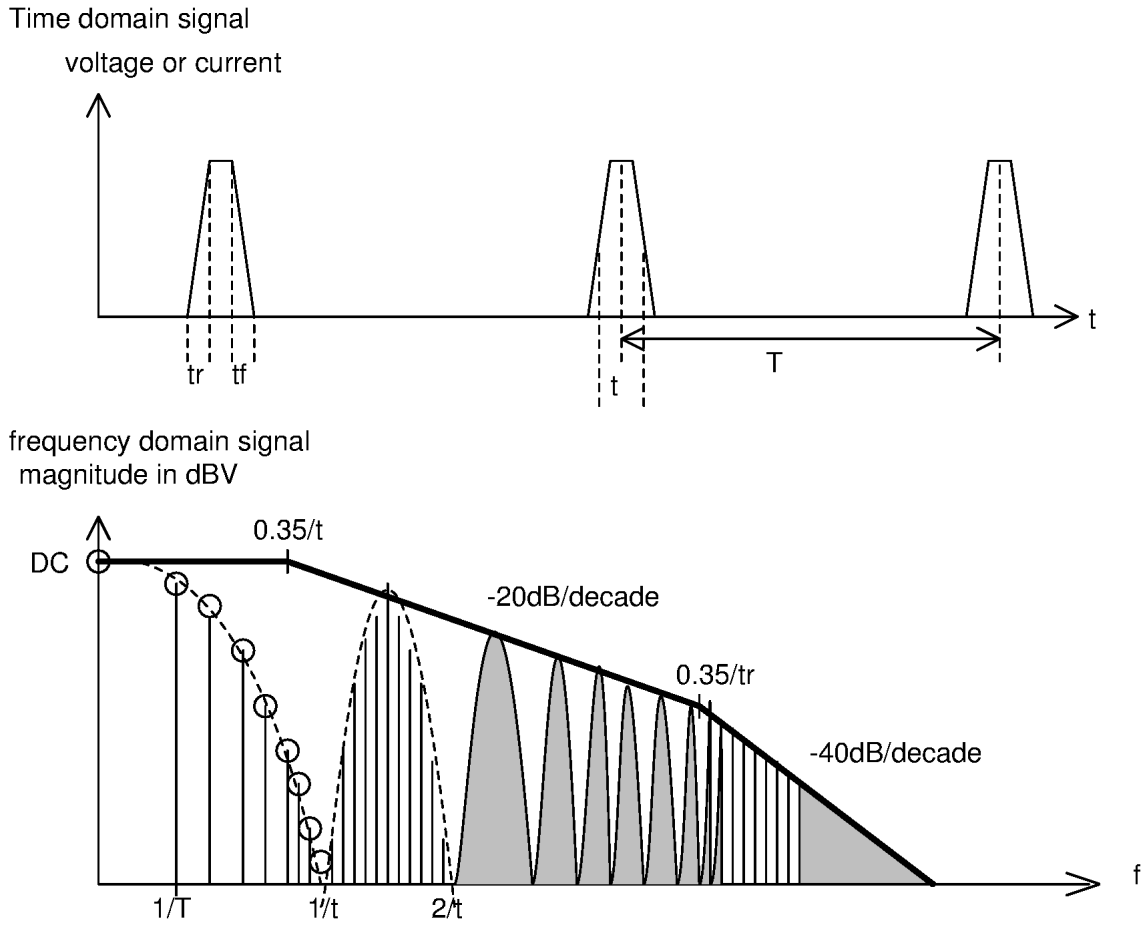


Figure 6.21: Roll off of the spectrum of a trapezoid signal

Where is GND [39]? The first conceptual thoughts must be spent on the question where the reference ground in the emission simulation will be. Since the model is used by board designers the ground node will most likely be the ground plane on the board. Chip ground is detached from board ground by the parasitic components of the package. This even applies to chip on board technology because searching for the microvolts of emission even some pH if the bumps of a flip chip mounting have an effect.

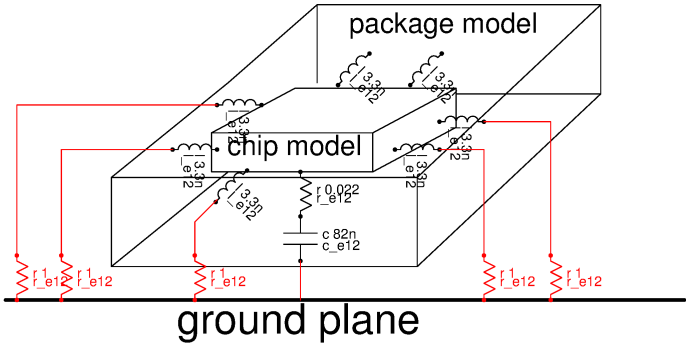


Figure 6.22: Concept of board ground, package model and chip model

As a consequence of taking the ground outside of the chip the globals 0 (node 0) and gnd! should be removed unless they refer directly to global gnound as a model reference (see later).

Package parasitcs [39]: Most pins of a package can be regarded as a series of the bond wire inductance and the pin inductance. Additionally the bond wire usually has a resistance of some 10 miliohm. Besides that there always is a capacity to the adjacent pin and to the ground plane underneath (assuming the customer uses a proper ground plane).

The quality of the inductance is limited. This can be modeled adding a parallel resistor as a damping element.

For convergence reasons floating pins should be avoided. Since most RF emission models only include the most significant sources of noise the risk of producing floating pins is high. Adding a path to ground with for instance 10G

Ohm is recommended (10G Ohm is high resistive compared to the RF impedances found. So it will not harm.)

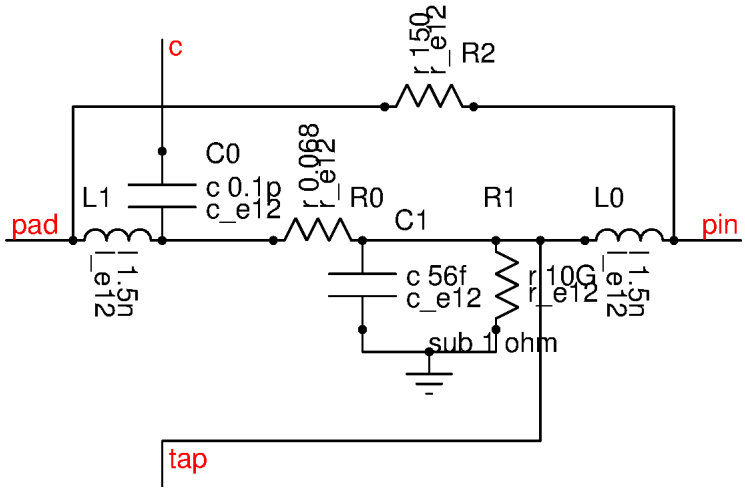


Figure 6.23: Model of a single pin of a TQFP package

In the above model the pin inductance and the bond wire inductance are represented by L0 and L1. The resulting resonant tank is damped by R2 that roughly corresponds the impedance of the pin 0.1mm over a ground plane. The resistance of the bond wire is represented by R0.

Each pin connects with node c to the node tap of the adjacent pin. The capacitive coupling to ground is represented by C1. R1 serves as a convergence aid for floating pins.

Using this pin subcircuit a TQFP32 package looks like this:

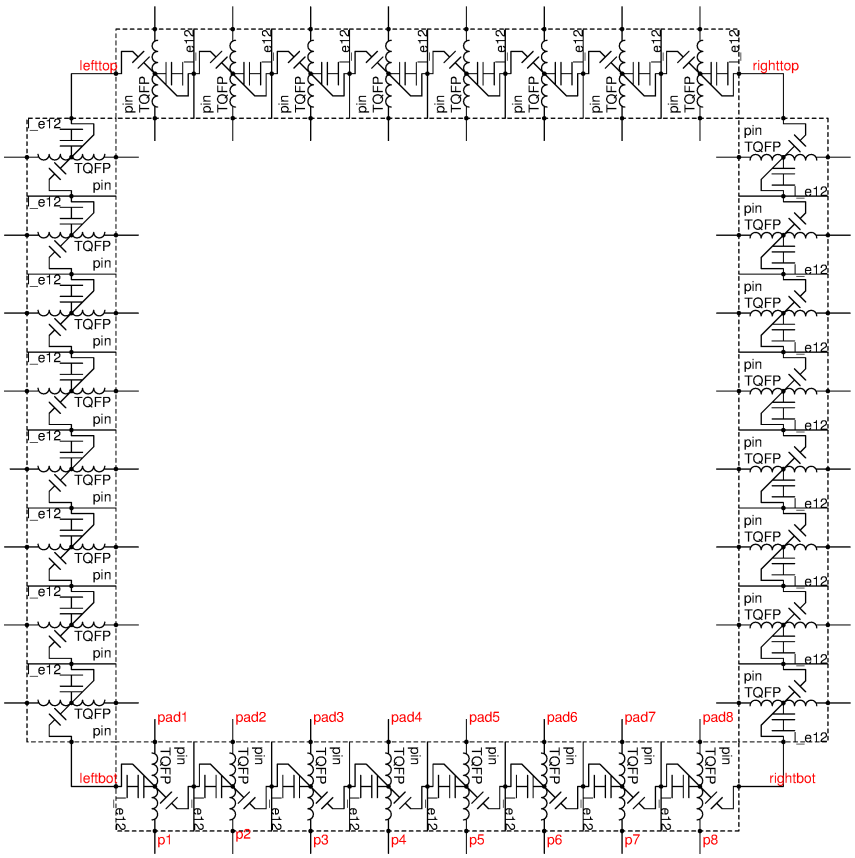


Figure 6.24: TQFP32 equivalent circuit

To get a first idea of the ground bounce inside the chip let us assume pins p3 (VDD) and p4 (VSS) supply a logic with 20mA average current consumption. The current consists of 1ns pulses with a repetition rate of 20MHz. (This is a reasonable assumption for a system clocked with 10MHz. Due to the short pulses the first cut off frequency of the spectrum is 500MHz. The on chip blocking capacity is assumed to be 300pF. Thus we get a first draft AC model (still neglecting substrate coupling. We will visit this detail later) of the logic as shown below.

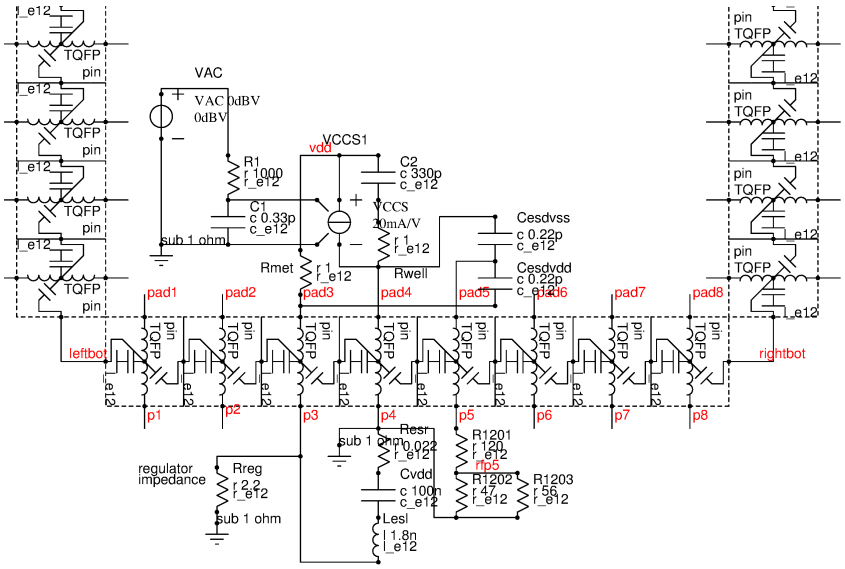


Figure 6.25: Package model combined with a first draft logic model and a model of the external blocking capacitor.

In this first model the logic is represented by an AC source and a low pass filter R1, C1 to model the shape of the envelope of the spectrum. VCCS1 converts the voltage into the current consumption of the logic. Well capacities inside the logic (and the associated well resistance) are represented by Rwell and C2. The metal path from the pad3 to the vdd of the logic is represented by Rmet.

The external blocking capacitor is assumed to have an ESR of 22 miliohm (Resr) and a parasitic inductance of 1.8nH (Lesl).

The logic is assumed to be supplied from a regulator with an RF output impedance of 2.2 Ohm (Rreg).

Although this model is still very simple it already provides a fairly correct envelope of the spectrum seen on the supply nodes of the logic.

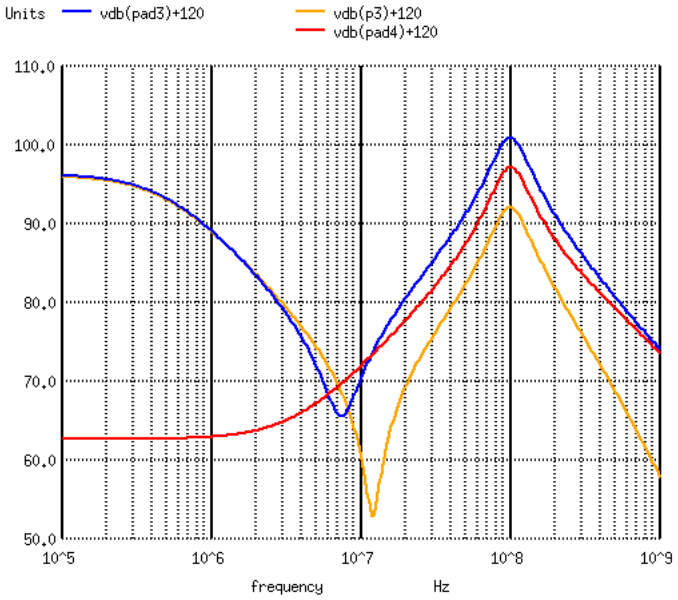


Figure 6.26: RF levels in dBuV at pad4 (chip ground pad), pad3 (chip supply pad) and pin p3 (supply pin on the board)

The RF present at pad3 and pad4 (supplies) also couples to other pins via the capacities of ESD protections. Here p5 is assumed to be floating but connected to ESD structures. It is measured with the usual 150 Ohm method set up. The capacity of the ESD protection is sufficient to produce up to 80dBuV at the measurement node rfp5 without any port activity!

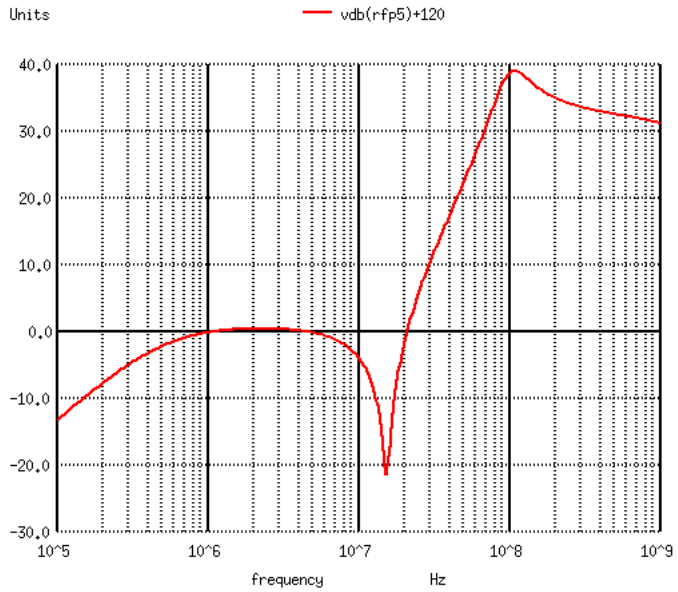


Figure 6.27: RF levels observed at p5 using the 150 Ohm direct coupling method

What is subst! The global net subst! is an approximation of something that in reality is an resistive network. The wafer usually is connected to ground at the edges (edge seal) and wherever a low resistive substrate contact is required. Position, size and metal path to the contact are designable parameters. In the following figure these are drawn in red.

Almost every component has a capacity to substrate (or handler wafer). The capacity and the resistance seen in series with the capacity mainly depends on the resistivity of the substrate material and the size of the component. (Csubst, Resrs).

From more or less middle of the chip to the back side there is a path depending on the resistivity of the sustrate material, the thickness of the chip and the area of the chip (Rvert). The inductance Lvert [41, 42] is caused by the magnetic field of the currents flowing vertically. The chip can be regarded as a conducting stub elevated over the ground plane. The magnetic field produces eddy currents in the conductive material of the substrate, dice pad and the ground plane. So Lvert is very lossy. This is represented by Rlosssubst.

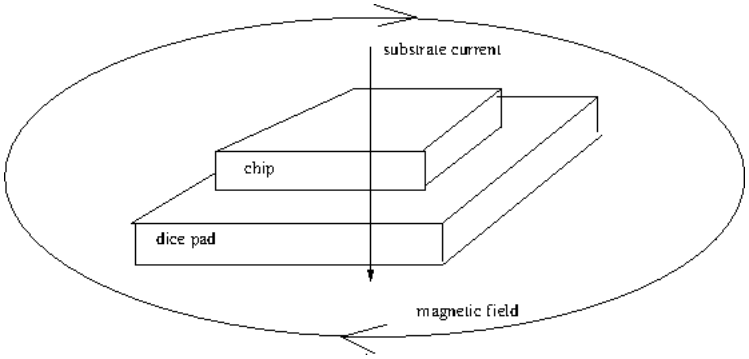


Figure 6.28: Substrate current and magnetic field surrounding chip and dice pad

If the back side of the waver is grinded there normally is a natural oxide of some nm (created simply by storing the wafers in normal atmosphere). The oxide usually gets scratched during the packaging process. So the capacitor Coxide is bypassed by a resistive path Rgrind.

In case a non conducting glue is used the back side capacity can become much lower (some hundred pF, depending on the thickness of the glue layer).

If the back side is soldered (gold plated back side) Rgrind is replaced by the resistivity of the soft solder (some micro Ohm).

The dice pad resistivity usually can be neglected. The inductance caused by the magnetic field around the metal normal to the ground plane of the board again is in the range of some pH with a lot of eddy losses (Lpad, Rlosspad). The connection of the dice pad to the ground plane only exists in the dice pad is soldered to the ground plane (exposed dice pad). If the dice pad is no soldered to ground there is a further capacity in series with the complete path.

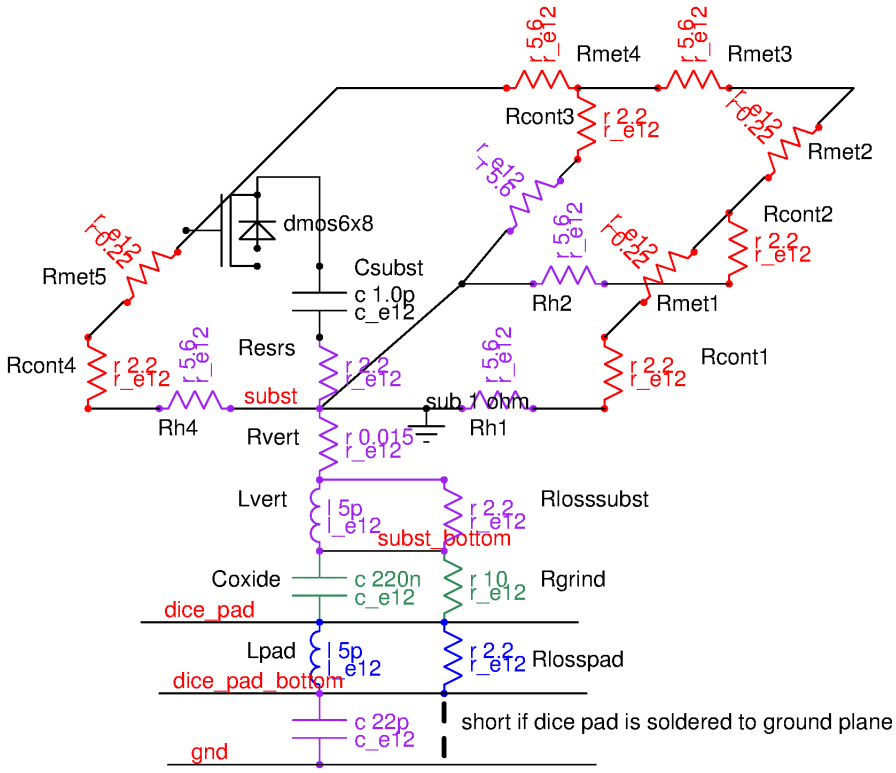


Figure 6.29: Conceptual substrate model

The way substrate is connected to board ground is decisive for the RF emission of most chips.

Furthermore it is possible to bond ground pads to the dice pad (net dice_pad) if an exposed dice pad is available. Since the number of down bonds can be high this is a very good method to create a low resistive path from chip ground to board ground even for high frequencies.

Model refinement: Now having a substrate model the most important capacities to substrate can be added. Usually these are nwells and high voltage components with significant substrate capacity as well as ESD protection structures. Include coupling of very active structures (high voltage swing combined with high bandwidth) even if these only have some hundred fF. We are looking for the uV levels in the substrate!

If multiple supplies are present (port supply to decouple logic noise etc.) these must be added together with RC models of the associated ESD protections.As an example let us have a look at the RF model of a digital IO cell. The cell consists of a push pull driver stage and the ESD protections. To get a better idea of the capacities the cross section of the technology is needed.In the following figure a typical mixed signal technology with junction isolation to the substrate and trenches between the epi tubs is shown. For better readability in the cross section only the biggest capacities are shown but not the contact spread resistances.

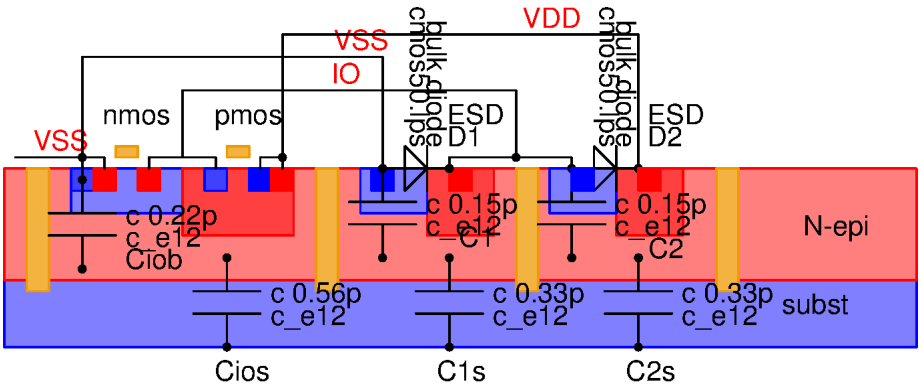


Figure 6.30: Example of a IO cell cross section

The corresponding schematic is shown below. In the schematic the contact spread resistances are included.

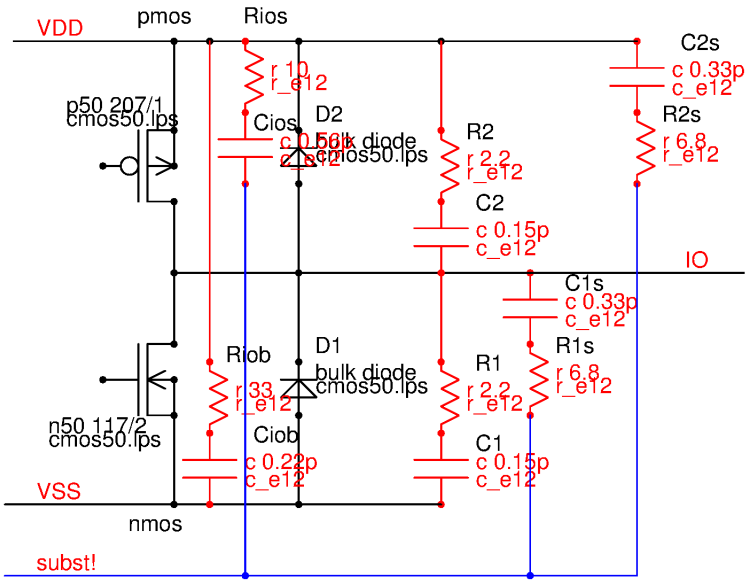


Figure 6.31: Transistor level schematic of an standard IO including parasitic capacities and contact spread resistances

In the RF equivalent circuit the transistors are replaced by voltage controlled voltage sources, representing the envelope of the spectrum of the output voltage. Each of the sources has a series resistance of double the Rdson of the transistors. The diodes are open circuits because in normal port operation they do not conduct.

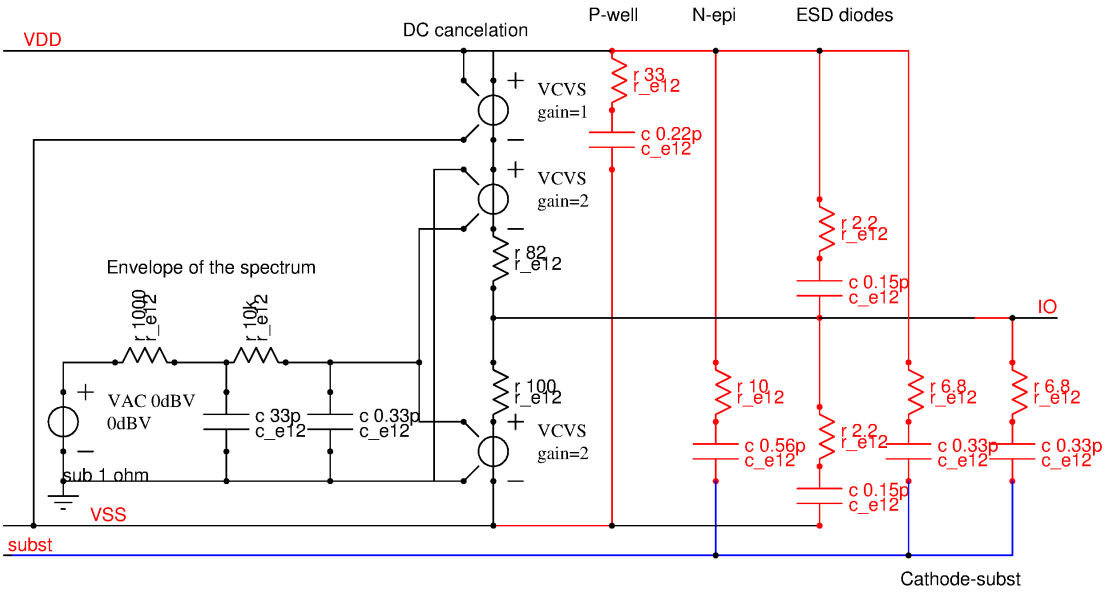


Figure 6.32: The RF model of the port

Running an AC sweep the left part of the circuit produces the envelope of the spectrum. The behavior of the output stage is simulated by the two voltage controlled voltage sources operating in push pull mode. The resistors in series with the voltage controlled voltage sources mimic the Rdson of the output transistors. Since each transistor as an average is on during 50% of the time the resistor values are twice the value of the Rdson they represent. The DC voltage is canceled by the first voltage controlled voltage source (having a gain of 1). The colored components on the right side are the parasitc capacities of the output transistors and the ESD protection.

Multiple uncorrelated RF sources: Most chips have more than one source of RF. AC simulation of most SPICE variants only sweeps one source at a time. SPICE dialects with the capability of sweeping multiple sources at a time assume they all have the same phase. This leads to adding (or in case of opposite phase subtracting) of VOLTAGES. In systems with independent functions of course there is no phase locking (Example: A chip with several port pins usually has different data streams on every pin.). To correctly model the statistical distribution of the phase of different pins using a noise analysis is the best idea. The only question is how to produce a noise source with $0dBV/\sqrt{Hz}$?

Here comes a proposal using a resistor and an ideal voltage controlled voltage source valid at 27 degrees Celsius:

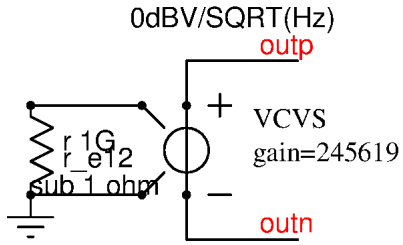


Figure 6.33: A noise source with $0dBV/\sqrt{Hz}$

The disadvantage of using a noise analysis is that per simulation run only one node can be observed. To compare different nodes multiple simulation runs are required. But since the noise simulation usually is running fast this is not too much of a problem. In case you want a SPICE netlist of the model:

```
X1 vn ref noise0dbv1
.subckt noise0dbv1 nplus nminus
rnoise n1 n0 1e9
enoise nplus nminus n1 n0 245619
rgnd n0 0 1
.ends noise0dbv
```

Resistor rgnd is part of the ground symbol to avoid conflicting net names (n0 and 0).

6.5.6 Thermal simulation

There are specific thermal simulators available on the market. Complete thermal simulations taking into account even effects of changing phase of materials (melting, vaporization) are very complex and will under normal circumstances not be used for IC development. The use of such tools normally is limited to chemical or physical process development.

If the heat flow from the package into the environment is part of the thermal simulation many different heat flow paths existing in parallel must be considered:

Convection: Heat transportation by flowing air. The air flow is caused by the temperature of the component cooled by convection. The amount of power transported by convection depends on the air pressure! Designing an IC for use in high altitude (aircraft applications) requires taking into account the reduced convection cooling by the lower air pressure. The heat flow due to convection is about proportional to the temperature. So convection can be approximated as a linear thermal resistance.

Forced air: Forcing an air flow by a ventilator will increase the heat transport compared to convection significantly. Forced air can increase the heat flow by up to one magnitude!

Heat conduction: Most of the heat gets transported out of the package by the heat conduction of the pins. The heat flows through the pins to the board. Usually the heat flow from the board to the environment is mainly by convection (see above). Heat conduction can be regarded as a more or less linear thermal resistance (There is some temperature dependence of the thermal conductivity of many materials, but usually this is in the range of some 10% within the temperature range of interest.)

Heat radiation: At very high temperatures (above 400K) heat radiation will significantly contribute to the total heat transport. Heat radiation depends on the color of the surface of the package. In space applications heat radiation may become the most important contributor of the total heat transport. Heat radiation is very nonlinear!

$$P_{radiated} = e * \sigma * A * (T^4 - T_{amb}^4) \tag{6.2}$$

with: $\sigma = 5.6703 * 10^{-8} W/m^2K^4$ (Stefan's constant). e=0..1 depending on blackness of the surface. A=radiating area.

Example: A radiating black surface with $1cm^2$ at a temperature of 400K (127 Celsius) in an environment with 300K radiates only 99mW.

Analog simulators such as spice or spectre permit building simple thermal RC models of a thermal network. This can work reasonably well as long as the range of temperatures is narrow and the heat transport is dominated by heat conduction and convection. Thermal properties of most materials (thermal conductivity, thermal capacity) are temperature dependent.

If heat radiation plays a significant role the simple linear resistor approach will fail. For this more complex situations including the heat transfer from the board to the air or even with a significant contribution of heat radiation the simple RC netlist becomes more and more inappropriate and dedicated heat simulators should be preferred.

The following piece of code shows how to include a thermal simulation in spice code.

```
* thermal simulation of a power transistor using SPICE
* basic idea: The power dissipation is described by a current source.
vgate gater 0 pulse 0 10 0 1u 1u 10u 50u
rg gater gate 10 vs vbat 0 dc 12
rload drain vbat 10m
x1 drain gate 0 tdrain tamb iplu300n4
* ambient temperature is represented by a voltage source
* attached at the bottom of the dice pad
vamb tboard 0 25 rth tamb tboard 40
**********************************
* models needed for the subcircuit *
**********************************
* the thermal resistance is described as a resistor network
* the thermal capacity is described as capacitors
* thermal resistances
* silicon: 150W/mK leads to  .model rthsi r rsh=0.00666667 tc1=1.3333e-05
* input parameters are area of the transistor in mm2 and length of the
* segment in um
* dice pad made of iron .model rthfe r rsh=0.013514
* input parameters are area of the transistor in mm2 and thickness
* of the dice pad in um
* dice pad made of copper .model rthcu r rsh=0.0025
* input parameters are area of the transistor in mm2 and thickness
* of the dice pad in um
* thermal capacities
* Si: 760 Ws/kgK * Fe: 452 ws/kgK * Cu: 385 Ws/kgK
* Specific weights
* Si: 2300kg/m3 * Fe: 7400kg/m3 * Cu: 8960kg/m3
* Since we have area in mm2 and length in um as input we need
* thermal capacity per volume
* Si: 1748000 Ws/m3K * Fe: 3344800 Ws/m3K * Cu: 3449600 Ws/m3K
*Since out input unit are mm2 and um we have to divide y 10e12
* Si: 1.7480e-06 Ws/(mm2*um*K) * Fe: 3.3448e-06 Ws/(mm2*um*K) * Cu: 3.4496e-06 Ws/(mm2*um*
.model cthsi c cj=1.7480e-06
.model cthfe c cj=3.3448e-06
.model cthcu c cj=3.4496e-06
* electrical models
.model rdex r rsh=0.54m tc1=2e-5
.model mint nmos tox=40n vto=3 nsub=6e+16
+ cgdo=3.3e-10 cgso=3.3e-10 cgbo=1e-10 js=1.107e-4 cj=1.284e-4
+ cjsw=4.43e-11 uo=513 kf=1e-28
************************************
* subcircuit including thermal model *
************************************
.subckt iplu300n4 drain gate source tdrain tamb
vtest drain drainr dc 0 rd drainr drainint
rdex w=25 l=12
m1 drainint gateint source source mint w=10 l=1u
rg gate gateint 2
rds drainint source 1e8
*power calculation
bbip 0 tdrain i=v(drain,source)*i(vtest)
*****************************************************
* here comes the transistor described as an RC network *
*****************************************************
* area is 25mm3 * thickness is 50um
* dice pad is made of copper, 200um thick
* surface in net drain
* power dissipation is entirely at the surface
* which is the worst case
* we use 10 segments for the silicon
```

```
rth0 tdrain down5u rthsi w=25 l=5
cth0 down5u 0 cthsi w=25 l=5
rth1 down5u down10u rthsi w=25 l=5
cth1 down10u 0 cthsi w=25 l=5
rth2 down10u down15u rthsi w=25 l=5
cth2 down15u 0 cthsi w=25 l=5
rth3 down15u down20u rthsi w=25 l=5
cth3 down20u 0 cthsi w=25 l=5
rth4 down20u down25u rthsi w=25 l=5
cth4 down25u 0 cthsi w=25 l=5
rth5 down25u down30u rthsi w=25 l=5
cth5 down30u 0 cthsi w=25 l=5
rth6 down30u down35u rthsi w=25 l=5
cth6 down35u 0 cthsi w=25 l=5
rth7 down35u down40u rthsi w=25 l=5
cth7 down40u 0 cthsi w=25 l=5
rth8 down40u down45u rthsi w=25 l=5
cth8 down45u 0 cthsi w=25 l=5
rth9 down45u down50u rthsi w=25 l=5
cth9 down50u 0 cthsi w=25 l=5
* now we reach the dice pad split in 2 layers
rth10 down50u middle rthcu w=25 l=100
crh10 middle 0 cthcu w=25 l=200
rth11 middle tamb rthcu w=25 l=100 .ends iplu300n4
.end
```

Normally such code is generated automatically from a simple description holding information such as materials, wafer thickness, power transistor sizes etc. Here it is shown in detail to demonstrate the concept.

Perl is a simple programming language optimized for handling textual information. To write such automatic netlisters using perl or similar languages is suggested.

6.5.7 Using digital simulators for system simulation

The use of verilog or vhdl is very attractive to achieve fast simulation runs. The analog blocks in this case must get a digital representation. Some examples are:

- replace charge pumps by up/down counters
- Use integers instead of real numbers for analog systems such as delta sigma converters to accelerate simulation (for instance 1 represents 1mV, 1000 represents 1V)

One of the issues using digital simulators for analog problems is the accumulation of truncation or rounding errors especially in intergrating systems.

6.6 Analytical solvers

Analytic solvers try to solve mathematical expressions in an analytical way rather than in a numeric way. Typical examples are Maxima, wxMaxima, Mupad. Some tools are available for simple questions in the internet such as Wolfram Alpha.

6.7 Waveform viewers

To view the simulation results you need a waveform viewer. There are various waveform viewers available for various formats. In addition commercial tools such as Cadence virtuoso bring their own built in waveform viewers.

Certain programming languages such as python have powerfull plot commands that allow writing your own viewer in a reasonably easy way.

6.7.1 dinotrace

Dinotrace is an open source viewer for digital simulation results. It can read the formats tra (DECIM), bt* (Tempest CCLI), vcd (Verilog change dump), vpd (Verilog VPD+). The following figure shows an example of using dinotrace.

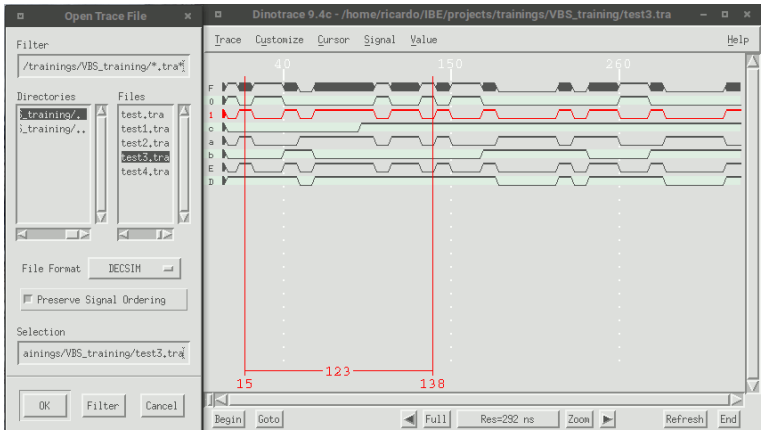


Figure 6.34: Dinotrace

6.7.2 gaw

gaw is a simple open source viewer for analog wave forms. It can be found at http://gaw.tuxfamily.org/. Here it is what it looks like:

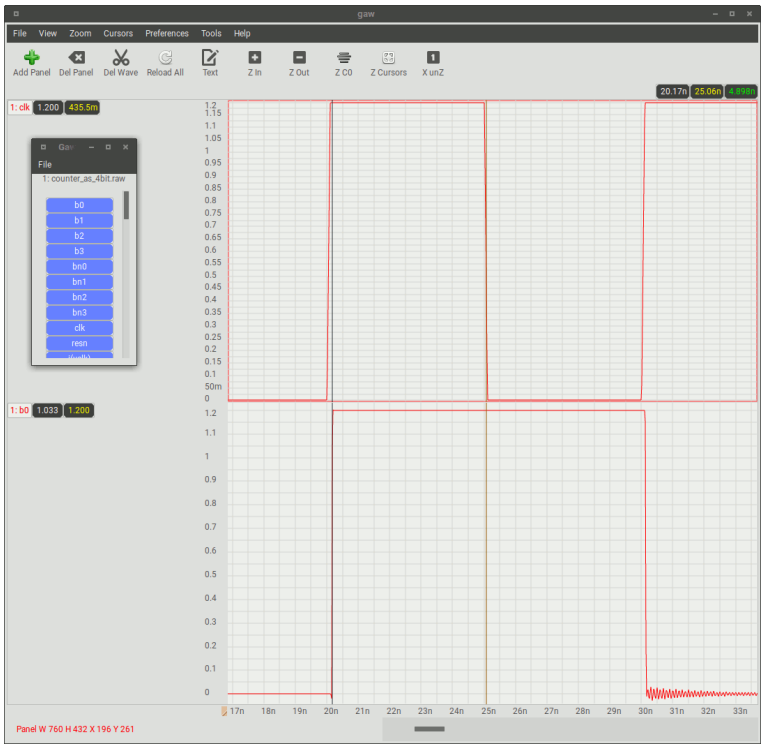


Figure 6.35: Screen of gaw

For details how to setup SPICE to create the correct format for gaw see section 6.2.12.

6.7.3 gnuplot

gnuplot is a powerful tool to plot almost everything that is saved in some kind of ASCII dump. Typically gnuplot needs a file in which the data is arranged in columns. gnuplot doesn't care if the data is coming from a simulator of from measurements or an editor. If it is ASCI and in columns gnuplot can handle it. Different from EXEL gnuplot hase no line limit (at least I didn't find a limitation plotting up to 8 million lines of sampling points of an oscilloscope)

gnuplot is controlled from a console with text mode commands. Typically it is launched with a command like:

```
ricardo@jupiter:~/projects/book> gnuplot
```

If gnuplot is started you are in an interactive shell. gnuplot expects command now. The help function can easily be called:

```
gnuplot> help
```

This command takes you to a general help function showing you how to get started. To get to the next help menu just type “q”. This takes you to something looking like this:

```
Help topics available:
    2D                  datafile            introduction        rectangle
    3D                  datastrings         iteration           rgbcolor
                .                                   .                       .
Help topic:
```

To get a specific information just type in one of the suggested key words.
To exit the help function type in “^C”. This takes you back to the command line of gnuplot.
since you probably don’t want to learn the complete usage in the first session here are some of the most frequently needed commands to get started:

```
gnuplot> plot log(x)
```

gnuplot has some built in mathematical functions such as sin(x), cos(x), tan(x), exp(x), log(x)... and many more. Nice to get a first idea of a mathematical function.
You may want to have a grid:

```
gnuplot> set grid
gnuplot> replot
```

Just setting a grid won’t do anything. You have to replot to make it visible.
Now the scale is a bit inconvenient. How about:

```
gnuplot> set logscale y
gnuplot> replot
```

May be you need an other range than the default:

```
gnuplot> set xrange [-20:20]
gnuplot> replot
```

This simple example yield a window like this:

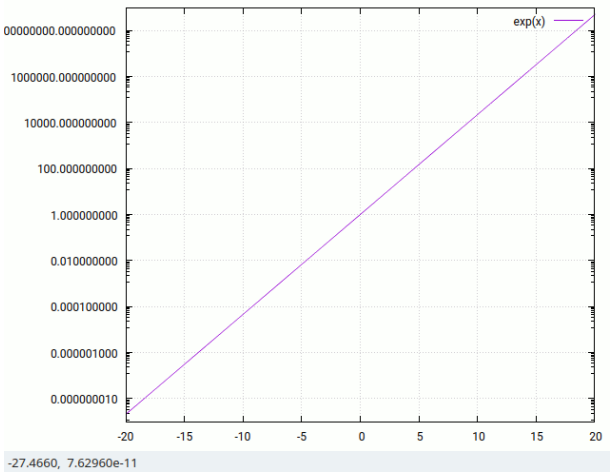


Figure 6.36: A typical gnuplot graph

Using built in functions is nice, but for viewing simulation results we want to see a file:

```
plot "gnuplot_test.txt" using 1:2 with linespoints
```

Here we go:

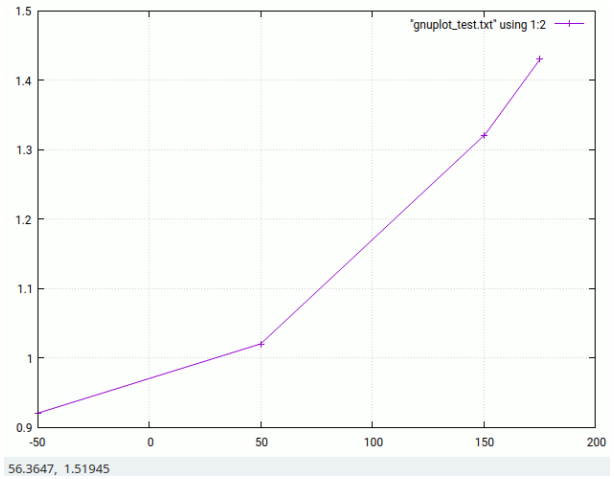


Figure 6.37: Reading a numerical file with gnuplot

The command simply toot the column 1 for the X- and column 2 for the Y-coordinates. The plot style is linespoints. It can just as well be points or lines only. Of course the file must have a certain format. It looks like this:

```
#test of gnuplot
#temp    Vos
-50      0.92
50        1.02
150      1.32
175      1.43
```

Using gnuplot to view simulation results you have to make your simulator dump the results in such a format. In spice this is done using the print command with the option "no page break". Here we go using ngspice:

```
ngspice 1040 -> tran 0.1n 40n
ngspice 1041 -> option nobreak
ngspice 1042 -> print v(1) > spiceout.txt
```

The resulting file is close to what we need. We only may have to add the gnuplot comment sign "#" to comment out the header (just use an ASCII editor before trying to plot. Often this isn't even needed because gnuplot is quite tolerant). For plotting be aware that the first column in spice is the sample number. The second column is the time and the third column is the signal.

To create meaningful plot some labels for the axes would be nice. Here is an example:

```
plot "spiceout.txt" using 2:3 with lines
gnuplot> set xlabel "time"
gnuplot> set ylabel "volt"
set label 1 "node 1" at 2.3e-8,1.1
gnuplot> replot
```

The position of xlabel and ylabel is automatic. The position of a label refers to the coordinates of the plot.

Voila:

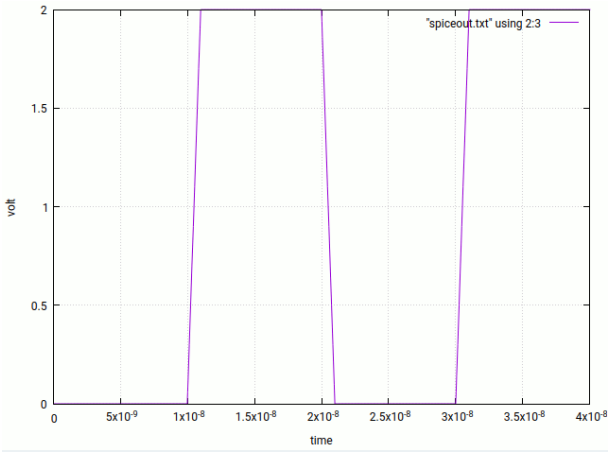


Figure 6.38: Example of plotting the ngspice result with gnuplot

gnuplot can be controled by scripts as well. There are two ways of running scripts:

- From the command line gnuplot can be called with something like:

  ```
  gnuplot file -
  ```

The file holds the commands gnuplot is expected to execute. Here comes an example:

```
plot "cross_g3_v5p_0x1f_0n_Ch4.txt6" using (1e9*$1):2 with lines title "V5P voltage, 0nF at G3"
replot "cross_g3_v5p_0x1f_1n_Ch4.txt6" using (1e9*$1):2 with lines title "V5P voltage, 1nF at G3"
replot "cross_g3_v5p_0x1f_3n_Ch4.txt6" using (1e9*$1):2 with lines title "V5P voltage, 3nF at G3"
replot "cross_g3_v5p_0x1f_8n_Ch4.txt6" using (1e9*$1):2 with lines title "V5P voltage, 8nF at G3"
# now time scale is ns set xlabel "time in ns"
set ylabel "V5P versus HVGND in V"
set grid set title "V5P versus HVGND at varying load at G3, current 0x1f"
set label "control file: cross g3 v5p 0x1f On.gnucmd" at 2000,0.9
set xrange [1200:3200] replot
```

The example shown plots the results of 4 ascii dumps of an oscilloscope, converts the time scale from seconds to nanoseconds and places some labels and a headline.

After plotting gnuplot must remain in interactive mode. The "-" trailing the last file name takes gnuplot into interactive mode instead of quitting immediatelly.

The result of the little script loos like this:

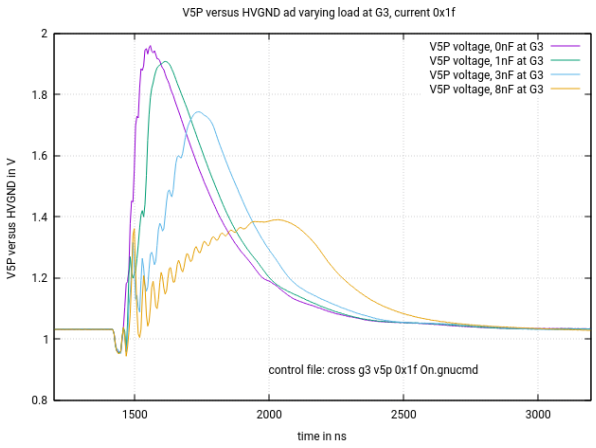


Figure 6.39: result of the little plotting script

- If gnuplot is already running the script can be launched by:

  ```
  load "file"
  ```

After the interactive load command gnuplot expects a variable name. If you want to directly enter the name of the file you need the " ". There is no more "-" needed because we are already in interactive mode of gnuplot.

6.7.4 gtkwave

gtkwave is a viewer for digital signals similar to dinotrace. It can be found at http://gtkwave.sourceforge.net/ . Many Linux repositories already hold precompiled versions of gtkwave. Here it is what it looks like:

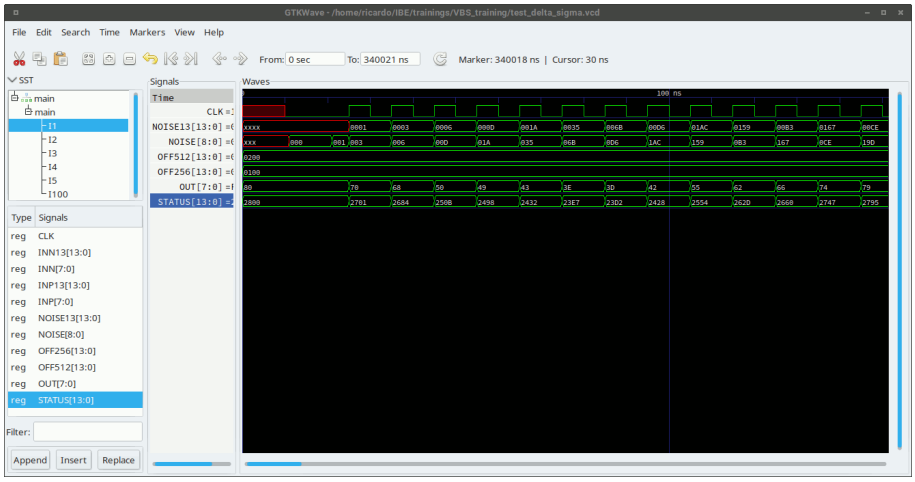


Figure 6.40: Example of using gtkwave

The usage (how to make verilog and VHDL dump the data in the right format) is described in section 6.3.1 and 6.3.2.

7 Basic Circuits

Before starting the real part of the book lets have a look at the most important principles of robust design.

There are tolerances everywhere! This applies to the technology (almost every parameter has an absolute spread of several ten percent!), to the matching (The smaller the device the bigger the impact of minor deviations of size) and to the operating conditions (temperature, supply voltage ...). Whatever you design must be checked for the impact of tolerances. The more complex your design gets the more difficult it becomes to understand all the influences. Thus the first rule is:

- Keep it simple

Device matching is magnitudes better than the spread of absolute technology parameters.

- Rely on matching but not on absolute parameters of the technology

The more sophisticated a component is, the more technology parameters will contribute to the tolerances. You will quickly loose sight which parameter impacts what.

- Use the most generic devices you can. (basic bipolars, NMOS and PMOS without any “high voltage tricks”)

Regulation loops always have the risk of becoming unstable. The instability depends on internal poles as well as the load poles.

- Keep specifications wide enough to get along without regulation loops.
- Avoid implementing regulation loops if they are not needed.
- If you can't avoid them keep the regulation loops as small and simple as possible.
- Minimize the number of gain stages in the loop to keep the poles under control.
- avoid nested loops

The more complex the design gets and the more technology parameters it depends on the longer the design verification (simulation of all corners or Monte Carlo simulation) will take. Running a Monte Carlo transistor level simulation on a clocked regulation loop to poke out the impact of transistor mismatch is a nightmare!

But I must admit sometimes you only can reach the target parameters with such a complex design. In this case you will have to go through this simulation hell. This makes sense if the anticipated production volume of your design is high. If the production volume is low and the main cost is the development going through this simulation hell may not be worth while.

Regarding production only chip real estate matters. If you can reduce the chip area using a complex CMOS design in stead of an area consuming bipolar design it may in fact make sense to go through the hassle of a complex design if the production volume is big enough.

Trimming often is proposed to improve accuracy. Use it with care because every non volatile memory cell I know has certain quality risks (loss of stored data). This becomes especially true at high operating temperature, operating in corrosive environment and in aerospace applications (erasure of the memory by radiation).

- Avoid trimming as long as you can to achieve long term stability of the design in application.

A word about simulation: In the simulation chapter I have shown you some software bugs I have found in commercial tools. Of course there ate thousands of further bugs. The only suggestion I can give is to first calculate each and every parameter solving the equations. If your simulation deviates more than some percent (that can be explained by 2nd order effects) from the calculation check for software bugs.

- check correctness of the netlist
- Are parameters in the netlist consistent with the parameters in the GUI?
- Are the models used correct (confusing an M with an m in the model file happens more often than you will beleave it!)
- are the pin orders correct (I have already seen netlists with confused pin orders!)
- Is there a ".end" somewhere in the middle of the netlist? (whatever comes behind the ".end" will be ignored!)
- Are the models complete? (Often parasitic transistors are neglected!)

It isn't uncommon to find out after weeks that all you simulated was in vain due to a software bug. Always calculate first and check for plausibility!

7.1 Voltage dividers

Building good voltage dividers is an art of its own! The diffculties usally are hidden in the following fields:

- Good matching requires carefull design using equal modules for all resistors involved
- Resistors available on integrated circuits often are voltage dependent
- Resistors available on integrated circuits change with temperature
- Electrical over stress can change resistor values
- Every resistor has parasitic capacitive coupling to the substrate under it, to neighbouring components and to wires passing nearby

7.1.1 Design for good matching

The width of a resistor usually is not exactly the same as the opening on the mask. Depending on the photolithography and in some cases outdiffusion and scattering of implants the width on the silicon can significantly differ. Well matching resistors must consist of unit cells af the same width.

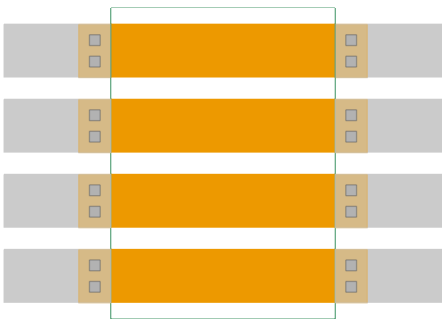


Figure 7.1: Well matching resistors

The resistors shown above are expected to match well for the following reasons:

1. All resistors have the same drawn width. If the width on the silicon differs the error is the same on all resistors.
2. In the contact area the current flows in an inhomogenous way. All resistors have the same heads.

3. Salicide stop mask does not cover the heads of the resistors. So the contact resistivity is low compared to the (more precise) resistor body.
4. The resistor body is long compared to the width (W/L>6 is recommended by several publications [39])
5. Metal traces have different mechanical properties. Since all metal traces have equal width the mechanical stress on all resistors is equal
6. No metal traces are crossing the resistors (I you have to cross one resistor add dummy crossings so all resistors are exposed to the same stress.
7. The resistors are close to each other so doping gradients are expected to have a low impact.
8. Due to proximity all resistors have a similar temperature.
9. For high precision (matching better than 1%) use the inner resistors only because the outer resistors may be exposed to outdiffusion of adjacent heavily doped areas (Sinkers, P+ or N+ doped areas)
10. If the width of the resistors approaches the wave length of light the inner resistors are in a periodic pattern. So the influence of interference of the light on the neigbouring structure is equal. This does not apply to the resistors at the edge. So edge resistors should be used as dummies but not included in precision dividers.

In the following figure some severe violations of good matching are shown:

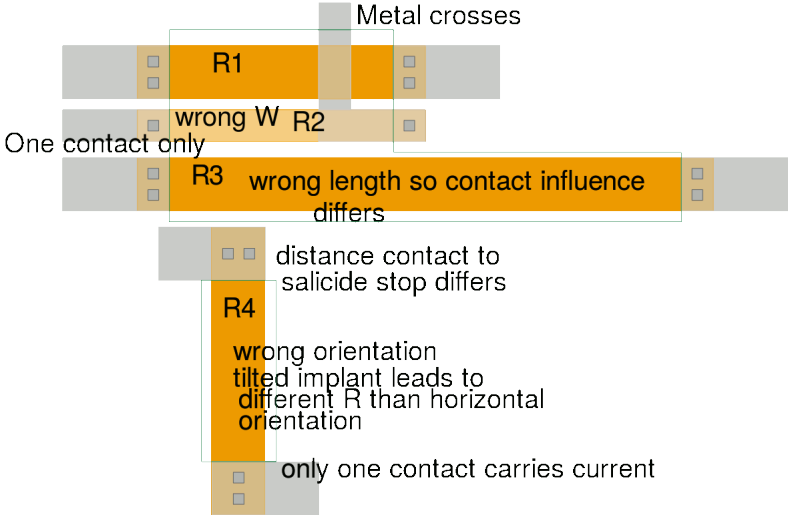


Figure 7.2: Matching violations

With the layout shown the following matching errors must be expected:

R1 and R2 missmatch due to the different width can be as high as 30%. Since at the edges of resistors often the poly silicon doping is dilluted the temperature coefficients of R1 and R2 differ additionally.

R1 and R3 are not factor two different. R3 holds only 2 contact areas per 2 unit lengths. This can lead to about 15% mismatch of the ratio 2.

Orientation errors between R1 and R4 can lead to about 30% deviation. Additionally R4 has different contact shapes and spacings between contacts and salicide stop mask. Expect 5% more errors for the contact deviations.

Anisotropic material: Diffused resistors were reported to differ depending on the direction the current flows through them, The differences observed in 1:2 voltage divides were 0.3%. (Measured using P-body resistors as reference dividers on a 4 channel valve driver, 1000 samples, technology BCD2)

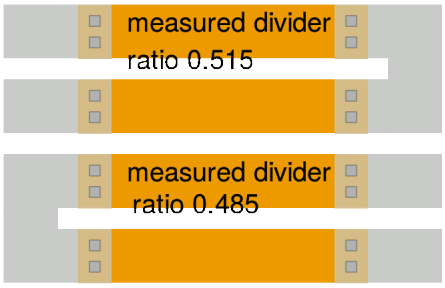


Figure 7.3: Impact of different current direction in diffused resistors

On poly silicon resistors such current direction dependence was not yet reported (at least I haven't heard any reports of such effects on poly silicon)

Self heating: Poly silicon resistors and metal resistors are embeded in silicon oxide. Thermal conductivity of SiO2 is about 0.9W/m*K compared to silicon with 110W/m*K (at 300 deg. C) and even 150W/m*K at room temperature. As a consequence poly silicon resistors tend to heat up and the 1:3 divider can deviate due to the different temperatures of the resistors.

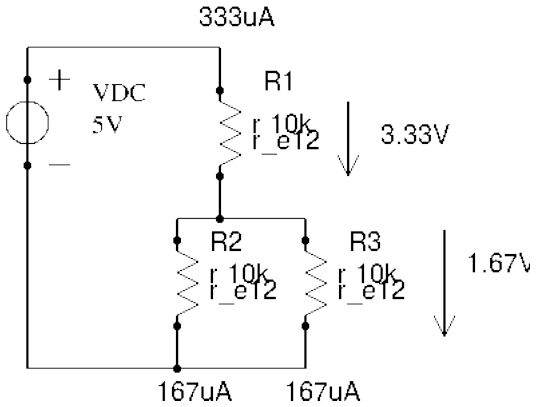


Figure 7.4: Self heating because R1 dissipates 1.11mW while R2 and R3 each dissipate 0.277mW

As a precaution against self heating try to operate all resistors at the same current density (so self heating is equal). Keep current densities low because even operation at equal currents will lead to a different temperature of the inner resistors compared to the edges of the array.

Example: Let us assume R1 is $20\mu m$ long and $2\mu m$ wide. The silicon oxide between the poly silicon and the substrate is $0.5\mu m$ thick. Rth becomes

$$R_{th} = \frac{d}{W * L * \lambda_{siO2}} = \frac{0.5\mu m * 10^6 \mu m * K}{2\mu m * 20\mu m * 0.9W} = 13.9K/mW$$

So R1 heats up about 15.44K while R2 and R3 only heat up 3.86K. The resulting error depends on the temperature coefficient of the resistor material. Several percent of error at a temperature difference of 10K are not unusual.

High voltage dividers: Resistors (even poly silicon resistors!) are sensitive to the voltage drop across the resistor and the voltage between the resistor and the substrate or the well underneath. For good linearity over a wide voltage range the resistors must be made of equal segments with wells under each segment. The wells under each of the segments must be driven such that the voltage between the resistor segments and the wells underneath are equal.

The wells can be taken from taps of the divider. But this leads to a capacitve loads attached to the divider. In a feedback loop this can be fatal. Alternatively the wells can be driven from a second, possibly lower precision, divider to avoid the capacitive load.

7.2 Current mirrors

In integrated circuits the good matching of transistors can be used to build current mirrors.

7.2.1 Bipolar current mirrors

Bipolar curret mirrors are used in technologies offering bipolar transistors only. As soon as a technology offers CMOS transistors it is in most cases advantageous to implement current mirrors using the CMOS transistors. Nevertheless it is a good idea to looks at these early current mirrors to investigate some of the principles.

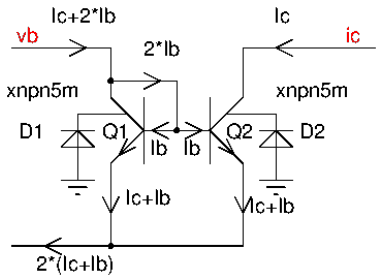


Figure 7.5: Basic bipolar current mirror

The transistors Q1 and Q2 act as a current mirror. The substrate diodes D1 and D2 under normal circumstances are not important (at least as long as nobody pulls the emitters of Q1 and Q2 more than 2 forward voltages below substrate level).

Q2 performs a logarithme operation of the current offered at pin vb. Neglecting the base currents we can approximate the base voltage.

$$I_C = I_{CS} * (e^{\frac{V_{BE}*e}{k*T}} - 1) \tag{7.1}$$

For $V_{BE} >> \frac{k*T}{e}$ the above equation can be simplified

$$I_C \approx I_{CS} * exp(\frac{V_{BE}}{V_T}) \tag{7.2}$$

with

$$V_T = \frac{k*T}{e} = 26mV$$

at room temperature. Usually Q1 and Q2 are operated at about $V_{BE} \approx 600mV$. This is high enough to justify this simplification. The simplified equation can be rearanged to:

$$V_{BE} = V_T * ln(\frac{I_C}{I_{CS}}) \tag{7.3}$$

Note that I_{CS}has a stron temperature dependence dominating the temperature proportionality of V_T. Usually V_{BE} decreases with about -2mV/K at constant collector current.

Transistor Q2 performs the inverse mathematical operation. The collector current of Q2 aproximatelly is an exponential function of the base voltage. Thus we get (neglecting losses caused by base currents):

$$I_{CQ2} = I_{CS} * exp(\frac{V_T * ln(\frac{I_{CQ1}}{I_{CS}})}{V_T}) = I_{CQ1} \tag{7.4}$$

Well, did you expect anything else?

Things are getting more interresting looking at the errors of the circuit.

Base current error: Since the input has to provide the collector current of Q1 plus both base currents we will find:

$$I_{in} = I_C * (1 + \frac{2}{B}) \tag{7.5}$$

B is the current gain of the transistors. Bipolar transistors (intentional ones, parasitc transistors hopefully are much weaker) found in integrated circuits usually have gains of $B = 10$ to $B = 400$. So the ratio of the current mirror typically becomes:

$$K_1 = \frac{I_{CQ2}}{I_{in}} = \frac{B}{B+2} \tag{7.6}$$

for a mirror with equally sized transistors. As soon as the output transistor is bigger (emitter erea N times bigger on the output side) the output current will increase and at the same time the base current increases as well. This leads to the ratio:

$$K_N = \frac{B*N}{B+N+1} \tag{7.7}$$

Looks harmless, but let us calculate for a mirror intended to be N=10 at a gain of B=100:

$$K_{10-100} = \frac{100 * 10}{100 + 10 + 1} = 9.009$$

So we have to think about reducing the base current error especially for big ratios N. But before jumping into a new circuit let us have a look at the statistical errors too.

Early effect Due to base with modulation bipolar transistors are not ideal current source. The collector currenthas a certain dependence on the collecto emitter voltage. So if the output voltage of a current mirror is varied the current ratio changes.

$$I_{out}(V_{ce}) = I_{in} * (1 - \frac{2}{B}) * (1 + \frac{V_{ce} - V_{be}}{V_{early}}) \tag{7.8}$$

The change of the collector current with the collector voltage is shown in the folloing figure.

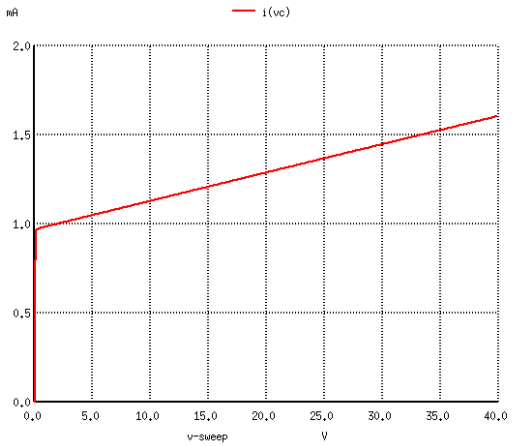


Figure 7.6: Dependence of the collector current on the early voltage

The output impedance of a current mirror calculates

$$R_{out} = \frac{dI_C}{dV_{ce}} = \frac{I_c(V_{be})}{V_{early}} \tag{7.9}$$

Improved current mirrors There are several possible enhancements for bipolar current mirrors.

Boosted current mirror The boosted current mirror uses a transistor in common collector topology to provide the current needed to drive the base of the transistors in the current mirror. The error current needed to drive the base of the current mirror transistors is divided by the gain of the booster transistor. This topology is especially interresting for current mirrors with a ratio k>1.

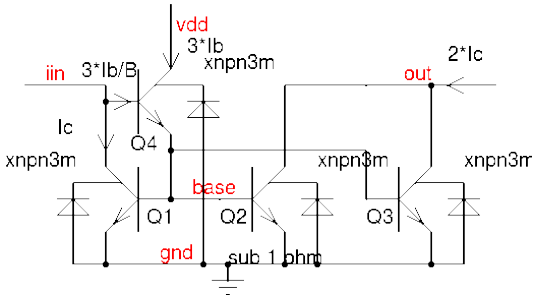


Figure 7.7: Boosted current mirror

The avove example shows a mirror with k=2. The input iin current calculates as

$$I_{in} = I_c + (k+1) * \frac{I_b}{B} = I_c * (1 + \frac{(k+1)}{B^2}) \tag{7.10}$$

Example: B=100,K=2 leads to

$$\frac{I_{out}}{I_{in}} = k * \frac{1}{1 + \frac{3}{100^2}} = 1.994$$

Without the booster it would have been 1.9417.

One of the disadvantage of the boosted current mirror is that now we have an emitter follower driving the diffusion capacity of Q1 to Q3. This creates one pole of the regulation loop. The transistor Q1 can be regarded as an amplifier stage driving the collector capacity of Q1 plus the base capacity of Q4. Thus we find two poles in the loop. As a consequence boosted bipolar current mirrors must carefully be checked for stability. If required the pole caused by the diffusion capacities of Q1 to Q3 can easily be shifted to higher frequencies adding a dummy load. The dummy load increases the emitter current of Q4 and reduces the output impedance of Q4. But this adds errors again....

The early effect however is still not solved. This can be done using cascoded mirrors.

Cascode current mirror for K=1 In case of a current mirror with ratio K=1 there is a very elegant solution to solve the base current error and the early error [4].

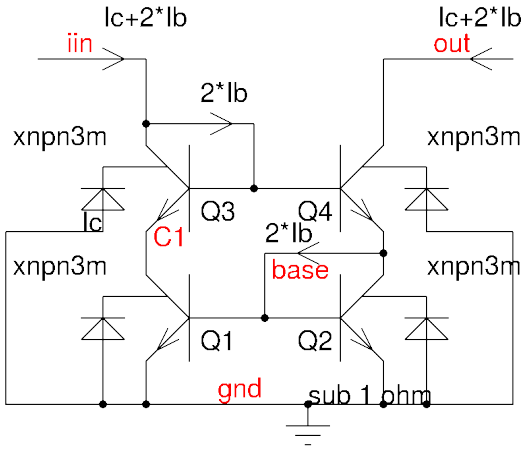


Figure 7.8: Current mirror with cascodes

Now we have the same base current error on both sides. Thus it cancels. Transistors Q1 and Q2 operate at (almost) the same collector voltage. A variation of Vout will change the emitter voltage of Q4 by

$$dV_{e4} = dV_{out} * \frac{I_c}{gm * V_{early}} \tag{7.11}$$

gm is the transconductance of Q4. It is a function if the collector current.

$$gm = \frac{I_c}{V_T} \tag{7.12}$$

Thus dV_{e4}becomes

$$dV_{e4} = dV_{out} * \frac{V_T}{V_{early}} \tag{7.13}$$

So Q1 and Q2 will see a difference of the collector voltage of dV_{e4}wich is the difference of the collector voltages of Q3 and Q4 reduced by factor $\frac{V_T}{V_{early}}$. The output impedance of the current mirror becomes

$$R_{out} = \frac{V_{early}^2}{I_c * V_T} \tag{7.14}$$

Example: using bipolar transistors with an early voltage of 20V the output impedance increases by factor 20V/26mV=769.

Using bipolr transistors this nice solution only works well for current mirrors with ratio K=1. For big current mirrors or current generators a different solution must be used.

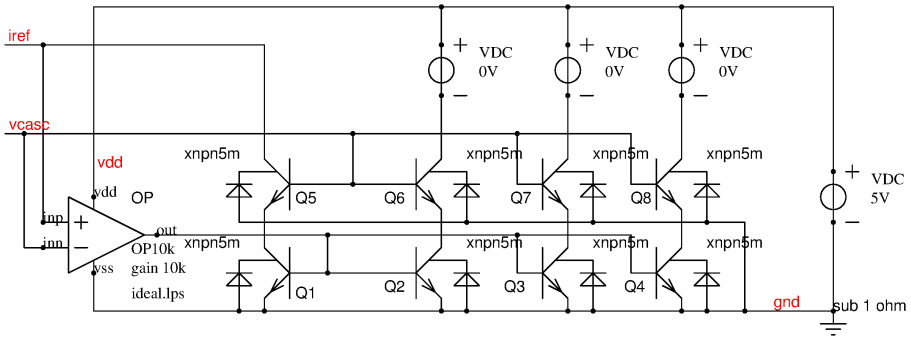


Figure 7.9: Example of a precision current generator

In the above example the cascodes are driven from a deicated supply that provides the base currents of Q5 to Q8. Typically such a supply is about 2*Vbe. The regulator OP regulates the base voltage of Q1 such that the collector voltage of Q5 becomes equal to the vascode supply vcasc. Q1 as well as Q2, Q3 and Q4 have a collector current of iref+Ib. So the output current flowing into the test sources is equal iref (because Q6, Q7 and Q8 have the same base current as Q5. The input bias current of OP must be magnitudes lower than iref.

One disadvantage of the circuit is the limited speed of the regulation loop. So this kind of precision current generator can not follow fast changes of iref.

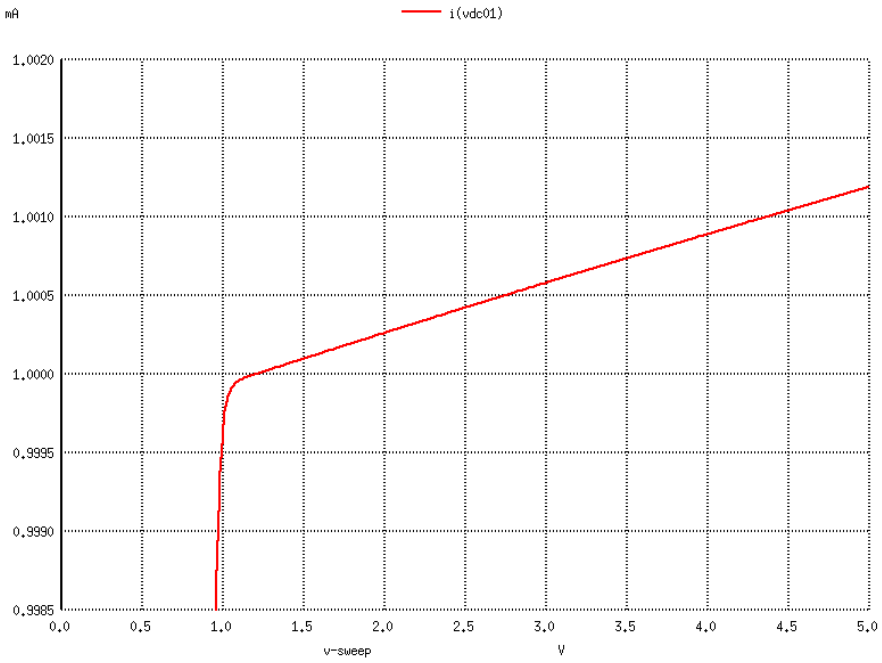


Figure 7.10: Sweeping the collector voltage of Q6 from 1.1V to 5V the current only changes from 0.9998mA to 1.0012mA at a reference current of 1.000mA

The use of cascodes of course limits the available voltage swing at the collectors of Q6..Q8.

Statistical errors: Statistical errors of current mirrors usually are caused by offset voltages. Offset voltages between transistors result from doping gradients and mask inaccuracies. Usually the statistical offset of transistors is noted in the design manual. (Usually the 1σ spread). As long s the spread is small compared to V_T we can linearize the problem and simply add error contributions.

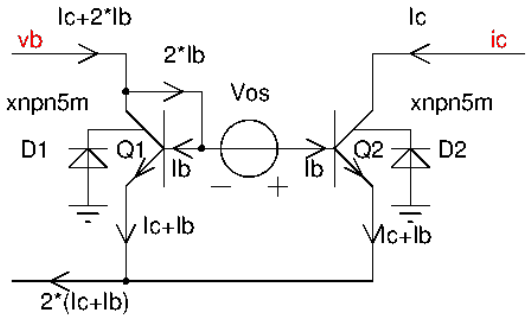


Figure 7.11: Current mirror with offset

The derivative of the current is:

$$\frac{dI_C}{dV_{BE}} = \frac{I_C}{V_T} = gm \tag{7.15}$$

Thus the linearized current error becomes

$$\Delta I_C = V_{os} * gm \tag{7.16}$$

Usually we are interrested in the relative change if the current

$$\frac{\Delta I_C}{I_C} = \frac{V_{os} * gm}{I_C} = \frac{V_{os}}{V_T} \tag{7.17}$$

Example: Working with bipolar transistors having an offset of 0.5mV the expected 1σ spread is $\pm 2\%$. To achieve good production yield the design should be built to be able to handle $6\sigma = 12\%$ spread of the mirror.

Usually it helps to increase the emitter area. This can be done using several transistors in parallel or using transistors with enlarged emitters. The enlargement is n.

$$\frac{\Delta I_{Cn}}{I_{Cn}} = \frac{V_{os}}{\sqrt{n} * V_T} \tag{7.18}$$

Looks like we have to find ways to reduce gm! An easy way to reduce gm is to add resistors in the emitter paths of the transistors.

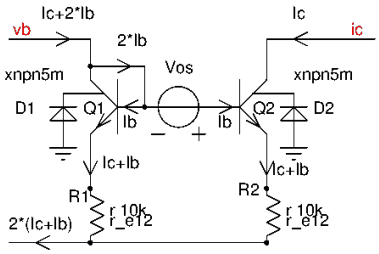


Figure 7.12: Current mirror with gm-degradation

gm of the bipolar transistor can be regarded as the emitter output impedance. After adding the resistors $R_1 = R_2 = R$ the resulting gm_r becomes

$$gm_r = \frac{1}{R + \frac{1}{gm}} \tag{7.19}$$

The current error caused by the offset of the transistor reduces to

$$\frac{\Delta I_C}{I_C} = \frac{V_{os} * gm_r}{I_C} = \frac{V_{os} * \frac{1}{R + \frac{V_T}{I_C}}}{I_C} \tag{7.20}$$

Usually the drop over the resistor is chosen at least factor 5 higher than the temperature voltage V_T and the above equation can be aproximated by

$$\frac{\Delta I_{Cos}}{I_C} = \frac{V_{os}}{I_C * R + V_T} \approx \frac{V_{os}}{V_R} \tag{7.21}$$

with V_R being the voltage drop over the resistor. A nice improvement but we added more components. The spread of the resistors will start to contribute to the spread of the current mirror. In the above equation we are observing the

current error caused by the offset alone. This is why now we call the change of the current ΔI_{Cos} . The deviation of the resistor in independent of the offset of the transistor. So we can add the "energy" of the errors.

$$\frac{\Delta I_C}{I_C} = \sqrt{(\frac{V_{os}}{V_R})^2 + (\frac{\Delta R}{R})^2} \tag{7.22}$$

Adding resistors only makes sense as long as the resistors match significantly better than the transistors. Fortunatelly the resistors on a chip can be made almost unlimited accurat by increasing the size. (At least as long as we don't have doping gradients killing resistor matching.)

Knowing how to improve the matching by emitter degradation it is time to come back to the systematic errors.

7.2.2 MOS current mirrors

As soon as MOS transistors are available current mirrors can be implemented smaller using the CMOS transistors. Furthermore MOS current mirrors don't suffer from errors caused by base currents - at least as long as the gate oxides are thick enough to allow neglecting gate tunneling currents. The most simple current mirror is shown below.

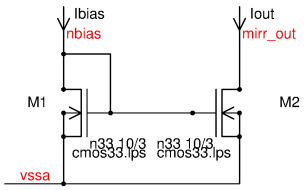


Figure 7.13: Most simple NMOS current mirror

Since we have no gate current the currents Ibias and Iout are equal (assuming perfect matching of the transistors). The current mirror shown has a ratio of 1. If other ratios are required transistor M2 can be scaled without DC drawbacks like in the bipolar case. Since MOS transistors have slightly different thresholds at the edges of the channel the transistors should be designed in equal modules. The following figure shows variants of the current mirror halfing or doubling the current. There however is one detail to be considered: Placing several transistors in series changes the early voltage the combination M21, M22. Therefore the currents will not match exactly anymore unless cascodes are used to limit the impact of the changed early voltage.

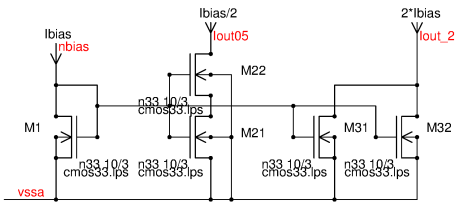


Figure 7.14: Current mirrors with ratios different from 1

The ratio of the current simply follows the number of transistors in the reference (M1) and the output (example: M31 and M32 are two transistors in parallel, so we get a ratio of 2). For very big current ratio mirrors this simple approach gets fairly expensive. To achieve a high ratio at reasonable cost a gate voltage shift is a reasonable approach.

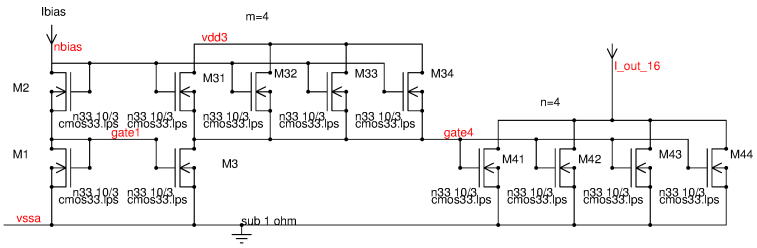


Figure 7.15: m*n current mirror using a gate voltage shift.

Using a straight forward design the current multiplication by factor 16 requires 17 transistors (one MOS diode plus 16 output transistors). The circuit shown above requires only 11 transistors. The trick is the reduced current density M31 to M34 are operating with. This creates an increase of the voltage of gate4 versus gate1. M1 and M2 operate at the same current density. So they both have the same (effective) gate-source voltage

$$V_{gs_{eff}} = \sqrt{\frac{I_{nbias} * L}{g_m * W}} \tag{7.23}$$

M3 forces the same current through M31 to M34. Since these are 4 devices in parallel (lower current density in M31 to M34) the voltage at gate 4 becomes:

$$V_{gs4_{eff}} = 2 * V_{gs_{eff}} - \sqrt{\frac{I_{nbias} * L}{m * g_m * W}} \tag{7.24}$$

$$V_{gs4_{eff}} = V_{gs_{eff}} * (2 - \frac{1}{\sqrt{m}}) \tag{7.25}$$

The current flowing into the transistors M41 to M44 becomes:

$$I_{out} = n * \frac{g_m * W * V_{gs4_{eff}}^2}{L} \tag{7.26}$$

$$I_{out} = n * \frac{g_m * W * V_{gs_{eff}} * (2 - \frac{1}{\sqrt{m}})^2}{L} \tag{7.27}$$

$$I_{out} = n * (4 - \frac{4}{\sqrt{m}} + \frac{1}{m}) * I_{nbias} \tag{7.28}$$

If m gets very high the term $(4 - \frac{4}{\sqrt{m}} + \frac{1}{m})$ approaches 4. This becomes very attractive due to the following behavior:

- with increasing m the spread of m has a decreasing influence on the multiplier 4
- m can be made big not only by increasing the number of transistors M31... but also by decreasing the length of M31..34
- Instead of using modules M31..34 can be given a very high aspect ratio by just making it a single transistor with high W/L
- m can further be further increased increasing the length of M3

These measures are very cheap in terms of silicon real estate. The circuit may for instance finally look like this without getting much more spread:

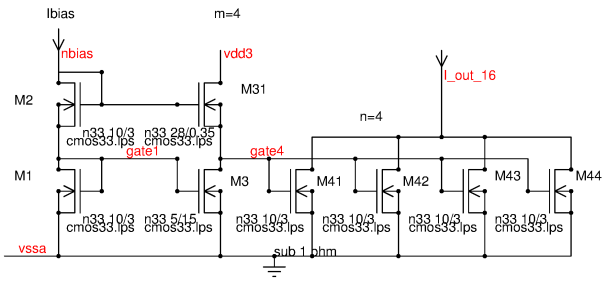


Figure 7.16: Area optimized version of the m*n current mirror

The circuit is almost identical with the previous circuit except that now M31 is compact and short while M3 is made longer and more narrow. The circuit achieves a current ratio of 14.98 (simulated with ngspice at Ibias=100μA), which is already very close to the theoretical limit of 4*n=16.

Of course making M31 long and M3 wide and short we also can build a current reducing mirror. This is (theoretically) valid down to m=0.25. At m=0.25 the effective gate voltage V(gate1) becomes 0 and we are definitely in a range where the equation isn't valid anymore (M41..M44 than are in weak inversion. The equation only is valid for strong inversion!). Values on m<0.25 are meaningless (deep weak inversion). The closer we get to m=0.25 the more sensitive the current mirror becomes to even the smallest production spread and mismatch of M3 and M31. Therefore using this kind of mirror for current reduction can not be recommended.

7.2.3 High voltage current mirrors

As soon as higher voltages than some volt are required the low voltage current mirrors must be protected from the high voltage using cascodes. Of course current mirrors theoretically can be built using high voltage transistors alone. But in most technologies the channel length of the high voltage transistors strongly depends on outdiffusion of drain or source implants. This makes the high voltage transistors fairly inaccurate. Using high voltage transistors as current mirrors usually leads to about +-10% spread more or less independent of the gate area involved. Therefore wherever possible using cascoded low voltage mirrors is preferred.

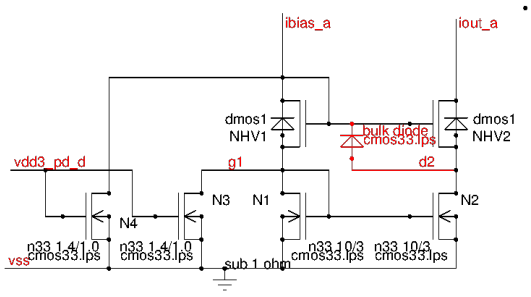


Figure 7.17: Current mirror using low voltage transistors combined with a high voltage cascode

The current mirror shown relies on the good matching of the low voltage transistors N1 and N2. Cascode NHV2 serves as an overvoltage protection for N2. So the net iout_a may be exposed to much higher voltages than the low voltage net d2. The gate voltage of N1 and N2 is defined by the properties of N1 and the current applied to net ibias_a. The voltage at net d2 follows the voltage of net g1 as long as the currents through NHV1 and NHV2 track. The voltage at net ibias_a depends on the properties of N1 and NHV1. In most technologies the voltage at nets g1 and d2 is about 600mV to 2V. The voltage at net ibias_a usually is chosen in the range 1.5V to 3.3V.

One problem arises if N3 and N4 are turned on in power down mode. N2 and NHV2 turn off except for leakage currents. Usually NHV2 is significantly bigger then N2. So the leakage current of NHV2 is higher than the leakage current of N2. This makes net d2 float up in stand by modes. Without the red protection diode the drain voltage of N2 would increase until the drain-bulk junction of N2 breaks or the gate oxide of NHV2 breaks. The red protection diode prevents the break down conducting the leakage current of NHV2 to net ibias_a and the drain of pull down transistor N4. Since leakages are in the range of a few nA a simple antenna diode is enough to protect net d2.

The protection diode does not necessarily connect to net ibias_a. It can just as well connect to any other net (that has a path to ground during stand by) with a voltage low enough to prevent a break down of N2 or the gate of NHV2.

Alternatively the net d2 can be pulled down by a third pull down transistor. This offers a cheap solution that does not require an nwell at net ibias_a.

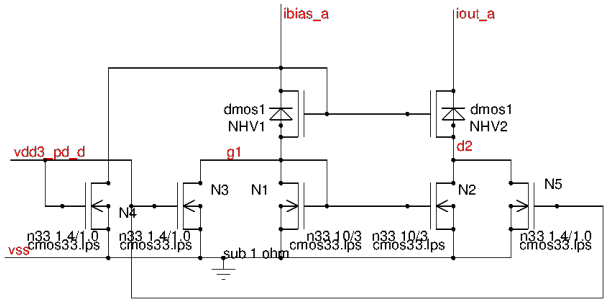


Figure 7.18: Alternative leakage protection using a drain pull down N5

There is one remarkable difference between the two protection implementation:

In the circuit using a diode as a leakage protection we also get a little bit of ESD protection for free. In case of an ESD event at the drain of NHV2 the source of NHV2 will be pulled up by the drain-bulk capacity of NHV2. The diode will turn on N2 and up to a certain energy level the circuit is self protecting.

The second implementation using N5 as a leakage protection does not self protect the circuit. Capacitive feed through of a pulse applied at iout_a will pull up the drains of N2 and N5 until one of the low voltage transistors breaks down. If N2 breaks but does not get destroyed thermally such ESD events can lead to a drift of the threshold of N2 (hot carrier stress). So the parameters of the current mirror will change. Usually such events increase the threshold of N2 leading to a mirror ratio of less than 1 after the pulse. In many cases hot carrier stress of N2 can be annealed by high temperature storage. (if you suspect hot carrier stress of N2 try storing the chip at 200°C for 24h. If the current mirror returns to the original ratio this is a strong indication of hot carrier stress in N2)

7.2.4 MOS current mirrors with gm degradation

The transconductance of MOS transistors can be altered changing the length of the device. The longer the channel the lower the transconductance and the higher the gain overdrive. Looks like a good concept but in fact using resistors may be more ecconomical!

Some technologies offer very well matching resistors with matching values of less than $5\%\mu m$while at the same time thin gate oxides below 15nm still lead to high transconductance even if the W/L is below 1/5.

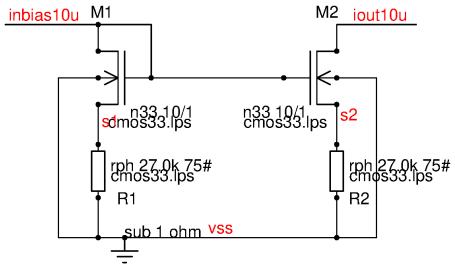


Figure 7.19: NMOS mirror with resistive gm degradation

Calculation of the errors: Using a 3.3V technology with 7nm gate oxide a typical offset error of the NMOS transistors is:

$$Vos = 7mV\mu m/\sqrt{W * L} = 2.2mV$$

Operating at 10uA the voltage drop over R1 and R2 is in the range of 270mV. This leads to an error contribution of the NMOS transistor offset of

$$Err_{nmos} = 2.2mV/270mV = 0.82\%$$

Matching constants for resistors depend on the process used. LBC9 offers very good values in th range of $2\%\mu m$at a resistor width of $0.35\mu m$. (other processes such as BCD9 are more in the range of $8\%\mu m$. So the process really matters!). Assuming 75# the length becomes $26.25\mu m$. Still an affordable size. The resulting resistor matching becomes:

$$Err_{res} = 2\%\mu m/\sqrt{0.35\mu m * 26.25\mu m} = 0.65983\%$$

The total one sigma error calculates as

$$Err = \sqrt{Err_{nmos}^2 + Err_{res}^2} = 1.0525\%$$

Trying to build the same performance with a gate overdrive of 270mV using long channel NMOS transistors we would end up with about $W = 17\mu m$,$L = 16\mu m$.

In most cases the sweet spot for this kind of gm degradation is in the range of 300mV to 400mV drop over the resistors.

7.3 Differential amplifier

The differential amplifier stage is the core of almost all analog circuits. No matter if we build it using bipolar transistors or MOS transistors it is always the same base topology.

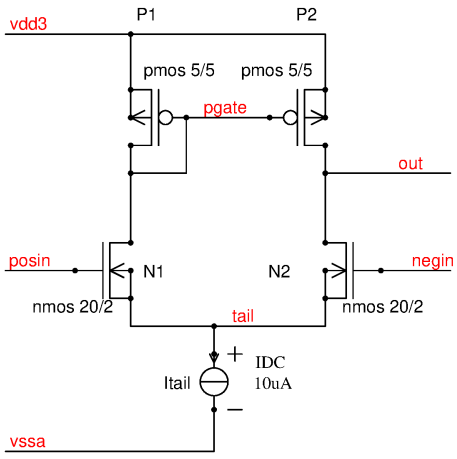


Figure 7.20: Generic differential amplifier with current mirror

The transistors N1 and N2 "measure" the differential voltage between nodes posin and negin. For a small signal analysis let's assume both inputs are at the same voltage. Each of the transistors will carry the same current. This is the ideal equilibrium operating point.

$$I_{N1} = I_{N2} = I_{tail}/2 \tag{7.29}$$

The same applies to the current mirror P1 and P2. The output current becomes:

$$I_{out} = I_{P2} - I_{N2} = 0 \tag{7.30}$$

But hey, This isn't an voltage amplifier! This is an OTA (operational transconductance amplifier)! It only turns into a voltage amplifier if we know the impedance at note out. We will come back to that point later.

In the following let us assume the current mirror P1 and P2 is perfect. No matter what happens $I_{P1} = I_{P2}$. This is a simplification that can be justified as long as we are in the small signal range changing the drain voltage of P2 just some 100mV and always staying outside of the triode region.

For small signal calculation let us further assume transistors N1 and N2 behave linear. $gm_{N1} = gm_{N2} = gm$. This is more or less true as long as the transistors match and as long as the gate voltages only differ significantly less than Vt (26mV at 300K). In other words:

$$|V_{posin} - V_{negin}| < 26mV$$

To maintain a constant tail current an increase of one input voltage must be compensated by a decrease of the other input. I we apply a differential voltage between the inputs we will observe the following effect:

$$dV = V_{posin} - V_{negin} \tag{7.31}$$

The voltages at the input nodes respond symmetrically:

$$V_{posinnew} = V_{posinold} + dV/2 \tag{7.32}$$

$$V_{neginnew} = V_{neginold} - dV/2 \tag{7.33}$$

The current flowing through N1 changes by:

$$dI_{N1} = \frac{gm * dV}{2} \tag{7.34}$$

At the same time the current through N2 changes as well:

$$dI_{N2} = -\frac{gm * dV}{2} \tag{7.35}$$

Since the current through N1 gets mirrored to the output by P1 and P2 the output current changes by:

$$dI_{out} = dI_{N1} - dI_{N2} = gm * dV \tag{7.36}$$

Operating in strong inversion: Nice, but what is the value of gm? We can look it up at section 4.6.1 NMOS transistor. There we find for strong inversion:

$$I_d = k' * \frac{W}{L} * V_{gseff}^2$$

with parameter k'

$$k' = \frac{\mu * \varepsilon_{sio_2}}{2 * n * t_{ox}}$$

Creating the derivative yields

$$\frac{dI_d}{dV_{gseff}} = k' * V_{gseff} * \frac{W}{L} \tag{7.37}$$

with $I_d = I_{tail/2}$we get

$$V_{gseff} = \sqrt{\frac{I_{tail} * L}{2 * k' * W}} \tag{7.38}$$

combining it all this leads to

$$gm_{si} = \sqrt{\frac{k' * I_{tail} * W}{2 * L}} \tag{7.39}$$

To see all possible influences let's expand k'.

$$gm_{si} = \sqrt{\frac{\mu * \varepsilon_{sio_2} * I_{tail} * W}{4 * n * t_{ox} * L}} \tag{7.40}$$

Now we have full visibility of technology parameters and design parameters on the transconductance of the little amplifier stage. To increase the gain there are several possible measure. The increase of gain follows a square root function of:

- carrier mobility
- dielectric constant of the gate oxide
- tail current
- aspect ratio W/L
- inverse gate oxide thickness $1/t_{ox}$

Operating in weak inversion: BUT: this equation applies to strong inversion. As soon as the gate voltage drops below about 3*$V_t > V_{gseff}$ we are entering weak inversion. The drain current starts to follow diffusion laws. This changes the equations to:

$$gm_{wi} = \frac{I_{tail}}{2 * n * k * T/q} \tag{7.41}$$

I weak inversion the transconductance becomes independent of most of the technology parameters! The factor $k * T/q$is called the thermal voltage. It is a function of the Boltzmann constant k, the absolute temperature T and the electron charge q. There is nothing we can do about it except changing the temperature. For simplicity at room temperature we can use:

$$V_{t_{300K}} = \frac{k * T}{q} \approx 26mV$$

Replacing N1 and N2 by bipolar transistors: So what changes if we replace N1 and N2 by bipolar transistors? A bipolar transistor can be regarded as a MOS transistor with gate oxide thickness 0. It is "always in weak inversion", this means it is completely controlled by diffusion. The coupling factor n becomes 1 and the equation of the transconductance further simplifies:

$$gm_{bipolar} = \frac{I_{tail}}{2 * k * T/q} \tag{7.42}$$

Achievable voltage gain: The achievable voltage gain depends on the impedance at node out. If we have a known impedance Z the voltage gain simply becomes

$$v = Z * gm \tag{7.43}$$

The interesting point here is, that Z can be complex. If the load of the differential amplifier stage consists of a network of resistors, inductors and capacitors this can be expressed in Z.

In most amplifiers Z is determined by the input capacity of the next stage (the gate capacity of the next transistors and/or the frequency compensation capacity) and the early voltages of the shortest transistors connected. The DC impedance simply becomes

$$Z(f = 0) = \frac{2 * V_{early}}{I_{tail}} \tag{7.44}$$

As a rule of thumb we can assume an early voltage of N2 of about $10V/\mu m$channel length. (as long as the transistors used don't have any halo implant. Transistors with halo implant have much lower early voltages!). This way we can estimate the achievable DC voltage gain for the different cases.

Strong inversion DC voltage gain: Strong inversion DC gain can be calculated multiplying the impedance with gm.

$$v_{si}(f = 0) \approx \frac{L_{N2} * 20V}{\mu m} * \sqrt{\frac{\mu * \varepsilon_{sio_2} * W}{n * t_{ox} * L * I_{tail}}} \tag{7.45}$$

A shocking result! Increasing the tail current reduces the DC voltage gain. High tail currents only are beneficial to increase the bandwidth of an amplifier.

Operation at high current densities far in the range of strong inversion is mainly used for RF amplifiers because here gm is more important than the achievable DC gain. The gain bandwidth product operating with a capacitive load calculates as.

$$GBW = \frac{gm}{2 * \pi * C_{out}} \tag{7.46}$$

In MOS amplifiers the GBW follows the square root of the tail current.

Using LC tanks as a load operation at even higher frequency is possible, but in this case we only have gain in a narrow frequency range (at resonance).

Weak inversion DC voltage gain: In weak inversion the DC voltage gain roughly becomes

$$v_{wi}(f = 0) \approx \frac{L_{N2} * 20V}{\mu m} * \frac{1}{n * k * T/q} = \frac{L_{N2} * 20V}{\mu m} * \frac{1}{n * V_t} \tag{7.47}$$

This is the highest achievable DC voltage gain for MOS amplifiers. For this reason usually an operation point exactly at the transition between weak inversion and strong inversion is chosen for low frequency amplifiers. This operating point offers almost the weak inversion voltage gain and still supports an acceptable gain bandwidth product.

Bipolar amplifier voltage gain: Bipolar amplifiers behave similar to weak inversion operation of MOS transistors. The achievable voltage gain becomes

$$v_{bipolar}(f=0) \approx \frac{V_{early}}{k*T/q} = \frac{V_{early}}{V_t} \tag{7.48}$$

In this equation V_{early} is the lower one of the early voltages of the current mirror and the transistors used for the differential amplifier stage.

Bipolar amplifier stages keep their exponential characteristic even at high bias currents. The GBW is proportional to the tail current. (at least as long as the emitter spread resistance doesn't become the limiting factor). This still makes bipolar technologies an attractive choice in the GHz range.

Thermal behavior: With increasing temperature the thermal voltage $V_t$increases and the carrier mobility μ decreases. Without special counter measures the DC voltage gain of the differential amplifier decreases with temperature. One possible counter measure working well for weak inversion and for bipolar differential amplifiers is to increase the tail current proportional to the junction temperature.

Improved stages: All these equations apply to the simple input stage without cascodes. Adding cascodes the DC impedance at node out can be increased significantly. This requires a higher supply voltage to add room for the cascodes. The following figure shows the concept. Increasing the DC voltage gain by up to 2 magnitudes is possible with cascodes.

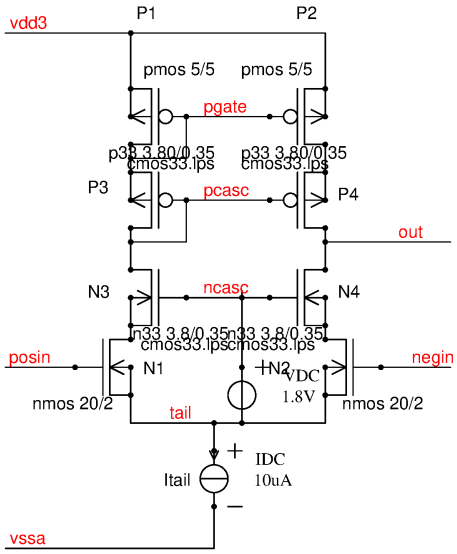


Figure 7.21: Same amplifier as before but with additional cascodes to increase the DC impedance and the DC voltage gain at node out.

Amplifier white noise (thermal noise): To get a rough idea about the white noise performance of the amplifier we can approximate the simple differential amplifier by the following equivalent circuit.

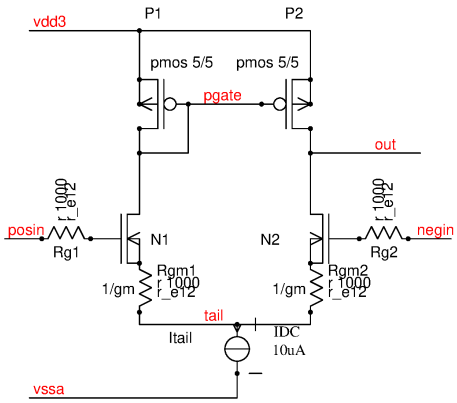


Figure 7.22: Noise equivalent circuit for white noise

In this equivalent circuit N1 and N2 are assumed to be "perfect" transistors with infinite gm and without noise. The transconductance of the whole stage is determined by the resistors Rgm1 and Rgm2. At the same time the

two resistors are the source of white noise generated by the transconductance. The noise generated by each of the resistors is described in section 4.3.2.

$$V_{ngm} = v_{ngm} * \sqrt{BW} = 2 * \sqrt{gm^{-1} * k * T * BW} \tag{7.49}$$

The resistance of the gate poly silicon also contributes white noise.

$$V_{nRg} = 2 * \sqrt{R_g * k * T * BW} \tag{7.50}$$

Since all 4 noise sources add in an uncorrelated way the total differential noise voltage becomes

$$V_n^2 = \sum V_n^2 \tag{7.51}$$

$$V_n = 2 * \sqrt{k * T * BW * (R_{g1} + R_{g2} + 2/gm)} \tag{7.52}$$

Well, not quite complete! To be exact we also have to replace P1 and P2 by noise equivalent circuits. But assuming a reasonable choice of transistor geometries the contribution of P1 and P2 to the total noise should be negligible. (less than 10%). So for most applications the simplification shown is good enough.

As a rule of thumb for a resistance of $1k\Omega$ at 300K we can expect a noise density of $v_{n300K1k\Omega} = 4nV/\sqrt{Hz}$. Let's calculate a simple example:

BW=1MHz, gm=0.1mA/V, Rg1=Rg2=1k yields

$$V_n = (4nV/\sqrt{Hz}) * \sqrt{10^6 Hz} * \sqrt{1 + 1 + 1/0.1 + 1/0.1} = 4nV * 10^3 * \sqrt{22} = 18.76\mu V$$

This is an estimate of the white noise at a bandwidth of 1MHz. It does not yet include the $1/f$ noise.

What we can learn from the above equations:

- To reduce the noise make gm as high as possible (or as high as affordable in terms of current consumption)
- Keep the gate resistance low

1/f noise of an amplifier $1/f$ noise (sometimes called flicker noise) depends on various process parameters. In most cases it is modeled as a frequency dependent noise density

$$v_f = \frac{K_f}{c_{ox} * W * L * f} \tag{7.53}$$

The constant $K_f$is process dependent. Usually it is determined empirically. $1/f$ noise usually only contributes significantly to the total noise up to some kHz, sometimes some hundred kHz. $1/f$ noise can be reduced making the gate area as big as possibles - but that makes the amplifier slow.

A good way to get rid of $1/f$ noise is building a chopped amplifier. But that is going a bit too far to describe it right here.

Large signal considerations For large signal applications the following parameter have to be taken into account:

1. The maximum input differential voltage is limited by the break down voltage of the gates of N1 and N2
2. The maximum output voltage swing is limited by the supply voltage
3. Signal compression starts when the transistors (current mirrors as well as differential stage) reach triode region. (Usually at peak to peak voltage $V_{pp} < vdd3 - 1.5V$)
4. Common mode range V_{cm} is

   $$V_{itail} + V_{gsN1N2} < V_{cm} < vdd3 - V_{gsP1P2} - V_{gsP3P4} + V_{gsN1N2}$$

   V_{itail} is the minimum operating voltage the tail current sink needs.
5. maximum possible output current is $\pm I_{tail}$
6. Slew rate is $dV_{out}/dt = I_{tail}/C_{out}$

Large signal considerations For large signal applications the following parameter have to be taken into account:

1. The maximum input differential voltage is limited by the break down voltage of the gates of N1 and N2
2. The maximum output voltage swing is limited by the supply voltage
3. Signal compression starts when the transistors (current mirrors as well as differential stage) reach triode region. (Usually at peak to peak voltage $V_{pp} < vdd3 - 1.5V$)
4. Common mode range V_{cm} is

$$V_{itail} + V_{gsN1N2} < V_{cm} < vdd3 - V_{gsP1P2} - V_{gsP3P4} + V_{gsN1N2}$$

   V_{itail} is the minimum operating voltage the tail current sink needs.
5. maximum possible output current is $\pm I_{tail}$
6. Slew rate is $dV_{out}/dt = I_{tail}/C_{out}$

7.3.1 Input protection circuits

Input stages of differential amplifier usually are built using low voltage components because in most technologies the low voltage components match best and offer the best speed/current ratio. The inputs of amplifiers may however be exposed to many sources of stess such as:

- ESD pulses.
- charging during plasma etching.
- RF injection.
- Transients during operation.
- DC overvoltage due to unintentional short circuit of the input to a high voltage rail.

There are two basic concepts how to protect the input stage. Either it can be an external protection using clamps on board level or it can be on chip protections. All protections have in common that they limit the voltage swing the differential pair is exposed to.

Antenna stress protection: During plasma etching of the metal traces the traces can be charged up to some 10V. In old CMOS processes with gate oxides in the range of 20nm or more this was not much of a problem. With gate oxides below 10nm the situation is so critical that the gate oxides can completely break. Even far below the level where the gates break there already is a risk of charge accumulation in the oxide between the gate and the channel. The amplifier will have a significantly higher offset than anticipated. Typically this offset created by antenna stress can partly be annealed storing the devices at high temperature for some hours.

To build well matching CMOS input stages the use of antenna diodes is even more important than inside the logic!

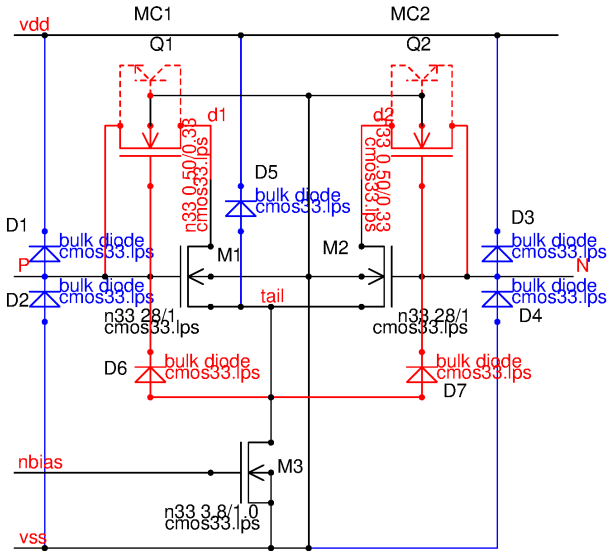


Figure 7.23: Antenna protections of a differential amplifier input

In the above figure two levels of antenna protections are shown. The blue colored diodes show the standard level of potection. D1..D4 keep the gate voltages of M1 and M2 between -Vbe and vdd+Vbe. Additionally D5 prevents the tail from floating above vdd.

If the matching requirements are very critical adding the red components is suggested. D6 and D7 limit the gate minimum voltage at V(tail)-Vbe. MC1 and MC2 additionally limit the gate voltages to V(dx)+Vth. Since M1 and M2 turn on if the gates get charged positive, the drain voltage and the tail voltage will be almost equal. So while the chip is unsupplied MC1 and MC2 will limit the positive gate voltage to V(tail)+Vth. MC1 and MC2 are very area efficient because they can be placed in the same well as the transistor they have to protect. As long as the bulks of MC1 and MC2 are connected to vss diodes D2 and D4 are already provided by the bulk diodes of MC1 and MC2.

One side effect of the protection transistors MC1 and MC2 is that we are adding two lateral NPN transistors Q1 and Q2 drawn in dotted lines. If the input voltage gest negative wersus the bulks of MC1 and MC2 these lateral bipolar transistors will reverse the signal of the differential stage. This will become especially tricky if the bulks of MC1 and MC2 are connected to node tail in stead of vss.

All the red and blue traces must be routed in metal 1! (Otherwise they are open during metal 1 etch and the protection doesn't work.)

Note that antenna protection diodes are very small, fast and perfect RF rectifiers! Keep RF away from differential amplifier inputs with antenna diodes.

Substrate PNP input stage: Substrate PNP transistors usually are quite robust. The base is a low doped epi region while the emitter and the collector are high doped P-regions. The break down voltages Vbebr and Vbcbr are defined by the epi doping. Usually this is the highest break down voltage a process offers.

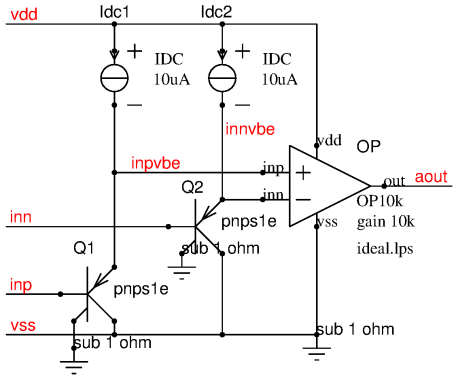


Figure 7.24: PNP input protection

In the above figure OP is a low voltage operational amplifier. High positive voltage at the base of the PNP transistors are isolated from the sensitive input. Negative voltages will be clamped to -Vbe at nodes inp and inn. The voltage range at nets innvbe and inpvbe is limited to about 0V to vdd.

If the gates of the input pair of OP are connected to the emitters of the PNP transistors by metal 1 the PNP transistors additionally serve as antenna protections during plasma etching of the metal.

One big drawback of the PNP input stage is the rectification of RF. If RF is applied at one of the inputs the PNP transistor emitter will follow the falling edges very fast. This even applies close to or above the transit frequency of the transistor (RF will still be rectified at the emitter even if the gain is below 1!) The rising edges simply turn off the PNP transistor and the rising slope at the emitter is determined by the pull up current and the input capacity of the amplifier differential stage. In other words the PNP transistors are negative peak rectifiers! (For more details about RF injection through parasitic capacitors and RF rectification please see the chapter I/O cells.)

A second draw back is the base current of the bipolar transistor. This kind of input stage is only of limited use in switched capacitor circuits.

The PNP input stage contributes an additional offset. Typical 1s spread of matched PNP transistors is in the range of 0.4mV for most processes commonly used.

$$V_{os}^2 = V_{osOP}^2 + V_{osPNP}^2 + (V_T * ln(I_1/I_2))^2 \tag{7.54}$$

The current generators contribute with logarithme of the ratio of the currents. Ideally I1 and I2 are equal. Usually the ratio of the current is in the range of 0.9..1.1 or better leading to ln(I1/I2) in the range -0.1 to +0.1 or less.

Cascode input protection: Cascodes using high voltage transistors can be used to protect the input of a low voltage differential amplifier too. If a HVPMOS transistor and a HVNMOS transistor are combined even negative voltages can be applied without destroying the circuit.

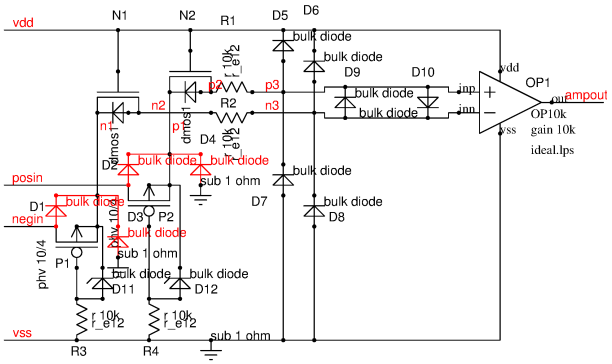


Figure 7.25: Input protection with HVMOS transistors

In this figure the amplifier OP1 is assumed to be low voltage CMOS amplifier with antenna diodes D5 to D10. Resistors R1 and R2 limit the current that can flow into protection diodes D9 and D10.

P1 and P2 limit the voltage at nets p1 and n1 so that these nets can never become negative. Since P1 and P2 have parasitic bulk diodes D1 and D2 nets p1 and n1 can still reach high positive voltages. Therefore the gates of P1 and P2 may not be hard wired to ground. They must be protected by zener diodes D11 and D12 and current limiting resistors R3 and R4.

Transistors N1 and N2 limit the positive swing of nets p2 and n2 to one threshold below vdd.

The common mode range of this amplifier is restricted to:

$$V_{thpmos} < V_{cm} < vdd - V_{thnmos}$$

Inside the common mode range the transfer characteristic of the common mode voltage is linear. The differential voltage at nets p3 and n3 is limited by diodes D9 and D10. In the middle between vdd and vss no rectification is expected. (Signals exceeding the range vss to vdd however still suffer from rectification effects.) This makes the circuit much more RF robust than the PNP input stage. Furthermore the protection doesn't add additional offsets to the amplifier.

Resistive input protection: Amplifiers that are exposed to excessive RF injection must be protected with linear devices because every active device might include a non linear characteristic that can act as an RF rectifier. Here the only solution is a resistor attenuator that reduces the level of disturbance to values that are non destructive to the amplifier input.

The drawback of such a resistor attenuator is the reduction of the signal. Offsets of the amplifier will be multiplied by the attenuation factor of the protection.

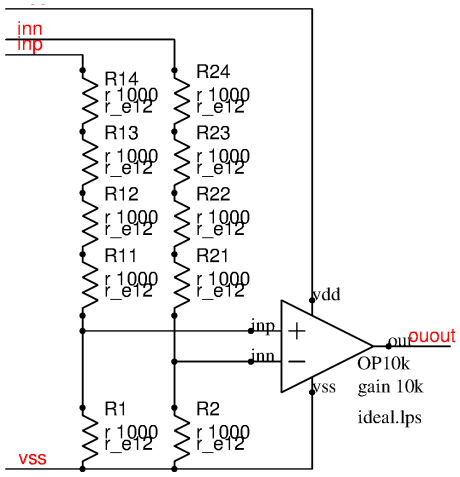


Figure 7.26: Resistor input protection

The layout of the attenuator must be done with extereme care not only perfectly matching the resistors but also the parasitic capacities. A layout scheme such as R1-R2-R1-R2 already holds a capacitive asymmetry the gets more noticable the more resistor segments are used! Better use a fully symmetrical scheme such as R1-R1-R2-R2 for stray capacitor symmetry.

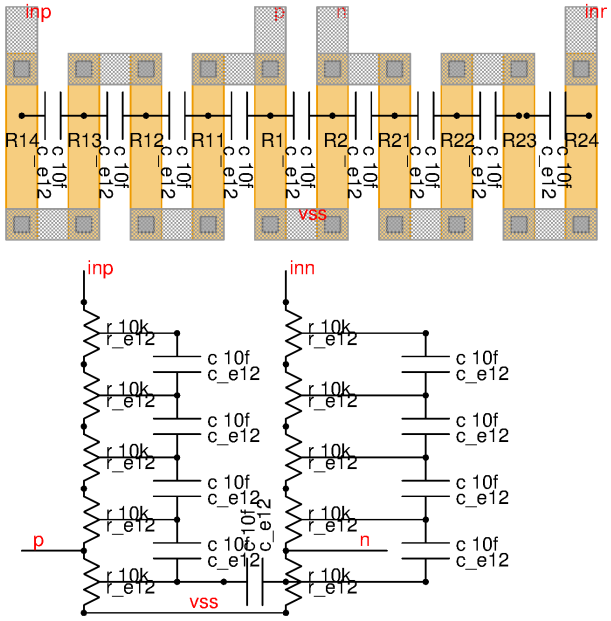


Figure 7.27: Layout with good capacitive matching

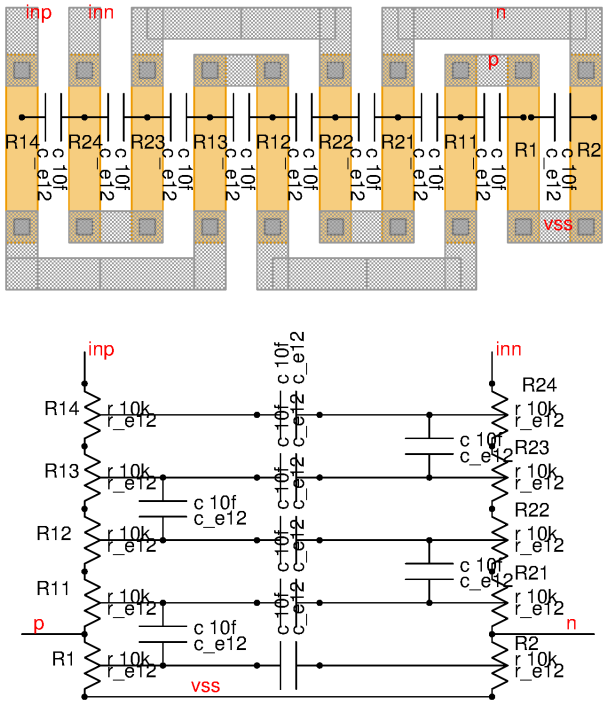


Figure 7.28: The same circuit but different layout breaking the capacitive symmetry.

This kind of layout damages common mode rejection at high frequency. The capacitive neigborhood of R1 and R2 as well as R14 and R24 differs! Measurements on CAN transceivers showed common mode rejections of "only" 50dB at 30MHz due to such capacitive asymmetries leading to RF sensitivity at 10..30MHz that was not visible in the original simulation that neglected the lateral capacitive coupling of resistors.

Dummies next to one branche but not present next to the other branche can make capacitive mismatch even worse!

Norton amplifier inputs: A norton amplifier is an elegant variation of the voltage divider input. It substitutes the resistors R1 and R2 by a current mirror. The output of the norton amplifier is already single ended. Ideally the output transistor should operate with a current density similar to the current mirror. As long as speed is not the main design target the current mirror and the output transistor will operate in weak inversion or in the transition zone between weak inversion and strong inversion.

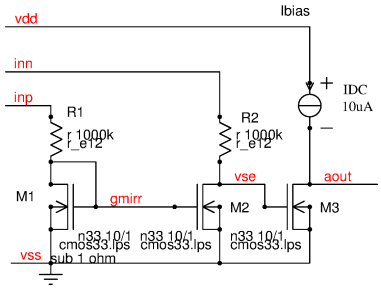


Figure 7.29: Norton input stage

Assuming a perfect mirror M1, M2 and equal resistors R1 and R2 the voltage at vse becomes.

$$V(vse) = V(inn) - (V(inp) - Vgs_{M1})$$

$$V(vse) = Vgs_{M1} + (V(inn) - V(inp))$$

As long as M3 operates at the same current density as M1 and M2 the gate overdrive of M3 simply is

$$Vgodrv_{M3} = V(inn) - V(inp)$$

Assuming operation close to weak inversion the offset mainly depends on the ratio of the current densities of M1 and M3.

$$V_{os} = V_T * k * ln(\frac{(V(inp) - Vth)/R1}{I_{DC}}) \tag{7.55}$$

If this systematic and common mode dependent offset can't be accepted the norton amplifier can be improved operating M3 at the same current density as M1 and M2.

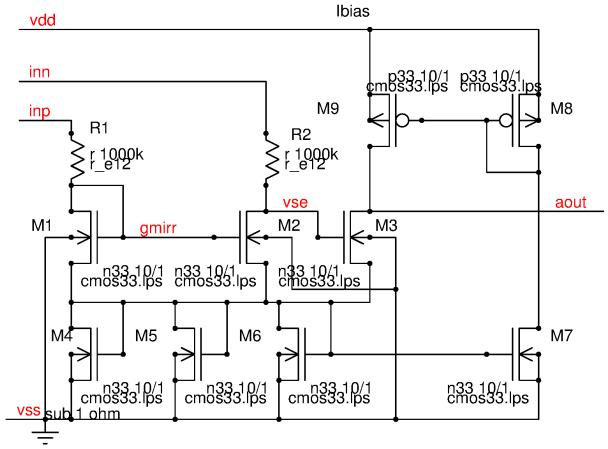


Figure 7.30: Norton amplifier with tracking bias current of the output transistor to reduce the systematic input offset

The voltage gain of both variants of the norton amplifier is

$$gain = gm_{M3} * Z_{out} \tag{7.56}$$

Z_{out} is the impedance of the node aout. As long as there is no resistive load the DC gain is limited by the early voltages of the current source M9 and transistor M3 and the current flowing through M3, M9.

Noise of Norton amplifiers: As soon as the input signal is attenuated by resistor dividers or by the resistors of a norton amplifier the noise of the resistors become the dominant noise source. For low noise applications norton amplifiers can't be recommended.

7.4 Bandgap circuits

There are hundreds of possible impementations of bandgaps. But they all follow the same concept. A bandgap reference adds the forward voltage of a bipolar diode and the temperature voltage created by the difference of two forward voltages operated at different current densities. The forward voltage of a bipolar diode (operated at a constant current) has a negative temperature coefficient of about -2mV/K. The difference of the forward voltages

depends on the absulte temperature, the Bolzmann constant, the electron charge and the logarithme of the current density ratio.

Ideally at 0K the forward voltage of a bipolar diode exactly corresponds the bandgap energy of the semiconductor while the temperature voltage (the difference of two forward voltages) becomes zero (in other words: as long as the voltage is below the bandgap voltage all carriers are in the valence band. As soon as we exceed the bandgap voltage all carriers jump into the conductive band at 0K. The ideal diode is infinitely conductive at 0K).

Note: The theoretical bandgap voltage of silicon at 0K is 1.11V. Building bandgaps operating at room temperature however we usually end up with a bandgap voltage of about 1.16V to 1.25V. 1,23V usually is a good initial guess for most technologies.

$$V_t = \frac{k * T}{e} \tag{7.57}$$

$$k = 8.617332478 \times 10^{-5} eV/K$$

$$e = 1.60217656535x10^{-19} C$$

At 300K we get:

$$V_{t300k} = 25.852mV$$

Operating two diodes at different current densities we get:

$$\Delta V_{be} = V_t * ln(\frac{i_1}{i_2}) \tag{7.58}$$

The temperature coefficient becomes:

$$\frac{dV_t}{dT} = \frac{k}{e} = 86.1733\mu V/K$$

$$\frac{d\Delta V_{be}}{dT} = ln(\frac{i1}{i2}) * \frac{k}{e} \tag{7.59}$$

7.4.1 The Widlar bandgap

To copensate the -2mV/K of a diode we have to multipy this temperature coefficient to reach +2mV/K. The most simple circuit doing this job is the Widlar bandgap.

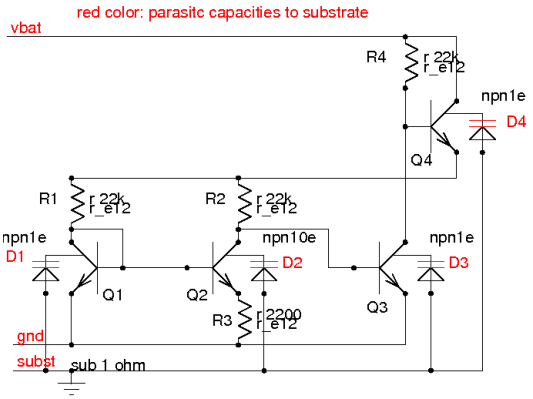


Figure 7.31: Widlar Bandgap

Ideally Q3 is operated at exactly the same current density as Q1. (Later we will have a look at errors caused by violating this condition.) So the voltage at the base of Q1 and the base of Q3 are equal and the currents trough E1 and R2 are equal. Since Q2 has more emitters than Q1 it operates at a different current density. (in out case the current density is 10 times less.) So at R3 we will find a voltage of:

$$V_{R3} = V_T * ln(n) \tag{7.60}$$

$$n = 10$$

$$V_{R3} = 25.852 * ln(10) = 59.526mV$$

at 300K. The coltage drop at R1 and R2 becomes:

$$V_{R1} = V_{R2} = 10 * V_{R3} = 595mV$$

The bandgap voltage becomes:

$$V_{bg} = V_{be} + V_{R2} \tag{7.61}$$

Ideally this should exactly be the bandgap voltage of the chosen material (1.23V in case of silicon). Note that the base emitter voltage of the above equation is part of the regulation loop. It is the base emitter voltage of Q3!

Practical considerations of the simple Widlar bandgap

Most important systematic errors Up to now we followed the assumption that the current through R4 is about equal to the current trough R1 and R2. If the supply voltage changes this assumption is no more true. With increasing supply voltage Q3 has to carry an increasing current. So the base emitter voltage of Q3 increases as well. The current flowing in R1 and R2 can be calculated (neglecting base currents) as:

$$I_{Q2} = \frac{V_T * ln(n)}{R_3} \tag{7.62}$$

While the current through Q3 is:

$$I_{Q3} = \frac{V_{bat} - V_{bg} - V_{beQ4}}{R_4} \tag{7.63}$$

Thus the difference of the base emitter voltages of Q3 and Q1 becomes:

$$V_{errR4} = V_T * ln(\frac{R_4 * V_T * ln(n)}{R_3 * (V_{bat} - V_{bg} - V_{beQ4})} \tag{7.64}$$

and the deviating bandgap voltage calculates as:

$$V_{bg} = V_{bgideal} + V_{errR4} \tag{7.65}$$

Even worse the error is supply voltage dependent. A simple octave script will calculate the error versus Vbat.

```
vbat=[1.4:0.1:12]
R1=22.000
R2=22.000
R3=2.200
R4=22.000
n=10
vt=25.8
vbe=0.65
vbg=1.23
vr3=vt*0.001*log(n)
ir3=vr3/R3
vr4=vbat-vbe-vbe
ir4=vr4/R4
verr=vt*log(ir4/ir3)
plot(vbat, verr)
```

The resulting plot shows the error voltage im mV versus the supply voltage vbat in V.

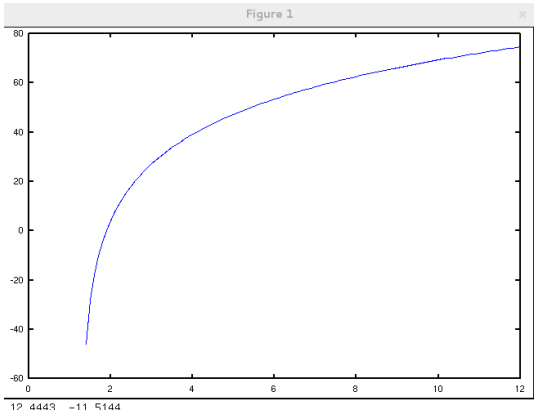


Figure 7.32: Error of the simple Widlar bandgap with pure resistor bias.

Most important statistical errors The production spread of the bandgap depends on the following factors:

1. Matching of the transistors Q1 and Q2
2. Matching of transistors Q1 and Q3
3. Matching of resistors R3 versus R1 and R2

Mismatch of the base emitter voltage (at the same current density) of Q1 and Q2 will lead to a deviation of the current flowing in R2 and R3. So the error contribution of an offset to the bandgap voltage will be:

$$\Delta V_{bg1} = \Delta V_{be12} * \frac{R_2}{R_3} \tag{7.66}$$

An offset of Q3 will simply be added to the bandgap voltage:

$$\Delta V_{bg2} = \Delta V_{be13} \tag{7.67}$$

The offset of resistors typically is a relave error given in %. This error usually becomes big for small resistors (having a small silicon area). In most cases it is sufficient to obseve the error of R3 while R1 and R2 occupy much bigger areas and their deviation can be neglected.

$$\Delta V_{bg3} = \frac{\Delta R_3}{R_3} * V_T * ln(n) * \frac{R_2}{R_3} \tag{7.68}$$

Since these are statistical errors that should be independent from each other we have to add the power:

$$\Delta V_{bg} = \sqrt{\Delta V_{bg1}^2 + \Delta V_{bg2}^2 + \Delta V_{bg3}^2} \tag{7.69}$$

EMC performance of the Widlar bandgap The Widlar bandgap is fairly robust against transients on the supply vbat. The only way to couple transients into the bandgap voltage is the miller capacity of Q4. Since the bandgap must have a frequency compensation anyway (for stability of the regulation loop) it is a good idea to use a parallel compensation (simply a capacitor from the base of Q4 to ground). On the other hand a miller compensation using a capacitor between the collector of Q3 and the base of Q3 is cheaper (but is less performant with regards to transients on Vbat).

The big drawback of the Widlar bandgap is its sensitivity to substrate noise. Since Q2 is the biggest active component it also has the biggest substrate capacity (collector of Q2 to substrate). Substrate noise picked up by the substrate capacity of Q2 will be peak rectified by the strongly non linear characteristic of Q3. So substrate noise will turn on Q3 and turn off the bandgap! The substrate noise sensitivity can be improved if the emitters of Q1 and Q3 and resistor R3 are tied to substrate rather than to metallic circuit ground. Modern CAD tools however will flag this design style as soft connect errors.

If other components in the neighbourhood of the bandgap inject electrons into the substrate the big area of Q2 will pick up these electrons. (It becomes the collector of a parasitic lateral NPN transistor. Minority carrier injection into the substrate close to Q2 will turn off Q3 and the bandgap voltage increases in an uncontrolled way!

Bottom line we must state that the Widlar bandgap can not be recommended for applications with high RF or electron injection into the substrate.

7.4.2 The Brokaw Bandgap

The Brokaw bandgap [13]solves the problem of matching the currents of Q3 of the Widlar bandgap with the currents inside the bandgap. This kind of bandgap became very attractive with the introduction of lateral PNP transistors into semiconductor design libraries end of the 1960s. (Brokaw published 1974). The most simple way of building a Brokaw bandgap is shown below:

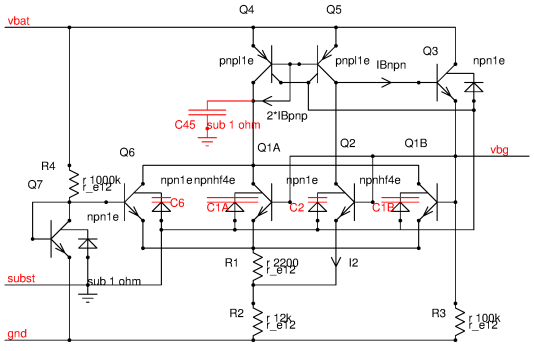


Figure 7.33: The most simple version of the Brokaw bandgap

The bandgap published by Brokaw was a bit more complex solving errors caused by the base currents of the transistors. This simplified version however serves well to explain the concept. If current mirror Q4, Q5 and ouput stage Q3 is designed using MOS transistors this will solve most of the systematic errors too.

Different from the Widlar bandgap the Brokaw bandgap needs a starter (R4, Q7, Q6). Once the bandgap is running the emitter of Q6 is lifted to about 600mV and starter Q6 turns off.

R2 now carries twice the bandgap current. This reduces the are needed for high precision resistors. R3 and R4 are simple pull up or pull down resistors that can be designed using minimum width devices. Q1 is split in two devices (Q1A abd Q1B) placed left and right of Q2 for better matching. In this example the ratio of the current densities becomes 4+4=8. So the ideal bandgap voltage calculates:

$$V_{R3} = V_T * ln(n) \tag{7.70}$$

In the example shown we have:

$$n = 8$$

So the voltage at R1 becomes:

$$V_{R1} = V_T ln(n)$$

At 300K we get:

$$V_{R1300K} = 25.852mV * ln(n) = 53.758mV$$

Neglecting base currents assuming an ideal current mirror Q4, Q5 we will find twice the current of R1 flowing in R2. So the voltage accross R2 is:

$$V_{R2} = 2 * \frac{R_2}{R_1} * V_T * ln(n) \tag{7.71}$$

In the example shown this calculates:

$$V_{R2} = 586.45mV$$

Assuming Vbe=650mV we get an ideal bandgap voltage of:

$$V_{BGideal} = V_{be} + V_{R2} = 650mV + 586.45mV = 1.2365V$$

To calculate the frequency compensation it is necessary to know the transconductance of the bandgap. To concentrate on the essential things in the following figure the current mirror is built using PMOS transistors. So we don't have any inluence of base currents in the mirror. The starter has be omitted as well. To make things a bit more conveniant the output current of the bandgap is defined to flow into the circuit. Since the current mirror doesn't loose any current the output current becomes

$$I_{out} = I_2 - I_1 \tag{7.72}$$

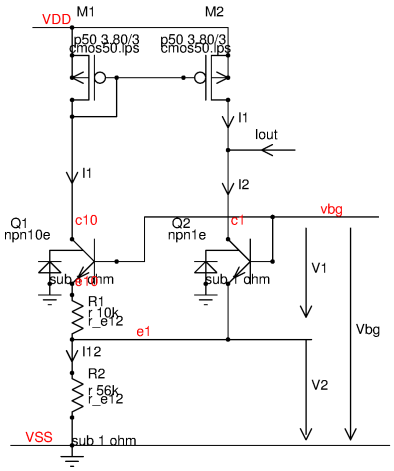


Figure 7.34: The core of the Brokow bandgap

To calculate the transconductances it is more conveniant to start with V1. In the operating point the currents through the transistors are equal.

$$I_1 = I_2 = \frac{I_{12}}{2} = I \tag{7.73}$$

Since both transistors operate at the same current we can simpy assume they have the same transconductance gm.

$$gm = \frac{I}{V_T} = \frac{I * e}{k * T} \tag{7.74}$$

For Q2 that has no resistor in the emitter we find the small signal transconductance

$$\frac{dI_2}{dV_1} = gm \tag{7.75}$$

At Q1 we have an emitter impedance in series with R1 leading to

$$\frac{dI_1}{dV_1} = \frac{1}{\frac{1}{gm} + R_1} \tag{7.76}$$

Thus the small signal transconductance of the bandgap becomes

$$\frac{dI_{out}}{dV_1} = \frac{dI_2}{dV_1} - \frac{dI_1}{dV_1} = gm - \frac{1}{\frac{1}{gm} + R_1} \tag{7.77}$$

This is an important result. The higher we chose R1 the higher the loop gain we can achieve. Of course we can not chose R1 arbitrarily. To be able to use a high R1 we have to use a high size ratio between the two transistors. On the other hand R1 following the logarithme of the areas of the emitter the increase of R1 is limited. Practical values for the size ratio of the transistors are about 8 to 15.

The second important observation is that the transconductance is a difference between a bipolar transconductance and a degraded transconductance (resistor in the emitter of the bigger transistor). The transconductance of the Brokaw bandgap is always lower than the transconductance of a bipolar differential amplifier. Therefore it is very important to have a good current mirror. In the circuit shown this current mirror is built using MOS transistors in stead of bipolar PNP transistors. This avoids the systematic error of the PNP transistor base currents. If the MOS current mirror is replaced by bipolar PNP transistors we add a significant systematic error!

Up to now we refered everything to V1 in stead of the bandgap voltage. So we have to find the ratio (small signal) between V1 and Vbg. The impedance at the emitter is

$$Z_e = \frac{1}{gm} \tag{7.78}$$

The (small signal) addmittance found from node vbg to node e1 becomes

$$Y = \frac{1}{Z_{e1}} = gm + \frac{1}{\frac{1}{gm} + R1} \tag{7.79}$$

$$Z_{e1} = \frac{1}{gm + \frac{1}{\frac{1}{gm} + R_1}} \tag{7.80}$$

Now the divider consisting of the impedance at e1 and R2 can be calculated. Since this is a small signal consideration the equation is written in its differential form.

$$\frac{dV_1}{dV_{bg}} = \frac{Z_{e1}}{Z_{e1} + R_2} = \frac{1}{1 + R_2 * gm + \frac{R_2}{\frac{1}{gm} + R_1}} \tag{7.81}$$

Refering everything to Vbg the transconductance of the bandgap becomes

$$\frac{dI_{out}}{dV_{bg}} = \frac{dI_{out}}{dV_1} * \frac{dV_1}{dV_{bg}} = \frac{gm - \frac{1}{\frac{1}{gm} + R_1}}{1 + R_2 * gm + \frac{R_2}{\frac{1}{gm} + R_1}} \tag{7.82}$$

Using the expression for the delta Vbe (depending on the emitter ratio n) and the current we can determine R1.

$$R_1 = \frac{V_T}{I} * ln(n) \tag{7.83}$$

Furtehrmore replacing gm by the current and Vt simplifies the equation for the transconductance of the bandgap

$$\frac{dI_{out}}{dV_{bg}} = \frac{I}{V_T} * \frac{1 - \frac{1}{1 + ln(n)}}{1 + \frac{I * R_2}{V_T} * \frac{2 + ln(n)}{1 + ln(n)}} \tag{7.84}$$

R2 is intentionally left in the equation because sometimes we might want to build a bandgap that has a temperature coefficient or someone wants a double bandgap. This can at low cost be built stacking further diodes and multiplying up R2. But this of course reduces the transconductance.

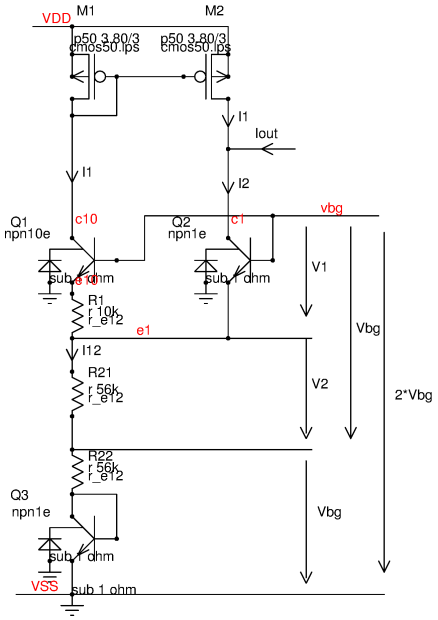


Figure 7.35: Stacking a diode more to create a double bandgap at low cost

Supply transient response of the Brokaw bandgap: The bigger NPN transistor with the multiple emitters has a higher capacity to ground than the smaller single emitter transistor. Therefore at a fast rising edge of the supply the Brokaw bandgap usually produces an overshot at the output. At a fast falling edge the bandgap turns off for some microseconds.

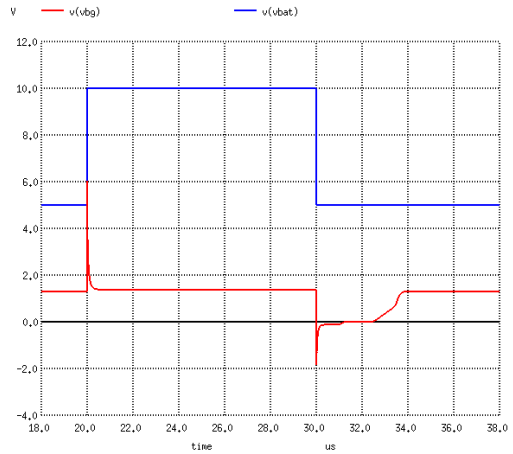


Figure 7.36: Response of the Brokaw bandgap to a transient on the supply rail

Replacing the PNP transistors by PMOS transistors helps to reduce the parasitic capacity to ground at the base of the current mirror. But the improvement is limited because we still have the capacity of the big NPN transistor that can't be avoided in the brokaw bandgap.

EMC performance of the Brokaw bandgap: The Brokaw bandgap is most sensitive to RF injected into the emitters of the bandgap transistors. Q1 and Q2 are extremely fast rectifiers. Usually the base of both transistors is more or less low resistive tied to the ground node of the bandgap (often there is a frequency compensation tying the base to circuit ground).

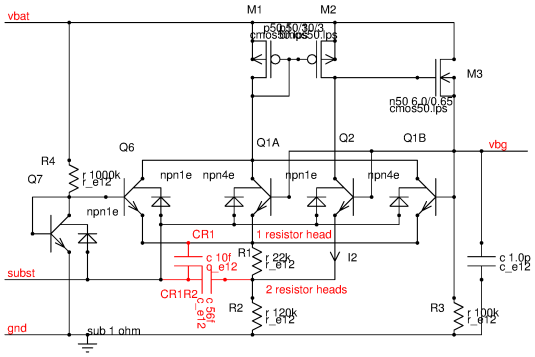


Figure 7.37: Parasitic capacities CR1 and CR1R2 can couple sbstrate noise into the emitters of the bandgap transistors ifthe resistors are placed over substrate.

Substrate noise coupled into the emitters will be amplified and peak rectified (negative peak) by Q1 and Q2. Usually Q2 wins because it sees the capacity of two resistors. This will pull down the bandgap even if the injected RF frequency is higher than the transit frequency of the transistors (common base circuit like the UHF tuners of the old days!).

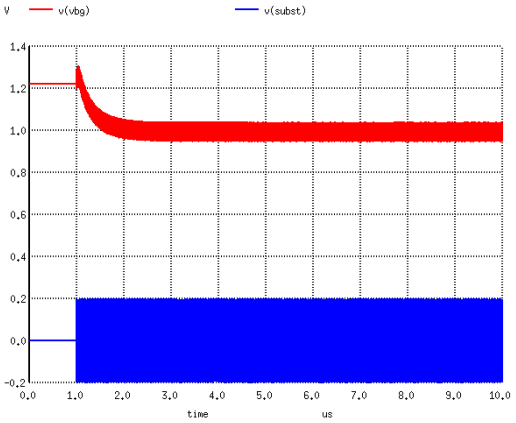


Figure 7.38: 200mV of substrate bounce at 200MHz pulls down the bandgap

The effect of RF coupling through the parasitic capacity at the bottom side of the resistors becomes stronger when the circuit is designed with higher impedances (bigger resistor area AND higher node impedance square the problem reducing current consumption of the bandgap!)

It is good engineering practice to use resistors over wells to isolate the resistors from substrate noise.

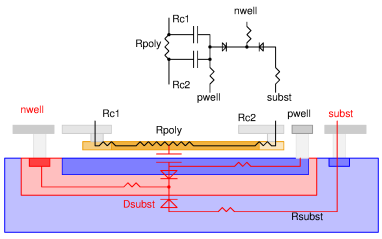


Figure 7.39: double well isolation of the poly silicon resistor

pwell has to be connected to the ground node of the bandgap (exactly where the bandgap voltage refers to). nwell usually is connected to the supply of the bandgap to prevent latch up. This way substrate noise is attenuated twice before it reaches the back side capacity of the resistor.

7.4.3 The Barba CMOS bandgap

Well, the following bandgap is nice and cheap BUT start up is unreliable. Therefore it can be found in hundreds of books but it is barely used in real chip design. Nevertheless project managers love it to tell designers they are wasting space using something else.....

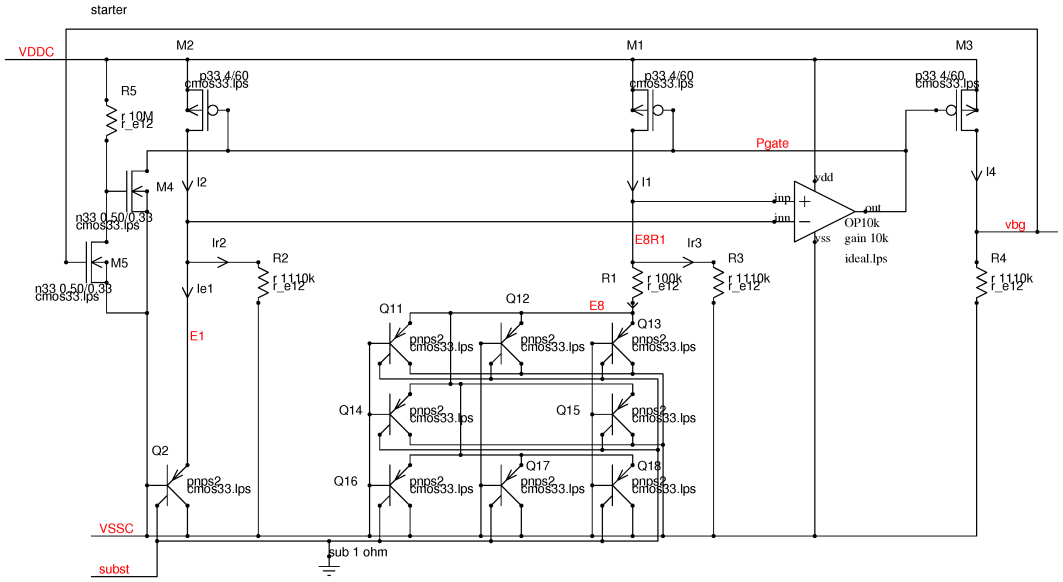


Figure 7.40: The Barba bandgap

In most CMOS processes the only bipolar component available is the substrate PNP transistor. Sad but true the collector usually is not available for circuit design. It is coincident with the substrate node. So the only thing to be done is to connect the base to ground (metallic circuit ground, reference ground, analog ground or whatever else the top level designer called the ground all precision stages refer to). The emitters are freely available. So the resistors can be connected to the emitters. The temperature voltage is the difference between the single emitter transistor and the multi emitter transistor.

Practical consideration for design anarchists: Your technology does not have this bipolar transistor? Don't worry. Look at the cross sections of the PMOS transistors. Just misuse the PMOS transistors et voila, you can build a bandgap. If you can simulate it depends on the detail level of the model of the PMOS transistors. And if you can't simulate it do it like in the old days. Forget simulation. Use a spread sheet.

So how does that beast work? Assuming M1 and M2 are matching (quite a challenge at low currents) the sum of the currents through Q2 and R2 is equal to the current flowing through Q11 to Q18 and R3. Furthermore the operational amplifier regulates the voltage at R2 and R3 to exactly the same value. Thus the currents through R2 and R3 become equal. Since I1 and I2 are equal as well the currents into Q2 and into Q11 to Q18 must be equal as well. As a consequence we find the old equation again:

$$V_{R1} = V_T ln(n) \tag{7.85}$$

So the currents flowing into R1 and into Q2 become:

$$I_{R1} = I_{Q2} = \frac{V_T * ln(n)}{R_1} \tag{7.86}$$

The parallel path R2 and R3 carries a current of:

$$I_{R2} = I_{R3} = V_{be}/R_2 = V_{be}/R_3 = V_{be}/R \tag{7.87}$$

assuming R2=R3=R. So the current flowing in M1 and M2 is:

$$I_{M1} = I_{M2} = \frac{V_{be}}{R} + \frac{V_T * ln(n)}{R_1} \tag{7.88}$$

Now R1, R2 and R3 must be chosen in a way that the current flowing through M1 and M2 becomes constant (except for the temperature coefficient of the resistors). So let us have a look at the derivatives:

$$\frac{dI_{M1}}{dT} = \frac{dI_{M2}}{dT} = \frac{dV_{be}}{dT * R} + \frac{dV_T * ln(n)}{dT * R_1} = 0 \tag{7.89}$$

In other words:

$$-\frac{R_1}{ln(n)} * \frac{\frac{dV_{be}}{dT}}{\frac{dV_T}{dT}} = R = R_2 = R_3 \tag{7.90}$$

Still looks terrible? No! We can take it from some lines up:

$$\frac{dV_{be}}{dT} = -2mV/K \tag{7.91}$$

(what did you expect?) and

$$\frac{dV_T}{dT} = \frac{k}{e} = 86.1733\mu V/K$$

plugging the numbers in we get:

$$R = R_2 = R_3 = R_1 * \frac{23.209}{ln(n)}$$

With this sizing of R2 and R3 the current through M1 and M2 gets constant. So we have the same constant current (or a multiple of a constant current depending on the size of M3) flowing into R4. As a result the voltage at R4 becomes constant. If R4=R3=R2 and M3 is equal M1 and M2 we just will find the bandgap voltage at R4.

The starter problem of the Barba CMOS bandgap: The parallel resistor bandgap suffers from the starter problem. As long as there is no current flow through the bipolar transistors (V(E1) = V(E8) < Vbe) the start up depends on the resistors R2 and R3 only. Ideally R2=R3 and the lop gain is exactly one. So the circuit is weak stable between 0V and Vbe. This means t will neither start nor turn off.

In practice the resistors are slightly different. If R2 is slightly bigger than R3 the operational amplifier will see a slightly higher voltage at the negative input. Provided the current mirror matches and the amplifier has no offset the circuit will start.

The opposite happens if R2 is slightly less than R3. Now the positive input is slightly higher. In the following figure this is the range where the blue line (representing the inverted current I1) is above the red line (representing the inverted current I2). In stead of turning on the amplifier will turn off the current mirror. Thus the bandgap will not start but turn off! Only when the current through the bipolar components increases the voltage at E8 will be lower than the voltage at E1. In the appendant figure this is the range where the blue line is below the red line.

Bottom line we have 3 stable operating points:

1. P1: V(E1)=V(E8R1)=0: The bandgap is off. This point is strong stable if R2<R3.
2. P2: V(E1)=V(E8R1)=V(E8)=Vbe: This is a weak stable point where the amplifier input voltage becomes exactly 0. There is no current flowing through the bipolars yet.
3. P3: V(E1)=V(E8R1) but due to current flowing V(E8R1)>V(E8). This is the desired operating point of the bandgap.

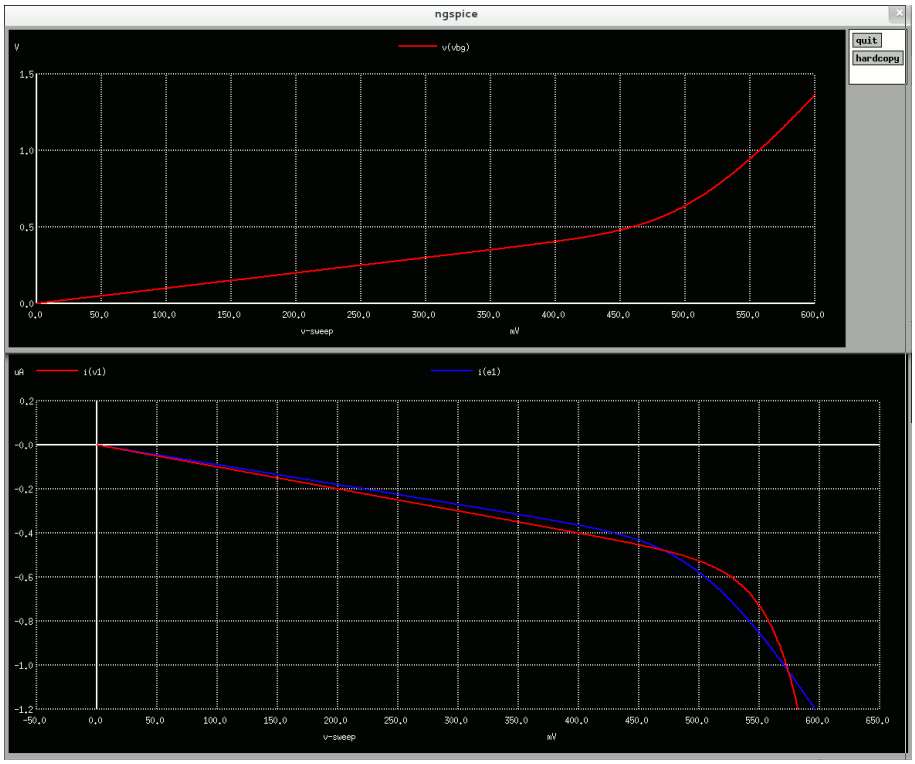


Figure 7.41: Start up conditions if R2 is less than R3

In the above figure the blue line above the red line means we are in a range the bandgap voltage will decrease (move to the left). If the red line is above the blue line the bandgap voltage will increase (move to the right). Looking at the bandgap voltage the weak stable point P2 is found at Vbe. The desired operating point P3 is at bandgap voltage.

For the design of the starter this means the starter may not turn off below P2 (Vbe) but must turn off before reaching P3 (Vbg). Thus the starter MUST correlate with Vbe. In the circuit shown above this means the threshold of M5 must under any operating conditions satisfy the condition:

$$V_{be} < V_{thM5} < V_{bg}$$

Since MOS transistors have a considerable production spread and M5 operates through a wide range of current flowing through starter resistor R5 this simple starter can not be expected to do the job sufficiently reliable! This is why I called it the academic CMOS bandgap. If we have a starter that correlates with Vbe the problem can be solved. But this makes the once very simple bandgap as complex as all the others and the benefit of the simple circuit is lost. (to be precise: due to component leakages P3 may exist in other bandgap topologies too. But there P3 is somewhere around Vout $<$ 10mV and we have no problems with the spread of the threshold of M5.)

Statistical errors of the Barba CMOS bandgap: The bandgap shown above requires the matching of the following components:

1. The bipolar transistors Q2 and Q11 to Q18 must match.
2. The offset of the operational amplifier must be small compared to Vt.
3. The resistors R1, R2, R3 and R4 must match.
4. Current mirrors M1, M2 and M3 must match.

Let us calculate the propagation of errors of these components into the bandgap voltage.

An offset of the bipolar transistors will be multiplied by the ratio of R1 and R4 and multiplication factor of M3. Assuming M3 has the same size as M1 and M2 the multiplication factor becomes 1. Thus we find:

$$\Delta V_{bgbip} = V_{osbip} * \frac{R_4}{R_1} \tag{7.92}$$

An offset of the operational amplifier propagates the same way (again assuming W/L of M3 is equal to W/L of M1 and M2)

$$\Delta V_{bgOP} = V_{osopamp} * \frac{R_4}{R_1} \tag{7.93}$$

Resistor spread usually is the bigger the smaller the resistors are. In most designs the area of R1 is the smallest and the spread of R1 is the most significant of all the resistor errors. It propagates as follows:

$$I_{R1} = \frac{V_T * ln(n)}{R_1} \tag{7.94}$$

$$\frac{dI_{R1}}{dR_1} = -\frac{V_T * ln(n)}{R_1^2} \tag{7.95}$$

Multiplying this expression with the change of R1 we get:

$$\Delta I_{R1} = -\frac{\Delta R_1}{R_1} * \frac{V_T * ln(n)}{R_1} \tag{7.96}$$

Now we just have to multiply the change of the current caused by the deviation of R1 with R4 to find the change of the bandgap voltage due to R1 spread.

$$\Delta V_{bgR1} = -\frac{\Delta R_1}{R_1} * \frac{R_4}{R_1} * V_T * ln(n) \tag{7.97}$$

As long as offsets are caused by area mismatch of bipolar transistors or MOS transistors operating in weak inversion (This usually is the case at well designed operational amplifier input stages) these offsets follow VT. The same applies to the offset caused by a deviation of R1. So these errors can be trimmed without producing severe problems with temperature coefficients. Looking at the matching of M1 to M3 the situation becomes different. Since we want to have good current matching these transistors usually are designed to operate in strong inversion. We have a parabola shaped characteristic in stead of an exponential one. So the offset caused by M1, M2 and M3 mismatch will not follow Vt. The offset errors of the current mirror can only be trimmed out for one temperature but the temperature coefficient remains!

For the calculation of the error contribution of M1, M2, M3 we need to know the operating point of the transistors and the technology properties (gm, tox, matching coefficient..). Empirically the matching coefficients of MOS transistors follow about tox.

$$match \approx \frac{t_{ox}}{nm} * mV * \mu m$$

To calculate the offset voltage we must know the gate area of the MOS transistors.

$$V_{osMOS} = \frac{match}{\sqrt{W * L}}$$

Since we are interested in the current error in stead of the offset voltage we have to figure out the transconductance of the transistors in their operating point.

$$I_d = g_m * \frac{W}{L} * V_{gs}^2 \tag{7.98}$$

$$\frac{dI_d}{dV_{gs}} = 2 * g_m * \frac{W}{L} * V_{gs} \tag{7.99}$$

$$\Delta I_d = \frac{dI_d}{dV_{gs}} * V_{osMOS} = 2 * g_m * \frac{W}{L} * V_{gs} * V_{osMOS} \tag{7.100}$$

$$\frac{\Delta I_d}{I_d} = \frac{V_{osMOS}}{V_{gs}} \tag{7.101}$$

If the current through M1 deviates from the current through M2 we will see a deviation of the bandgap output voltage of:

$$\Delta V_{bgM1} = R_4 * I_d * \frac{V_{osM1}}{V_{gs}} \tag{7.102}$$

The same applies if M3 deviates from M1 and M3.

$$\Delta V_{bgM3} = R_4 * I_d * \frac{V_{osM3}}{V_{gs}} \tag{7.103}$$

Assuming all these errors are statistically independent of each other we have to add the squares and take the square root of them.

$$\Delta V_{bg} = \sqrt{\Delta V_{bgbip}^2 + \Delta V_{bgOP}^2 + \Delta V_{bgR1}^2 + \Delta V_{bgM1}^2 + \Delta V_{bgM3}^2} \tag{7.104}$$

As mentioned before only the first 3 errors can be trimmed over temperature. So the following is a reasonable worst case estimation for the trimmed bandgap:

$$\Delta V_{bgtrimmed} = \sqrt{\Delta V_{bgM1}^2 + \Delta V_{bgM3}^2} \tag{7.105}$$

Note: These are one sigma errors! Usually you want to have good production yield and the 1s error should be about 5 to 6 times lower than the specified error (5 sigma design or 6 sigma design strategy).

7.4.4 CMOS bandgap with improved accuracy

Th CMOS bandgap shown before has two issues:

1. The poor accuracy due to the high number of error sources.
2. The critical start up circuit.

The following bandgap reduces the number of error sources and solves the start up problem. As a disadvantage it does not provide a reference current for free. It automatically provides a bias current with positive temperature coefficient (ptat current).

$$I_{ptat} \sim V_T/R \tag{7.106}$$

To create a temperature constant current a current with negative temperature coefficient must be added (ntat current). In the following circuit the ntat current is created using Vbe and a resistor.

$$I_{ntat} \sim V_{be}/R \tag{7.107}$$

At the end the desired current generator temperature coefficient can be chosen by the ratio of the two currents added.

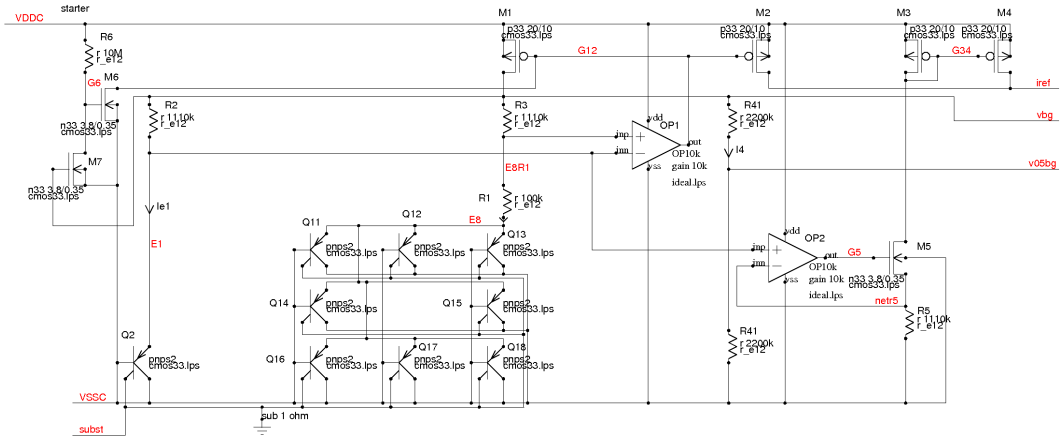


Figure 7.42: Improved CMOS bandgap

Now the bandgap voltage is created inside the regulation loop. The matching errors of the PMOS transistors do not influence the bandgap voltage anymore. The only parameter still affected by the mismatch of the PMOS transistors is the reference current.

The bandgap shown can be regarded as a Brokaw bandgap turned upside down. In stead of using the bipolar transistors as an amplifier stage we now have to use the operational amplifier OP1 because in most CMOS technologies the collector of the PNP transistors is coincident with the substrate. We find the usual equations:

$$V_{R1} = V_T * ln(n) \tag{7.108}$$

$$V_{R2} = V_{R3} = \frac{R_2}{R_1} * V_T * ln(n) \tag{7.109}$$

$$V_{bg} = V_{be} + V_{R2} \tag{7.110}$$

As usual we have to chose the ratio of R1 and R2 in such a way that the temperature coefficients cancel.

$$\frac{R_2}{R_1} * \frac{k}{T} * ln(n) = 2mV/K$$

Propagation of errors to the bandgap voltage: We have 3 main errors affecting the bandgap voltage:

1. The bipolar transistors Q2 and Q11 to Q18 must match.
2. The offset of the operational amplifier must be small compared to Vt.
3. The resistors R1, R2, and R3 must match.

The offset of the bipolar transistors propagates to the output voltage of the bandgap:

$$\Delta V_{bgbip} = V_{osbip} * \frac{R_2}{R_1} \tag{7.111}$$

An offset of the operational amplifier propagates the same way:

$$\Delta V_{bgOP} = V_{osopamp} * \frac{R_2}{R_1} \tag{7.112}$$

Resistor spread usually is the bigger the smaller the resistors are. In most designs the area of R1 is the smallest and the spread of R1 is the most significant of all the resistor errors. It propagates as follows:

$$I_{R1} = \frac{V_T * ln(n)}{R_1} \tag{7.113}$$

$$\frac{dI_{R1}}{dR_1} = -\frac{V_T * ln(n)}{R_1^2} \tag{7.114}$$

Multiplying this expression with the change of R1 we get:

$$\Delta I_{R1} = -\frac{\Delta R_1}{R_1} * \frac{V_T * ln(n)}{R_1} \tag{7.115}$$

Now we just have to multiply the change of the current caused by the deviation of R1 with R2 to find the change of the bandgap voltage due to R1 spread.

$$\Delta V_{bgR1} = -\frac{\Delta R_1}{R_1} * \frac{R_2}{R_1} * V_T * ln(n) \tag{7.116}$$

The total 1s statistical error becomes:

$$\Delta V_{bg} = \sqrt{\Delta V_{bgbip}^2 + \Delta V_{bgOP}^2 + \Delta V_{bgR1}^2} \tag{7.117}$$

All these errors follow Vt. So trimming the bandgap can be expected to be efficient over the whole temperature range. This is a significant difference to the equations of the bandgap with parallel resistors at the bipolar transistors.

Propagation of errors to the reference current: The reference current is affected by the following statistical errors:

1. The error of the ptat current
2. The error of the ntat current
3. the error of current mirror M1, M2
4. The error of the current mirror M3, M4
5. The absolute value of R1 and R5

If the spread of the reference current is a problem depends on the application. The spread of the absolute value of the resistors is in the range of +-20% (at one temperature!). There are two cases to be distinguished:

1. If the reference current is used to supply other cells on the same chip that require a defined voltage drop across resistors of the same type the spread of the resistors absolute value cancels.
2. If the reference current is compared to current flowing outside of the chip the absolute value of the current must be trimmed.

Error of the ptat current: The ptat current error has 3 main contributors: The offset of the bipolar transistors, the offset of the amplifier and the deviation of R1.

$$\frac{\Delta I_{ptatbip}}{I_{ptat}} = \frac{V_{osbip}}{V_T * ln(n)} \tag{7.118}$$

$$\frac{\Delta I_{ptatOP}}{I_{ptat}} = \frac{V_{osopamp}}{V_T * ln(n)} \tag{7.119}$$

The error contribution of R1 is a function of the deviation of R1 due to shape inaccuracies compared to an ideal R1 with perfect shape. $m_R$is the matching number of the resistor R1. It usually is scaled in $\%\mu m$. Typical values for modern technologies are around 3%μm to about 10%μm.

$$\frac{\Delta I_{ptatR1}}{I_{ptat}} = \frac{\Delta R_1}{R_1} = \frac{m_R}{\sqrt{W_{R1} * L_{R1}}} \tag{7.120}$$

Error of the ntat current:
The most important error of the ntat current generator is the offset of the opamp OP2.

$$\frac{\Delta I_{ntat}}{I_{ntat}} = \frac{V_{osOP2}}{V_{be}} \tag{7.121}$$

Since Vbe usually is one magnitude bigger than the voltage at R1 the offset requirements of OP2 are much less difficult to be met than the offset requirements of OP1. If OP1 is designed for 0.25mV offset OP2 can typically be designed for 2mV offset spread. Thus the area of the input transistor gates of OP2 can be about a factor 50..100 smaller than the area of the gates of the input transistors of OP1.

Error of the PMOS current mirrors: The same calculation as all current mirrors.

$$V_{osMOS} = \frac{match}{\sqrt{W * L}}$$

"match" is the matching coefficient of the PMOS transistors. Normally this number can be found in the technology documentation. If there is no information to be found a rough guess is:

$$match \approx \frac{t_{ox}}{nm} * mV * \mu m$$

The current error of the mirror is:

$$\frac{\Delta I_d}{I_d} = \frac{V_{osMOS}}{V_{gs}} \tag{7.122}$$

The gate overdrive Vgs can be calculates as:

$$V_{gs} = \sqrt{\frac{I_d * L}{gm * W}} \tag{7.123}$$

A lot of equations... Well, probably the most pragmatic idea is to plug it all into a spread sheet. Of course you can just as well design something suitable as a first guess design and run Monte Carlo simulations until things look nice. But a spread sheet is faster and it pinpoints what is causing most of the trouble much better than a simulation!

And now including the starter: At the end the amplifier OP1 has to be broken down to the transistor level to see if the starter works. Here comes an example:

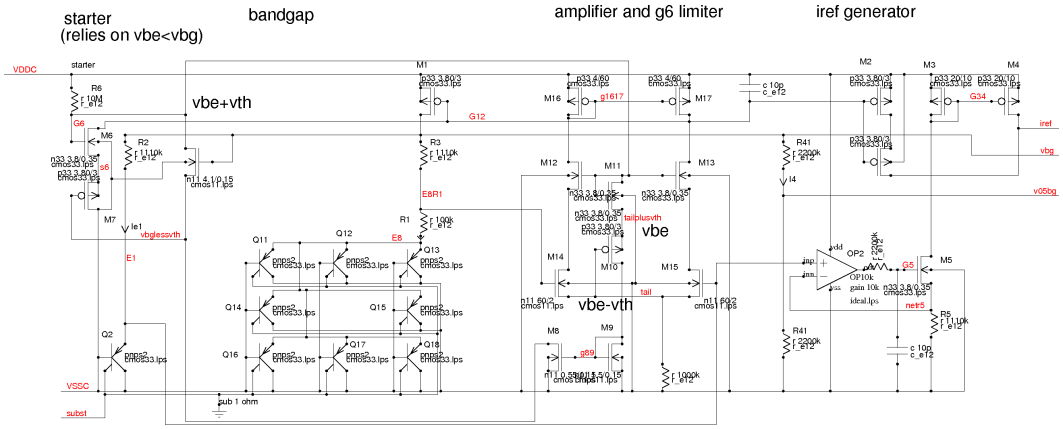


Figure 7.43: The CMOS bandgap including the starter

In the example shown the starter also biases the cascodes. This way in the amplifier transistors with a thinner gate oxide can be used (type n11, M14, M15, M8, M9). This offers a lower offset voltage with less area penalty. (This approach makes sense in modern technologies having single gate oxide transistors for the 1.5V logic and analog circuits and double gate oxide transistors for more rugged I/O cells.)

M8, M9 recycles the current flowing through the clamp M10 and M11. This simply is cheaper than adding an other resistor of several Meg. Ohms.

7.4.5 Bruno's Bandgap

The following bandgap was proposed to me by Bruno Murari 1991. At that time we didn't have any information if it ever had been used before. Later I found a remark in Hans Camencind´s book that the same topology has already been used by Bob Widlar [32]. It can be regarded as a variation of the Brokow bandgap just taking the delta Vbe into the base branch. This little change nevertheless has several consequences.

1. The loop gain increases and current mirror errors don't propagate as much as in other bandgap designs.
2. The base current produces an additional error. This error can be canceled for typical production runs, but the cancellation doesn't work for corner runs (e.g. low B)
3. We get an NPN current generator for free.

Bottom line a nice design for low performance applications but not suited for high precision.

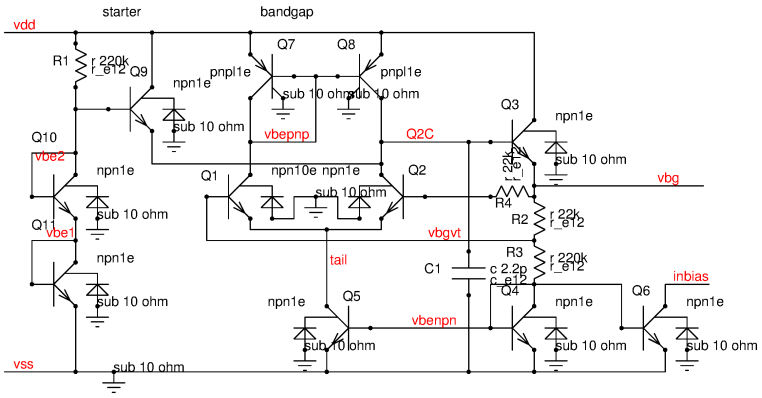


Figure 7.44: Bruno Murari's bandgap proposal

Using this bandgap take care that Q6 always has enough collector voltage. If Q6 goes into saturation it will steal the base current of Q5!

Assuming an infinite gain of the bipolar transistors the voltage across R2 becomes:

$$V_{R2} = V_T * ln(n) \tag{7.124}$$

The resulting current at inbias becomes

$$I_{inbias} = V_{R2}/R_2 = \frac{V_T * ln(n)}{R_2} \tag{7.125}$$

Assuming n=10 at room temperature this current would be 2.72μA.

The base current of Q1 produces a systematic error. The base current calculates as:

$$I_{Q2base} = \frac{V_T * ln(n)}{2 * B * R_2} \tag{7.126}$$

The error current flows through R2 but not through R3. To compensate the additional drop at R2 R4 is added to the base path of Q2. The exact bandgap voltage is at the base of Q2. For pragmatic reasons in stead of providing the voltage at the base of Q2 to other blocks usually the voltage at the emitter of Q3 is getting distributed (lower impedance, less sensitive to capacitive loads). The drop over R4 (and Rthe additional drop at R2) is the systematic error of the bandgap voltage:

$$V_{errorQ2} = \frac{V_T * ln(n)}{2 * B} \tag{7.127}$$

As an example let us assume a gain of the NPN transistors of B=100.

$$V_{errorQ2} = \frac{26mV * ln(10)}{2 * 100} = 0.3mV$$

Compared to the error caused by the base current of Q7 and Q8 these 0.3mV are negligible (provided R4 is really present. If R4 is missing the error caused by the base current will be multiplied by the ration of R2 and R3 and will get a negative sign)). Q1 has to carry the collector current of Q7 plus the base currents of Q7 and Q8. Furthermore the base current of Q3 reduces the current to be carried by Q2. Adding all these effects and assuming a gain of the PNP transistors of Bpnp=50 the current ratio of Q1 and Q2 can easily reach 1.05 (in stead of the ideal 1.00). The systematic error caused by this current mismatch becomes:

$$V_{errormirror} = V_T * \frac{R_2 + R_3}{R_2} * ln(\frac{1 + 2/B_{pnp}}{1 - 2/B_{npn}}) \tag{7.128}$$

Note: Q3 operates at twice the current of Q7 and Q8. So the base current error gets a higher weight. Let's plug in some typical numbers:

$$V_{errormirror} = 26mV * 11 * ln(\frac{1.04}{0.98}) = 17mV$$

In case of the original Brokaw bandgap the systematic error of the bipolar current mirror would even be worse because the resistor in the emitter path reduces the differential transconductance of Q1 and Q2 (by typically 30% for about 7..10 emitters). For a pure bipolar technology shifting the delta Vbe into the base side can make sense. We are accepting an additional systematic error of 3mV to achieve a higher loop gain that reduces the error propagation of

the current mirror. The poorer the performance of the PNP transistors the higher the benefit of Bruno's bandgap compared to the Brokaw bandgap.

If Q3, Q7, Q8 are replaced by MOS transistors (That have no base current) the original Brokaw bandgap becomes the better choice. The same applies if the PNP current mirror is replaced by a boosted current mirror and Q3 is replaced by a darlington transistor.

EMC performance: Bruno Murari's bandgap is sensitive to substrate noise. There are 3 almost unavoidable substrate capacities:

- The collector of Q2
- The collector of Q1
- The base of Q7 and Q8

The following describes a variant of the Murari bandgap that caused a lot of EMC problems. The modification versus the original were the following:

- The regulator transistor in the variant was a PMOS P3 to make it a low drop design
- The bipolar current mirror was replaced by a PMOS current mirror P1 and P2 because the technology didn't offer lateral PNPs.

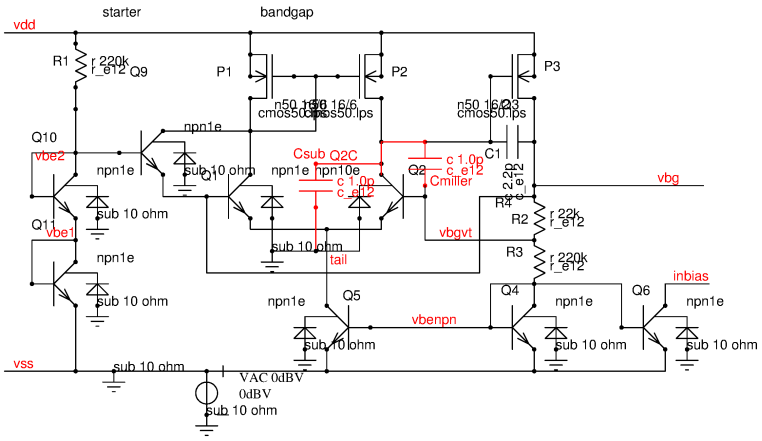


Figure 7.45: A variant of Bruno Murari's bandgap that caused severe electromagnetic susceptivity problems

In this modified version the collector of Q2 is the biggest capacity to substrate (Csub). If the substrate is connected to a different ground than the vss of the bandgap (IC packages with exposed dice pad!) the substrate voltage at high frequency can differ significantly from vss. Relative to vss the substrate can easily bounce several hundred mV!

This substrate noise reaches the collector of Q2 via the substrate capacity Csub. The collector impedance of Q2 is high because the frequency was too high for the regulator transistor to have a significant gain. So the collector follows this substrate bounce. Since we are looking at frequencies close to the transit frequency of Q2 or even above the transit frequency the RF reaches the base of Q2 via the miller capacity Cmiller. At the base the RF gets rectified. During the positive half wave the base-emitter diode of Q2 conducts. Since we are close to the transit frequency the collect current decreases in stead of increasing during the positive half wave!

$$I_{CQ2} = I_{tail} - I_{base} \tag{7.129}$$

The current flow charges the miller capacity Cmiller.

During the negative half wave the base emitter diode blocks. The charge inside Cmiller remains stored. (except for a little bit of current flowing into the feedback resistors, but this usually is negligible.). In addition during the negative half wave Q2 is off (well, at these frequencies we rather say 'less on' than in DC operation). So there is no current discharging the collector of Q2.

Due to this charge pumping into Cmiller and the base-emitter diode of Q2 the average collector voltage of Q2 moves up and the PMOS transistor P3 is regulated back. The bandgap voltage drops until the peak voltage at the base falls below the base voltage of Q1 (Then the charge pumping into Cmiller stops).

The problem could be eliminated in simulation assigning the substrate node of Q2 to vss. (This of course only is possible in simulation). Eventually the problem could be solved adding a capacity between the base of Q2 and vss. This capacity attenuates the RF flowing via Cmiller. The drawback is that we add one more pole to the regulation loop. This can partly be fixed adding a zero to the miller compensation (A resistor in series with C1).

7.4.6 Weak Inversion Bandgap

In principal every kind of bandgap built with bipolar transistors can just as well be built using CMOS transistors operating in weak inversion. In bipolar bandgaps the voltage with the positive temperature coefficient is constructed with a delta Vbe.

$$\Delta V_{be} = \frac{k * T}{e} * ln(m) \tag{7.130}$$

with k being the Boltzmann constant, T being the temperature in K and e being the electron charge. m is the ratio of the current densities flowing in the bipolar transistors. Using MOS transistors operating in weak inversion the equation looks very similar.

$$\Delta V_{gs} = \frac{k * T}{e} * n * ln(m) \tag{7.131}$$

The only difference is the factor n. n is the inverse coupling factor of the gate voltage to the channel.

$$n = \frac{C_g + C_{bulk}}{C_g} \tag{7.132}$$

Next step the gate to channel capacity C_g and the channel to bulk capacity C_{bulk} must be calculated. (The calculation can be found in the chapter describing the MOS transistor. Here only the result is shown.)

$$n = 1 + \frac{\varepsilon_{si} * t_{ox}}{\varepsilon_{sio2} * \sqrt{\frac{2 * \varepsilon_{si} * (\Phi - V_b)}{q * N_b}}} \tag{7.133}$$

Looks a bit ugly. Bottom line we have a multiplication factor n in the range of 1.1 to 1.6 for most technologies. What is not so nice is that this factor is not fully constant! It changes a little bit with the doping of the bulk and the work function Φ (at the surface of the silicon) and the bulk voltage (back gate). As a consequence bandgap circuits using MOS transistors in weak inversion by concept are less precise and more technology dependent than their bipolar counterparts. Literature reports weak inversion bandgaps with a 1σ spread in the range of 1% to 2%. The thinner the gate oxide the better it gets. (Well, a bipolar transistor can be regarded as a MOS transistor with a gate oxide thickness of $t_{ox} = 0$. So n approaches 1 and we exactly get the equation of ΔV_{gs} .)

The temperature coefficient of V_{gs} becomes:

$$\frac{dV_{gs}}{dT} = \frac{k * n}{e} * ln(m) \tag{7.134}$$

At 300K this is

$$\frac{dV_{gs}}{dT} = n * ln(m) * 86.1733 \mu V/K$$

To compensate n the resistor ratio just has to be adjusted a little bit (compared to the bipolar bandgap).

Since we have shown that a weak inversion bandgap by concept is less precise than a bipolar bandgap, why do people still use it? The answer is simple: MOS transistors in modern technologies are smaller than bipolar ones. So a weak inversion bandgap is just cheaper than a bipolar one. That is the only attractive feature of it.

7.4.7 Open loop weak inversion bandgap

The most simple example of a weak inversion bandgap is shown below.

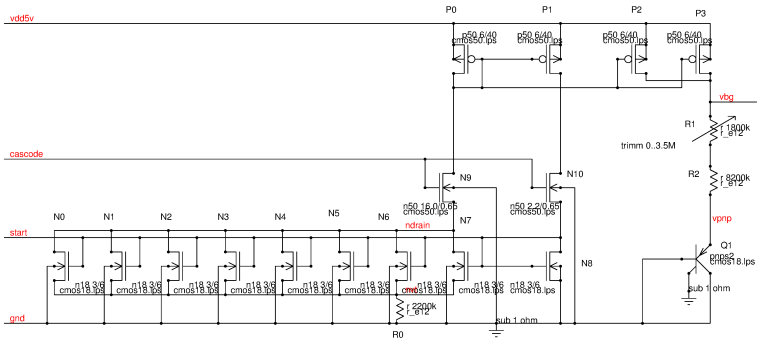


Figure 7.46: Example of a simple weak inversion bandgap

N0 to N8 operate in weak inversion. The voltage drop across R0 is in the range of 60mV. The current flowing is 30nA. The current flowing through R1 and R2 is 60nA.

Weak inversion bandgap using MOS transistors only: Q1 can be replaced by an NMOS transistor. In weak inversion the threshold voltage like Vbe of a bipolar transistor has a negative temperature coefficient. Normally a copy of N0 to N8 will be used as a "diode". Unfortunatelly Replacing Qi by a MOS diode adds further error sources to the design. The threshold of a MOS transistor can be engineered (intentionally as well as unintentionally) by the doping of the gate poly silicon and by the doping at the surface of the channel.

$$V_{th} = -\varphi_{bi} + 2 * \varphi_b + \sqrt{4 * \epsilon_{si} * q * N_a * \varphi_b}/c_{ox} \tag{7.135}$$

In this equation $\varphi_{bi}$is the built in voltage that depends on the work function of the gate material. In case of a poly silicon gate this value can be adjusted in a certain range by the gate doping. For n-doped poly silicon gates of a NMOS transitor $\varphi_{bi}$is in the range of -0.1V to -0.2V [69] . Using p-doped poly silicon for the NMOS transistor the built in voltage φ_{bi}can becomes positive leading to a threshold of 0V or even a negative threshold. (This would be a somewhat unusual process, but it can be done.) If the gate material differs from silicon (for instance if a real metal gate is used) the difference of the work functions of the gate material and silicon has to be used. This means transistors using a gate that is not poly silicon may have completely different thresholds! (The work function of silicon is 4.05eV. The built in voltage is the difference between the work function of the gate material and silicon. A nice table can be found at [70])

φ_b is the "distance" between the Fermi level of the doped bulk semiconductor and the intrinsic Fermi level.

$$\varphi_b = V_{th} * ln(\frac{N_a}{n_i}) \tag{7.136}$$

N_a is the acceptor doping of the substrate. q is the elementary charge. The choice of the substrate doping can modify the threshold in a range of 1V to 2V. ($c_{ox}$is the specific capacity t_{ox}/ϵ_{ox} of the gate dielectric.)

Since there are so many possibilities to modify the gate voltage this is done to tune the thresholds for the best performance of the logic. For the weak inversion bandgap this means the bandgap voltage depends on the way the threshold of the transistor used is tuned. Here are some examples of my own experience:

Table 29: Weak inversion MOS bandgap voltages (personal experience)

vdd	tox	Vth	Vbg	remark
5V	25nm	1.7V	3.9V	simulated, not used in a product due to spread seen in corner simulation
5V	15nm	1.3V	3.3V	only used for cascode bias. Never used as a reference.
3.3V	7nm	0.9V	2.3V	very poor accuracy
1.2V	3nm	0.4V	1.4V	used transistor without halo implant

The table already shows that the weak inversion bandgap voltage deviates a lot from the bandgap voltage found using bipolar transistors. Replacing Q1 of the figure above by a MOS diode means you are at the mercy of the process engineer. If the process is tuned for better digital performance (I will be tuned. You can bet on it!) your bandgap will change it's behavior and it's typical voltage.

7.5 Current generators

Almost every circuit needs a bias current generator. The only exception I am aware off are pure CMOS logic ICs.

In many cases the currents can - more or less for free - been taken out of the bandgap. This works nicely for currents in the μA range to the some $10\mu A$ range. In case your bandgap doesn't provide the current you need for free, here are some ideas for local current generators.

7.5.1 VBE over R generator

Probably the most trivial one.

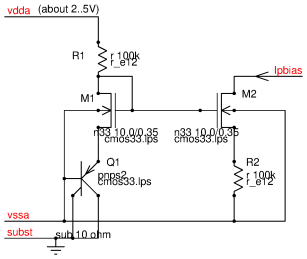


Figure 7.47: VBE over R current generator

The voltage drop over R2 follows the base-emitter voltage of Q1. The drain current of M2 is

$$I_{pbias} = \frac{V_{be}}{R2}$$

7.5.2 Vth over R generator

Just as trivial as before.

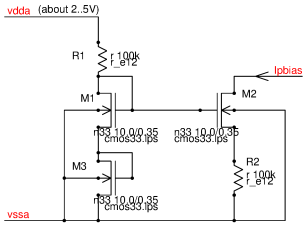


Figure 7.48: Vth over R generator

As you already will guess:

$$I_{pbias} = \frac{V_{th}}{R2}$$

Both, the Vbe over R and the Vth over R generator usually provide a current with negative temperature coefficient. The Vth over R generator's temperature coefficient can be tuned to a certain extent by the choice of the current density in M3. The higher the current density in M3 the less negative the temperature coefficient. Tuning of the temperature coefficient however relies on the technology parameters of the NMOS transistors. If the technology changes the tuning of the temperature coefficient has to be adjusted again.

7.5.3 Vt over R current generator

Building a Vt over R current generator is simple as long as a bandgap reference is available. A bandgap is always a sum of a Ptat voltage and an Ntat voltage. The Ptat voltage that is always a multiple of the temperature voltage $V_t = k*T/e$ can simply be retrieved subtracting an base-emitter voltage from the bandgap voltage. If the technology offers NPN transistor this is very easy.

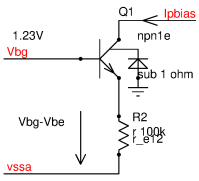


Figure 7.49: Vt over R current generator using an NPN transistor

The current calculates as

$$I_{pbias} = \frac{V_{bg} - V_{be}}{R2}$$

The temperature coefficient of this current is positive.

In many CMOS technologies there is no NPN transistor available for such a circuit. In this case the voltage difference $V_{bg} - V_{be}$ has to be generated in a different way. Here comes a proposal how to do it:

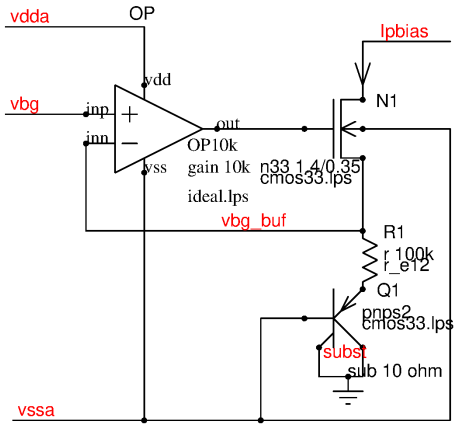


Figure 7.50: Vt over R generator without NPN transistor

The amplifier OP and N1 create a replica of the bandgap voltage. The voltage drop over R1 again is the bandgp voltage minus Vbe.

$$I_{pbias} = \frac{V_{bg} - V_{be}}{R1}$$

In a certain way this circuit takes us to a hen and egg problem: What biases the operational amplifier?

7.5.4 Vref over R current generator

Very simple - provided the reference is already there.

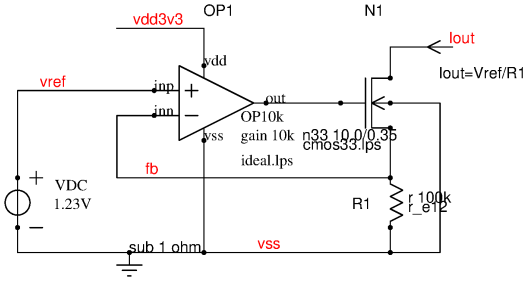


Figure 7.51: Vref over R current generator

7.5.5 NPN only ring current generator

The ring current generator provides about Vt over R as an output current.

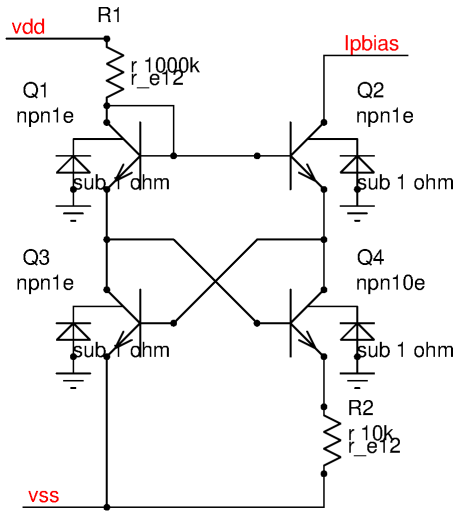


Figure 7.52: NPN ring current generator

In the NPN ring current generator the two transistors Q1 and Q2 enforce about the same operating point at the collectors of Q3 and Q4 and at the bases of Q3 and Q4. Since Q3 and Q4 have an emitter size ratio of n (in the

drawn circuit n=10) the voltage drop over R2 is

$$V_{R2} \approx V_t * ln(n)$$

The collector current of Q4 and Q2 becomes about

$$I_{pbias} \approx \frac{V_t * ln(n)}{R2} \tag{7.137}$$

(There are some error contributions such as the base currents of Q3 and Q4 that add to the current. But since usually the NPN transistors have a gain in the range of 100 or more this error is small compared to the spread of the resistors)

In the operating point the loop gain becomes 1. So the circuit is stable.

If Q1 and Q2 are replaced by MOS transistors the loop gain in the operating point increases and stability has to be checked carefully.

The base currents of Q3 and Q4 create an important start condition. Replacing Q3 and Q4 by NMOS transistors adds two more weak stable operating points to the design: fully ON and fully OFF. This means the circuit won't start in a reliable way anymore!

7.5.6 MOS ring current generator

The following circuit works reasonably well with bipolar transistors as well as with MOS transistors. Take care about the starter circuit. Without the starter it may not start correctly.

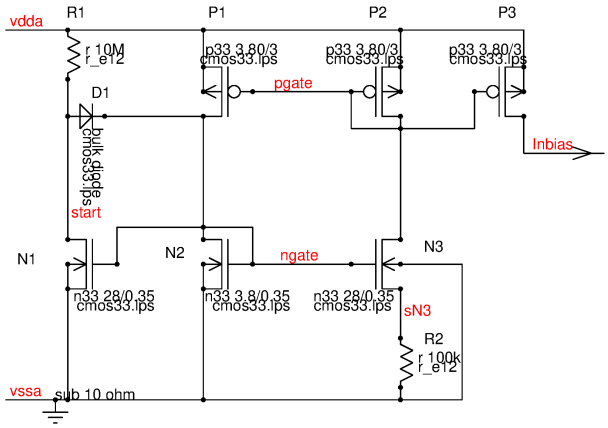


Figure 7.53: MOS ring current generator

Since N3 operates at a lower current density than N2 it has a lower Vgs. The difference of the gate source voltage of N2 and N3 together with R2 determines the current flow through P2 and the output current Inbias flowing through P3.

If N2 and N3 operate in weak inversion and mirror P1, P2 has a ratio of 1 the current calculates as

$$I_{nbias} = \frac{n * V_t}{R2} * ln(\frac{W_3}{W_2}) \tag{7.138}$$

n depends on the gate capacity and the bulk capacity. Usually $n = 1.2..1.6$.

To start the ring current generator R1 and diode D1 provide a start up current. Once the ring current generator is running N1 pulls down node "start" and diode D1 turns off. In the operating point the loop current gain is 1.

If N2 and N3 operate in strong inversion the calculation of the voltage drop over R2 becomes more difficult. Using a simulator in fact is the best way of determining the currents then.

7.5.7 Weak inversion current generator

Weak inversion current generators are a cheap way to produce small pull up or pull down currents in the nA to 500nA range. The basic idea is to operate a MOS transistor exactly at the threshold. Using the pure equation for the quadratic characteristic of the strong inversion this is the operating point the transistor (according to the strict strong inversion calculation) has a current of 0. In reality at this operating point the transistor still operates in weak inversion. The current produced there has a lot of spread but it never fully disapears. One possible implementation of such a weak inversion current sink is shown in the next figure.

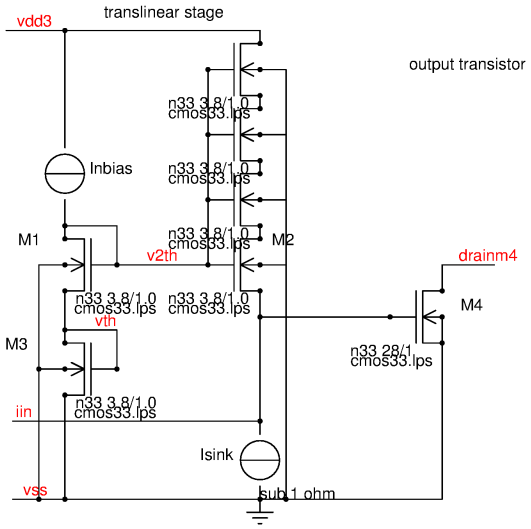


Figure 7.54: Weak inversion current generator

Implementing the circuit in a technology that offers monte carlo simulation parameters and model levels higher than 1 (usually BSIM3 or better) the distribution of such a current generator looks like this:

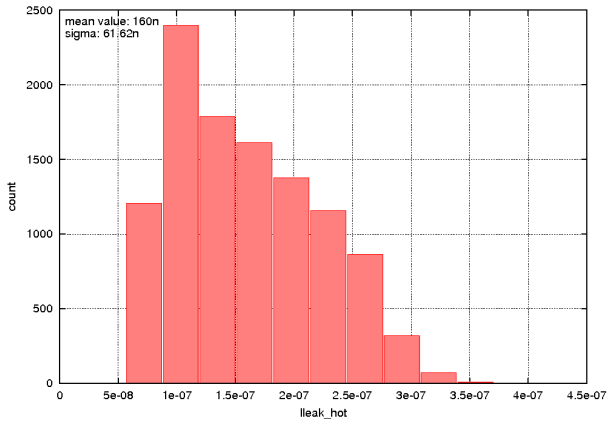


Figure 7.55: Distribution of a weak inversion current generator

In the linear scale the spread looks very asymmetrical. The reason is the exponential change of the current with the spread of Vth. So for upward deviations the distribution looks stretched while for downward deviations the distribution gets compressed. Rescaling the X-axis logarithmically the distribution starts to look much more normal again.

7.6 Oscillators

Oscillators are used to create AC signals (for instance to be used in charge pumps) and to clock all kinds of sequential logic. In RF systems oscillators and mixers often are used to shift signals into an other frequency domain. In analog systems clock signals often are needed for chopping and auto zero amplifiers.

Frequency stability requirements and cost constraints can vary a lot from system to system.

7.6.1 Phase shift oscillators

All linear oscillators are based on a feedback path that shifts the phase with frequency and that has a certain amplitude characteristic. The feedback condition was first formulated by Heinrich Barkhausen in 1921 (Barkhausen stability criterion). For a stable oscillation the Barkhausen criterion says, the sum of the phase shift in the amplifier and in the feedback path must be 0 (or 2*π).

If at that frequency the gain of the complete loop (losses of the network multiplied with the gain of the oscillator) is above 1 the circuit will oscillate.

If at that frequency the gain is below 1 the circuit will not oscillate.

The frequency stability of an oscillator is determined by the steepness of the transition (dφ/ df) and the change of the propagation delay of the amplifier with temperature, process spread, supply voltage. To build an oscillator with good stability a fast amplifier and a feedback path with a steep phase change is required. The following figure shows a conceptual example.

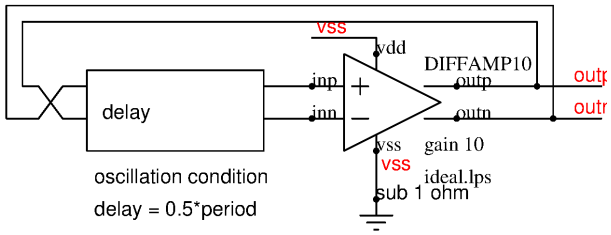


Figure 7.56: Most simple concept of an oscillator

The feedback path is crossed. So the amplifier works as an inverting amplifier. The amplifier has a phase shift of π (or 180 degrees) plus it's propagation delay. So the feedback network must have a phase shift slightly below π . The inverting amplifier offers the advantage that the inversion can be used to stabilize the DC operating point. In addition an inverting amplifier can easily be built by a simple inverter or a single transistor.

Before having a detailed look at the oscillators themselves it is worth while looking at the most useful feedback paths and their phase behavior. The following figure shows the most frequently used feedback topologies.

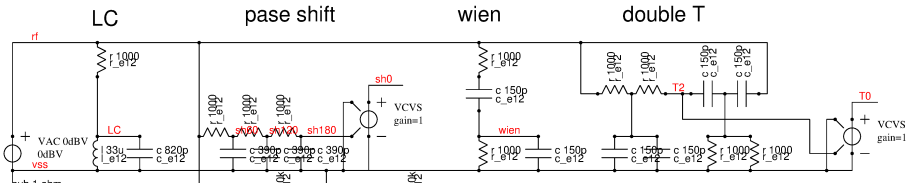


Figure 7.57: A test bench for the most frequently used feedback networks

What is interesting us most is the phase versus frequency behavior.

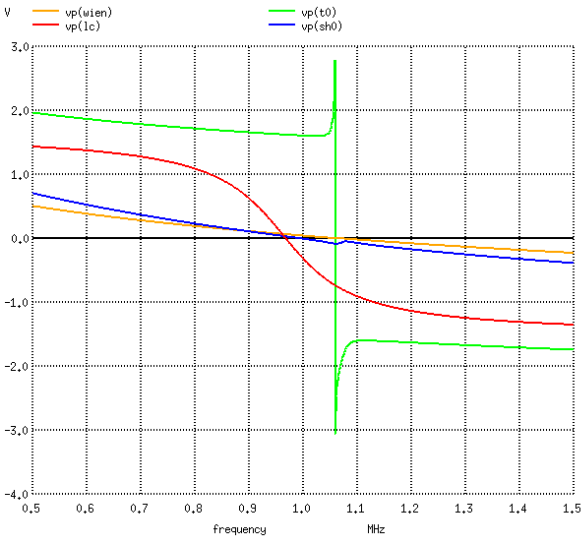


Figure 7.58: Phase versus frequency of the LC resonator, the phase shift circuit, the Wien network and the double T filter

The strange dip of the phase shift exactly where the double T filter has it's phase jump indicates there is something numeric happening in the simulation. The amplitude plot will show it.

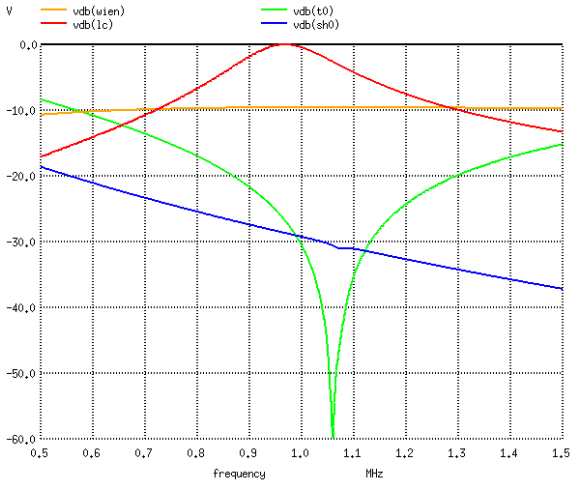


Figure 7.59: Transfer functions scaled in dB of the analyzed networks

Caught!. The double T filter is a notch. The output signal becomes zero at the phase jump. (Simulators don't like divide by zero. This is the reason of the strange looking phase of the phase shift network.) This notch behavior makes using the double T for an oscillator a bit difficult.

The Wien bridge oscillator [77] can be regarded as a differential stage as well.

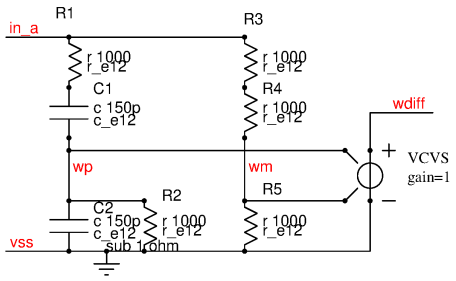


Figure 7.60: Wien bridge network with perfect tuning regarded as a differential network

At the resonant frequency the phase of both signals, wp and wm is equal. At wdiff the phase crosses 0 degrees. The change of the phase versus frequency depends on the correctness of the resistor ratio of R3 to R5. If the ration is exactly 3 the change of the phase is abrupt. The following plot shows phase and amplitude rot R5=1000Ω and for R=900Ω .

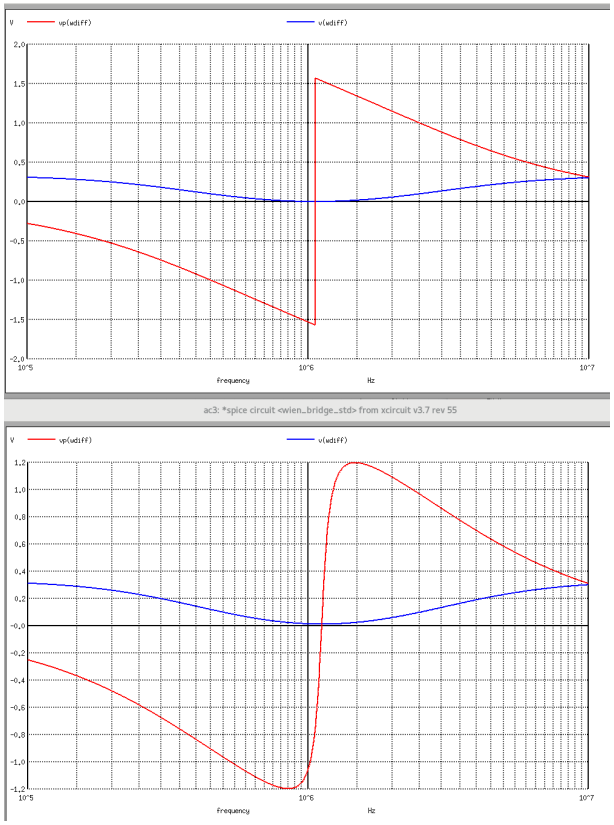


Figure 7.61: AC Transfer functions for R5=1000Ω (top) and R5=900Ω (bottom)

To build a good Wien bridge oscillator the resistor ratio of divider R3, R4, R5 must be as close to 3 as possible.

One transistor phase shift oscillator

A phase shift oscillator usually consists of an odd number of inverting amplifier stages and a delay network providing an additional phase shift of 180 degrees to satisfy the oscillation condition. Early descriptions of phase shift oscillators can for instance be found in [8] on page 279 (Well, I guess there are even older ones to be found).

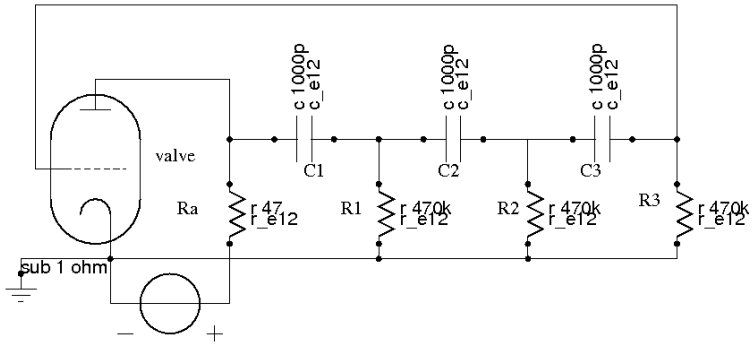


Figure 7.62: Phase shift oscillator built with a valve

Of course this kind of oscillator can just as well be built using MOS transistors. Here it is drawn with a valve because the original schematic was done with a valve too.

Normally the resistors R1 to R3 are choose equal. The same applies to the capacitors C1 to C3.

$$R1 = R2 = R3 = R$$

$$C1 = C2 = C3 = C$$

The oscillation condition is satisfied at:

$$\omega = \frac{1}{\sqrt{6} * R * C} \tag{7.139}$$

A possible solution to integrate this circuit on a chip could look like this:

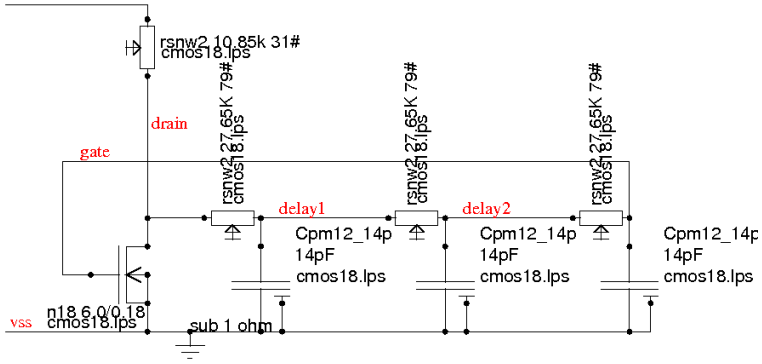


Figure 7.63: Phase shift oscillator with low pass

Now the capacitors and resistors have changed places. The solution of the oscillation equation becomes [9]

$$\omega = \sqrt{6}/RC \tag{7.140}$$

Ring oscillators

Even more area efficient for integration are ring oscillators. Here the resistors are replaced by the output impedance of the inverters. Since the output impedance is a function of the early voltage of the transistors and the current flowing through the inverter in the small signal operating point the frequency of the oscillation is technology dependent and varies with temperature. So these ring oscillators are typically used as VCOs - voltage controlled oscillators - (or better CCOs - current controlled oscillators) in PLL systems (phase locked loop systems).

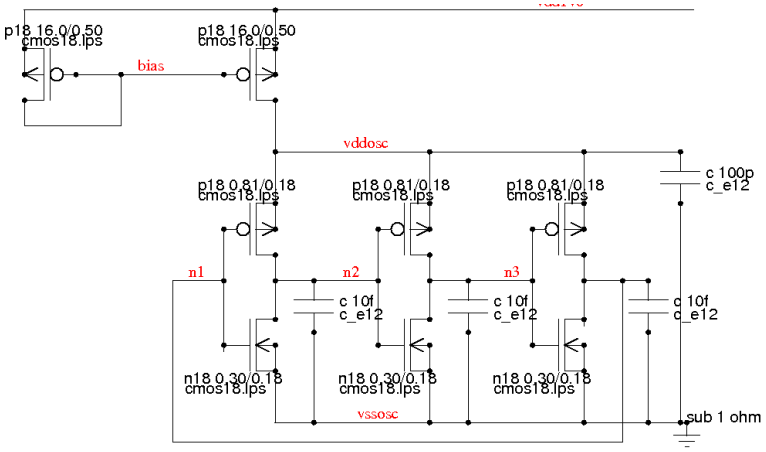


Figure 7.64: ring oscillator using 3 inverters

Using a 1.8V technology with about 4nm gate oxide such oscillators can operate up to about 500MHz reliably. Typical simulations (room temperature, typical models, bias current at net bias of about 0.1mA) show oscillation up to about 2GHz. Since this oscillator implementation is very technology dependent this circuit is one of the most frequently used structures to characterize technology parameters (especially speed of a technology) on test chips.

Besides the implementation using CMOS inverters further implementations using current mode logic stages (CML), emitter coupled logic (ECL) or fully differential designs using differential amplifier stages in stead of inverters can be found as well. These topologies offer better control of the parameters but the speed achieved is not higher than using CMOS inverters. See [10] pages 63ff. The reason is that the CMOS inverter offers the highest gain per current. Small signal gain of a CMOS inverter operated as an operational transconductance amplifier (OTA) is:

$$gm = gmpmos + gmnmos \tag{7.141}$$

So the (complex) voltage gain becomes:

$$gain = gm * Zload \tag{7.142}$$

The gain bandwidth product of one stage becomes

$$GBW = \frac{gm}{2 * \pi * Cload} \tag{7.143}$$

Cload consists of the sum of the wire capacities, the gate-source capacities of the following stage and double the gate-drain capacities of the following stage.

Most other topologies like CML, ECL and most differential amplifier stages only use the gm of one transistor type (mainly the NMOS in stead of taking benefit of the gain of both transistor types. This leads to a lower gain bandwidth product (GBW) than using CMOS inverters at the same current. (Typically half of the GBW of a CMOS gate.)

The level shift problem Ring oscillators running in current starved mode will have a different supply voltage and signal levels than the logic driven by them. The classic approach done as long as current consumption is not the limiting factor is interfacing the oscillator and the logic with a single phase level shift. The input stage of the level shift is running with a replica of the oscillator voltage. It should not be supplied by the oscillator voltage itself because in this case the current consumption of the level shift would affect the operation of the oscillator.

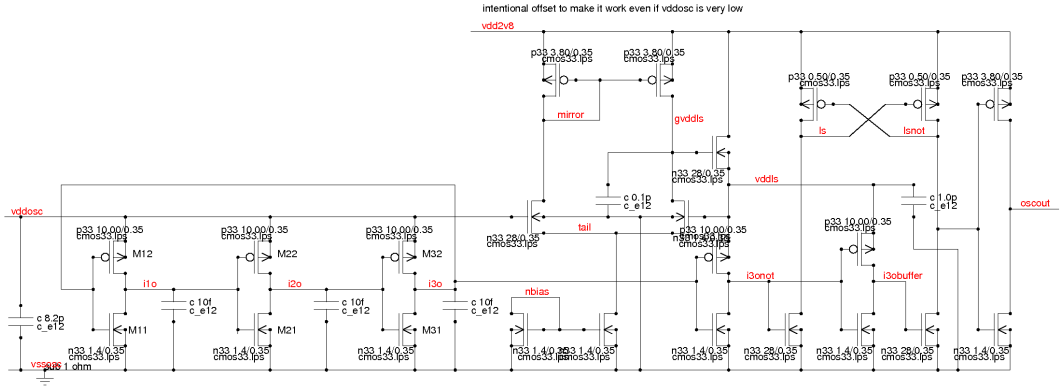


Figure 7.65: ring oscillator with level shift attached

This approach looks straight forward. Nevertheless it has some issues:

1. if the oscillator is running at a low frequency vddosc (it is driven by a current source!) can become very low. As a consequence level shift NMOS transistors driving signals ls and lsnot are operated with a very little gate overdrive. But they still have to be stronger than the PMOS latch to successfully switch the level shift. So their W/L must be designed much bigger than the PMOS W/L.
2. On the other hand the PMOS transistors can not be made too short because this would lead to very slow rising edges of the level shift stage.
3. These two limitations usually lead to tremendous cross current in the level shift stage.

The energy loss caused by the cross conduction when the NMOS transistors have to overdrive the PMOS transistors can easily become the dominant current consumption of the whole oscillator stage. Besides that the spikes are a significant source of RF emission.

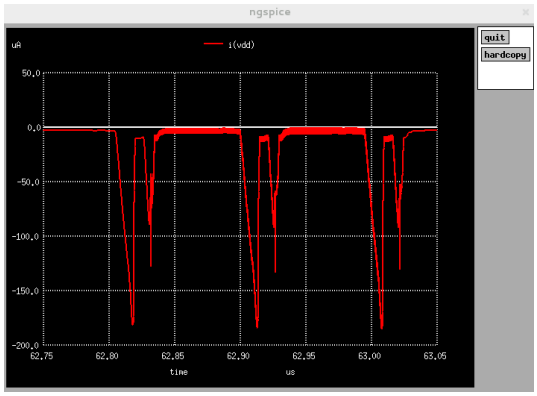


Figure 7.66: simulated current consumption of a ring oscillator together with the level shift stage. The ring oscillator is running with a bias current of 1.4μA. The spikes are produced by the level shift stage.

7.6.2 LC oscillators

An LC oscillator consists of a resonant tank and an amplifier. On chip only a few pH can be integrated. For this reason LC oscillators often use external inductors.

Armstrong oscillator: Using voltage transformation of the resonant tank may be included in the feedback path. The probably oldest LC oscillator is the Armstrong or Meisner oscillator named after the inventors. Armstrong and Meisner both invented this circuit independently in the years 1912 (Armstrong) and 1913 (Meisner).

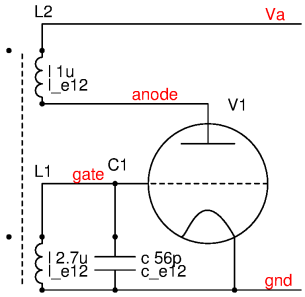


Figure 7.67: Armstrong (or Meisner) oscillator

The Armstrong oscillator requires a transformer in the feedback path. It was a common topology in the early days of radio. The drawback is the transformer. Two windings to close the feedback path simply are more expensive than just one simple inductor. Sometimes this kind of oscillator is built involuntarily due to parasitic inductors!

Clapp oscillator: The Clapp oscillator was published 1948 by James Kilton Clapp. The topology however is older. It already was used by Geoffry George Gouriet who worked for the BBC in 1938. The design only requires a voltage follower stage (unity gain) with low output impedance. Today an important advantage compared to the Armstrong oscillator is that this design only requires a single inductor. There is no transformer feedback creating additional cost for a complex inductive component.

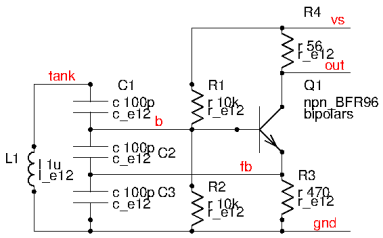


Figure 7.68: Clapp oscillator

A nice feature of the Clap oscillator is that the RF signal can be picked up at R4 at the collector of the transistor. This way the oscillator becomes fairly insensitive to the properties of the next stage.

Colpitts oscillator: the Colpitts oscillator was invented by Edwin H. Colpitts in 1918. It uses an amplifier with voltage gain and closes the feedback loop with a capacitive tap. The amplifier doesn't require a current gain. Therefore a grounded base circuit can be used as an amplifier. The transistor can be operated above it's transit frequency using a grounded base circuit. This opens the door to very high frequencies. The Colpitts oscillator often is used up to the GHz range.

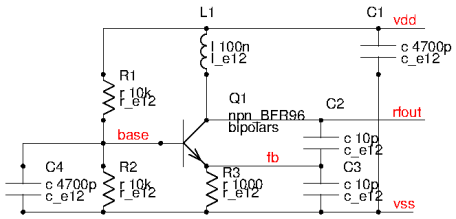


Figure 7.69: Colpitts oscillator

Hartley oscillator: The Hartley oscillator follows the same concept as the Colpitts oscillator but taps the inductance instead of a capacitive divider.

7.6.3 Wien oscillator

The Wien oscillator offers the advantage (compared to phase shift oscillators) that only two components need to be tuned synchronously to change the frequency. For this reason it is frequently used as a sine wave oscillator for low

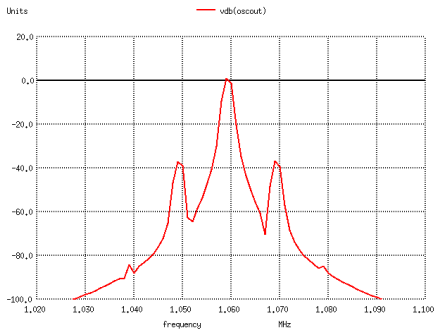


Figure 7.72: Spectrum of the same oscillator with a 10kHz noise source in the feedback path

cost test equipment. (For high frequency stability in the AF region today digital synthesizers with a quartz reference are used.). For extremely low distortion HiFi tests the Wien oscillator still has it's place because in contrast to digital synthesis there is no quantization noise in the spectrum.

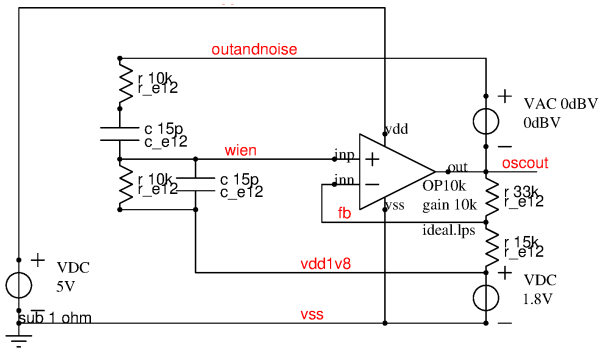


Figure 7.70: Wien oscillator

The Wien oscillator is extremely sensitive to the correct adjustment of the gain. In practical applications the resistive feedback network is part of an amplitude regulation loop. Above that the slightest clipping of the amplifier becomes visible immediately.

The difficulty of regulating the gain in a linear way is one of the reasons why the Wien oscillator is barely used on chips.

The AC source in the feedback path was added to simulate the sensitivity of the oscillator to out of band noise added. This is a typical test to estimate how well the feedback network filters signals that are not at the desired operating frequency. In most cases the noise signal will become visible as side bands of the spectrum of the oscillator.

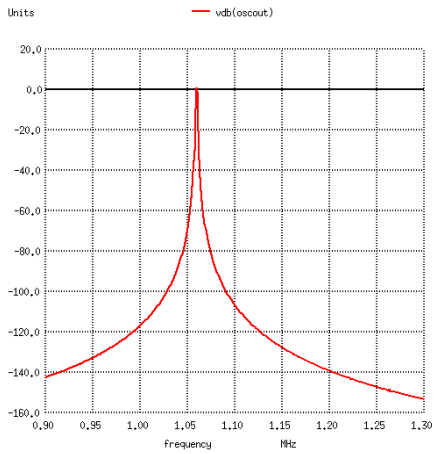


Figure 7.71: Spectrum of the oscillator without noise

Turning on the noise source with a frequency of 10kHz leads to the following distorted spectrum with side bands 10kHz away from the nominal oscillation frequency.

7.6.4 Crystal oscillators

Almost every oscillator that can be built with an LC tank can also be implemented with a quartz. A quartz is a mechanical resonator that has multiple oscillation modes.

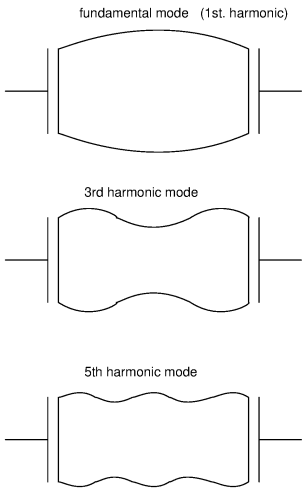


Figure 7.73: Mechanical oscillation modes of a quartz

Due to the various oscillation modes of a quartz the impedance of a quartz has a periodic pattern. Basically it looks like this:

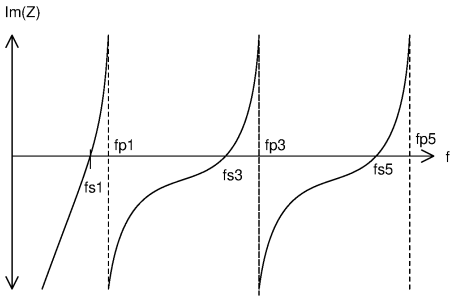


Figure 7.74: Imaginary part of the impedance of a quartz

The quartz can either be operated at the serial resonance (fs) or at the parallel resonance (fp). The serial resonance and the parallel resonance usually only differ by a few kHz. Most quartz manufacturers specify the serial resonance and optimize the quartz for operation at the serial resonance. At which resonance (mode) the quartz will operate is usually determined by additional filters in the circuit.

Inverter quartz oscillator: If no special measures are taken the the frequency characteristic of the amplifier used in the quartz oscillator will in most cases enforce operation at the fundamental (fs1 or fp1) frequency. A typical example is the classical inverter oscillator found in most digital chips.

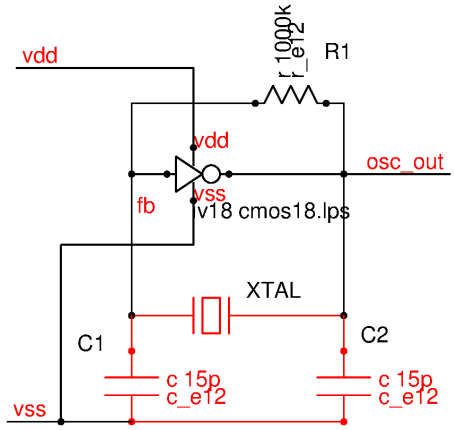


Figure 7.75: Inverter oscillator

In the simple inverter oscillator the low pass created by the inverter output impedance and the capacitor C2 provides a gain decay with increasing frequency. This forces the oscillator into the fundamental mode.

At the resonance frequency a high current flows in the red colored loop. This resonant current is magnitudes higher than the bias current flowing through the inverter. For EMC reasons the red colored loop must be designed as compact as possible.

Clapp quartz oscillator: The inverter oscillator needs two pins. If a one pin oscillator is needed the quartz can be operated in parallel resonance. In this case the Clapp oscillator is a possible choice.

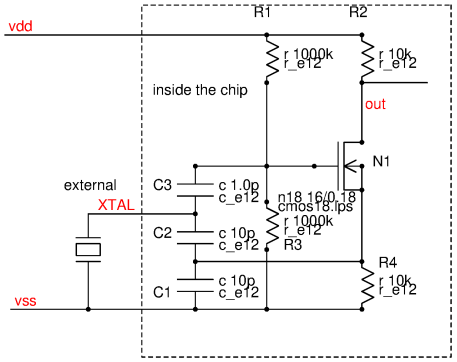


Figure 7.76: Clapp quartz oscillator

Operating a quartz in parallel resonance must be done with precaution. Some performance loss must be accepted!

- The parallel resonance frequency deviates from the frequency specified by the quartz manufacturer.
- node XTAL is a high impedance node. This makes it very sensitive to noise coupling of adjacent pins.
- since XTAL has a high impedance exactly at the operating frequency the resistive noise of R1 and R3 becomes part of the oscillator signal.
- Scaling the circuit for low supply voltage worsens the signal to noise ratio.

Quartz oscillators operating at higher harmonics To force oscillation at higher harmonics the lower harmonics must be suppressed by filters. Usually this is done designing an LC oscillator using the quartz in the feedback path. The LC filter determines which harmonic is used and the quartz in the feedback path provides the frequency stability.

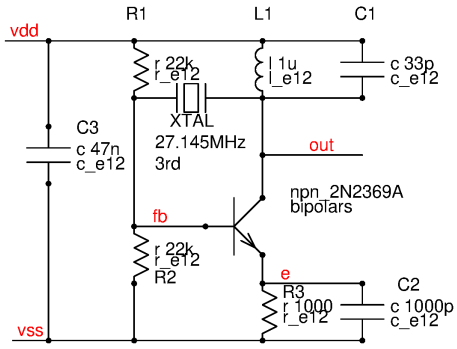


Figure 7.77: A typical oscillator using a quartz at the 3rd harmonic

In the circuit shown the LC tank L1 and C1 is tuned to the 3rd harmonic of the quartz. Oscillation at the first and the fifth harmonic is suppressed.

Amplitude regulation: Modern SMD quartzes only permit a limited level of RF power. Most quartz manufacturers recommend limiting the RF amplitude to about 1V. Exceeding this level for a long time (several hours) may burn the contacts of the quartz or lead to mechanical damage. For this reason modern oscillators usually have an amplitude regulation. If the amplitude regulation is unstable or tends to ringing the amplitude regulation may lead to unexpected side bands.

Exceeding the recommended limit for short time (seconds to minutes) usually won't harm the quartz.

To speed up start up of the oscillator sometimes the amplitude regulation is inactive for the first millisecond after power up.

Older quartz designs are less sensitive to overloading. But today these old designs are not in use for new production anymore. If you use designs of old books (for instance [36] pages 266 and 267) be aware that you have to use the old bulky but robust quartzes of the 1960s.

7.6.5 Relaxation oscillators

Most ring oscillators strongly depend on technology parameters such as gate propagation delays.

LC oscillators suffer from the fact that the inductor can not be integrated (unless the frequency is in the GHz range). Replacing the inductors by capacitors and gyrators brings back the problem of spread because the gyrator gains are temperature dependent.

Crystal oscillators are not well appreciated by customers because crystals cost money.

So the relaxation oscillators try to solve the problem of acceptable accuracy at low cost providing circuits that only depend on very few well controllable components.

Astable multivibrator: The probably best know relaxation oscillator simply consists of only two transistors. This used to be a bread and butter oscillator of the 1960s when transistors still were expensive.

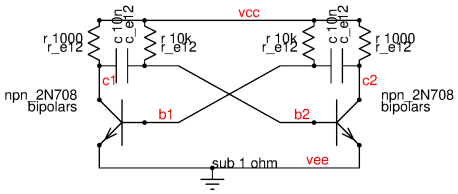


Figure 7.78: Simple relaxation oscillator

The oscillator frequency is determined by the 10K base resistors and the two capacitors. The voltage swing at the collectors runs from the saturation voltage of the transistors (V_{sat}) to the supply voltage (V_{cc}). The base voltage is limited by the base-emitter diode. The highest voltage at the base is V_{be}. The voltage swing at the base is the same as at the collectors. So the minimum voltage at the base becomes:

$$V_{bemin} = V_{be} - V_{cc} + V_{sat} \tag{7.144}$$

The voltage drop across the 10K resistors ranges from

$$V_{rmin} = V_{cc} - V_{be} \tag{7.145}$$

$$V_{rmax} = V_{cc} - V_{bemin} = 2 * V_{cc} - V_{be} - V_{sat} \tag{7.146}$$

The following simulation shows the signals.

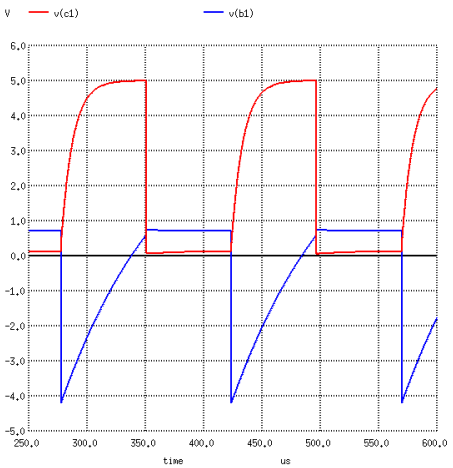


Figure 7.79: Signals of the 2 transistor relaxation oscillator

Now we can calculate the time needed to charge the capacitor from one voltage level to the other (This calculation neglects the base currents and the turn off delay of the transistors!).

$$t_{charge} = R * C * ln(\frac{2 * V_{cc} - V_{be} - V_{sat}}{V_{cc} - V_{be}}) \tag{7.147}$$

The oscillator frequency becomes:

$$f_{osc} = \frac{1}{2 * t_{charge}} = \frac{1}{2 * R * C * ln(\frac{2*V_{cc}-V_{be}-V_{sat}}{V_{cc}-V_{be}})} \tag{7.148}$$

In many books the saturation voltage and the base emitter voltage are neglected. For this reason in most of the books you will find:

$$f_{osc} \approx \frac{1}{2 * ln(2) * R * C} \approx \frac{0.7}{R * C} \tag{7.149}$$

The circuit suffers from several issues:

- The negative peaks at the base limit the supply voltage to about 5V because most transistors have a base-emitter break down of -7V
- The rising edge of the collector voltage is limited by the collector resistors that have to charge the capacitors
- The transistors operate in saturation. This leads to t turn off delay of the bipolar transistors
- Due to the turn off delay the frequency deviates significantly if the circuit is used in the MHz range
- In most technologies turn off delay of saturated bipolar transistors isn't modeled correctly
- The base currents of the transistors slightly change the frequency of the oscillation

In addition this oscillator has the risk of not starting! There is one weak stable operating point: Both transistors are fully on and since they are in saturation the loop gain collapses (This can happen if the base resistors are too low resistive or if the current gain is higher than anticipated). This simple circuit must be verified thoroughly to be sure it starts and operates correctly in all process corners.

ECL multivibrator: For high frequencies the following variant of the multi vibrator works better. In this circuit the currents and the voltage swing are limited by the currents sinks I1 and I2.

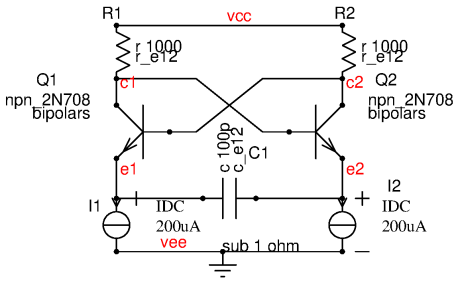


Figure 7.80: ECL relaxation oscillator

The frequency is determined by the capacitor and the resistors R1 and R2. The voltage swing at the resistor is limited by the two current sinks. The current must be limited in a way that the amplitude at the collector remains below Vbe to prevent saturation.

$$V_{dropR} = R_{12} * (I_1 + I_2) < V_{be} - V_{sat} \tag{7.150}$$

Typically the currents are scaled for a voltage swing of 200mV to 400mV. The charge time of the capacitor becomes

$$t_{charge} = V_{dropR} * C_1 / I_{12} \tag{7.151}$$

Since the voltage drop follows the current the frequency becomes:

$$f_{osc} = \frac{1}{4 * C_1 * R_{12}} \tag{7.152}$$

A variation of the current sinks changes the voltage drop over the resistors and at the same time the charge current of the capacitor. This way the currents cancel in the equation.

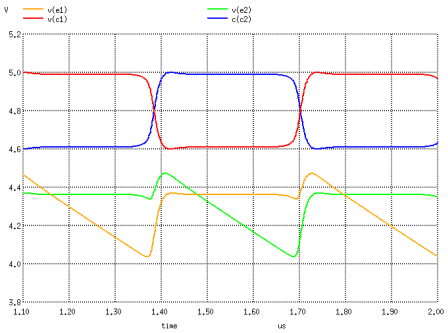


Figure 7.81: Signals of the ECL multivibrator

The little kicks up at each transistor turn off already shows that the practical swing is slightly higher and the real frequency achieved is a little bit lower than the calculated one. So this oscillator isn't too accurate either. The kick up is caused by the charges stored in the transistor before turn off. Nevertheless the achievable high frequency makes it attractive for designing PLLs.

To tune the oscillator either the capacitor or the resistors must be tuned. For high frequency applications replacing the capacitor by a varactor to tune the frequency is a reasonable option. For low frequencies tuning the resistors is preferred. The tuning range however is limited because the bipolar transistors should not be driven into saturation.

Schmitt Trigger oscillator: The Schmitt trigger oscillator avoids the risk of not starting of the simple 2 transistor multivibrator. There are many variants of using Schmitt triggers for oscillators. Basically most of them consist of capacitors that are charged or discharged between two levels by a resistor or a current source. Some nice circuits are shown in [14].

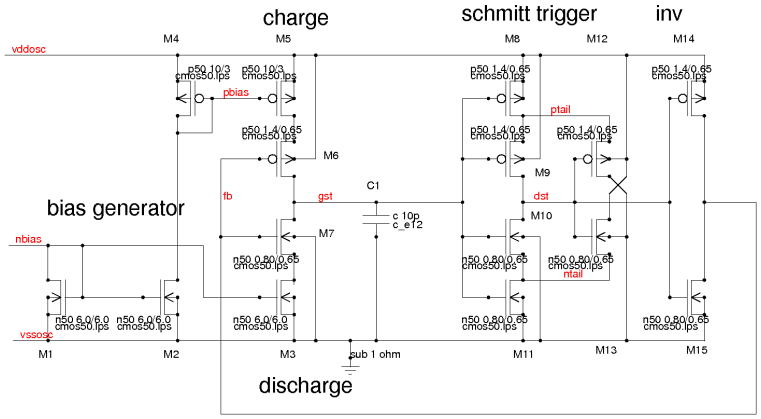


Figure 7.82: One example of the oscillator implementations proposed by [14]

Neglecting the propagation time of the Schmitt trigger the frequency of this oscillator is a function of the hysteresis Vhyst, the bias current Ibias and the capacity C1.

$$f_{osc} = \frac{I_{bias}}{2 * C_1 * V_{hyst}} \tag{7.153}$$

The problem is hidden in Vhyst. The hysteresis of the Schmitt trigger depends on the supply voltage applied at pin vddosc and on the properties of the transistors M8 to M13. Since 3 of these transistors are PMOS and the other three are NMOS (that do not match!) there is a lot of production spread in the design even if the supply voltage vddosc is kept constant.

2 Capacitor oscillators: Mansour Izadinia and Tamas Szepesi show a better solution getting rid of the hysteresis [15]. The price of this solution is that the duty cycle of the oscillator varies with the change of the thresholds of the PMOS and NMOS transistors. The frequency however can be made very stable. The current used must be proportional to the supply voltage. Alternatively the current generators can be replaced by resistors. In this case the cancellation is supply variations is not perfect but still quite good.

The idea of this circuit is that the sum of the charge times of C1 and C2 is constant no matter what is the trip point of the inverter M13, M14. If the trip point is close to vssosc the charge time of C2 gets shorter while the charge time of C1 gets longer by the same amount of time. If the trip pint is close to vddosc charge time of C1 decreases and the charge time of C2 increases. This works well as long as the currents flowing through M6 and M4

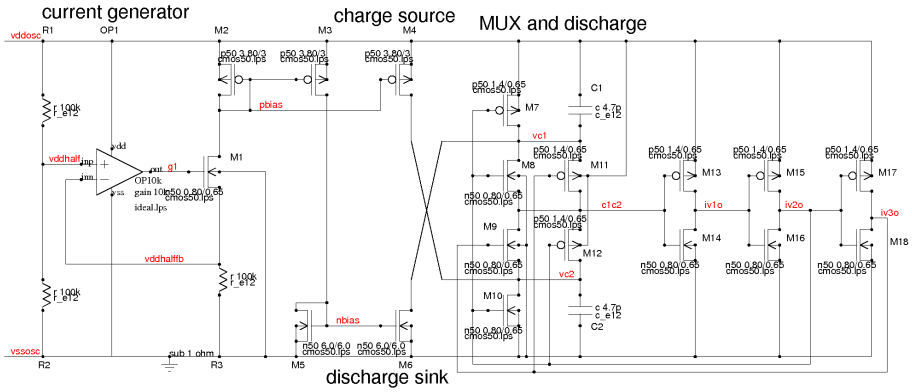


Figure 7.83: Precision oscillator avoiding hysteresis impact on the frequency

are equal and C1=C2=C. Compared to the current flowing in M6 and M4 the resistances of M10 and M7 should be low. (The circuit can be further improved placing additional switches in series with M4 and M6.)

$$f = \frac{R_2}{(R_1 + R_2) * C * R_3} \tag{7.154}$$

As long as V(c1c2) is below the trip point of M13, M14 signal iv1o is logic 1 turning on M9, M12 and M7. C2 is getting charged by M4 while C1 is shorted by M7. Reaching the trip point the multiplexer switches to C1 turning on M8 and M11. M7 turns off and M10 turns on. Now C2 is shorted and C1 gets pulled down by M6.

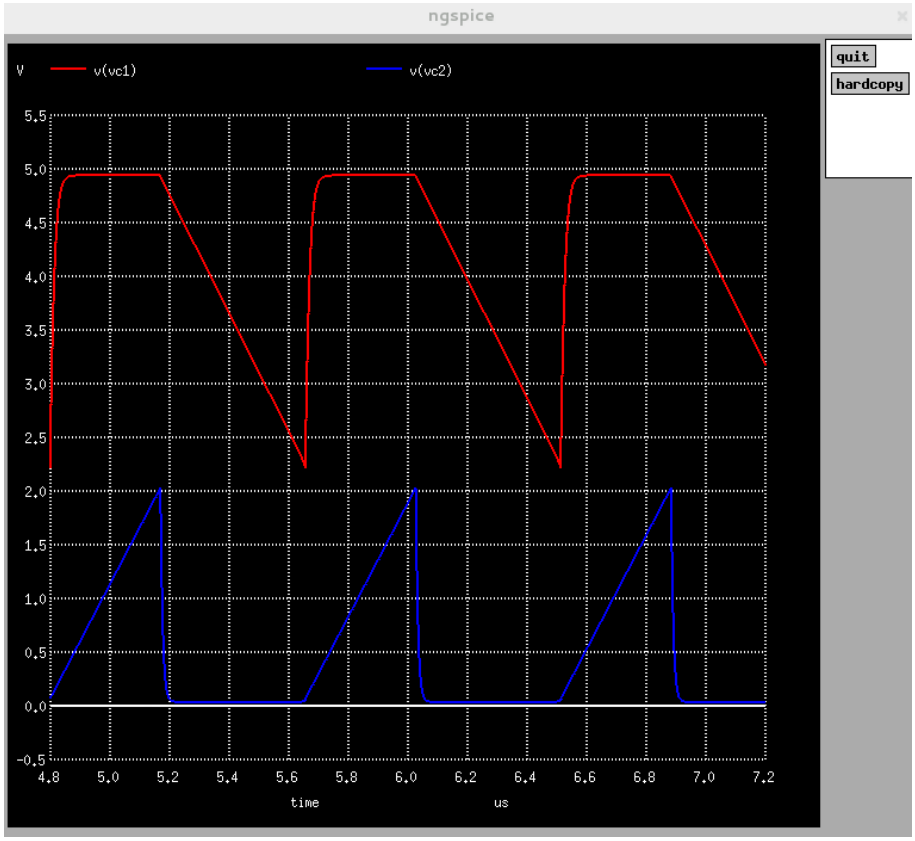


Figure 7.84: Signals at the two capacitors of the oscillator

The inverter used to evaluate the voltage at the timer capacitors is the fastest solution possible (because we add the transconductance of the NMOS transistor M9 and PMOS transistor M8 with a very high current flowing at the trip point). Furthermore the transistor can be made very small without affecting the frequency. So this is probably the relaxation oscillator with the least possible error caused by gate delays and comparator delays. Nevertheless the delays are visible looking at the overshot of the saw tooth signals.

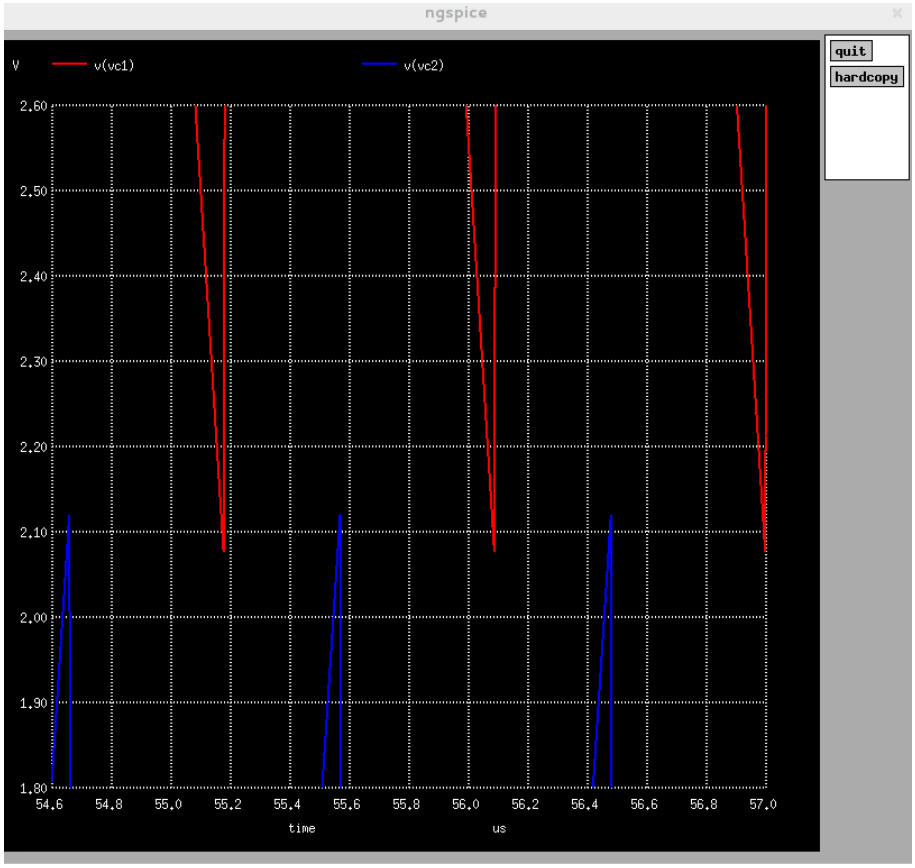


Figure 7.85: Zoom into the trip points. the sum of the overshots is about 20mV

Compared to a signal swing of 5V (both triangles) the 20mV correspond an error of 0.4%.

Calculation of the frequency error caused by the delay of the first amplifier stage The first amplifier stage in the case shown above is a simple inverter consisting of M13 and M14. It has to drive the capacities associated with the node iv1o. These are the drain capacities of M13 and M14, the gate capacities of M15 and M16 and the wire capacities of the metal trace connecting the components. The gate capacities of M15 and M16 require special care because there is a capacity from the gates to the sources plus a capacity from the gates to the drains. Since signal iv1o has the opposite phase of the gate voltage of M13 and M14 we have to double the drain-gate-capacities of M15 and M16!

$$C_{iv1o} = C_{dM13} + C_{dM14} + C_{wire} + C_{gsM15} + C_{gsM16} + 2 * (C_{gdM15} + C_{gdM16}) \tag{7.155}$$

The voltage at the output of the inverter starts to change as soon as the input voltage exceeds the threshold of the inverter. The inverter can be linearized for small signals.

$$I_{iv1o}(t) = (V_{c1c2}(t) - V_{th}) * gm \tag{7.156}$$

In the equation above $V_{th}$is the threshold of the inverter. gm is the transconductance of the inverter exactly at the trip point (The sum of the gm of the NMOS and the gm of the PMOS). Assuming only small overshots the current $I_{iv1o}(t)$can be assumed to be more or less triangular. Assuming the inverter M15, M16 is designed in a reasonable way (threshold in the middle of the supply rails) we can estimate the time needed to charge or discharge the capacity C_{iv1o}.

$$\int I_{iv10}(t)dt = C_{iv1o} * \frac{V_{vddosc}}{2} \tag{7.157}$$

Above the trip point of M13, M14 the current available is:

$$I_{iv1o}(t) = t * \frac{dV_{c1c2}}{dt} * gm \tag{7.158}$$

with:

$$\frac{dV_{c1c2}}{dt} = f_{ideal} * V_{vddosc} \tag{7.159}$$

leading to:

$$f_{ideal} * V_{vddosc} * gm * \int tdt = C_{iv1o} * V_{vddosc}/2 \tag{7.160}$$

Solving the integral the overshot time becomes:

$$t = \sqrt{\frac{C_{iv1o}}{gm * f_{ideal}}} \tag{7.161}$$

Since the overshot takes place twice per period the total error becomes:

$$t_{error} = 2 * \sqrt{\frac{C_{iv1o}}{gm * f_{ideal}}} \tag{7.162}$$

Usually we are interested in the relative deviation of the oscillator.

$$err_{rel} = \frac{t_{error}}{T} = 2 * \sqrt{\frac{C_{iv1o} * f_{ideal}}{gm}} \tag{7.163}$$

Let's have a look at an example:
We want to build a 20MHz oscillator, have an inverter with a gm of 50μA/V and a capacity $C_{iv1o}$of 20fF.

$$err_{rel} = 17.9\%$$

This means the oscillator is running 17.9% slower than the ideally calculated frequency.
What can we conclude from this result?

1. The supply voltage cancels as long as gm is supply independent (in case of a simple inverter acting as an amplifier this is NOT the case!).
2. the transconductance gm must be made as big as possible. The structure with the highest possible gm versus bias current is the simple inverter.
3. The load of the first inverter stage C_{iv1o}must be minimized.
4. The lower we choose the (ideal) frequency the more accurate we can build the oscillator.
5. If we want to build an oscillator with low frequency drift gm must be kept constant over temperature. (So the pure inverter is NOT a good idea because gm of the inverter decreases with temperature.)
6. Fast accurate oscillator will consume a lot of current to provide a high gm of the first amplifier stage.

EMC considerations: RF superimposed on the supply will be coupled into the ramp generators M4 and M6. Therefore M4 and M6 should be designed for low Cds (compared to C1 and C2).

Low power relaxation oscillator: Building a low power oscillator high cross conduction currents as in the fast oscillator shown before can not be tolerated. Here the concept is:

1. Move as few nodes as possible
2. avoid static current consumption. Don't use standard comparators with permanently flowing tail currents
3. prevent cross conduction. Don't use logic inverters as amplifiers

One possible solution is shown in the next figure.

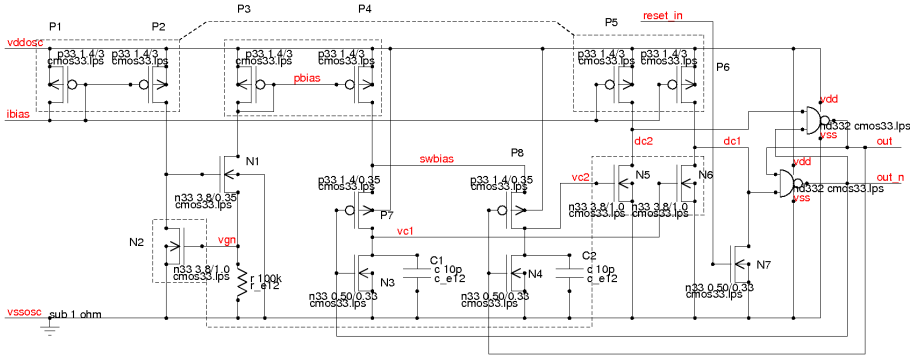


Figure 7.86: Low power oscillator

The oscillator only has four constantly flowing bias paths:

1. ibias flowing through P1.
2. The current producing the charge current flowing through P2 andN2.
3. The reference current corresponding Vth/R flowing through N1 and P3.
4. The charge current flowing through P4 either charging capacitor C1 or capacitor C2.

The currents through P5, N5 and P6, N6 only are active for s short time. Cross conduction in the logic only flows at the instant of switching the latch.

The current charging the capacitor is defined by the threshold of N1.

$$I_{charge} = V_{th}/R \tag{7.164}$$

The trip point of the amplifiers N5 and N6 correspond their thresholds when the drain current exceeds the current provided by P5 and P6. Since this is the same current N2 is operated at we get the same thresholds. So one charge time is defined by:

$$t_{charge} = C_1 * V_{th}/I_{charge} = C_2 * V_{th}/I_{charge} = C_1 * R = C_2 * R \tag{7.165}$$

(assuming C1=C2). The ideal oscillation frequency (neglecting the propagation time of the path N5, N6, logic, N3, N4, P7, P8) becomes:

$$f_{osc_{ideal}} = \frac{1}{2 * R * C} \tag{7.166}$$

Errors caused by the limited speed of N5, N6, P5,P6 can be calculated in a similar way as before.

Calculation of the oscillator error caused by delays Similar to the fast oscillator shown before the amplifier stages N5 and N6 have to discharge the capacities at the drains on N5, P5 and N6, P6. These capacities are the wiring capacity (C_{wire}), the drain capacities ($C_{d_{N56}}, C_{d_{P56}}, C_{d_{N7}}$) and the input capacities of the logic gates ($C_{in_{logic}}$).

$$C_{load} = C_{wire} + C_{d_{N6}} + C_{d_{N7}} + C_{d_{P6}} + C_{in_{logic}} \tag{7.167}$$

Usually these capacities are in the range of some 10fF.

Since we want to have short cross conduction times inside the logic N5 and N6 should be designed for maximum gain. This means N5 and N6 at the trip point should work in the transition zone between weak inversion and strong inversion. (Making N5, N6 bigger leads to necessary capacities, Making N5, N6 too small reduces the ration gm/Id and the switching slopes get slower due to lack of voltage gain.) The assuming weak inversion the gm of N5 and N6 can be calculated:

$$I_d = I_0 * \frac{W}{L} * exp(\frac{k * Vgs_{eff}}{V_t}) \tag{7.168}$$

$I_0$is a technology dependent factor (can be calculated knowing gate charge, carrier mobility). But since we are interested in gm we don't need it.

$$gm = \frac{dI_d}{dV_{gs}} = I_0 * \frac{W}{L} * exp(\frac{k * Vgs_{eff}}{V_t}) * \frac{k}{V_t} \tag{7.169}$$

$$gm = I_d * \frac{k}{V_t} \tag{7.170}$$

Factor k depends on the capacities between the channel and the gate and the capacity between the channel and the bulk. Typical values of k are around 0.7.

$$k = \frac{C_{gate-channel}}{C_{gate-channel} + C_{channel-bulk}} \tag{7.171}$$

So the estimation for gm becomes:

$$gm \approx 0.7 * \frac{I_d}{V_t} \tag{7.172}$$

The rest of the calculation follows the same procedure as done before at the example of the fast oscillator. The subtle difference is hidden in the amplitude of the triangular signals that now follows the threshold $V_{th_{NMOS}}$ of the NMOS transistors.

$$\frac{dV_{c1c2}}{dt} = 2 * f_{ideal} * V_{th_{NMOS}} \tag{7.173}$$

Assuming the threshold of the logic gates is in the middle of the supply rails this leads to:

$$2 * f_{ideal} * V_{th_{NMOS}} * gm * \int tdt = C_{load} * V_{vddosc}/2 \tag{7.174}$$

Solving for the delay time we get:

$$t = \sqrt{\frac{C_{load} * V_{vddosc}}{2 * f_{ideal} * V_{th} * gm}} \tag{7.175}$$

Since this delay takes place twice per oscillator period we can calculate the error time:

$$t_{error} = \sqrt{\frac{C_{load} * V_{vddosc}}{f_{ideal} * V_{th} * gm}} \tag{7.176}$$

Including the expression for the transconductance gm we get:

$$t_{error} = \sqrt{\frac{C_{load} * V_{vddosc} * V_t}{f_{ideal} * V_{th} * I_d * k}} \tag{7.177}$$

and the relative error of the frequency:

$$err_{rel} = \sqrt{\frac{C_{load} * V_{vddosc} * V_t * f_{ideal}}{V_{th} * I_d * k}} \tag{7.178}$$

Let's have a look at an example:

We want to build a 1MHz oscillator, have bias current of N5 and N6 of 1μA and a capacity C_{load} of the amplifier stages of 20fF. The gm is about 700nA/V. The threshold of the NMOS is about 600mV. For the supply of the oscillator we use 3.3V. The oscillator operates at room temperature and Vt is 26mV.

$$err_{rel} = 4.51\%$$

This means the oscillator is running 4.5% slower than the ideally calculated frequency.

What can we conclude from this result?

1. The load capacity of N5 and N6 must be minimized.
2. The supply voltage should be kept low to minimize the charge to transferred into the load of N5, N6
3. The error follows the square root of the frequency.
4. The error increases with temperature because gm decreases and Vth decreases.
5. If possible use transistors with a high threshold.
6. Scaling to smaller technologies only leads to limited improvements (Cload) because threshold usually are scaled with the supply voltage.
7. Supply the amplifier stages with a ptat current to reduce temperature drift.

EMC considerations The basic concept of the oscillator refers every signal to vssosc. Modulation of the supply voltage vddosc will change the frequency of the oscillator via the early effect of the current generators P4, P5 and P6. This applies to DC changes as well as fast changes of the supply voltage.

Fast changes of the supply voltage can propagate into the ramps of the oscillator via the Cds of P4. To make the oscillator insensitive to RF on the supply the capacity of P4, P5, P6 should be kept as low as possible. (The ratio of Cds of P4 and the oscillator capacitors C1 and C2 is decisive!). The worst case relative pulling range is:

$$\frac{\Delta f_{osc}}{f_{osc}} = \frac{V_{0p} * Cds}{V_{th} * (C + Cds)} \tag{7.179}$$

In this equation V_{0p} is the peak voltage of the RF on vddosc (measured versus vssosc). $V_{th}$is the trip point of the transistors N5 and N6. Cds is the drain-source (including drain-bulk) capacity of P4. The equation holds the assumption that C1=C2=C.

7.6.6 PLL

PLL stands for phase locked loop. Typically a phase locked loop is used to synchronize two oscillator. A classical use is a high frequency oscillator that is getting synchronized with a crystal oscillator running at a lower frequency. The phase locked loop in most cases consists of a VCO (voltage controlled oscillator), a divider, a phase comparator, a reference oscillator and an integrator providing the dominant pole of the regulation loop.

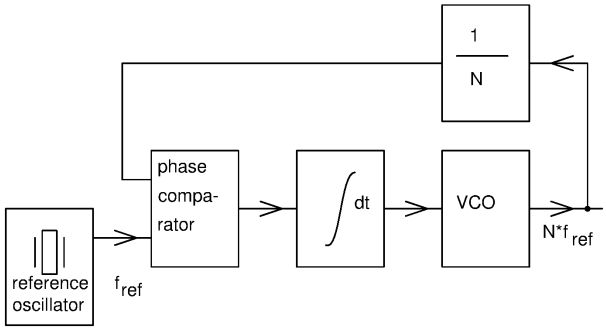


Figure 7.87: concept of a PLL (phase locked loop)

The most important part of the PLL is the phase comparator. A simple NAND gate is a fine phase comparator as long as the frequency of the reference oscillator and the divided VCO are close to each other. If the phase between the signals shifts too much the simple NAND gate acting as a phase comparator will fail and we run into a saw tooth signal at the output of the integrator. The simple NAND gate PLL will become unstable.

The simple NAND gate phase comparator on the other hand has some clear merits: It has by concept no hysteresis!

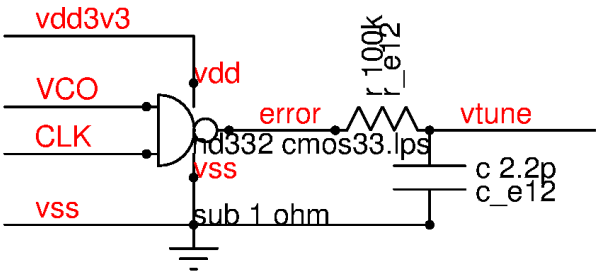


Figure 7.88: A simple NAND phase comparator

The following plot shows the behavior of the phase comparator if the two input signals V(VCO) and V(CLK) differ by 4% in frequency.

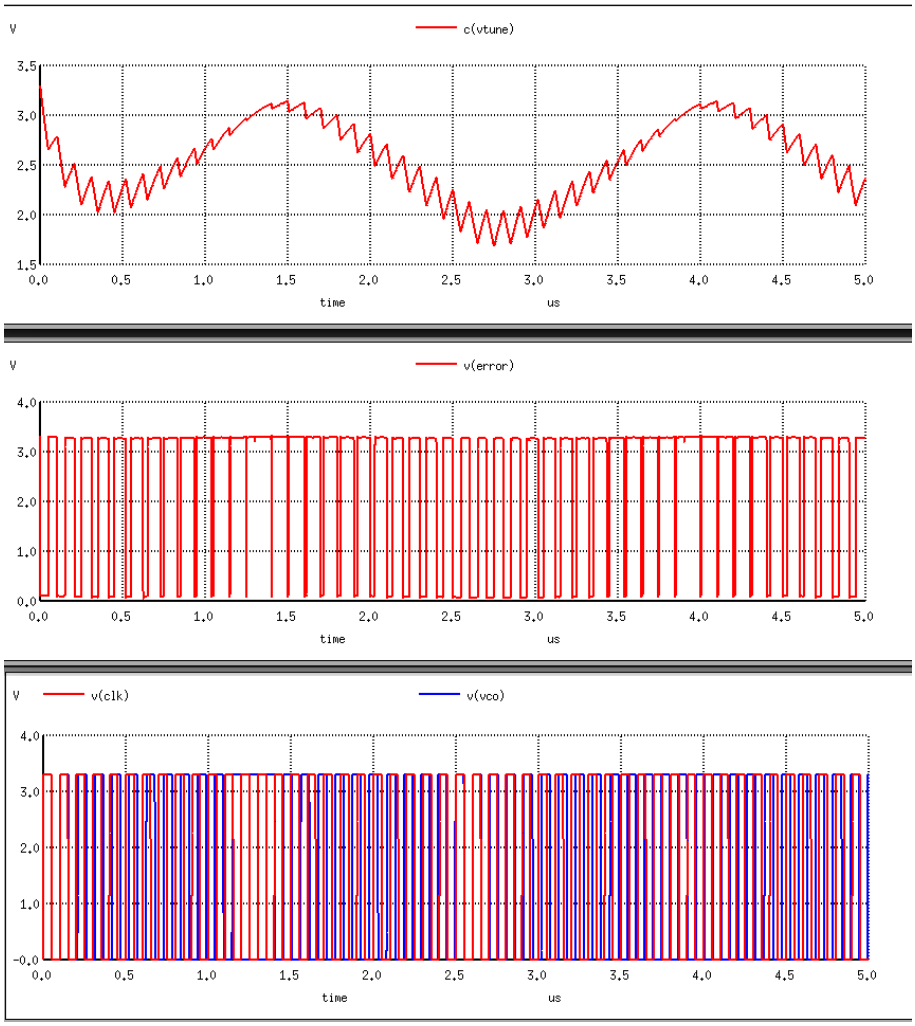


Figure 7.89: Behavior of the NAND phase comparator

If the phase between the VCO and the reference oscillator is shifted by 180 degrees the correction signal reaches its maximum of 3.3V. When the VCO and the reference oscillator are in phase the duty cycle of the error signal becomes 50% and the signal at vtune is about 1.8V. The problem of the NAND gate phase comparator is that if f(VCO) is twice the reference frequency we again get a vtune voltage close to 2V which would normally mean we are close to synchronization while in reality both oscillators are a factor 2 different in frequency.

Second problem of a NAND gate used as a phase comparator is the periodic response. If the frequency of the oscillator to be synchronized doesn't get locked within the time

$$T_{NANDPLL} < \frac{0.5}{f_1 - f_2} \tag{7.180}$$

the PLL using the simple NAND gate will just keep slipping. In the above equation f_1 and f_2 are the frequencies of both oscillators if they are left free running without any synchronization. Equation (7.181) means the integration time of the PLL filter must be less than $T_{NANDPLL}$. Due to this restriction this simple circuit is barely used in practical design. Nevertheless it is good to be aware of this simple synchronization mechanism because it sometimes happen unintentionally if several independent oscillators are placed on the same chip.

An improved version of the phase detector measures the time difference between the edges of the VCO signal and the reference oscillator using two flip flops.

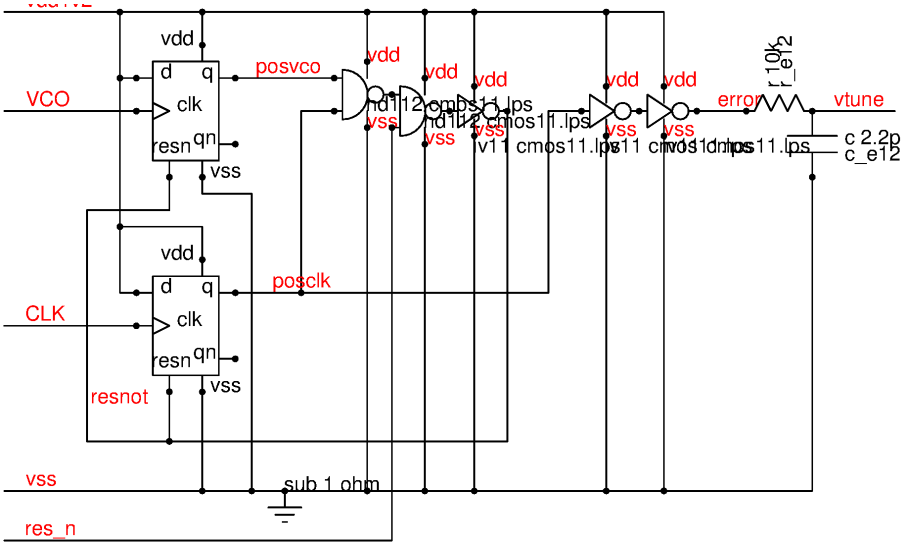


Figure 7.90: Improved phase comparator that tolerates big frequency differences

Instead of slipping over the desired operating point the output signal simply goes into saturation if the frequencies of pin VCO and CLK deviate. This saturating behavior leads to more freedom designing the regulation loops as slow as required.

7.6.7 Comparison of start up behavior of different oscillator types

We can distinguish between 3 classes of oscillators:

1. Harmonic oscillator
2. Relaxation oscillators
3. PLL systems

These designs differ significantly at start up.

Harmonic oscillators (ring oscillators, phase shift oscillators, LC oscillators, quartz oscillators) start with a small amplitude that slowly increases. The start up time takes hundreds to thousands of periods. Converting the signal of a harmonic oscillator into a system clock relies on Schmitt trigger circuits. During start up the signal of the oscillator can be just at the trip point of the Schmitt trigger and pulses may have wrong timings or can simply loose pulses. Usually the logic is protected from such wrong clock signals by simply gating the clock until the oscillator has been running for a certain number of periods.

The start up time of relaxation oscillators is very short. In most cases the start up time of a relaxation oscillator is simply the sum of the time needed for the bias generators to settle and one period of the oscillator. Most relaxation oscillators are running correctly after a few micro seconds.

The start of a PLL mainly depends on the time constant of the low pass filter. Of course the reference signal CLK must be coming from somewhere. Usually this is a quartz oscillator. The complete start up time of a PLL system therefore is in the range of some thousand periods of the quartz oscillator plus the time constant of the loop filter.

Since relaxation oscillators start quickly while PLL and harmonic oscillators start slowly it is common practice the use a relaxation oscillator to check the frequency of the (usually much more precise) harmonic oscillator for plausibility. If the harmonic oscillator frequency is out of a certain range the relaxation oscillator will take over and the system will go into a safe state with reduced functions. (some microcontroller manufacturers call this “limp home mode”)

7.6.8 Clock distribution

The signal coming from the oscillator usually is not yet ready to clock a significant amount of logic.

- Each logic gate adds to the capacity to be driven by the clock driver
- Very often a clock and a clock_not with inverted phase is needed
- Many circuits require a symmetrical clock signal (50% duty cycle)
- Sometimes certain edge conditions such as non overlapping clocks are needed (e.g. for sample & hold circuits)

Most oscillators do not immediately meet these requirements.

The most direct way to drive a big capacitive load is to simply build a strong buffer and to distribute the clock with low resistive wires in one of the upper metal layers (minimum stray capacity to ground or substrate). Ideally the strength of the inverters increases by factor e (2.7183) per stage. Using standard gates the sizing of the inverters is not completely free. So scaling of factor 2 to 4 per stage are common.

One of the limitations of this approach is the resistance of the clock line. Together with the distributed parasitic capacities and the gate capacities long clock lines create a significant delay differing from flip flop to flip flop in the logic. For synchronous logic this delay should be kept below the propagation delay of the fastest flip flop (otherwise the "always at clock edge" approach used in synthesized logic is no more valid because some logic signals already change before the last FF gets the clock edge needed to read the current state!) As a consequence modern logic synthesis tools create distributed clock paths with almost equal delay by deliberately plugging in delay inverters in the paths that are too fast. This is done automatically in the back end (layout synthesis).

Even symmetric oscillators do not well enough meet a 50% duty cycle. The duty cycle of oscillator always suffers from mismatch breaking the ideal symmetry. Ring oscillator almost always have odd numbers of inverters. So the signal can even be asymmetric by concept! Running the oscillator at double the clock frequency and using a divider stage is mandatory if 50% duty cycle has to be met. But even then the clock is not perfect. Most flip flops have different delays for rising and falling edge.

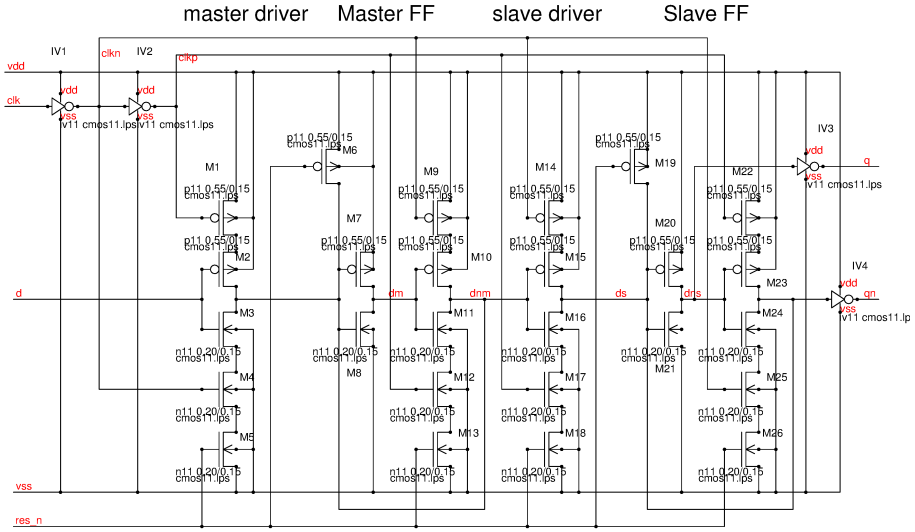


Figure 7.91: Data flip flop (DFF)

The delay times of output q and qn differ by the delay of the gate consisting of M22, M23, M24, M25 and M26. This delay can be made visible simulating the flip flop operating as a clock divider.

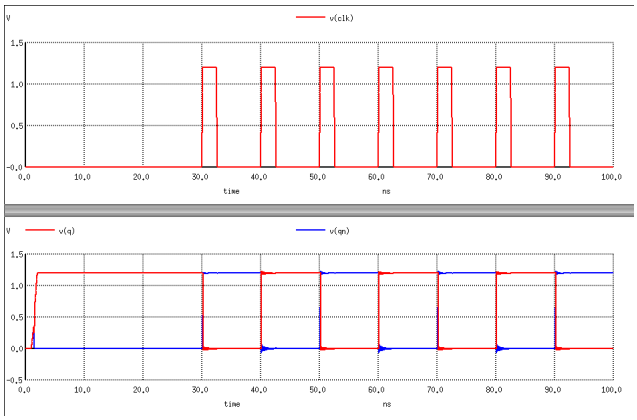


Figure 7.92: Flip flop operating as a clock divider to produce a symmetrical signal

The differences in the propagation delay of about 40ps become visible zooming into the edges of q and qn.

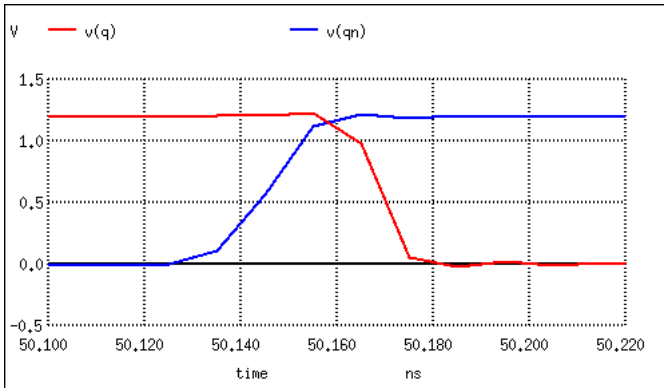


Figure 7.93: Zoom into the edges of q and qn

To solve these asymmetries a chain of inverters and latches can be used.

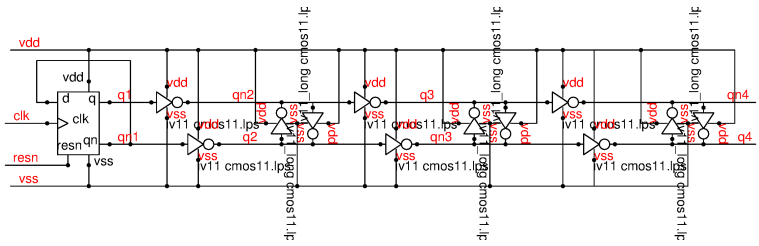


Figure 7.94: Divider with successing synchronizer latches

The long channel inverters act as a load for the path with the leading edge and as a speed up for the path with the trailing edge.

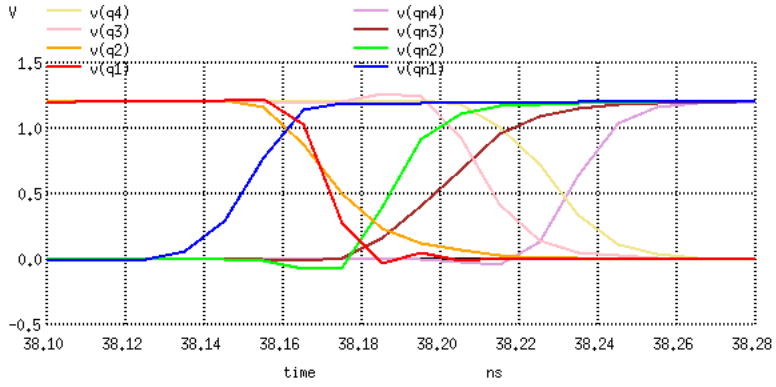


Figure 7.95: Simulation of the synchronizer chain

The figure above shows how the synchronizer moves the crossing of qx and qnx more and more into the middle with every stage. At the output of the flip flop the crossing was at 1V (falling edge was later than the rising edge.) After the 4th stage the crossing of the opposite signals is already at 0.5V.

7.7 Amplifiers

In section 7.3 the input stage of a differential amplifier was already discussed because some of these concepts were already required to understand the design of bandgaps and reference circuits. A full blown amplifier however consists of more than just a simple input stage. In fact typically an amplifier design starts with the load it has to drive. This section is dedicated to the design of an amplifier considering the requirements of the load and distortions. We are more or less starting the other way round: From output to input.

7.7.1 The output stage first

Why do we have a look at the output stage before looking at the input stage? The output stage is a common block for all kinds of amplifiers no matter what the input stage looks like. Usually we want the output stage to be broad band and with a low output impedance. Especially power amplifiers require big high voltage output transistors with

high gate capacities. These big transistors limit the bandwidth of the whole amplifier and produce several kinds of distortions. Some of these distortions can be reduced by feedback, but complete elimination of the distortions isn't possible. Therefore a detailed look at the properties of output stages is worth while!

Amplifier classes: Power amplifier designers distinguish between different classes of amplifiers. To understand this classification the amplifier must be drawn in a very theoretical way.

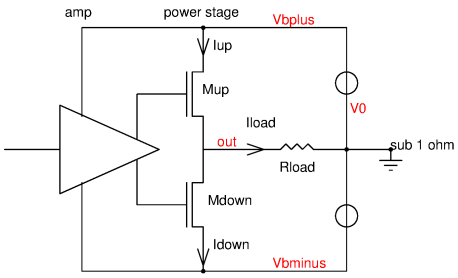


Figure 7.96: Strongly simplified power output stage

The most classical output stage consists of two power transistors. Mup us pulling up the output for positive output signals and Mdown is pulling down the output for negative output signals. The amplifier amp provides the drive signals for the two power transistors. In the following the conceptual ideas of the different amplifier classes are explained without any details of the implementation.

To understand the distortions in a more quantitative way the transistor level topology must be discussed. This will be done much later after discussing the properties of source follower concepts versus grounded source concepts.

class A amplifier: In a class A amplifier the currents Iup and Idown both never drop to 0. Iload is a linear sum (with Idown having a negative sign) of Iup and Idown.

$$I_{load} = I_{up} - I_{down} \tag{7.181}$$

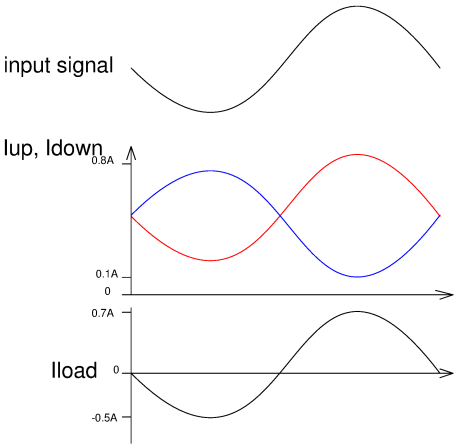


Figure 7.97: Currents in a class A amplifier

The signals shown above illustrate the currents flowing in a class A amplifier. If there is no input signal Iup and Idown settle in the middle of the range at about 0.5A in our example. This means there is a lot of cross conduction. The efficiency of a class A amplifier is very low because it consumes a lot of current even if there is no signal.

class B amplifier: In a perfect class B amplifier the pull up current is delivered by Mup and the pull down current is delivered be Mdown. The control by the preamplifier is designed such that Mup exactly starts to conduct wen Mdown reaches 0A and Mdown starts to conduct when Mup exactly hits 0A.

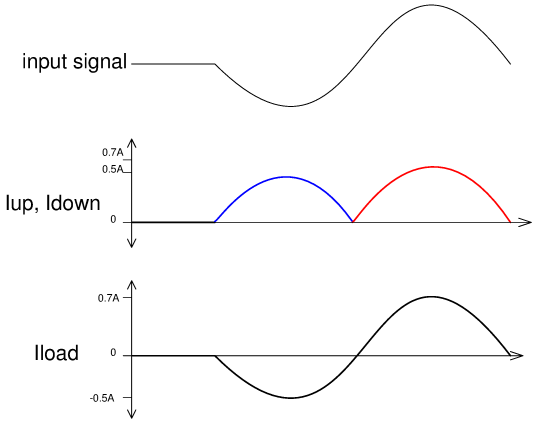


Figure 7.98: Currents flowing in a perfect class B amplifier

This design would work best if the turn on threshold of the transistors are known perfectly well and the transconductance of the transistors is infinite once they turn on. Real transistors have a transconductance that is a function of the current. They don't turn on in such an abrupt way. A more realistic picture of a real life class B Amplifier looks more like this:

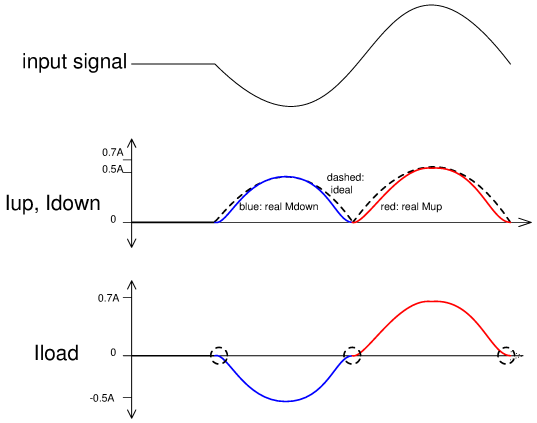


Figure 7.99: Limited transconductance of the power transistors leads to a less than perfect signal of real class B amplifiers

The main distortions are in the cross over range. These distortions are called cross over distortions. But this is not the only problem of class B amplifiers. The threshold of the transistors has a certain production spread. So trying to build class be amplifier in volume production we will find samples where the transistors turn on too late (one side is already completely off before the other side turns on) and samples where the transistors turn on too early (So the opposite transistor turns on before the other one really is off. In other words: Designing a real class B amplifier is impossible.

Real amplifiers are either optimized for low distortion operating between class A and the idealized class B or they are optimized for low (or no) bias current operating in a range where the current of both transistors falls to 0 in the transition phase.

class AB: Working between class A and class B is called class AB. In the transition zone both transistors are conducting simultaneously to linearize the transition. Usually this is a reasonable trade off between cross over distortion and current consumption while there is no signal.

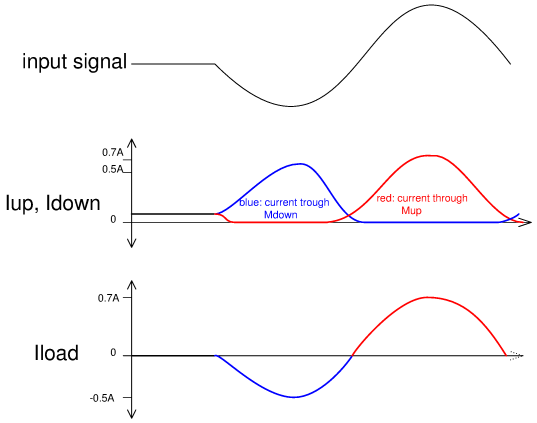


Figure 7.100: currents flowing in a class AB amplifier

The cross over distortions are significantly reduced (compared to class B) because at cross over both transistors contribute to the signal. During the transition the amplifier behaves similar to a class A (although the transconductance of the transistors at the the transition zone usually differs from the transconductance in large signal range).

class C: The target of a class C amplifier is to reduce or even completely eliminate the bias current that is flowing in a class AB stage while there is no input signal. This minimizes stand by current consumption. Using a class C approach cross over distortions can't be avoided.

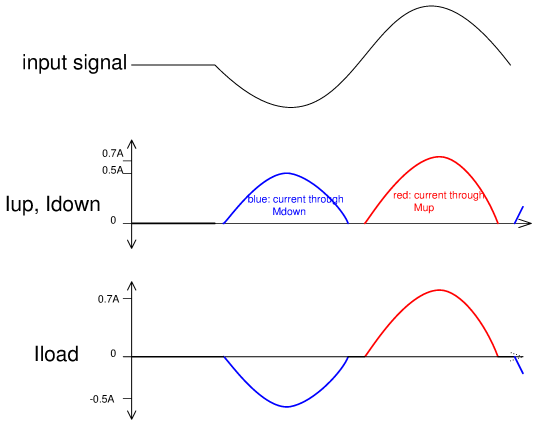


Figure 7.101: currents of a class C amplifier

Typical applications of class C amplifiers are RF power amplifiers that have filters successing the amplifier stage to remove the distortions. For audio applications class D amplifiers are limited to low cost systems or low quality systems such as megaphones etc.

class D: With the advent of fast switching transistors it became possible to represent the output signal of a power stage by a PWM. (pulse width modulation). The analog signal has to be reconstructed by an LC low pass filter following the output stage. This approach is called class D amplifier. The advantage of using a PWM is that at the switch there either is current, but almost no voltage across the switch, or the switch is open having a voltage across it but no current. The distortions of a class D amplifier are determined by the linearity of the pulse width modulator and the speed and edge symmetry of the power switches.

At the output of a class D amplifier we find the original signal folded around 0Hz and images of the original spectrum folded around the PWM frequency and all it's harmonics. The spectra folded around the PWM frequency and multiples of the PWM frequency have to be eliminated by filters.

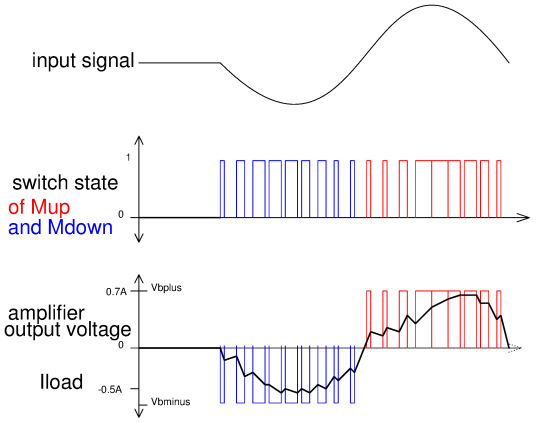


Figure 7.102: Signals of a class D amplifier

In the figure above the signals of a class D amplifier are shown. At the output of the half bridge we will find a PWM with positive and negative pulse trains. The inductance of the load will low pass filter the PWM and the current flowing in the load will approximately follow the black curve.

class E and class F: The concept of class E and class F is to shape the current and the voltage at the power transistor such that there either is a current but no voltage drop or a voltage but no current. The concept is discussed in literature for RF amplifiers with very complex matching networks. But it seems not to be usable for classical analog amplifiers.

class G: class G amplifiers attempt to boost efficiency switching between a low supply voltage for small signals and a high supply voltage for large signals. It has been used by Hitachi audio amplifiers in the 1970s and 1980s. (US patent 4100501). The switching between different supply rails on the fly (during the sine wave) however adds distortions due to the limited CMRR of the power stage.

Implementation of the power stage: There are basically two possible implementations of the power stage:

1. The load is connected to the drains of the power transistors. (grounded source or grounded gate/base)
2. The load is connected to the source of the power transistors. (grounded drain)

Each of these structures has specific merits and demerits. Before being able to analyze the distortions in a quantitative way these aspect have to be understood.

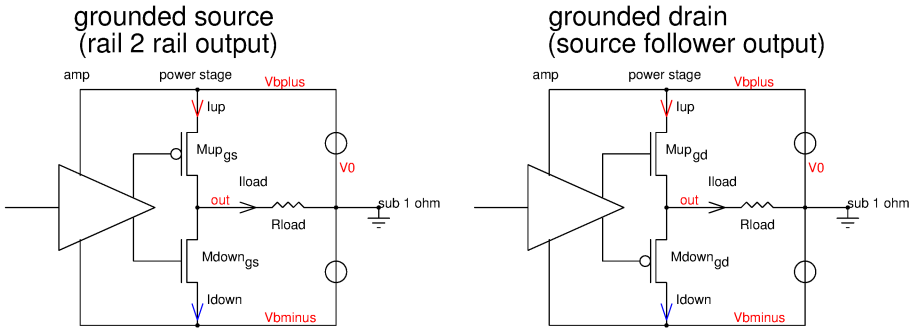


Figure 7.103: Comparison of grounded source and grounded drain topology

The following table summarizes the most important difference between the power stages. In fact, running the two topologies open loop means we are looking at dual circuits!

Table 30: Comparison of grounded source and grounded drain output stages

topology	grounded source	grounded drain
output behavior	similar to a current source	similar to a voltage source
output impedance	high, $Z_{out} = \frac{Vearly}{I_{drain}}$	low, $Z_{out} = 1/gm$
voltage gain	$gain = gm * Z_{load}$	$gain \approx 1$
resonant behavior	high power at parallel resonance	high power at serial resonance
signal swing	Vb_{minus} to Vb_{plus}	$Vb_{minus} + Vth_{pmos}$ to $Vb_{plus} - Vth_{nmos}$
stability in feedback loop	depends on Z_{load}	for reasonable Z_{load} load independent

For stability reasons and load independence the grounded drain design clearly offer a lot of advantages provided the supply voltage is significantly higher then the threshold voltage of the MOS transistors. The grounded source topology only can be recommended for low supply voltage or if the load impedance is really well defined. To make a grounded source stage a useful voltage amplifier having a feedback loop is a must. A grounded drain stage can be operated open loop as well. Therefore we first focus on the grounded drain (or source follower) stage first. The grounded source stage will be discussed later using it as a driver stage for a source follower (in this case the load impedance is well determined and (almost) independent of the load. To simplify things the discussion will start with a single ended source (or emitter) follower before adding the complication of making it a push-pull stage.

Emitter follower and Source follower stages: The classical design is an emitter follower or a source follower output. The output impedance in open loop operation is $1/\mathrm{gm}$. The most simple design is either an NPN transistor with a current sink or a resistor acting as a load.

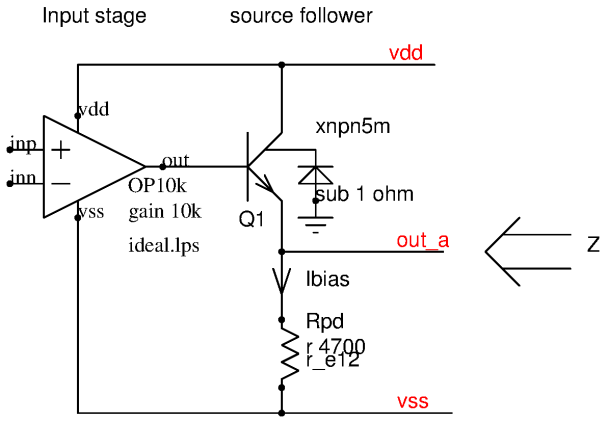


Figure 7.104: Bipolar source follower output

As long as the gain B of the transistor is not becoming the limiting factor the small signal output impedance of a bipolar source follower is:

$$R_{out} = \frac{k * T}{e * I_{bias}} = \frac{V_T}{I_{bias}} \tag{7.182}$$

Example: Assuming operation at room temperature $V_T = 26mV$ and a bias current of 1mA the ideal output resistance is 26Ω. In practical designs the achievable output impedance is a bit higher because of the emitter resistance of the transistor and the base resistance and the output resistance of the input stage.

$$R_{out} = \frac{V_T}{I_{bias}} + \frac{(R_b + R_{op})}{B} + R_e \tag{7.183}$$

Especially in low current consumption designs the amplifier can be built quite high resistive and the output resistance gets limited by $R_{op}$and current gain B.

MOS transistors operating in weak inversion have an exponential characteristic too. The performance does not quite reach the performance of a bipolar transistor due to the capacitive input divider consisting of the gate capacity and the capacity between the channel and the bulk. To operate in weak inversion usually only very low bias currents can be used.

An advantage of using MOS transistor stages is that we can assume B to approach infinite because (ideally) there is no gate current. In stead of the emitter resistance we have to take into account the source contact resistance R_s.

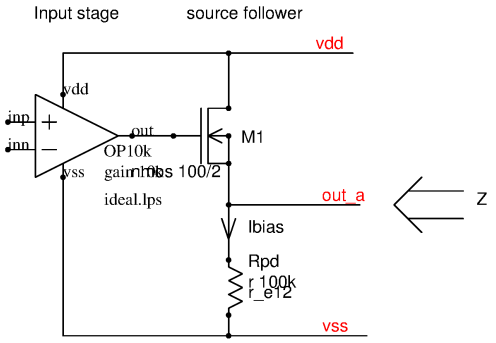


Figure 7.105: NMOS source follower output

Typically the degradation factor comparing with a bipolar follower is about:

$$K_C = \frac{C_{tox} + C_{bulk}}{C_{tox}} = 1.2...1.4 \tag{7.184}$$

and the output impedance in weak inversion becomes:

$$R_{out} = \frac{k * T * K_C}{e * I_{bias}} + R_s \tag{7.185}$$

Example: Operating at 300K with a bias current of $10\mu A$ and $K_C = 1.3$ we get:

$$R_{out10u} = \frac{26mV * 1.3}{10\mu A} = 3.4K$$

To decrease the output impedance the bias current must be increased. This will take the NMOS follower into strong inversion with a quadratic characteristic in stead of the exponential one. The derivative of the source current in strong inversion is:

$$\frac{dI_s}{dV_{gs}} = \frac{2 * K' * V_{gs} * W}{L} \tag{7.186}$$

with

$$V_{gs} = \sqrt{\frac{I_s * L}{K' * W}} \tag{7.187}$$

Combining the equations we get

$$R_{out} = \frac{1}{2} * \sqrt{\frac{L}{K' * W * I_s}} \tag{7.188}$$

K' depends on the type of transistor (NMOS or PMOS) and the gate oxide thickness. Typical values for NMOS transistors are about

$$K'_{nmos} = \frac{2000 \frac{\mu A * nm}{V^2}}{t_{ox}}$$

PMOS transistors have lower values because the mobility of holes is lower than the mobility of electrons. Her we usually have values around

$$K'_{pmos} = \frac{850 \frac{\mu A * nm}{V^2}}{t_{ox}}$$

Example: $t_{ox} = 7nm$, $W = 1000\mu m$, $L = 0.6\mu m$, $I_s = 2.5mA$ leads to:

$$K' = 285 \frac{\mu A}{V^2}$$

$$R_{out} = 35\Omega$$

What can be done to achieve a lower output impedance?

1. Increase K' by reducing the gate oxide thickness - Well, usually we are limited by the supply voltage range and this approach only is possible in very few cases.
2. Increase the width of the transistor.
3. Reduce the length of the transistor - again we are limited by supply voltage requirements.
4. Increase the bias current.
5. Use a push - pull configuration.

The push - pull configuration looks most promising. Here we simply add the gm of an NMOS transistor and a PMOS transistor. If the W/L of the PMOS transistor is chosen about twice as big as the W/L of the NMOS transistor the performance doubles or the same performance can be reached using half the bias current.

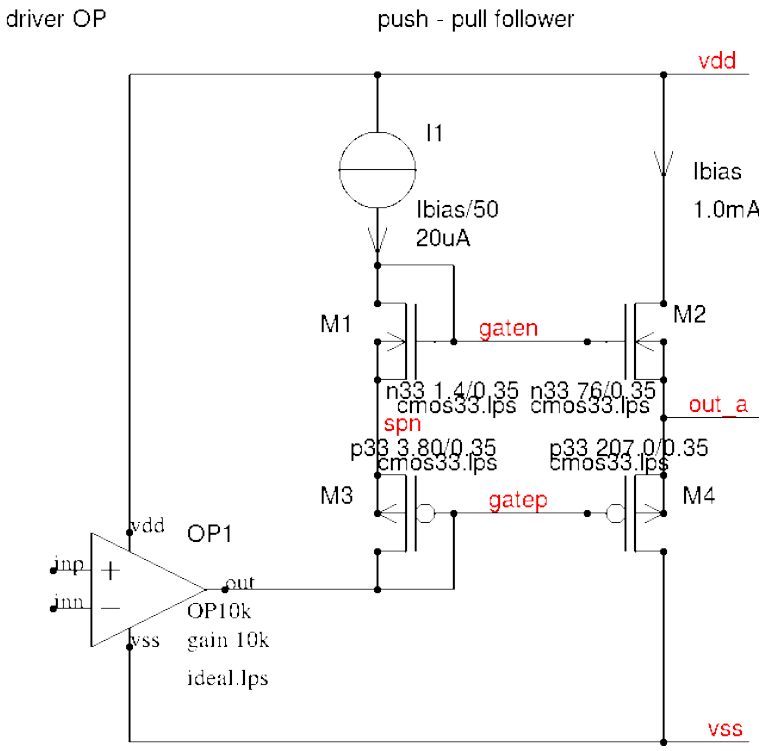


Figure 7.106: Push Pull MOS follower

Let's estimate the small signal output impedance assuming we have 7nm gate oxide:

$$R_{outM2} = \frac{1}{2} * \sqrt{\frac{0.35\mu m}{285\frac{\mu A}{V^2} * 76\mu m * 1mA}} = 64\Omega$$

$$R_{outM4} = \frac{1}{2} * \sqrt{\frac{0.35\mu m}{121\frac{\mu A}{V^2} * 207\mu m * 1mA}} = 59\Omega$$

$$R_{out} = \frac{R_{outM2} * R_{outM4}}{R_{outM2} + R_{outM4}} = 30\Omega$$

Now let us use transistors with a thinner gate oxide. The circuit could look like this:

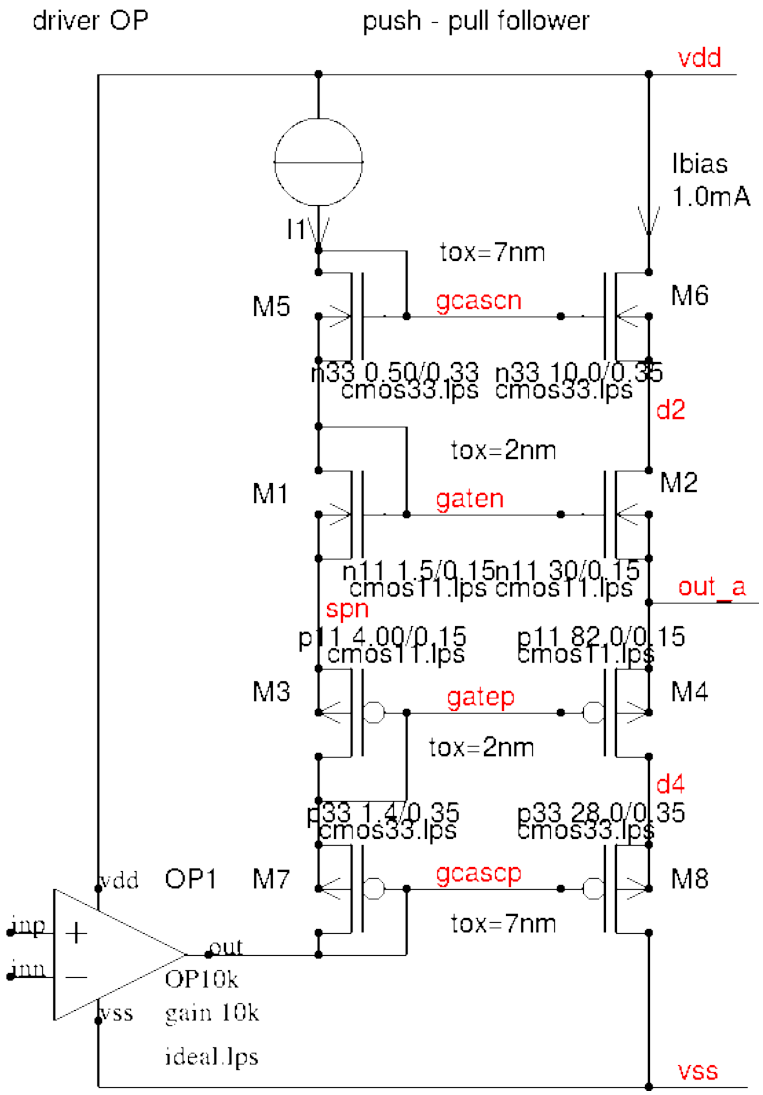


Figure 7.107: Source follower using thin gate oxide transistors and cascodes to lower the open loop output impedance

In this circuit output voltage swing was sacrificed to insert protection cascodes for M2 and M4. This way very thin gate oxides can be used for M2 and M4. Assuming we keep the bias current of 1mA flowing through M2, M4, M6 and M8 we can achieve an open loop output resistance of

$$R_{outM2} = \frac{1}{2} * \sqrt{\frac{0.15\mu m}{1000\frac{\mu A}{V^2} * 30\mu m * 1mA}} = 35\Omega$$

$$R_{outM4} = \frac{1}{2} * \sqrt{\frac{0.15\mu m}{425\frac{\mu A}{V^2} * 82\mu m * 1mA}} = 32\Omega$$

$$R_{out} = \frac{R_{outM2} * R_{outM4}}{R_{outM2} + R_{outM4}} = 17\Omega$$

M2 and M4 are already in the transition zone to weak inversion. So we are in fact close to the theoretical limit of the open loop output impedance!

$$R_{outmin} = \frac{1}{2} * \frac{k * T * K_c}{e * I_{bias}} \tag{7.189}$$

Depending on the gate coupling non ideality factor $K_c$this limit at room temperature is between about $13\Omega mA/I_{bias}$and $18\Omega mA/I_{bias}$. As long as pin out_a is inside the integrated circuit the solution is well usable. As soon as out_a is connected to a pad the thin gate oxide transistors M2 and M4 are exposed to ESD (electrostatic discharge). Such transistors are extremely vulnerable! Adding ESD protection resistors between the sources of M2 and M4 and the pin out_a however will destroy the performance of the circuit again.

Distortions of a source follower or emitter follower stage: To estimate the distortions let's start with the most obvious. A class C amplifier stage operating in open loop configuration has the highest distortions of the analog follower stages (class D needs a more complex discussion about PWM linearity, this is an other can of worms than class A to class C). A class C stage typically is unbiased. The MOS version looks like this:

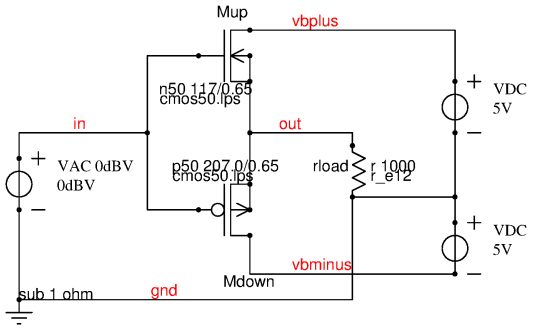


Figure 7.108: Simulation schematic of a class C amplifier

The most important deviation between the input signal and the output signal is caused by the thresholds of the transistors. The following plot shown the input signal v(in), the output signal v(out) and the error of the amplifier v(in,out). v(in,out) is the distortion of the class C amplifier stage. It is clearly visible that this distortion is almost rectangular and has an peak to peak amplitude of the twi thresholds.

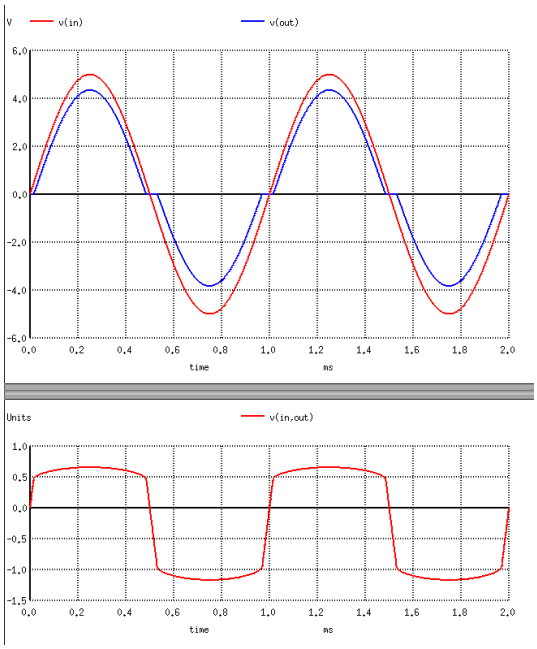


Figure 7.109: Input signal, output signal and error of a class C amplifier

The conclusion out of this is: To reduce the distortions of a class C amplifier the dead zone defined by the two thresholds of the transistors must be reduced. The lower the thresholds the better the result gets. Usually MOS thresholds are higher than the base emitter voltage of bipolar transistors. So bipolar transistors operating as a class C amplifier usually have less distortions than MOS transistors. This observation was already mentioned by [50]. Knowing this almost rectangular error signal we can almost immediately determine the spectrum at the output. It simply is the sine wave we are exciting the amplifier with plus the spectrum of the error v(in, out).

Adding a floating voltage source reduces the distortions. We are getting closer to a class B amplifier behavior.

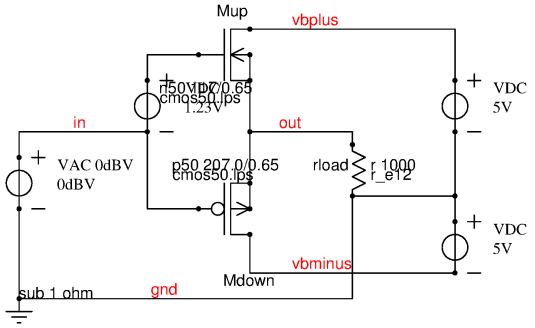


Figure 7.110: Simulation schematic of a class B amplifier

On the first glance we expect to get rid of all distortions if we exactly match the voltage of the floating DC source

with the thresholds of the MOS transistors. Surprisingly this is not the case! The orange curve shows the resulting error.

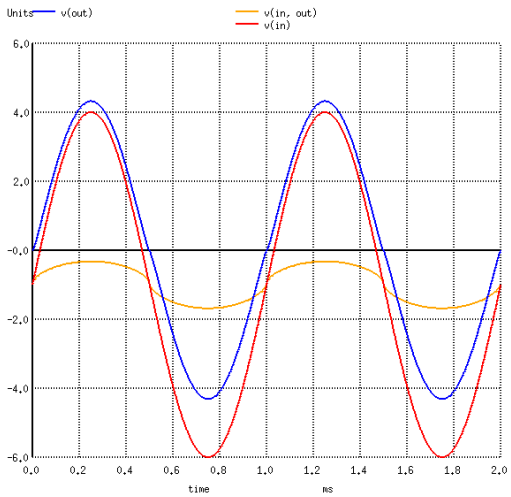


Figure 7.111: Simulation result of a class B amplifier after perfectly tweaking the floating bias voltage exactly to the thresholds described in the transistor models

The orange curve now describes the gate overdrive we need for the transistors to provide the current needed by the load resistor. For MOS transistors the half waves of the orange signals are proportional to the square root of the sine wave! In fact, this distortion already was present in the class C amplifier. There the roofs of the rectangular distortion are not completely even. They are rounded in the same way. So going from class C to class B we significantly reduced the cross over errors but there is a second error. This is a signal compression error!

The peak to peak amplitude of this signal compression can be estimated. This doesn't give nice to calculate equations but at least this will show what are the major influencing parameters. We have to look at both half waves separately because they are produced by different transistors. The current of a MOS transistor is determined by the gate overdrive ($V_{gseff} = V_{gs} - V_{th}$, how much higher the gate voltage is than the threshold). In an ideal class B amplifier the assumption is, that the transistors are prebiased exactly at the threshold voltage. (And there is already the first problem! In this case the transistor reaches substhreshold operation and we still have the sbthreshold current flowing through the power transistors. Looking as aspect ratios W/L in he range of 100000 or even higher the substhreshold current will NOT be negligible anymore. The end of the practical implementation of the ideal class B amplifier!)

$$I_d = k' * V_{gseff}^2 * \frac{W}{L} \tag{7.190}$$

The gate overdrive at the peak of the sine wave becomes:

$$V_{gseff} = \sqrt{\frac{I_{peak} * L}{k' * W}} \tag{7.191}$$

Off course the signal exists twice and the peak to peak voltage of the deviation becomes:

$$V_{distpp} = \sqrt{I_{peak}} * (\sqrt{\frac{L_{Mup}}{k'_{nmos} * L_{Mup}}} + \sqrt{\frac{L_{Mdown}}{k'_{pmos} * L_{Mdown}}}) \tag{7.192}$$

What are the learnings of these equations:

1. distortions increase with the square root of the peak current.
2. doubling the peak current will increase the distortions by 3dB
3. the aspect ration must be as big as possible to achieve a high gm
4. doubling W/L will reduce distortions by 3db
5. Since k' reduces with temperature distortions will go up with temperature.
6. Using transistors with higher gm will improve circuit performance

Looking at the symmetry of the signal we find out that due to different carrier mobilities in NMOS and PMOS transistors the distortion either gets asymmetric or we have to scale the PMOS transistors between 2 and 3 times bigger than the NMOS transistors. Asymmetric signals always will lead to even harmonics. If we want to reduce even harmonics we must match the NMOS and the PMOS transistors. This however will never work reasonably well!

Point 6 of the listing above gives an important hint. We can replace the MOS transistors by bipolar transistors to achieve a higher transconductance. For bipolar transistors the collector current calculates as:

$$I_c = I_0 * exp(V_{be}/V_T) \tag{7.193}$$

with $V_T = k * T/e$. At 300K $V_T \approx 26mV$. Even nicer: This equation applies to NPN transistors as well as PNP transistors. So we get better symmetry for free!

$$V_{be} = V_T * ln(I_c/I_0) \tag{7.194}$$

The distortion peak to peak voltage now becomes:

$$V_{distpp} = V_T * (ln(I_{peak}/I_{0npn}) + ln(I_{peak}/I_{0pnp})) \tag{7.195}$$

The only not so nice thing is that I_0 is very low and we basically end up at about $V_{distpp} = 2 * V_{be}$. Fortunately this can be fixed taking the bipolar amplifier into class AB mode. The distortion of a bipolar transistor stage in class AB mode drops depending on the bias current.

$$V_{distpp} = V_T * 2 * ln(I_{peak}/I_{bias}) \tag{7.196}$$

Making I_{bias}about 10% of the peak current gives nicely low distortions in the range of only 120mV peak to peak (at room temperature).

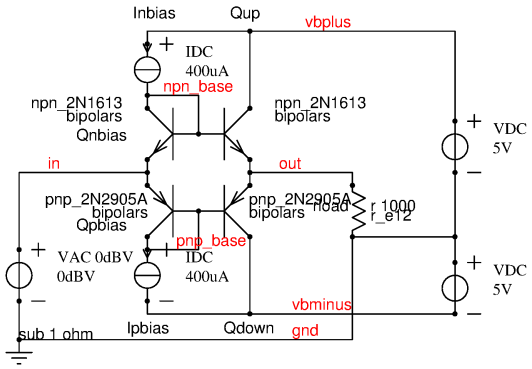


Figure 7.112: A bipolar stage operating in class AB mode including a bias current generator providing a bias current of 10% of the peak current

The performance of this class AB stage is convincing! The distortions are really down to about 120mV peak to peak as estimated before.

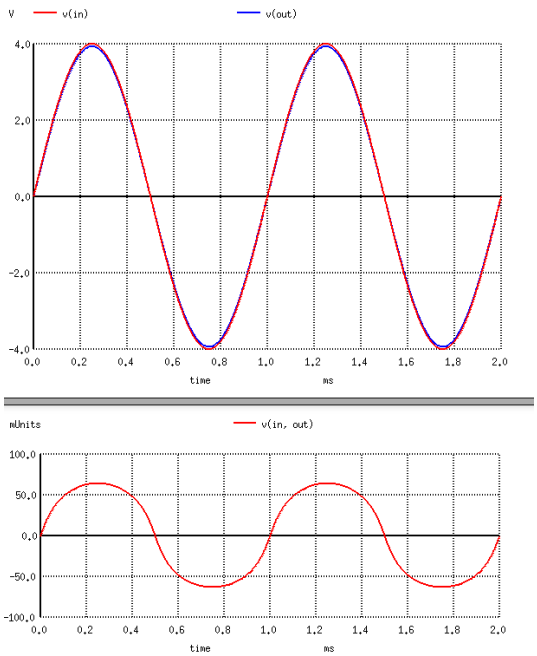


Figure 7.113: Simulation result of the bipolar class AB amplifier shown above

The distortions can be further reduced by increasing the bias current. In the end this would in fact take us to class A, the holy grahl of HiFi enthusiasts. But since we are discussing chips we won't want to afford the heat sins such a class A amplifier would require.

Even worse: Modern main stream technologies are optimized for MOS transistors. We don't even have high voltage bipolars anymore. This forces us to build the class AB stage with MOS transistors. The distortion equation thus becomes:

$$V_{distpp} = (\sqrt{I_{peak}} - \sqrt{I_{bias}}) * (\sqrt{\frac{L_{Mup}}{k'_{nmos} * W_{Mup}}} + \sqrt{\frac{L_{Mdown}}{k'_{pmos} * W_{Mdown}}} \tag{7.197}$$

So what scales so nicely with bias current using bipolar transistors will barely improve using MOS transistors unless we really go to class A just because the bipolars have a nice exponential control function while the poor MOS transistors fall back to a quadratic control function. This is the mathematical explanation of Douglas Self's observation that deviating from literature statements in the 1980s in fact MOS power amplifiers will not perform better than bipolar power amplifiers. They perform worse!

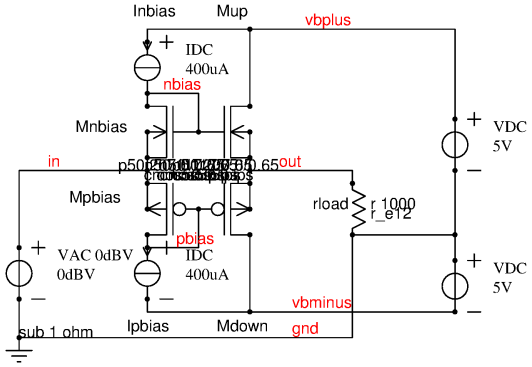


Figure 7.114: class AB amplifier with $I_{bias} = 0.1 * I_{peak}$ used for distortion simulation

The signals and the deviation between the input and the output signal of this stage is shown below.

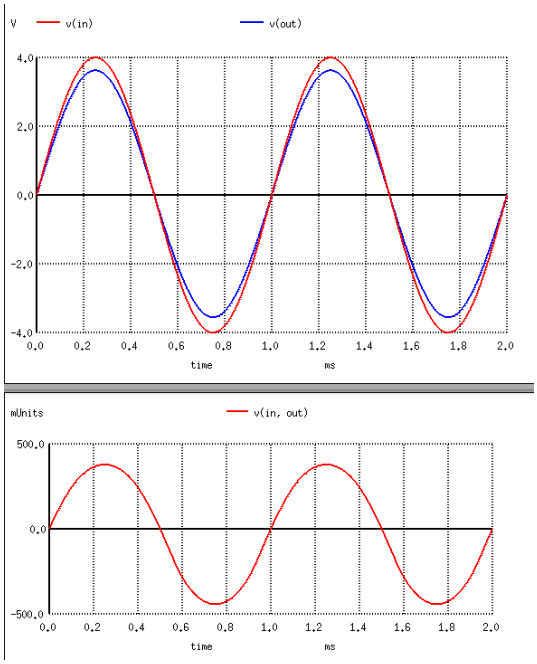


Figure 7.115: Signals of the class AB MOS amplifier and the deviation between input and output

As we can see the MOS implementation has factor 10 higher deviations. But at least the deviation curve looks a bit more sinusoidal than the deviation of a bipolar amplifier.

Closing the loop: After having analyzed distortions of the open loop topology let us close the feedback loop.

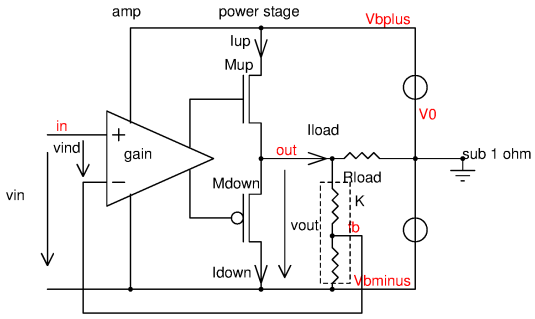


Figure 7.116: Complete power amplifier including the feedback loop

To calculate how much the feedback loop reduces the distortions we have to calculate the output voltage of a real amplifier (with limited gain) and compare it with the output signal of an ideal amplifier with unlimited gain. The feedback divider has the ration K=R2/(R1+R2). This is a constant number usually smaller than 1. The output voltage of an ideal (unlimited gain) amplifier calculates as:

$$V_{outideal} = V_{in}/K \tag{7.198}$$

The output voltage of a real amplifier with limited gain calculates as:

$$V_{out} = gain * V_{ind} \tag{7.199}$$

With

$$V_{ind} = V_{in} - V_{out} * K \tag{7.200}$$

This leads to:

$$V_{out} = V_{in} * \frac{gain}{1 + K * gain} \tag{7.201}$$

The difference between the real output voltage and the ideal output voltage becomes:

$$V_{error} = V_{in} * (\frac{gain}{1 + K * gain} - \frac{1}{K}) \tag{7.202}$$

$$V_{error} = -\frac{V_{outideal}}{1 + K * gain} \approx \frac{V_{outideal}}{gain/gain_{closedloop}} \tag{7.203}$$

In other words: The regulation loop reduces the errors by the ration between the gain of the amplifier amp and the closed loop gain (This applies as long as gain is significantly higher than 1/K).

What can we conclude for our power amplifier?

1. The higher we make the gain of the amplifier amp the better the feedback loop will reduce distortions
2. Since the gain of amplifier amp rolls off with increasing frequency the rejection of distortion rolls off with typically -20dB/decade as well.
3. Since harmonics roll off the decrease of the rejection and the harmonic roll off compensate.
4. All harmonics will approximately have the same amplitude until we reach the unity gain bandwidth of the amplifier.
5. Above unity bandwidth we reach at the open loop distortion level of the amplifier.
6. Above unity bandwidth we will observe the natural roll off of the harmonics (usually -20dB/decade)
7. Avoid capacitive loads of the amplifier to allow a higher gain bandwidth product of the complete amplifier.

The following plot shows an example of a closed loop amplifier spectrum with a fundamental at 125kHz

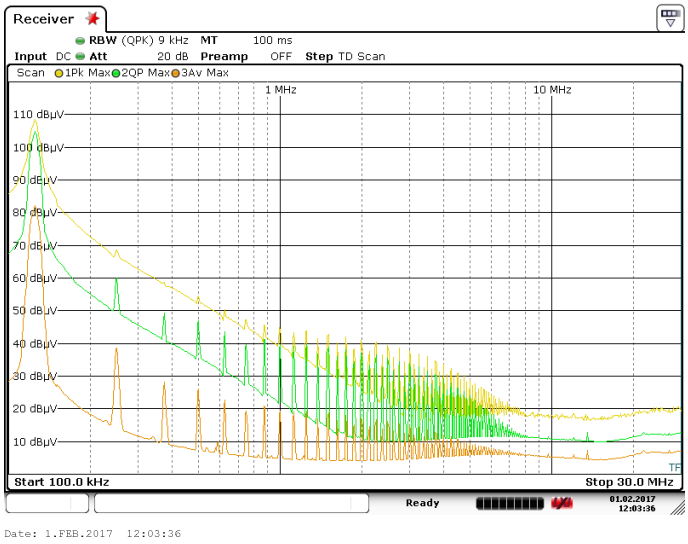


Figure 7.117: Example of a real class AB amplifier with closed loop operation

As described the spectrum of the distortions in fact flattens out between 400kHz and about 4MHz.

Production cost considerations: Chip costs are related to the die size. Building big PMOS transistors or PNP transistors is more expensive than scaling NMOS or NPN transistors for the same current. This leads to replacing PNP and PMOS transistors by quasi PMOS darlingtons and quasi PNP darlingtons. The control characteristic is determined by a small PMOS or PNP transistor while the high current is handled by the NMOS or the NPN transistor. This trick is common since the 1960s.

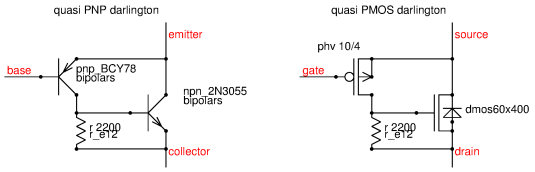


Figure 7.118: replacing PNP and PMOS power transistors by quasi darlington stages

Replacing one side of the amplifier (the pull down transistor) by a quasi darlington leads to an asymmetric capacitive load of the preamplifier because the small input transistors require less base current (bipolar BCY78) and have less gate capacity (phv 10/4) compared to the originally used power transistor. So in quasi darlingtons are used to save cost on the pull down side the capacity and the base drive has to be symmetrized again using NPN or NMOS darlingtons on the high side as well. Otherwise we will run into a frequency compensation chaos later.

Quasi complementary power output stages: Quasi complementary output stages are used for power amplifiers if there is enough supply headroom available. Originally this has been a concept used for bipolar audio amplifiers using discrete transistors. The first publications date back to the 1960s.

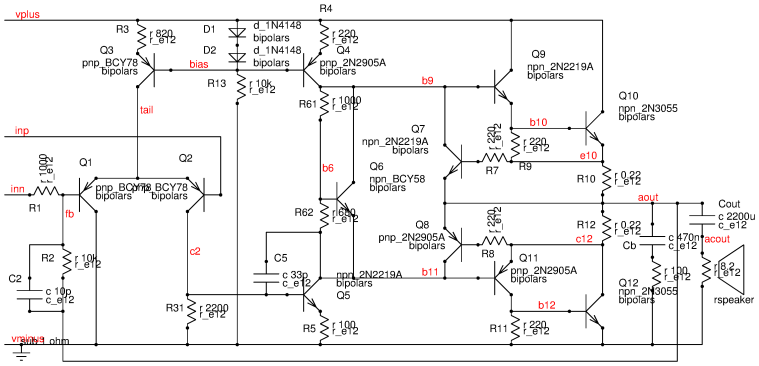


Figure 7.119: Classical bipolar audio amplifier with quasi complementary output stage

The circuit is very classic. There are various implementations of this circuit shown in [36, 34]. The design looks simple. Nevertheless some critical thermal couplings must be taken into consideration:

Q6 together with Vbe multiplication by R61 and R62 determines the DC operating points of Q10 and Q11 (Q12 just amplifies the current flowing in Q11). Ideally Q11 and Q12 are operated at a DC current in the range of 10mA to 50mA. The higher the current the more linear the amplifier gets. But making the current too high may produce unacceptable power dissipation. Q6 must thermally be coupled to Q10 and Q11.

Q7 and Q8 act as current limiters. Thermal coupling of Q7 and Q10 reduces the current limit the hotter the power amplifier gets. In the same way Q8 should be thermally coupled to Q12.

The current through Q5 is limited by R5 and the the ratio of R31/R3 and the forward voltage of D1.

Loudspeakers usually are complex loads. Depending on the frequency and the back-emf of the moving coil they can move from inductive behavior to capacitive behavior. To maintain a certain minimum real part of the impedance for all frequencies the output of audio amplifiers usually needs to be damped with an RC bypass of the loudspeaker.

Frequency compensation of the amplifier usually is done adding a miller capacity to Q5. In some case the feedback network also has a phase feed forward capacitor.

The same concept can be realized using MOS transistors. Since electron mobility is higher than the mobility of holes NMOS transistors can (at the same Rdson) be built smaller than PMOS transistors. For slow signals the high side stage can be built using a simple NMOS power transistor. The drawback of this simple approach is the high gate capacity of the high side stage.

Building an NMOS darlington to get rid of the high capacitive load of the preamplifier is tempting, but this adds one more loss of a threshold of the power transistors. To avoid adding one more threshold in the high side stage it is much better to use a folded darlington.

The driver stage output must be rail to rail capable.

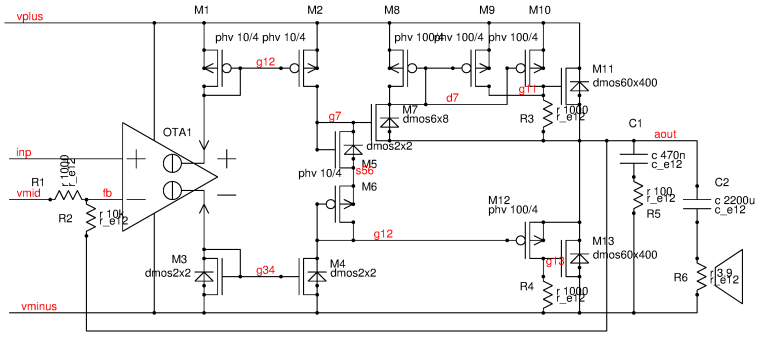


Figure 7.120: CMOS quasi complementary power amplifier

The driver stage is a fully differential operational transconductance amplifier.

The current limitation of the power amplifier was omitted for simplicity.

Practical designs often replace the current mirrors (M1, M2 and M3, M4) by translinear stage to combine more gate drive current with low current consumption.

Replacing current mirrors M3, M4 and M1, M2 by translinear stages: For power amplifiers as well as operating with low supply voltages it is important to exploit the available supply as good as possible. The swing of the output signal should get as close to the supply rails as possible. There are two basic concepts how to build a rail to rail output stage:

1. Source follower stage with driver supply from a bootstrap stage or a charge pump.
2. Drain output stages (low side NMOS, high side PMOS)

The source follower stage with bootstrapped driver supply requires external capacitors. This approach is limited to few applications where the additional components can be accepted (from cost point of view). This approach was done frequently in the design of amplifiers using discrete transistors. On chip the capacitors required to store the energy to pump up the supply voltage can't be implemented.

The drain output stage is the classical approach found in almost all low voltage operational amplifier stages. As a starting point let us analyze one half of the classical output stage. It basically is a current controlled current source with moderate nonlinearity.

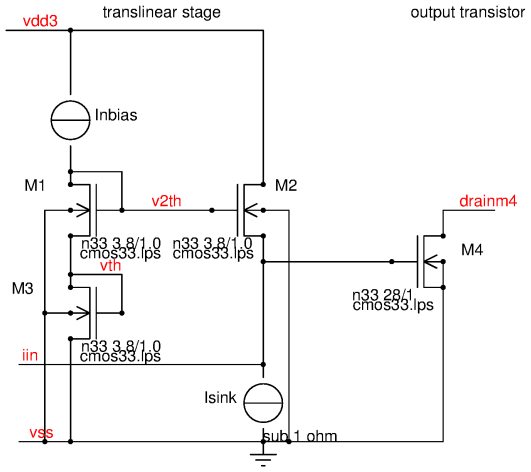


Figure 7.121: One half of the output stage of a typical rail to rail output

As a simple start let's assume the current source Isink and Inbias both have the same current and no input current is present at pin iin. In this case the operating points are easy to calculate. The voltage at iin (gate of the output transistor) and at vth become equal. So the bias current flowing in the output stage (if pin drainm4 is connected to a voltage higher than Vdssat of M4) M4 becomes:

$$I_{drainm4} = I_{nbias} * \frac{W_{M4} * L_{M3}}{W_{M3} * L_{M4}} \tag{7.204}$$

Example: Isink=Inbias=10μA, W/L of M4 is 28, W/L of M3 is 3.8 leads to an operating point of M4 of

$$I_{drainm4} = 10\mu A * \frac{28}{3.8} = 73.6\mu A$$

If a current is applied at pin iin the small signal gain becomes:

$$gain = \frac{gm_{M4}}{gm_{M2}} = \frac{W_{M4} * L_{M2}}{W_{M2} * L_{M4}} \tag{7.205}$$

In our example the small signal current gain is 7.36. This gain however is non linear! With an increasing current flowing into the circuit M2 will operate with decreasing current. So the gm of M2 decreases and the impedance increases accordingly. When the current exceeds Isink it even becomes infinite. (then we start to pull up the gate of M4 even when M2 is fully off. gm of M3 drops to 0.)

A general calculation including different sizes for all transistors leads to more complex equation than the more or less trivial case shown above. The effective gate voltage available at net iin becomes:

$$V_{gseffm4} = V_{gsm4} - V_{th} = \sqrt{\frac{I_{bias}}{gm}} * (\sqrt{\frac{L_3}{W_3}} + \sqrt{\frac{L_1}{W_1}}) - \sqrt{\frac{I_{sink} * L_2}{gm * W_2}} \tag{7.206}$$

To make life a bit easier let us substitute the constant part of the equation.

$$V_{gseffOP0} = \sqrt{\frac{I_{bias}}{gm}} * (\sqrt{\frac{L_3}{W_3}} + \sqrt{\frac{L_1}{W_1}}) \tag{7.207}$$

This is the theoretical operating point at Isink=0 that would be established at the gate of M4 if M2 would not have a substhreshold slope. Due to M2 entering subthreshold operation at very low sink currents this operating point does not really exist. Entering substhreshold operation the equations shown here do not apply anymore. But since normally building output stages we are in strong inversion we can use the quadratic characteristics with negligible error.

The current flowing into the drain of M4 becomes

$$I_{drainm4} = \frac{gm * W_4}{L_4} * (V_{gseffOP0} - \sqrt{\frac{I_{sink} * L_2}{gm * W_2}})^2 \tag{7.208}$$

or if we multiply it out:

$$I_{drainm4} = \frac{W_4}{L_4} * (I_{bias} * (\frac{L_1}{W_1} + 2 * \sqrt{\frac{L_1 * L_3}{W_1 * W_3}} + \frac{L_3}{W_3}) - 2 * \sqrt{\frac{I_{bias} * I_{sink} * L_2}{W_2}} * (\sqrt{\frac{L_1}{W_1}} + \sqrt{\frac{L_3}{W_3}}) + I_{sink} * \frac{L_2}{W_2}) \tag{7.209}$$

This is the equation of a parabola. But of course we are only interested in one half of it: The side with positive Isink (The other half M4 is off and the equation does not apply anymore). To design an amplifier the most interesting parameter is the gain.

$$gain = \frac{dI_{drainm4}}{dI_{sink}} = \frac{W_4}{L_4} * (\frac{L_2}{W_2} - V_{gseffOP0} * \sqrt{\frac{L_2 * gm}{W_2 * I_{sink}}}) \tag{7.210}$$

$$gain = \frac{W_4}{L_4} * (\frac{L_2}{W_2} - \sqrt{\frac{I_{bias} * L_2}{I_{sink} * W_2}} * (\sqrt{\frac{L_3}{W_3}} + \sqrt{\frac{L_1}{W_1}})) \tag{7.211}$$

Since we are only using one side of the parabola we are only interested in the side where the gain is negative (The side with positive gain is exactly the side where M4 is off and the equations do not apply). The most interesting observation is that the gain roughly follows

$$gain \approx 1 - 2 * \sqrt{\frac{I_{bias}}{I_{sink}}}$$

To get a better feeling of these complex looking equations let us have a look at the example of above:

$$L_1 = L_2 = L_3 = L_4 = 1\mu m$$

$$W_1 = W_2 = W_3 = 3.8\mu m$$
$$W_4 = 28\mu m$$
$$I_{bias} = I_{sink} = 10\mu A$$

This leads to the following results:

$$I_{drainm4} = 28 * (10\mu A * \frac{4}{3.8} - 2 * 10\mu A * \sqrt{\frac{1}{3.8}} * 2 * \sqrt{\frac{1}{3.8}} + 10\mu A * \frac{1}{3.8}) = 73.684\mu A$$

$$gain = 28 * (\frac{1}{3.8} - \sqrt{\frac{1}{3.8}} * 2 * \sqrt{\frac{1}{3.8}}) = -\frac{28}{3.8} = 7.3684$$

(Well, this is just the general equation of exactly the trivial case we already calculated before. But now we can also plug in other currents and transistor sizes to get the complete curve.)

There are two interesting cases in this equation.

1. $I_{sink}$is in the denominator of the equation of the gain. If I_{sink}approaches 0 the gain becomes infinite!
2. The gain can also become 0! This is the case when the effective gate voltage of M0 reaches 0

The second case can be constructed as a circuit as well. Here it is:

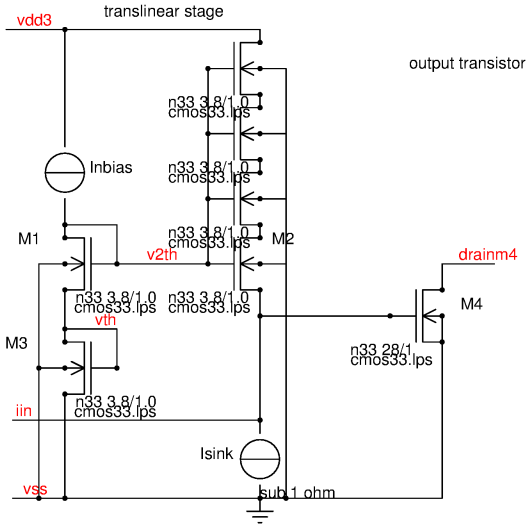


Figure 7.122: Output stage without any gain

So what happened? M2 is 4 times longer than M1 and M2 but operating at the same current. The nice squaring function creates a difference between the voltage of net vth and net iin that exactly corresponds $\sqrt{4} * Vgseff$. So

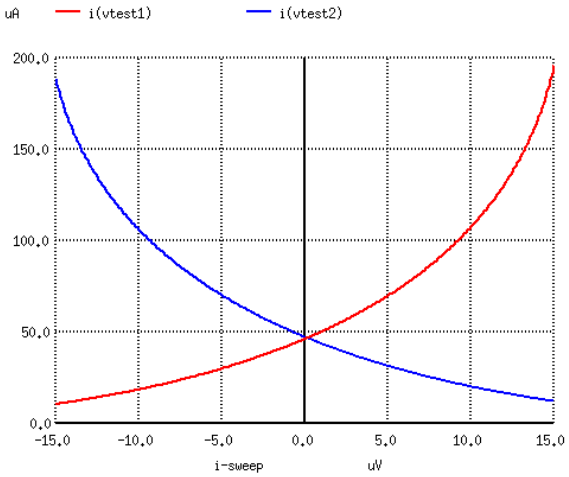


Figure 7.124: Quadratic characteristic of the current flowing through P6 and N6 sweeping the input current linearly

M4 is operated exactly at it's threshold. Using the pure strong inversion calculation M4 is exactly off. Well, this is not quite the case. M4 in fact is in weak inversion and will just have a drain current of some hundred nA.

Theoretically I expected I can use this circuit to produce a constant current of some nA. But over temperature it was not stable. I ended up with something that - depending on temperature - produces between 100nA and 500nA. It can be improved making M4 wider and adding a resistor in the source of M4. This basically is a gm reduction of M4. So non ideal effects won't get amplified with the exponential function of the weak inversion operation anymore. But this is not an analytic circuit anymore!

Example of a rail to rail output stage: The following circuit shows a simple rail to rail output stage. Since we use a translinear stage the input signal consists of two currents. To boost the gain part of the currents flowing in the translinear stages is cross coupled. Don't cross couple 100% of the current because then the loop gain of the cross coupling reaches 1 and we get a latch! P5 must always be smaller than P4 and N5 must always be smaller than N4.

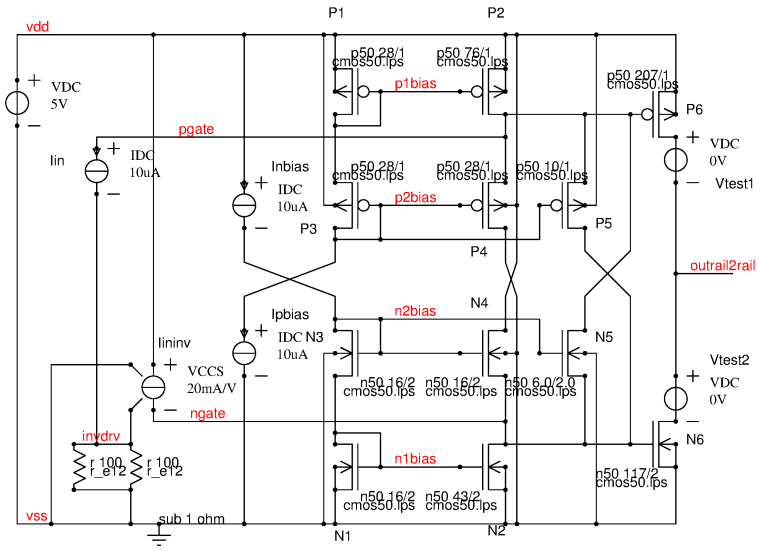


Figure 7.123: Rail to rail output stage with gain boost P5 and N5

To characterize the DC transfer function Iin is sweeped while at the same time the voltage controlled current source VCCS perform exactly the opposite sweep. Measuring the currents flowing in the output transistors using test sources Vtest1 and Vtest2 shows the quadratic behavior of the output current. (The output is forced to 2.5V during the test).

The resulting current flowing out of the circuit is the difference of both currents. Exactly where we may have to deal with cross over distortions the gain is lowest. At the two ends of the range the gain reaches its maximum. Therefore for a rail to rail output stage stability must be verified at the extremes of the output signal range!

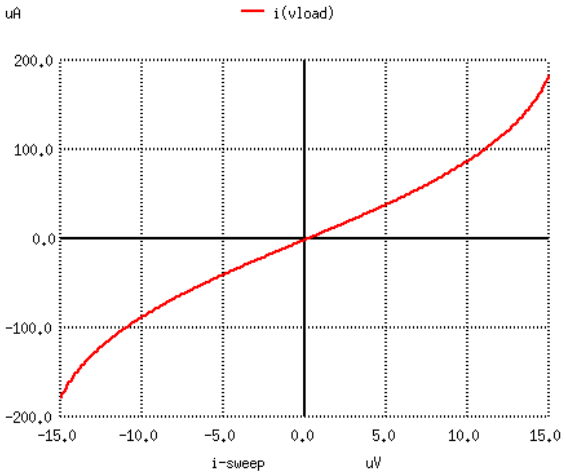


Figure 7.125: DC transfer characteristic sweeping Iin and measuring the current flowing into a 2.5V source

The more cross conduction we allow in the middle of the range the more linear the stage becomes. Usually the permissible cross conduction is limited by thermal considerations.

Since the output consists of two current stages the output pole is determined by the load resistance and the load capacity. The impedance seen at the output is the load resistance divided by the (frequency dependent) gain margin of the amplifier. To obtain the best possible pole splitting even at changing load capacities the only usable way of compensating the rail to rail amplifier is to use a miller compensation. Tricks like replica output stages common for source output stages in most cases can't be used for rail to rail outputs.

Power dissipation of the output stage: The power dissipation of the output stage is a function of the load current, the supply voltage and the output voltage. The difference V_{ds} between the supply voltage and the output voltage is the voltage drop over the power transistor. In a class B or class C amplifier only one of the power transistors is conducting at a time.

$$V_{ds} = V_b - V_{out} \tag{7.212}$$

The power calculates:

$$P(t) = V_{ds}(t) * I_{out}(t) \tag{7.213}$$

As a simple case let us assume the load is resistive. The load current flowing becomes:

$$I(t) = \frac{V_{out}(t)}{R_{load}} \tag{7.214}$$

While the pull up transistor is on the output voltage can vary from 0V to $V_b - V_{dsmin}$. $V_{dsmin}$is the minimum drain source voltage the power transistor can reach. Usually the limit for resistive load is:

$$V_{dsmin} = I_{loadmax} * R_{dson} \tag{7.215}$$

If no bootstrap voltage or charge pump is available the limiting factor can also be the required gate voltage needed to drive the output stage. In this case we get:

$$V_{dsmin} = \sum V_{gs} \tag{7.216}$$

(If the output stage is a darlington transistor several gate voltages ma need to be stacked.)

In most cases class B linear amplifiers are used for audio applications. So the calculation becomes most interesting for sinusoidal signals.

$$V_{out}(t) = A * sin(\omega t) \tag{7.217}$$

$$I_{out}(t) = \frac{A}{R_{load}} * sin(\omega t) \tag{7.218}$$

$$V_{ds}(t) = V_b - A * sin(\omega t) \tag{7.219}$$

choosing the maximum possible signal (where we just don't clip) the amplitude becomes:

$$A = V_b - V_{dsmin} \tag{7.220}$$

In this case the voltage drop over the power transistor becomes:

$$V_{ds}(t) = V_b * (1 - sin(\omega t)) + V_{dsmin} * sin(\omega t) \tag{7.221}$$

The power dissipation becomes:

$$P(t) = \frac{sin(\omega t)}{R_{load}} * [V_b^2 - V_b V_{dsmin} - sin(\omega t) * (V_b - V_{dsmin})^2] \tag{7.222}$$

For very large amplitudes $V_{dsmin} << V_b$ the power P(t) approaches

$$P(t) \rightarrow \frac{V_b^2}{R_{load}} * sin(\omega t) * (1 - sin(\omega t)) \tag{7.223}$$

In this case the power peaks at sin(t)=0.5 leading to

$$P_{max} = \frac{V_b^2}{4 * R_{load}} = \frac{V_b * I_{peak}}{4} \tag{7.224}$$

Example: $V_b = 20V$, R=20Ω

$$P_{max_{20}} = 5W$$

As soon as current and voltage are not in phase anymore (Load is not resistive but is a reactance) things get more critical! Worst case is a reactive (capacitive or inductive) load.

$$P_{reac}(t) = \pm cos(\omega t) * I_{peak} * [V_b * (1 - sin(\omega t)) + V_{dsmin} * sin(\omega t)] \tag{7.225}$$

The sign depends on the the kind of load. For a capacitive load the current starts to flow through the pull up transistor at $\omega t = -\frac{\Pi}{2}$ and disappears at $\omega t = \frac{\Pi}{2}$. The sign is positive. If $V_{dsmin} << V_b$the equation simplifies to the approximation:

$$P_{reac}(t) = \pm cos(\omega t) * I_{peak} * V_b * (1 - sin(\omega t)) \tag{7.226}$$

$$P_{reac}(t) = \pm I_{peak} * V_b * (cos(\omega t) - \frac{sin(2\omega t)}{2}) \tag{7.227}$$

To find the peaks we have to solve

$$\frac{dP_{reac}(t)}{dt} = \pm I_{peak} * V_b * (-sin(\omega t) - cos(2\omega t)) = 0 \tag{7.228}$$

The peak power dissipation is about

$$P_{reacmax} \approx 1.3 * I_{peak} * V_b \tag{7.229}$$

This is about 5.2 times more peak power dissipation than what we observe with a real (resistive only) load. The following figure shows the function sin(t)+cos(2t) where it is positive

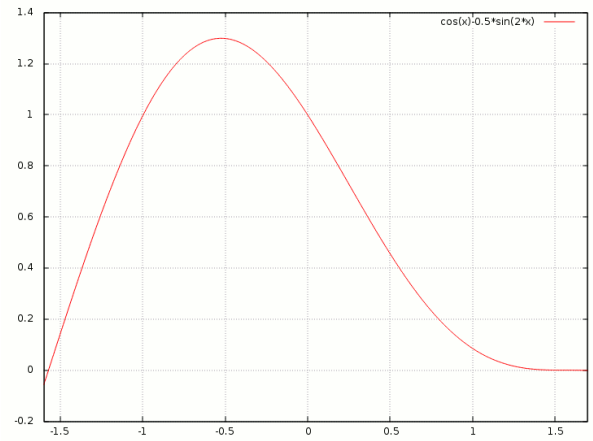


Figure 7.126: Power dissipation of the output stage driving a reactive load. (The value of this function must be multiplied with $I_{peak} * V_b$)

Efficiency of a power amplifier: The efficiency calculation doesn't make sense for a pure reactive load. Usually efficiency is calculated for ideal resistive loads. The output power of a class B amplifier operating with a sine wave signal is:

$$P_{out} = \frac{I_{peak} * V_{peak}}{2} \tag{7.230}$$

Average current consumption (of the power stage alone, excluding the driver stage consumption) is

$$I_{baverage} = \frac{I_{peak}}{2 * \pi} * \int sin(\omega t) d\omega t \tag{7.231}$$

Integration of the sinus runs from 0 to Πonly because the second half wave is flowing through the other side of the half bridge. This leads to

$$I_{baverage} = \frac{I_{peak}}{\pi} \tag{7.232}$$

Since the amplifier needs two supplies of $+V_b$ and $-V_b$to allow positive and negative half waves the total supply voltage becomes $2 * V_b$ and the average power consumption ideally becomes:

$$P_{average} = \frac{2 * I_{peak} * V_b}{\pi} \tag{7.233}$$

The resulting best case efficiency (for sine wave operation and V_{peak} reaching the supply rails) becomes:

$$\eta = \frac{P_{out}}{P_{average}} = \frac{\pi * V_{peak}}{4 * V_b} \tag{7.234}$$

Best case $V_{peak} \rightarrow V_b$the best case efficiency becomes:

$$\eta_{opt} = \frac{\pi}{4} = 78.5\%$$

Practical values are always below 78% because the peak voltage of the output signal of the amplifier never reaches the supply rail voltage of $\pm V_b$ and because the driver stage of the power transistors consumes additional current. Introducing a minimum drop at the power transistor the equations start to change a little bit.

$$P_{out} = \frac{I_{peak} * (V_b - V_{dssat})}{2} \tag{7.235}$$

The efficiency of the amplifier with a minimum drop becomes

$$\eta = \frac{P_{out}}{P_{average}} = \frac{\pi * (V_b - V_{dssat})}{4 * V_b} \tag{7.236}$$

Example:

An amplifier has a symmetrical supply of +-20V, The minimum drop V_{dssat} over the power transistor is 3V. This leads to an efficiency of

$$\eta_{example} = \frac{\pi * 17V}{4 * 20V} = 66.7\%$$

To estimate the cooling effort the losses of an amplifier are of primary interest:

$$P_{loss} = P_{out} * \frac{1 - \eta}{\eta} = P_{average} * (1 - \eta) \tag{7.237}$$

Example:

Using the example of before and the requirement $I_{peak} = 1A$ we can calculate the losses:

$$P_{average_{example}} = \frac{2 * 1A * 20V}{\pi} = 12.732W$$

$$P_{loss_{example}} = P_{average_{example}} * (1 - 66.7\%) = 4.24W$$

$$P_{out_{example}} = P_{average_{example}} * \eta_{example} = 8.499W$$

The efficiency calculation can nicely be done by a simple octave script.

```
% calculation of efficiency and losses
% depending on the margin we need
Vsup=40;
Vh=Vsup/2;
Vmargin = 0:0.5:20;
Ibias=0.02;
Ipeak=1;
Iav=Ipeak/pi + Ibias;
Vpeak=Vh-Vmargin;
Pout=Vpeak*Ipeak/2;
Pin=Vsup*Iav;
n=Pout/Pin;
figure(1);
plot(Vmargin,n*100);
```

```
xlabel('Vmargin in V');
ylabel('efficiency in %');
grid;
Ploss=Pin*(1-n);
figure(2);
plot(Vmargin,Ploss);
xlabel('Vmargin in V');
ylabel('Ploss in W');
grid;
```

For the example above the little script plots the efficiency and the power dissipation of an amplifier supplied with either 40V or a symmetrical supply of +-20V versus the minimum drop required over the power transistor. The script has been enhanced taking into account the DC biasing of the amplifier too (See fourth line of the code "Ibias=0.02".

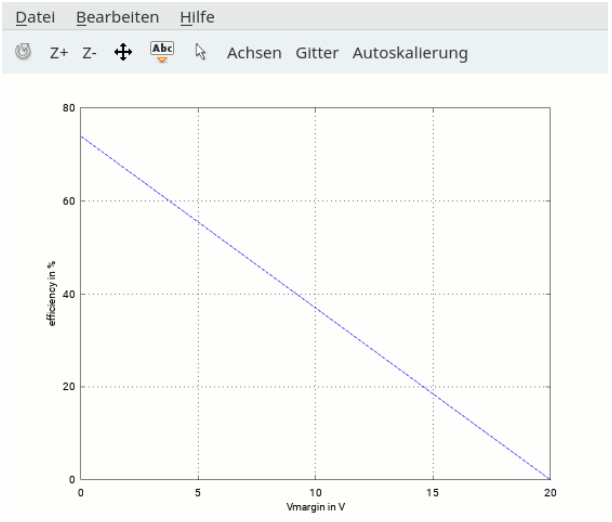


Figure 7.127: Efficiency of a linear power amplifier versus drop over the power transistors at the peak of the sine wave

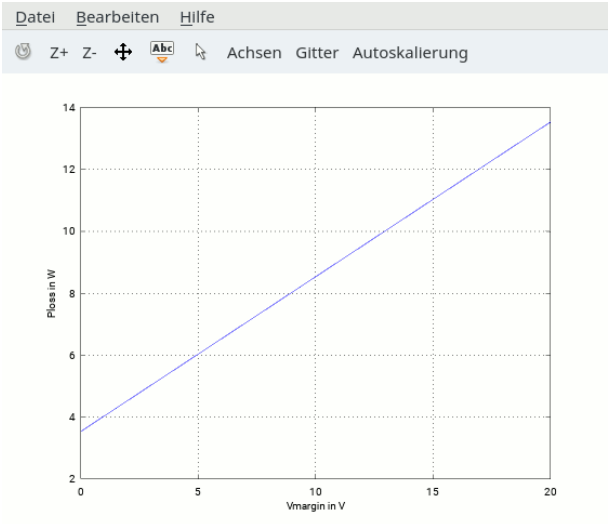


Figure 7.128: Losses of an amplifier supplied with 40V operating at a peak current of 1A versus drop of the power transistors at the peak of the sine wave

The graphs shown above apply to a load current that is independent of the output amplitude. In most cases the load is (more or less) resistive. Thus the load current decreases with the reduction of the amplitude. This leads to a reversed parabola characteristic of the power dissipation of the amplifier. The following plot shows the example of an amplifier designed for 1A peak current operating with different load resistors ranging from 15 Ohm to 19 Ohm and having a supply voltage of +-20V.

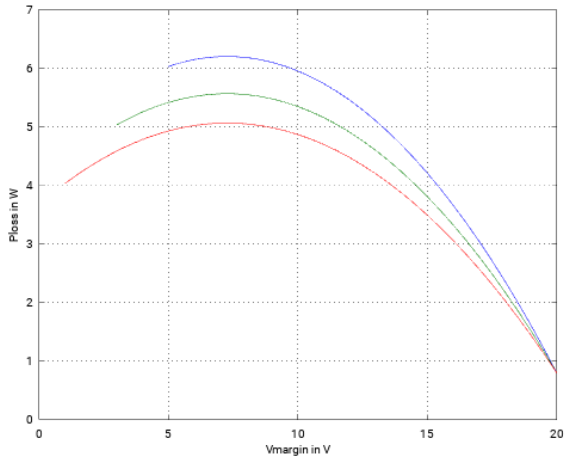


Figure 7.129: Losses of a power amplifier with resistive load (19Ω, 17Ω, 15Ω) versus the minimum voltage drop required at the power transistor operating at a supply voltage of +-20V

Since the amplifier's peak output current is limited at 1A the curves don't reach a drop voltage of 0V.

There is a trick that can be used to improve efficiency. The power amplifier can (for low frequencies) be built with multiple output transistors supplied from stacked supplies. This approach boosts efficiency over the theoretical value achieved by a simple design but switching between different stages adds distortions. (class G amplifier, proposed by Hitachi 1976, US patent 4100501)

7.7.2 Trans impedance amplifier (TIA)

A trans impedance amplifier converts an input voltage into an output current. A MOS transistor in grounded source configuration can be regarded as the most simple form of a trans impedance amplifier. Such simple single transistor trans impedance amplifiers are used for RF amplifiers in the GHz range.

In the low frequency range more symmetrical designs are preferred because the symmetry provides a better common mode rejection.

7.7.3 Operational transconductance amplifier (OTA)

This basically is the input stage of the operational amplifier. It consists of a differential pair and either current mirrors at the output or an open drain output. The CA3080 is a classical OTA with folded current mirror output. Here comes a simplified version (The real CA3080 used cascode current mirrors. But to understand the concept a simple mirror is OK.)

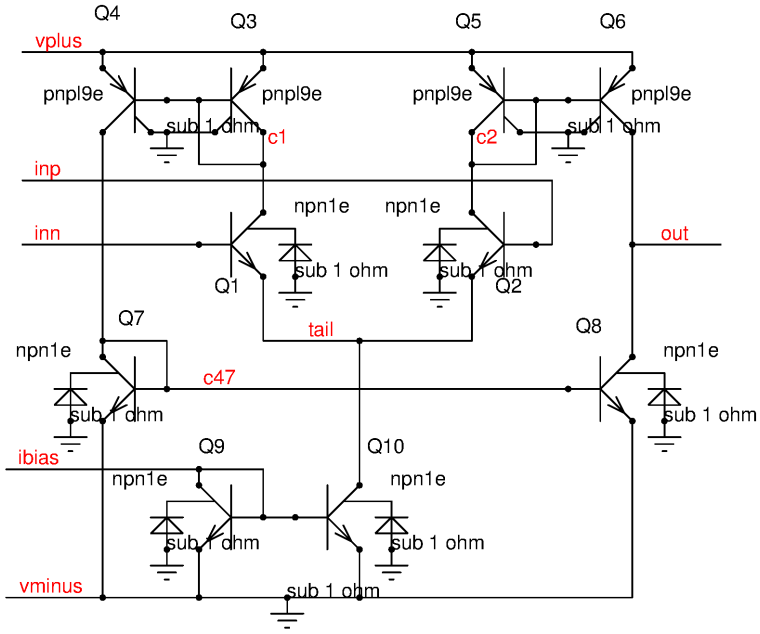


Figure 7.130: A bipolar operational transconductance amplifier (OTA)

The bias current is defined by the current fed into pin ibias. Q10 provides the bias current to the differential pair Q1 and Q2. So Q1 and Q2 each operate with half of the tail current. The differential voltage applied at the pins inp and inn is shared between both base emitter junctions of Q1 and Q2. Assuming the positive node of V_d is at inp and the negative node is at inn we find:

$$\frac{dI_{Q2}}{dV_d/2} = \frac{I_{Q2}}{V_T} \tag{7.238}$$

and

$$\frac{dI_{Q1}}{dV_d/2} = -\frac{I_{Q1}}{V_T} \tag{7.239}$$

The reason for $dV_d/2$ in the denominator can be seen in the symmetry if we split the differential input voltage in two voltages of half the amplitude.

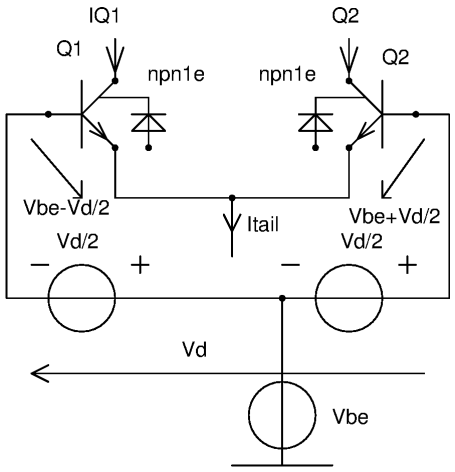


Figure 7.131: Split of Vd to calculate the transconductance of the differential stage using the symmetry

For a small differential voltage (some mV or less) Q1 and Q2 operate at about

$$I_{Q1} \approx I_{Q2} \approx \frac{I_{Q10}}{2}$$

and the linearization of the exponential function of the collector current is acceptable. Neglecting the current loss caused by the base currents we just assume the current flowing into Q3 equals the current flowing in Q4 (This is a simplification. but the error is the same as the error we make at Q5 and Q6. So it cancels). We do the same simplification for Q7 and Q8 (Well, here we really make an error. Let us come back to that later). So we can approximate the output current

$$\frac{dI_{out}}{dV_d} = \frac{dI_{Q2}}{dV_d} - \frac{dI_{Q1}}{dV_d} = \frac{I_{Q2}}{2 * V_T} + \frac{I_{Q1}}{2 * V_T} = \frac{I_{Q10}}{2 * V_T} \tag{7.240}$$

Example: We use a tail current of 26μA (So each transistor operates at 13μA). This leads to a transconductance of 26μA/52mV =0.5mA/V. Operating the amplifier with a differential input voltage of 2mVpp and a load resistor of 1MΩ the expected output signal is 1Vpp. Taking into account the losses of the current mirrors (base currents making the mirror ratio less than 1) and some non ideal emitter resistance we have to expect a little bit less than 1Vpp.

Usage of OTAs: OTAs are voltage controlled current sources. Using a frequency dependent resistor as a load (combinations of reactances) a filter can be built in an easy way.

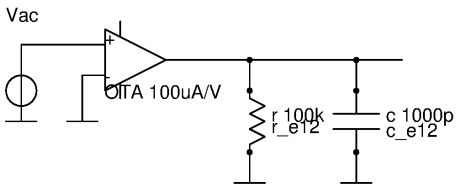


Figure 7.132: A low pass filter using an OTA

The gain of the filter can be adjusted changing the bias current.

Limitation of this first draft OTA: The OTA with the simple current mirrors suffers from the base current. The current needed to drive the collector current of Q6 is:

$$I_{Q2} = I_{Q6} * (1 + \frac{2}{B_{pnp}}) \tag{7.241}$$

The same applies to Q1:

$$I_{Q1} = I_{Q4} * (1 + \frac{2}{B_{pnp}}) \tag{7.242}$$

Additionally the path from Q1 to the output has a second current mirror. So the collector current of Q4 calculates as:

$$I_{Q4} = I_{Q8} * (1 + \frac{2}{B_{npn}}) \tag{7.243}$$

As a consequence in equilibrium (no output current, $I_{Q6} = I_{Q8}$) the ratio of the collector currents of Q1 and Q2 becomes:

$$\frac{I_{Q2}}{I_{Q1}} = \frac{B_{npn}}{B_{npn} + 2} \tag{7.244}$$

Using the Ebers Moll exponential expression of the collector current as a function of Vbe the offset becomes:

$$V_{os} = \frac{k * T}{e} * ln(\frac{B_{npn}}{B_{npn} + 2}) \tag{7.245}$$

Example: Assuming room temperature with k*T/e =26mV and a gain of the NPN transistors of 100 the systematic offset becomes

$$V_{os} = 26mV * ln(100/102) = -0.515mV$$

To avoid this systematic offset the CA3080 uses a different current mirror.

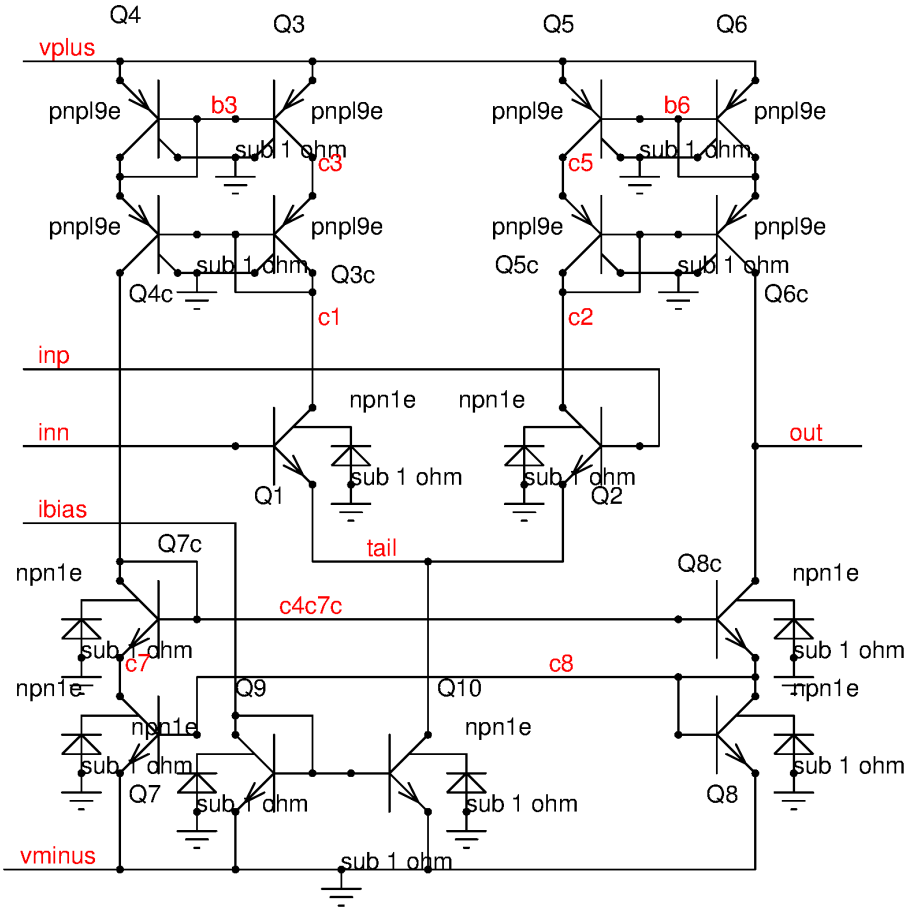


Figure 7.133: The real CA3080 including the cascode current mirrors

Besides the cancellation of the base currents this cascode current mirror has a second well appreciated side effect: The cascode makes the current source behavior more ideal. The output impedance at node out increases by several magnitudes!

Of course the same topology can be built using MOS transistors as well. Since MOS transistors have no base current the design even gets easier. The simple topology shown first will not suffer from the systematic offset caused by the base currents anymore. Operating the differential stage in weak inversion will even lead to the same exponential equations and the same linear gain dependence on the bias current. The only thing to be added to the equations in weak inversion is the little correction factor describing the gate coupling and the back gate coupling to the channel. The early voltage of the output transistors can be optimized adjusting the length of the current mirror transistors. (Current mirrors working in strong inversion do not harm). Increasing the length only is possible in a certain range. If a higher output impedance is needed cascodes still are the best option.

7.7.4 Operational amplifiers (OPAMP)

An operational amplifier is a DC amplifier with a differential input stage and an output that can either be single ended or differential. On board level the single ended output is more common. Differential output are more common for on chip usage.

7.7.5 Input stage:

The classical input stage of an opamp is a bipolar input. Most semiconductor technologies offer NPN transistors that are optimized for speed but have a low reverse base-emitter break down voltage of approximately $V_{bemin} = -7V$. If this reverse input voltage is exceeded the base emitter junction breaks down and the transistor's gain B will degrade (due to hot carriers) within a few us! For this reason NPN input stages either require special protection or they can only be used for low voltages.

PNP transistors in most cases use a low doped nwell as the base. The break down voltage (now positive because we are looking at a PNP) is much higher. $V_{bemax} = 40V$ is quite a common rating for lateral PNP transistors. The drawback of using PNP inputs often is a lower speed.

Matching of bipolar transistors in most technologies is in the range of +-0.5mV or less. Sometimes the spread mainly is determined by contaminations or inaccurate shapes of the emitter corners. Therefore some manufacturers use octagon shaped emitters to achieve better matching. (Be careful with mask grows. This can make the corners snap to the next grit point and the transistor no more has an octagon shaped emitter!). Round emitters would be even better but this leads to off grit structures and at mask making the round structure will be randomly deformed. So round emitter can not be recommended anymore. Regard the round emitter as a historical shape of the wild 1960s when the masks were cut manually.

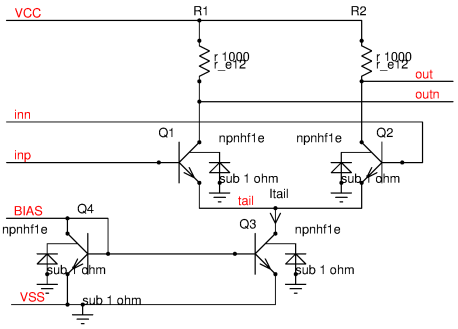


Figure 7.134: Basic NPN differential amplifier input

The transconductance of one transistor is:

$$gm = \frac{I_{tail}}{2 * V_T} \tag{7.246}$$

The voltage difference of the two inputs is shared by both transistors. One side moves up half the differential voltage while the other side moves down half the differential voltage. So the single ended gain (assuming R1=R2=R) becomes:

$$gain_{singleended} = \frac{1}{2} * R * gm \tag{7.247}$$

Since the output voltage is composed of the voltage change of both resistors the differential gain becomes:

$$gain_{diff} = R * gm = R * \frac{I_{tail}}{2 * V_T} \tag{7.248}$$

A very elegant method to produce a fixed gain amplifier stage is to create the bias current with a delta Vbe. This leads to a bias current that is proportional to VT.

$$I_{tail} = I_{bias} = ln(k) * \frac{V_T}{R_{bias}} \tag{7.249}$$

Now all parameters cancel except for the ratio of the resistors and the emitters (k) of the bias generator.

$$gain_{diff} = \frac{R}{R_{bias}} * \frac{ln(k)}{2} \tag{7.250}$$

These geometrical parameters can be kept under tight control. The concept of a delta Vbe bias generator together with a bipolar differential amplifier and resistive load is frequently used for RF and beat frequency amplifiers such as the TBA120 and it's derivatives.

The same concept can be used for a MOS differential amplifier if the MOS transistors of the bias current generator and the differential amplifier are operated in weak inversion.

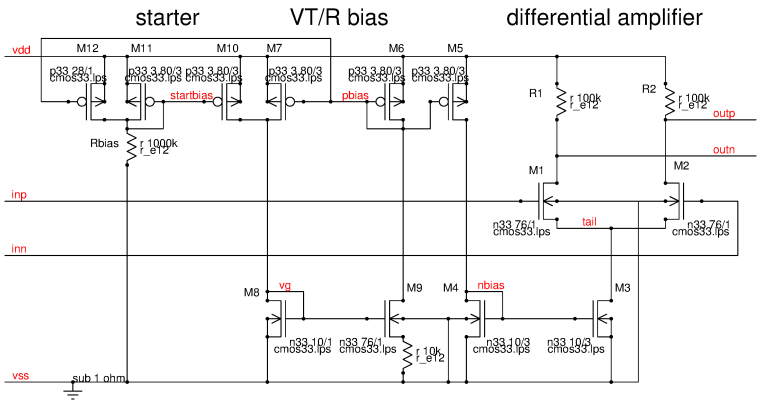


Figure 7.135: NMOS differential amplifier biased for a fixed gain

High voltage input stage: Building a high voltage amplifier input stage is always a compromise. Typically the input is built using low voltage components for better matching and higher gm. These transistors have to be protected by cascodes. The following circuit shows a possible implementation. Gate protections and antenna diodes were omitted for simplicity. The stage can handle high common mode signals. Differential mode signals may not exceed the Vgs limits of N1 and N2.

Note: The bulks of N1 and N2 are tied to the tail of the differential stage! (Otherwise the difference between bulk and source could reach break down of the junction.)

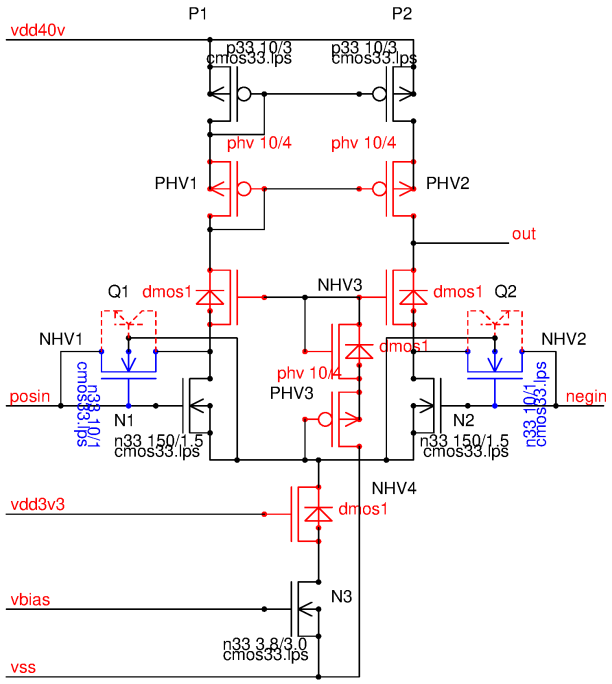


Figure 7.136: An opamp input stage designed for a wide common mode swing

The red colored transistors are the cascodes protecting the 3.3V transistors. The blue colored transistors serve as gate protections (in stead of classical antenna diodes). The main parameters (offset, gm) mainly depend on the tail current flowing through N3 and the feature sizes of N1, N2.

Warning: protection transistors NHV1 and NHV2 include two parasitic NPN transistors Q1 and Q2. These parasitic transistors will turn on if the input voltage at posin or negin drops below the bulk voltage. In this case the signal of the amplifier reverses!

rail to rail input stage: The lower the supply voltage of an operational amplifier the more attractive it gets to build a rail to rail input stage to maintain a common mode range that is as wide as possible. The rail to rail input basically is a composition of two OTAs (operational transconductance amplifiers) and an addition of the output currents. One of the OTAs has a PMOS input that is ground compatible. The other one of the OTAs has an NMOS input that is supply compatible.

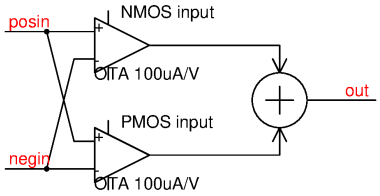


Figure 7.137: concept of a rail to rail input

Both input stages are folded cascode differential stages to make the common mode range reach the supply rails. The amplifier has 3 operating ranges:

1. Below about 1V only the amplifier with the PMOS input is operating.
2. Between about 1V and vdd-1V both amplifiers are operating and the transconductance of both stages will be added. The total gate area of the differential stage is bigger than in the other operating ranges. Usually the statistical offset is significantly better in this operating range.
3. Above vdd-1V only the amplifier with the NMOS input is working.

Having different gains in the different operating ranges leads to issues regarding the frequency compensation. For this reason usually the bias current of the still operating path is increased below 1V and above vdd-1V to circumvent this unwanted behavior. In addition the aspect ratios of the NMOS transistors and the PMOS transistors of the input stages are adjusted to compensate the different mobility of holes and electrons.

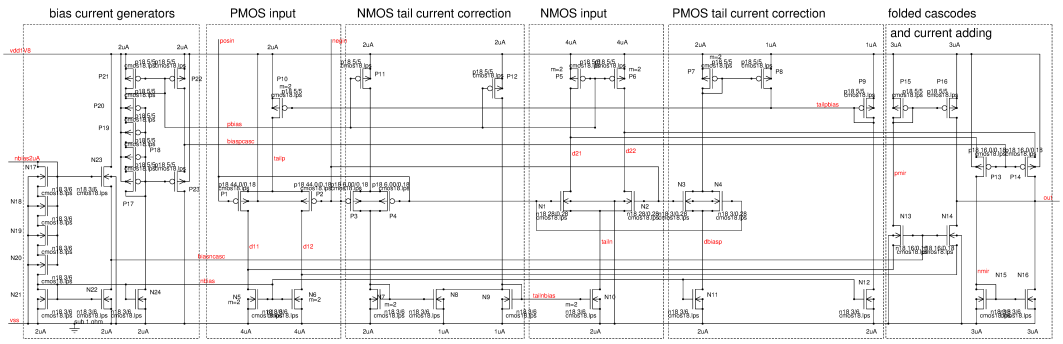


Figure 7.138: Rail to rail input

The rail to rail input shown above consists of the two differential stages P1, P2 and N1, N3. The transistors P3, P4 and N3, N4 are required to measure the common mode voltage. Current mirrors N7, N8 and P7, P8 reduce the tail currents of the differential stages as long as both differential stages operate within their common mode ranges. If the PMOS input stage leaves its common mode range P3, P3, N7, N8 turn off and the tail current of the NMOS stage increases to compensate the loss of gm. Vice versa if the NMOS input stage leaves the common mode range N3, N4 and P7, P8 turn off and the currents flowing in the PMOS stage increase.

The output currents in this simple example are added at node out. In a full blown amplifier this adding of currents is a bit more complex. In stead of just adding in one node the output currents of the folded cascode stages are used to drive a translinear output stage. This translinear stage has a quadratic characteristic and operates almost in class AB mode. This way a higher gain and higher output drive capability can be achieved. But including the translinear stage in this drawing would have made the whole circuit a bit too complex to serve as a simple example to explain the concept.

The scaling of N5, N6 and P5, P6 is not accidental. The ratio of the currents between the tail and the current mirrors must be bigger than the current increase to keep the folded cascode operating even if the tail currents reach its extremes. This however increases the impact of the current mirror mismatch on the input offset of the amplifier. The input stage operates at 1/4 of the currents flowing in the folded cascode bias generator! The relative error of

the current mirrors will be multiplied by 4 instead of the ideal factor 2 that can be chosen without the tail current correction.

Calculation of statistical errors is still possible similar to the simple folded cascode amplifier. But since now we have two signal paths and a total of 4 current generator pairs, two differential stages and 2 folded cascode pairs involved it makes more sense to use a spread sheet than to write everything into a single equation that requires several lines.

Comparing the most simple amplifiers used in 2.5V to 5V technologies, folded cascodes amplifier used for lower supply voltages and full blown rail to rail amplifiers the efforts are:

Table 31: Comparison of complexity of OPAMP topologies

type	simple	folded cascode	rail to rail
transistors	8	19	48
relative	1	2.4	6

May be this looks a bit too simple because not all transistors have the same size. But since the error propagation gets worse going to a rail to rail design the cost of a rail to rail topology will in most cases be even more then only 6 times higher than a simple low cost amplifier without rail to rail capability. Adding the effort of the translinear stage the ratio for the rail to rail amplifier approaches values in the range of 8..12 compared to the most simple design possible. Rail to rail amplifiers should only be used where it really is necessary.

7.7.6 The bread & butter OPAMPs

This section holds a brief overview of the most frequent OPAMP designs found on most mixed signal chips. The main design target is low cost and low silicon real estate. Typical supply voltages range from 2.5V to 5V depending on the technology used.

Here is an example using a 3.3V technology.

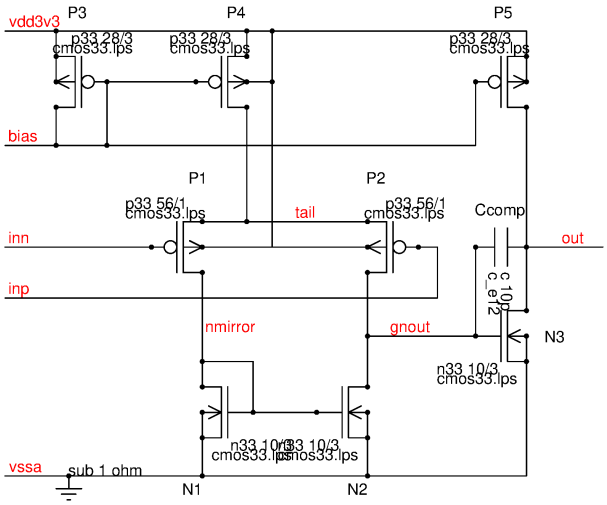


Figure 7.139: A very simple operational amplifier frequently used in mixed signal chips

Looking at the circuit we can directly see some of its limitations:

common mode range: The input stage can only be operated in a linear way in the following range:

$$V_{gsN1} + 4 * V_t - V_{gsP1} < V_{cm} < vdd3v3 - V_{gsP1} - V_{gsP4} + V_{thP4}$$

The left side of the equation determining the lower end of the common mode range assumes we need at least 4 times k*T/e as a drains-source voltage of the differential pair. This is about true if the differential stage operates close to weak inversion. The right side of the equation determines the upper limit of the common mode range. Since the current generators typically don't operate close to weak inversion here we really have to apply the gate overdrive of P4 ($V_{gsP4} - V_{thP4}$). Using a 3.3V technology a gate overdrive of about 0.5V is a usual engineering practice. Threshold voltage in a 3.3V process is typically 0.8V at room temperature. Including spread and cold temperature (-40 deg. C) the thresholds can reach up to 1.2V. Assuming the thresholds of the NMOS transistors and PMOS transistors match with an error of +-200mV the resulting common mode range becomes:

$$300mV < V_{cm} < vdd3v3 - 1.7V$$

This equation already shows why this simple circuit is not usable for supply voltages below 2.5V anymore. For lower supply voltages more sophisticated circuits with folded cascodes are required. Therefor current consumption starts to increase again for the analog parts using technologies with less than 2.5V supply voltage.

output drive: The output has a strong pull down N3 able to pull down several $100\mu A$. The pull up current is determined by the current generator P5. Typical bias currents are in the range of $1..50\mu A$. As long as the output just has to drive some CMOS gates or a high resistive feed back divider this isn't a problem. If low impedance loads have to be driven the circuit already needs some enhancements.

Maximum differential input voltage: The maximum permissible differential input voltage is limited by the dielectric strength of the gate oxide of P1 and P2. Up to about 5V the gate oxide dielectric strength normally follows the process voltage. Above 5V this is no more the case.

Slew rate: The slew rate is determined by the bias current and the frequency compensation capacitor Ccomp.

$$dV_{out}/dt = I_{bias}/Ccomp \tag{7.251}$$

Example:

with a bias current of $20\mu A$ and a frequency compensation of 10pF the slew rate becomes $2V/\mu s$.

Input offset: The statistical input offset mainly depends on the gate area of P1 and P2 and the matching parameters. As a rule of thumb (using SiO2 gate oxides) the matching is about:

$$Vos_{proc} \approx 1mV * t_{ox}/nm \tag{7.252}$$

The oxide thickness approximately scales with the process voltage.

$$t_{ox} \approx \frac{V_{gsmax}}{0.5V/nm} \tag{7.253}$$

Example:

A 3.3V process typically has a 7nm gate oxide. This leads to a matching constant of about $Vos_{3v3proc} = 7mV\mu m$.

The 1 sigma input offset voltage calculates:

$$V_{os1s} = \frac{Vos_{proc}}{\sqrt{W * L}} \tag{7.254}$$

Example:

The OPAMP shown has a 1s offset of about $V_{os1s} = 7mV\mu m/\sqrt{56\mu m * 1\mu m} = 0.935mV$. Since the manufacturer of the chip wants a good production yield and due to the number of amplifiers on one chip the specification usually states a 6 sigma value of 5.6mV.

To improve the offset, the input pair can be made bigger. Half the offset means 4 times the area and 4 times the current consumption to achieve the same speed again!

DC voltage gain: The DC voltage gain of the amplifier is the product of the gain from the input to node gnout (gain1) and from gnout to node out (gain2). Assuming the load impedance is significantly lower than the output impedance of N3 and P5 gain2 becomes:

$$gain2 = R_{load} * gm_{N3}$$

The transconductance of N3 can roughly be estimated from the aspect ratio W/L, the gate oxide thickness t_{ox}, electron mobility μ_{esi} of silicon and the bias current I_{bias}. (The equation is just the result of solving the classical MOS transistor equations for strong inversion)

$$gm_{N3} = \frac{dI_d}{dV_{gs}} = 2 * \sqrt{\frac{W}{L} * I_{bias} * \frac{\mu_{esi} * \varepsilon_{sio2}}{2 * n * t_{ox}}} \tag{7.255}$$

The factor n is the gate coupling factor. for most technologies it is in the range of 1.2..1.6. In the following let's assume n=1.4. The dielectric constant of silicon oxide is $\varepsilon_{sio2} = 0.34pAs/Vcm$. Electron mobility of silicon is about $\mu_{esi} = 600cm^2/Vs$.

Example:

Using our example of a 3.3V process and 7nm gate oxide and a bias current of $20\mu A$we get

$$gm_{N3} = 2 * \sqrt{\frac{10}{3} * 20\mu A * \frac{600cm^2 * 0.34 * 10^{-12}As}{Vs * Vcm * 2 * 1.4 * 7 * 10^{-7}cm}} = 385\frac{uA}{V}$$

Assuming we have to drive a load of $100K\Omega$ the voltage gain of the output transistor becomes 38.5 .

In a similar way we can calculate gain1. First we need the impedance of node gnout. For DC the impedance is determined by the early voltage of the shortest transistor connected to this node. Normally (like in the example) this is P2. The DC impedance depends on the early voltage. As a rough estimation the early voltage is about:

$$V_{early} = L * 10V/\mu m$$

The resulting DC resistance becomes:

$$R_{gnout} = V_{early}/I_{P2} = 2 * V_{early}/I_{bias} \tag{7.256}$$

Example:

In out example we get $R_{gnout} = 10V/10\mu A = 1M\Omega$.

To calculate the gain we have to use the same equation as before with two modifications:

1. We have to use the mobility of holes in stead of electrons $\mu_{hsi} = 200cm^2/Vs$
2. We have to use the W and L of P1 and P2

$$gain1 = R_{gnout} * 2 * \sqrt{\frac{W_{P1P2}}{L_{P1P2}} * I_{bias} * \frac{\mu_{hsi} * \epsilon_{sio2}}{2 * n * t_{ox}}} \tag{7.257}$$

Example:

$gain1 = 10^6\Omega * 2 * \sqrt{\frac{56}{1} * 20\mu A * \frac{200cm^2 * 0.34 * 10^{-12} As}{Vs * Vcm * 2 * 1.4 * 7 * 10^{-5} cm}} = 39.42$

The total gain of out bread & butter amplifier eventually is

$$gain = gain1 * R_{load} * gm_{n3} = 38.5 * 39.4 = 1516$$

or 64dB.

Not much compared to a high performance OPAMP. But does it really harm? Let's operate this bread & butter OPAMP in a closed loop and sweep the target gain of the closed loop from 1 to 1000 and observe the relative error.

$$Err_{rel} = \frac{1}{1 + gain_{amp} * gain_{fb}} \tag{7.258}$$

with $gain_{fb} = 1/gain_{ideal}$

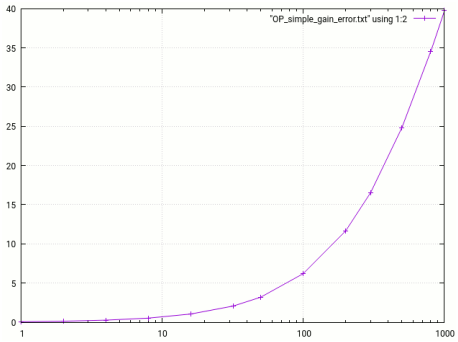


Figure 7.140: Gain error of the closed loop in % sweeping the closed loop target gain from 1 to 1000

The error versus target closed loop gain shows that as long as the closed loop is kept in the range of 1..10 the error is in the range of 1% or less. Using the bread & butter OPAMP can be justified. Using such a low performance OPAMP for higher closed loop gains will lead to problems.

Gain bandwidth product: The calculation of the gain applies to higher frequencies as well. The gain of the first stage has to drive the miller capacity Ccomp.

$$gain(f) = gm_{P1P2}/j\omega C_{comp}$$

$$gain(f) = \frac{1}{j * \omega * C_{comp}} * 2 * \sqrt{\frac{W_{P1P2}}{L_{P1P2}} * I_{bias} * \frac{\mu_{hsi} * \epsilon_{sio2}}{2 * n * t_{ox}}} \tag{7.259}$$

The unity gain frequency or gain-bandwidth-product is

$$f_{unity} = \frac{1}{2 * \pi * C_{comp}} * 2 * \sqrt{\frac{W_{P1P2}}{L_{P1P2}} * I_{bias} * \frac{\mu_{hsi} * \epsilon_{sio2}}{2 * n * t_{ox}}} \tag{7.260}$$

In our example this is $f_{unity} = 314kHz$.

Not really a race horse, but for standard applications such as a unity gain buffers for slow signals or even DC this simple amplifier is good enough and - from sales point of view much more important - it is small and cheap.

Bread & butter OPAMP for low supply voltage: Reducing the supply voltage below about 2.5V narrows down the common mode range too much. Nevertheless there is a need for OPAMPs with only about 1.2V to 2V supply voltage. The solution is two fold.

- Find a circuit that operates for common mode voltages down to 0V to not loose the 200mV at the bottom.
- Use transistors with low threshold voltages to achieve a higher upper limit.
- Lower the threshold using short channel effects.
- Operate the bias generator with a lower gate overdrive.

A folded cascode allows operating the differential stage at a drain voltage of just a few hundred mV. This way the lower limit of the common mode range can be taken down to 0V. The folded cascode design however requires more current.

Modern technologies use halo implants to make keep the threshold of the transistors constant even for short channels. In many low voltage technologies the halo implant can be masked. This masking of the halo implant leads to a lower threshold if the channel is designed very short. Some technologies additionally offer the flexibility to change the gate doping or the bulk doping to create low Vt transistors. The drawback of all these tricks is weak inversion leakage.

The next trick is to make the aspect ration very big to intentionally operate in weak inversion. This means the transistor operates at a Vgs that is lower than the threshold.

Current generators that in 3.3V technologies usually are operated at Vgs-Vth=0.5V can be operated with lower gate overdrive voltages. This lowers Vdssat at well (theoretically Vdssat=Vgs-Vth). The price for this measure is a loss of current generator accuracy.

Here comes a prototype of this design style.

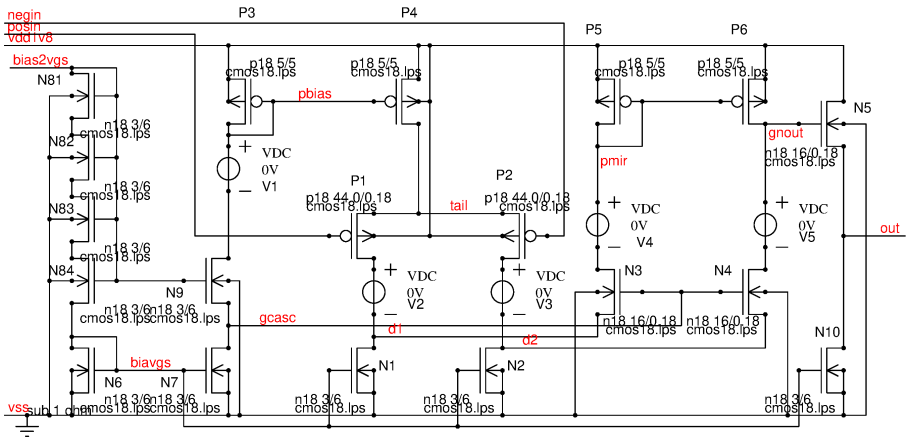


Figure 7.141: The low voltage bread & butter OPAMP

The circuit shown here includes the biasing because now the bias block (N6, N7, N8, N9) matters because it provides the gate voltage of the folded cascode. There also are some 0V sources included. These don't have a function for the circuit. Their purpose is to measure the currents flowing in the transistors. In the final circuit these voltage sources will be replaced by wires. (Well, you can just as well do the following trick: give the voltage sources the property LVS_exclude and program the cds-netlister to replace everything that has this property by a short to make LVS match).

The bias current is intended to be $2\mu A$. So each of the differential stage transistors will have a drain current of $1\mu A$. The aspect ratio W/L is very big (30/0.15=200). P1 and P2 are operating in weak inversion. This reduces the gate voltage and increases the upper end of the common mode range. Normally P1 and P2 will be designed as multiple modules or multiple fingers. Depending on the process properties this may be a critical issue. In some processes the edge of a transistor has a lower threshold than the middle. This leads to a subthreshold hump and degrades matching [24, 25]. Building differential amplifiers operating in weak inversion requires checking the layout with technology specialists to design the best modules sizes for good matching.

In the example here the W/L is simply the sum of all transistors. Assuming we did the best possible layout and we have a gate oxide of about 4nm and a matching of about $4mV\mu m$ the expected 1σ offset of the input stage becomes:

$$V_{osP1P2} = \frac{4mV\mu m}{\sqrt{W * L}} = 1.42mV$$

Bias currents and operating points: Since we plan to bias node bias2vgs with $2\mu A$the currents through N7, N1, N2 and N10 will be $2\mu A$ as well. The same applies to the PMOS mirror P3 and P4. As a consequence the transistors N3 and N4 have to supply nodes d1 and d2 with $1\mu A$. (This is not exactly true because N1 and N2 operate at a very low Vds, but as a first guess this is a good starting point.)

The ideal model of a MOS transistor distinguishes between two operating ranges.

1. Saturated operation: $V_{ds} > V_{gs} - V_{th} = V_{gseff}$
2. Triode mode: $V_{ds} < V_{gs} - V_{th} = V_{gseff}$

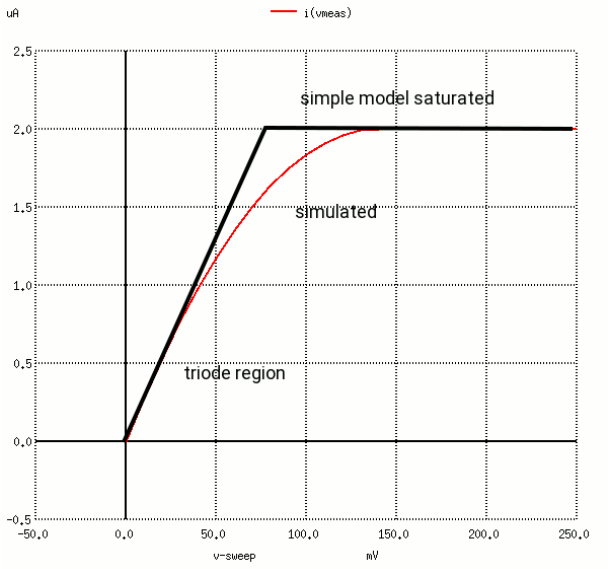


Figure 7.142: Operating ranges of an NMOS transistor working as a current generator

The black curve shows the very simple calculation with a saturated operating range and a triode region with abrupt transition between the two ranges. Real transistors have a more soft transition between the two operating regions. (red simulated curve)

The operating range is decisive for the impedance at the drain of the transistor. Ideally in saturated mode the impedance is very high because the transistor acts as a current source. In triode mode the transistor acts as a resistor and the impedance is low.

$$R_{triode} \approx \frac{V_{dssat}}{I_{dsat}} = \frac{V_{gseff}}{I_{dsat}} \tag{7.261}$$

Regarding the example amplifier operating in triode region would mean an impedance of only about $50K\Omega$ killing the gain of the first amplifier stage! For gain reasons N1 and N2 should be operated in saturated mode. On the other hand the DC voltage of nets d1 and d2 should be as low as possible. As a consequence we have to find an operating point of the folded cascode N3, N4 that takes the drain voltage of N1 and N2 slightly above V_{dssat} . This is why the bias block is part of the circuit.

Since N8 consists of 4 serial modules (N81 to N84) and the transistors N6 to N9 operate in strong inversion (quadratic characteristic) the voltage of net gcasc becomes:

$$V_{gcasc} = V_{th} + 2 * V_{gseff}$$

N3 and N4 have a much bigger aspect ratio than N6 and N7 and operate at half the current ($1\mu A$ compared to $2\mu A$ of N7) the voltage of nodes d1 and d2 becomes slightly higher than V_{dssat}. Assuming N3 and N4 operate in weak inversion the voltage can be approximated

$$V(d1) \approx V_{gseff} + V_t * ln(2 * \frac{16/0.18}{3/6}) = V_{gseff} + 26mV * 5.87 = V_{gseff} + 152mV$$

N1 and N2 operate about 150mV above the ideal transition point from triode region to saturated operation.

Having calculated the operating points with some simplifications of of the behavior of MOS transistors let's check by simulation how well the calculations fit. In the first test bench the inputs are tied to 0V.

```
vdd vdd1v8 vss dc 1.8
vinp posin vss dc 0
vinn negin vss dc 0
Ibias vdd1V8 bias2vgs dc 2u
```

The voltages at the tail node and at nodes d1 and d2 is shown below

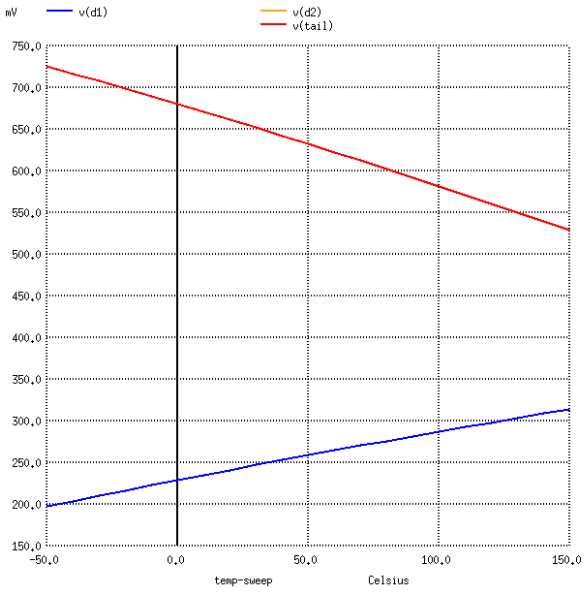


Figure 7.143: Temperature sweep of the source and drain voltage of the differential amplifier

The plot shows that the manual calculation in spite of the simplifications is close to the simulation. The distance of the red and the blue curve is the available Vds of the differential pair at a common mode input voltage of 0V.

Contribution of the current generator errors to the offset of the amplifier: To make the circuit work down to an input common mode voltage of 0V the current generators N1 and N2 must have a W/L big enough to achieve a gate overdrive of only about 100mV. This either makes the current generators fairly big (long and wide) of leads to a high current error (short, but wide). The mismatch of the current generators can be calculated.

$$I_{error} = V_{os} * gm = \frac{V_{osproc}}{\sqrt{W * L}} * 2 * \sqrt{\frac{W * K' * I_d}{L}} = 2 * V_{osproc} * \frac{\sqrt{K' * I_d}}{L} \tag{7.262}$$

with $K' = \frac{\mu_{esi} * \varepsilon_{sio2}}{2 * n * t_{ox}}$ and $\mu_{esi} = 600cm^2/Vs$, $\varepsilon_{sio2} = 0.34pAs/Vcm$, $n \approx 1.4$.

Assuming a gate oxide of $t_{ox} = 4nm$ and a process matching of $V_{osproc} = 4mV\mu m$ the 1 sigma current error of N1 and N2 becomes 25.4nA.

The current error of the PMOS mirror P5, P6 can be calculated in a similar way. The only difference is the mobility of holes in the PMOS transistor. It is about $\mu_{hsi} = 250cm^2/Vs$. P5 and P6 operate at half the bias current. So we have to calculate for $I_d = 1\mu A$. This leads to a 1 sigma error of the PMOS mirror of 11.6nA.

The total statistical current errors must be summed in non correlating way (summing power, not absolute values):

$$I_{errorPN} = \sqrt{I_{errorN}^2 + I_{errorP}^2} = 27.9nA$$

To calculate the propagation of this current error into the input offset this current error must be divided by the transconductance of the input differential stage. The input transistors work in weak inversion. This leads to a very simple equation of the error propagation.

$$V_{osmirrors} = V_t * ln(\frac{I_{errorPN} + I_{tail}/2}{I_{tail}/2}) = 26mV * ln(1.0279) = 0.72mV$$

The total input offset consists of the statistical offset of the input pair and the offset propagation of the current generators.

$$V_{os} = \sqrt{V_{osP1P2}^2 + V_{osmirrors}^2} = 1.59mV$$

This calculation shows that due to adding the errors of 2 current generators and at the same time operating the differential stage transistors with the difference of these currents the error propagation of the current mirrors is much higher than in an amplifier that avoids folded cascodes. Even worse the NMOS current sinks must be operated with a low gate overdrive! As a consequence the folded cascode amplifier requires area consuming current sinks (that can even become bigger than the input transistors) and in addition has a poor performance regarding offset errors.

Calculation of the DC gain: The DC gain is determined by the gm of the input stage and the impedance of node gnout. Under normal circumstances the impedance there mainly depends on the early voltage of the PMOS mirror. Using a channel length of $5\mu m$ we can roughly expect $V_{early} \approx 50V$. P6 operates at about half the bias current. In our example this is $1\mu A$ leading to a DC impedance of $50M\Omega$. The DC voltage gain of our example amplifier becomes:

$$gain = \frac{I_{tail}}{V_t} * \frac{V_{early}}{I_{tail}/2} \approx 2000$$

Expressed in dB this is 66dB.

Output voltage swing: Since we operate everything with low currents to achieve weak inversion operation of P1 and P2 node gnout is too high resistive to use it to drive a feedback network. A follower stage becomes mandatory. The follower stage N5 limits the output voltage swing to 0V to $vdd1V8 - V_{th}$.

Calculation of the frequency compensation: Since the highest resistive node is gnout it is usually a reasonable idea to use this node for the frequency compensation. The most simple approach is a parallel compensation with a capacitor between gnout and vss. This way we loose the pole splitting of a classical miller compensation. As a consequence the compensation must be well optimized for the load capacity (that is creating a second pole together with the output impedance at node out). Different from a miller compensation finding the best possible compensation starts to depend on the load of the amplifier as well as internal parasitic capacities. For this reason in the simple example no compensation is shown. It becomes too application specific to provide a standard solution.

Comparison of the "bread & butter OPAMP" using 2.5V to 5V supply and the "bread & butter" OPAMP using less than 2.5V supply voltage: Comparing both circuits shows:

1. Both amplifiers have similar gain in the range of 60dB
2. Below 2.5V supply the design complexity increases significantly
3. The better matching of thinner gate oxides is getting absorbed by the additional errors of the current mirrors (spread is adding while the differential stage operates with a difference of two currents)
4. The analog functions don't scale with voltage anymore
5. Making the input transistors smaller to achieve a better GBW reduces the upper limit of the common mode range
6. In most cases analog design only benefits from sub micron technologies if massive digital post processing is required for other reasons anyway

7.7.7 Instrumentation amplifiers

Most instrumentation amplifiers are based on operational amplifiers. The most common instrumentation amplifier consists of an opamp and a feedback network.

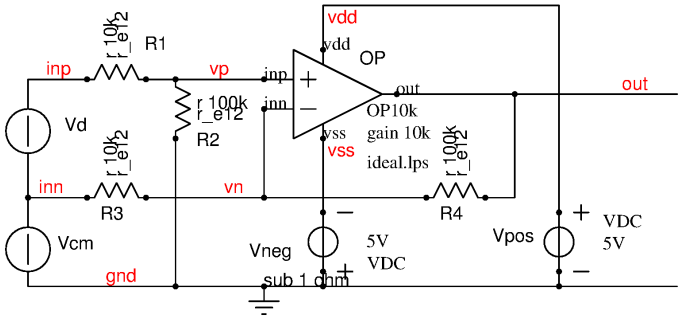


Figure 7.144: Instrumentation amplifier using one OPAMP

Ideally the gain of the operational amplifier is unlimited (or at least several magnitudes higher than the gain defined by the negative feed back. So the amplifier will always regulate vn to be equal to vp. Assuming gnd is our ground reference we can set up the equations.

$$vp = vn \tag{7.263}$$

$$vp = (Vcm + Vd) * \frac{R2}{R1 + R2} \tag{7.264}$$

$$vn = Vout + (Vcm - Vout) * \frac{R4}{R3 + R4} \tag{7.265}$$

Solving these equations for Vout yields:

$$Vout = Vcm * (\frac{R2}{R1 + R2} - \frac{R4}{R3 + R4}) * \frac{R3 + R4}{R3} + Vd * \frac{R2}{R1 + R2} * \frac{R3 + R4}{R3} \tag{7.266}$$

The most interesting case is

$$\frac{R2}{R1 + R2} = \frac{R4}{R3 + R4} \tag{7.267}$$

In this case the common mode signal Vcm will not be amplified and the output voltage of the instrumentation amplifier becomes

$$Vout = Vd * \frac{R2}{R1 + R2} * \frac{R3 + R4}{R3} = Vd * \frac{R4}{R3} = Vd * \frac{R2}{R1} \tag{7.268}$$

To obtain the same impedance at node inp and inn usually the resistors are chosen equal

$$R1 = R3 \tag{7.269}$$

and

$$R2 = R4 \tag{7.270}$$

Noise of an ideal instrumentation amplifier: Below about 10Hz the 1/f noise usually is dominant

For higher bandwidth the thermal noise of the resistors and the noise of the amplifier dominate the overall noise level. The power of all noise sources is added. Besides the resistors the input noise voltage V_N of amplifier and the noise current of the amplifier $I_N$contributes the the total noise level. Referred to the positive input inp the noise of the amplifier becomes [47]:

$$V_{N_{RTI}} = G * \sqrt{BW} * \sqrt{(V_N^2 + 4kTR_{12} + 4kTR_3(\frac{R_4}{R_3 + R_4})^2 + I_{Ninp}^2 R_{12}^2 + I_{Ninn}^2 R_{34}^2 + 4KTR_4(\frac{R_3}{R_3 + R_4})^2} \tag{7.271}$$

G is the attenuation factor of the input divider.

$$G = \frac{R_1 + R_2}{R_2} \tag{7.272}$$

BW is the bandwidth of the amplifier.

All noise voltages under the square root are noise densities in $V/\sqrt{Hz}$.

The impedance seen at the positive input of the amplifier is:

$R_{12} = (R_1 + R_2)/(R_1 * R_2)$.

Similarly the impedance seen at the negative input of the amplifier is:

$R_{34} = (R_3 * R_4)/(R_3 + R_4)$.

To refer noise produced by R_3 to the input it must be multiplied by the transfer function $R_4/(R_3 + R_4)$. This is the inverse of the noise gain. In a similar way the noise produced by R_4must be referred to the input by multiplier$R_3/(R_3 + R_4)$.

The output referred noise must be calculated using the noise gain and the inverse attenuation 1/G.

$$V_{N_{RTO}} = \frac{1}{G} * (1 + \frac{R_4}{R_3}) * V_{N_{RTI}} \tag{7.273}$$

Instrumentation amplifier with input buffer: Sometimes this is not sufficient. To achieve a higher input impedance the inputs can be buffered and even preamplified.

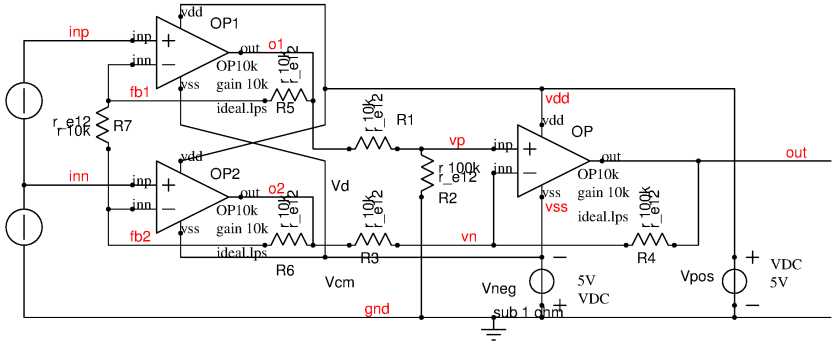


Figure 7.145: Instrumentation amplifier using three OPAMPs

The input stage consisting of OP1, OP2, R5, R6, R7 offers several features:

1. The input impedance at nodes inp and inn becomes almost infinite.
2. The common mode gain of the preamplifier stage is 1 and the common mode gain remains as before

$$gain_{cm} = 1 * (\frac{R2}{R1+R2} - \frac{R4}{R3+R4}) * \frac{R3+R4}{R3} \tag{7.274}$$

3. The differential gain of the input stage is

$$Vo1 - Vo2 = Vd * \frac{R5+R6+R7}{R7} \tag{7.275}$$

   as a consequence the total differential gain becomes

$$gain_d = \frac{Vout}{Vd} = \frac{R5+R6+R7}{R7} * \frac{R2}{R1+R2} * \frac{R3+R4}{R3} \tag{7.276}$$

4. Since the differential gain increases while the common mode gain remains as before the common mode rejection is improved.

Putting a significant part of the differential gain in the first stage dramatically relaxes the matching requirements of R1 to R4.

The common mode range of OP1 and OP2 requires attention. The output voltage swing of OP1 and OP2 is the common mode voltage Vcm plus the differential voltage Vd multiplied with the gain of the first stage! The input common mode range required is the common mode voltage Vcm plus the differential mode voltage Vd.

The common mode range of the second stage (OP) is a little bit relaxed compared to the first stage because R1, R2 (depending on the values chosen) reduce the voltage levels at nodes vp and vn.

7.7.8 Fully differential amplifiers

Fully differential amplifiers have an differential input (like a normal OPAMP) and a differential output. The most simple way to build a fully differential amplifier is the simple resistor negative feedback. One of the most famous fully differential amplifiers is the LM733 or MC733 (These are the same chips but manufactured by different companies) [29]. It was developed in the mid 1960s and it still is in use.

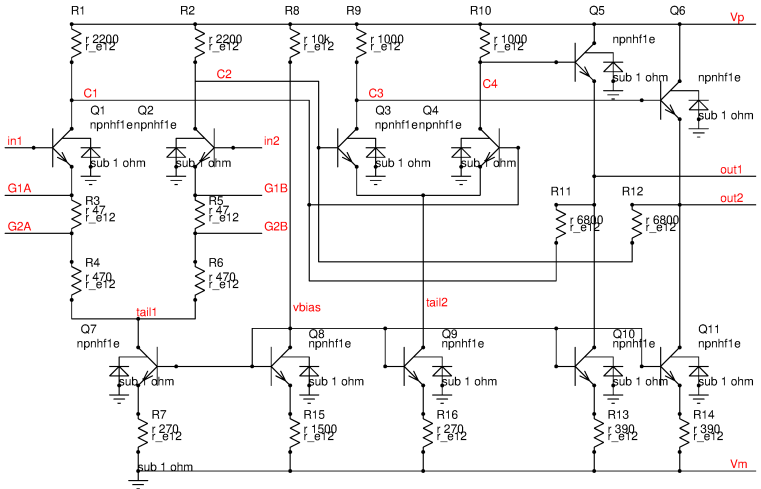


Figure 7.146: LM733 as an example of a fully differential amplifier

The gain of the amplifier depends on the ratio of the resistors R3,4,5,6 and resistors R11, R12. To adjust the gain the pins G1A, G2A, G1B, G2B can be shorted. This leads to the following gains:

Table 32: Gain settings of the LM733 amplifier

short	calculation	gain
all open	R11/(R3+R4+Vt/Ic)	10
G2A, G2B	R11/(R3+Vt/Ic)	100
G1A, G1B	R11*Ic/Vt	400

Ic is the current flowing through Q1 and Q2.

The common mode output voltage at the outputs out1 and out2 depends on the current flowing through R9 and R10 and Vbe of the transistors. It is approximately in the middle between Vp and Vn.

This kind of topology using only NPN transistors was chosen in the early times of IC design because the NPN transistors were much faster than the lateral PNP transistors. Fully differential amplifiers often are used for RF amplifiers. If the required bandwidth even doesn't allow a feedback over one stage (The voltage gain of the LM733 is mainly provided by Q3 and Q4. Input Q1 and Q2 is just an OTA converting an input voltage into a current driving R11 and R12) an even more simple approach has to be used.

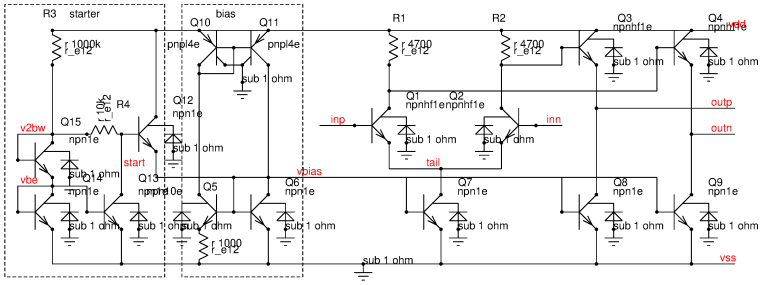


Figure 7.147: Fully differential amplifier with gain defined by resistors and gm

In this design the gain is defined by:

$$gain_{diff} = R1 * I_{tail}/V_t$$

On the first glance this looks as if the gain will decrease with temperature because $V_t$increases. The trick is to make the bias current proportional to the temperature voltage.

$$I_{tail} = ln(K) * V_t/R_4$$

Now we have an amplifier with a gain that depends on the ratio of the resistors R_1, $R_2$and $R_4$and the emitter ratio of $Q_5$and Q_6. The achievable gain of such a single stage amplifier mainly is limited by the voltage drop over the collector resistors $R_1$and R_2. The resistors can only be increased as much as the supply voltage of the amplifier and the required input common mode range permits. If more gain is needed than one stage can provide further stages can be added. This is a common approach building RF amplifiers.

The concept can be used with MOS transistors as well. But then the transistors of the bias generator as well as the transistors used for the differential stage must be operated in weak inversion to achieve an exponential characteristic like when using bipolar transistors. This however reduces the achievable bandwidth of a MOS amplifier with constant gain.

Simply operating an exponential characteristic with a bias current produced by the logarithm of the same exponential function of course only works for small signals. As soon as big input signals are present the transfer function must be linearized adding emitter resistors. Now the transconductance is defined by the resistors rather than the bias current.

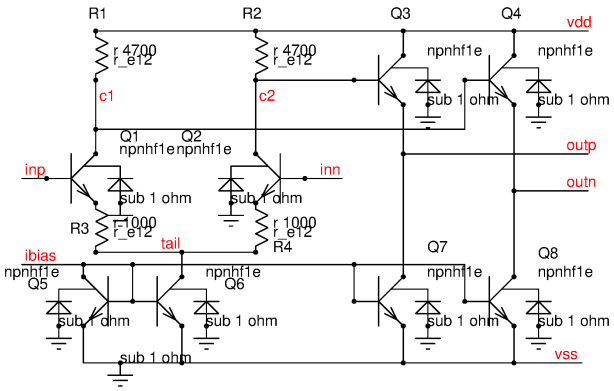


Figure 7.148: Differential amplifier with gain defined by emitter resistors

If the bias current is chosen such that the voltage drop over R2 and R3 is about 100mV or more the change of the temperature voltage only has am minor influence on the gain of the amplifier.

The two amplifiers shown have a common mode output voltage that depends on the supply voltage. This sometimes isn't desired. Often the common mode voltage at the amplifier output has to be matched to the requirements of the next stage. A common mode voltage regulation will solve this problem. In the most simple form the common mode voltage regulation simply tunes the supply voltage of the differential amplifier.

All these amplifier shown have in common that the common mode output voltage depends on the supply voltage and the Vbe of the transistors used. With decreasing supply voltages this dependence of the common mode voltage on transistor parameters could not be accepted anymore. Regulation loops were added to exactly keep the common mode voltage under control. A basic example is shown in the following figure.

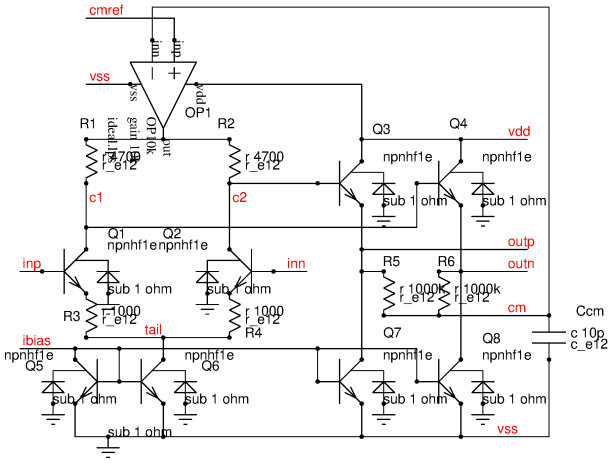


Figure 7.149: fully differential amplifier with common mode regulation

The resistors R5 and R6 measure the output common mode voltage. The operational amplifier regulates the supply of the differential stage to reproduce the common mode reference cmref as the common mode voltage at outp and outn. Ideally both amplifiers don't influence each other. If there is a resistor mismatch the common mode regulation however will see some of the differential signal. For this reason the common mode regulation usually is designed significantly slower than the differential amplifier itself to keep the pole of the common mode regulation as the dominant pole no matter what happens in the differential amplifier.

In MOS circuits we usually don't want to rely on the weak inversion characteristics (mainly for speed reasons). Here usually different circuit topologies are chosen. Here is a simple conceptual example using an OTA (operational transconductance amplifier) with differential output. The OTA always forces a differential voltage of 0V between nodes np and nn. Using equal resistors this just is a differential buffer. Since the differential voltage at np, nn is 0V this buffer has an input current acting as a load at the driving stage.

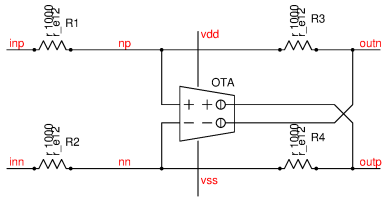


Figure 7.150: fully differential buffer using an OTA

If we want to make the input high resistive we have to add the inverted current at the input side. This cancels the current flowing into R3 and R4.

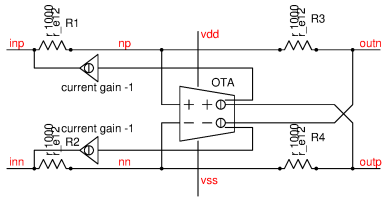


Figure 7.151: Fully differential amplifier with current cancellation

The circuit looks kind of strange, but the transistor implementation using current mirrors is surprisingly easy. Here comes a first conceptual try not yet including cascodes and common mode regulation.

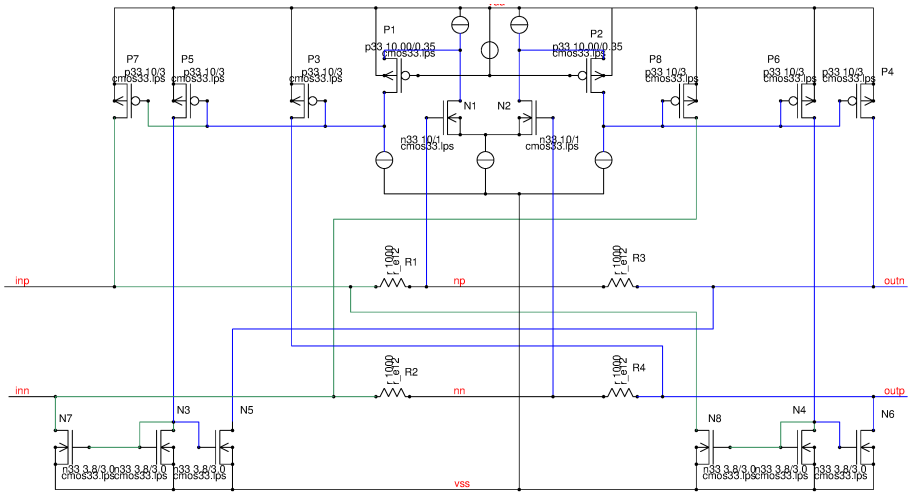


Figure 7.152: Fully differential amplifier using a CMOS OTA and current compensation to make the inputs high resistive

In the CMOS circuit the signal path is colored blue while the current compensation path is colored green.

7.7.9 Stability of an amplifier inside a regulation loop

Most amplifiers are operated inside some kind of a feedback loop. The loop consists of a forward gain of the amplifier and a transfer function of the feedback loop. Additionally the gain may be affected by the load impedance (e.g. if the forward path is an OTA!). A simple example driving a capacitive load may look like this:

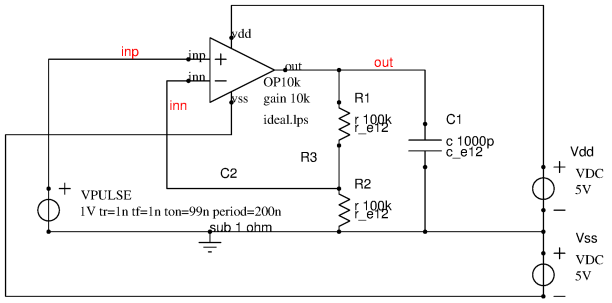


Figure 7.153: A typical application of an operational amplifier

The closed loop gain can be described as

$$gain_{closed} = \frac{gain_{amp}}{1 + \frac{R_2}{R_1+R_2} * gain_{amp}} \tag{7.277}$$

The system becomes unstable when the denominator becomes 0. To make life a bit easier the feedback network transfer function can be replaced by β.

$$\beta = \frac{R_2}{R_1 + R_2} = \frac{1}{k} \tag{7.278}$$

$$gain_{amp} = -\frac{1}{\beta} \tag{7.279}$$

In other words the amplifier must shift the phase by 180 degrees in excess of the inversion. Replacing the transfer function of the amplifier by a system of two poles the gain depending on the complex frequency s becomes:

$$gain(s) = \frac{gain_0}{(1 + \frac{s}{\omega_{p1}}) * (1 + \frac{s}{\omega_{eq}})} \tag{7.280}$$

s is the complex frequency

$$s = j * \omega \tag{7.281}$$

The second pole ω_{eq} is an equivalent pole approximating a multiple pole system.

$$\frac{1}{\omega_{eq}} = \frac{1}{\omega_{p2}} + \frac{1}{\omega_{p3}} + ... \tag{7.282}$$

For $\omega >> \omega_{p1}$ the equation can be approximated by:

$$gain(s) = \frac{gain_0 * \omega_{p1}}{s * (1 + \frac{s}{\omega_{eq}})} \tag{7.283}$$

The product $gain_0 * \omega_{p1}$is called the gain bandwidth product GBW.

$$GBW = gain_0 * \omega_{p1} \tag{7.284}$$

$$gain(s) = \frac{GBW}{s * (1 + \frac{s}{\omega_{eq}})} \tag{7.285}$$

The closed loop gain becomes:

$$gain_{closed}(s) = \frac{1}{\beta} * \frac{1}{1 + \frac{s}{GBW} + \frac{s^2}{GBW * \beta * \omega_{eq}}} \tag{7.286}$$

$$gain_{closed}(s) = \frac{k}{1 + \frac{s}{Q * \omega_0} + \frac{s^2}{\omega_0^2}} \tag{7.287}$$

with:

$$\omega_0 = \sqrt{\beta * GBW * \omega_{eq}} \tag{7.288}$$

and

$$Q = \sqrt{\frac{\beta * GBW}{\omega_{eq}}} \tag{7.289}$$

The loop gain is the feedback factor β multiplied with the gain of the amplifier $gain_{amp}$:

$$\beta * gain_{amp}(s) = \frac{\beta * GBW}{s * (1 + \frac{s}{\omega_{eq}})} \tag{7.290}$$

The phase margin has to be calculated at the cross over frequency ω_t where the loop gain becomes 1.

$$\beta * gain_{amp}(\omega_t) = 1 \tag{7.291}$$

$s = j\omega_t$ leads to expression:

$$1 = \frac{\beta * GBW}{j\omega_t * (1 + \frac{j\omega_t}{\omega_{eq}})} \tag{7.292}$$

Solving for the amplitude we get

$$\beta^2 * GBW^2 = \omega_t^2 * (1 + \frac{\omega_t^2}{\omega_{eq}^2}) \tag{7.293}$$

Taking the square root and dividing by ω_{eq} yields:

$$\frac{\beta * GBW}{\omega_{eq}} = \frac{\omega_t}{\omega_{eq}} * \sqrt{1 + \frac{\omega_t^2}{\omega_{eq}^2}} \tag{7.294}$$

This is exactly the square of the quality factor Q. So Q can be expressed as

$$Q = \sqrt{\frac{\omega_t}{\omega_{eq}} * \sqrt{1 + \frac{\omega_t^2}{\omega_{eq}^2}}} \tag{7.295}$$

The phase margin of the system is:

$$phasemargin = -\frac{\pi}{2} - atan(\frac{\omega_t}{\omega_{eq}}) \tag{7.296}$$

To make life a bit easier here is a little octave script calculating the relationship:

```
% Overshoot as a Function of Phase Margin
% x = wt/weq
x = .27:.01:1.0;
pm = 90-(180/pi)*atan(x);
q=sqrt(x.*sqrt((1+x.^2)));
os=100*exp(-pi./(sqrt(4*q.^2-1)));
plot(pm,os)
title('Overshoot');
xlabel('Phase Margin (degrees)');
ylabel('Percent Overshoot');
grid;
```

The script produces a nice graph of the overshoot versus the phase margin. So a closed loop stability test can use the overshoot of the system to estimate the phase margin.

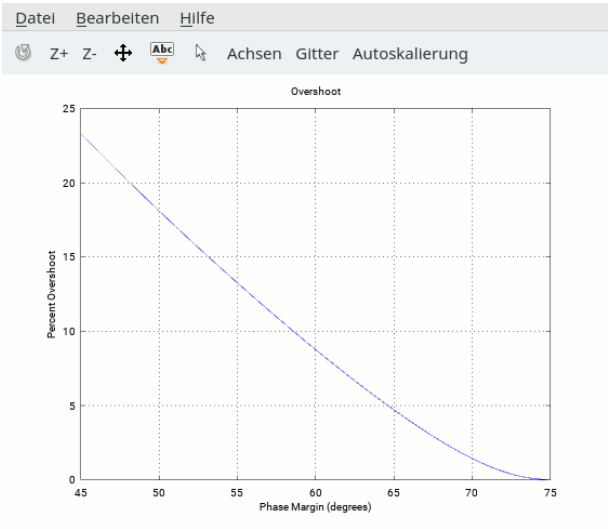


Figure 7.154: Overshoot versus phase margin plot

For most applications a phase margin of 60 degrees is desired. In extreme cases (high capacitive load) 45 degrees may be acceptable in the worst case corner.

Stability of multi stage systems If the regulation loop consists of more than one gain stage things become more difficult. Let's assume we have an operational amplifier followed by a second gain stage. The second gain stage could for instance be a high voltage output stage. Here is a simple example:

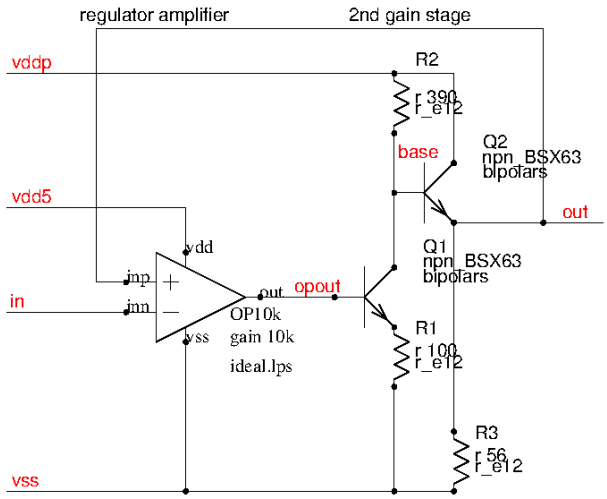


Figure 7.155: Two gain stages in one loop

In this example the OPAMP is supplied from a low power, low voltage supply but the power stage is supplied from a high voltage supply. So the miller compensation may not be connected to signal base or signal out. The frequency compensation is completely inside the OPAMP.

Since R1 and R2 have a ratio of 3.9 the OPAMP must be stable down to a gain of 1/3.9. This produces a significant problem! A classical miller compensation consists of an OTA (operational transconductance amplifier) and an output stage driven by the OTA. This kind of compensation works well as long as the gain is above 1. When the gain drops below 1 the AC path through the compensation becomes dominant over the gain of the output stage - but the phase then turns 180 degrees.

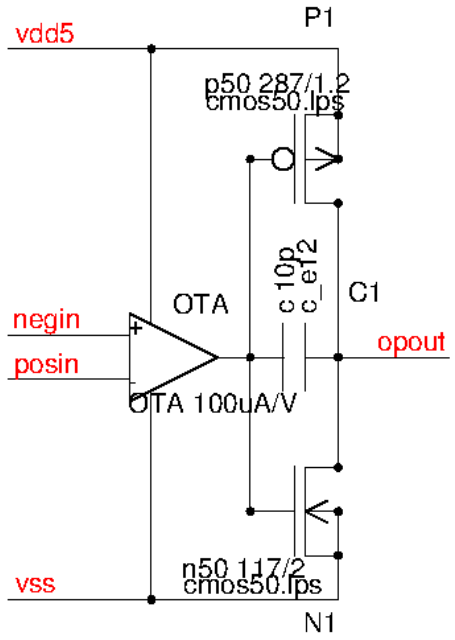


Figure 7.156: Standard OPAMP design for unity gain stability

This standard design can't be made stable in a loop with an additional stage as shown before. We need an amplifier that doesn't have a path bypassing N1 and P1 when the gain drops below 1. The capacitor must be decoupled from the output.

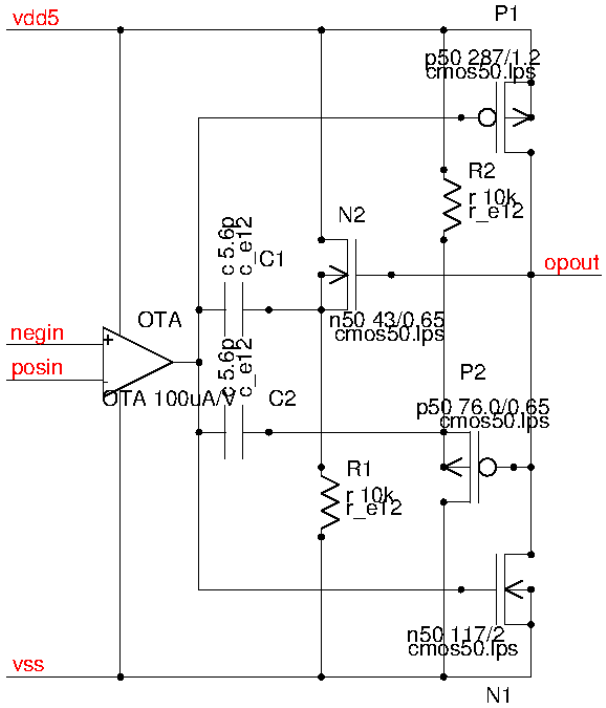


Figure 7.157: Amplifier with buffered miller compensation

The buffer N2, P2, R1, R2 prevents bypassing N1 and P1 with the wrong phase via C1 and C2. The buffer must be very fast in order not to add phase shift in the miller path. In addition the output impedance of the buffer has to be very low. Typically a simple source follower is used. The current consumption of the buffer can even be higher than the current consumption of the OTA! Therefore buffered miller compensation isn't a standard feature of most OPAMPs. It is a specialized solution for applications that can't avoid the second gain stage.

Simulation of stability Stability and phase margin can just as well be analyzed in frequency domain directly. In this case the feedback path is simply cut open. In simulation this can be done easily while opening the loop of a real design may be a major problem because pads are needed to insert the blue components of the following figure. These pads have a significant parasitic capacity and therefore will affect the measurement.

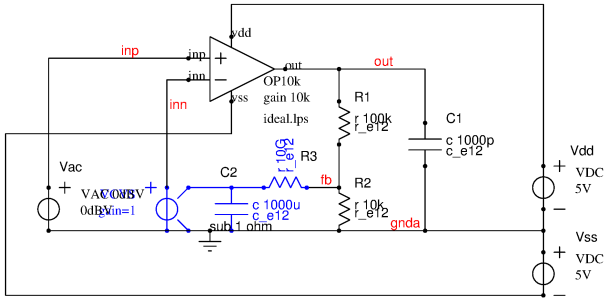


Figure 7.158: Cut of the feedback loop to measure gain and phase directly in an AC simulation

The AC simulation linearizes the whole circuit at the operating point. Giving the source Vac a magnitude of 0dBV will directly output the open loop gain of the amplifier at note out.

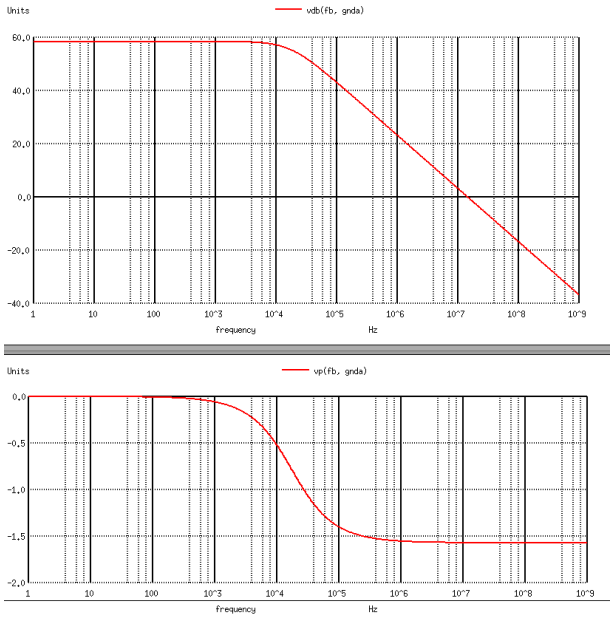


Figure 7.159: Gain (in dB) and phase (in rad) of an opamp with capacitive load

As soon as the gain rolls off the opamp together with the load acts as an integrator. So the phase shifts be -90 degrees or $-\pi/2$. A 180 degree shift that would make the amplifier with feedback unstable corresponds -180 degrees or $-\pi$. Thus the phase margin at 14MHz (here the loop gain is 0dB or 1) is 90 degrees or $\pi/2$.

(0dB correspond a gain of 1, for stability the tap between R1 and R2 - that is connected to inn in the application schematic - is of interest.)

7.7.10 Comparators and Schmitt trigger circuits

Comparators and Schmitt trigger circuits have an analog input and a digital output. Ideally the digital output can only have two states: A logic one or a logic 0. These two states in most cases are represented by voltage levels. If the input voltage is above the trigger level the output of the comparator is in one state. If the input voltage is below the trip point the output is in the opposite state.

Ideal Schmitt trigger: An ideal comparator can be thought of as an amplifier with unlimited gain. The problem is to achieve an unlimited gain. In practical design this is not possible and other states than the logic 1 and he logic 0 become possible due to the limited gain. Thus the unlimited gain must be established adding a positive feed back.

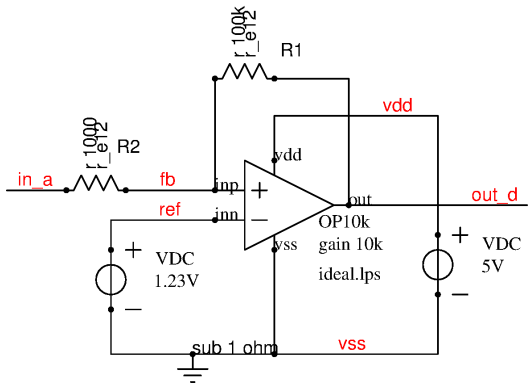


Figure 7.160: Concept of a comparator using an amplifier with limited gain and a positive feed back

Let's assume we have a ramp from 0V to 5V applied at pin in_a. The output out_d will change from 1 to 0 when the voltage at node fb crosses 1.23V (the reference voltage applied at pin ref). Due to the divider R1 and R2 and the initial output voltage of 0V this will happen when the voltage at in_a crosses:

$$V_{tripH} = V(ref) * \frac{R1 + R2}{R1} \tag{7.297}$$

In the above example this happens at $V_{tripH} = 1.01 * 1.23V = 1.2423V$.

Next step let us assume we have a falling ramp from 5V to 0V. Since we are starting from a high level the initial voltage at he output out_d is equal V(vdd). This increases the voltage at node fb.

$$V(fb) = V_{in} * (1 - \frac{R2}{R1 + R2}) + vdd * \frac{R2}{R1 + R2} \tag{7.298}$$

The trip point for the falling edge becomes:

$$V_{tripL} = V_{ref} + \frac{R2}{R1} * (V_{ref} - vdd) \tag{7.299}$$

In our example this falling trip point calculates as $V_{tripL} = 1.23V + 0.01 * (1.23V - 5V) = 1.1923V$.

Well, in this simplified calculation we assumed the gain of the amplifier to be much higher than the ratio $\frac{R1}{R2}$. This is a simplification that will not contribute too much error as long as the gain of the amplifier is at least 2 magnitudes higher than the resistor ratio.

The amplifier in combination with a positive feed back is called a Schmitt trigger.

The difference of the trip points V_{tripH} and V_{tripL} is called the hysteresis of the Schmitt trigger.

$$V_{hyst} = V_{tripH} - V_{tripL} = \frac{R2}{R1} * vdd \tag{7.300}$$

The simple topology shown here has the disadvantage that the positive input has to provide the current for the resistor network R1 and R2. If the source driving in_a is high resistive the source resistance must be added to R2. So the source resistance changes the hysteresis and the trip points!

To overcome this draw back other topologies are used in practical design.

A second problem can be the slow edges at the output. For this reason it is common practice to make the switching edges as fast as recommended for the logic the comparator has to drive.

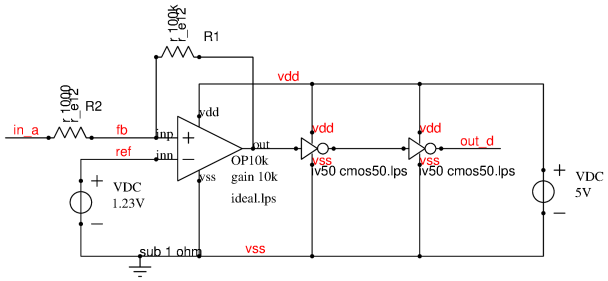


Figure 7.161: Ideal comparator with buffer to drive the logic with faster edges

2 comparator Schmitt trigger: The 2 comparator Schmitt trigger is a standard approach for precision designs. One of the first implementations was the famous 555.

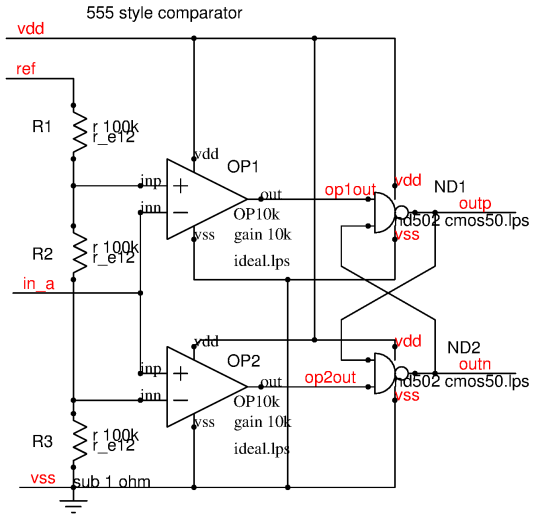


Figure 7.162: Precision Schmitt trigger with two amplifiers

This kind of Schmitt trigger offers a well defined hysteresis but (ideally) has no feed back on the input voltage applied at in_a. In case of the original NE555 of 1971 [32] the nodes ref and supply vdd are shorted internally. The original NE555 used a bipolar input stage. So deviating from ideal there still in an input current flowing into the base of the differential transistor bases. But this current is already several magnitudes lower than the current flowing through R1 and R2 of the conceptual implementation shown before.

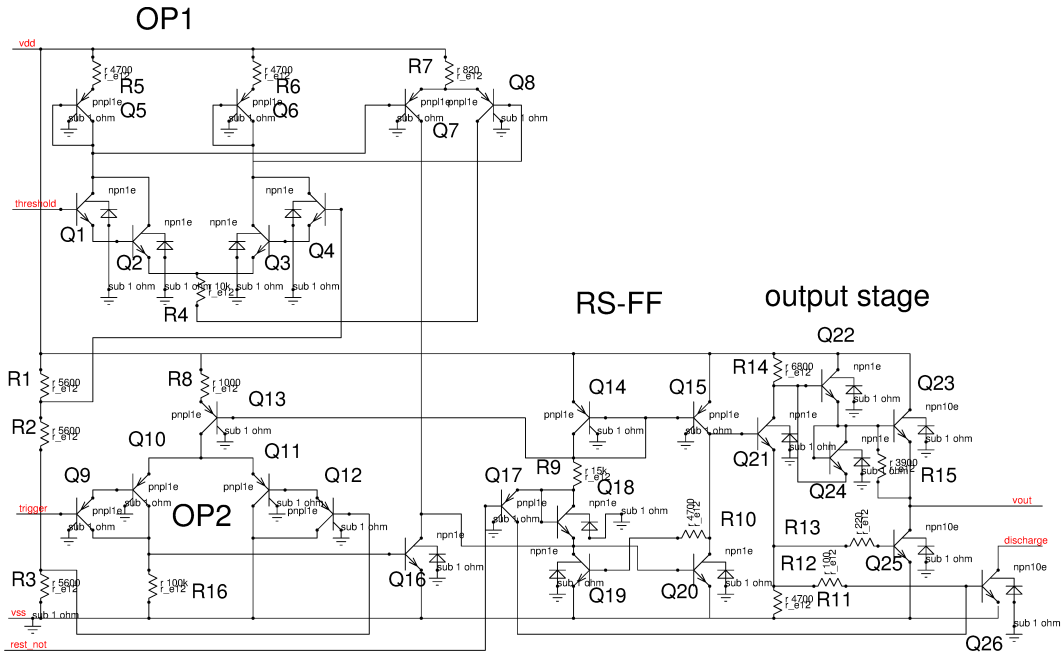


Figure 7.163: Original NE555 circuit

Low current consumption 2 comparator Schmitt trigger: The same concept can be built using CMOS components. Even better: It is possible to build a precision comparator that stops consuming current when the input voltage approaches the supply rails. The basic idea is to only activate the amplifier needed for the next switching event. The second amplifier (that had triggered the switching event before) is deactivated to reduce current consumption.

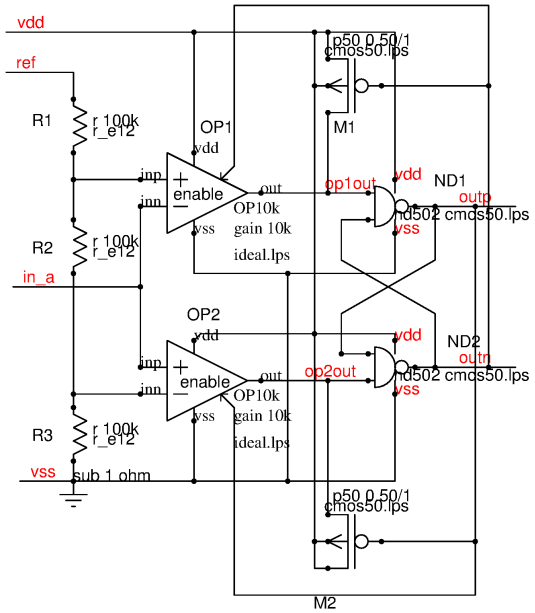


Figure 7.164: Low consumption precision Schmitt trigger

OP1 is designed such that at low input voltage it does not consume current. Close to the trip point and above of course there is current consumption. As soon as OP1 sets the latch OP1 gets turned off to reduce current consumption and M1 pulls the input of the latch beck into the non dominant state.

After turning off OP1 the opposite OP2 is activated. It is designed in a way that at high input voltage it does not consume current. Current consumption of OP2 starts approaching the lower threshold. As soon as OP2 resets the latch OD2 is disabled and M2 pulls the (now floating) output of OP2 into a non dominant state again.

Care must be taken dimensioning the hold transistors M1 and M2. Turn on of M1 or M2 acts as a negative feed back. The set and reset inputs of the latch may not be pulled up again before the latch is fully flipped. This can be achieved by the following means:

1. Make M1 and M2 long to allow overriding them as long as the amplifiers still are active (during the turn off event)
2. Turn on M1 and M2 with a certain delay to guarantee sufficient hold time for the inputs of the latch.
3. Place a buffer with hysteresis between the drains of M1 and M2 and the inputs of the latch.

The following circuit shows one practical implementation of the circuit.

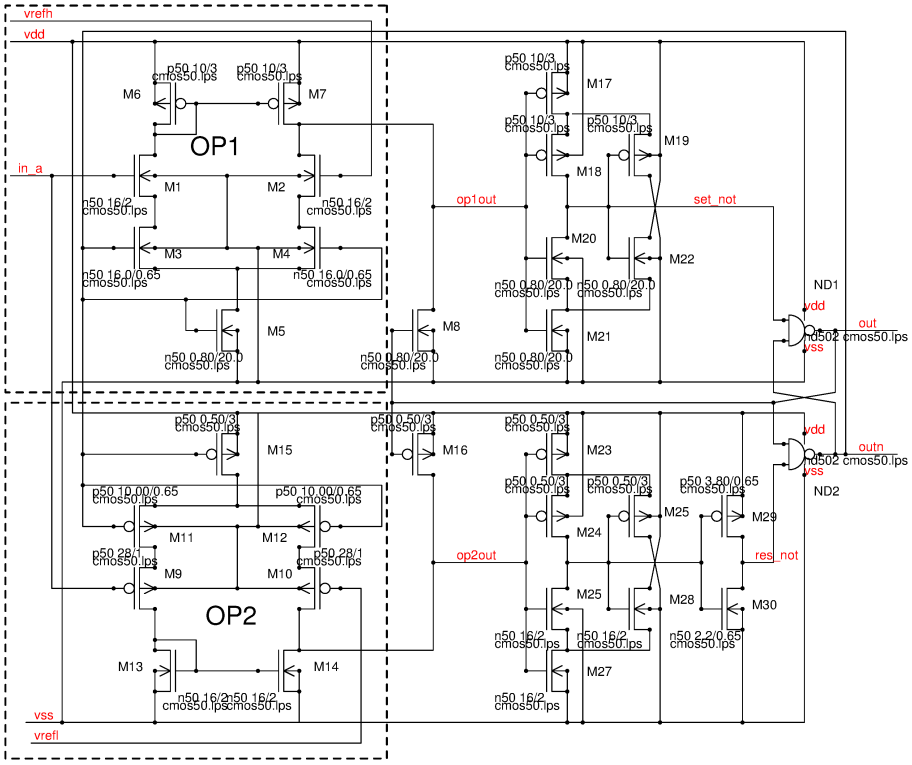


Figure 7.165: CMOS comparator without current consumption when the input signal approaches the supply rails

The two input amplifiers OP1 and OP2 are exactly complementary. OP1 is designed not to consume current as long as the input signal is below the threshold of M1. OP2 is designed not to consume current as long as the input signal is above vdd-Vth of M9. Since OP1 is turned off after crossing the threshold vrefh we need a hold transistor that pulls down the (now floating) net op1out.

OP2 is treated in the same way. crossing vrefl OP2 gets turned off and M16 pulls up the (now floating) net op2out.

Inside the amplifier stages switches M3, M4 and M11, M12 are needed to prevent current flow from the hold transistors into the MOS diodes M13 and M13 (via the transistors of the differential amplifiers that are still driven at the gates).

To prevent any ambiguous analog voltages at the input of ND1 and ND2 the res_not and set_not signals are interfaced to op1out and op2out by inverters with hysteresis.

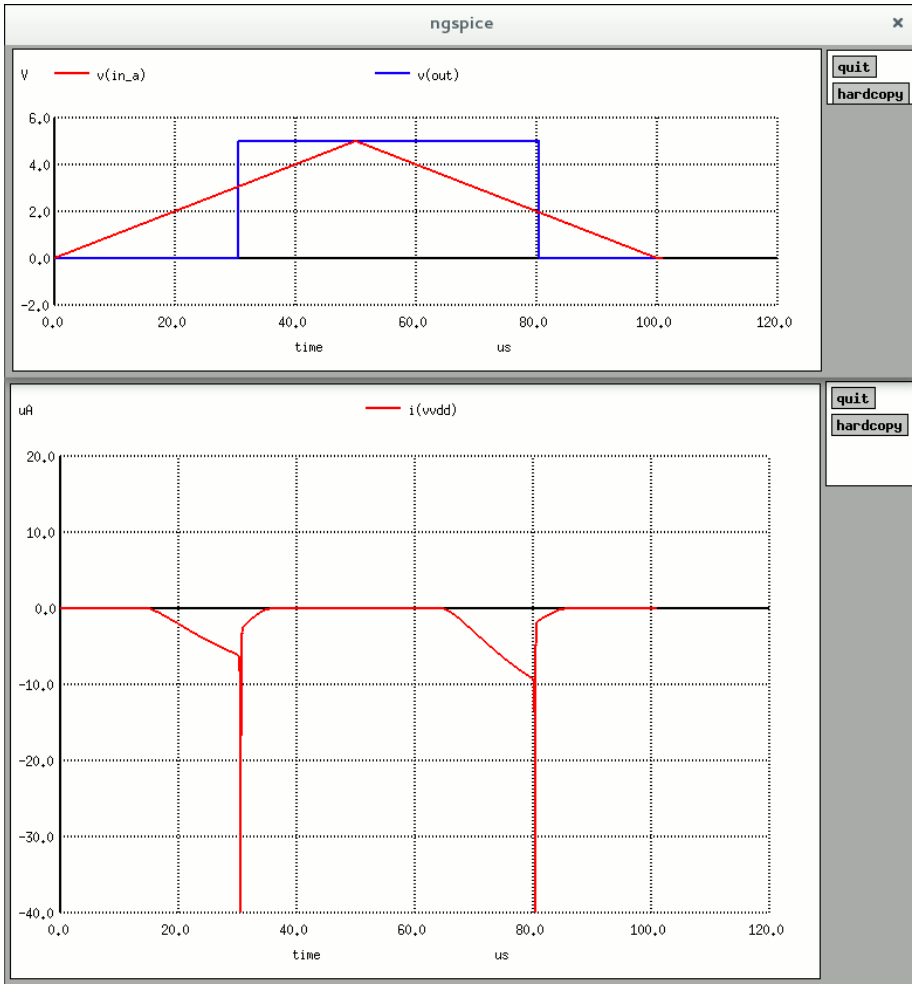


Figure 7.166: Simulation result of the comparator

In the simulation the current consumption peaks at the trip points. Ramping up node in_a the current consumption is 0 up to about 1.4V. Exceeding 3V OP1 gets deactivated and we only have the current consumption of OP2 until note in_a reaches about 3.6V. Ramping down the current consumption of OP2 starts at about V(in_a)=3.6V. At 2V OP2 gets deactivated. From 2V down to 1.4V OP1 consumes current. Below about 1.4V the current consumption of OP1 drops to zero.

Cost reduction: Building two amplifiers in stead of one consumes more silicon real estate than necessary. If current consumption can be tolerated the number of amplifiers can be reduced to 1 using transmission gates.

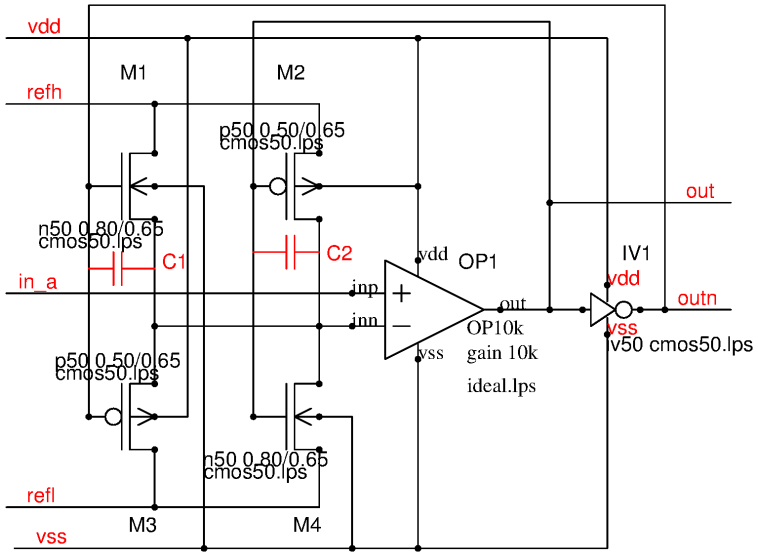


Figure 7.167: Schmitt trigger using one amplifier and transmission gated to produce the hysteresis

In this fairly simple circuit a risk of oscillation is hidden! The miller capacity of the transistor (parasitic capacitors C1 and C2) can act as a feedback that makes the Schmitt trigger oscillate close to the trip points! To prevent oscillation the Schmitt trigger it is suggested to:

1. Make the hysteresis as big as possible.
2. Use low resistive sources for refl and refh (RF impedance!).
3. Use minimum transistors for M1 to M4.
4. Match the sizes of the gates (gate capacities) of M1, M4 and M2, M3.
5. A capacity from the negative input of the amplifier to ground reduces feed through of the switching signal into the amplifier.
6. Don't make the amplifier faster than required by the application.
7. Use an amplifier with inherent hysteresis that is bigger than the capacitive feed through spike ant the negative input.
8. If nothing helps add a low pass filter between the gates of M1 to M4 and the driving amplifier and inverter.

Propagation delay of a comparator: The circuit shown above is a nice example to show the change of the comparator delay with the overdrive. The overdrive is the voltage the input signal exceeds the threshold. For a better understanding what happens lets replace the ideal amplifier by a transistor level design. The replacement is in the dashed box.

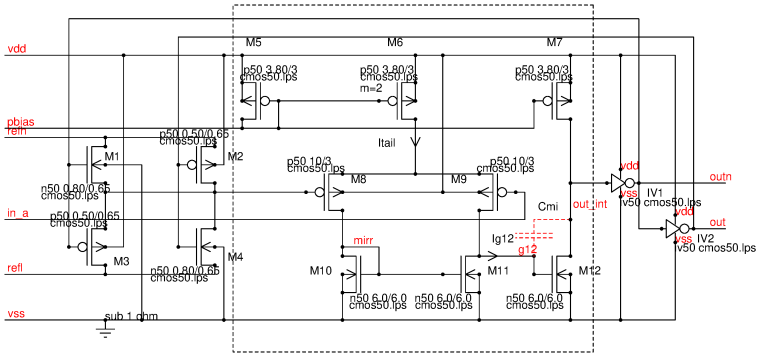


Figure 7.168: Replacing the ideal amplifier with a transistor level circuit

The miller capacity Cmi is the most important capacity limiting the speed of the comparator. Cmi is a parasitic capacity (between the drain and the gate of M12). This is why it is drawn dashed.

To flip the hysteresis the node out_int must either be charged from 0V to about vdd/2 (rising edge) or from vdd down to vdd/2 (falling edge). The time needed to charge the capacity Cmi is the dominant delay of the comparator. charging time becomes

$$t_{charge} \approx \frac{C_{mi} * vdd}{2 * I_{g12}} \tag{7.301}$$

The capacity Cmi is non linear. This is why the equation holds a $\approx$ instead of a $=$ sign. To get a feeling what happens we can calculate with some kind of an average capacity.

The current driving the node g12 is the difference between the drain current of M9 and the drain current of M8. Exactly at the trip point both currents are equal and the comparator will be infinitely slow. As soon as the signal at node in_a drops slightly below refl the current through M9 exceeds the current flowing through M8. This difference depends on the voltage difference between refl and in_a and the transconductance of the differential amplifier M8, M9.

Vice versa if the voltage if in_a is slightly higher than the voltage of refh the current through M8 will exceed the current through M9.

The minimum delay is reached when either M8 or M9 takes over the complete tail current Itail. Neglecting the propagation delay of the two inverters we get:

$$t_{dmin} = \frac{C_{mi} * vdd}{2 * I_{tail}} \tag{7.302}$$

The current through M8 and M9 is determined by the gate voltage of the transistors, the aspect ratio and the technology parameter k'. (k' depends on thee oxide thickness and the carrier mobility)

$$I_d = k' * \frac{W}{L} * V_{gseff}^2$$

With $V_{gseff} = V_{gs} - V_{th}$and $k' = \mu\varepsilon_{sio_2}/2nt_{ox}$, n=1.2..1.6. At the equilibrium point M8 and M9 both carry the same current. the resulting gate overdrive at this point becomes:

$$V_{gs0} = \sqrt{\frac{I_{tail} * L}{2 * k' * W}} \tag{7.303}$$

or reordered:

$$I_{tail} = 2 * k' * \frac{W}{L} * V_{gs0}^2$$

The gate voltage for other operation points can be described by this equilibrium voltage and an overdrive voltage V_{od}.

$$V_{gseff} = V_{gs0} + V_{od} \tag{7.304}$$

The tail current must always match the sum of the currents through M8 and M9. The output current charging or discharging the parasitic capacity Cmi is the difference of these two currents.

$$I_{tail} = I_{M8} + I_{M9} \tag{7.305}$$

$$I_{g12} = I_{M9} - I_{M8} \tag{7.306}$$

The current through M8 can be replaced by the current through M9 leading to

$$I_{g12} = 2 * I_{M9} - I_{tail} \tag{7.307}$$

Describing the current through M9 by the equilibrium voltage and the overdrive yields

$$I_{g12} = 2 * k' * \frac{W}{L} * (V_{gs0} + V_{od})^2 - I_{tail}$$

$$I_{g12} = 2 * k' * \frac{W}{L} * (V_{gs0}^2 + 2 * V_{gs0} * V_{od} + V_{od}^2) - I_{tail}$$

This simplifies to

$$I_{g12} = 2 * k' * \frac{W}{L} * (2 * V_{gs0} * V_{od} + V_{od}^2) \tag{7.308}$$

Since the current through M8 and M9 can only range from 0 to the tail current the overdrive voltage is limited to the following range

$$V_{od_{min}} = 0 \tag{7.309}$$

The minimum case means there is no current charging or discharging Cmi and the delay becomes infinite.

The maximum case means the complete tail current is flowing into the capacitor.

$$I_{g12} = I_{tail} = 2 * k' * \frac{W}{L} * V_{gs0}^2$$

Combining the equations:

$$2 * k' * \frac{W}{L} * V_{gs0}^2 = 2 * k' * \frac{W}{L} * (2 * V_{gs0} * V_{od} + V_{od}^2)$$

Solving for V_{od}yields

$$V_{od_{1/2}} = V_{gs0} * (-1 \pm \sqrt{2})$$

Since we defined out overdrive as positive the valid solution is

$$V_{od_{max}} = V_{gs0} * (\sqrt{2} - 1) \tag{7.310}$$

Higher overdrive voltages won't make the comparator any faster because the complete tail current already flows through one of the two transistors. The delay time for overdrive voltages below $V_{od_{max}}$will lead to a delay time of

$$t_d(V_{od}) = \frac{vdd * C_{mi} * L}{4 * k' * W * (2 * V_{gs0} * V_{od} + V_{od}^2)} \tag{7.311}$$

This looks like a complex equation while the accuracy of the result is limited by the poor knowledge of the average capacity Cm1. The big benefit of this equation is that now we can clearly see which design parameter has which

influence on the delay of the comparator. The following plot displays the comparator delay versus the overdrive voltage for tail currents $1\mu A$and $10\mu A$.

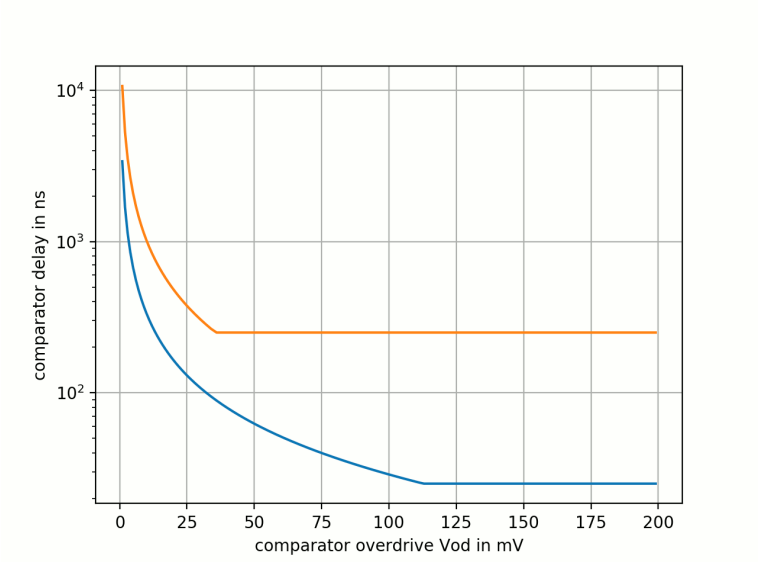


Figure 7.169: Comparator delay versus overdrive voltage for tail currents $1\mu A$(orange) and $10\mu A$ (blue)

Each of the curves has two ranges. Below $V_{od_{max}}$the delay is dominated by the transconductance of the differential stage and the overdrive voltage. Above $V_{od_{max}}$the delay is determined by the tail current alone. On the left side of the knee separating the two ranges the differential stage operates in a more or less linear way. On the right side of the knee the differential stage can be regarded as a switch that simply switches the tail current to the miller capacity or to the current mirror diode. The knee itself moves to the right increasing the tail current (as long as we leave the aspect ratio of the differential stage untouched).

For high performance comparators (for instance used in ADCs) we typically want to have the knee as far to the left as possible and the tail current as high as we can afford. This means the aspect ratio W/L must be scaled with the tail current according to the application. This works as long as the differential pair operates in strong inversion. Moving the knee down to less than about 20mV isn't possible because the input stage approaches weak inversion then. Since we can only push the knee to the left in a limited way the design of ADCs with an LSB of less than about 25mV (8 Bit at 5V supply) either has to be paid by a significant reduction of speed or we have to consider more complex multi stage comparators with a limited output swing of the first stage. (So the first stage can operate with a low overdrive similar to a linear RF amplifier while the 2nd stage operates with an overdrive in the 50mV range again to switch the output transistor.)

The calculations reducing everything to the miller capacity Cmi and the current charging or discharging the gate of M12 is a rough simplification because the miller capacity is voltage dependent. Nevertheless we can see which design parameter propagates into the circuit performance in which way.

Since the overdrive voltage is limited to less than V_{gs0} the dominating part of the denominator is the factor $2 * V_{gs0} * V_{od}$.The propagation delay approximately follows

$$t_d(V_{od}) \sim \frac{1}{V_{od}}$$

The following relations show which circuit change has the highest impact on the speed of the comparator:

$$t_d \sim vdd$$

$$t_d \sim C_{mi}$$

for low overdrive voltages (left side of the knee):

$$t_d(V_{od}) \sim \frac{L}{W}$$

If the overdrive is high (right side of the knee) the aspect ratio of the differential stage doesn't matter anymore:

$$t_{d_{min}} \sim \frac{1}{I_{tail}}$$

Knowing these propagations we can see much better which optimization has the best effect than simply designing by try and error and simulating with SPICE or SPECTRE.

Comparator with inherent hysteresis provided by the current mirrors: This is a very efficient solution for building a Schmitt trigger. Additionally the inherent hysteresis can be used to eliminate capacitive feed through coming from transmission gates. There are many ways to build an inherent hysteresis. The most common way is to add a loop gain higher than 1 to the current mirrors.

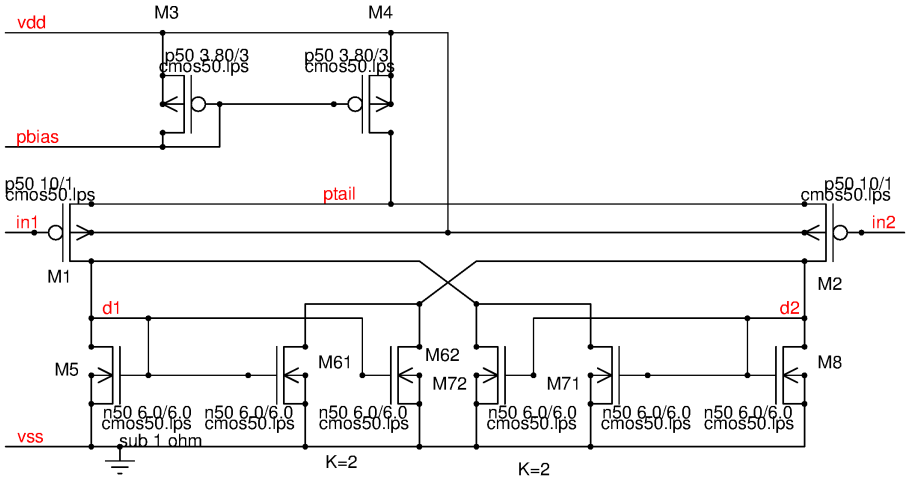


Figure 7.170: Input stage of a comparator with inherent hysteresis

Exact calculation of the hysteresis is somewhat cumbersome. The equations for the currents of the transistors are:

$$I_{M1} = I_{M5} + I_{M7x} \tag{7.312}$$

and

$$I_{M2} = I_{M8} + I_{M6x} \tag{7.313}$$

Plugging in the factor K (ratio of M5 and M6x and M8 and M7x yields:

$$I_{M1} = I_{M5} + K * I_{M8} \tag{7.314}$$

and

$$I_{M2} = I_{M8} + K * I_{M5} \tag{7.315}$$

The important question is: What is the condition needed to flip the current mirrors. Let us assume that the current of M2 flows into diode M8. So M7x can sink more K times more current than M2 provides. M1 only can flip the mirrors when it exceeds the current flowing through M2 by a factor K. The trip points becomes

$$I_{M1trip} = K * I_{M2} \tag{7.316}$$

The reverse way round works in the same way:

$$I_{M2trip} = K * I_{M1} \tag{7.317}$$

For M1 and M2 operating in weak inversion the hysteresis of the comparator becomes:

$$V_{hyst} = 2 * \frac{k * T * n}{e} * ln(K) \tag{7.318}$$

(n is the gate coupling factor as described in chapter 4

$$n = 1 + \frac{C_{bulk}}{C_g} = 1 + \frac{\varepsilon_{si} * t_{ox}}{\varepsilon_{sio_2} * \sqrt{\frac{2 * \varepsilon_{si} * (\phi - V_b)}{q * N_b}}} \tag{7.319}$$

discussing the MOS transistor equations)

Weak inversion leads to a low hysteresis but this hysteresis unavoidably has a positive temperature coefficient!

To calculate the hysteresis assuming M1 and M2 operate in strong inversion we need to know the bias current (flowing through M4). So at the trip point one transistor operates at

$$I_{M1} = \frac{I_{M4}}{(1 + K)} \tag{7.320}$$

while the second transistor operates at

$$I_{M2} = \frac{K * I_{M4}}{(1 + K)} \tag{7.321}$$

$$V_{hyst} = 2 * \Delta V_{gs} = 2 * (\sqrt{\frac{K * I_{M4} * l}{(1+K) * w * k'}} - \sqrt{\frac{I_{M4} * l}{(1+K) * w * k'}}) \tag{7.322}$$

$$V_{hyst} = 2 * (\sqrt{K} - 1) * \sqrt{\frac{I_{M4} * l}{(1+K) * w * k'}} \tag{7.323}$$

In this equation l is the length, w the width of the transistors M1 and M2. k' is the transconductance of this specific transistor type.

From the equation we can see that the hysteresis in strong inversion changes with the bias current and k' (which is a function of the gate oxide thickness). Thus the inherent hysteresis provided by a current mirror with positive feedback (K>1) is not a good solution for for high precision applications. To a certain extent the hysteresis can be tweaked to be more constant giving the bias current I_{M4} a negative temperature coefficient (because k' usually decreases with temperature due to the decrease of the mobility of the carriers at increasing temperature).

Note that for very high numbers of K the ratio $\frac{\sqrt{K}-1}{\sqrt{1+K}}$ saturates and the hysteresis can never exceed

$$V_{hystmax} < 2 * \sqrt{\frac{I_{M4} * l}{w * k'}} \tag{7.324}$$

(This is twice the gate overdrive the transistors of the differential amplifier need to take over the complete bias current.)

Bandgap comparators: Up to now we relied on a reference voltage being present to provide the trip point of the comparators. This is not always the case. Reference voltage generators require a bias current. In case of stand by modes with a wake up pin it is desirable to turn off the reference generator to reduce current consumption. The comparator itself must have an inherent reference generator that only is activated in the state of the circuit that is allowed to consume current. The solution is to build an open loop bandgap. In stead of having a feedback we operate the bandgap as an amplifier. The circuit can be built using MOS transistors operating in weak inversion or using bipolar transistors. Since the historically first implementations used bipolar transistors here the bipolar version is shown.

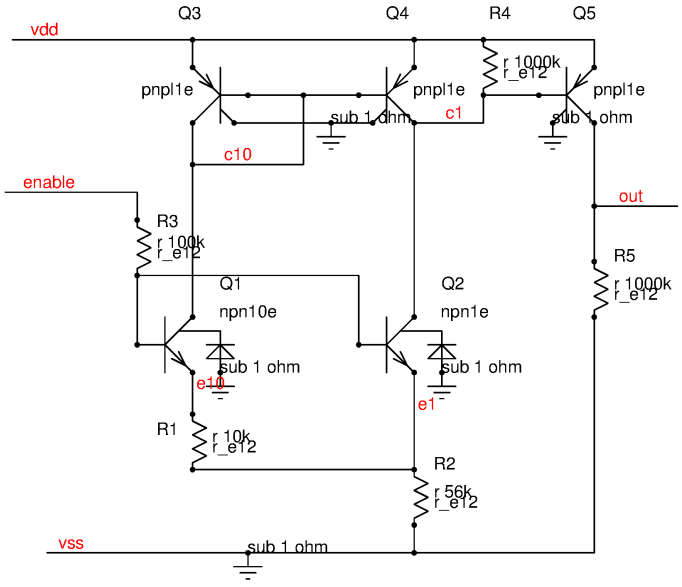


Figure 7.171: Comparator without reference input using a bandgap topology

As long as the input voltage applied at pin enable is below Vbe there is no current consumption. Reaching Vbe Q1 conducts more current than Q2 until the bandgap voltage is reached. Crossing about 1.2V Q2 will turn on Q5 and the output goes HIGH. In this state there is current flow in all branches. Resistor R4 is not needed for the bipolar implementation. In case Q3 to Q5 are replaced by MOS transistors R4 is required to prevent floating nets when the input voltage is below Vbe. R3 is needed to prevent the collector of Q2 from going HIGH due to the current flow through the base-collector diode of Q2. (The parasitic substrate PNP of Q2 sinks most of the current flowing through R3.)

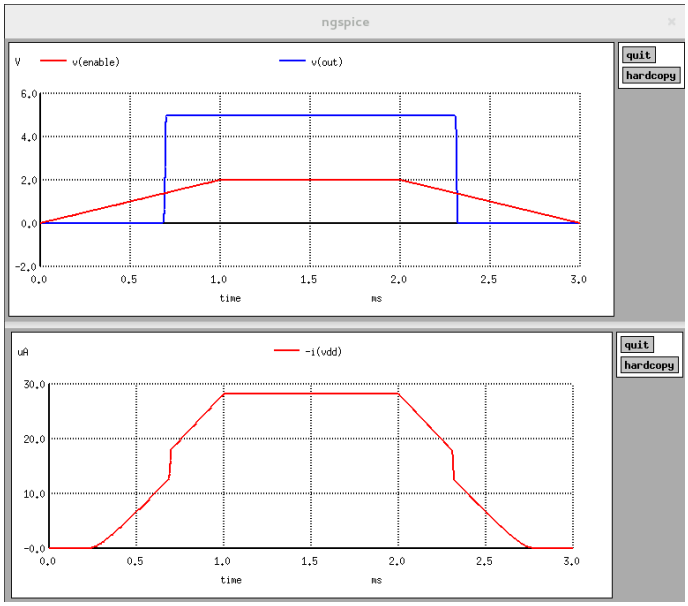


Figure 7.172: Simulation of the response of the bandgap comparator to a ramp at the input

The simulation shows the switching of the comparator and the current consumption.
There are several possible modifications of the circuit:

1. Stacking additional diodes and resistors trip points at integer multiples of the bandgap voltage can be realized. 2*Vbg leading to about 2.46V is a common application.
2. switching the resistors with the output signal can be used to provide a hysteresis.
3. In stead of using a Brokaw bandgap using a Widlar bandgap is a option when the input is driven low resistively.
4. Many CMOS technologies only offer PNP transistors. In this case the input signal must be connected to the emitters of the PNP transistors via a resistor network.

Minimum component design Schmitt Trigger: For many applications an accurate threshold is not required. The only thing needed is an amplifier with hysteresis that has a logic 1 or a logic zero at the output, but no analog values in between. This kind of circuit usually serves as an interface between an analog signal and a digital input. Depending on the technology used the circuit usually is designed using bipolar transistors and resistors or CMOS transistors without the need of resistors.

Bipolar Schmitt Trigger: The design used the earliest to interface analog signals and digital inputs. It only requires two transistors. This made the circuit very attractive even before integrated circuits became available at low cost. Applications of the circuit are described for instance at [33]. [34] already describes the same Schmitt trigger using the integrated circuit TAA151.

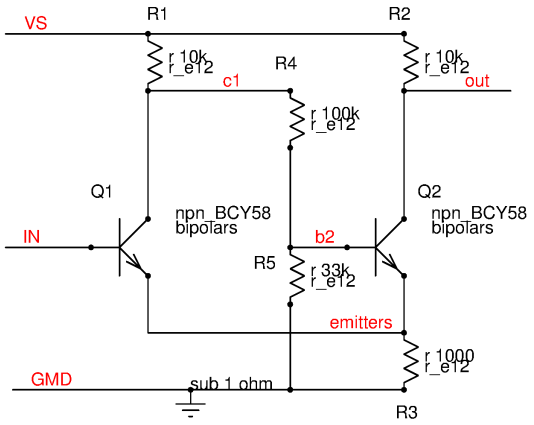


Figure 7.173: Most simple bipolar Schmitt Trigger

The circuit has two possible states:

1. Q2 is on. Neglecting the V_{CEsat} of Q2 the voltage at node "emitters" is about

$$V(emitters) \approx VS * \frac{R3}{R2+R3} * (1+\frac{1}{B}) \tag{7.325}$$

To turn off Q2 the base voltage of Q1 must reach

$$V_{tripp} \approx Vbe + VS * \frac{R3}{R2+R3} * (1+\frac{1}{B}) \tag{7.326}$$

Q2 is off, Q1 is conducting. Q2 will turn on when node "c1" reaches

$$V(c1) \approx (Vbe + V(emitters)) * \frac{R4+R5}{R5} \tag{7.327}$$

Since Q1 is on at the trip point the current flowing out of the emitter of Q1 is about

$$Ie1 = (1+\frac{1}{B}) * \frac{VS - Vc1}{R1} \tag{7.328}$$

$$V(emitters) \approx R3 * (1+\frac{1}{B}) * \frac{VS - Vc1}{R1} \tag{7.329}$$

Including the equation of Vc1 and adding Vbe we get the negative trip point voltage of

$$V_{tripn} \approx Vbe + \frac{R3 * (1+\frac{1}{B}) * (VS - Vbe * \frac{R4+R5}{R5})}{R1 - R3 * (1+\frac{1}{B}) * \frac{R4+R5}{R5}} \tag{7.330}$$

Well, doesn't look to intuitive. Even worse we are assuming a current independent Vbe, which of course is not true. Furthermore we assumed Vce to be 0V for the transistors that are on. Not quite correct either! This is why these equations are approximations. So let us calculate out example at a supply voltage of 5V: R1=R2=10K. R3=1K, R4=100K, R5=33K, Vbe=0.65V, B=300, VS=5V. The theoretical calculation yields:

$$V_{tripp} \approx 0.65V + 5V * \frac{1}{11} * 1.003 = 1.1059V$$

$$V_{tripn} \approx 0.65V + \frac{1K * 1.003 * (5V - 0.65V * 4)}{10k - 1k * 1.003 * 4} = 1.0454V$$

This solution is not quite correct. There is a nonlinear transition when Q1 and Q2 both conduct part of the current. A simulation takes care of this non linear behavior by an iteration. So here comes the simulation result:

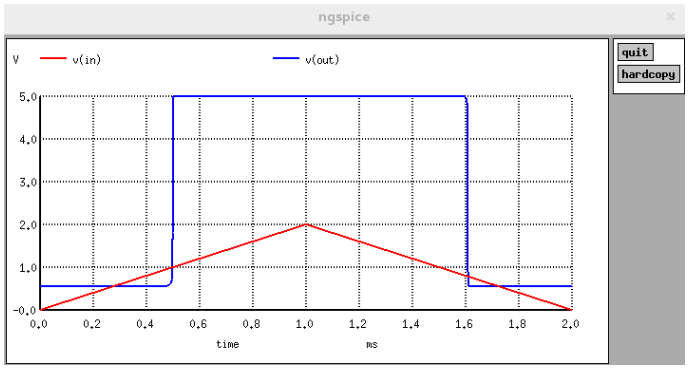


Figure 7.174: Simulation of the 2 transistor Schmitt trigger

The simulation shows lower trip points of 990mV and 780mV with a slightly higher hysteresis of 210mV. Part of the deviation is a result of the assumption that Vbe is 650mV in the calculations. But since we are operating the transistors at very low currents (The BCY58 is designed for 100mA but we work at 0.5mA) the lower Vbe is not too surprising. The calculated hysteresis is 70mV compared to 210mV in simulation. Part of the increase of the hysteresis is caused by the emitter impedances of the transistors ($V_t/I \approx 100\Omega$) reducing the gain of the stages. It is strongly suggested for such non linear circuits to first try to prove analytically that there is a hysteresis and then simulate its the accurate value.

CMOS Schmitt Trigger: The CMOS low cost Schmitt trigger can be regarded as a complementary differential amplifier with positive feed back. One of the first descriptions can e found in [35]. In the following figure it intentionally is drawn a bit different than in most books to make the differential stages better recognizable.

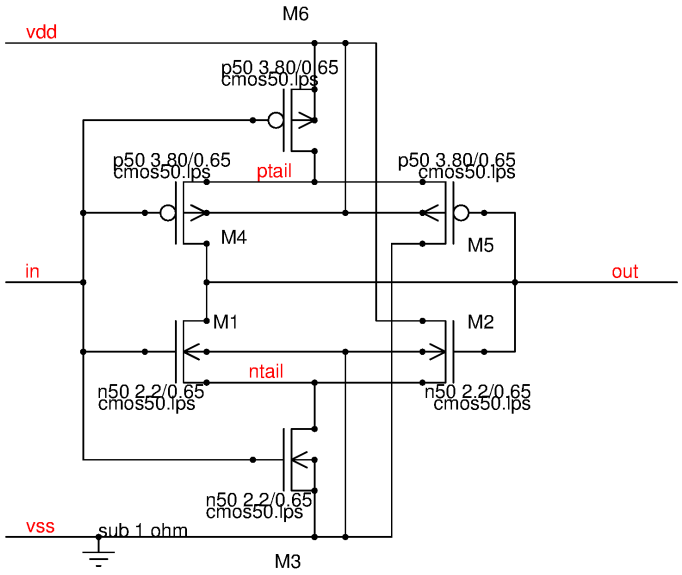


Figure 7.175: CMOS schmitt trigger

Different from classical analog amplifiers the bias current generators M3 and M6 are not driven from a constant bias but from the input signal itself. This offers the advantage that the circuit gets currentless when the input voltage reaches vdd-Vth or falls below Vth. So M3 and M6 can more or less be regarded as resistors close to the trip points. M1, M2 and M4, M8 are the complementary differential stages.

The circuit is extremely non linear. Furthermore it is sensitive to the matching of NMOS and PMOS transistors (which of course never match because they use different bulks an have different carrier mobility!) So the only thing we definitely know is that the trip points are between Vth of the NMOS transistors and vdd-Vth of the PMOS transistors.

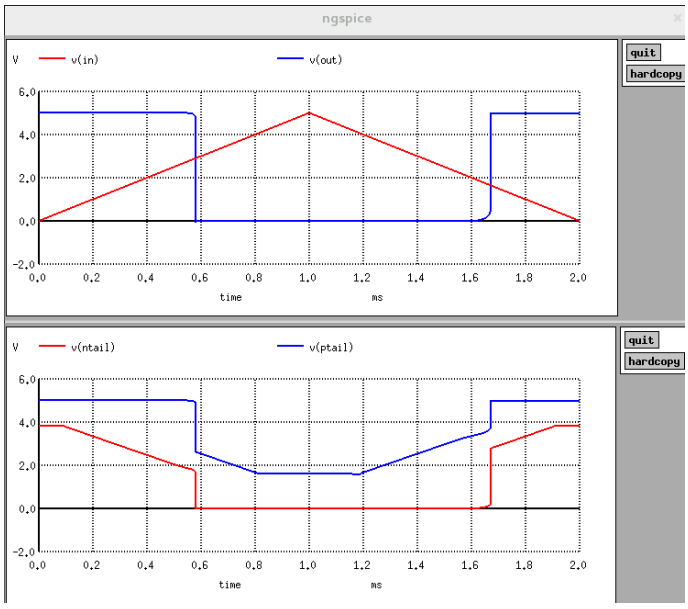


Figure 7.176: Simulation of the CMOS schmitt trigger

ESD recommendations of this circuit: Using the circuit please be aware the drains of M5 and M2 are directly connected to supplies. This is an ESD weakness in case the supplies can be attacked by ESD pulses. For charged device ESD it is strongly suggested to either use the circuit with a local supply (that is not connected to a pad) or to add resistors between the drains of M2 and M5 and the corresponding supply rails. The resistors limit ESD current flowing into the bulk diodes (that may reach zener break down) of the transistors during overvoltage at the supply rails.

7.7.11 Clocked comparators

Clocked comparators make sense for all kind of applications that are only running while the system clock is available. Typical applications are:

- ADCs
- Synchronized I/Os with voltages deviating from the standard logic 1 and logic 0 levels
- All kinds of sampling systems
- Low power applications with a very slow clock

The benefit of clocked comparators is the high speed of response when the clock is applied. (While there is no clock edge there is no response to changes at all!)

The calculation of the delay of a single gain stage continuous time comparator showed a delay of

$$t_d(V_{od}) = \frac{vdd * C_{mi} * L}{4 * k' * W * (2 * V_{gs0} * V_{od} + V_{od}^2)} \tag{7.331}$$

for low overdrive voltages. For high overdrive voltages the delay is limited by the bias current of the differential amplifier stage.

$$t_{dmin} = \frac{C_{mi} * vdd}{2 * I_{tail}} \tag{7.332}$$

Both delays have the factor vdd/2 representing the voltage swing the miller capacity has to be charged or discharged with. A clocked comparator offers the possibility to bring the circuit exactly to the trip point before releasing it to measure the input signal. Ideally the term vdd/2 disappears from the equations. In practical designs there still is some voltage swing required. Speed improvements in practical circuits are in the range of one magnitude over the speed of the corresponding continuous time comparator.

The second option is to temporarily increase the bias current I_{tail}at the sampling event.

For low power applications with long idle time it even is possible to turn off the bias current while the comparator function isn't needed. The bias current only is turned on again some μs before the comparator will be needed and will be turned off again after the measurement.

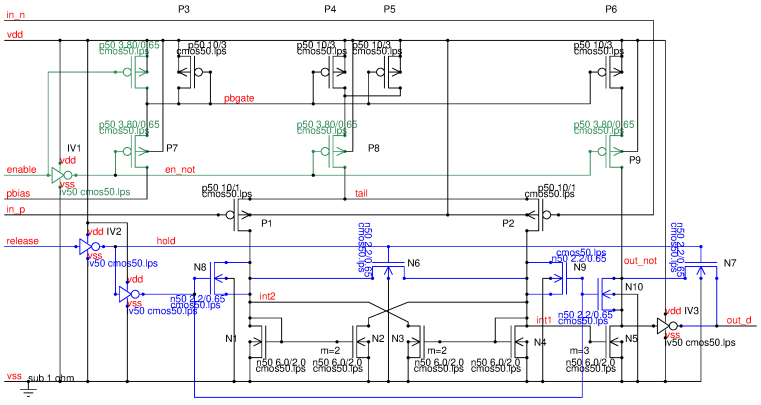


Figure 7.177: clocked comparator

The comparator shown above is the basic comparator with inherent hysteresis with the following enhancements:

- IV1, P7, P8 can be used to turn off the bias current while the comparator is not in use
- IV2, N6, N7 force the comparator into the equilibrium state before the signal is sampled.
- N8, N9, N10 compensate the charge injection of N6 and N7

At the rising edge of signal release transistor N6 turns off and the hysteresis becomes activated. The nodes int1 and int2 both are at about Vth of transistors N1 to N4. At the same time N7 turns off. The node out_not has an initial voltage that exactly corresponds the threshold of inverter IV3. It will only take fractions of ns until the node out_not will move either above the threshold of IV3 or below the threshold of IV3 after the rising edge of the signal release. N6 must be sized such that the tail current only creates a drop of a few mV over the transistor. The same applies to N7. N7 must be low resistive compared to N5 at the operating point when N6 short circuits the current mirrors N1 to N4. To test the comparator an overdrive of 5mV was used.

```
vrelease release vss pulse 5 0 5u 1n 1n 2u 20u
vref in_n vss dc 1.23
vin in_p vss dc 1.225
```

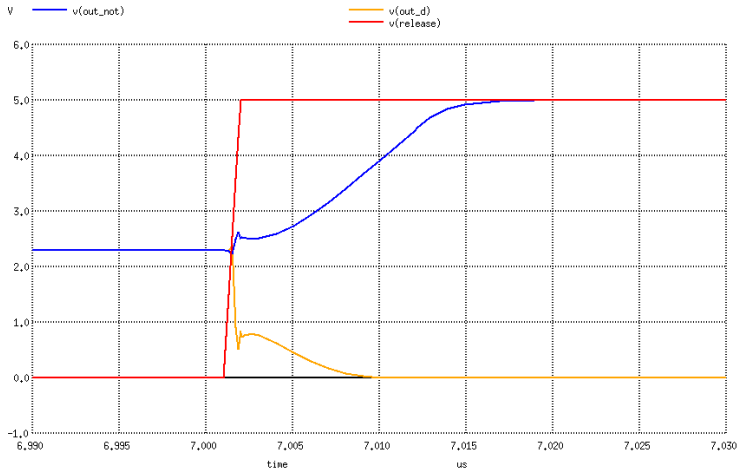


Figure 7.178: Transient simulation of the clocked comparator

The signal at node out_not shows that the comparator is already at the trip point when the release signal goes up. From the rising edge of signal release to the settling of the comparator it only takes about 10ns using a tail current of $20\mu A$. The hysteresis only is visible in the holding mode of the comparator after about 20ns. Immediately after the rising edge the hysteresis isn't active because the comparator is starting out of the balanced state.

Operating the same comparator in continuous mode the hysteresis becomes visible. The propagation delay increases by factor 4 although the overdrive was increased significantly! (see the last two lines of the stimulus shown below)

```
vdd5 vdd vss dc 5
rbias pbias vss 400k
venable enable vss DC 5
vrelease release vss DC 5
vref in_n vss dc 1.23
vin in_p vss pulse 1.15 1.31 2u 1n 1n 5u 10u
```

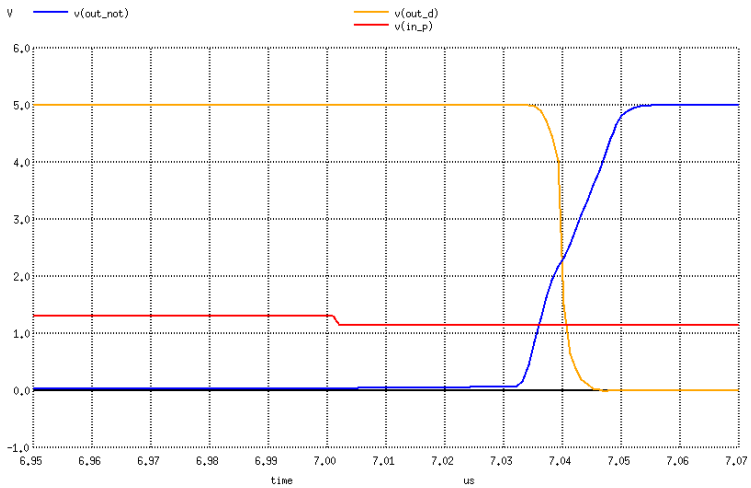


Figure 7.179: Same simulation as before but the comparator is operated in continuous mode and the input signal is increased significantly

Conclusion: Main differences between clocked comparators and continuous time comparators are:

1. The hysteresis disappears in clocked comparators.
2. The delay from the rising edge of the release signal until the output of the comparator is valid is significantly shorter than the time a continuous mode comparator needs to respond to a change of the input signal.
3. clocked comparators can only be used for systems with the clock running while measurements are taken.

These differences make clocked comparators the preferred solution for ADCs and sampling applications.

Low supply voltage: Building comparators for low supply voltages leads to similar considerations as building OPAMPs. If classical differential input stages are intended to be used the common mode range decreases reducing the supply voltage. This leads to designs using folded cascodes and rail to rail topologies. Folded cascodes and rail to rail topologies lead to complex circuits and many transistors contributing tho the offset similar to rail to rail amplifiers.

Low cost: There is one way out of the dilemma: Using a CMOS inverter as an amplifier. CMOS inverters used as linear amplifiers have a poor power supply rejection of only bout 6dB. Therefore using inverters as linear amplifiers requires a very stable supply voltage. The following circuit shows the concept.

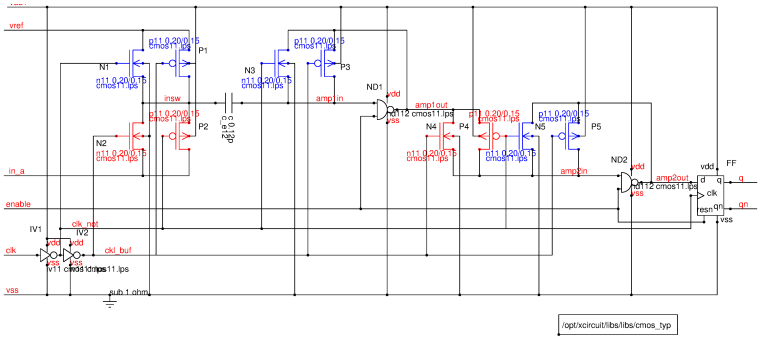


Figure 7.180: NAND gates used as amplifiers for a comparator

The amplifier has 2 operating states. In one state the reference voltage vref is measured. During the reference measurement the blue colored transistors are on. Both NAND gates are shorted between the output and the input. The voltage of the node amp1in is exactly the equilibrium voltage. The difference between this equilibrium voltage and the reference vref is stored in the 100fF capacitor. To measure the input voltage the blue switches are opened and the red switches are closed. If the input voltage at in_a is higher than the reference the node amp1in moves up and both inverters amplify the signal. The output amp2out moves changes to HIGH. If the input voltage at node in_a is lower than the reference the node amp1in moves down and output amp2out becomes LOW.

Switches N4 and P4 can also be replaced by a capacitor. This is advantageous if the operating points of the two gates ND1 and ND2 differs significantly while the switches N3, P3 and N5, P5 are closed. On the other hand the size of a capacitor is bigger than the size of the two minimum transistors N4, P4. It is a trade off between offset and chip real estate.

At the falling edge of the clock the result of the measurement is stored in the flip flop FF.

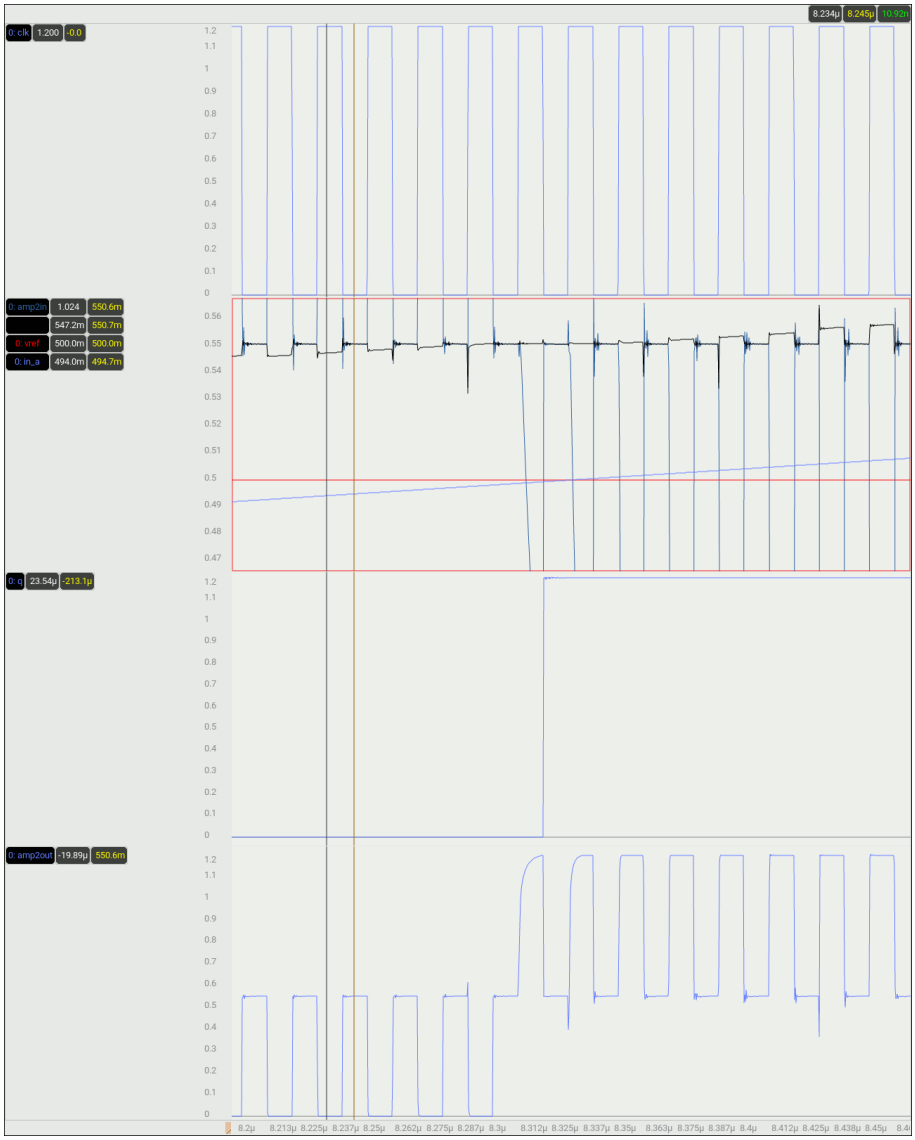


Figure 7.181: Simulation of the clocked comparator using 2 NAND gates

In the second strip the increasing voltage V(in_a) (blue) and the reference V(vref) (red) is shown. The signal is amplified by the two NAND gates while the clock signal is logic 1. The output amp2out toggles either between 0 and the equilibrium voltage (V(in_a)<V(vref)) or between 1 and the equilibrium voltage (V(in_a)>V(vref)). The flip flop samples the state of amp2out at every falling edge of the clock.

The performance of the circuit is limited by the clock feed through of the switches into the capacitor. There are several options to improve the circuit:

- Use same size NMOS and PMOS transistors to compensate the clock feed through
- Add dummy transistors of half the size of the switches connected to the opposite phase of the clock to compensate clock injection
- Increase the capacity to reduce the clock injection

Wire coupling and layout requirements All the analog traces are low level signals that MUST be protected from noisy signals. The worst aggressor in this circuit are the clock lines that have a high voltage swing (typically 5V or 3.3V) and fast edges (rise and fall times in the range of 100ps). The analog signals and the clock lines may under no circumstances be routed in close proximity. The recommended solution is to route the clock lines perpendicular to the analog lines. This minimizes the parasitic capacities.

In addition every analog line that is crossed by a clock line must also be crossed by an inverted clock line to compensate the clock injection.

If clock lines have to be routed in parallel with analog lines the clk and clk_n line must be designed like a twisted pair to achieve the same parasitic capacity between signal and clk as well as between signal and clk_n.

7.7.12 Interfacing comparators with the logic

Most comparators are used to interface analog signals coming from outside of the chip or from analog functions with the logic. It depends on the function how the comparator is connected to the logic.

Unsynchronized interface: Directly connecting a comparator output to the logic without any synchronization is very unusual. It only is done for applications that must be functional even if the system clock isn't running anymore. Typical application of this kind are under voltage lock out, wake up or system reset. Unsynchronized interfaces must be described carefully in the logic design because most logic synthesis tools automatically insert a synchronization at every interface! It is strongly recommended to check manually each interface that is unsynchronized intentionally. (I have already seen wake up inputs that didn't end STOP mode because some logic synthesis added a synchronization. So it was possible to send the CPU into STOP mode, but waking it up again didn't work because in STOP mode no more clock was running.)

Most asynchronous logic functions (for instance asynchronous reset) require a certain minimum pulse time. This is needed to guarantee that setup and hold times of flip flops aren't violated.

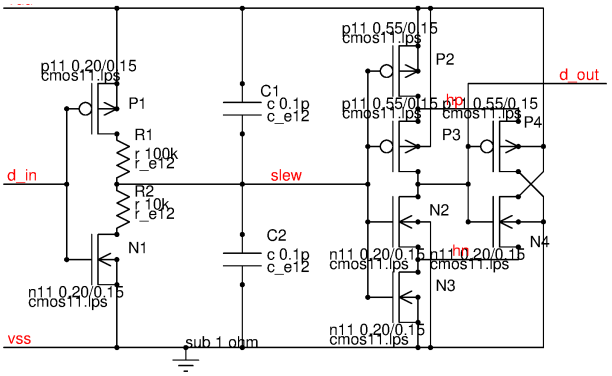


Figure 7.182: Minimum pulse generator

The minimum pulse generator provides either no pulse at all or a pulse with a certain minimum HIGH or LOW time. The resistors can of course be replaced by current sources as well. (In many cases this is cheaper).

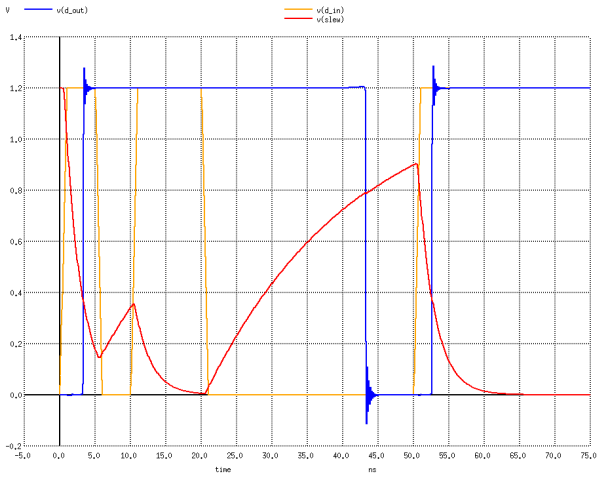


Figure 7.183: Simulation of the minimum pulse width generator

In the plot above a short pulse from 6ns to 10ns will not trigger a change of signal v(d_out) while a long drop of V(d_in) from 20ns to 50ns switches V(d_out). The hysteresis of the 6 transistor Schmitt trigger can be seen comparing V(slew) rising crossing V(d_out) and V(slew) falling crossing V(d_out). The hysteresis is about 400mV. The minimum LOW pulse width is determined by the capacitors and R2. It is about 2.6ns. For a HIGH pulse the minimum pulse width is determined by R1 and the capacitors. It is about 26ns.

Synchronized interfaces: This is the standard. A comparator output can change exactly at the clock edge. Reading such a signal leads to violations of setup and hold times of the flip flops inside the logic. Such violations can have two effects:

- The flip flop doesn't detect the signal as expected. (Best case it will lead to a detection one clock edge later. For many applications this isn't a problem - but for testing it is a severe one because the whole pattern fails depending on spread of the delay!)
- The flip flop starts to oscillate as a ring oscillator when the data input has an edge that is coincident with the clock edge. Usually these oscillations end after a few cycles. But for a counter this can already be a disaster.

The most common approach to prevent such situations is double buffering with two flip flops.

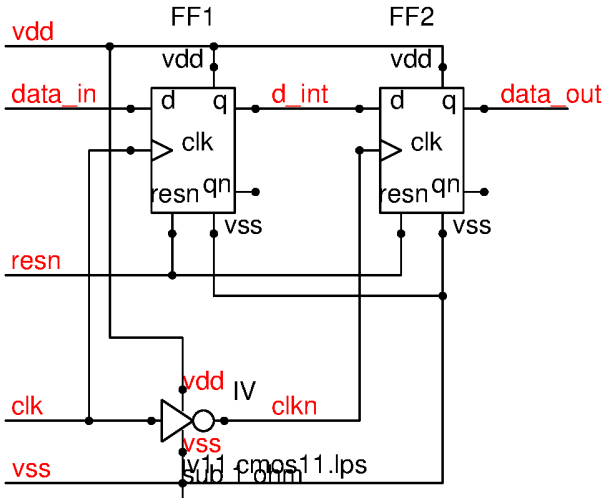


Figure 7.184: Typical clock synchronization circuit

Even if there is a setup time or hold time violation at FF1 and FF1 may either start to oscillate or have a very wide spread of delay times the second flip flop stabilizes the timing again. For a synchronous logic synchronizing with only one stage isn't good enough. The second flip flop will provide a stable output signal. Hitting such a condition in simulation is extremely difficult. For this reason a little hardware setup using a standard CMOS 4035 shift register is used to demonstrate the effect (I admit I first tried to simulate it but in simulation I didn't get it as nicely as on the hardware test bench!).

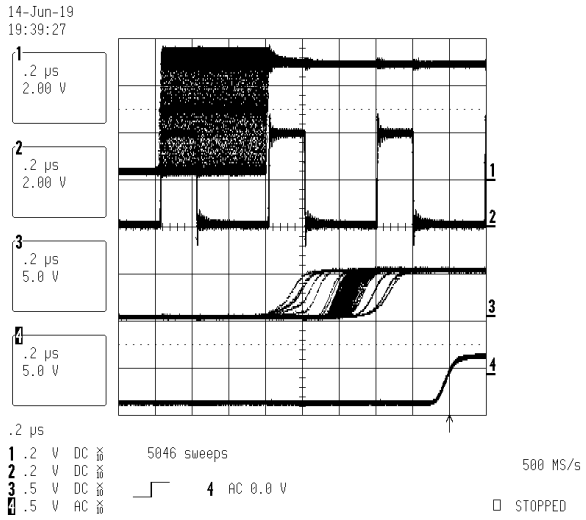


Figure 7.185: Demonstration of the behavior of a double synchronization

Trace 1 is the asynchronous input signal of the synchronizer. The clock is shown on trace 2. After the first flip flop the sampled signal has a very wide timing spread due to setup and hold time violations (trace 3) . After the second flip flop the signal is stable (trace 4).

One of the drawbacks of the synchronization is the delay added to the signal path. Using clock and inverted clock as shown in the circuit above the worst case delay is one clock period + the propagation delay of the flip flops. Sometimes the circuit is implemented using clk (instead of clkn) for the second flip flop as well. (This is the case in the example measurement using the 4035 shift register as a synchronizer.) In this case the delay time can reach 2 clock periods + the propagation delay of the flip flops.

To work properly the propagation delay of the flip flops must be less than half the clock period using the implementation shown in the circuit above. If both flip flops work on the same edge of clk the requirements for the propagation delay relax by factor 2.

The delay isn't critical for most applications. In extreme cases (for instance PWM control of a switched mode power supply or fast ADCs) even this short delay may in fact matter! In this case it may make sense to build a self contained asynchronous design and only provide the status information to the logic via a synchronizer.

7.8 Low side power output stage

Most low side power output stages are built simply driving a power transistor either with a switch to a low voltage supply and a switch to ground at the gate or with current sources pulling the gate to about 5V or ground.

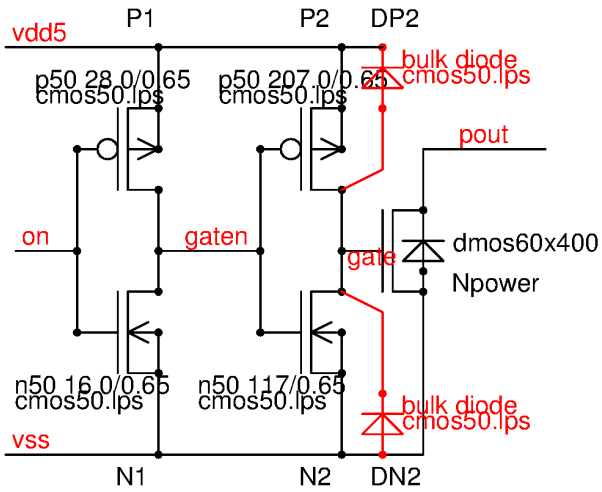


Figure 7.186: Most simple low side power driver

The driver shown above is hard switching and has no protections against overvoltage or overcurrent. Practical designs usually require protections. There is a certain priority of the protections. Overvoltage is destructive within a couple of ns. Overcurrent doesn't directly harm. The destruction is thermal. Depending on the power density destruction due to overcurrent in most cases takes some us. (May be less than a us in case of SiC and GaN transistors due to the high power density achieved there). Thermal sensors usually have a response time in the range of some ten to hundred us. So the current must be limited to limit the speed of the temperature increase in a way that the thermal protection responds before the transistor reaches about 400 deg. C. This leads to the following hierarchy of protections:

1. over voltage protection
2. over current protection
3. temperature protection

7.8.1 Over voltage protection

Over voltage protection is mainly required operating a low side driver with inductive loads. If the load is well known it may be sufficient to limit dI/dt and/or dV/dt. The most simple way to do this is to drive the power transistor gate with limited current.

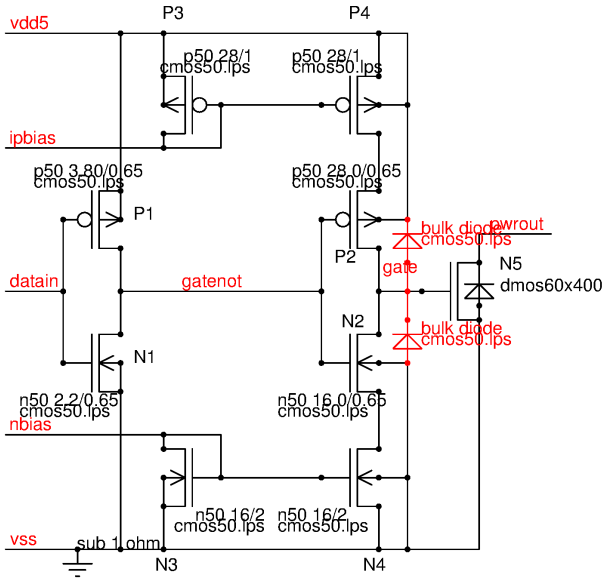


Figure 7.187: Low side power driver with dV/dt limitation by the miller capacity of the power transistor

The peak voltage at pin pwrout depends on the inductance of the load. The following 2 plots show a turn off with 1uH inductance in series with 10 Ohm and with 100uH inductance in series with 10 Ohm. (Power supply was 12V, so peak current was 1.2A)

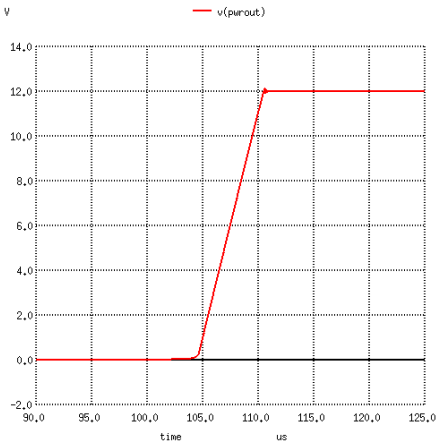


Figure 7.188: Turn off of a slew rate limited driver with a load of 10 Ohm in series with 1uH (Supply of the load is 12V)

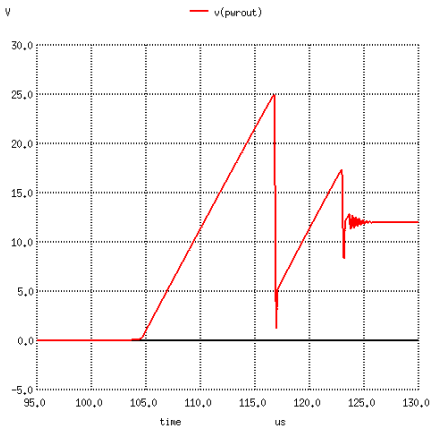


Figure 7.189: Turn off of a slew rate limited driver with a load of 10 Ohm in series with 100uH (Supply of the load is 12V)

The rising slew rate is limited by the driver. The falling edges of the ringing is very non linear depending on

capacities because the power transistor can't do more than turn off. Further increase of the load inductance will lead to even higher overshots until the power transistor gets destroyed. Therefore if the load in unknown an additional zener clamp must be added. The zener clamp requires a diode in series to prevent draining the gate voltage through the diode.

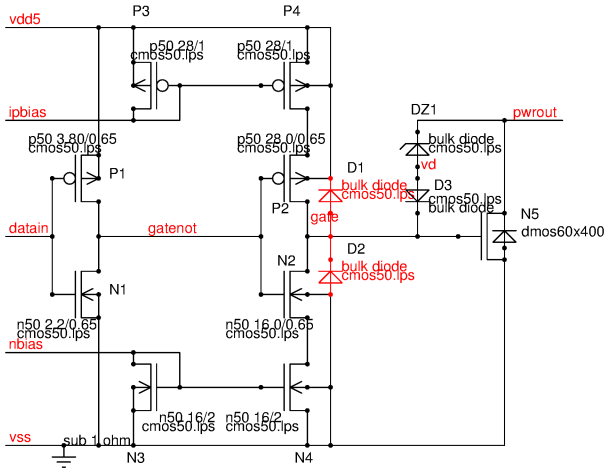


Figure 7.190: Low side power stage with overvoltage protection

In the circuit shown above the zener diode DZ1 limits the voltage at the drain of N5.

$$V_{clamp} = Vz + V_{fD3} + V_{thN5} \tag{7.333}$$

To operate correctly the zener diode must be low resistive compared to the impedance at the drain of N2 and P2. As an example the same load as before (100uH, 10 Ohm) is simulated using a 12V zener diode as a clamp.

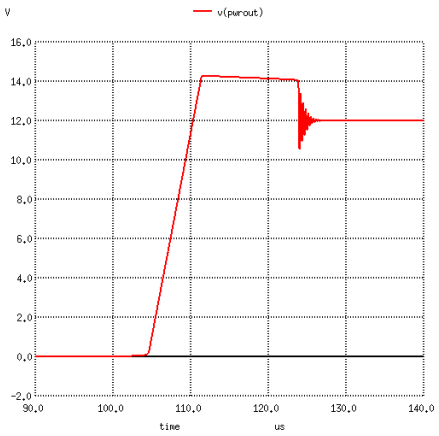


Figure 7.191: Limiting the voltage with a 12V zener diode

The energy stored in the inductor is mainly dissipated by the power transistor. Only a little fraction remains in the real part of the load impedance.

7.8.2 RF sensitivity of a low side driver with overvoltage protection

The gate node of the over voltage protected driver is high resistive (compared to the RF impedance of the miller capacity of N5). So RF applied at the drain of N5 while N5 is off will propagate to the gate via the Cdg of N5. At the gate this RF will be rectified by the bulk diode D2. This leads to an increase of the gate voltage and will turn on N5 until it becomes low resistive enough to attenuate the RF. Typically N5 will conduct some 10mA to some 100mA if RF is applied. To turn on N5 the rectified RF voltage must be higher than:

$$V_{rect} > V_{th} + V_f \tag{7.334}$$

Since the voltage at the gate is attenuated by the divide consisting of Cdg and Cgs the required peak to peak RF voltage becomes:

$$V_{pp} > \frac{C_{dg}}{C_{dg} + C_{gs}} * (V_{th} + V_f) \tag{7.335}$$

RF applied while N5 is already on will be shorted by N5.

At very high frequencies RF applied at the drain of N5 will be shorted by the Cdb of N5.

To reduce RF rectification the RF must be kept away from D2. This can be done increasing Cgs or adding a low pass to the circuit. On the gate side (right side) of the resistor no parasitic diodes are permitted. The resistor must be a poly silicon resistor. Diffused resistors won't work.

Since the power transistor acts as an RF clamp too the effect is limited. Due to the clamping of the power transistor itself the steady state signal at the gate becomes an oscillation with:

$$V_{ppg} = 2 * V_{th} \tag{7.336}$$

This oscillation swings around the DC level provided by the drier stage N2, P2. Even if the driver pulls down to 0V at the right side of the resistor we will have an signal ranging from -Vth to Vth periodically turning on N5. This still is true even if the RC filter doesn't allow RF to reach D1 and D2. Thus the improvement only is:

$$improvement = 20 * log10(\frac{2 * V_{th}}{V_f + V_{th}}) \tag{7.337}$$

Normally this is in the range of a few dB. The signal at the drain of the power transistor when the stage fails becomes:

$$V_{pp} = 2 * V_{th} * \frac{C_{gd}}{C_{gd} + C_{gs}} \tag{7.338}$$

The best solution is to increase Cgs and in a way that it becomes significantly larger than Cdg. But this is an expensive solution.

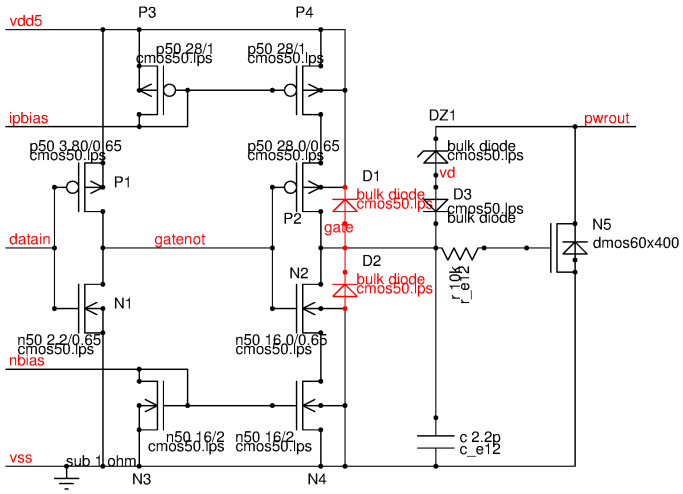


Figure 7.192: Low side driver with RF filter to keep RF away from the parasitic diodes D1 and D2. Nevertheless positive RF half waves will still turn on N5

7.8.3 over current protection

The purpose of the over current protection is to limit the power dissipation in a way that the power transistor doesn't heat up too fast for the thermal protection to react. The current itself isn't destructive (except for metal migration, but that takes weeks to years).

A very common way to protect power stages is a delta Vbe comparator. The threshold has a thermal temperature coefficient but since the metal resistors have a positive temperature coefficient this doesn't harm as long as both transistors of the comparator are on an isotherm or the temperature gradient is decreasing the threshold.

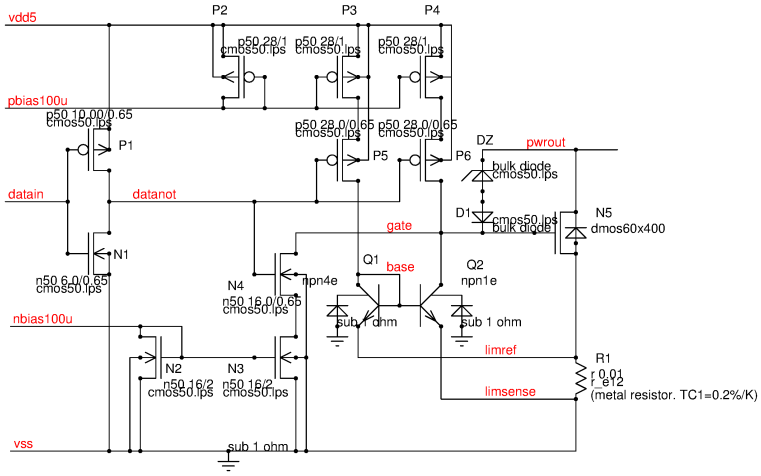


Figure 7.193: The most frequently found implementation of a current limit

The two NPN transistors Q1 and Q2 operate at the same current. Since the transistor emitter ratio is n=4 the current limit is about

$$I_{lim} = ln(n) * \frac{V_T}{R_1} \tag{7.339}$$

With n=4, V_t=26mV, R_1=10mΩ we get a current limit of 3.6A (assuming both NPN transistors operate at the same temperature). If Q2 is closer to the power transistor than Q1 the current limit will be lowered depending on the temperature gradient of the chip. Placing Q1 closer to N5 than Q2 will most likely lead to a destruction at short circuit!

If there are no bipolar transistors available in the process used the simple NPN comparator can either be replaced by an NMOS delta Vgs comparator operating in weak inversion or by a full blown operational amplifier.

Loop stability usually is determined by the gate capacity of N5 acting as the dominant pole.

7.8.4 RF sensitivity of the current limit circuit

RF injected into net pwrout will propagate to the gate of N5 via the gate capacity of N5. If the power stage is off the voltage at net gate is close to 0V. As a consequence Q2 is operating in saturation. The diffusion capacity from the base of Q2 to the collector can easily be in the range of some pF! So RF injected will reach the base of Q2. If the RF level is high enough Q2 will turn on and override the current provided by P4, P6. So different from the unprotected stage that turns on the current limited stage will be disabled by RF.

The situation gets even worse if a boosted current mirror is used. In this case the boost transistor Q3 acts as an extremely aggressive RF rectifier. Making thing even worse the only pull down at the base of Q1 and Q2 is the base current.

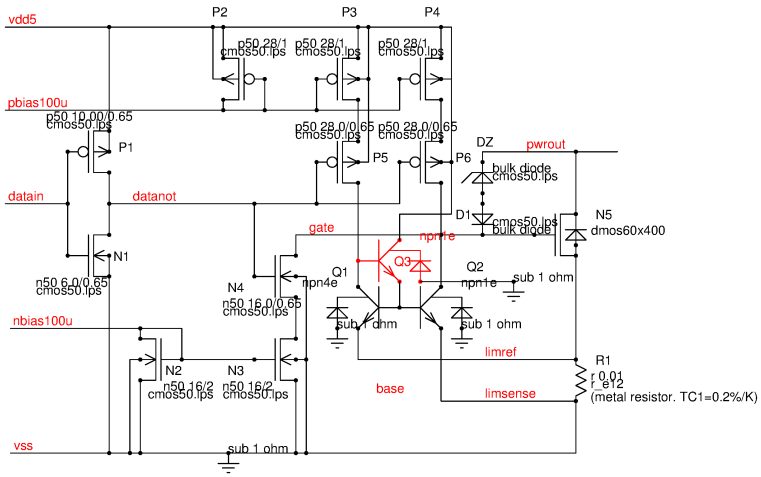


Figure 7.194: The worst case implementation of the current sense regarding RF injection

There is only one comment about Q3: DON'T DO IT THAT WAY! The gain of accuracy is insignificant but the EMI trouble keeps you busy for a long time.

To improve the circuit for better RF injection ruggedness the following modifications are suggested:

1. Add an RF filter between the gate of the power transistor and the current limiting NPN transistor

2. Remove the booster that acts as a perfect rectifier.

This leads to the following circuit:

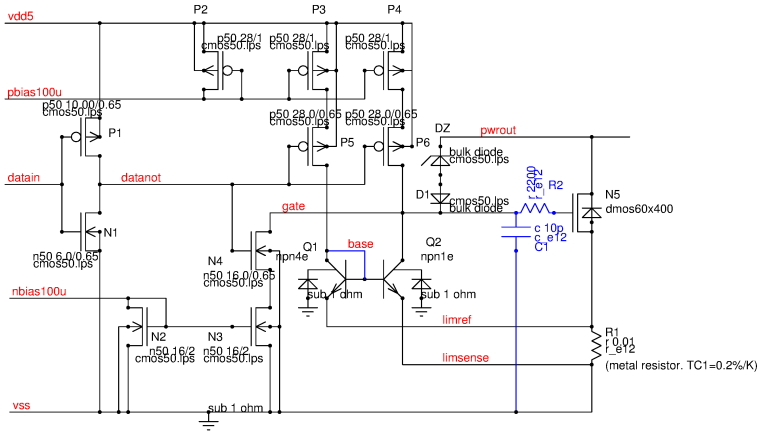


Figure 7.195: Low side driver with current limit and RF filter

This RF filter solves the problems caused in the current limiting circuit. It however doesn't solve the weak turn on of the power stage at every positive half wave. But at least we come back to the old limit of:

$$V_{pp} = 2 * V_{th} * \frac{C_{gd}}{C_{gd} + C_{gs}} \tag{7.340}$$

At this RF level the power transistor starts to softly turn on.

7.9 High side power output stages

Most high side switches use NMOS transistors and charge pumps to drive them. The reason for using NMOS transistors is the higher carrier mobility. PMOS power output stages for only are used if no charge pump can be accepted (charge pump noise) or if the currents required are fairly low. The following figure shows a typical high side driver. Protection circuits for short circuit and overvoltage are omitted for simplicity.

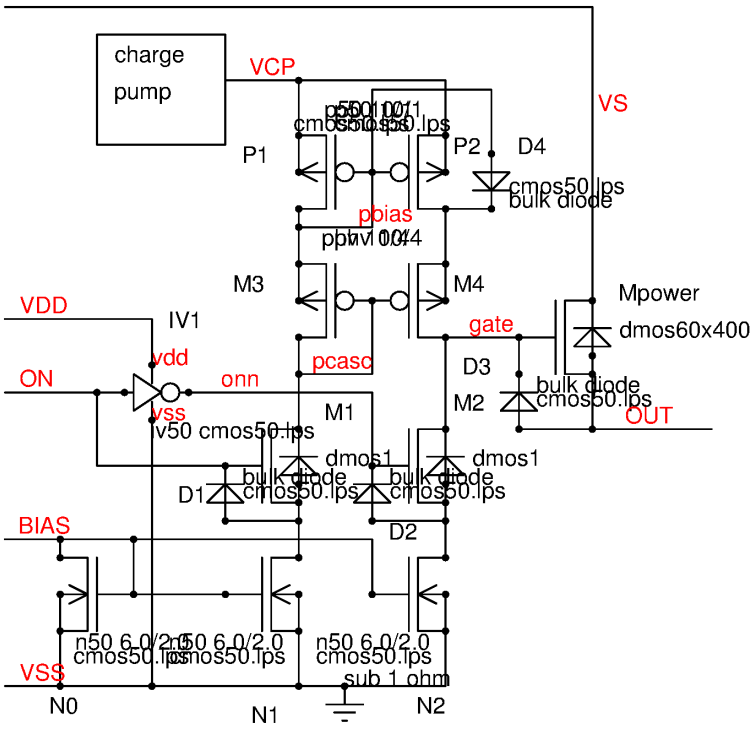


Figure 7.196: The most simple high side switch

Diode D3 protects the gate of Mpower in case a capacitive load is driven. If the gate of Mpower is pulled down the diode clamps the gate voltage before a destruction of the gate oxide takes place.

Diodes D1, D2 and D4 are leakage protections. They clamp the drain voltage of the low voltage transistor in case the high voltage cascodes are leaky.

The switching speed of the power stage is limited by the miller capacity of Mpower and the bias current flowing in N2, M2 and P2, M4. Since the miller capacity changes with Vgd the slope usually is not linear. Additional to the slope time the stage has a delay turning on and turning off caused by the gate charge time from 0V to the threshold voltage (turning on) and from the chargepump voltage to the threshold voltage (turning off). This delay mainly depend on Cgs of Mpower and the bias current. To achieve symmetrical delays the charge pump voltage should roughly be double the threshold voltage of Mpower. If switching delay symmetry is not of interest the charge pump voltage usually is chosen as high as possible to achieve a low Rdson. in the following simulation the stage is optimized for Rdson rather than symmetrical delay time.

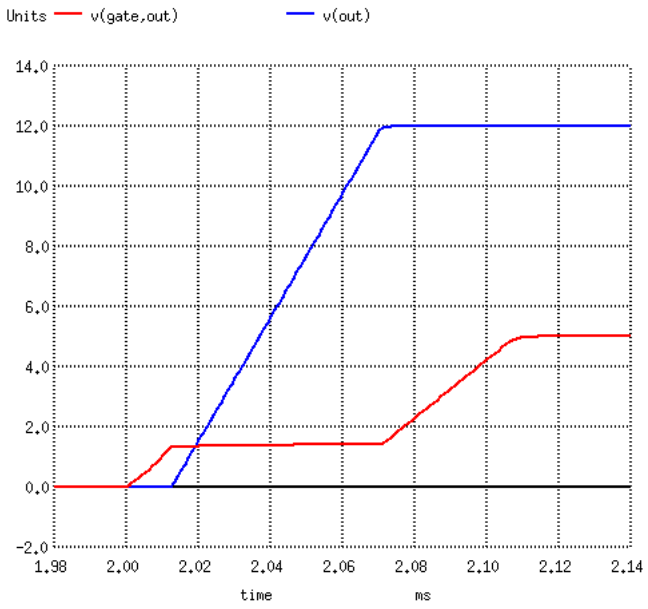


Figure 7.197: Vgs and Vout of the simple driver shown above

The red curve illustrates the charging of the gate until the threshold is reached. At the threshold Vgs stays almost constant while the charge current mainly flows into the miller capacity. This is called the gate voltage plateau. At the end of the slope the gate finally gets charged to the full charge pump voltage.

Driving a resistive load the maximum power dissipation is reached when the output voltage of the driver is 50% of the supply voltage.

$$P_{maxR} = V_s^2/4R_{load} \tag{7.341}$$

Normally this peak dissipation only is present for a very short time (compared to the thermal time constant of the power transistor). Integrating the product of current and voltage over the rise time we can calculate the switching loss. Assuming a linear edge simplifies the calculation.

$$V_{out}(t) = V_s * \frac{t}{t_r} \tag{7.342}$$

$$P(t) = (V_s - V_{out}(t)) * \frac{V_{out}(t)}{R_{load}} \tag{7.343}$$

$$E_{switch} = \int_0^{tr} P(t)dt \tag{7.344}$$

$$E_{switch} = \frac{V_s^2 * t_r}{6 * R_{load}} \tag{7.345}$$

The average power dissipation of the power transistor is more interesting. To calculate the average dissipation the duty cycle D, the Rdson, the load current and the switching losses are required. For slow switching applications and static on (D=1) the losses in the Rdson are dominant and the switching losses are neglected in the following equation.

$$P_{on} = D * V_s^2 * \frac{R_{dson}}{(R_{dson} + R_{load})^2} \tag{7.346}$$

For higher frequencies the switching losses can no longer be neglected. Since in every period we have a rising edge and a falling edge both losses have to be added.

$$P_{switch} = \frac{V_s^2 * f * (t_r + t_f)}{6 * R_{load}} \tag{7.347}$$

The total average power dissipation becomes

$$P_{av} = P_{on} + P_{switch} = V_s^2 * (\frac{D * R_{dson}}{(R_{load} + R_{dson})^2} + \frac{f * (t_r + t_f)}{6 * R_{load}}) \tag{7.348}$$

with $D = f * t_{on}$.

Note that the rise and the fall time usually are a function of the supply voltage Vs! Te supply dependence of the slope losses can be reduced making the bias current a function of Vs.

Example: A resistive heating is driven using a 5 milliohm transistor. The heater has a resistance of 0.2 Ohm. Supply voltage is 12V. tr=tf=10us, f=100Hz

Power dissipation at D=100% (permanently on):

$$P_{100\%} = 144V^2 * 0.005/0.205 = 3.512W$$

Power dissipation operating at 50% duty cycle:

$$P_{50\%} = 144V^2 * (0.5 * 0.005/0.205 + \frac{100Hz * 20\mu s}{6 * 0.2\Omega}) = 1.7561W + 0.24W = 1.9961W$$

7.9.1 High side driver operating with partial capacitive load

If he driver is operated with a capacitive load in parallel to the resistive load the current flowing through the witch during the edges is no more exactly proportional to the output voltage.

$$I(t) = \frac{V_{out}(t)}{R_{load}} + C_{load} * \frac{dV_{out}(t)}{dt} \tag{7.349}$$

The additional dissipation needed to charge the capacitor (area of a triangle) is:

$$E_{charge} = \frac{1}{2} * C * V_s^2 \tag{7.350}$$

This energy has to be dissipated by the switch once every period.

After turning on the capacitor stores an energy of:

$$E_{cap} = \frac{1}{2} * C * V_s^2 \tag{7.351}$$

The energy stored by the capacitor will flow into the load when the high side switch turns off again.

The total energy taken from the supply voltage to charge the capacitor is the sum of the stored energy and the energy dissipated by the switch.

$$E_{Csupply} = C * V_s^2 \tag{7.352}$$

The total dissipation of the high side switch driving a resistor and a capacitive load becomes

$$P_{av} = P_{on} + P_{switch} + P_{charge} = V_s^2 * (\frac{D * R_{dson}}{(R_{load} + R_{dson})^2} + \frac{f * (t_r + t_f)}{6 * R_{load}} + \frac{f * C}{2}) \tag{7.353}$$

7.9.2 High side driver operating with an inductive load

The losses of the power driver operating with an inductive load strongly depend on the operating conditions. Applications that are not designed for PWM operation typically work without a free wheeling diode. At turn on the current through the inductive load is 0A. So turn on losses usually can be neglected in discontinuous mode. After turn on the current increases until the resistive part of the load limits the current. At turn off the inductive part of the load keeps the current flowing and pulls the output of the driver below ground. Since the gate voltage is pulled down to 0V the source of the power stage will clamp the negative swing to -Vth if the power transistor.

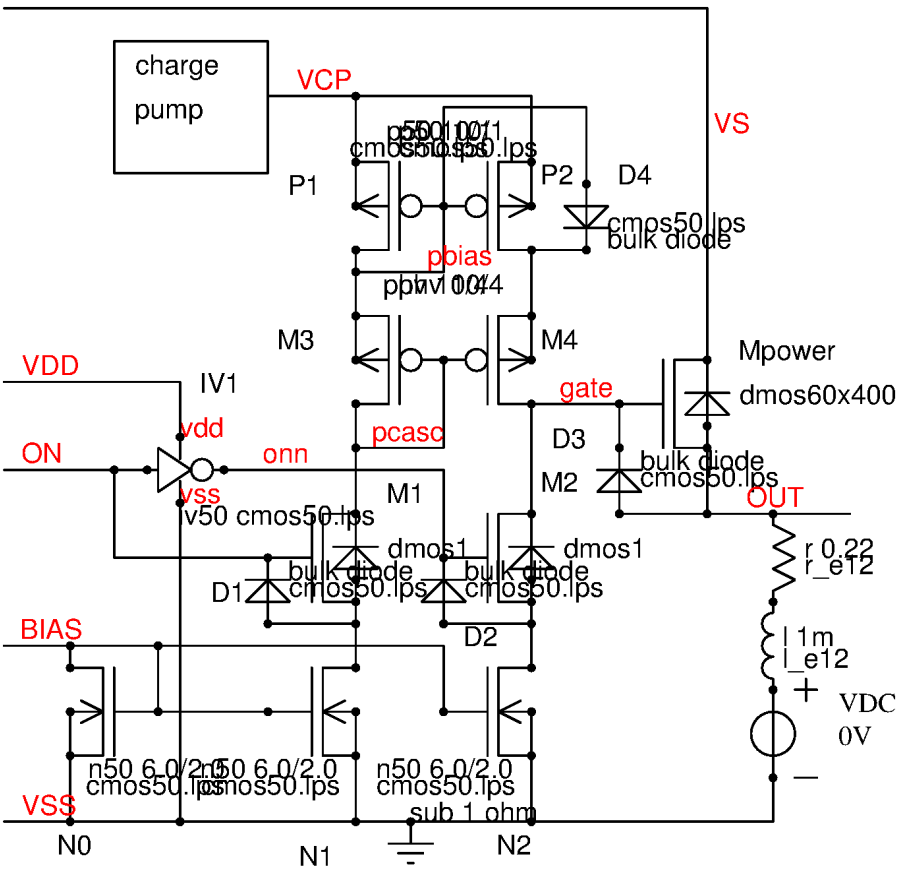


Figure 7.198: High side driver with partially inductive load

The voltage source VDC is only needed to measure the current in the simulation. The following figure shows the voltage at net OUT and the current flowing through VDC.

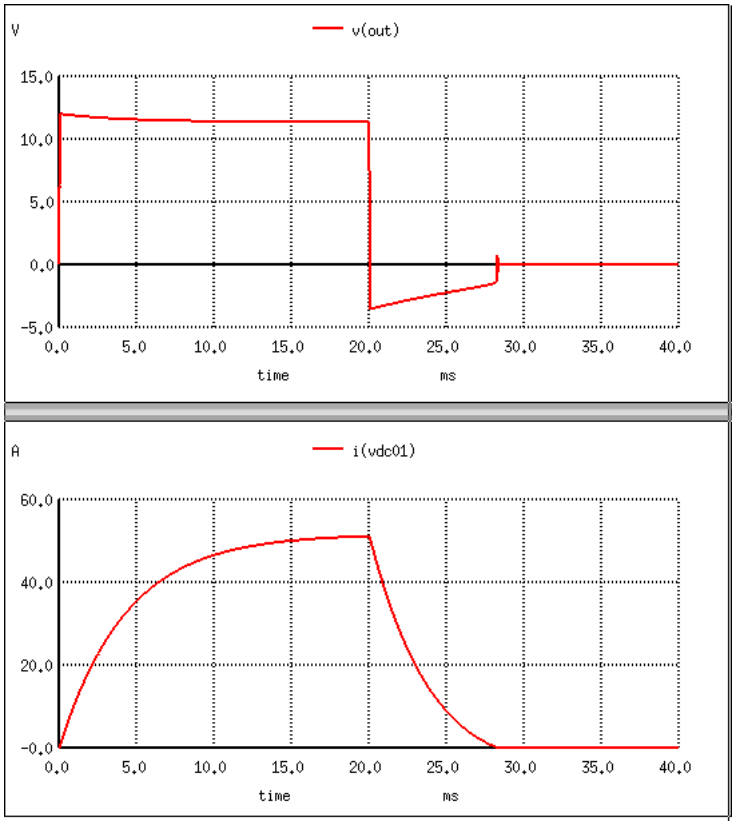


Figure 7.199: Driver output voltage and current flowing in the load

During the falling edge of the current the power transistor has a Vds of about 16V. The peak power dissipation

is 50A*16V=800W! Even a power transistor with 4mm^2 is already close to immediate destruction exposed to such high power levels. The following plot shows the power dissipation (bottom trace, red) and the simulated junction temperature (upper trace, blue).

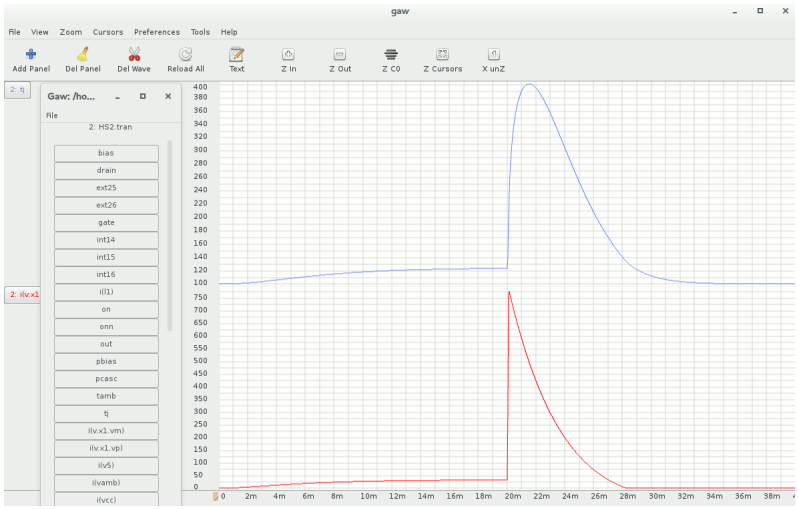


Figure 7.200: Power dissipation in W and junction temperature in deg. C assuming a transistor area of 4mm^{2}of the power transistor

Even if the metal traces do not melt yet (AlSi alloy melts at about 420 deg. C) temperature swings of more than 100K within less than a ms are critical because thermal expansion of the hot silicon will (e.g. if it happens repetitively) break the oxide isolation (that has a different thermal expansion coefficient).

To reduce the thermal impact on the power transistor a free wheeling element should be added. The voltage drop over the free wheeling element is much lower (about 1V to 1.5V if a simple diode is used) than the drop over the power transistor. Adding the free wheeling diode the size of the power transistor may be reduced (unless there is a second limitation in static on state) much more than the free wheeling diode will cost.

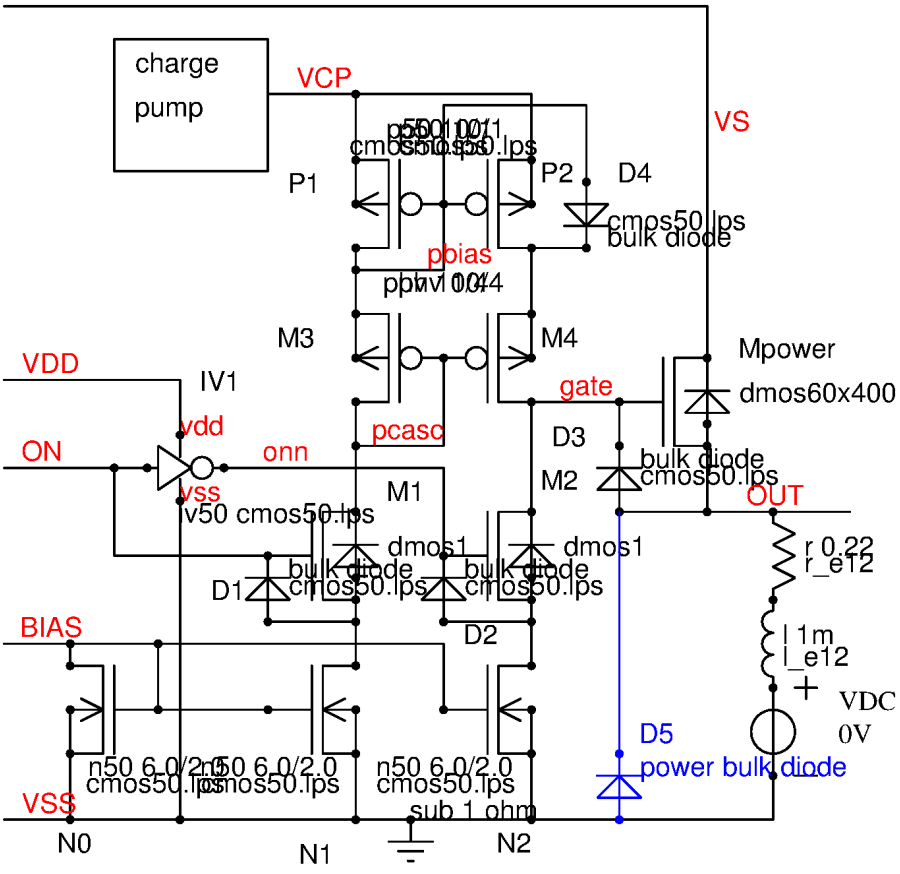


Figure 7.201: High side driver including the free wheeling diode for inductive loads

With the free wheeling diode the output is clamped to about -1.2V. This leads to a slower decay of the current

but the power dissipation inside the diode is one magnitude less than the power dissipation inside the power transistor before. Assuming a diode area of 0.5mm^{2}the peak temperature reached is 270 deg. C (initial temperature 100 deg. C). Now Mpower can be optimized without needing the thermal capacity for the turn off.

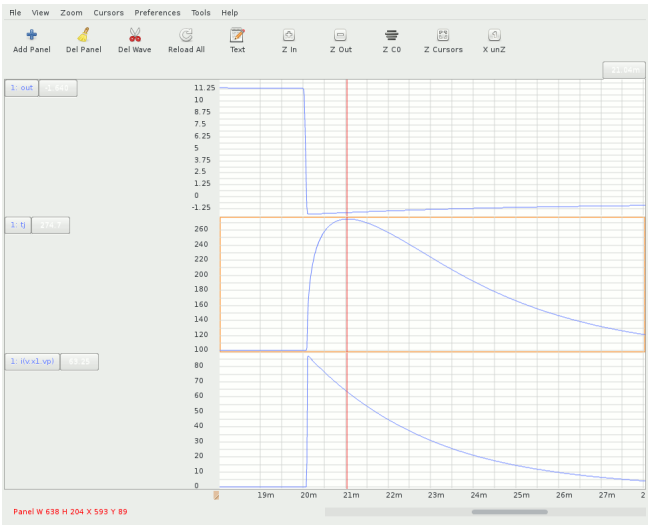


Figure 7.202: Turn off energy provided by free wheeling diode

7.9.3 RF injection into the NMOS HIGH side driver:

The standard gate protection used for most of the high side drivers acts as an RF rectifier. As soon as an RF signal is applied at the output the positive half waves open the diode and charges the capacities of net "gate". (Cgd of Mpower, Cds of M2). This turns on the output stage. To avoid this effect the gate protection can use anti serial zeners in stead of using single diodes.

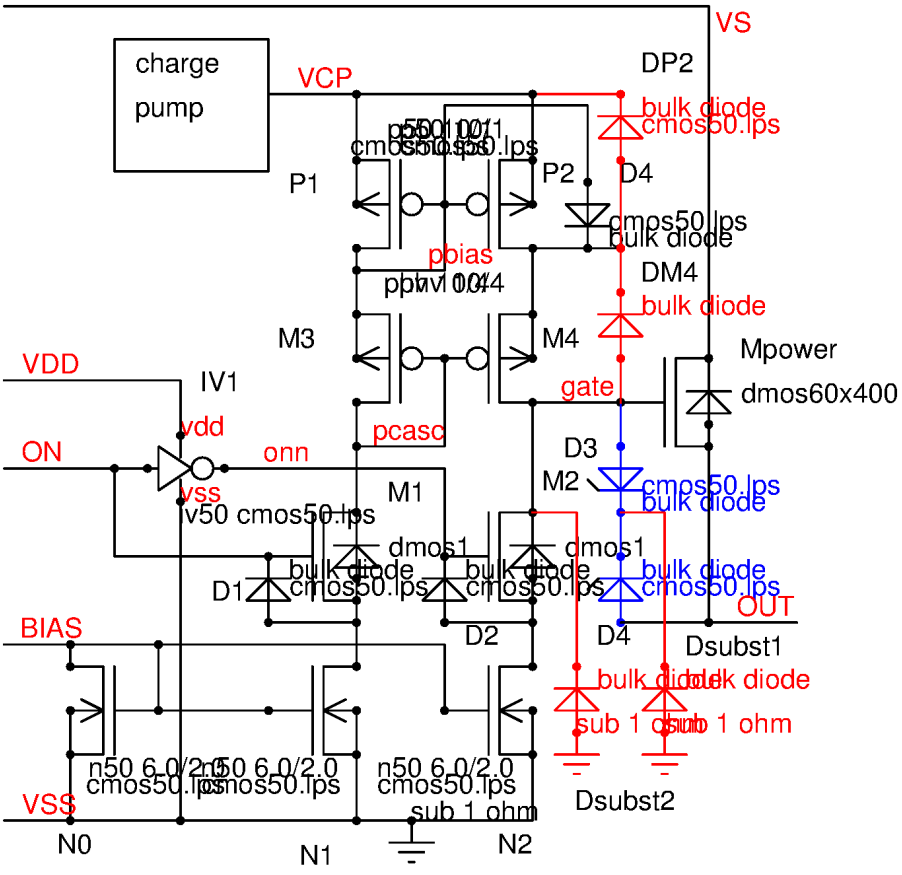


Figure 7.203: RF hardened high side driver

The gate protection now is built with two diodes. The unavoidable substrate diode (Dsubst1, red color) is connected to the middle between the two zener diodes.

The next parasitic rectifiers that will be activated are the substrate diode of M2 (Dsubst2, red color) and the bulk diode of M4 (DM4). In OFF state the gate capacity of Mpower carries the RF signal to Dsubst2. Dsusbst2 will clamp the negative half waves at about -0.7V. So the output voltage at the source of Mpower is clamped to about -2V (depending on the threshold of Mpower). The positive half wave of the RF gets divided by the capacitive divider Cgs and Cgd of Mpower. So Mpower will turn off during the positive half wave. The energy pulling up the output comes from the RF source.

In ON state the diodes D4 and DM4 will clamp the positive half waves (provided the charge pump is low resistive). This degrades the gate drive of Mpower. The drop over Mpower will increase due to the rectification. To push up the RF level, at which the rectification in DM4 and D4 kills the performance the charge pump voltage can be increased (at least as far as the maximum Vgs of Mpower allows it).

Trouble maker current limitation: Additional to the gate protection current limiting circuits can produce a lot of RF sensitivity trouble. The classical current limitation circuit uses a delta Vbe circuit. This is one of the best rectifiers you can build! Furthermore there is a emitter current flowing. So blocking the emitters with on chip RC filters is almost impossible! (You would have to make the resistors very low and the capacitors extremely big.)

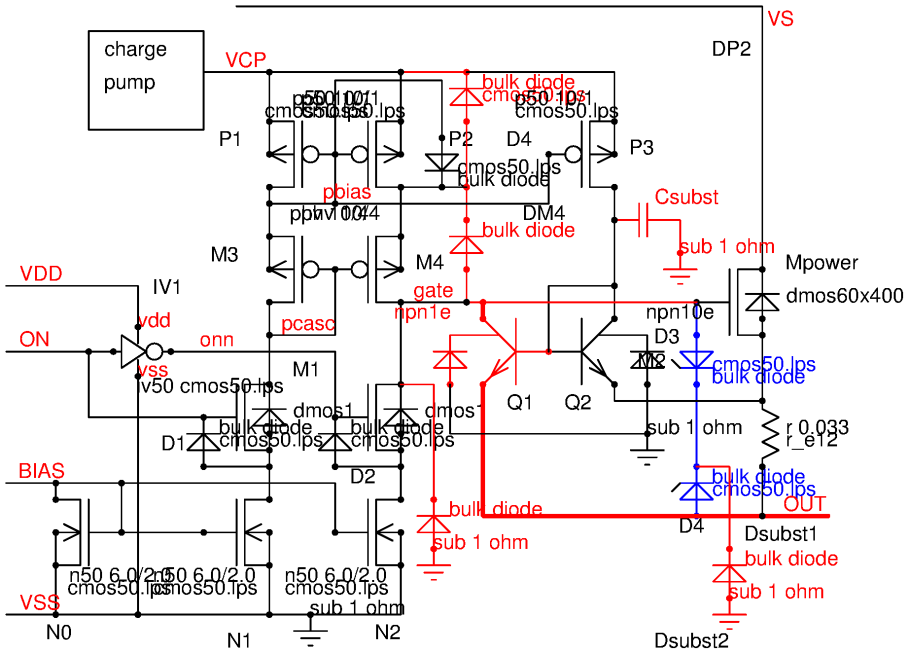


Figure 7.204: High side power stage with delta Vbe current limit circuit

RF injected into the net OUT will unavoidably be rectified by Q1. Thus Q1 pulls down the gate of Mpower. The base charge reservoir during the rectification is the substrate capacity of Q2. The only way out is to prevent RF injection into the emitter of Q1. Since Q1 must be connected to net OUT with a low impedance protecting Q1 on chip is impossible. Typically such circuits already fail at RF power levels of 10dBm to 20dBm. Using boosted NPN current mirrors (driving the base of Q1 and Q2 with an NPN emitter or an NMOS source) makes things even worse!

The only way to build an RF insensitive current limiter is to use a real operational amplifier with offset (and high input impedance!) and connect it to net OUT via an RC low pass filter. But this is a significantly more complex concept and requires more silicon real estate.

7.9.4 Bipolar solutions

Today using bipolar transistors has become very uncommon. Historically bipolar high side power stages usually were built with a PNP driver and an NPN output stage making it a quasi-PNP darlington transistor. This was a usual compromise accepted in the 1980s to take benefit of the (at those times) cheaper pure bipolar technologies.

7.9.5 Floating switch driver stages

Floating switches can easily be implemented using CMOS technologies as long as the maximum permissible gate voltage is less than the supply voltage. This is used in all the standard transmission gates used for 3.3V or 5V.

As soon as the range the switch can float exceeds the maximum gate voltage the gates of the switches must be protected against destruction. Furthermore floating switches must be built in a way that the parasitic diodes don't bypass the switch. In most cases this leads to an anti serial configuration of two switches.

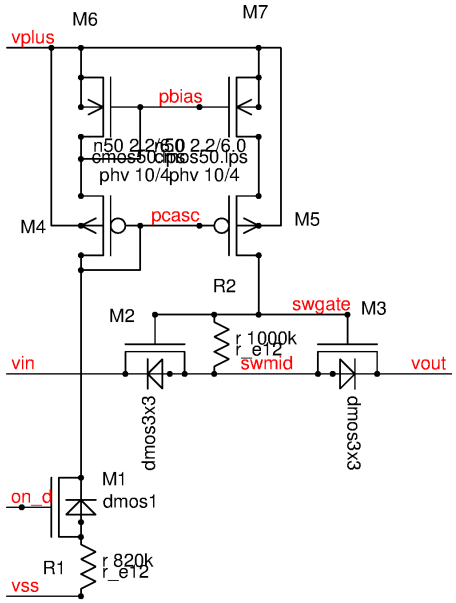


Figure 7.205: floating high voltage switch

This is the most simple variant of a floating high voltage switch. To make it work correctly Vplus must at least be one threshold voltage higher than the voltages applied at vin and vout. The voltage at the gate of M2 and M3 is limited by the logic signal applied at on_d and the ratio of the two resistors R1 and R2. The switch as shown has two weaknesses:

1. The current flowing in R2 is influencing the voltage being switched. As long as the signals applied are low resistive and the required accuracy is low this is not a problem.
2. In order not to influence the switched signal too much the current flowing in R2 usually is chosen low. This makes the switch slow

7.10 Power bridges

Power bridges are used for high current bipolar PWM control circuits. Typical applications are electronic controls of 3 phase electrical motors. The following picture shows the concept of a 3 phase bridge. Often it is called a 6B configuration because it uses 6 power transistors in the bridge.

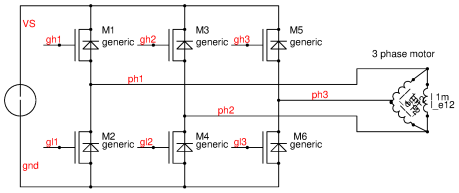


Figure 7.206: B6 3phase bridge

The transistors create the required sinusoidal current by a PWM. The duty cycle of the PWM represents the current. The current flowing in the motor is the integrated PWM voltage. The gate drive pattern looks something like this:

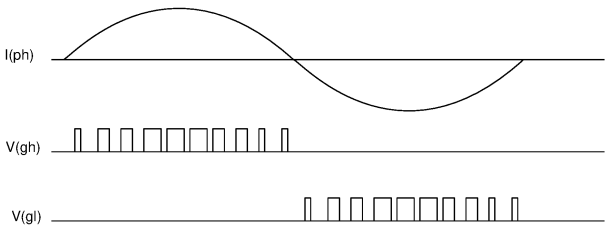


Figure 7.207: Current flowing in the motor and gate drive signals of one half bridge

High power bridges are designed for up to several hundred amperes. Switching such high currents the parasitic inductances between the power transistors and the blocking capacitors can't be neglected anymore. Adding the parasitic inductances the half bridge (one leg) becomes something like this:

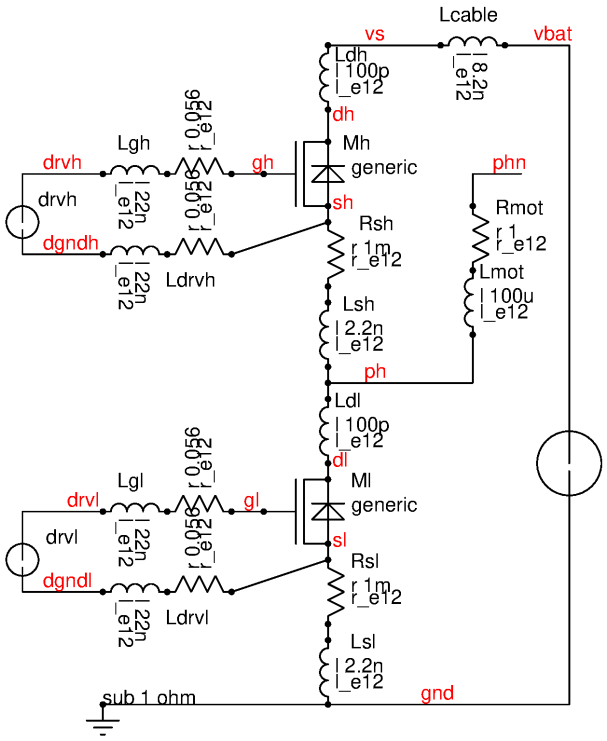


Figure 7.208: Half bridge including parasitic inductances

Most loads are more or less inductive with inductances magnitudes higher than the parasitic indictances. So for a single PWM switching event the load can be approximated as a constant current source.

To start let's assume the load current is flowing into node ph. There are two possible ways the current can flow:

1. Ml is ON: the current flows through Ml to gnd.
2. Ml is OFF: the current flows either through the actively turned ON transistor Mh (active rectifier) or through the body diode on Mh into node vbat.

Things are getting interesting when the state of the transistor Ml changes. The current remains the same but the path it is flowing changes.

Ml turns off: The current has to be taken over by the path Lsh, Rsh, Mh (usually the body diode during the break before make gap), Ldh, Lcable. The voltage overshot at node ph becomes:

$$V_{os_{off}} = I * R_{sh} + \frac{dI}{dt}(L_{sh} + L_{dh} + L_{cable}) + V_f$$

Example: I=100A, dI/dt=10A/ns, $V_f = 3V$(SiC transistor), $L_{sh} + L_{dh} + L_{cable} = 10.5nH. => V_{os_{off}} = 118V$

At the same time the current through Ml decreases. The source of Ml will move negative (versus gnd) and the drain of Ml will move positive versus node ph.

$$V_{sl} = -\frac{dI}{dt} * L_{sl}$$

$$V_{dl} = V(ph) + \frac{dI}{dt} * L_{dl}$$

So the drain source voltage of Ml peaks a little bit higher than the phase voltage.

$$V_{dsMl} = Vbat + V_f + I * R_{sh} + \frac{dI}{dt} * (L_{sh} + L_{dh} + L_{cable} + L_{sl} + L_{dl}) \tag{7.354}$$

Example: all conditions as before and additionally $L_{sl} = 2.2nH$, $L_{dl} = 100pH. => V_{dsl} = Vbat + 141V.$

To protect the transistors against overshots caused by the parasitic inductances the current sloped dI/dt must be limited. Usually this is done by the limitation of the gate drive current turning off and the gate capacity (in combination with the transconductance gm of the transistor being turned off). To design a power stage in a proper way the following parameters must be known:

- All parasitic inductances
- the gate capacity and the transconductance gm of the power transistor (this determines dI/dt)

- The gate pull down current (or plateau voltage and impedance of the driver stage)

Since the gate drive must be fitted to the power stage parameters most gate drivers are intended to be configurable with external resistors or the gate drive pull down currents are made programmable.

Alternatively a fast clamping circuit limiting V_{ds}can be used. However the fast clamping circuit usually relies on zener diodes. Zener diodes tend to age and change their zener voltage at high currents. For this reason a zener clamp can be regarded as a last resort for short circuit turn off but not for normal operation at nominal load current.

Ml turns on: Let's assume the current is flowing through Mh and Ml is off as an initial state. When Ml turns on the current through Mh gets reduced. Neglecting the parasitic inductances the node ph would remain high until Ml has taken over the complete load current.

As soon as we don't neglect the parasitic inductances anymore things start to change! As long as Ml hasn't taken over all the current and there still is some current flowing through the body diode of Mh the voltage at the drain of the high side transistor is:

$$V_{dMh} = Vbat - \frac{dI}{dt} * (L_{dh} + L_{cable})$$

As long as there still is current flowing through the bulk diode (Ml has not yet taken over the whole current) the source voltage of Mh remains at:

$$V_{sMh} = Vbat - \frac{dI}{dt} * L_{dh} + V_f$$

The voltage at node ph thus becomes:

$$V_{ph} = Vbat + V_f - \frac{dI}{dt}(L_{dh} + L_{sh} + L_{cable}) - I * R_{sh}$$

In extreme cases (fast turn on of the low side transistor) the voltage at node ph can already be close to 0V long before Ml has taken over the complete current.

As soon as the bulk diode of Mh turns off dI/dt suddenly approaches zero. The drain of Mh will shoot up and the source will be pulled down. This leads to extremely fast voltage slopes at the (passive) high side transistor. The dV/dt at the passive transistor Mh is faster than the dV/dt seen at the active (tuning on) low side transistor Ml. There are 3 limitations of dV/dt of the passive transistor:

- the drain source capacity C_{ds} limits the speed but produces overshots by ringing. (no risk as long as $V_{dsmax} > 2 * Vbat$ because V_{ph} at the moment the diode current gets disrupted never can be below 0V (gnd))
- The transistor Mh can turn on due to it's miller capacity and the limited pull down capability of the driver stage (harmless, but produces power dissipation. May even be desired to protect Mh against ringing overshots)
- The parasitic NPN turns on because C_{db}provides enough base current to overcome the bulk resistance (usually destructive because the parasitic NPN suffers from current crowding)

To protect against too high dV/dt that might turn on the parasitic NPN the turn on speed of the power transistors must be limited as well.

Note 1: In SiC the hole mobility is about factor 6 lower than the electron mobility. The resistance of the P-bulk can be significant.

Note 2: Often the V_{dsmax} is chosen less than 2*Vbat to benefit from a lower $R_{dson}$and to use a cheaper transistor technology.

Note 3: Symmetrical parasitic inductances ($L_{sh} + L_{dh} + L_{cable} \approx L_{sl} + L_{dl}$) can help to limit the ringing peak over Mh to about 1.5*Vbat. The crux often is the short cable from the power bridge to the blocking capacitors that breaks the symmetry of the parasitic inductors.

Since the voltage overshots strongly depend on switching speed, and switching speed depends on transistor parameters and driver parameters together with parasitic inductances, the power bridge must be optimized for the worst case corners (usually the fastest transistors and the highest currents). As a consequence part of the efficiency must be sacrificed for the nominal components.

Even worse: Usually there are no "worst case fast samples" to develop the power bridge by experiment. Instead the design will have to rely on simulation using corner models!

Currently (2019) SiC transistors have a lot of surface states at the interface between the SiC and the gate isolation (CVD oxide or nitride because there is no natural thermal oxide like in standard Si technologies). This leads to a stress induced threshold shift of about $\pm 0.5V$. This drift depends on the gate voltage applied during the last couple of ms and it is reversible as well.

To build efficient power bridges the spread of the power transistors and the driver stages must either be under tight control or the design must be trimmed. Trimming in volume production however is a nightmare. Finding solutions that are self trimming is advantageous.

7.10.1 Power transistor spread considerations

Big power bridges usually consist of discrete power transistors and integrated driver stages. To keep ringing under control and to prevent voltage overshots that are destructive the driver stage must be tuned to match the transistor parameters as well as the parasitic inductancees. Parasitic inductances usually are well under control because these are mainly determined by mechanical parameters (wire length, permeability of near by materials, inductive coupling between bond wires etc.). For the transistor parameters this is not the case. There are a lot of parameters that have a considerable spread and a lot of influence on the circuit behavior:

1. Threshold voltage of the transistor depends on bulk doping, surface states and gate oxide thickness.
2. transconductance gm depends on Cox, W, L and carrier mobility μ
3. Miller capacity depends on the capacity between the gate and the drain extension or on the parameters of the shallow trench isolation (if poly over STI is used as a field plate)

Warning: mobility μ at the surface can differ significantly from the mobility μ inside the semiconductor usually measured to characterize a material. The deviation strongly depends on the surface states. SiC surface mobility deviates more than the surface mobility of Si.

Most semiconductor manufacturers offer a nominal model and in some cases corner models 'fast' and 'slow'. Regarding inductive overshots the 'fast' model at cold usually is the most critical one. To design a driver stage that works well in all corners the optimization has to be done with 'fast' models. This however leads to a power stage that is too conservative for nominal components. The efficiency using the center of the distribution is lower than desired. Even worse all the cooling of the power stage must be designed for the even slower 'slow' models.

The next undesired side effect: The design can't be verified by measurement because corner models usually are so far at the edge of the distribution that these components will never be encountered in engineering sample production.

All this leads to a design style of using simulation rather than measurement and once production is running every field return has to be verified if there are devices faster than 'fast' or slower than 'slow'. If such devices are found in a field return the models have to be adjusted and the power bridge has to be redesigned using the adjusted corner models. This is a time consuming development process with possibly poor failure rates as long as there is limited production experience.

7.10.2 Protection of the driver stage

High power bridges with modern IGBTs or with SiC transistors can carry currents of hundreds of A during normal operation. Under short circuit conditions currents can reach several thousand Amperes! Worst case usually is a short 2 (the transistor is already on when the short circuit takes place). The voltage drop over the bond wires can reach several hundred volt for up to 100ns! For this reason the driver stage must float up with the source of the power transistor. The input of the driver stage must permit this floating up.

Power bridges with low voltage supply in the range of up to 60V can work without galvanic isolation. The interface between the logic and the driver stage is a levelshift providing the required protection. To illustrate the ground bounce taking place at a short the following simplified schematic of a stepper motor driver bridge is used.

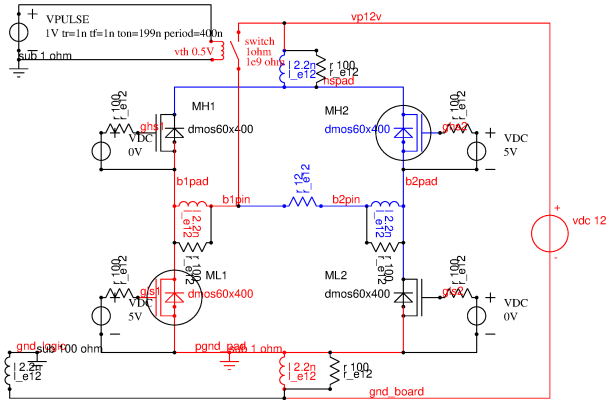


Figure 7.209: A typical power bridge of a fully integrated stepper motor driver

Let us assume transistors ML1 and MH2 are on. The current through the load before the short circuit is in the range of 1A. The short circuit is applied with the switch. During the short circuit the current flows along the red path. The simulation show the ground bounce between the power ground (pgnd_pad) and the logic ground (gnd_logic) when an ideal short takes place. The ground bounce is determined by the ratio of the inductances of the pins b1pin and pgnd plus some ringing.

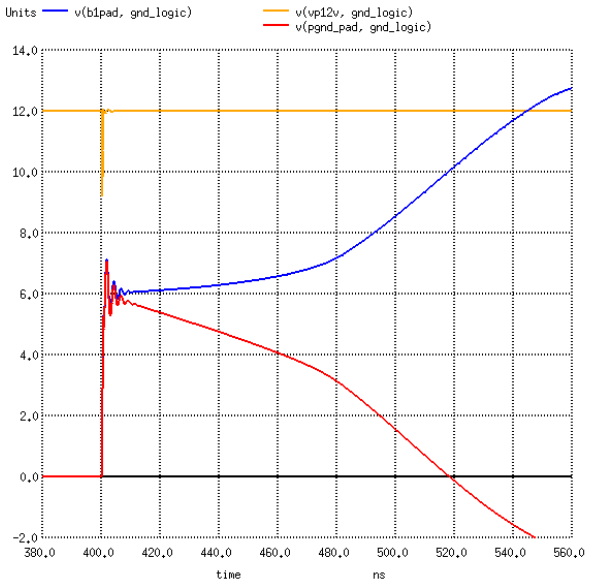


Figure 7.210: Simulation of an ideal short 2

The red curve shows that the levelshift between the logic and the gate driver of the low side must be able to survive 7V offset between the two ground domains. Ruggedness of the levelshift can be estimated as:

$$V_{bounce} > V_{12} * \frac{L_{pgnd}}{(L_{pgnd} + L_{bx})}$$

Ideally the driver of the low side should be separated from the logic by a levelshift able to handle a ground bounce of at least 50% of the maximum supply voltage of the power stage!

Alternatively the drivers of the low side could be grounded on the logic side. This limits the maximum source voltage that can be reached at the node pgnd_pad. This approach limits the switching speed of the power stage to

$$\frac{dI}{dt} = \frac{V_{dr} - V_{th}}{L_{pgnd}}$$

in this equation $V_{dr}$is the supply voltage of the driver stage, $V_{th}$is the gate voltage of the power transistor, L_{pgnd} is the inductance between the source of the power transistor and the ground of the driver stage. Especially for applications switching hundred's of Amperes normally nobody will accept such a slow switching because the switching losses increase significantly. A typical design of a high voltage power bridge uses galvanic separation for the high side stages as well as for the low side stages.

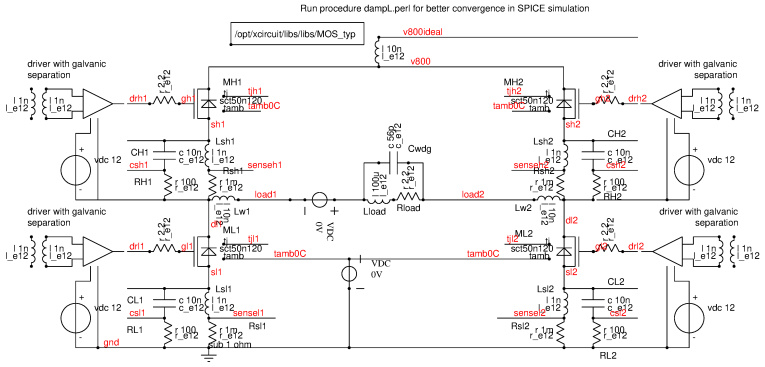


Figure 7.211: High voltage power bridge for 800V, 65A

Each of the drivers can measure the current flowing using the voltage drop over the bond wire. The inductivity of the bond wire is compensated by the capacitor of the read out circuit.

Important: Current measurement is always differential between pin sh1 and csh1 or sl1 and csl1 (voltage accross the capacitor CH1 or CL1). In some cases the ground of the driver stage is connected to sh1 and sl1 to allow faster switching. Details of the implementation often depend on the mechanical design of the power stage and resulting bond restrictions.

In addition often the voltage drop over the power transistor is measured (desat current detection) and the temperature of the complete system is measured. (Theoretically a temperature sensor can be implemented on the

power transistor. But the small signals in the mV range are too distorted in fast switching applications to build a reliable temperature sensor)

The isolation requirements of the galvanic separation are typically 1000V DC, 6000V pulse. So the oxide separating the wires of the transformers must be about $6\mu m$ to $10\mu m$ thick.

7.11 Temperature sensors

There is a big variety of temperature sensors in use. Which type of sensor is used depends on the requirements.

7.11.1 delta Vbe temperature sensor

The most accurate temperature sensors rely on a delta Vbe. Two bipolar junctions are operated with a different current density. The forward voltages of the diodes differ by:

$$\Delta V_{be} = \frac{k * T}{e} * ln(\frac{i_1}{i_2}) \tag{7.355}$$

There are many different ways to create the different current densities. Very often it is the easiest approach to just operate two identical diodes with different currents. The two sensing diodes must be thermally well coupled to the power device to be monitored. In case of trench isolated technologies the thermal coupling is not trivial. Ideally the diodes should be placed in the same trench as the power transistor. In most cases this leads to a sharing of diffusions. If a DMOS transistor is to be monitored the drain of the DMOS transistor often is coincident with the collector of the NPN transistor measuring the temperature!

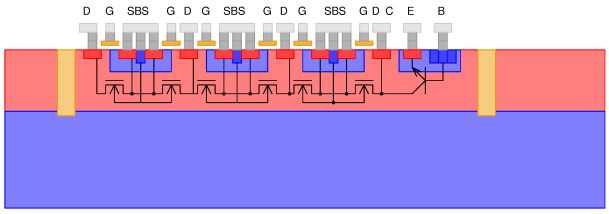


Figure 7.212: Power DMOS transistor and temperature sensor NPN sharing the same trench (nwell)

In the example shown the collector of the sensor is connected to the drain of the power device. In case of a high side driver the temperature sensor circuit is either supplied from the power transistor supply or it can be supplied by a lower supply depending on the circuit driving the base of the NPN transistor and the properties of the NPN transistor itself. If the NPN transistor is not capable of blocking the maximum permissible supply voltage the sensor circuit must follow the drain voltage of the power transistor.

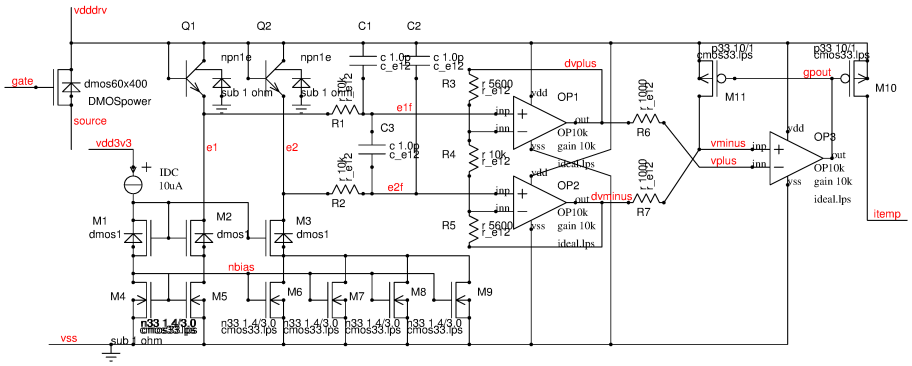


Figure 7.213: Delta Vbe temperature sensor sharing the well of a power DMOS with current output

The sensor consists of two current generators with a current ratio of 4, an RF filter (R1, R2, C1..C3), a differential amplifier with a gain of 2 (OP1, OP2, R3..R5) and a voltage to current converter (OP3, R7, M10, M11). The output current is:

$$I_{temp} = \frac{K * T}{e} * ln(4) * \frac{R_3 + R_4 + R_5}{R_4 * R_7} \tag{7.356}$$

The accuracy of the sensor strongly depends on the offset spread of the amplifiers OP1 and OP2. The offset of OP3 has a lower impact because OP1 and OP2 already have a differential gain determined by R3..R5. Usually this gain is chosen about factor 2.

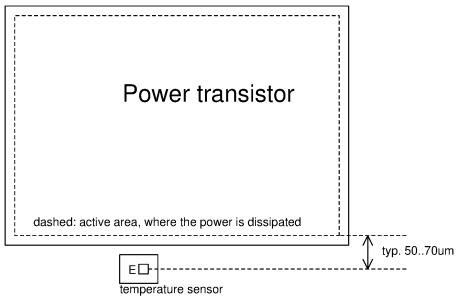


Figure 7.214: Typical placement of a transistor acting as a temperature sensor

7.11.2 Vbe temperature sensor

Most integrated circuits hold a bandgap. So it is very common to compare a bandgap voltage with the forward voltage of a diode. If the same diodes are chosen to build the bandgap and to measure the temperature the difference between the bandgap voltage and the diode simply becomes:

$$V_{temp} = V_{bg} - V_{be} \tag{7.357}$$

Since a bandgap voltage is the sum of a Vbe and a multiplied delta Vbe this equation can be rewritten:

$$V_{temp} = V_{be} + m * \frac{k * T}{e} - V_{be} = m * \frac{k * T}{e} \tag{7.358}$$

Assuming a well designed bandgap (optimized for low temperature coefficient) m is in the range of 25.

If the temperature sensing diode can be placed arbitrary (no fast response of the sensor required, no thermally isolating trenches) this kind of temperature sensor can be implemented with significantly lower cost than building a full blown delta Vbe sensor. Regarding the bandgap as already given the offset requirements for the amplifiers are reduced by factor m!

For simplicity often a diode outside of the power transistor is used to measure the temperature. This way the circuit reading the temperature doesn't need to follow the drain voltage of the power transistor. This reduces the common mode rejection requirements of the temperature measurement system. The draw back of moving the temperature sensor diode out of the power transistor is the delay between the local temperature increase inside the power transistor and the temperature sensor measuring the temperature sensor. Even if the diode sensing the temperature is only $50\mu m$ away from the junction heating up inside the power transistor this delay can already be in the range of about $100\mu s$. Regarding the power density of modern transistors this is about 1 to 2 magnitudes too slow to protect the power stransistor. So relying on the protection of a temperature sensor alone is completely insufficient using modern devices with high current densities. A temperature sensor outside of the power device must always be combined with a current limitation circuit to stretch the time until the power device reaches melt down temperature.

This kind of protection with a sensor about 70um away from the power dissipating are works reasonably well up to a power density of $30W/mm^2$. At higher power densities the power device heats up too fast and when the sensor reaches thermal shut down temperature the center of the power transistor is already at 400 deg. Celsius.

7.11.3 Using the bandgap as a temperature sensor

The bandgap provides a current proportional to Vt for free. The current in the bandgap resistor simply calculates as:

$$I_{bg} = \frac{k * T}{e} * \frac{ln(n)}{R_{bg}} \tag{7.359}$$

n is the ratio of the current densities between the two sides of the bandgap.

Using the bandgap itself as a temperature sensor is a good option if the temperature distribution on the chip is uniform. Power chips that might have hot spots can't be monitored by simply using the bandgap current (Usually the bandgap is not placed close to power devices where the temperature gradients are high).

7.11.4 Modeling the thermal path

Since in most technologies we can't directly obeserve the temperature of the power device the best thing we can do is to calculate the temperature. A very nice proposal how to do this is described in [78] on page 3-23 to page 3.26. In stead of relying on a temperature sensor only we use the signal of the temperature sensor to correct the "far end" of the thermal path replica. The transfer function of the thermal path must be known. To model the temperature the thermal path is modeled in the electrical domain by an RC network having the same (or almost the same) transfer function (thermal path replica).

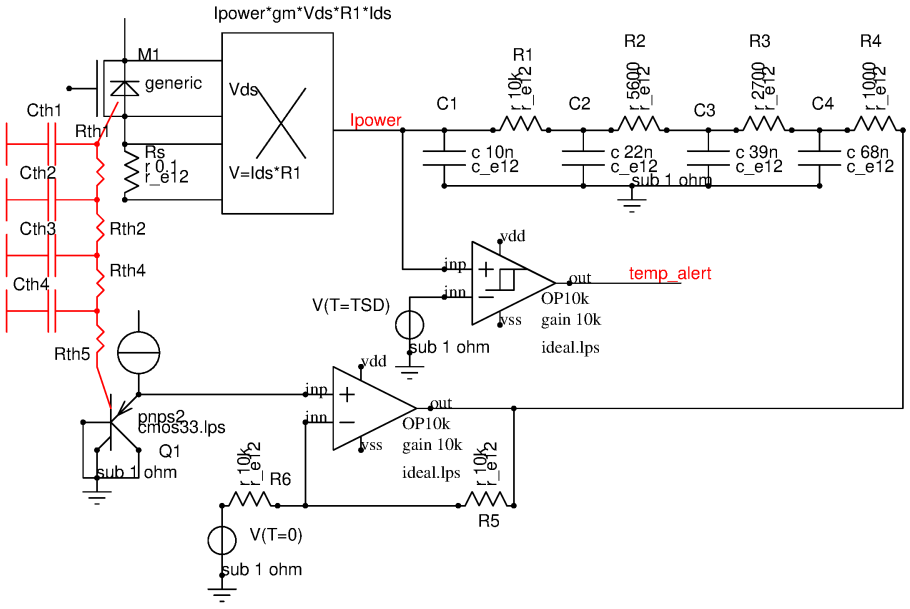


Figure 7.215: Thermal shut down using a replica of the thermal path

The multiplication circuit multiplies Vds and the voltage drop over Rs, the sense resistor. The output of the multiplication is a current. The network R1..R4, C1..C4 has the same transfer function as the thermal path (drawn in red) Rth1..Rth4, Cth1..Cth4 from the "thermal center of gravity" of the transistor to be protected (M1) to the temperature sensor (Q1). The OPAMP simply is needed to correct the difference in gain between the multiplier and the base emitter characteristic of the sensor and to provide a low impedance to drive R4. At the node Ipower we will have a replica of the temperature under the power transistor represented as a voltage. This replica is used by the comparator to produce the temp_alert signal.

Alternative implementations:

Building an analog multiplication is somewhat cumbersome. If the process offers enough logic density to replace the analog circuit by some ADCs and a little ALU (arrithmetic Unit) the multiplier, the path replica and the comparator can be implemented as a hard wired calculation.

If a CPU is available the multiplication and the path model can also be implemented as software. But a software solution already requires quite some CPU performance to do he real time calculation!

The concept of this circuit is very old. I learned it during my education as an engineer at the Technical University Munich in 1985. Later I have never seen any more publications about this concept called "Stoergroessenbeobachter" in German anymore.

7.12 Overvoltage protection

Overvoltage can lead to an avalanche break down of the power transistor. If during the snap back caused by the avalanche the power transistor can draw current from a powerful reservoir (big capacitor) the high amount of available energy will lead to a rapid destruction. This becomes especially critical if current crowding concentrates the energy into a very small area. For this reason in most cases avalanche break down must be avoided. (Avalanche break down may be acceptable if the energy is limited for instance by an inductance between the energy reservoir and the power transistor). Overvoltage protection in most cases has the highest priority (high current usually can be tolerated for a longer time than overvoltage).

The most simple way to limit the voltage is a zener diode stack. The energy can be dissipated completely in the zener diodes. In this case varistors are often used because they offer a higher thermal capacity than the junction area of a zener diode.

In the figure above D1 to D4 protect the transistor against a Vds break down. Diode Dg protects the gate. This is the most reliable protection but the price is high. Diodes D1 to D4 must be deigned to absorb the whole energy.

To reduce the size of the zener diodes the power transistor can be turned on intentionally before an avalanche break down takes place. This works well for older technologies where the transistors can be operated above the "current stable point". In modern technologies where transistors sometimes are operated at low Vgs a local hot spot lowers the threshold and the current at the hot spot increases. This makes an active turn on less reliable than a zener parallel clamp.

In the active turn on protection the zener diodes must override the driver stage. For this reason R1 is added to the circuit. To protect the gate clamp Dg the current through the zener diode chaim must be limited as well. This is done by R2.

Integrated zener diodes often include parasitic PNP transistors that act as a load of the driver stage. (floating nwell of diode D2 is the base of a parasitic substrate PNP transistor. Cathode of D2 is the emitter. D3 adds one

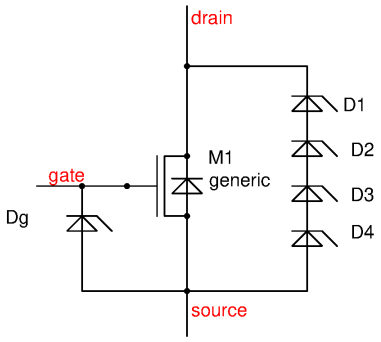


Figure 7.216: Parallel zener protection

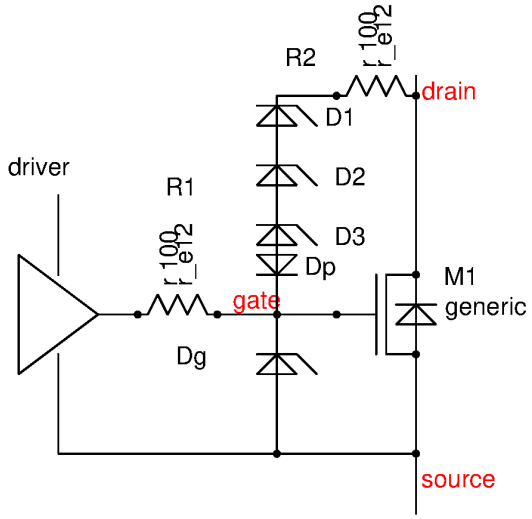


Figure 7.217: Active turn on

more susbtrate PNP transistor.. The longer the zener chain the worse it gets!) Diode Dp disconnects this parasitic PNP darlington transistor from the node gate.

If M1 is a big power transistor the high gate capacity will requires fairly large zener diodes. In this case the zener diode chain may as well just turn on the driver stage in an analog way. The following figure shows an example using a BICMOS technology.

In this circuit the current through the zener diode is amplified by the NPN transistor Q1. This reduces the cost of the zener diodes but has some other side effects to be taken into account:

- The circuit only works as long as the collector of Q1 is supplied.
- To guarantee functionality when supply vdd5 is missing dupply diode Ds2 has to be added.
- Q1 and diode Ds1 must be capable of handling the full voltage swing of the power transistor
- Inverter iv1 is likely to have a parasitic diode (drawn in red). This diode will reverse supply vdd5. vdd5 must be protected against being pulled up

Most likely your circuit won't remain as simple as this conceptual circuit once you start adding all the protections needed! Which protections you will finally need depends on the capability of the process you are using. The final circuit will become very technology specific. This leads to solutions that can hardly be reused in other technologies than the one it was designed for.

7.12.1 Shared protection

If several pins need to be protected several power transistors can share the protection.

This sharing of overvoltage protections works well as long as there are no significant parasitic impedances in series with D1 to D3 and as long as all transistors have a common node (in the example shown the sources are the common node).

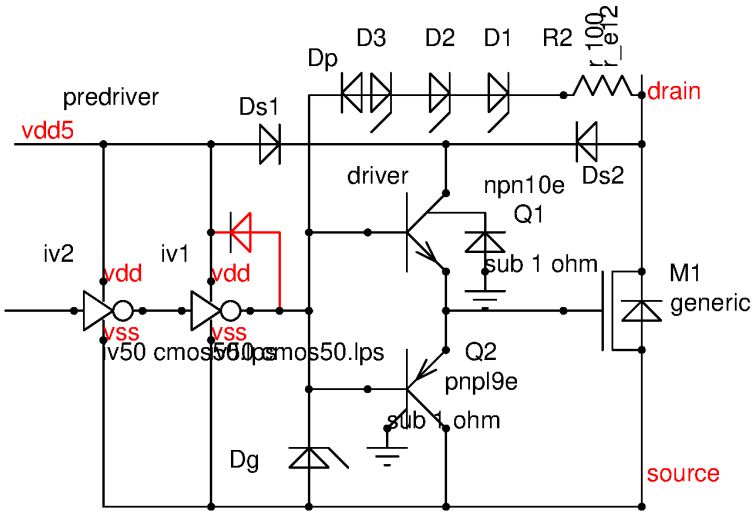


Figure 7.218: Overvoltage protection using the driver to reduce zener diode area

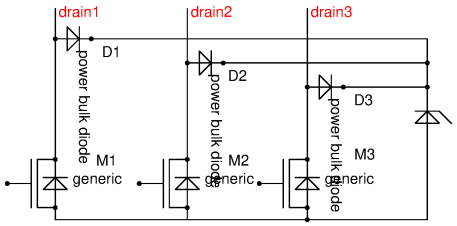


Figure 7.219: sharing overvoltage protection

7.12.2 Tolerances of overvoltage protections

Calculating all the tolerances can become disapointing! The high spread limit of the zener diodes must be below the low spread limit of the power transistors. The tolerance of most zener diodes including ageing and temperature coefficients can be very high. I experienced +-1V for 7V zener diodes. We have to satisfy:

$$n * max(V_z) < min(V_{dsbr})$$

n is the number of zener diodes we are stacking.

Example: Let's assume we want to protect a 60V power transistor. To handle 60V the typical break down of the transistor usually has to be designed for about 75V. This means we are using a 75V device to guarantee 60V break down.

The zener diodes having 7V nominal voltage can get as high as 8V. To protect the 60V minimum break down we can only stack n=7 diodes ending up at a maximum clamp voltage of 56V. The typical clamp voltage is expected to be 49V and the lowest possible clamp voltage becomes 7*6V=42V. These 42V are the highest operating voltage we can specify although we are using a typ. 75V (minimum 60V) power device!

7.13 Overcurrent protection

Power stages need to be protected against overcurrent for the following reasons:

1. Protection against thermal destruction (high accuracy needed)
2. Protection of the bond wires (med. accuracy is good enough)
3. Protection of the metal traces on the chip (med. accuracy is good enough)
4. limitation of flyback energy

Accuracy requirements of current limits: Power stages using a current limit can operate at the limit for a long time. So the limiting value must be matched well with the capabilities of the metalization (electromigration limit). Typically a power driver is rated for a certain life time at a specified current (application current). The current limit is designed higher than the maximum application current. So operating at the current limit leads to a shorter life time. Usually specifications hold a note that the operation at the current limit is permitted only for a reduced time.

Current limits MUST always be combined with a thermal shut down protection because the power dissipation persists after detecting the overload!

Accuracy requirements for over current shut down: If the current limiter doesn't keep the device in operation but turns off the transistor a long term operation at the overcurrent limit is unlikely to happen. Therefore overcurrent shut down circuit usually are permitted to have a bigger spread than limiting circuits. This allows using other current detection methods than only sense transistors and current sense resistors.

Depending on the target of the protection the requirements in terms of speed and accuracy of the turn off limit or the current limit differ. The following table tries to summarize the merits and draw backs of different current measurement methods.

Table 33: Accuracy of over current shut down circuits

method	accuracy	influence	complexity	cost
sense resistor	good	res. spread	low	resistor
sense cell	good	transistor match.	medium	sense transistor
desat sensor	medium	Rdson spread	medium	HV protection
2D method	spread	L, integrator	high	signal processing

7.13.1 Current measurement using a sense resistor

Thermal destruction requires a high current flowing through the power transistor while a high voltage drop is present. Thus the power dissipation is high and the power transistor heats up until it becomes self conducting and the metalization melts.

For resistive loads and switches operating at a low supply voltage using a simple current limit can be appropriate. Wort case the power dissipation becomes:

$$P_{diss} = I_{limit} * V_s \tag{7.360}$$

Just limiting the current of a power device leads to a more or less slow heating up of the power device (depending on transistor area and thermal properties of the package and the chip). A simple current limit usually requires a thermal protection in addition. This kind of approach often is used for analog regulators that are intended to work at a high voltage drop over the power stage as a regular operating mode (for instance linear voltage regulators). The most simple implementation can be a simple resistor and a bipolar transistor acting as the most simple current limiter. This very simple approach even offers the nice feature of a decreasing limit with increasing temperature if the bipolar transistor is thermally coupled to the power transistor.

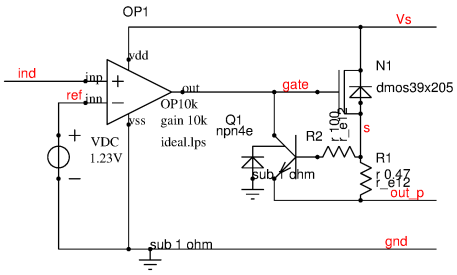


Figure 7.220: A simple Vbe/R current limit circuit

The pull down transistor Q1 must be stronger than the drive capability of the driver amplifier OP1.
The circuit as shown here has two disadvantages.

1. The current limiter requires a drop of one Vbe - about 500..900mV depending on temperature - over R1.
2. Since Vbe changes dramatically over temperature the specification doesn't look "nice". Although technically this isn't a problem for most applications it is a disadvantage for selling the product. (Customers are willing to pay money for things that technically doesn't make sense. Therefore marketing departments tend to sell useless thing just because they look nicer.)
3. Spread of R1 (usually +-20% for most technologies) can't be trimmed out in an easy way.

To reduce the drop needed to measure the current using amplifiers with a designed trip point of just a few milivolts often is preferred over the simple Vbe over R limitation.

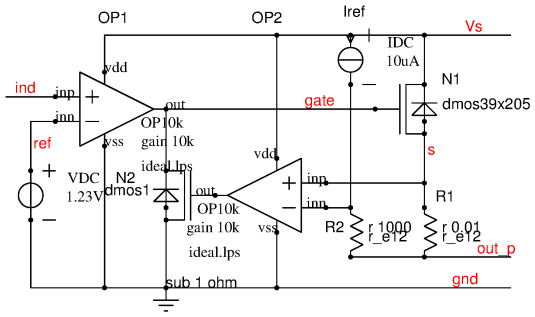


Figure 7.221: Current limitation using an opamp

In the circuit shown above the amplifier OP2 compares the drop over R2 with the drop over R1. In the example the drop over R2 is 10mV. A significant reduction versus 0.5..0.9V of the Vbe limiter. The accuracy of the limit depends on the matching of R1 and R2 and the accuracy of the reference current source Iref. Trimming Iref the accuracy of such a circuit looks much nicer than the wide range of the Vbe over R limitation. Well, this is not really a big advantage for many applications in real life, but is sells much better!

There is one reason why an accurate limit may be needed: If it is not really the power transistor to be protected but some other device in the system such as the cables or the supply source providing Vs. Complex systems often have nested limits such that the switch shown here has to limit current consumption before Vs breaks down due to the limitation of the supply one hierarchy level above.

The solution using an opamp as a current limiter offers the possibility to place an EMC filter between the load (Path N1, R1) and the inputs of the amplifier.

The price to be paid for OP2 is high! The reasons are:

- OP2 must have a very wide common mode range that covers the complete output voltage swing of out_p
- The inputs of OP2 are directly connected to a power output stage. It must be protected against all possible transient voltages.
- The regulation loop N1, R1, OP2, N2 can be quite tricky to compensate. (Leads to expensive capacitors on chip)
- The reference source Iref must be able to operate at the lowest drop possible over N1. (In extreme cases this can be a few mV!)
- Designing matched resistors with several magnitude difference in resistance can be come extremely expensive (try to build a 1K metal resistor to make R1 cheap or a $10m\Omega$ poly silicon resistor to make R2 cheap. it simply won't work well!)

For applications with high supply voltage a current limiting approach can't be done anymore. Instead of limiting the current the power transistor must be turned off before it gets thermally destroyed. (A temperature sensor that normally is combined with a current limit can't follow the temperature if the temperature increases faster than 100K/ms). The current regulator will be replaced by a Schmitt trigger and a latch. The power stage is turned off when the current limit is exceeded for a certain time.

Having a filter time (blanking time) is mandatory for an overcurrent turn off to prevent premature turn off just because of EMC or inrush currents of capacitive loads.

Implementing an on chip blanking time in combination with a Vbe/R sensor is a problem because the base node of a bipolar transistor is fairly low resistive and the filter capacitor would become big and expensive. Therefore Vbe/R circuits are very uncommon for overcurrent shut down applications.

Filtering inrush currents using a comparator (with high input impedance) as a sensing element works better. The following figure shows a possible solution.

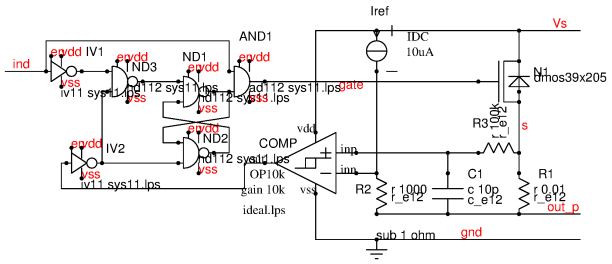


Figure 7.222: Simple overcurrent shut down

The circuit consists of a filter (R3, C1) to filter inrush currents, a comparator COMP (replacing the former opamp) and an error latch (ND1, ND2). The real implementation additionally needs level shifts between the comparator and the logic and between the logic and the gate drive signal gate. These were omitted for simplicity.

To turn on the power transistor the logic input ind must be HIGH.

If an overcurrent is detected the output of the comparator becomes HIGH. Output of ND3 becomes HIGH independent of the state of ind to make the short detection dominant over any other control signal. At the same time the output of ND2 becomes HIGH and output of ND1 becomes LOW. The over current detection is latched and the signal gate is pulled down.

Reset of the latch is possible as soon as the comparator COMP doesn't detect any more overcurrent. Now the reset path via ND3 is sensitive again. A LOW signal at ind will reset the latch again (until the next overcurrent event is detected).

The next rising edge of ind will turn on the power transistor again.

The error latch logic can be implemented in several ways each having it's pros and cons.

For logic test coverage reason logic designers will tend to pull the error latch into the logic. But latches are not synthesizable. Instead a synthesized logic will have some kind of clocked finite state machine representing the function of the latch. Short circuits often lead to dramatic ground bounce issues on the chip. The clock oscillator is one of the most sensitive functions. Implementing the error latch as a finite state machine can lead to a disaster if the clock stops at short circuits or the supply of the level shifts collapsed during a short circuit.

The more fool proof way is to handcraft the error latch as a real error latch outside of the logic (inside of the power driver stage). If done in a clever way this approach also avoids or at least reduces the risk of level shifts not being supplied due to ground offsets between the power stage and the logic. This however leads to local logic that isn't covered by scan test. A second good reason for having these few gates inside the power stage is that you can use 5V or 3.3V gates to achieve a better noise margin than you would get in the synthesized 1V logic.

In extreme cases the error latch can be represented by a thyristor discharging the gate! The advantage: As long as there is gate charge the thyristor will work (even if nothing else is supplied anymore because the power stage is so strong that the power supply collapses). When the gate voltage is gone you don't need the latch function anymore.

My personal experience: As soon as the first samples come back from the field because the short detection failed due to a loss of clock or a fail of a levelshifts your customer will make you move the latch into the analog part!

7.13.2 Current measurement using a sense transistor:

To get rid of the resistor matching problem the power transistor can be split into a main current path and a low current path. Both transistor segments have to be operated at the same voltage drop to achieve the same gate voltage and the same Vds (important in triode mode operation of the power stage).

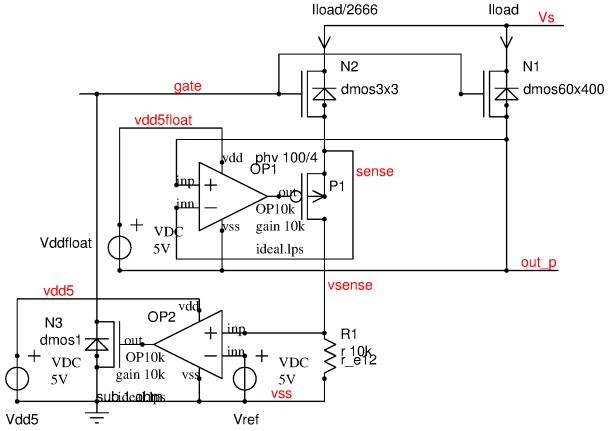


Figure 7.223: Current limitation using a sense transistor

In the circuit above the regulation loop OP1, P1 creates a copy of the source voltage of N1 at the source of N2. So N1 and N2 operate at the same current density. The current through N2 is a fraction of the current flowing through N1 according to the aspect ratio between N1 and N2. The ratio between N1 and N2 can be optimized to use the most economic sense resistor R1.

Since we replaced one opamp by two operational amplifiers the question is: What is the benefit?

1. The two amplifiers used now are both low voltage designs. This reduces cost significantly
2. The voltage representing the current at vsense has a high swing. So the precision requirements for OP2 are low.
3. The trip point of the current limit can be chosen arbitrarily. The common mode range requirements of OP2 are low.

4. The voltage vsense is referred to system ground. This opens the door to measuring the current with a simple ADC. (Very nice for digital regulations)

For applications with high Vds the current limit again must be replaced by a overcurrent shut down to prevent thermal destruction before the temperature sensor responds. Using a sense transistor an overcurrent shut down looks like this.

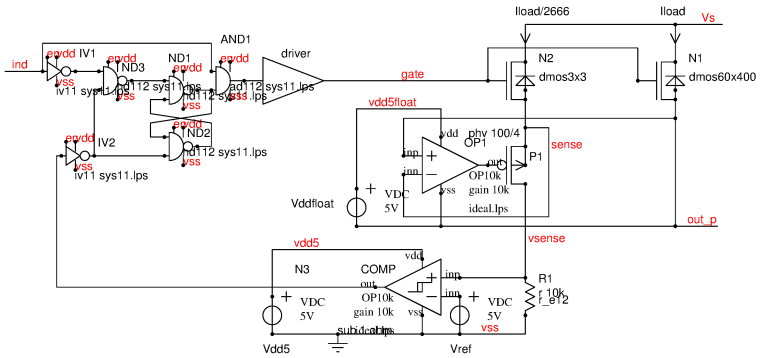


Figure 7.224: Overcurrent shut down using a sense transistor

The shift of the comparator into the low voltage side reduces the requirements for level shifts between the comparator output and the latch function. The driver shifting the logic signals up to the gate of the power stage is needed anyway because the logic gates can't drive a power gate directly.

7.13.3 Desat current sense:

The concept of a desat current sense (The expression 'desat' dates back to the bipolar transistors where saturation was the expression of operation with very low Vce - Using MOS transistors the expression 'desat' becomes a bit confusing I agree.) is that the Rdson of the power transistor is (at least approximately) known. The destruction mechanism of the power transistor is thermal (a certain drain-source voltage and high current at the same time). In stead of measuring the current with high precision the voltage drop over the transistor is measured. The measurement only makes sense when the transistor is known to be on.

If the transistor is off the drain voltage can be very high (1000V in case of IGBTs or SiC transistors!). Usually such high voltages are beyond the capabilities of most integrated circuit technologies. So external resistor networks in combination with clamp structures are needed to protect the amplifier inputs.

The benefit of desat current detection is that no sense resistor is required and no sense pin is needed. At hundreds of Amperes avoiding additional high current components is crucial. Furthermore sense transistors always create protection issues at high dI/dt and resulting high inductive voltage drops (that differ between the main transistor and the sense transistor).

The basic concept of a desat protection looks as shown below:

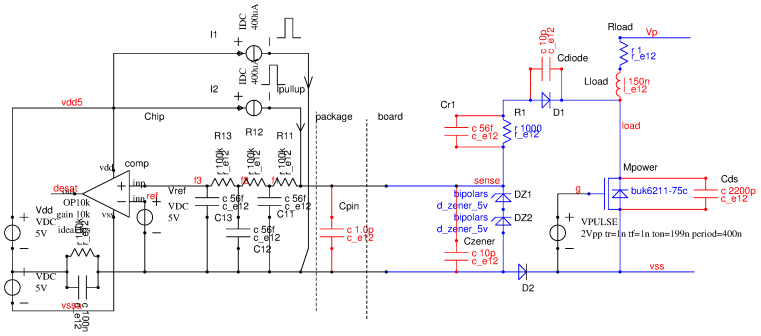


Figure 7.225: Desat protection

The circuit consists of components on the board drawn in blue color and components inside the chip drawn in black color. In addition there are parasitic components drawn in red color.

The power driver often is supplied from two supply rails, one about 5V below the source voltage of the power transistor, the other one about 15V above the source of the power transistor. This is represented by the voltage sources connecting the desat supply to the source of the power transistor via a 100K resistor with 100nF in parallel. (The capacitor must provide the energy for charging and discharging the power transistor gate from the same - more or less - floating supplies.) For typical motor controls the cable has a length of some cm to some 10cm. The parasitic inductance is in the range of some hundred nH. In fast switching applications (e.g. SiC transistors!) the

load inductance together with the parasitic capacity Cds of the transistor can produce severe ringing. This ringing can harm the performance of the desaturation detection. To filter the ringing the chip holds a low pass filter. The cut of frequency of the low pass filter should be tuned to the requirements of the application! (R11, R12, R13, C11, C12, C13)

To protect the chip from the high voltages at the drain of the power transistor the signal disconnected by diode D1. Since the diode has a significant junction capacity the signal additionally is attenuated by a resistor R1 and clamp diodes DZ1 and DZ2 on board level. If the power transistor is on the current of Ipullup will flow via R1 and D1 into the drain of the power transistor. If the drop over the power transistor increases diode D1 will block and the voltage of node sense increases.

For IGBT transistors the drop of diode D1 is negligible compared with the drop to be measured over the transistor. In this case diode D2 can be omitted. (Typical trip points using IGBT are in the range of 8V)

Using SiC power transistors the drop in on state usually is expected to be below 2V. In this case the diode drop can no more be neglected. To compensate the diode drop in the measuring path a second diode is added in the source path. Both diodes are driven by the same kind of current generator.

The current generators may only be on while the transistor is in ON state. The reason is that if the current generators would be on in off state the drain of the power transistor would slowly get charged. The current is coming from a floating power supply that can float up too if the source of the power transistor (High side stage) floats up. Therefore the current sources operate in pulsed mode with typical turn on times and turn off times in the range of 100ns. Accuracy and matching of these current sources usually in uncritical because the diodes have an exponential increase of the current when the drop over the diodes reaches the forward voltage.

The resistors have parasitic capacities bypassing them. For a 0603 resistor typical bypass capacities are in the range of 40fF [53].

Entering the chip the pin capacity Cpin additionally influences the circuit behavior.

Since the trip point of a desat detection usually is in the range of 5..10V this kind of protection serves well for hard short circuits but doesn't work well for soft shorts (Vds in the range of some V but currents are already in the hundreds of Amperes). Adding a thermal protection to the power transistors to detect soft short circuits is strongly recommended.

7.13.4 Current measurement using the bond wires

The concept is to measure the inductive voltage drop of a piece of wire (for instance the pin of the power transistor or the bond wire of the power transistor). The voltage drop is:

$$V_{ind} = L * \frac{dI}{dt} \tag{7.361}$$

If L is known the current can be calculated simply integrating the inductive voltage drop.

$$I = \frac{1}{L} * \int V_{ind} dt \tag{7.362}$$

This works well as long as the inductive drop is significantly higher than the resistive drop. If the resistive drop has a significant signal contribution the equations start to change:

$$V_{meas} = V_R + V_{ind} = I * R + L * \frac{dI}{dt} \tag{7.363}$$

The Laplace transformation is:

$$V_{meas}(p) = I(p) * R * (1 + p) \tag{7.364}$$

$$I(p) = \frac{V_{meas}(p)}{R * (1 + p)} \tag{7.365}$$

with $p = \frac{j\omega L}{R}$. To get back the current we just need a transfer function of

$$gain(p) = \frac{1}{R * (1 + p)} \tag{7.366}$$

This is nothing else than a lossy integrator with a resistor in parallel! So the current measurement looks like this:

The resistors and the inductors must match:

$$R_{int} = \frac{L_s}{C_{int} * R_s} \tag{7.367}$$

The resistor $R_g$simply defines the gain. It can be determined looking at the DC condition.

$$V_{out} = I_s * R_s * \frac{R_{int}}{R_g} \tag{7.368}$$

Usually the ground bounce of the sense resistor is so high that is simply makes no difference which polarity we use for the integrator. At the and we have to remove the common mode signal by a differential stage anyway. This leads to the following first concept of building a current sensor:

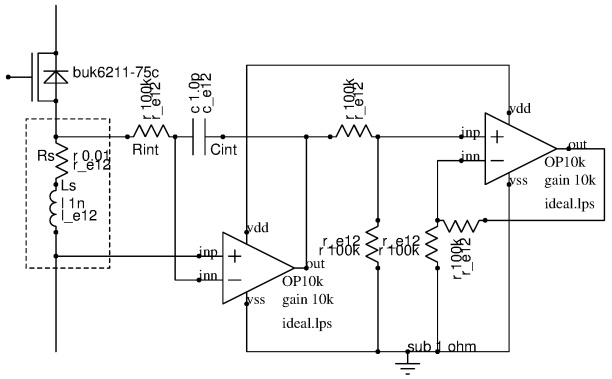


Figure 7.226: Current sense using a classical integrator

The straight forward looking circuit has a couple of problems. The first operational amplifier needs a fairly wide common mode range. In addition due to ringing the common mode range has to be extended to negative voltages. In addition the bandwidth requirements of the integrator amplifier are high.

Since we need a non ideal integrator we can just as well compensate the high pass characteristic of the bond wire inductance using a passive low pass filter with matching corner frequencies.

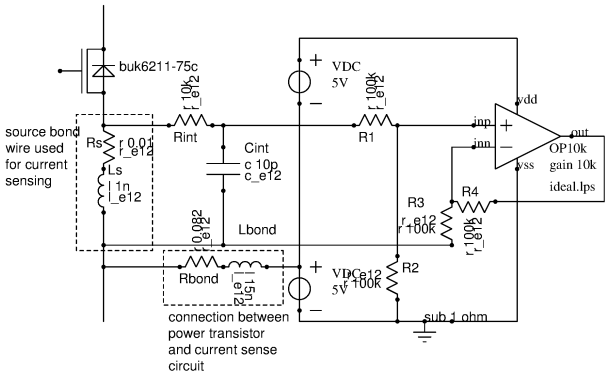


Figure 7.227: 2D current sense using a passive compensation network

The drop over the bond wire of the power transistor as before (for simplicity represented in the frequency domain) is:

$$V_{meas} = I(f) * (R_s + j\omega L_s) \tag{7.369}$$

The resistor R_{int} together with capacitor C_{int} can be regarded as a voltage divider with complex resistors.

$$V_{Cint} = V_{meas} * \frac{\frac{1}{j\omega C_{int}}}{R_{int} + \frac{1}{j\omega C_{int}}} \tag{7.370}$$

$$V_{Cint} = I(f) * (R_s + j\omega L_s) * \frac{\frac{1}{j\omega C_{int}}}{R_{int} + \frac{1}{j\omega C_{int}}} \tag{7.371}$$

$$V_{Cint} = I(f) * \frac{j\omega L_s + R_s}{j\omega R_{int} C_{int} + 1} \tag{7.372}$$

To achieve a flat frequency response the ratio between the imaginary part and the real part of the numerator and the denominator must be equal. This leads to the condition:

$$\frac{j\omega L_s}{R_s} = j\omega R_{int} C_{int} \tag{7.373}$$

We can easily calculate the integration capacity needed:

$$C_{int} = \frac{L_s}{R_s * R_{int}} \tag{7.374}$$

The transfer function of the passive network is flat starting from DC. So we simply get:

$$V_{Cint} = I * R_s \tag{7.375}$$

This works reasonably well as long as the resistors R1 and R2 are at least one magnitude bigger than the resistor R_{int} .

7.14 Save operation area protection (SOA protection)

Save operation area protection was common for class A and class AB amplifiers. The basic idea is to limit the power dissipation. The current limit changes with the voltage drop over the power transistor. Usually a SOA protection is combined with a temperature measurement. The power is limited to a value that the temperature sensor still can follow the increase of the temperature.

Today power densities of shorted transistors are so high that the temperature may already reach destructive values within a few μs. The temperature sensor usually is some $10\mu m$ away from the heat source and has a time constant in the range of $100\mu s$. This is long enough to exceed 400 deg. C and the power device will become self conducting. As a consequence it isn't possible to turn off the power device anymore. Therefore today SOA protections are barely found anymore.

7.15 Logic gates and flip flops

Throughout semiconductor history many different logic families have been invented. In the beginning the logic designs used the easier to produce bipolar technologies. For logic operations diodes (being the easiest to manufacture) were preferred. In the 1980s NMOS technology and CMOS technology replaced the bipolar logic designs. Since the 1990s CMOS is the dominant technology because it provides lowest possible current consumption (at least as long as the logic is in a static state) and at the same time offers the smallest transistors.

Current mode logic and ECL are still in use for very high frequencies (GHz prescalers etc.) but the field of application is getting more and more narrow.

7.15.1 Logic Synthesis

Today the most common design style for logic is logic synthesis. The logic equations are given in a verilog or VHDL file. The logic synthesizer tries to map the logic equations to a design library holding gates and flip flops. After having created a first gate level netlist a tentative placement of the cells in a layout takes place. In this tentative layout the trace length are determined to annotate the wire capacities to each gate output.

Most critical part of logic synthesis usually is the clock tree. The synthesis tool automatically inserts clock buffers where long clock wires are to be driven.

During wiring of the logic gaps between the gates have to be added for routing. These gaps are filled with so called filler cells. The choice of the correct filler cells can significantly affect EME of the logic (RF emission). There are filler cells with capacitors between the supply rails and filler cells without capacitors. Which type of filler cell is used depends on the parameters set for the place & route software. For low emission filler cells with blocking capacitors are to be preferred. Practical tests show that the on chip blocking can easily vary by a factor 3 depending on the types of filler cells chosen.

7.15.2 Inverters

In the following the basic inverter is used to illustrate the concept of each logic family.

DTL inverter: DTL (diode transistor logic) was the first logic family used in the beginning of the 1960s. It attempted to provide logic functions at the lowest possible transistor count. Diodes were used to provide the logic functions. The transistors were only used to recover the loss of the signal swing. Motorola offered one of the first DTL solid state IC families introducing the MC930 around 1964.

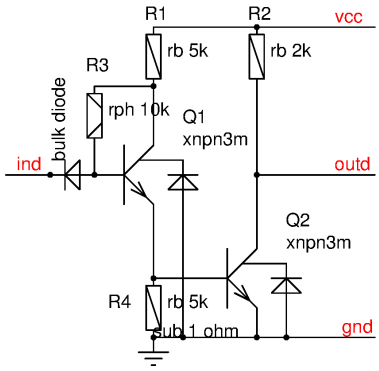


Figure 7.228: DTL inverter

In this DTL inverter Q2 operates in saturated mode. At turn off the delay of Q2 determines the speed of the logic. Using a low load resistance (R2) makes the logic faster but increases the current consumption.

Driving capacitive loads the falling edge at the output is faster than the rising edge.

TTL inverter: TTL (transistor transistor logic) was introduced a short time after DTL. The most important feature of TTL was a totem pole output stage. This output stage offered more pull up drive current than DTL and at the same time reduced power consumption while the output was LOW.

First TTL logic ICs were offered 1964 but it took until 1970 when Texas Instruments offered the 7400 Series, which was a low cost spin off of the 5400 military temperature range series.

Main idea of the TTL output stage was to make the switching speed independent of the edge. (See DTL inverter). Therefore a push pull stage was introduced. Since PNP transistors are slower than NPN transistors the complete gate only uses NPN transistors and resistors.

As an input stage a multi emitter NPN transistor is used. The base-collector junction of the input transistor is misused as a diode.

To improve the speed of the logic the transistors either were doped with gold to reduce the reverse recovery time (turn off time) or Shottky diodes were used to prevent saturation (74Sxx and 74LSxx series). One of the draw backs of gold doping was a significant decrease of the current gain of the transistors. Besides that gold doping created a high recombination noise in all the transistors. Production lines polluted with gold could never be used again for high performance analog products!

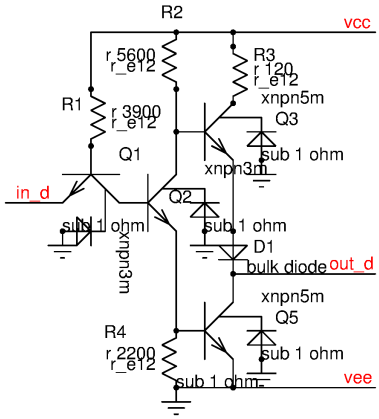


Figure 7.229: TTL inverter

In this logic Q1, Q2, and Q5 saturate. If Schottky diodes are used the base-collector junction of Q1, Q2 and Q5 is bypassed with diodes. Standard TTL propagation delay is in the range of 40ns. TTL-LS offers about 15ns and the lower resistive (more supply current) TTL-S family even achieved gate propagation delays of about 3ns.

I2L inverter: I2L (integrated injection) logic was an attempt to reduce component count and size. To reach this target the collector and the emitter of the NPN transistors was swapped. This allows using one big N-pocket and one base diffusion for multiple open collectors built with the N-emitter doping. In fact the NPN transistors were operated in reverse mode. To reduce chip real estate low gains of the reverse operated transistors (often in the range 4..10) had to be accepted. I2L was used for the first LSI (large scale intergration) chips in the 1970s. Using the NPN transistors in reverse mode limited the speed of I2L logic to about 5..40MHz (depending on the injection current).

In the following drawing the transistors N+ side is shown as an emitter although it is used as a collector.

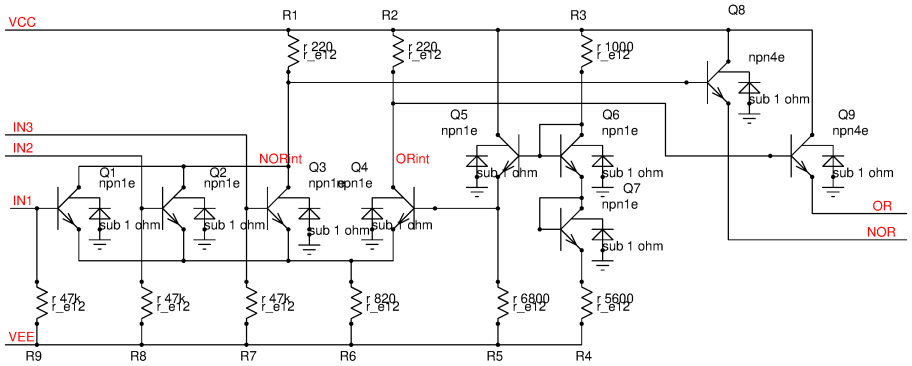


Figure 7.231: ECL OR NOR gate

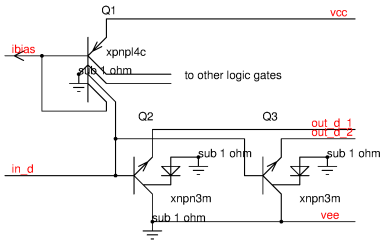


Figure 7.230: I2L inverter with 2 outputs

The open collector outputs of the I2L gates can be used to build wired AND structures. The wired AND is the only logic gate available in I2L logic. The cell shown consists of only two transistors (Q2 and Q3 are nothing else than a multi emitter NPN operated in reverse mode). Q1 even can be shared between 3 gates. So the number of N-epi regions per gate drops down to 1.33. The lowest die real estate achievable in bipolar logic design!

I2L mainly was used inside the chip. To interface to the outside world standard TTL output stages were used.

The gain of an NPN transistor operated in reverse mode strongly depends on the doping level of the N-epi, which is the collector of the NPN transistors when operated in forward mode. High voltage technologies having a low N-epi doping lead to a low gain of the I2L logic transistors. This is why I2L usually is incompatible with technologies offering high voltage NPN transistors. (Usually the limit is about Vcemax=20V. If I2L is needed inside a high voltage bipolar technology an additional N+ implant below the base of the I2L transistors is needed. This leads to one additional mask for the I2L logic inside a high voltage technology)

ECL inverter and NOR gate: ECL (emitter coupled logic) attempts to operate at the highest possible speed. To achieve high speed saturation of the bipolar transistors had to be avoided. First ICs (Motorola MECL1 series) were introduced 1962. The most successful series was Motorola's MECL10000 of 1971. [79]

The concept behind ECL was to reduce the voltage swing to avoid saturation of the bipolar transistors. Additionally the reduction of the voltage swing reduces the energy needed to charge the switching nodes. For high clock rates (100MHz and higher) ECL offered a better figure of merit (power consumption per MHz clock frequency) than TTL, TTL-S and TTL-LS. Propagation delays of ECL10000 are in the range of 2ns. ECL3, which was designed for highest possible speed, achieved about 1ns gate propagation delay (at about 3 times more current consumption than ECL10000) in 1962.

ECL flip flops implemented in OXIS technology in the 1980s already reached operating clock rates of more than 1.3GHz ([64], SDA4211). The concept of ECL later was used for extremely fast CMOS logic too. There it was called CML (Current mode logic) or serial low voltage differential interface.

ECL logic levels refer to the supply VCC. The HIGH level is VCC-Vbe. The LOW level is about VCC-Vbe-1100mV. This low swing in needed to keep the differential stage Q1 to Q4 out of saturation. The choice of the signal swing is a computerize between speed and supply noise immunity.

On chip ECL logic with well controlled supply rails uses a lower signal swing in the range of 150mV to 200mV for internal signals.

NMOS inverter: NMOS logic picks up the concept of DTL and reimplements it using NMOS transistors. This made sense in the early 1980s because CMOS was still to expensive to produce but NMOS transistors have already become smaller than I2L logic elements. Early samples of the 8080 microprocessor and the Z80 microprocessor were implemented in NMOS logic.

CMOS inverter: CMOS (complementary MOS) logic is today's (2015) main stream logic technology. It offers low current consumption (ideally no DC current consumption) and small transistor size. For most standard applications it is desirable to create a fast inverter using the minimum transistor length available.

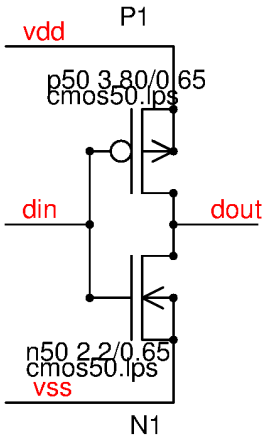


Figure 7.232: Standard CMOS inverter

As long as the input voltage is below the threshold of N1 or above vdd-Vth of P1 the circuit is currentless and only one of the transistors is on. If the input voltage at din is $Vth_{N1} < V(din) < (vdd - Vth_{P1})$ there is a significant cross conduction.

For some applications it is desirable to have inverters with low current or low speed (digital delay lines etc.) In this case it is common practice to stack the inverter transistors with current generators. The current generators can either be placed between the inverter cell and the supply rails or the inverter is taken apart and the current generators are placed in the middle. The following figure shows both flavors of starved current inverters.

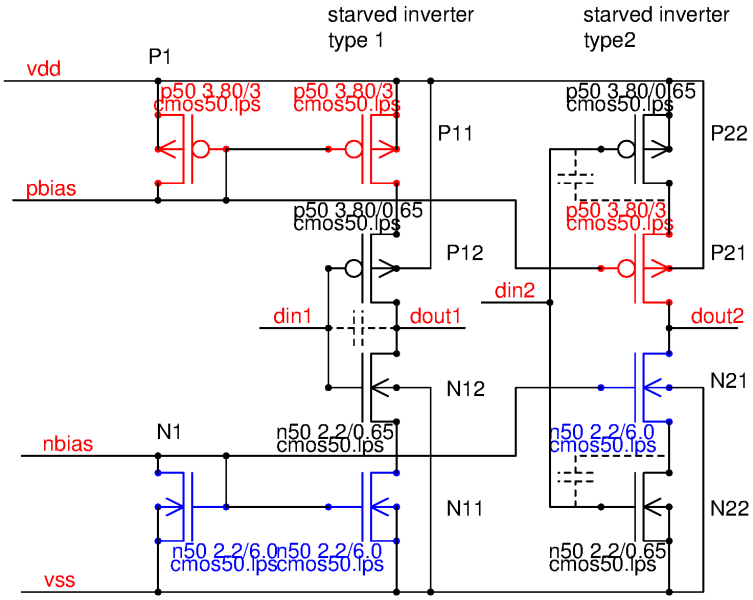


Figure 7.233: Two versions of the starved current inverter

The miller capacities of the inverter switches are symbolized by the dashed lines. The current generator transistors are colored (PMOS in red, NMOS in blue).

The miller capacities of the inverter stages feed the fast edge present at inputs din1 and din2 through the inverter stage directly to the output. In case of starved current inverter type 1 this leads to a spike before the inverter switches. In case of starved current inverter type 2 the spike reaches the current generator transistors but gets attenuated by the resistance of the current generator transistors P21 and N21.

In the following plot the inverters were operated without load capacity to make the charge injection well visible. In typical applications the starved current inverters drive a known capacitive load to obtain a well defined slope at the output.

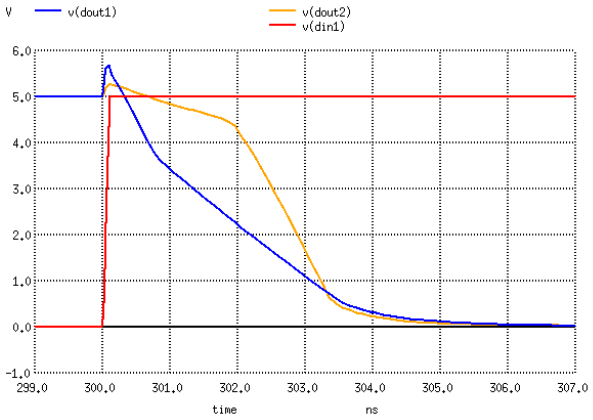


Figure 7.234: Different capacitive feed through of the two types of current starved inverters.

Placing the current generators between the inverter transistors leads to significantly less capacitive feed through and is clearly recommended for low noise applications (drivers for analog switches that are not allowed to inject RF noise via the gate capacity of the switch etc.)

Normally starved current inverters are operated with a defined load capacity. The following figure shows the same dual starved current inverter operating with a 100fF load.

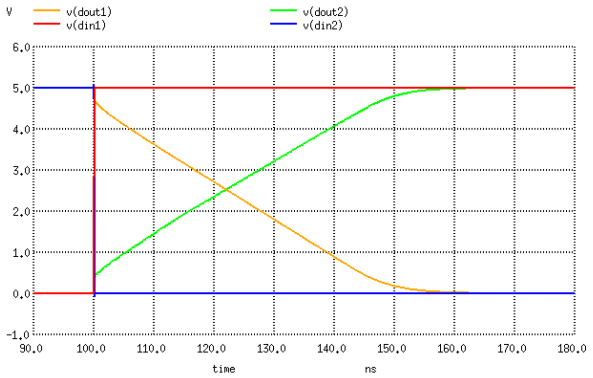


Figure 7.235: dual starved current inverter operating with a defined capacitive load of 0.1pF

7.15.3 NAND gates

In many technologies NAND gates are the preferred topology because they offer the highest speed (per current or per switching charge loss).

The more inputs a NAND gate has the slower it gets because the capacities to be charged increase and the driving impedances increase. So the number of inputs usually is limited to 4. Nevertheless there are exceptions to this rule accepting either a current consumption penalty or a speed penalty (example: 7430 is an 8 input NAND gate).

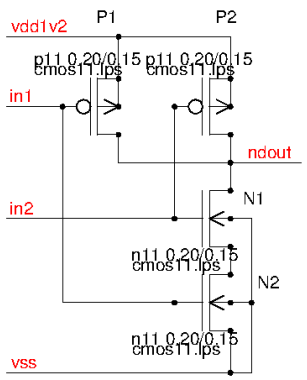


Figure 7.236: 2 input NAND gate in CMOS technology

Table 34: truth table of a NAND gate

in1	in2	ndout
0	0	1
0	1	1
1	0	1
1	1	0

7.15.4 NOR gates

Fast NOR gates require making the PMOS transistors significantly larger than the NMOS transistors. This leads to a a significantly higher silicon real estate than building NAND gates. For this reason logic design prefers NAND gates. NOR gate should be avoided in speed critical applications.

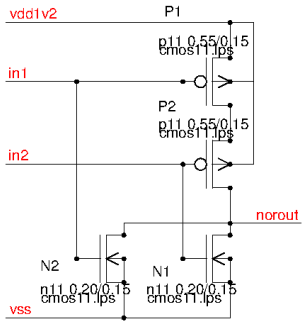


Figure 7.237: 2 input NOR gate in CMOS technology

Table 35: truth table of a NOR gate

in1	in2	norout
0	0	1
0	1	0
1	0	0
1	1	0

7.15.5 AND gates

In CMOS technology an AND gate consists of a NAND gate and an inverter.

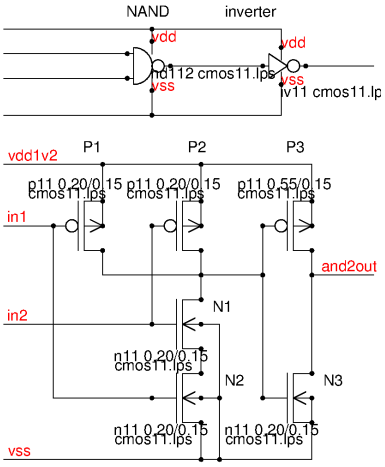


Figure 7.238: 2 input AND gate

Table 36: truth table of an AND gate

in1	in2	and2out
0	0	0
0	1	0
1	0	0
1	1	1

7.15.6 OR gates

Or gates in CMOS technology consist of a NOR gate and an inverter.

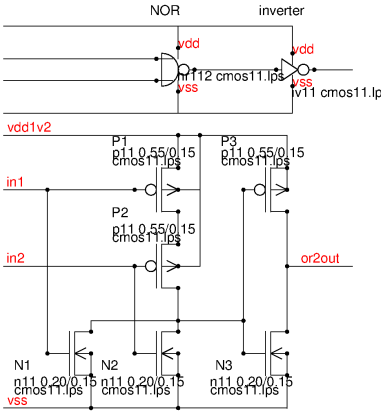


Figure 7.239: 2 input OR gate

Table 37: truth table of an OR gate

in1	in2	or2out
0	0	0
0	1	1
1	0	1
1	1	1

7.15.7 EXOR gates

EXOR gates either can be composed of NAND gates and inverter gates or they can be designed on transistor level. using transistor level is more efficient than composing them from basic gates.

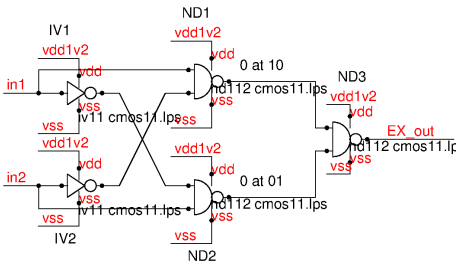


Figure 7.240: EXOR based on standard gates

The gate level design requires 16 transistors. The silicon real estate can be reduced to 12 transistors by the following circuit:

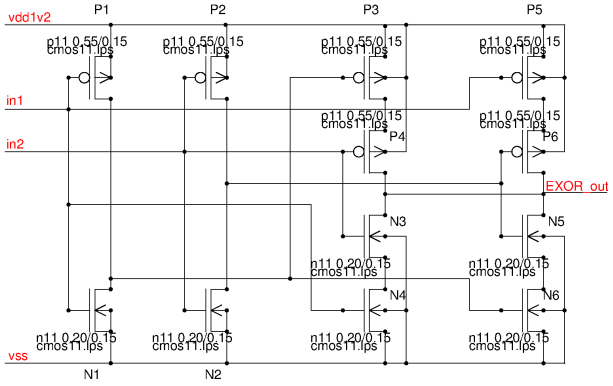


Figure 7.241: Transistor level based EXOR gate

The transistor level gate has the following behavior:
in1=0 and in2=0 turns on N5 and N6. EXOR_out becomes 0.
in1=0 and in2=1 turns on P5 and P6. EXOR_out becomes 1.
in1=1 and in2=0 turns on P3 and P4. EXOR_out becomes 1.
in1=1 and in2=1 turns on N3 and N4. EXOR_out becomes 0.

Table 38: truth table of an EXOR gate

in1	in2	EXOR_out
0	0	0
0	1	1
1	0	1
1	1	0

7.15.8 Multiplexers

Multiplexers are used to switch between two inputs. Either they are constructed using standard gates or area optimized transistor level designs are used.

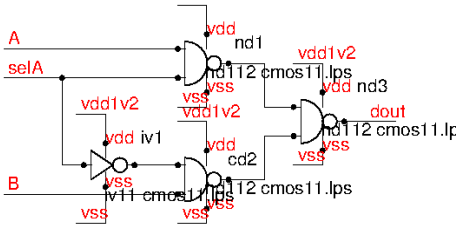


Figure 7.242: Multiplexer composed of standard gates

Looking at silicon real estate the multiplexer shown above requires 14 transistors. Area optimized designs look more like the EXOR shown above.

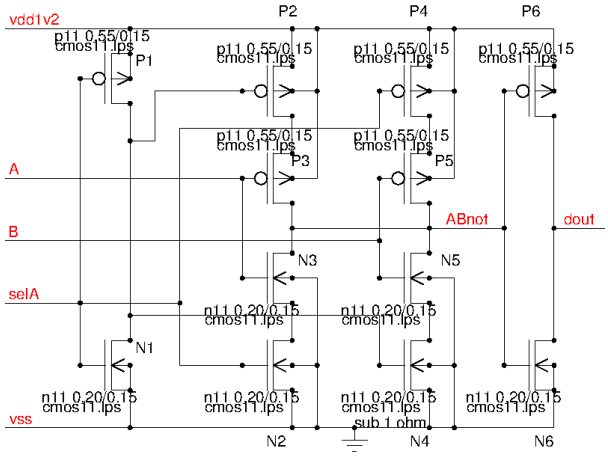


Figure 7.243: Multiplexer transistor level design

The transistor level design on the first glance seems to only save 2 transistors. The trick is the regularity of the structure. The following transistors can be merged using one active with two gates: p2 and P3, P4 and P5, N2 and N3, N4 and N5. These merged transistors are almost as small as a single transistor. So this design roughly consumes the area of an 8 transistor circuit.

This still doesn't look too interesting, but multiplexers are often used as shift elements in arithmetic logic units and counters. Since there you need one multiplexer per bit to be shifted an arithmetic logic unit easily requires some hundred multiplexers. Then the area saving multiplies up!

7.15.9 Latches

Latches are bistable circuits. This means the circuit can take one out of (usually) two possible states. Most latches consist of two inverting amplifiers. The most common way to implement a latch is using two logic inverters. To switch the latch a low resistive drive signal has to be connected for a short time. This low resistive signal forces the latch into one state. (provided the source connected is stronger than the holding capability of the two inverters.

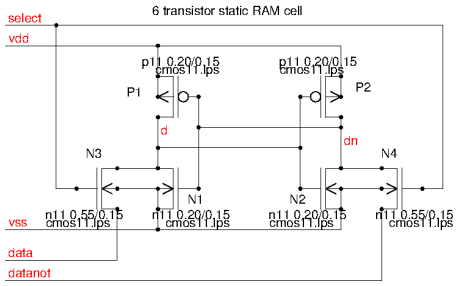


Figure 7.244: 6 transistor static RAM cell

The cell can be written if the signals data and datanot are driven by a low resistive driver and signal select turns on N3 and N4. Since N3 and N4 are stronger thanP1 and P2 the data will overwrite the nodes d and dn. When signal select goes to 0 the data remains stored in d and dn. To read out the latch the signals data and datanot are connected to a high resistive read amplifier and select goes to 1 again.

This is the typical kind of a latch used to store data in static RAMs.

If the signal select doesn't exist latches can be built using logic gates. The cheapest way to build latches with logic gates is to use either NAND gates or NOR gates.

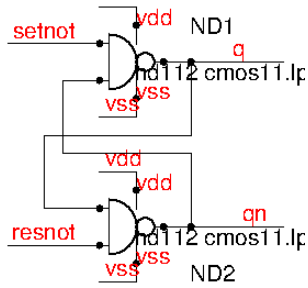


Figure 7.245: NAND latch

A logic 0 at input set will make q go to logic 1 and qn to logic 0. When set returns to logic 1 the data will remain stored (provided setnot is at logic 1). Vice versa if setnot goes to logic 0 the latch will be reset.

There is one forbidden code of the latch: setnot and resnot may not be logic 0 at the same time! If this forbidden code takes place both outputs q and qn will become logic 1 at the same time. The status that gets stored at the end depends on which signal (setnot or resnot) returns to logic 1 later. The later one will win. So it is very wise to design the circuits driving the latch such that this forbidden code never is reached.

Implementing a latch with NAND gates instead of NOR gates offers a higher speed due to the higher mobility of the electrons and the resulting smaller transistors sizes required.

The latch consisting of NAND (or NOR) gates already requires 8 transistors compared to the RAM cell shown before.

If the forbidden code b00 at the input can't be excluded the latch should be modified to make one of the two inputs dominant. This leads to a NAND latch with dominant input.

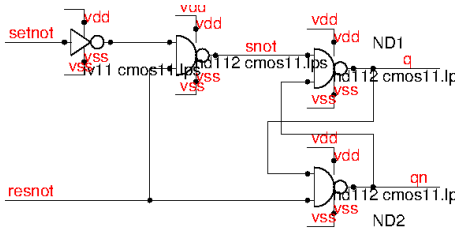


Figure 7.246: NAND latch with dominant resnot

In some applications it can't be predicted how long the resnot is logic 0. If the latch is supposed to get set at the rising edges of the two input signal (instead of the state) the need to build a 2 stage circuit. The first latch determines whether there was a rising edge or not and the second stage saves the data. This leads to the edge triggered latch.

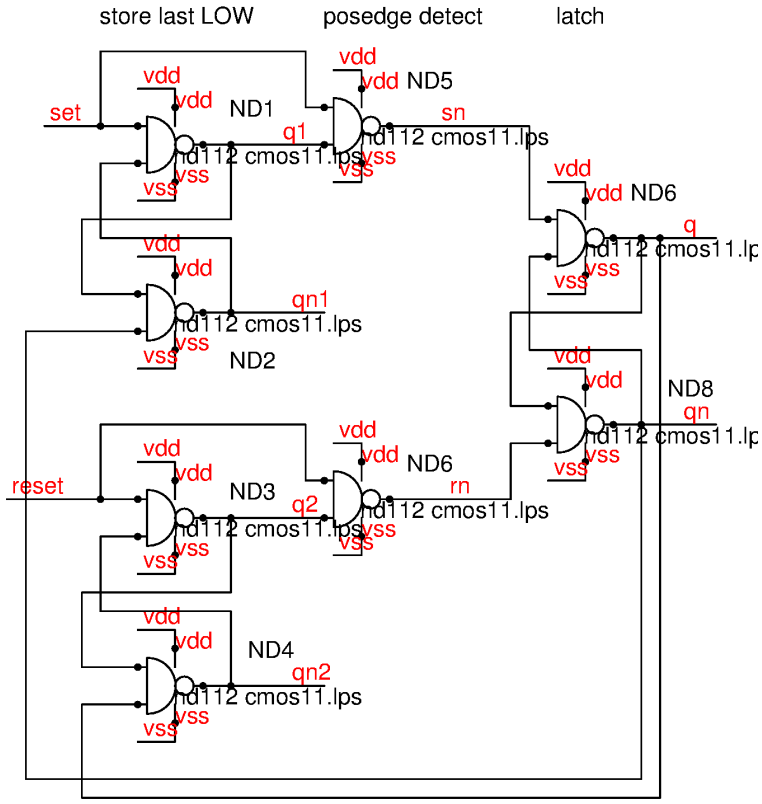


Figure 7.247: edge triggered latch

The edge triggered latch consists of 3 latches. The two input latches (ND1 to ND4) get set at the logic 0 of their inputs. ND5 and ND6 prevent the propagation to the 3rd latch. At the rising edge of the input set or reset the signal can propagate through ND5 or ND6 (provided there was a 0 before that is stored in the firs two latches.) At the rising edge of set the 3rd latch switches. As soon as the 3rd latch has switched the input latch is reset again to prepare the detection of the next rising edge.

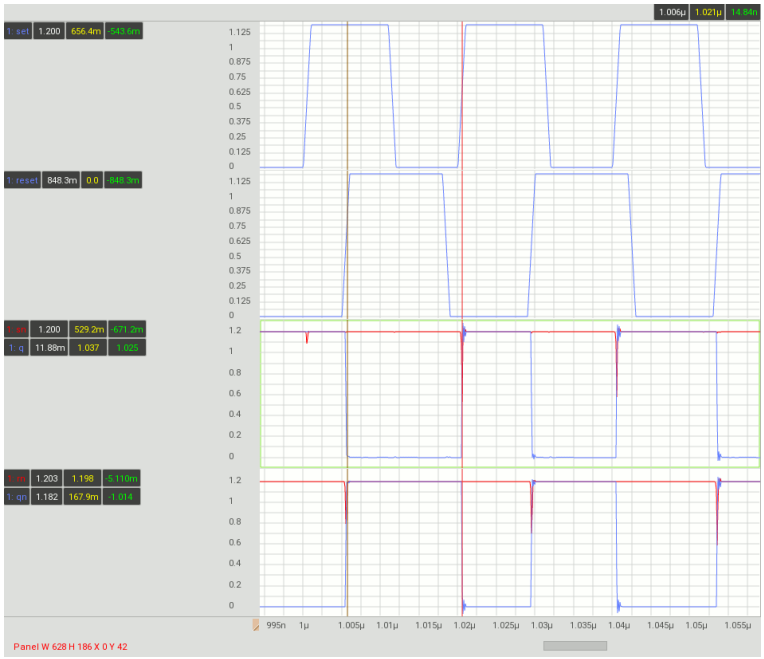


Figure 7.248: simulation of the edge triggered latch

The duration of the logic 0 states of signals sn and rn depends on the propagation delay of the 3rd latch and the reset time of the edge detection input latches.

Theoretically it would be sufficient to simply produce a short pulse by simply feeding back from sn into the reset of the set edge detection and from rn into the reset edge detection. But this more simple approach has a race. If the 3rd latch (ND6, ND8) is slower than the input latches the circuit wouldn't trigger correctly. To prevent this kind of race the more save approach is to reset the edge detection from the 3rd latch instead of doing it in a shorter loop.

Very often it is preferred to have an edge triggered input set but a state triggered dominant reset. This requires a slight modification of the circuit. Now only the set path has an edge detection latch. At reset this latch is cleared together with the output latch. The first latch stores the next logic zero as soon as the reset becomes logic 0 and set is logic zero.

If reset changes to 0 but set is already logic one nothing will happen because the first latch hasn't seen a zero since the reset.

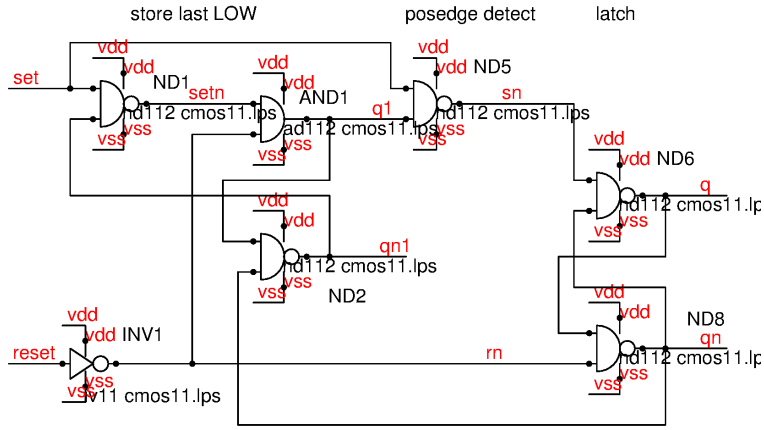


Figure 7.249: half edge triggered latch

The following figure shows the simulation. There only is a response if the rising edge of set takes place while reset is logic 0.

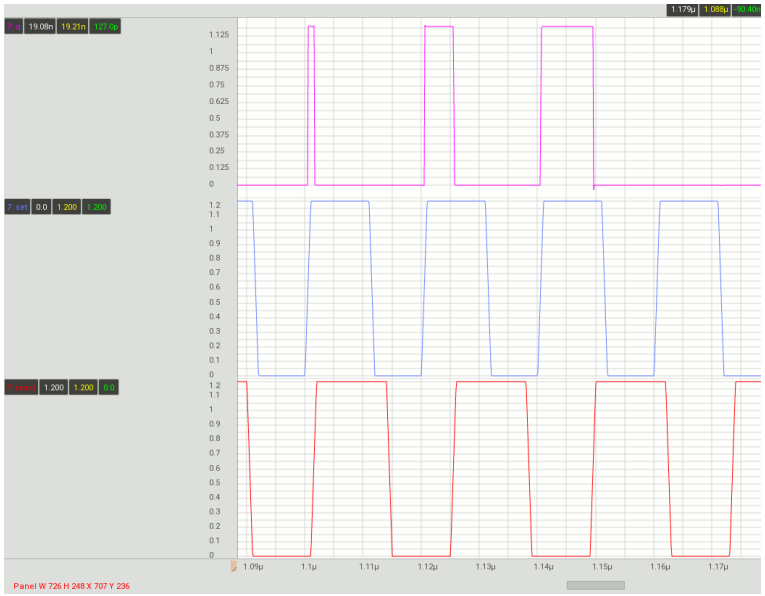


Figure 7.250: the half edge triggered latch only responds to rising edges of set while reset=0

Using latches: Latches are able to store events without having a clock signal. This violates the philosophy of classical logic synthesis (There a stored state may only change at a clock edge, but not at any other edge!). Building latches inside a synchronous logic therefore normally is forbidden. To minimize tool conflicts only use latches for the following purposes:

1. Memories (there transistor count is more important than clean synthesis).
2. If you have to capture events without having a clock.
3. Schmitt triggers often use latches for hysteresis switching.
4. Level shift circuits (Latches permit level shifts that have no static current consumption)

Building latches always means hand crafted asynchronous logic. Do it with special care! Transition from hand crafted latch based logic to synchronous logic always needs a synchronizer between the two design styles.

7.15.10 data flip flop (DFF)

A data flip flop (DFF) has 2 major inputs. The data input (d) and the clock input (clk). To define the initial state most data flip flops additionally have a reset (res) or a reset not pin (res_n). The data flip flop consists of two latches. The second latch reads the output of the first latch when the first latch goes into hold mode. This way it behaves like a data storage element that stores data exactly at the edge of the clock. This behavior can directly be seen looking at the most generic verilog code describing a data flip flop.

One possible transistor level representation is shown in the following figure:

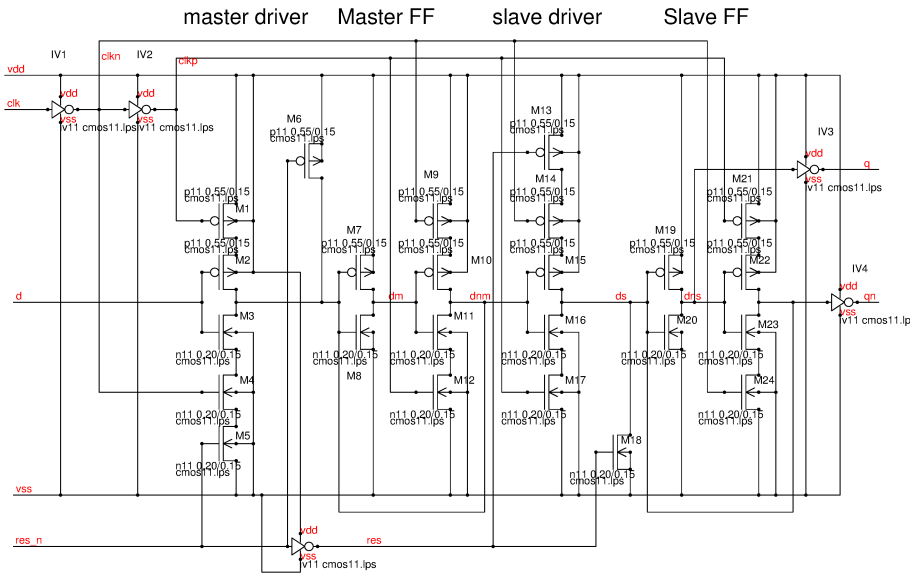


Figure 7.251: Transistor level of a data flip flop

The circuit shown here uses more transistors than if the classical transmission gate design style is used. This implementation nevertheless is very effective because many of the transistors share a common active area. As a consequence the layout can be done very compact in spite of the higher number of components.

7.15.11 flip flops for very high speed dividers

7.15.12 Counters

Counters consist of flip flops dividing the frequency. The most simple counter is a binary up counter.

Assynchronous counters: Assynchronous counters are very traditional designs used in hand crafted logic. The output of one flip flop toggles the next flip flop. The longer the counter the longer the delay until the last (most significant bit) toggles. The following figure shows an assynchronous counter.

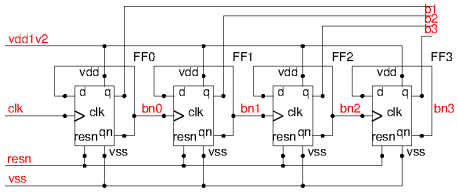


Figure 7.252: Assynchronous 4 bit counter

The plot shows the counter counting up. At t=170ns the counter reaches overflow at decimal 15 (b1111) and returns to 0.

Figure 7.253: counting from 0 to 15. At the 16th clock pulse the counter overflows

Zooming into the simulation shows the delay from one stage to the next. This delay best is visible when the counter overflows. Ideally at overflow all bits should change simultaneously. But in an assynchronous counter we observe the propagation delays of all flip flops summing up.

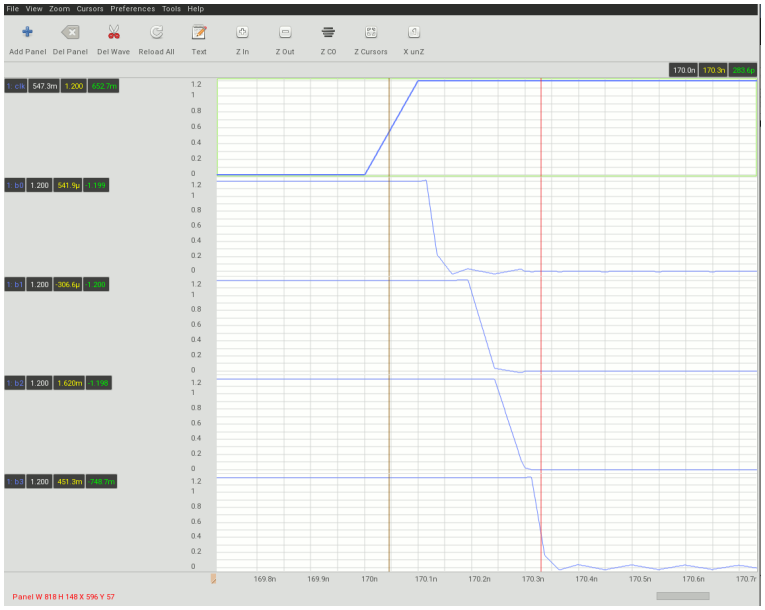


Figure 7.254: Zooming into the overflow the delays of the flip flops can be seen

The summing up of delays can create glitches if the counter is followed by a combinatoric logic (consisting of gates only). This can't be accepted for logic synthesis. Assynchronous counter are not appreciated in automated logic design because timings depend on the length of the counters.

An advantage of assynchronous logic is the lower current consumption. Only the firs flip flop (FF0) is operated at the full clock speed. So 50% of the current consumption are at FF0. The following FFs only contribute as a geometrical sum:

$$I_{total} = I_{FF0} + \frac{1}{2} * I_{FF0} + \frac{1}{4} * I_{FF0} + \frac{1}{8} * I_{FF0} .... = 2 * (1 - \frac{1}{2^{n-1}}) * I_{FF0} \tag{7.376}$$

with n being the length of the counter (in the example n=4).

If an overflow is not desired the clock has to be stopped reaching the final count. Here comes an example how to do this.

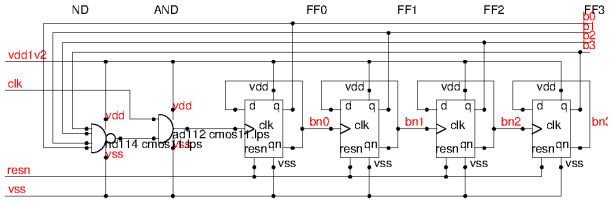


Figure 7.255: Counter without overflow

The NAND gate ND detects when the counter reaches b1111 and gate AND blocks the clock before an overflow takes place. This works well as long as the propagation time of all the flip flops together is shorter than the clock period.

Synchronous counters: Synchronous counters are better suited for design synthesis. Building a synchronous counter it is easy to add the up/down counting capability. The following figure shows a 4 bit up/down counter with overflow and underflow prevention.

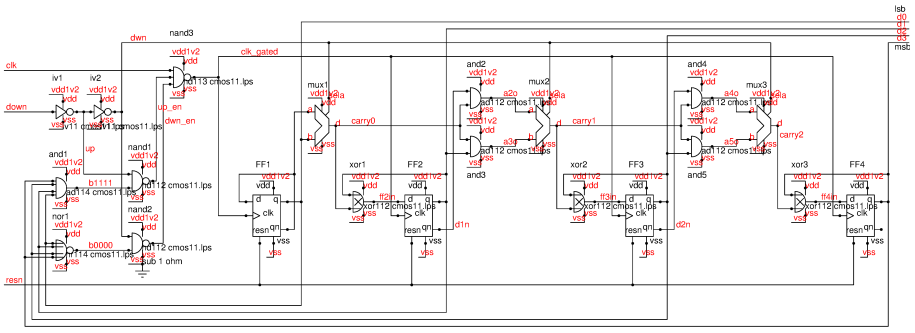


Figure 7.256: A 4 bit up/down counter

The following figure shows the simulation result of this counter.

Figure 7.257: Simulation result of the 4 bit up/down counter

7.15.13 Shift registers

The most simple shift register only shifts in one direction

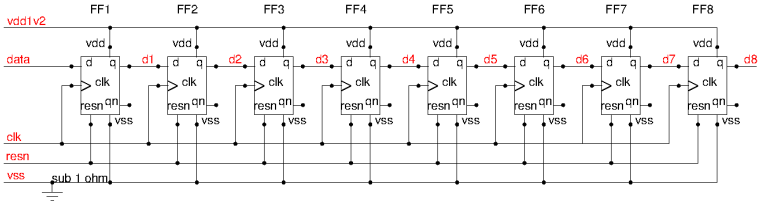


Figure 7.258: 8 bit shift register

To be able to change the shift direction multiplexers need to be added. This leads to the up/down shift register.

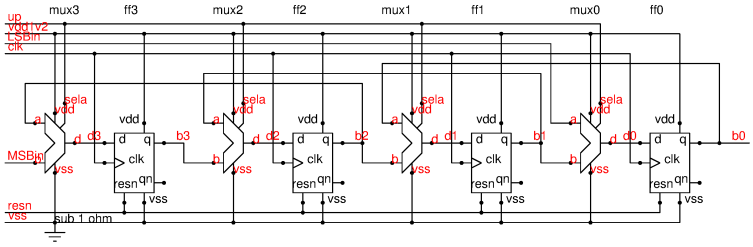


Figure 7.259: 4 bit up down shift register

In the following simulation the shift register is first filled with ones from the left side. Then the ones are shifted back, Later it is filled with ones and zeros from the right side.

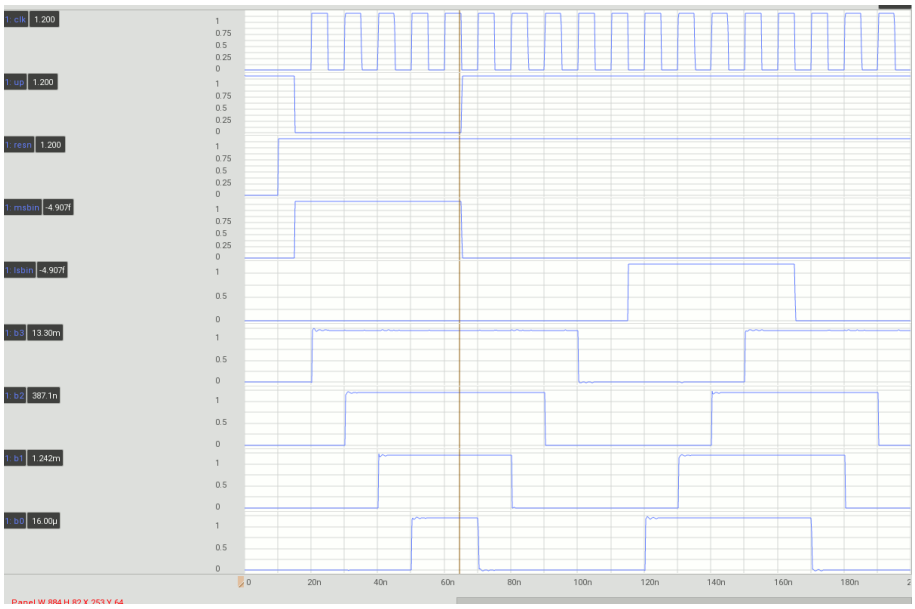


Figure 7.260: 4 bit shift register filling from the left and then from the right side

7.15.14 Level shift circuits

There are many different ways of building level shift circuits. Which kind of level shift is used depends on the application.

Shift down level shift Shifting down from a high supply level (for instance 3.3V) to a low supply level (for instance 1.2V) usually is done building a 3.3V inverter but supplying it with the low voltage rail. So the input can handle 3.3V while the output is in the 1.2V domain. Take care that the first inverter doesn't have an antenna diode between in3v and vdd1v2.

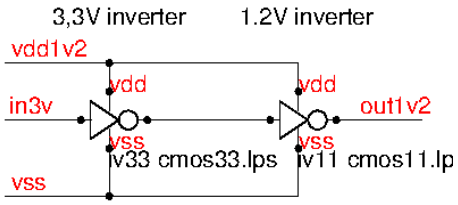


Figure 7.261: shift down from 3.3V to 1.2V

Shift up level shift without latch function Shifting from a low supply domain to a high supply domain typically requires a PMOS half-latch and NMOS switches. This kind of level shift is requires the least area. The disadvantage is that it becomes undefined if the supply on the input driver side gets lost.

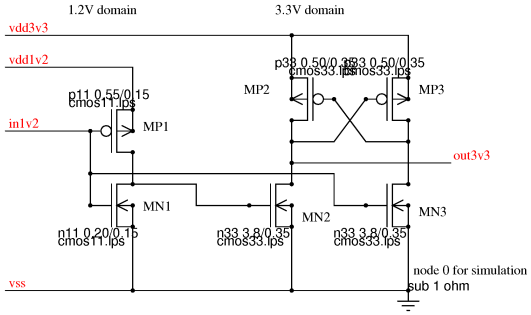


Figure 7.262: shift up level shift without latch function

Note that the NMOS transistors MN2 and MN3 are significantly stronger than the PMOS transistors MP2 and MP3. There are two reasons for it:

- MN2 and MN3 are driven with a lower gate voltage (coming from the 1.2V domain) than the PMOS transistors MP2 and MP3 (driven from the 3.3V domain)
- MN2 and MN3 must be stronger than MP2 and MP3 under all circumstances. Worst case usually are low 1.2V supply, high 3.3V supply, simulation corner sf (slow NMOS, fast PMOS)

If the 1.2V supply vdd1v2 is lost the gates of MN2 and MN3 are not driven. The previously on PMOS will remain conducting as long as the gate does not float up. If the gate of the conducting PMOS floats up (for instance due to leakage) the transistor turns off an the outputs of the level shift become undefined.

Shift up level shift with latch function The level shift is based on the shift up level shift without latch but has two additional hold transistors. These two hold transistors are added to maintain the last valid state even if the low voltage driver looses its supply. This way floating gates at the output of the level shift are prevented.

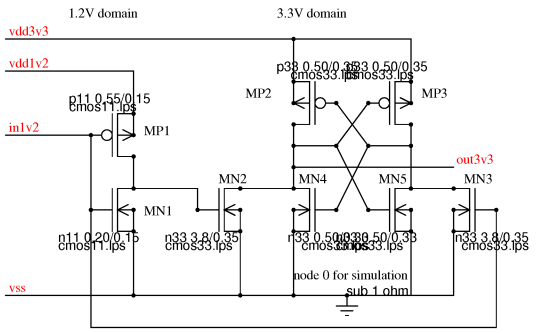


Figure 7.263: Level shift with latch function

If the supply of the 1.2V domain gets lost the transistors MP2, MP3, MN4, MN5 store the last state of the level shift while MN2 and MN3 are not driven.

Variants of the standard level shift: The following circuits show variants of the standard level shift with slightly different figures of merit.

High voltage level shift: The high voltage level shift is needed to transfer signals from one supply level to an other supply level that is further away than the break down voltage of a gate oxide. Typically high voltage level shifts are built with hold latches.

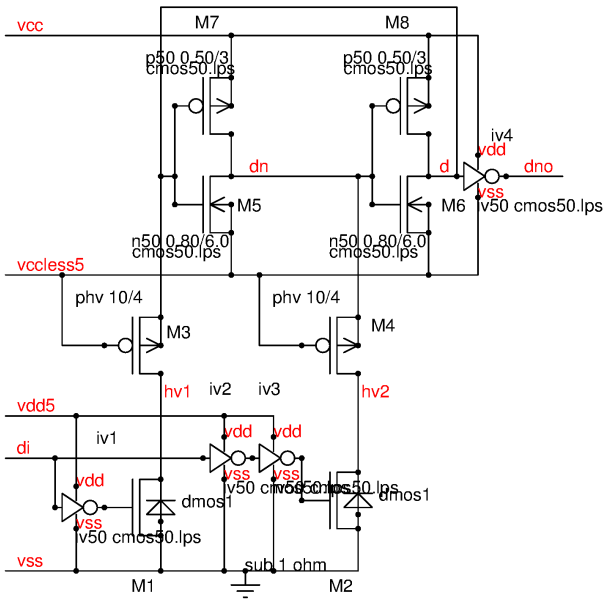


Figure 7.264: Standard high voltage level shift

In this level shift the components M1, M2, M3 and M4 are high voltage transistors that must sustain the voltage between net vcc and vss. M1 and M2 drive the long channel latch M5, M6, M7, M8. To protect the gate oxides of M5..M8 transistors M3 and M4 limit the voltages swing of nets d and dn. Transistors M1..M4 must be stronger than M7 and M8. Therefore M7 and N8 are long channel transistors. The threshold of inverters M5, M7 ad M6, M8 must be higher than the Vgs of M3 and M4. Therefore the bulks of M3 and M4 usually can not be connected to vcc (back ate would increase the threshold.

Since the pull down transistors are stronger than M7 and M8 the falling edges at nets d and dn can be designed fast. The rising edges usually are slow because of the limited strength of M7 and M8 and because of the high capacity of nets hv1 and hv2. The delay of the level shift differs for rising edges and for falling edges.

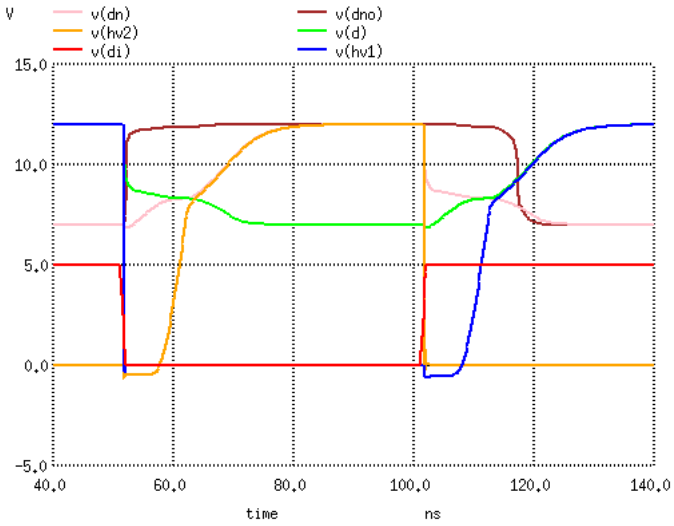


Figure 7.265: Simulation of the standard high voltage level shift

The simulation shows clearly the slow rise of the signals of the nets hv1 and hv2 and the long delay of the level shift from the rising edge of di to the falling edge of dno.

Fast high voltage level shift: To increase the speed of the levelshift circuit on both edges one possible solution is to always select the fast path provided by the falling edge of the signal at the high voltage transistors.

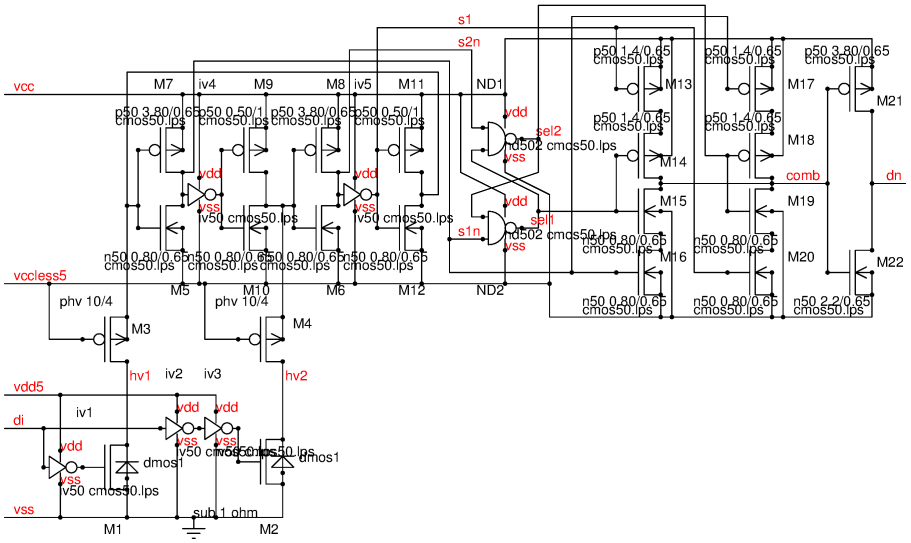


Figure 7.266: fast high voltage level shift with select latch selecting which path controls the data output dn

The input stage is exactly the same as the input stage used in the standard high voltage level shift circuit before. The receiver latch now consists of 3 inverters (M5, M6, M7, M8, M9, M10, M11,M12 and iv4, iv5). Since we want to have a fast falling edge at nodes hv1 and hv2 but we are not much interested in the rising edge the PMOS transistors M9 and M11 intentionally are designed weak. The read inverters should respond right at the beginning of the falling edges. Therefore the thresholds of inverters M7, M5 and M8, M6 are chosen higher than usually done (strong PMOS, weak nmos). The two more stages needed are standard cells (iv4, iv5).

The select latch (Nd1, ND2) responds delayed versus the data input signal because it is driven inverted (s1n and s2n are the inverted signals hv1 and hv2). Before the latch has flipped the fast path driven by the falling edges of hv1 and hv2 controls the combiner node comb.

Since the levelshift always responds to the falling edge it is sensitive to fast rising supply voltage VCC. If the supply voltage VCC moves up faster than the long channel PMOS transistors can charge the drain capacity of M1 and M2 the levelshift responds with a change of the output state. After a while the drain of the high voltage transistor that is off moves up again and the levelshift falls back into it's correct state. Typically such glitches are in the range of 10ns to 30ns.

Design risks of level shift circuits: Level shift circuits always depend on a strength hierarchy of transistors. More or less a level shift always consists of an inverter latch on the receiving side and a driving HV transistor that must be able to override the latch. The holding strength of the latch depends on the supply voltage of the latch, the aspect ratios and the threshold voltages of the transistors belonging to the latch. The strength of the driving HV transistor depends on the gate overdrive of the input side and the aspect ratio of the HV transistor. These parameters never match because the transistors involved are completely different. The full process spread plus the mismatch both impact functionality and parameters of the level shift. Monte Carlo simulation including variation of the supplies and temperature are a MUST designing level shift circuits. If a level shift fails it can get stuck at HIGH as well as stuck at LOW. Automatic error checking of the simulation must be able to detect both states. Since there only is pass or fail but no meaningful distribution (So you can't determine statistical parameters such as sigma s) the Monte Carlo simulation requires many hundred runs per corner to prove reliability of the level shift!

8 System building blocks

System building blocks consist of several, in some cases quite complex, analog functions. These may be combined with configuration logic and features such as trimming or auto zero. Settling time of bias current generators, bandgaps, auto zero functions etc. are long in comparison with the response time of the logic. For this reasons system building blocks should at least have one power enable signal and one output enable signal.

The power enable signal turns on the bias generators, references and whatever else is required to operate the block. The response time to the power enable signal is typically in the range of some micro seconds.

During power up there may be glitches taking place until the biasing is settled. To prevent propagation of these glitches to the logic most building blocks have an additional signal enabling the output.

To decouple the digital noise from the analog functions most analog functions have an analog supply rail (VDDA). In the same way most analog functions have a separate ground (VSSA) to prevent noise coupling via the ground resistance. High voltage chips may have even more supply rails.

In addition access to internal signals may be required for testing. Typically an analog test bus and a digital test bus is required.

Ideally the analog test bus should be 4 wires wide to allow differential measurements simultaneously with differential stimulation.

Comparators converting analog signals into digital signals require access to the comparator outputs using a digital test bus.

Table 39: Typical signals found in almost all mixed signal designs

signal	purpose		
enable_bias	turns on the bias		
enable_out	turns on the output when bias is settled		
VDDA	analog supply		
VSSA	analog ground		
atb<n>	analog test bus. n=4 if possible		
dtb<m>	digital test bus for comparators		

8.1 Amplifier applications

almost all analog circuits are based on amplifiers with or without feedback signal. For this reason it makes sense to first have a look at basic applications of amplifiers before going to the details of certain building blocks.

8.1.1 Amplifier requirements

There are certain classes of applications for amplifiers. Depending on the application the design targets can vary significantly. The solutions to reach those targets have changed from the time of tubes to modern power MOS FET transistors. Some basic concepts however didn't change.

Table 40: Amplifier topologies and usage

application	most important parameter	amplifier type
RF receiver	low noise, high bandwidth	open loop
DC regulation	low offset	closed loop
audio preamp.	low noise, low distortion	closed loop
audio power	low distortion, high efficiency	class AB to D
Servo amplifier	efficiency, adjustable compensation	class B, C
RF power	high efficiency, bandwidth	load is resonant
unity gain buffer	low output impedance, bandwidth	usually open loop

8.1.2 Open loop operation

Open loop operation means the amplifier is working without feedback.

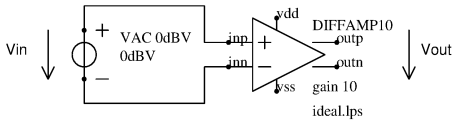


Figure 8.1: Open loop operation of an amplifier

This kind of operation of course only makes sense if the gain of the amplifier is well defined by design and if the gain is stable over temperature.

$$V_{out} = V_{in} * gain_{amp} \tag{8.1}$$

If the input signal has a swing such that the output signal would go beyond the supply rails the amplifier simply starts clipping the signal. This becomes especially important if the amplifier has a very high gain (for instance 1000) and the input signal has an amplitude of several 10mV.

If the input signal exceeds some mV a simple differential stage becomes too non linear. In this case resistors can be used to linearize the amplifier (at a loss of gm and noise performance).

The most important advantage of open loop operation is that the amplifier can be designed for a high bandwidth without stability problems because there is no feedback that suffers from the phase shift of the amplifier. Therefore open loop amplifiers are preferred for RF applications.

Amplifiers intentionally designed for open loop operation often can't be used in closed loop configurations.

8.1.3 Closed loop operation

In closed loop operation of an amplifier a part of the output signal is fed back to the input. To keep things simple in the following we use an operational amplifier with single ended output.

Usually operational amplifiers have a very high gain (at least for low frequencies).

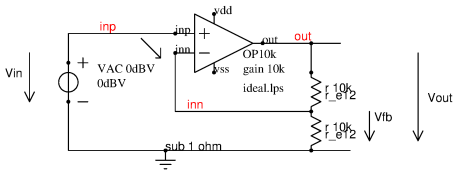


Figure 8.2: Amplifier in closed loop operation

We are interested in the output signal Vout. Of course Vout is the difference between V(inp) and V(inn) times the gain of the amplifier.

$$V_{diff} = V(inp) - V(inn) \tag{8.2}$$

$$V_{out} = gain_{amp} * V_{diff} \tag{8.3}$$

The voltage of node inn depends on the resistor divider and V_{out}

$$gain_{fb} = \frac{R_2}{R_1 + R_2} = \frac{1}{2}$$

$$V(inn) = V_{out} * gain_{fb}$$

$$V_{out} = gain_{amp} * (V_{in} - V_{out} * gain_{fb})$$

$$V_{out} = V_{in} * \frac{gain_{amp}}{1 + gain_{amp} * gain_{fb}} \tag{8.4}$$

Example: $gain_{amp} = 10000$, $gain_{fb} = 1/2$ lead to $V_{out}/V_{in} = 10000/(1 + 5000) = 1.9996$

The ratio

$$gain_{closed} = \frac{V_{out}}{V_{in}} \tag{8.5}$$

is called the closed loop gain. As long as the product $gain_{amp} * gain_{fb} >> 1$ the closed loop gain is close to $1/gain_{fb}$. An amplifier with unlimited gain would reach exactly this closed loop gain. The difference between the output signal of a (theoretical) amplifier with unlimited gain and the real amplifier (with limited gain) is the error signal.

$$V_{error} = V_{out_{ideal}} - V_{out}$$

$$V_{error} = V_{in} * \frac{1}{gain_{fb} * (1 + gain_{amp} * gain_{fb})} \tag{8.6}$$

Often we are more interested in the relative error $V_{error}/V_{out_{ideal}}$

$$Err_{rel} = \frac{1}{1 + gain_{amp} * gain_{fb}} \tag{8.7}$$

Example: $gain_{amp} = 10000$, $gain_{fb} = 1/2$ lead to $Err_{rel} = 1/(1 + 10000 * 0.5) = 1.9996 * 10^{-4}$

8.1.4 Noise and offset propagation in closed loop operation

To understand the noise and offset propagation let's add the error voltages to the closed loop circuit. To make things easier let's assume we have an ideal amplifier with an infinite gain. Thus the errors are multiplied according to the resistor ratios only.

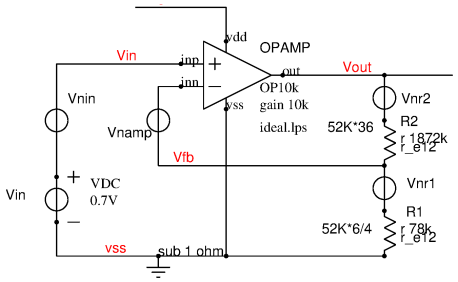


Figure 8.3: Noise propagation in a closed loop design

To get a rough overview let us assume the sources Vnr1 is the resistive noise of R1, Vnr2 is the resistive noise of R2, Vnamp is the noise of the input transistors of the amplifier and Vnin is the noise of the input signal. At room temperature the resitive noise is about:

$$V_{nrx} = \frac{4nV}{\sqrt{Hz * R/K\Omega}}$$

Let's further assume the bandwidth of interest is 10kHz. This way in the example the two resistive noise sources calculate as:

$$V_{nr1} = 3.533\mu V$$
$$V_{nr2} = 17.31\mu V$$

The noise of R2 will be amplified with a gain defined by the resistors R1 and R2. Since R1 connects to ground this is the signal gain. Assuming the amplifier has a unlimited gain the signal gain becomes:

$$gain_{signal} = \frac{R_1 + R_2}{R_1} = 25$$

The noise of R2 is in the feedback path. It won't get amplified.

The input noise of the amplifier V_{namp}will will be compensated producing the negative image of V_{namp} at node fb. This means the amplifier noise will be amplified with the signal gain. The noise or R1 will be amplified inverted with a gain of $gain_{signal} + 1$.

The noise of the input source will also be amplified by the closed loop gain. As long as all sources of noise are uncorrelated (noise adds as a power) this leads to an output noise of:

$$V_{nout} = \sqrt{V_{nr2}^2 + gain_{signal}^2 * V_{nin}^2 + (gain_{signal} + 1)^2 * (V_{nr1}^2 + V_{namp}^2)} \tag{8.8}$$

In our example even neglecting the amplifier noise and the noise of the source we get about $V_{nout} = 90.05\mu V$ (for 10kHz bandwidth). The noise of R1 clearly dominates because it will be amplified by factor 26 while the noise of R2 will not be amplified.

The ratio $R_2/R_1 = gain_{signal} + 1$ is called the noise gain because this is the gain most of the noise gets amplified with.

In practical applications the signal Vin often is already noisy due to the noise contributions of other blocks providing the signal!

8.1.5 AC characteristics of an amplifier with one pole with feedback

The amplifier as well as the feedback network usually have a frequency dependent transfer function. In stead of describing the gain and the feedback factor in real numbers a complex description can be used. In this description s is a standardized frequency.

$$s = \frac{j * \omega}{\omega_{ref}} \tag{8.9}$$

As a simple example let's assume the amplifier has a DC gain of $gain = 10000$ and a roll off of the gain of -20dB/decade starting at $\omega_{ref} = 100$. This leads to a transfer function of the amplifier of

$$g_{amp} = gain * \frac{1}{1 + s} \tag{8.10}$$

The closed loop gain assuming a feedback network that is constant over frequency having a feedback gain

$$gain_{fb} = \frac{1}{k} \tag{8.11}$$

Inserting the frequency dependent gain of the amplifier into equation (8.4) yields

$$\frac{V_{out}}{V_{in}} = \frac{gain * \frac{1}{1+s}}{1 + \frac{1}{k} * gain * \frac{1}{1+s}} = \frac{gain}{1 + s + gain/k} \tag{8.12}$$

Not yet looking familiar? Let's replace $K_{real} = 1 + gain/k$and $gain = k * (1 - K_{real})$ rewrite it

$$gain_{closedloop} = \frac{V_{out}}{V_{in}} = k * (K_{real} - 1) * \frac{1}{K_{real} + s} \tag{8.13}$$

Now it should be clear: It is a low pass! The constant $K_{real}$is the real part. s is the imaginary part $s \sim j\omega$. Equation (8.13) shows the low pass more clearly but equation (8.12) better shows what happens below the cut off frequency.

Since $K_{real}$is a real number and s is the imaginary normalized frequency the amplitude transfer function can be written as:

$$|gain_{closedloop}| = k * \frac{K_{real} - 1}{\sqrt{K_{real}^2 + \omega^2/\omega_{ref}^2}} \tag{8.14}$$

The factor K_{real}can be rewritten

$$K_{real} = \frac{k + gain}{k} \tag{8.15}$$

In typical applications the open loop gain of an OPAMP is in the range of $10^3$to 10^5while the closed loop gain k usually is kept in a range of 1..100. This justifies the approximation

$$K_{real} \approx \frac{gain}{k}$$

This is the ratio of the available open loop gain and the gain used in the application. Equation (8.13) can be approximated with a more handy expressions

$$gain_{closedloop} \approx k * \frac{K_{real}}{K_{real} + s} \tag{8.16}$$

$$|gain_{closeloop}| \approx k * \frac{K_{real}}{\sqrt{K_{real}^2 + \omega^2/\omega_{ref}^2}} \tag{8.17}$$

This means the amplifier in closed loop works with a gain of k defined by the feedback network. It starts to act as a low pass filter when s becomes dominant. The cutoff frequency is defined by

$$\frac{gain}{k} = \frac{\omega_g}{\omega_{ref}}$$

$$\omega_g = \omega_{ref} * \frac{gain}{k} \tag{8.18}$$

Looking at our example amplifier with a DC gain of 10000 the cut off frequency in closed loop operation is expected to be at

$$\omega_g = 100Hz * \frac{10000}{101} \approx 10^4 Hz$$

To demonstrate how close the exact calculation (8.14) and the approximation (8.17) are lets calculate the amplitude transfer function of the following circuit using both approaches:

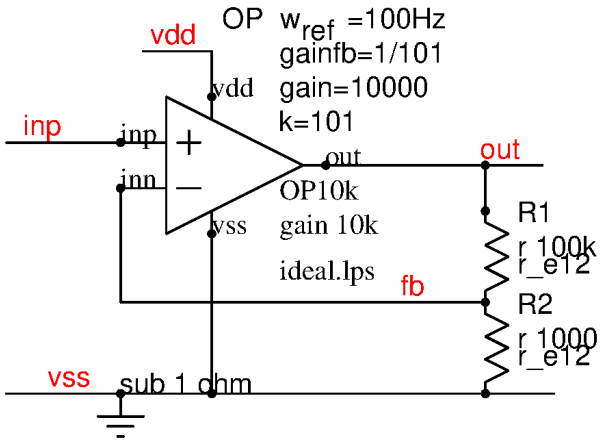


Figure 8.4: OPAMP in example application

The following plot shows the results doing the exact calculation (using equation 8.14) in red color and the approximated calculation (using equation 8.17) in blue color. The difference mainly is visible for low frequencies. The exact calculation yields a closed loop gain of 100 because the open loop gain only is factor 100 higher than the closed loop gain. The approximation yields a closed loop gain of 101 because it neglects subtracting the 1 from the ratio of the open loop gain and the closed loop gain.

The higher the ratio between open loop gain of the amplifier and the closed loop gain k adjusted by the feedback network the lower the differences between exact calculation and approximation will be.

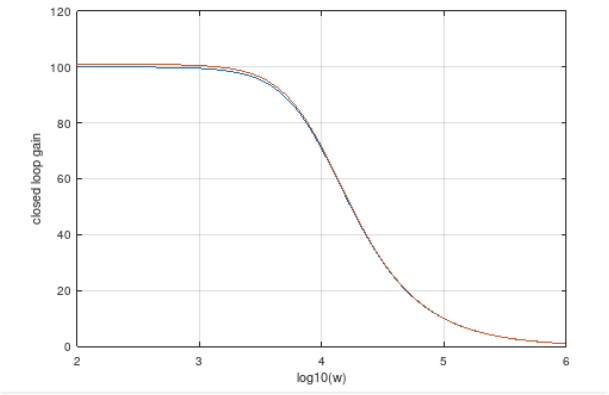


Figure 8.5: comparison of exact calculation and approximation

Using a logarithmic scale (in dB) for the closed loop gain the errors between the exact calculation and the approximation can barely be noticed!

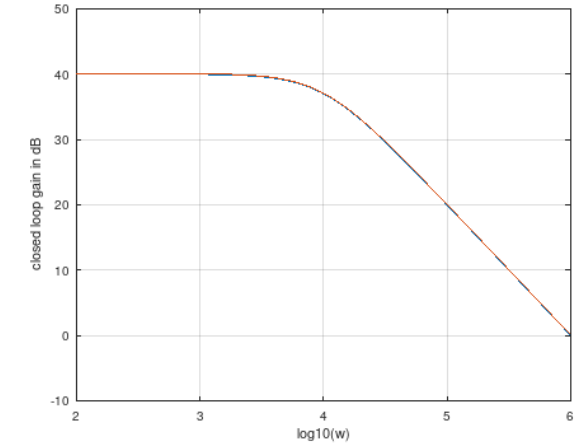


Figure 8.6: comparison of exact calculation and approximation scaling the gain in dB

Changing $k = 1/gain_{fb}$from 1 to 10, 100, 1000, 10000 at an amplifier $gain = 10000$ leads to the following plot:

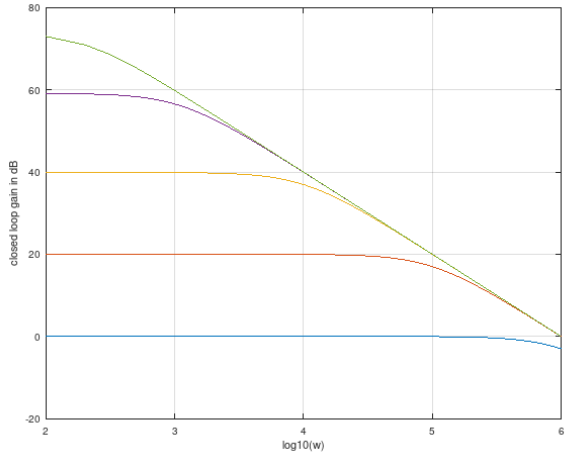


Figure 8.7: amplifier bandwidth using different gains

The lower the gain (k) the higher the bandwidth. At k=1 the bandwidth becomes 1MHz. At a gain of 10 (20dB) the bandwidth drops to 100kHz. At a gain of 100 (40dB) the bandwidth drops to 10kHz. At a gain of 1000 (60dB) the bandwidth drops to 1kHz.

Going up to k>10000 the gain saturates because the open loop gain falls below the intended closed loop gain. The product of gain and bandwidth is more or less constant. This is called the gain-bandwidth-product (GBW) of an amplifier.

8.1.6 Amplifiers with two poles

To increase the gain bandwidth products multiple stages are required. This can be done stacking stages with individual feedback. Each stage consists of a single pole amplifier and has its individual feedback network. To achieve a large bandwidth each stage is operated at a low closed loop gain.

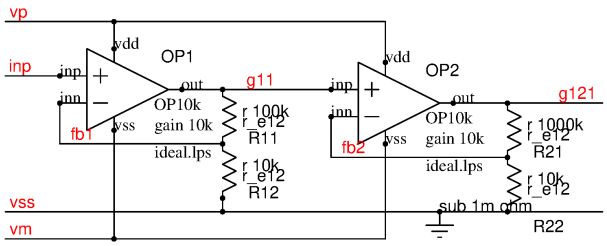


Figure 8.8: 2 stage amplifier with 40dB, 100kHz bandwidth

Each amplifier is using it's own feedback network. Up to about 100kHz the gain is flat (11*11=121). Above the cut off frequency the gain rolls off with -40dB/decade instead of -20dB/decade using a single amplifier with one pole. Ath the cut off frequency the phase shift (of both stages together) is already 90°. For $\omega >> \omega_{ref} * K_{real}$the phase shift (of both stages together) approaches 180°.

An alternative idea to increase the gain bandwidth product is to use multiple amplifier stages with one common feedback. The following figure shows the concept.

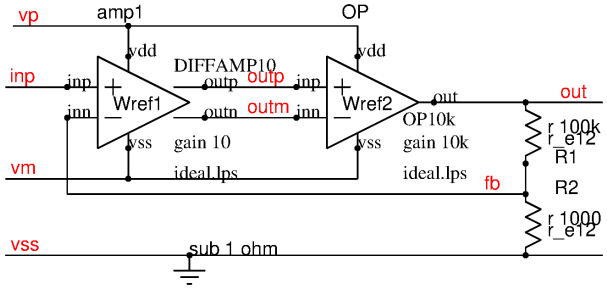


Figure 8.9: 2 stage amplifier with one feedback network

The basic idea is that now the open loop DC gain becomes

$$gain = gain1 * gain2$$

The AC open loop gain becomes

$$g_{amp} = gain * \frac{1}{(1+s_1)*(1+s_2)} \tag{8.19}$$

with $s_1 = j * \omega/\omega_{ref1}$ and $s_2 = j * \omega/\omega_{ref2}$.

Using this new expression to calculate the closed loop gain of the two stage amplifier application we get:

$$\frac{V_{out}}{V_{in}} = \frac{gain * \frac{1}{(1+s_1)*(1+s_2)}}{1 + \frac{1}{k} * gain * \frac{1}{(1+s_1)*(1+s_2)}}$$

Let's beautify it:

$$\frac{V_{out}}{V_{in}} = \frac{gain}{(1+s_1)*(1+s_2) + \frac{gain}{k}} \tag{8.20}$$

Here comes trouble! s_1 and s_2 are imaginary numbers! So there is a chance the denominator becomes 0. The frequencies at which the denominator becomes 0 are called the poles of the circuit.

In equation (8.21) $s_1$and $s_2$are normalized to different reference frequencies. This can be corrected referencing both to the same reference frequency. Assuming

$$\frac{\omega_{ref2}}{\omega_{ref1}} = m \tag{8.21}$$

and using $\omega_{ref1} = \omega_{ref}$, $\omega_{ref2} = m * \omega_{ref}$ for everything we can rewrite (8.21) using $s_2 = \omega/(m * \omega_{ref}) = s/m$and $s_1 = s$

$$\frac{V_{out}}{V_{in}} = \frac{gain}{(1+s)*(1+\frac{s}{m}) + \frac{gain}{k}} = \frac{m * gain}{(1+s)*(m+s) + \frac{m*gain}{k}} \tag{8.22}$$

From this formula we can derive a condition for stability: If the denominator becomes 0 the output signal becomes infinite - or in other words the system is instable. (Well, in real life the signal never reaches infinite. The real system starts clipping as soon as the output signal reaches the supply rails.)

Instability at s=0 means the system simply runs into the clipping state and stays there. This for instance happens in a schmitt trigger circuit. (In a schmitt trigger k is negative or the feedback goes to the positive input in stead of using the negative input)

Instability at a certain frequency means the system starts to oscillate. The condition for instability is:

$$(1+s)*(m+s)+\frac{m*gain}{k}=0$$

$$s^2+(m+1)*s+m*(1+\frac{gain}{k})=0 \tag{8.23}$$

Since s is imaginary ($s=j\omega/\omega_{ref}$) the real part will reach 0 at

$$|s_p|=\sqrt{m*(1+\frac{gain}{k})} \tag{8.24}$$

but with only two poles the imaginary part still exists:

$$denom(s_p)=j*(m+1)*\sqrt{m*(1+\frac{gain}{k})}$$

So even when the real part reaches 0 the closed loop gain remains at

$$\frac{V_{out}}{V_{in}}(s_p)=-j*\frac{m*gain}{(m+1)*\sqrt{m*(1+\frac{gain}{k})}}$$

Things get more interresting comparing this resonant gain with the DC gain. This is the gain peaking at the cut off frequency.

$$peaking=\frac{gain_{cl}(s_p)}{gain_{clDC}}=\frac{m*gain*(1+\frac{gain}{k})}{gain*(m+1)*\sqrt{m*(1+\frac{gain}{k})}}$$

$$peaking=\frac{\sqrt{m*(1+\frac{gain}{k})}}{(m+1)} \tag{8.25}$$

What can we see from this equation?

1. the further the two poles of the amplifier are appart (the higher ratio m) the lower the peaking gets.
2. The higher the closed loop gain (compared to the open loop gain) the lower the peaking gets.

The worst thing you can build is a unity gain buffer (k becomes 1) consisting of two identical amplifier stages (so m becomes 1). This desaster design has a peaking of:

$$peaking_{worst}=\frac{\sqrt{1+gain}}{2} \tag{8.26}$$

But real life is even worse! There always are further poles. The system does not only peak. Due to the higher order poles it WILL in most cases oscillate.

The phase shift of the closed loop amplifier simply calculates as

$$\varphi=arctan(Im(V_{out}/V_{in})/Re(V_{out}/V_{in})) \tag{8.27}$$

In the following plots the normalized frequency ω/ω_{ref} is sweeped.

Example 1: 2 stage amplifier with gain=10 and two identical poles. (m=1, gain=10, k=1). Peaking of about 5dB is unacceptable for most applications

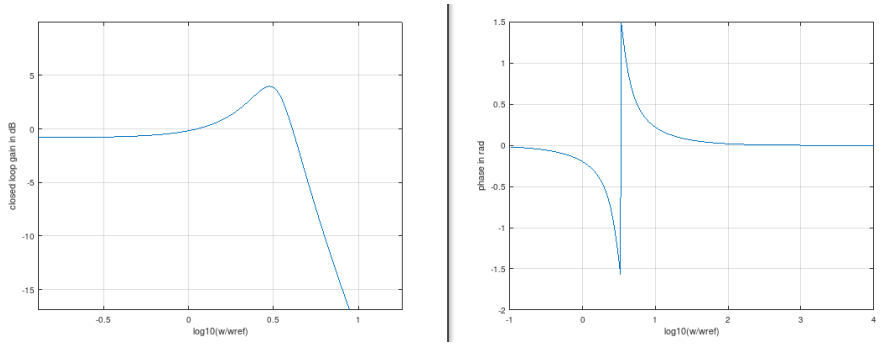


Figure 8.10: m=1, gain=10, k=1 unity gain buffer

Example 2: 2 stage amplifier with gain=10 and two poles one decade appart. (m=10, gain=10, k=1). Peaking almost disapears (reduced to 1dB) when m reaches about the gain margin gain/k.

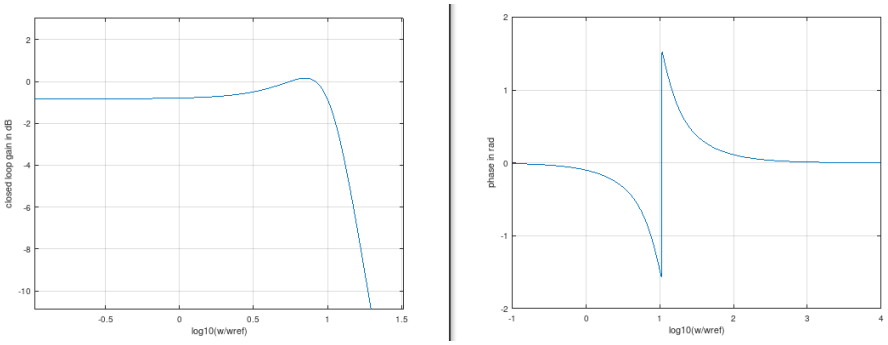


Figure 8.11: m=10, gain=10, k=1 unity gain buffer

Example 3: 2 stage amplifier with gain=10 and two poles double the gain margin appart. (m=20, gain=10, k=1). Now the peaking is gone. The gain characteristic is almost flat until about $5 * \omega_{ref}$. Chosing m=2*gain/k usually leads to very good performance of the amplifier in closed loop configuration.

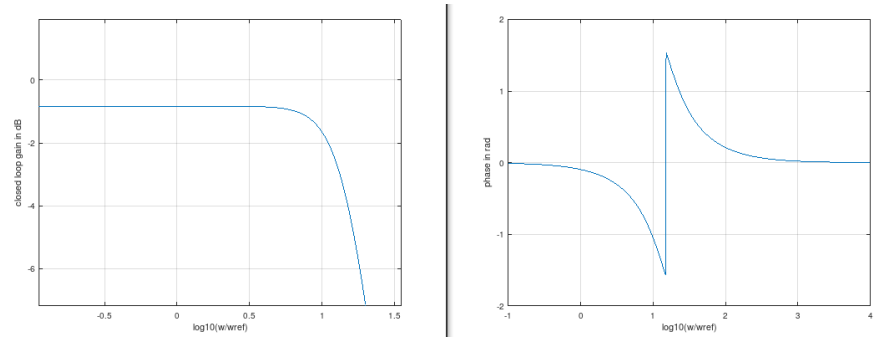


Figure 8.12: m=20, gain=10, k=1 unity gain buffer

Example 4: 2 stage amplifier with gain=10 and two poles 3 decades appart (M=1000, gain=10, k=1). Shifting the poles further appart than 2*gain/k doesn't lead to a significant improvement of the closed loop performance anymore. Shifting one of the poles to such a high frequency only costs current without leading to a significant improvement. Note that from about $gain * \omega_{ref}$ to $1000 * \omega_{ref}$ the voltage gain rolls off with -20dB/decade. Above the second pole the roll off becomes -40dB/decade.

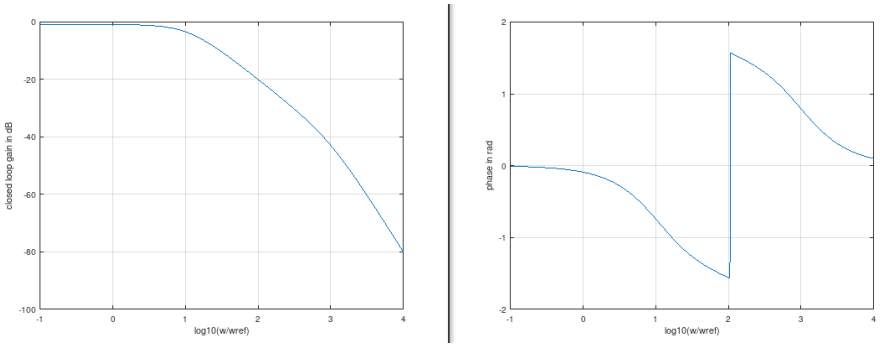


Figure 8.13: m=1000, gain=10, k=1 unity gain buffer

Example 5: Increasing the open loop gain improves the DC accuracy. This can be done if the two poles are far enough apart. This is the reason why most general purpose operationnal amplifiers are designed with a dominant pole in the singel Hz range and all higher order poles at m=2*gain. Using this scaling the general purpose amplifier can be used as a unity gain buffer with very good DC accuracy. The following example uses a closed loop gain k=1 while the open loop gain is 10000 (80dB) with two poles at $\omega_1 = \omega_{ref}$and $\omega_2 = 20000 * \omega_{ref}$. This is a typical pole distribution used on OPAMPs designed for unity gain stability.

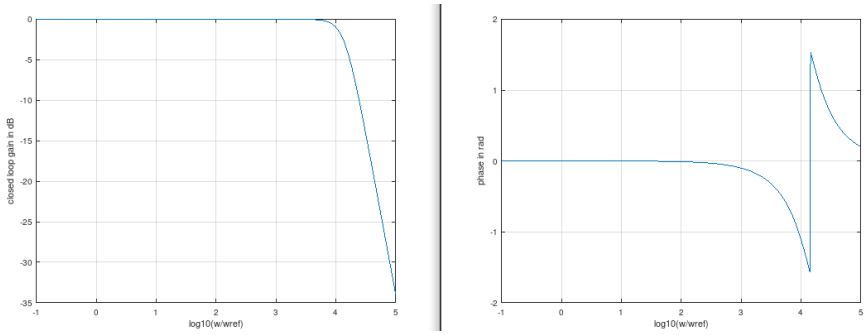


Figure 8.14: m=20000, gain=10000, k=1. An OPAMP used as a unity gain buffer

A typical example of such a design is the famous LM741. It has an open loop gain=50000..200000, a dominant pole at 2Hz and a second pole at 2MHz [84, 3-260]. The ratio between the frequency ratio m and the open loop gain is typically 5.

In stead of sweeping the frequency equation (8.24) can be solved directly for two poles.

$$ax^2 + bx + c = 0 \Rightarrow x_{1/2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

replacing $a = 1$, $b = m + 1$, $c = m * (1 + gain/k)$ leads to:

$$\omega_{1/2} = -\frac{m+1}{2} \pm \frac{\sqrt{m^2 + 2 * m + 1 - 4 * m * (1 + gain/k)}}{2}$$

$$\omega_{1/2} = -\frac{m+1}{2} \pm \frac{\sqrt{1 + m^2 - 2m - 4 * m * gain/k}}{2} \tag{8.28}$$

As soon as the expression $1 + m^2 - 2m - 4 * m * gain/k$ becomes negative $\omega_{1/2}$ describes an oscillation. The real part

$$Re(\omega_{1/2}) = -\frac{m+1}{2} \tag{8.29}$$

describes the damping of the oscillation. The higher the ratio gain/k gets the lower the damping and the more the amplifier will ring. The ringing frequency is described by the imaginary part.

$$Im(\omega_{1/2}) = \pm\frac{\sqrt{1 + m^2 - 2m - 4 * m * gain/k}}{2} \tag{8.30}$$

If the content of the square root becomes positive there will be no more ringing. The closed loop system will settle in an aperiodic way. The transistion from periodic to aperiodic behavior is met when the content of the square root $1 + m^2 - 2m - 4 * m * gain/k$crosses zero.

$$m_{1/2} = (1 + 2 * gain/k) \pm 2 * \sqrt{gain/k + (gain/k)^2}$$

Since this equation is only valid for $m \geq 1$ only one of the solutions is valid.

$$m = 1 + 2 * \frac{gain}{k} + 2 * \sqrt{\frac{gain}{k} * (1 + \frac{gain}{k})} \tag{8.31}$$

According to this (more pessimistic) equation for very high gains and k=1 the ratio between the frequencies of the dominant pole and the second pole should be higher than 5*gain. The designer of the LM741 obviously did an excellent job designing exactly a factor 5.

8.1.7 Amplifiers with low output impedance

In some cases amplifiers with very low output impedance over a wide frequency range are needed. Typical applications of such amplifiers are reference buffers for clocked loads such as ADC inputs. Most analog to digital converters draw a pulsed load current during signal sampling. The output of the amplifier may not drop significantly during this load current pulse.

The most common approach for building a low impedance output is using a source follower stage as already discussed in section 7.7.1.

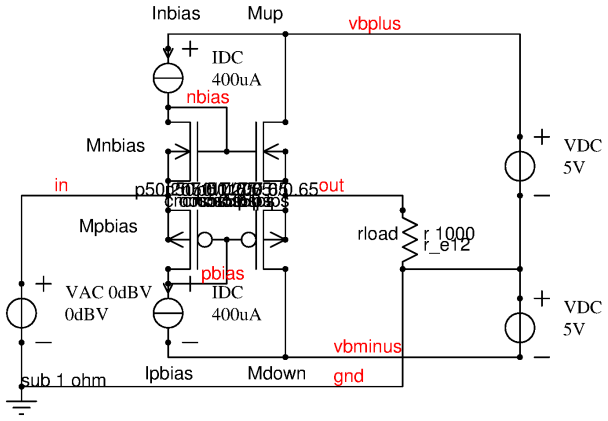


Figure 8.15: class AB push pull output stage

The small signal output impedance simply can be calculated using the transconductance of Mup and Mdown.

$$Z_{out} = \frac{1}{gm_{Mup} + gm_{Mdown}} \tag{8.32}$$

Ton reduce the output impedance the following measures are suggested:

1. Make the aspect ratio (W/L) of Mup and Mdown as big as possible
2. Increase the bias current

Not very elegant because the current consumption goes up dramatically!

A second option is to operate the output stage in a closed loop with an amplifier.

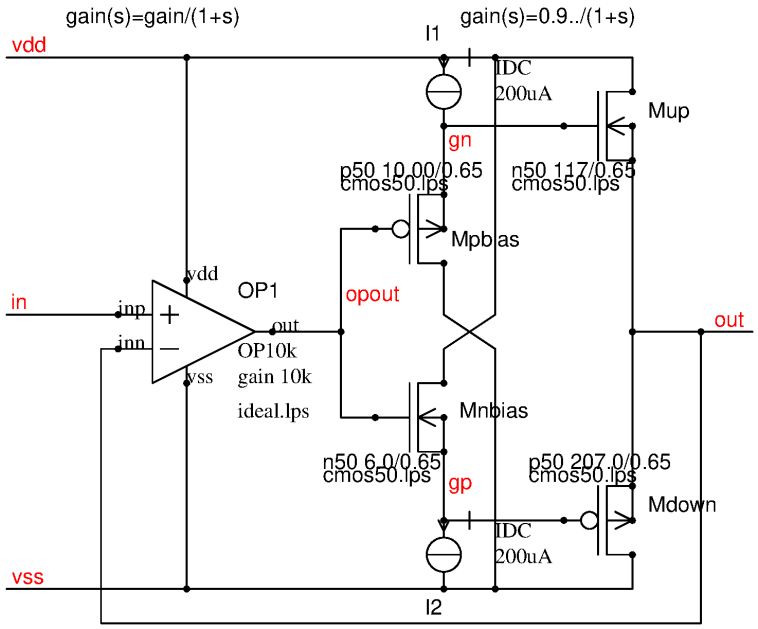


Figure 8.16: source follower in a closed loop

The output impedance of this closed loop configuration gets reduced by the gain margin of the amplifier. The stability of the design remains uncritical as long as the pole of the source follower stage is much higher than the pole of the operational amplifier. Neglecting the pole of the source follower stage $gain_f$ the output impedance becomes.

$$Z_{out} = \frac{1}{(gm_{Mup} + gm_{Mdown}) * gain_f * (1 + gain_f * gain(s))} \tag{8.33}$$

Usually the gain of the source follower is in the range of 0.7 to 0.95 over a wide frequency range. The reduction of the output impedance is proportional to

$$Z_{out} \approx Z_{openloop} * \frac{1}{gain(s)} \tag{8.34}$$

This approach works nicely for low frequencies. If the output impedance must be kept low up to several MHz the amplifier OP1 has to have a very high gain bandwidth product. Building an amplifier with high gain bandwidth product requires using small transistors. Using small transistors in the input stage of OP1 on the other hand leads to offset errors.

8.1.8 Regulation loops

Often amplifiers are used in regulation loops. The behavior of an amplifier in a regulation loop can be:

- proportional (P)
- integrating (I)
- differentiating (D)

In addition a more complex behavior can be constructed combining different kinds of regulators. In the most general case the behavior can be described as proportional-integrating-differentiating (PID).

Proportional (P): In a proportional stage the output follows the input signal with a constant gain. This is a classical OPAMP application with simple resistor only feedback. The following example shows a current regulator. The regulator is drawn in blue color.

Since the gain of the regulator is limited the current through M1 doesn't reach the target of 3A.

Ideally the transfer function in the frequency domain is constant.

$$gain(p) = K$$

In practical design the amplifier has a limited bandwidth. with $p = jf/f_g$ and $f_g$being the cutoff frequency of the amplifier (including the load capacity of the gate of the transistor it is driving!) the true frequency dependent gain of the regulator becomes:

$$gain(p) = K * \frac{1}{1 + p} \tag{8.35}$$

The regulation loop will have a (normally single) pole.

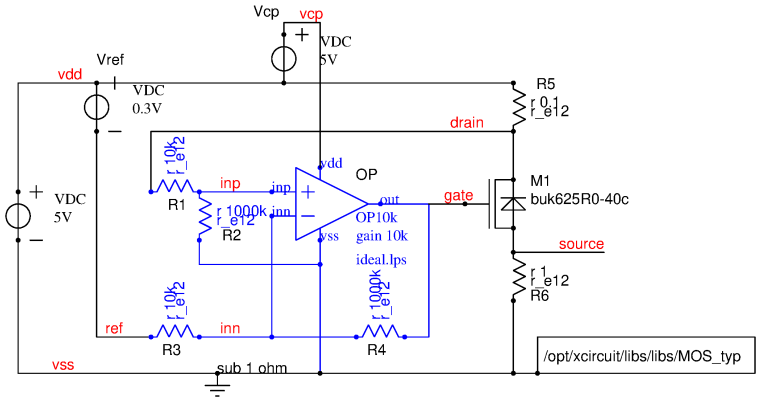


Figure 8.17: Example of a P-regulator

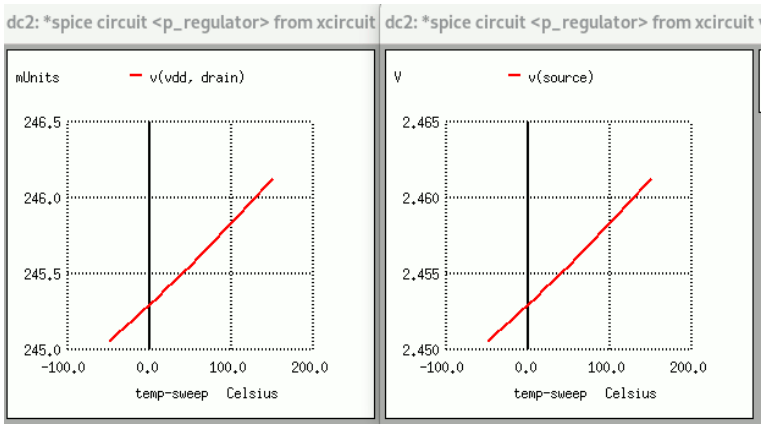


Figure 8.18: Simulation of the P-regulator with a gain of 100

integrating (I): An integrating (I) regulator integrates the deviation of the regulated parameter from the target. The more time is given to the regulator the more accurate the regulation result gets. Practical integrating regulators however are limited by the open loop gain of the amlifier. So even an integral regulator doesn't reach a perfect match between target and regulated parameter.

The DC performance can be seen in the following plot. The voltage drop over the resistor R5 is very close to the reference voltage of 0.3V and the output current is very close to the taget of 3A

The regulator's integrating behavior leads to the following frequency dependent gain:

$$gain(p) = \frac{1}{p} \tag{8.36}$$

In this equation $p = j * \frac{f}{f_1}$. $f_1$is the frequency at which the absolute gain of the regulator drops to 1. As long as the rest of the regulation loop has a high bandwidth an integrating regulator behaves reasonably stable.

Adding further poles to the loop quickly will lead to instability!

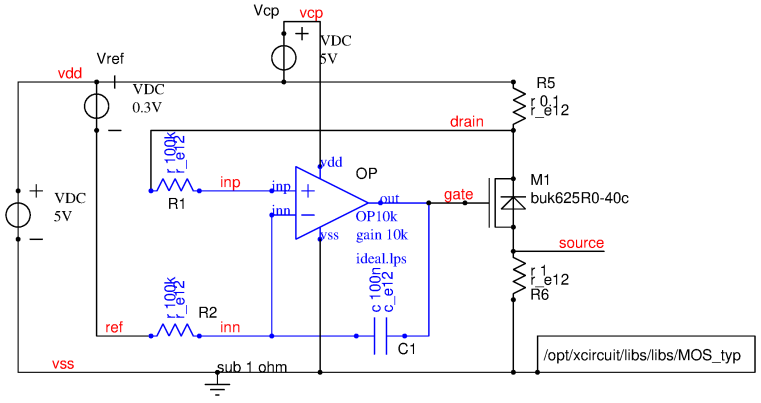


Figure 8.19: Example of an I-regulator

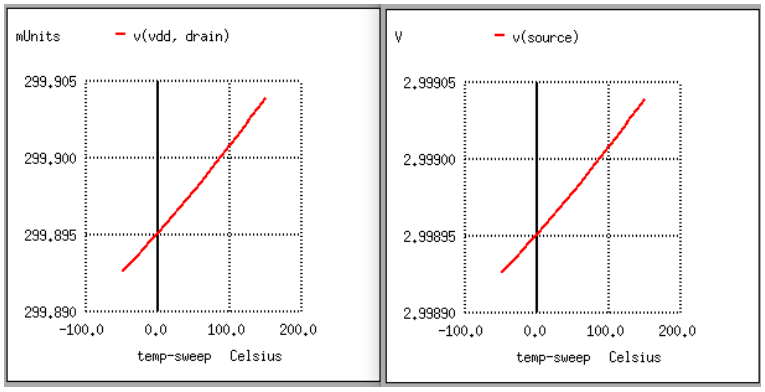


Figure 8.20: Simulation of the I-regulator

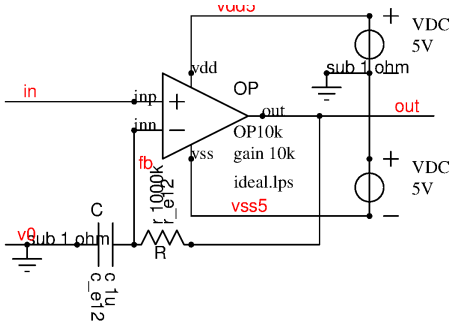


Figure 8.21: Example of a D-regulator

8.1.9 Differentiating (D)

Differeciating (D) regulators are somewhat uncommon. There are multiple reasons for this:

1. A close to ideal differenciating regulator requires an amplifier with extremely high bandwidth.
2. Differenciating regulators don't correct steady state deviations from the regulation target.

For this reason differenciating regulator stages usually only are found in combination with P and I regulators.

Ideally a differenciating amplifier has the following structure:

The AC transfer function is simple as long as the gain and bandwidth limitations of the amplifier are neglected (ideal amplifier).

$$gain(p) = -p \tag{8.37}$$

In this equation $p = j * \omega * R/C$. In time domain ist can be described as

$$V_{out}(t) = -R * C * \frac{dV_{in}(t)}{dt} \tag{8.38}$$

8.1.10 PID regulator

The PID regulator is the most general form of a regulator composed of a proportional path, an integrating path and a differentiating path. The signals of all three paths are summed by resistors (basically this is a voltage to current conversion). Since this addition reduces the signal level amplifier OPsum inverts and amplifiers the signal again. The main purpose of OPsum is to create a virtual ground at node sum. This is required to achieve a correct addition of the currents flowing through R5 to R7.

The transfer function of the PID regulator is:

$$gain(p) = K_p + K_i * \frac{1}{p} + K_d * p \tag{8.39}$$

8.2 Voltage regulators

Everything on a chip starts with the power supply. It MUST work or the chip will not work.

There are hundreds of possible solutions ranging from a simple regulator to a complex power management with switch mode power supply and charge pumps. In this chapter we proceed from simple to complex.

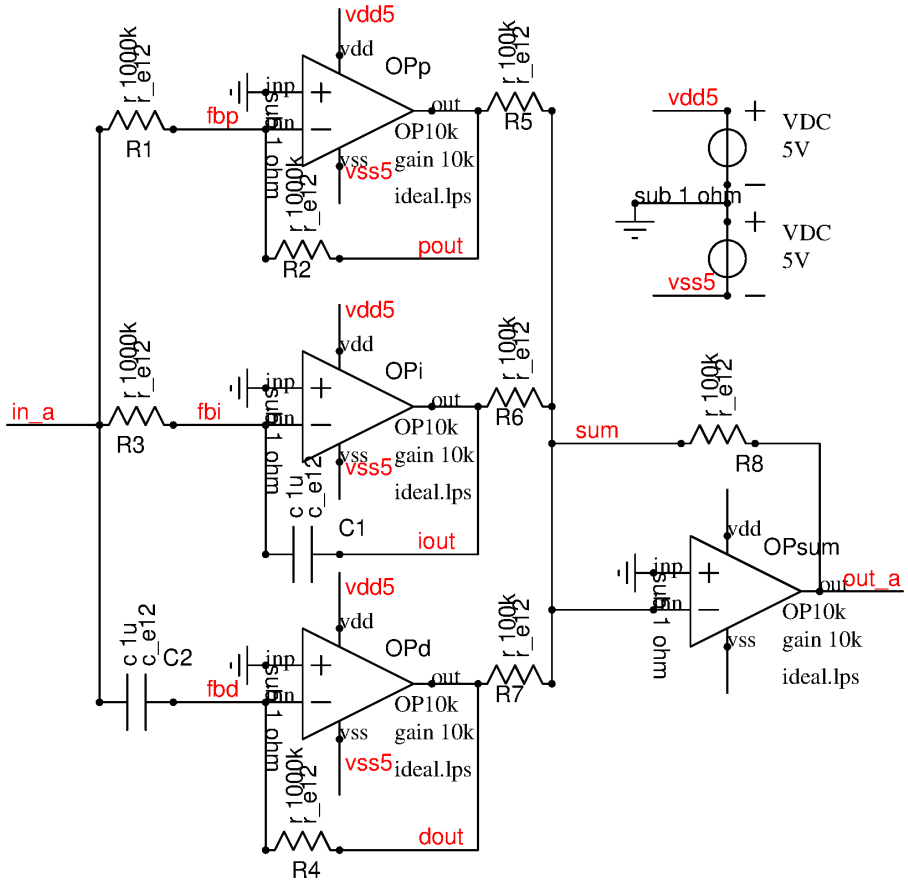


Figure 8.22: The most generic PID regulator topology

8.2.1 Unregulated prestabilizer

This is the most simple solution to provide a more or less stable supply. It consists of a voltage generator and a source follower or in case of a bipolar technology an emitter follower. The first question is where does the reference voltage come from. For a long time zener diodes have been a preferred choice. (See for example MC1723L in [29]). The first integrated implementations dating back to the 1960s looked like this (Well, this concept already was used with valves since the 1920s with stabilizer valves holding a gas that became conductive at a certain voltage. So this approach is much older than semiconductor electronics! See for instance page 5-19 of [30]):

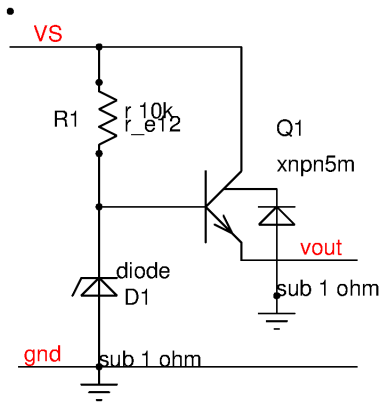


Figure 8.23: Prestabilizer using a zener diode and an emitter follower

This kind of prestabilizer is cheap and usually robust. But it has some important limitations:

1. The output voltage is a function of the break down voltage of the diode and the base emitter voltage of the transistor.

$$V_{out} = V_{br} - V_{be}(I) \tag{8.40}$$

Since V_{be} is a function of the current flowing the output voltage drops with increasing load current. Neglecting

the emitter resistance (ideal bipolar transistor) we will find:

$$V_{out} = V_{br} - V_f - \frac{k * T}{e} * ln(\frac{I_{out}}{I_0}) \tag{8.41}$$

2. The small signal output impedance (assuming a constant break down voltage of the diode) is:

$$R_{out} = \frac{k * T}{e * I_{out}} \tag{8.42}$$

3. Q1 can only deliver a correct output voltage as long as it gets enough base current trough R1. So we have the following limitation:

$$I_{out} < \frac{B * (V_s - V_{br})}{R_1}$$

4. The emitter of an NPN transistor is one of the smallest structures found in most technologies. This makes the emitter very ESD sensitive.

5. Break down of diodes always produces hot electrons. So the zener diode will age. Usually the break down voltage increases with operating time. Therefore usually zener break down is not long term stable. (buried zeners usually are a bit better than surface zeners but I still would not recommend it for high precision.)

6. A zener diode break down always involves avalanche effect. This can produce tremendous noise. Up to 1V peak to peak is not unusual! The voltage fluctuations have statistical timing and usually the edges have up to volts per ns! The higher the zener voltage the higher the noise voltage produced by the avalanche contribution gets.

7. Avalanching depends on the cleanliness of the process. The cleaner the process the worse it gets. So circuits working in a dirty old process (short free space for the avalanche to develop until it hits the next lattice defect) can easily fail if implemented in a clean modern process.

To solve at least the aging and the noise problems of zener diodes the zener diode can be replaced by a bandgap circuit. This leads to the following fairly simple circuit:

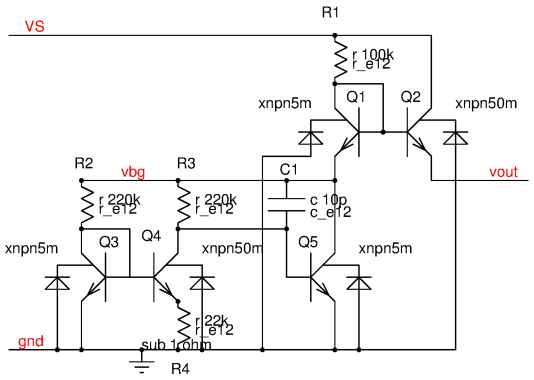


Figure 8.24: Prestabilizer using a bandgap in stead of a zener diode

This open loop regulator can be used well to understand some of the basic properties of voltage regulators (most of it applies to closed loop regulators too as soon as we are looking at frequency ranges beyond the cut of frequency of the regulator amplifier).

Most important standard voltage regulators and prestabilzers only can source current but not sink it. So if the load drops to zero the output voltage will increase. Of course the output voltage will not approach infinite as suggested by the equation above. Here the simplified model of a transistor being a current source becomes invalid. Nevertheless leakage of Q2 can lift the output up to VS.

There is a second limitation. If the output vout gets too high the base-emitter junction of Q2 will break down. This will limit the output voltage to about Vbg+7V (Most standard bipolar technologies have a base-emitter break down voltage of about 5V..10V). Operating Q2 in base emitter break down is fatal because the transistor will get damaged and B will decrease significantly within a few milliseconds. Therefore it is recommended to always have a certain minimum load connected to the emitter of Q2 to prevent it from floating up.

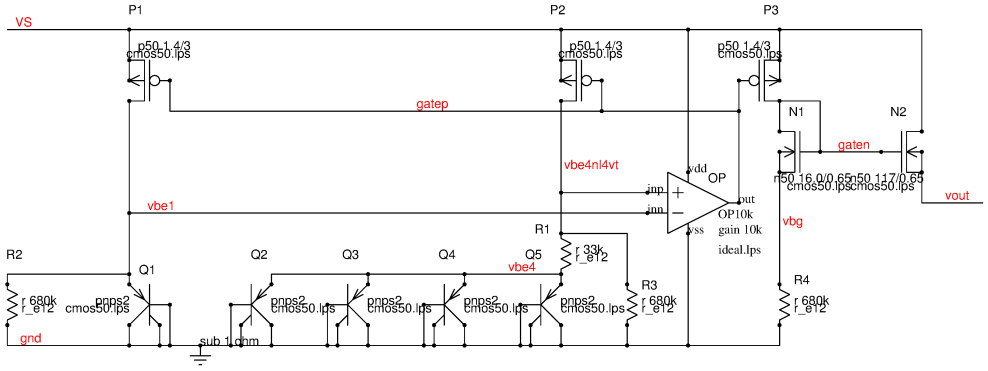


Figure 8.26: Prestabilizer using a 5V CMOS technology

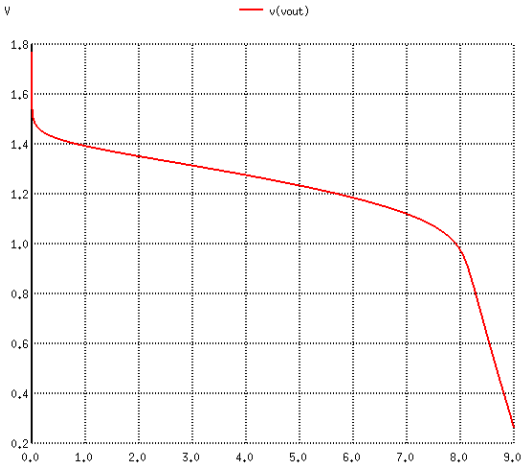


Figure 8.25: DC output characteristic sweeping the load current from 0 to 9mA

In the graph of the output voltage we see two areas where the prestabilizer will not work properly anymore. At very low current the output becomes too high due to leakage. At high currents (above 8mA) the resistor R1 and the current gain of Q2 limit the output current. Using the simple ebers moll formula for the current we would expect a logarithmic decrease of the output voltage between some microamperes and the limitation caused by the gain of the transistor. This simplification in fact holds until about 1mA load current. Above 1mA the ohmic resistance of the emitter limits gm and the load curve becomes more or less linear from 1mA to 6mA.

So you think bipolar is outdated? Well, same concept now with CMOS output stage and a CMOS bandgap. It is easier to linearize anyway.

Simulation shows a similar behavior to the bipolar counter part except that now the output voltage changes with the square root of the current.

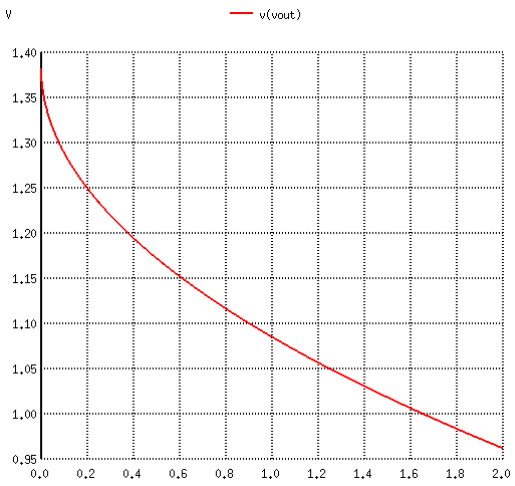


Figure 8.27: Output voltage of the CMOS prestabilizer sweeping the load current from 0 to 2mA

Well, we could have calculated in stead of running a simulation! And equations often give a better insight than

any simulation. We only have to consider that normally N2 is running in strong inversion and the current follows the square of the gate voltage.

$$V_{out} = V_{bg} + V_{GSN1} - V_{GSN2} \tag{8.43}$$

with

$$V_{GS} = V_{th} + \sqrt{\frac{L * I}{K' * W}} \tag{8.44}$$

we get:

$$V_{out} = V_{bg} + \frac{1}{\sqrt{K'}} * (\sqrt{\frac{L_1 * I_1}{W_1}} - \sqrt{\frac{L_2 * I_{out}}{W_2}}) \tag{8.45}$$

The output impedance of the prestabilizer calculates as:

$$R_{out} = \frac{1}{gm} = \frac{1}{2} * \sqrt{\frac{L}{W * K' * I_{out}}} \tag{8.46}$$

To make the prestabilizer work well we have to do the following:

1. Make W/L of the output transistor as big as possible (as long as other parameters don't suffer too much).
2. Choose a transistor type with high K' (This means thin gate oxide).
3. Don't allow the load current to drop too low.
4. Since K' is temperature dependent use similar current densities in N1 and N2. (Otherwise the output voltage becomes temperature dependent)

In many applications a low stand by current I_{out} is desired. This kills the performance of the power supply! Check carefully how much current the system may consume and then take as much of it as a minimum load (resistive minimum load) as you can to make the design robust.

Switching loads: Usually loads supplied by prestabilizers are not constant. The current consumption of the load may change within nanoseconds. A typical example is a CMOS logic (a finite state machine or a small micro controller). Today synchronous design is the standard design style. So all flip flops and all gates change state at the same clock edge. Switching times range from 100ps to about 1ns. The cross conduction during the state change typically is in the range of $100\mu A$ to 1mA (per gate). So even if there are only 1000 gates and flip flops in the logic the expected current spike in in the range of 10mA to 100mA. The prestabilizer can not deliver this current. So a blocking capacitor must be added at the output. The peak current is provided by the capacitor and the active component N2 will have to recharge this capacitor from one clock edge to the next.

Since the blocking capacitor has to carry a high peak current the equivalent series resistance (ESR) of the capacitor is a decisive design parameter. Typically we can accept a drop of 10% of the output voltage. We roughly split it into 50% on the ESR and 50% into discharging the capacitor. Thus we can allow about 60mV drop at the ESR and 60mV discharge. This leads to:

$$R_{ESR} < 0.05 * V_{out}/I_{peak}$$

and

$$C > t_{pulse} * I_{peak}/(0.05 * V_{out})$$

In out example with $I_{peak} = 100mA$,$V_{out} = 1.2V$,$t_{pulse} = 1ns$ we have to choose about $R_{ESR} < 0.6\Omega$ and $C > 1.66nF$. So this is our starting guess what we will need. Further details depend on the shape of the pulse. But this only changes some 10%. The initial guess is already quite close.

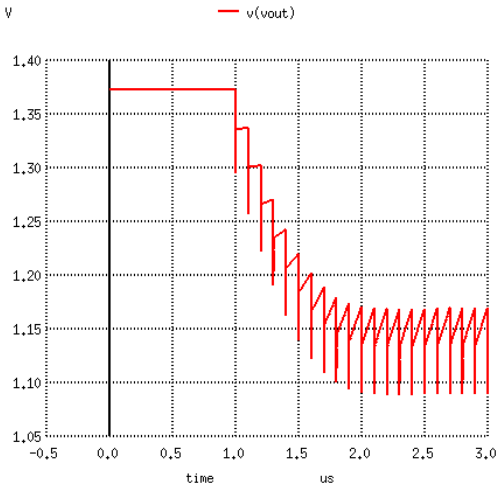


Figure 8.28: Response to a clocking load activated at $t = 1\mu s$

So the average reflects the average current of the load changing from $1\mu A$ to 1mA. The ripple consists of a triangular part showing the recharging of the blocking capacitor and a negative spike representing the voltage drop across the ESR.

The result may look a bit disappointing! It clearly shows that our prestabilizer works acceptably well for load changes of one magnitude. But it can not handle load changes of 2 magnitudes or more. This must be covered by a blocking capacitor of high quality.

Supply voltage transients: In most cases the supply voltage is not stable. Slow changes can be verified using DC simulations. Fast changes leading to currents through parasitic capacities of the circuit must be verified either by transient simulations or by AC simulations. To get a better understanding in the following the blocking capacitor is removed again. As a load a 100K resistor is used. (which corresponds a load current of $12\mu A$.) To achieve stability of the bandgap the PMOS current generator needs a dominant pole adding a 10pF capacitor. (otherwise the simulation oscillates due to an unstable loop inside the bandgap.)

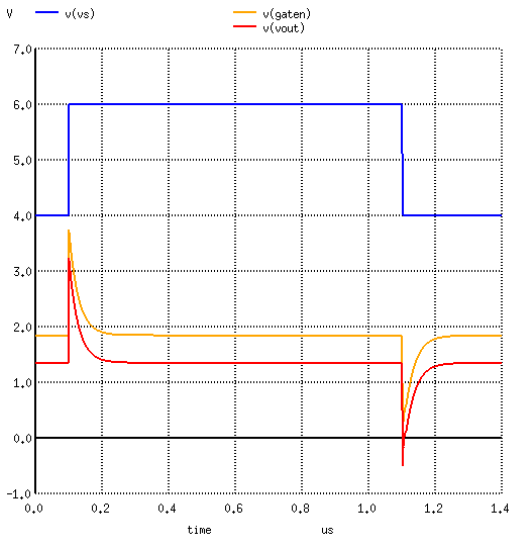


Figure 8.29: Transient rejection of the prestabilizer first draft not having any special precautions

In the above simulation the supply voltage rapidly changes from 4V to 6V and back to 4V. The result is disastrous. The transient almost without any attenuation propagates to nodes gaten and vout. Running an AC simulation does not make it look much better!

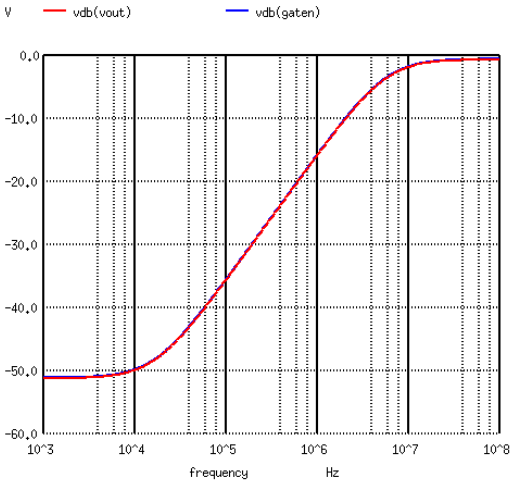


Figure 8.30: AC transfer function of the prestabilizer (Vertical axis is in dB)

This looks like a first order high pass between node VS and gaten. vout simply follows gaten. It is caused by the miller capacity of N2 and the drain-bulk capacity of P3. From the plot we can see that the cut off frequency is about 10MHz. Thus we can calculate the capacity.

$$C_m = \frac{1}{2 * \pi * f_g * R_4} \tag{8.47}$$

In our example we get $C_m = 0.023pF$. Not much, but due to the high impedance at net gaten the effect is tremendous. So what can be done to improve the circuit?

1. decrease the value of $R_4$to shift the cut off frequency to higher values (of course this requires increasing the current through P3 to keep the output voltage at the same value.
2. Add a filter capacity from gaten to gnd.
3. Make N2 and P3 as small as possible to minimize the parasitic capacities.

So here comes the improved circuit including the parasitic stray capacity drawn dashed red:

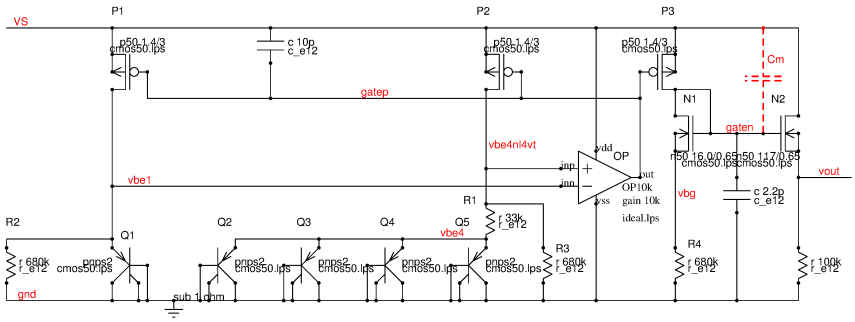


Figure 8.31: Prestabilizer with capacitive filter to improve the PSRR

Now we have created a capacitive divider for high frequency. The expectation is to achieve a flat characteristic above

$$f_g = \frac{1}{2 * \pi * R_4 * C} \tag{8.48}$$

Using C=2.2pF we can expect PSRR to become flat at about 100kHz with an attenuation of about 40dB (defined by the capacitive divider consisting of the miller capacity and the filter capacitor. An AC simulation confirms this expectation up to 100MHz. Above 100MHz we start to see the impact of the drain-bulk capacity of N2 bypassing out nice filter.

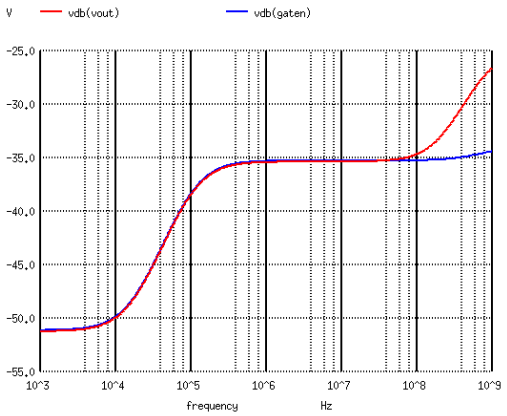


Figure 8.32: Improved prestabilizer's PSRR with supply noise filter at the the gate of N2. (Vertical axis is in dB)

To improve the behavior above 100MHz we have to re-introduce the output capacitor we already had to handle load transients. As long as out load transients are small it is sufficient to choose a capacity about 100 times bigger than $C_{db}$of N2. Since N2 is small a few pF will already do a good job. In the following example the capacity between vout and gnd in 2.2pF.

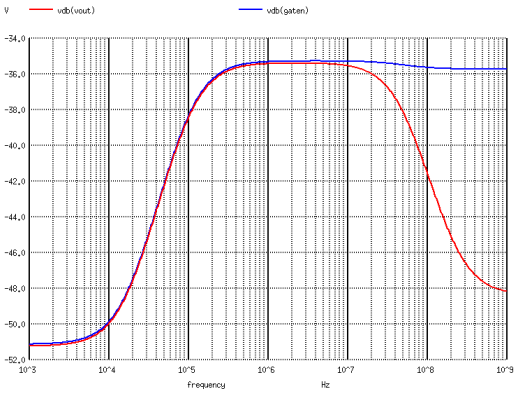


Figure 8.33:

Now we can observe that at the output we created a second low pass with

$$f_{g2} = \frac{gm}{2 * \pi * C_{out}} \tag{8.49}$$

But this is still not quite the truth. The gate of the power transistor has a path resistance. The gate resistance depends on the number of fingers and the resistivity of the gate poly. Making many short gate fingers in stead of one long one is a good idea.

Similar considerations apply to the capacitor connected to net gaten and to net vout. The lower the parasitic resistance the better.

The capacitor(s) connected to vout often are outside of the chip. Thus we additionally may have the parasitic pin inductance and bond wire inductance in series with the external capacitor. To get a conclusive idea of the power supply rejection ratio up to a GHz or even more all these parasitic components must be taken into account. We may end up with a circuit similar to the following schematic:

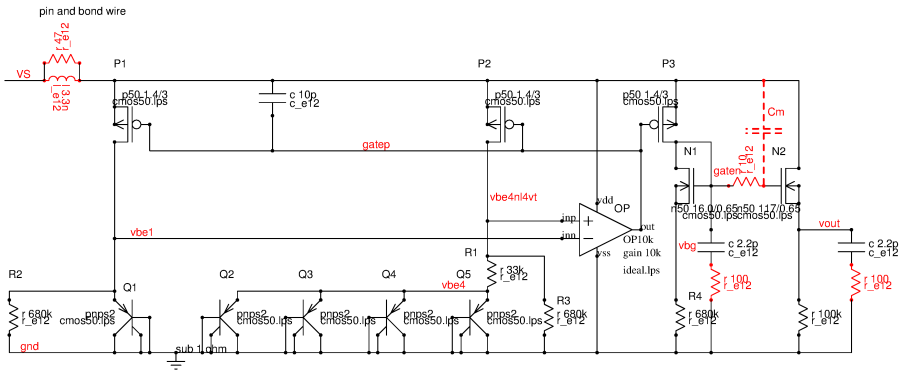


Figure 8.34: Prestabilizer including parasitic components drawn in red

Above about 200MHz the resulting PSRR differs quite a bit from the ideal PSRR we have seen before.

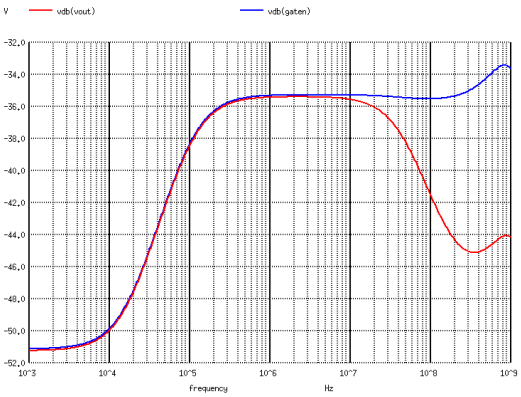


Figure 8.35: Parasitic components start to change the high frequency performance of the prestabilizer above about 200MHz

In our example the main impact is caused by the serial resistance of the gate capacitor. Theoretical calculation of the frequency the serial resistance becomes dominant over the capacity of 2.2pF shows:

$$f_{g3} = \frac{1}{2 * \pi * 100\Omega * 2.2pF} = 720MHz$$

The effect is limited because at the same time the pin and bond wire inductance starts to act as a low pass filter. Thus we rather see a reduction of the PSRR by about 2dB than the expected 3dB at 720MHz.

Rejection of RF applied on the supply side mainly depends on the ratio of the capacities of the power transistor and from the power transistor's gate to ground. To achieve a good PSRR at high frequencies the power transistor should be built as small as possible.

Next point of interest is how much does the input current change if we change the load of the regulator. For DC this is an easy question. The change of the load current will directly be seen as a change of the input current. For alternating current at high frequencies this may change because an increasing fraction of the current starts to flow in the output capacitor in stead of the power transistor. Furthermore we usually are not interested in the current itself but in the disturbance we will see at pin VS. So we need a clear understanding of the network applied to pin VS and its (complex) impedance to ground. This network also includes trace inductances and impedances of the board!

As simple example let us assume the chip is blocked from VS to gnd using a 100nF capacitor with an ESR of 0.05 Ohm and a parasitic inductance of 5nH. Furthermore we assume the cable inductance between the ideal supply and VS is $1\mu H$. The customer measures with a $120\Omega/50\Omega$ attenuator and a spectrum analyzer having an input impedance of 50Ω. This is a usual setup used in many EMC (electromagnetic compatibility) tests. As a load we assume a 1mA AC current sink and a 1mA DC current sink. (Thus the load varies from 0 to 2mA with an average of 1mA).

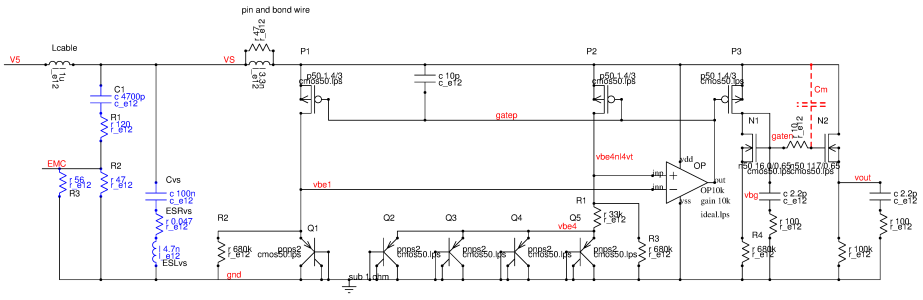


Figure 8.36: Test circuit for the rejection of load current changes (Blue components belong to the external circuitry and the standardized EMC test setup)

The circuit is excited by a load described in the SPICE setup:

```
rload vout gnd 1k
iload vout gnd ac 1m
```

In the following simulation result we see there is almost no attenuation by the prestabilizer. The performance depends mainly on the external components. The first resonance at 500kHz is created by the model of the supply wire ($1\mu H$) together with the external blocking capacitor (Cvs, 100nF). At about 7MHz the blocking capacitor (Cvs)

has a serial resonance with the parasitic inductance (ESLvs). Above 7MHz the behavior of the parasitic inductance dominates the behavior. Above about 100MHz we start to see some influence of the impedance of the power transistor (1/gm).

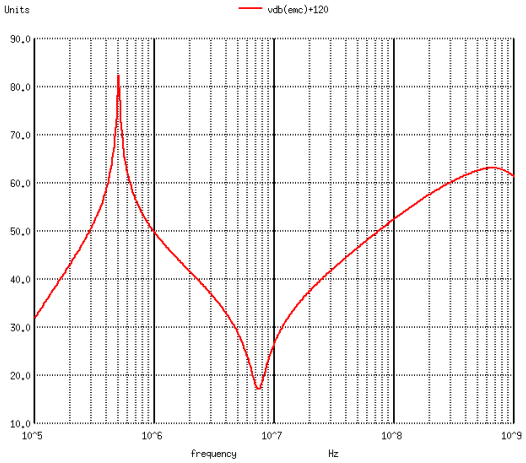


Figure 8.37: Propagation of load ripple (1mA magnitude) to the input vs. (vertical scale in $dB\mu V$)

Thus we can conclude an integrated prestabilizer has very low filtering performance for noise generated on the load side. Except for the source impedance of the power transistor the noise of the load passes the prestabiliser in reverse direction without significant attenuation.

Feeding an AC current into the output of the prestabilizer things get even more ugly. The prestabilizer can source current, but not sink! So we get a rectification. The negative half wave will see the nonlinear output resistance. The positive half wave will look into a MOS transistor in off state. So it will charge the output capacity of the regulator and the average output voltage increases because the output transistor clips the negative half waves. The prestabilizer can not sink the positive half waves.

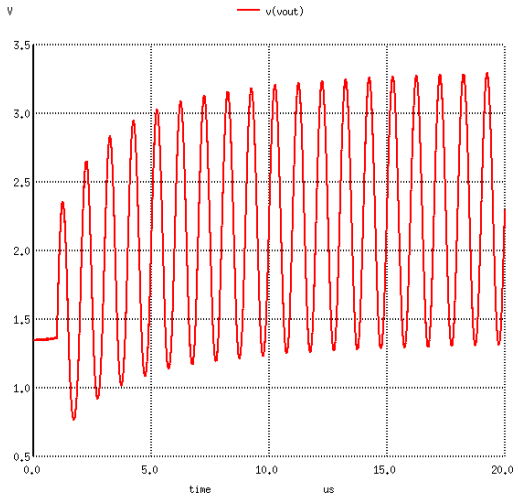


Figure 8.38: Rectification of RF fed into the output of the prestabilizer

The RF source is turned on at $t = 1\mu s$. Before that time we see the DC output voltage of the prestabilizer. A higher load current acts as a sink for the positive half waves. So increasing the load current makes the circuit more rugged against RF fed into the output.

8.2.2 Emitter follower voltage regulator

The emitter follower regulator is the oldest (and probably the best known) implementation of a voltage regulator. Very classical designs are the LM723 or the LM309. LM723 used a zener diode as a reference and a simple differential amplifier as a regulator amplifier. The following figure shows one of the first versions of the LM723 including the external circuitry (in blue) of the most simple application.

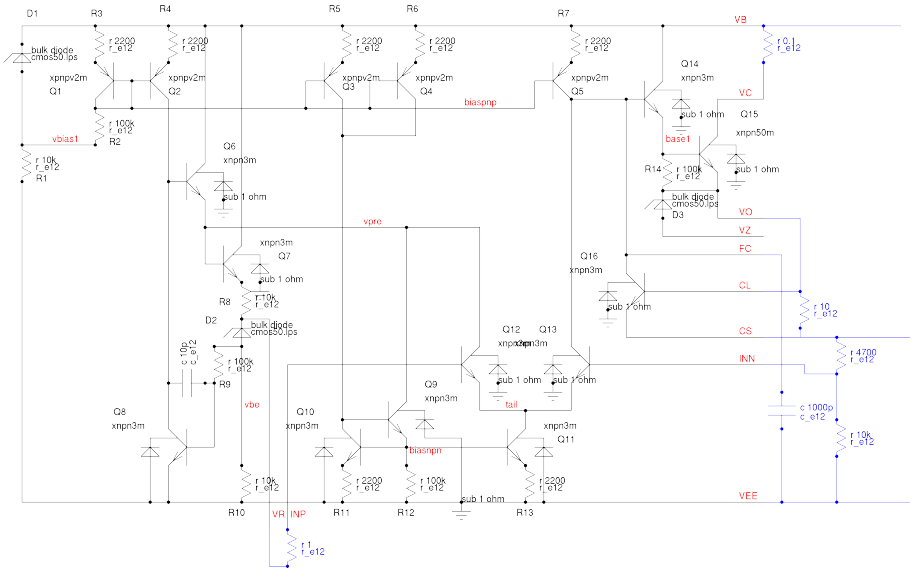


Figure 8.39: One of the first implementations of the LM723 at about 1978

Today the LM723 is still in production but in stead of using a zener diode as a reference most implementations today use a bandgap reference.

To better understand the behavior of the design it is a good idea to simplify it. R1 and D1 together with R2 provide a stabilized bias current to drive the current generators Q1 to Q5.

The reference voltage is provided by the zener diode D2 and the base-emitter junction of Q8. This trick is used because usually 7V zener diodes have a slightly positive temperature coefficient that is compensated by the negative temperature coefficient of the base-emitter junction of Q8.

The regulator amplifier is supplied by vpre. vpre is generated by the regulation loop D2, Q8, Q6, Q7. vpre is about 700mV higher than the reference voltage.

Q11 is a current sink used for the tail current of the regulator amplifier Q12, Q13.

The regulation loop driving pin VO consists of differential amplifier Q10, Q11 and the darlington output stage Q14 and Q15. Q16 and he external 10 Ohm resistor only is needed if a current limitation is required. The current limit in this example is Vbe/10Ω. Ideally Q16 is thermally coupled to Q15. Thus the current limitation decreases when Q15 (and Q16) heat up due to the power dissipation of Q15. This decrease of the current with increasing temperature reduces the risk of thermal destruction of the chip at short circuit operation.

To better analyze the regulator let us reduce it to the essential stuff. We simply regard bias current and reference voltage as a given. So here comes the regulation loop:

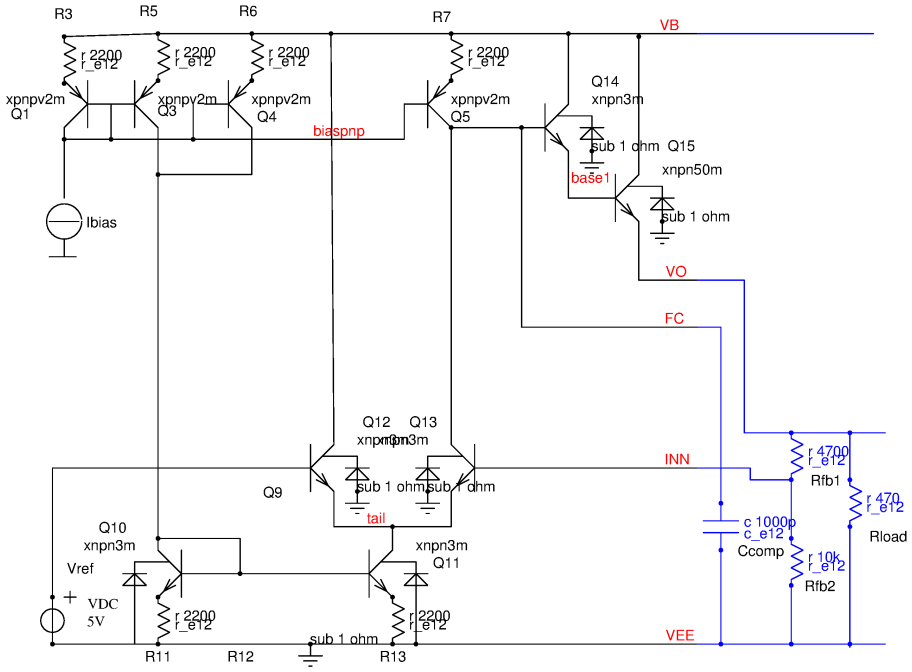


Figure 8.40: Simplified regulator of LM723

Now the regulation loop can be seen much better: Q13 (acting as the voltage amplifier, Q5, Q14, Q15, Rload, Rfb1, Rfb2 and frequency compensation Ccomp.

The minimum voltage drop between VB and VO can be calculated as:

$$Min(V_B - V_o) = 2 * V_{be} + V_{satQ5} \tag{8.50}$$

The base-emitter voltage of Q14 and Q15 depends on the temperature and the load current. To accommodate to a temperature range of -40°C to 150°C the dropout voltage (difference between VB and Vout) of such a regulator usually is specified as 2V.

(To reduce the drop out behavior it is possible to drive the current generator Q5 and the first darlington transistor Q14 from a separate supply that is about 2V higher than the collector voltage of Q15. Nevertheless this approach is not very elegant because the additional supply voltage needed has to carry a load current of I_{out}/B_{Q14} which means the charge pump needed will become strong and expensive. As a consequence bipolar regulators using a voltage follower output stage usually are not usable for very low drop applications.)

Ideally the regulation loop establishes:

$$V(INN) = Vref$$

This applies as long as the currents through Q12 and Q13 are equal. (So the base current of Q14 is much lower than the bias current provided by Q5.)

The input resistance of Q14 becomes:

$$R_{inQ14} = R_{load} * B_{Q15} * B_{Q14} \tag{8.51}$$

The DC gain of gain stage Q13 is:

$$gain_{DC} = R_{inQ14} * I_{Q13}/(2 * V_T) \tag{8.52}$$

The factor 2 in the denominator is caused by the emitter impedance of Q12. If the current trough Q13 increases the current through Q12 decreases and the emitter voltage of Q12 increases accordingly. This lowers the gain by factor 2 compared to the situation of attaching the emitter of Q13 to an ideal voltage source.

$$gain_{DC} = \frac{R_{load} * B_{Q15} * B_{Q14} * I_{Q13} * R_{fb2}/(R_{fb1} + R_{fb2})}{2 * k * T/e} \tag{8.53}$$

Any deviation of the output voltage from the regulation target will be amplifier and increases the base voltage of Q14. Thus the low frequency output impedance (small signal impedance) becomes:

$$R_{outDC} = \frac{2 * k * T/e}{I_{out} * (1 + gain_{DC})} \tag{8.54}$$

This is the output impedance of an unregulated darlington prestabilizer divided by (1+gain).

Stability of the emitter follower regulator: To obtain a stable regulation loop the phase shift of the complete regulation loop must remain below 2π. Since the regulator amplifier is inverting the remaining phase for the low pass filters is only π or 180 degrees. When a phase shift of the reactances of π or 180 degrees is reached the loop gain must be below 1. Thus we can regard a voltage regulator as an OPAMP with a single ended output and a capacitive load. There are 3 main contributors to the phase shift:

1. The load capacity together with the open loop output impedance.
2. The cut off frequency of the regulator.
3. The delay time of the regulator and the feedback divider.

Since these phase shifts are all in a feedback loop these low passes become poles of the regulation loop. The equivalent circuit of the regulation loop is shown in the following figure.

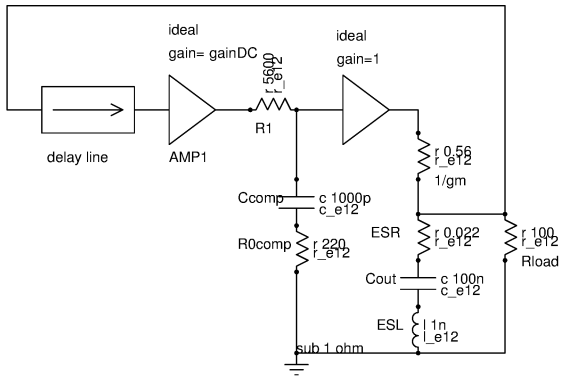


Figure 8.41: AC regulation loop

In the above figure the delay line represents the propagation delay of the regulator amplifier. AMP1 is an ideal amplifier of unlimited bandwidth and a gain equal the DC-gain calculated before. The impedance found at the pin FC is represented by R1 (R1 strongly depends on the implementation of the regulator amplifier. It can range from some $k\Omega$ to several $M\Omega$). The second amplifier stage is an ideal amplifier having a gain of 1 and unlimited bandwidth (In real life the bandwidth of the second amplifier is limited by the speed of the transistors - usually at some 10MHz to 1GHz depending on technology). The transconductance of the output transistors is represented by the resistor between the second amplifier and the output. Rload is the load connected to the regulator.

Usually voltage regulators have an output capacitor to provide pulse currents at frequencies above the cut off frequency of the regulation loop. This output capacitor is represented by Cout. Cout is never ideal. In real life such capacitors have a parasitic resistance (ESR or equivalent series resistance) and a parasitic inductance (ESL or equivalent series inductance). The equivalent series inductance even is present at SMD (surface mount device) capacitors because every component with a current flowing through it is surrounded by a magnetic field. An ESL of 0.5nH is typical for SMD capacitors of the size 805.Adding the traces on the board better assume something in the range of 1nH to 3nH.

The load pole: The output capacitor Cout together with 1/gm of the power transistor forms the load pole. The cut off frequency is:

$$f_{gload} = \frac{gm}{2 * \pi * C_{load}} \tag{8.55}$$

Note that gm is a function of the load current!

Above f_{gload} the loop gain decreases with increasing frequency. The phase margin decreases to $\frac{\pi}{2}$ because of the integrating behavior of C_{load} . If C_{load} were ideal (no ESR, no ESL) the circuit would work perfectly stable without Ccomp. Due to the parasitic components the integrating behavior gets lost at

$$f_{0load} = \frac{1}{2 * \pi * ESR * C_{load}} \tag{8.56}$$

When the ESR becomes dominant the load capacitor does not shift the phase anymore but the loop gain roll off ends as well.

Even worse if ESL becomes dominant the loop gain even increases with frequency if all other components of the loop are not bandwidth limited! The ESL becomes dominant above the resonant frequency of the capacitor.

$$f_{resload} = \frac{1}{2 * \pi * \sqrt{C_{load} * ESL}} \tag{8.57}$$

To get a better feeling of the load behavior let us calculate the example for a load current of 100mA.

$$f_{gload} = \frac{52mV/100mA}{2 * \pi * 100nF} = 820kHz$$

$$f_{0load} = \frac{1}{2 * \pi * 22m\Omega * 100nF} = 723MHz$$

$$f_{resload} = \frac{1}{2 * \pi * \sqrt{1nH * 100nF}} = 15.9MHz$$

So in our example the ESR is meaningless because the capacitor is already inductive long before the ESR starts to play a role! Between 820kHz and 15.9MHz the loop gain roll off is only 26dB.

If the gain of the amplifier AMP1 is higher than 26dB at 820kHz we need a second low pass to take over reaching the resonant frequency of the load capacitor. Theoretically this could be done with Ccomp. But placing the pole of Ccomp above the resonant frequency of C_{load}requires an extremely fast amplifier with a propagation delay in the range of a single ns or even less! The current consumption of such a fast regulator amplifier would be prohibitively high!

Alternatively the cut off frequency of Ccomp can be made lower than the cut off frequency of the load capacity. Assuming we have a DC loop gain of 60dB and we need to be at 26dB (or less) at 820kHz choosing a cut off frequency of Ccomp together with R1 in the range of 8.2kHz (so we have about 6dB amplitude margin) is a reasonable choice.

$$f_{gCcomp} < f_{resload}/gain_{DC}$$

$$f_{gcomp} = \frac{1}{2 * \pi * C_{comp} * R_1} \tag{8.58}$$

Making Ccomp ideal we run into a problem: at f_{gload}we have two poles shifting the phase. This can lead to instability of the loop if there is any further phase shift coming from the delay time of the amplifier. Even if the delay does not yet lead to oscillation the loop will surely tend to ring. To avoid ringing it is suggested to add a zero to the

compensation to reduce the phase shift of the compensation when the load starts to shift the phase. This intentional 0 in the frequency compensation is provided by R0comp.

$$f_{0comp} = \frac{1}{2 * \pi * C_{comp} * R_{0comp}} \tag{8.59}$$

Coming back to the little example circuit we find:

$$f_{gcomp} = \frac{1}{2 * \pi * 5.6K\Omega * 1nF} = 28kHz$$

$$f_{0comp} = \frac{1}{2 * \pi * 220\Omega * 1nF} = 723kHz$$

With the compensation used in the example the maximum allowed gain to still be stable is

$$gain_{DC} = 20 * log10(f_{resload}/f_{gcomp})$$

Filling in numbers we get:

$$gain_{DC} = 55dB$$

Ideally the expected bode plot looks like this:

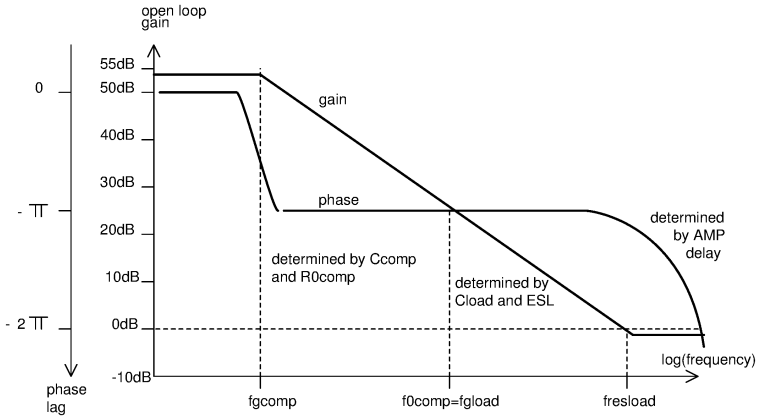


Figure 8.42: Ideal case of the Bode plot of the regulator

From f_{g0comp} until f_{ocomp} the frequency compensation determines gain and phase of the regulation loop. At $f_{gload}$the load capacity and the output impedance (emitter impedance of the power stage) acts as a second low pass. The frequency compensation stops to have an influence there. Above the resonant frequency $f_{resload}$ the load would theoretically be inductive and the gain increases again. Usually there the amplifier itself has a pole (at minimum 1st order, in most cases even higher). Thus above$f_{resload}$ the gain usually further decreases and remains below 1 (0dB) before the phase crosses -2*π .

The amplifier shown is an operational amplifier. Very often in practical implementations an OTA (operational transconductance amplifier) is used instead. In this case the open loop DC gain is determined by the base impedance of the power stage and the gm of the OTA. The gain bandwidth product is determined by the gm and the compensation capacity. Changing the impedance at the base of the power stage will at the same time change the DC gain and the cut off frequency f_{gcomp} . The plateau below f_{gcomp} will move up or down but the roll off curve remains and stability is not affected by the change of the input resistance of the power stage.

A change of gm of the OTA does have an effect on stability because the cut off frequency remains but the DC gain changes.

If anything can go wrong it will: Calculating the regulator for one load current is not sufficient. The load pole depends on the transconductance of the output transistor. This transconductance is a function of the load current. Increasing the load current shifts up f_{gload}! So the stability must be proven for the highest possible load current.

Optimizing for a high load current means taking the 0 of the frequency compensation to a higher frequency. Doing this we will lead to a low phase margin at low load currents. So operating the regulator at lower currents than it was designed for leads to ringing of the regulation loop.

A regulator design can only be a compromise that tries to provide a more or less stable solution for a certain load range. To design a good regulator it is mandatory to have sufficient information about the load and the load capacity including its parasitic components (ESL and ESR).

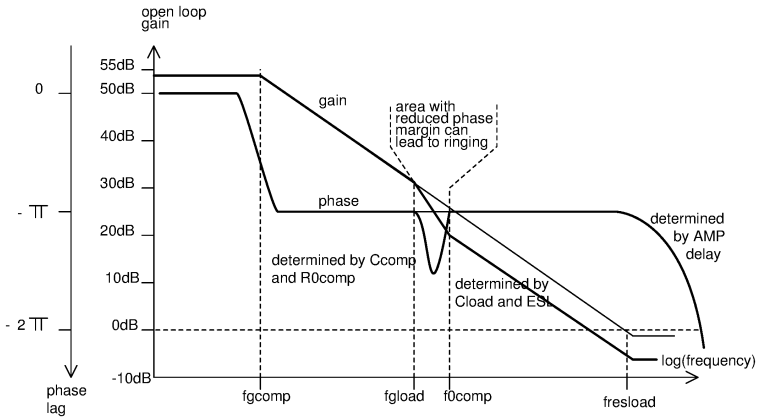


Figure 8.43: Same regulator discussed as before but operated at a lower load current

In the above figure the frequency compensation and the load pole overlap. This leads to a lower phase margin in the range of f_{gload} and f_{0comp} .

Large signal behavior of an NPN regulator: The large signal behavior of an NPN regulator is determined by the load capacity Cload, the load current and the current limitation of the regulator. Furthermore the speed of the regulator amplifier determines the reaction time.

If the output voltage falls significantly below the regulation target the amplifier will increase the base drive of the power stage until the power transistor reaches its current limitation. In most cases an intentional current limitation is implemented to protect the regulator against short circuits. The regulator will recover with a slew rate of:

$$\frac{dV_{recover}}{dt} = \frac{I_{limit} - I_{load}}{C_{load}} \tag{8.60}$$

As soon as the output voltage crosses the target voltage the regulator amplifier reduces the base drive. Since the amplifier is completely overdriven this requires a certain reaction time depending on the design of the amplifier (mainly determined by internal capacities and the bias current of the amplifier). The overshot becomes:

$$V_{overshot} = \frac{dV_{recover}}{dt} * t_{delay} \tag{8.61}$$

When the regulator amplifier has turned off the power transistor the load current discharges the load capacity until the regulator falls below the regulation target again.

$$\frac{dV_{discharge}}{dt} = -\frac{I_{load}}{C_{load}} \tag{8.62}$$

As a result the regulator recovers from large signal distortions with a saw tooth signal. At the end of the saw tooth signal the regulator amplifier remains in the linear range. Thus the regulator will start to show the small signal response then.

This kind of large signal response applies to almost all voltage regulators.

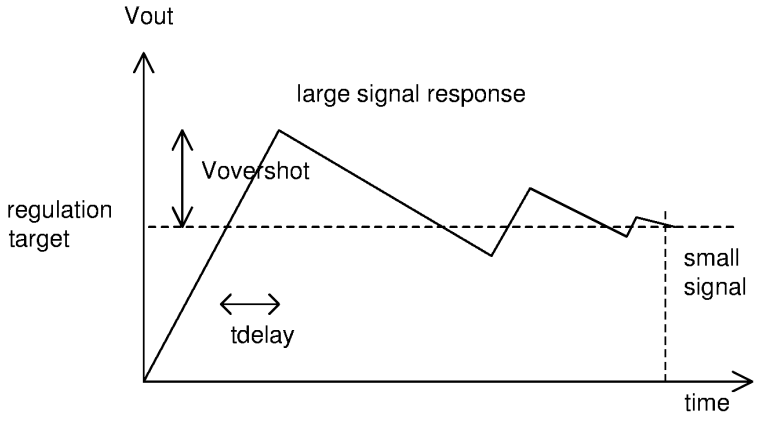


Figure 8.44: Large signal response of a voltage regulator

To reduce the overshot the following measures are recommended:

1. Keep the current limitation as low as possible.

2. Make the regulator amplifier as fast as possible.
3. Clamp the overdrive of the amplifier as close to the normal operating range as possible.
4. Use a sufficiently large output capacitor Cload.

One of the advantages of bipolar output stages is the base-emitter junction of the power transistor acting as a clamp for the regulator amplifier. Especially in combination with a diode (Anode at the emitter of the power NPN transistor, cathode at the base) the regulator amplifier will nicely be kept close to the desired operating point. Compared to MOS power transistors that do not inherently offer this clamping behavior most NPN regulators have a nicer large signal response at the same speed of the amplifier.

Over-engineering a regulator offering a higher current limit than needed for the specific application and using a low current consumption design (this leads to a slow amplifier with long delay time) without need will lead to poor large signal performance. Operating this regulator with a low load capacity will further escalate the problems.

8.2.3 Source follower voltage regulator

Using a power MOS transistor operating as a source follower leads to a similar behavior as the one discussed above using an NPN power transistor. The open loop output impedance of the power stage slightly increases compared to bipolar output stage, but most of the calculations already shown for the NPN regulator still apply. As long as the power transistor is operated at very low current densities it is in weak inversion. The open loop output resistance seen at the source becomes:

$$R_{OLwi} = \frac{1}{gm} = n * \frac{K * T}{e * I_{out}} \tag{8.63}$$

In this equation factor n is the gate coupling factor

$$n = \frac{C_g + C_{bulk}}{C_g} \tag{8.64}$$

Typically n is in the range of 1.2 to 1.6. Thus at low currents the load pole of the regulator operating in weak inversion moves down 20% to 60% compared to an equivalent NPN output stage.

Operating in strong inversion the open loop output impedance of the power transistor changes.

$$I_d = I_s = 2 * k' * \frac{W}{L} * V_{gseff}^2 \tag{8.65}$$

$$R_{OLsi} = \frac{1}{dI_d/dV_{gs}} = \frac{L}{2 * k' * W * V_{gseff}} \tag{8.66}$$

Still not very convenient because we don't know V_{gseff} yet. Let us expressV_{gseff} using the drain current.

$$V_{gseff} = \sqrt{\frac{L}{W}} * \frac{1}{\sqrt{I_d * k'}} \tag{8.67}$$

So the open loop output impedance in strong inversion calculates as

$$R_{OLsi} = \frac{1}{\sqrt{I_d * k'}} * \sqrt{\frac{L}{W}} \tag{8.68}$$

This means the load pole moves to higher frequencies following $\sqrt{I_d}$ as soon as we are in strong inversion. A change of the current from 1mA to 100mA changes the pole only by factor 10 (compared to a factor 100 using a bipolar transistor) provided the power stage is already in strong inversion at 1mA load current. On the other hand a fast change of the load current (faster than the regulation loop will react but too long to buffer from Cout) will lead to a higher short term impact on the output voltage than using an NPN output stage.

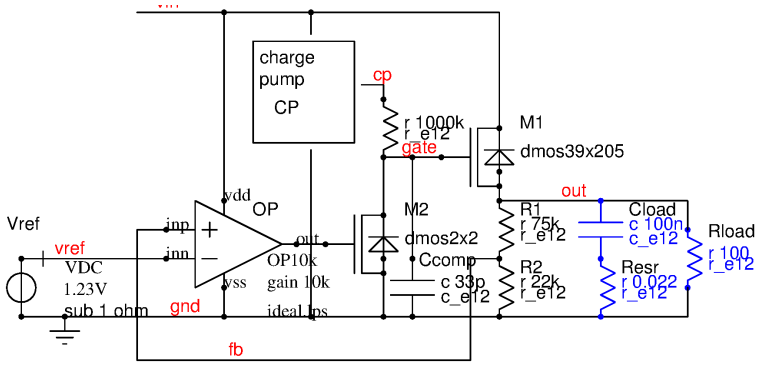


Figure 8.45: Voltage regulator with NMOS output stage

Looking at the schematic above we see 2 significant differences compared to the NPN regulator discussed before:

1. The gate voltage of M1 is supplied by a charge pump and a pull up resistor.
2. The regulator OP indirectly drives M1 using high voltage transistor M2 as an interface.

Using a charge pump offers very low drop capability. This means the voltage drop required between the drain and the source of M1 only depends on the Rdson of M1.

$$V_{inmin} = V_{out} + R_{dsonM1} * I_{load} \tag{8.69}$$

Since we only have to supply a few μA from the charge pump the low drop capability has become affordable using an NMOS output stage.

To keep the cost of the charge pump low the regulator amplifier intentionally is not supplied from the pump. In many cases the regulator amplifier even is supplied from a prestabilizer providing about 3..5V (for simplicity this has been omitted in this conceptual circuit). Supplying the regulator with only 3..5V offers two advantages:

1. Low voltage transistors usually match much better than HV-transistors with their inhomogenous drain extensions.
2. Low voltage transistors are smaller and faster (lower parasitic capacities).

These optimizations of the topology have their price:

Now we have more than one voltage gain stags in the regulation loop (OP and M1). In most cases a parallel frequency compensation as shown in the conceptual circuit above will not work anymore. In stead we will need a nested miller compensation of possibly both: M1 and OP.

Using a nested miller compensation we loose the nice gate voltage storage capability of the parallel compensation (see discussion of the NMOS prestabilizers a couple of pages back).

Combining a reduced parallel compensation with a nested miller compensation in many cases leads to a reasonable compromise between stability and supply rejection. The path of optimizing this kind of circuit for all load cases, temperatures and technology corners is long and cumbersome. The following picture shows the resulting - already uncomfortably complex - circuit.

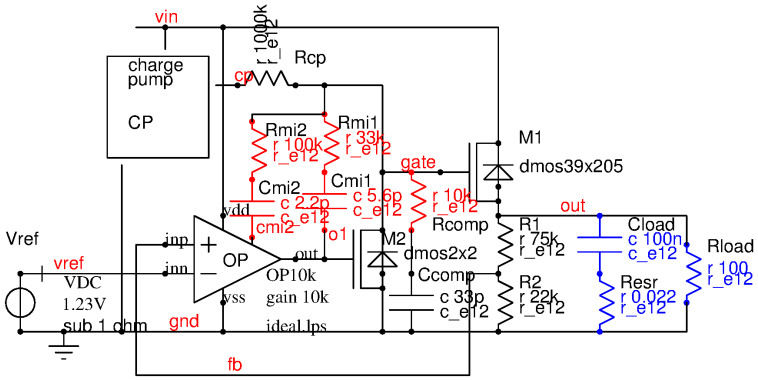


Figure 8.46: Same regulator as before now including the nested miller compensation (additional components drawn in red color)

Since the miller compensation paths connect directly to the nodes o1 (output) and cmi2 (compensation) of the regulator amplifier itself we have to dive down into the details of the amplifier OP.

8.2.4 Source follower regulator with stacked output transistors

For applications without blocking capacitor the output stage should have the following properties:

1. High gm (low output impedance
2. low Cgs (to minimize pulling the gated down when the load turns on)
3. low Cgd (for high supply rejection)

If there is enough headroom available stacking a low voltage transistor and a high voltage transistor offers an interesting topology. The low voltage transistor can be designed with such a big width that it is operating close to weak inversion. (highest possible gm) and the parasitic capacities are still acceptable. The stacked high voltage transistor operates in strong inversion (lower gm) and mainly has to protect the low voltage transistor against a too

high Vds. The high voltage transistor can even be driven by a buffer stage to keep the impedance at the gate of the high voltage transistor low.

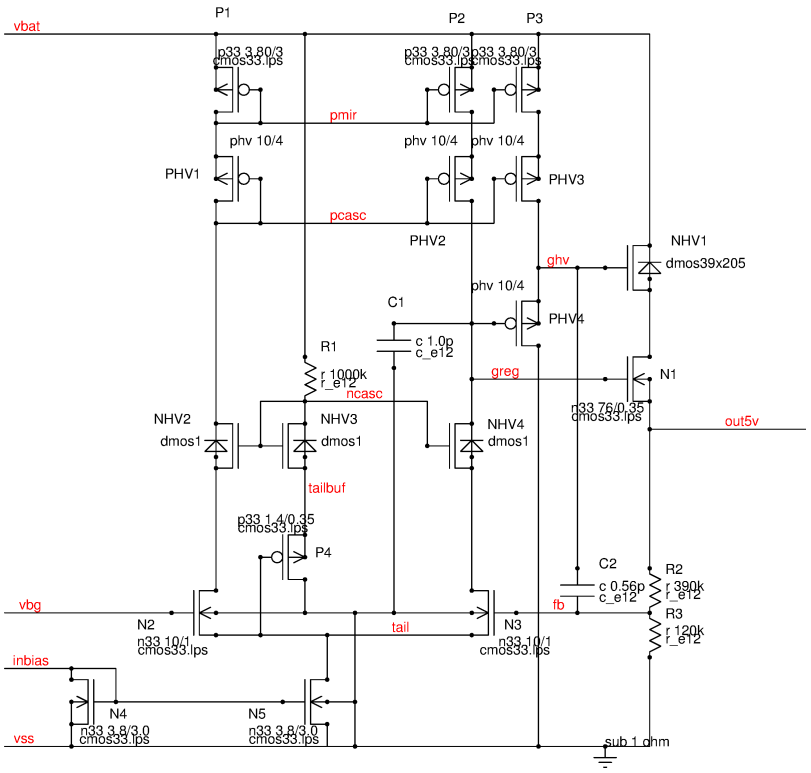


Figure 8.47: NMOS regulator with stacked output transistors

The circuit shown is without charge pump. Adding a charge pump is possible supplying the current mirror P1, P2, P3 from the charge pump.

C1 provides the dominant pole. If there is a load capacity connected to the output of the regulator C1 may need a resistor in series for better phase margin. But even without resistor the regulator is stable due to the phase feed forward provided by C2. C2 is intentionally driven from the source of PHV4 to not slow down the regulation loop by the capacitive load of C2 and to be sure that even with an infinite load capacity there still is a resistive component (1/gm) between the phase feed forward and the external capacitor that acts as an integrator.

The biasing of the cascodes follows the tail voltage. So the differential stage always operates at Vds=Vgs of P4. (Alternatively the gates of NHV2 and NHV4 can also be driven from about 3-4V if such a cascode bias is available.)

The dropout voltage of the regulator (without charge pump) is determined by the Vdssat of N1, Vgs of NHV1 and the Vdssat of P3 and PHV3. Typically about 2..3V can be expected (worst case at cold, slow models).

8.2.5 Low drop regulators with PNP power transistors

Low drop regulators (LDO) using PNP transistors or PMOS transistors as power devices offer a low voltage drop between the input and the output of the regulator without the need of a charge pump. Therefore PNP or PMOS LDOs are a prefered solution to provide a moderate current with low noise (no charge pump noise).

Basic concept of a PNP LDO: The LDO using PNP transistors is the most classical implementation of a low drop regulator. It consists of a regulator amplifier, a high voltage driver stage converting the output voltage of the regulator amplifier into a current and an output transistor. To become independent of changes of the gain B of the power transistor the output staage often is implemented as a current mirror.

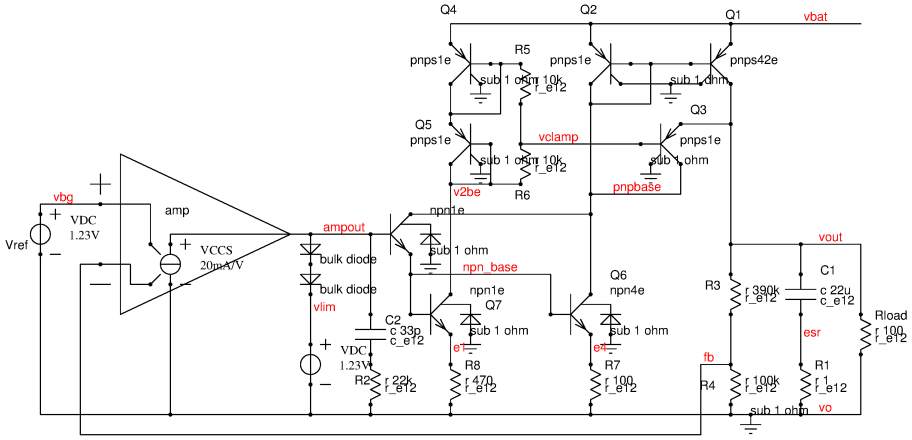


Figure 8.48: LDO using PNP transistors

The circuit shown above can be regarded as a standard LDO of the 1980s and 1990s. Devices like the L4949, L4938, TLE4269 and others used more or less the same topology. There only were minor modifications of the circuit depending on technology details (mainly the use of vertical PNPs in the L49.. family versus use of lateral PNPs in the TLE42.. family).

The output of the regulator is the collector of Q1. So the output signal driving net vout is a CURRENT! The current is converted into a voltage by integrating it in capacitor C1. Usually C1 together with the load impedance Rload provides the dominant pole of the regulation loop. Compared to an NPN regulator the two poles of the loop have changed their places.

$$f_{p1} = \frac{1}{2 * \pi * C_1 * R_{load}} \tag{8.70}$$

Example: $R_{load} = 100\Omega$ and $C_1 = 22\mu F$ provides $f_{p1} = 159Hz$

Usually C1 is an electrolytic capacitor. In the 1990s mass production aluminium electrolytic capacitors used for such applications had an equivalent series resistance (ESR) in the range of 0.1Ω to 10Ω. So the integrating behavior already stopped again between 1.6KHz and 160kHz depending on the ESR. For the regulation loop this is a zero.

$$f_{zero1} = \frac{1}{2 * \pi * C_1 * R_1} \tag{8.71}$$

To keep the regulation loop stable an other pole had to take over to bring down the loop gain to 0dB before the regulator amplifier's (amp) internal parasitic poles significantly shift the phase. So the internal pole defined by the impedance of net ampout and C2 is placed at the same frequency as the zero. Since for general purpose regulators there is a big variation of the capacitors used the 2nd pole is made a bit faster than the worst case zero. 2kHz is a typical choice. This way the pole-zero cancelation works reasonably well for high ESR values and the bandwidth requirements for the amplifier and the power transistor still are in reach of a standard 40V bipolar technology.

Assuming an ideal OTA as a regulator amplifier the impedance of net ampout is mainly determined by R7 and the gain of the darlington transistor.

$$R_{ampout} = R_7 * B^2 \tag{8.72}$$

Example: B=150 and $R_7 = 100\Omega$ provides $2.25M\Omega$.

Practical values (especially at low load current) are a bit higher due to the emitter impedance 1/gm of Q6 that must be added to R_7.

The frequency compensation capacitor C_2 calculates as:

$$C_2 = \frac{1}{2 * \pi * R_{ampout} * f_{p2}} \tag{8.73}$$

with $f_{p2} = 2kHz$. In our design example this leads to $C_2 = 36pF$.

If the load capacitor has a lower ESR the regulation loop will see a reduction of the phase margin between $f_{p1}$and f_{zero1}. Reducing the ESR thus is critical for stability of the regulator. A regulator designed for a wide range of external capacitors having a wide range of ESRs is always a trade off between stability and DC accuracy.

Usually the first parasitic pole is provided by the base capacity of Q1 and the impedance at the base. Since the transistor is working as a mirror (in the case shown having a ratio of 42) the base small signal resistance is about:

$$R_{basepnp} = \frac{K * V_t}{I_{out}} \tag{8.74}$$

with K=42, $V_t = 26mV$ and $I_{out} = 50mA$we get a small signal base impedance of about $22K\Omega$. Typical capacities of such power PNP transistors are around 200pF. So the pole produced by the PNP mirror is at:

$$f_{pmirror} = \frac{1}{2 * \pi * R_{basepnp} * C_{base}} \tag{8.75}$$

In the example shown we get $f_{pmirror} = 36kHz$. The pole however moves with the load current because the changing current changes the real part of the small signal input resistance of the PNP mirror. So the pole-zero compensation attempted with R2 always is only a rough optimization for a typical operating point.

$$R_2 = \frac{1}{2 * \pi * C_2 * f_{pmirror}} \tag{8.76}$$

This way R_2becomes about $123K\Omega$.

Due to the movement of the poles it is common practice to let the loop gain cross the 0dB line at or before reaching the PNP mirror pole. This is the stability limit of the transimpedance of the regulator amplifier. Assuming the poles and zeros cancel well we will find a gain roll off of -20dB/decade. (This would exactly be the case if the ESR of the blocking capacitor matches the the internal pole of 2kHz). The DC loop gain calculates as:

$$gain_{DC} = R_{load} * K * \frac{1}{R_7} * R_{ampout} * gm_{amp} * \frac{R_4}{R_3 + R_4} \tag{8.77}$$

Tos reach stability the DC lopp gain must be below:

$$gain_{DC} < \frac{f_{pmirror}}{f_{p1}} \tag{8.78}$$

In our example this is: $gain_{DC} < 36kHz/159Hz = 226$.

The transconductance of the amplifier must be below

$$gm_{amp} < \frac{f_{pmirror} * (R_3 + R_4) * R_7}{R_{load} * K * R_{ampout} * R_4 * f_{p1}} \tag{8.79}$$

replacing the term $R_{load} * f_{p1}$ with $1/(2 * \pi * C_{load})$ the equation becomes independent of R_{load} .

$$gm_{amp} < \frac{2 * \pi * C_{load} * R_7 * (R_3 + R_4) * f_{pmirror}}{K * R_{ampout} * R_4} \tag{8.80}$$

The equation shows that the gain of the regulator is directly limited by the first parasitic pole $f_{pmirror}$ and by the external load capacity C_{load} . If we want to increase the loop gain to enhance accuracy this must either be paid by a bigger external capacitor or by using a faster technology for the output transistor.

The DC loop gain determines the DC output impedance of the voltage regulator. It is:

$$R_{outDC} = R_{load}/gain_{DC} \tag{8.81}$$

Well, a bit fast.. Let's look at it in more detail.

$$R_{outDC} = dV_{out}/dI_{out} \tag{8.82}$$

The change of the current provided by Q1 is:

$$dI_{out} = dV_{out} * \frac{R_4}{R_3 + R_4} * gm_{amp} * R_{ampout} * \frac{1}{R_7} * K \tag{8.83}$$

or if reordered:

$$\frac{dV_{out}}{dI_{out}} = \frac{R_7 * (R_3 + R_4)}{R_4 * gm_{amp} * R_{ampout} * K} = R_{load}/gain_{DC} \tag{8.84}$$

Looking at out example regulator the output impedance is slightly below 0.5Ω. Well, the load rejection of L4949 is specified to be 20mV/50mA=0.4Ω . If we want to improve the load rejection we have to increase $gm_{amp} * R_{ampout} * K$. As we have seen before increasing the loop gain either requires a lower dominant pole (bigger C_{load}) or a higher parasitic pole of the output stage (faster technology). Load rejection, required load capacity and speed of the technology thus determine the performance of a low drop regulator. The consequences of this observation are:

- Don't make the output transistor any bigger than needed.
- Use the fastest technology available for the voltage range the regulator is intended for.
- Make the load capacity as big as possible
- Keep the ESR of the load capacity under tight control to prevent loss of phase margin in the middle of the operating frequency range. (Specify the ESR range)

Eventually the design of regulators with PNP power transistors gets limited by the mobility of the minority carriers inside the transistors. Vertical PNPs are a bit better due to the thinner base, but even vertical PNP power transistors usualy have a transit frequency in the range of only some 10 MHz.

Further tricks needed:

The PNP transistors dump a lot of current into the substrate if they operate is saturation. Therefore the base drive must be reduced approaching low drop. This is done by Q3 and the diode stack Q4, Q5. Resistors R5 and R6 fine tune the level at which the saturation protection reduces the base drive.

Technologies using lateral PNP power transistors can use sense collectors (an outer collector ring surounding the main collector) to pull up the base before saturation is reached.

Current limitation of the power stage relies on the aspect ratio of Q1 and Q2 and the base voltage clamp at the base of the darlington NPN transistor.

The L4949 transistor level: As a very simple practical implementation let's have a look at the L4949.

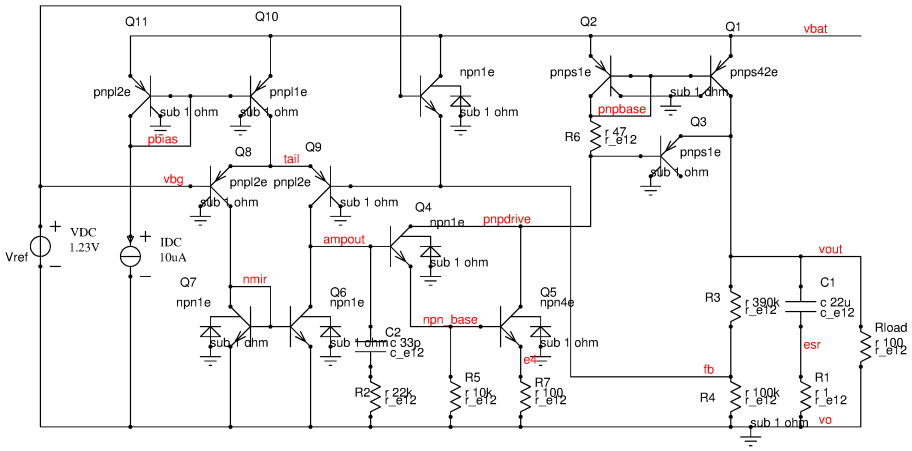


Figure 8.49: The regulator part of the L4949

The current limitation doesn'r require a clamp. It simply relies on the limited voltage swing of net ampout. The reference voltage of 1.23V (bandgap voltage) clamps the voltage of net tail at about 1.9V. This way the voltage at ampout can't get mich higher than 1.23V+Vbe-0.1V. This limits the volage available at R7 to about 400mV (at cold) to about 800mV (at hot).

The current limit has a fold back characteristic because for very low output voltages the voltage at tail drops to values between as low as 2*Vbe (if the output voltage is 0V). In this case the current through Q5 is close to 0 and only Q4 will carry current (flowing through R5). For pure resistive loads this may still be acceptable. If the load is a constant current sink this fold back characteristic may lead to an output voltage stuck close to 0V. The effect can be reduced clamping the base voltage of Q9 slightly below Vbg (for instance by using an emitter follower with the base connected to Vbg+Vbe-0.1V and the emitter connected to net fb)

To eliminate this fold back effect L4938E used a slightly different amplifier structure.

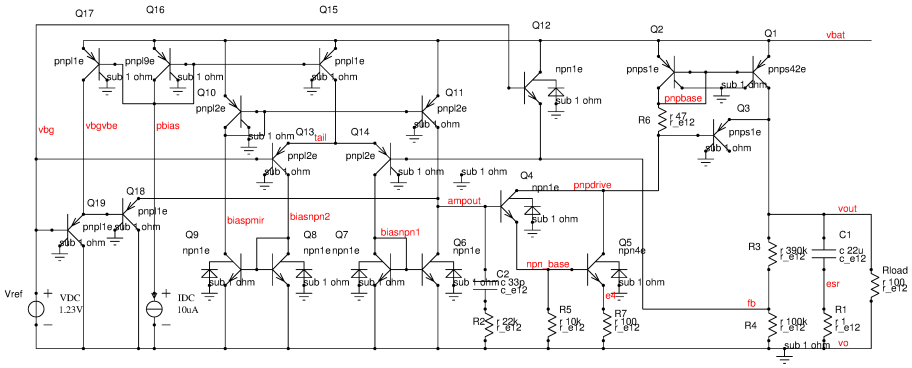


Figure 8.50: Regulator amplifier of the L4938E

Here the signal ampout is limited to Vbg+2*Vbe by the reference voltage and the transistors Q18 and Q19. This limits the voltage drop at R7 to a very stable vbg and the current limit will have significantly less spread.

Limitations of bipolar technologies: Bipolar transistors usually are significantly larger than 5V or 3.3V CMOS transistors. Thus they have significantly higher substrate capacities than their CMOS counter parts. The parasitic

poles of the regulator amplifiers of L4949 and L4938 were determined by the bias currents of the differential stage and by the substrate capacities of the transistors rather than by the transit frequency of the transistors themselves. The lower the bias currents get the more the substrate capacities become a stability problem. Replacing the bipolar regulator amplifiers by CMOS transistor amplifiers offers significant bandwidth improvemeents.

8.2.6 Low drop regulators with PMOS power transistors

In principal the same topology can be used to build an LDO with a PMOS power stage. But high voltage PMOS transistors have a significantly bigger drain area than the collector area of their bipolar counter parts. This leads to a much higher miller capacity of the power stage. Using a high impedance driver stage this miller capacity degrades the PSRR of the regulator. Therefore PMOS regulators usually have a low impedance buffer amplifier between the regulator amplifier and the power transistor. This buffer decouples the frequency compensation from the miller capacity (it can be regarded as a capacity multiplication). The buffer amplifier can only be omitted if the PMOS transistor is very small. (Low current required or the regulator is optimized for a low drop between input and output.) Let's start with the more general case of a large PMOS power transistor.

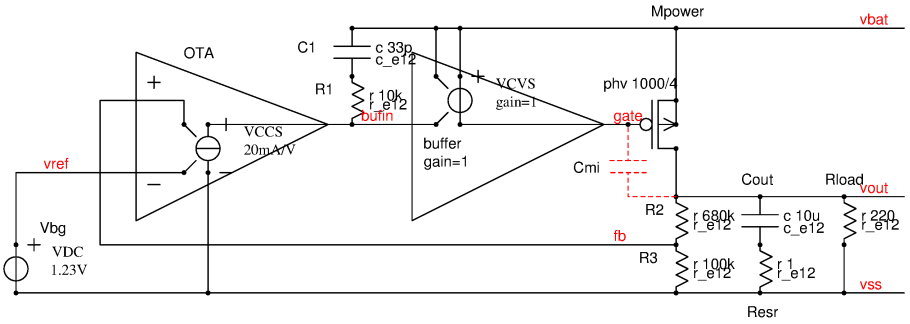


Figure 8.51: PMOS regulator with buffer amplifier

The figure above shows the concept of the L4952 regulator. The miller capacity of the large power transistor (Cmi) is almost taken out of the regulation loop by the low resistive buffer stage. This way a fairly small capacity C1 can be used in the frequency compensation. One of the disadvantages of this concept is that now we have the delay of two amplifier stages in the regulation loop. Usually the buffer is designed for highest possible speed and lowest possible output impedance driving the gate of Mpower. In case of the L4952 the buffer is an amplifier with feedback to achieve a low impedance. To achieve enough speed the buffer amplifier stage consumes 80% of the whole supply current of the regulator!

8.3 Chargepump

Charge pumps are used to provide supply voltages exceeding the supply rails of the chip. Under normal circumstances charge pumps are not able to provide high currents. So their use normally is limited to supply gate drivers or write (or erase) stages for non volatile memories. The basic concept alway is an AC voltage source, a pump capacitor and a rectifier stage. Using on chip capacitors the currents provided are in the range of some hundred μA or less.

Using external capacitors the current range of charge pumps can be extended to some mA or some 10mA.

8.3.1 Simplified chargepump with ideal rectifier and ideal switches

This very much simplified circuit is only for instructive use. Of course in real life ideal switches do not exist. In spite of all the simplifications it provides first insight into the behavior of a charge pump.

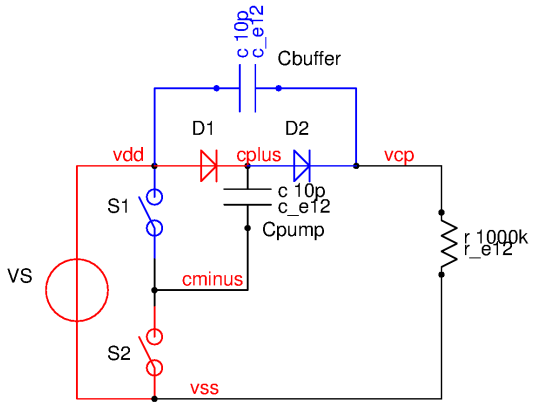


Figure 8.52: Concept of a charge pump

The charge pump has two states:

- S1 opens, S2 closes. The current flows the red path through VS, D1, Cpump, S2. This way the capacitor Cüump is getting charged.
- S1 closes, S2 opens. The current flows the blue path through S1, Cpump, D2, Cbuffer. This way the energy stored in Cpump is getting transfered into Cbuffer.

Assuming a low load current through Rload Cbuffer can be charged almost to the level of the supply VS. The output voltage of the pump is ideally becomes

$$V_{00}(vcp) = 2 * V(vdd) \tag{8.85}$$

V_{00} is the open load voltage of the charge pump measured versus ground. The open load output voltage measured versus vdd is

$$V_0(vcp) = V(vdd) \tag{8.86}$$

The load of the charge pump can either be connected from vcp to vss or from vcp to vdd. In the following connecting the load from vcp to vdd is considered (This is the more straight forward case because the load current strictly becomes a function of Cpump, the frequency and voltage change over Cpump during one period. The (average) short circuit current shorting the pump output vcp to vdd can be calculated:

$$I_{sh} = C_{pump} * V(vdd) * f \tag{8.87}$$

Thus the output resistance of the ideal charge pump becomes:

$$R_{out} = V_0(vcp)/I_{sh} = \frac{1}{C_{pump} * f} \tag{8.88}$$

If higher voltages are required several stages must be stacked. This approach has been used at least since the 1960 to provide the CRT (cathod ray tube) acceleration voltage of TV receivers.

8.3.2 Charge pump with resistive switches and rectifiers

If we use resistive switches the output resistance of the charge pump changes. As long as we are using integrated capacitors with some 10pF the impact of the switch resistance usually is negligible. Using external capacitors that can have several nF and load currents in the mA range these switch resistances can no more be neglected.

In an integrated circuit design using state of the art technologies the metal paths can be as resistive as the switches themselves! So in the following drawing the path resistances are taken into account as well.

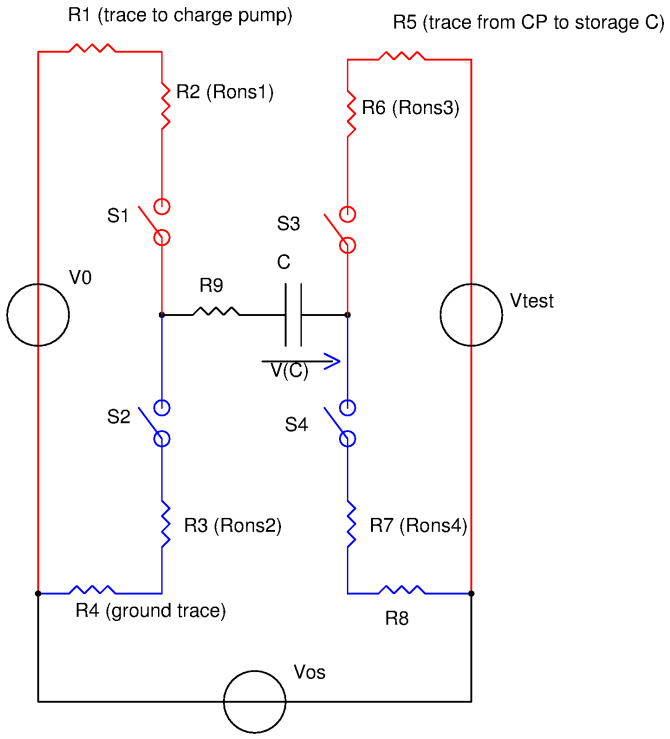


Figure 8.53: Charge pump with resistive switches

Now the driver stage is represented by switches S1 and S2 that have an ON-resistance (R2 and R3) plus path resistances to supply (R1) and ground (R4). R9 is the resistance of the trace from the driver stage to the pad connecting the (external) pump capacitor. On the rectifier side we also have the resistance of the switches S3 and S4 acting as synchronous rectifiers (R6 and R7) as well as the metal paths (R5 and R8).

As before the open load output voltage (measured at Vtest is Vtest is not allowed to drain current) is equal V0.

Aproximation for low capacitor values (change of the voltage of the pump capacitor is bigger than the resistive drop): The short circuit current is flowing if Vtest is 0V and the (average) current is flowing through Vtest. Now the voltage swing is distributed over the resistors and capacitor C. During charging of capacitr C the switches S2 and S4 are closed while S1 and S3 are open. The voltage over C increases. Assuming C was fully discharged the voltage drop remaining on the resistive part of the path (R4, R4, R9, R7, R8) becomes

$$V_{rch} = V_0 * exp(\frac{-t}{R_{ch} * C})) \tag{8.89}$$

with

$$R_{ch} = R4 + R3 + R9 + R7 + R8 \tag{8.90}$$

The time t can be represented by the charge duty cycle $m_{ch}$and the frequency f of operation of the pump.

$$t = \frac{m_{ch}}{f} \tag{8.91}$$

We now can rewrite the voltage loss of the charge path at the end of the charge time.

$$V_{rch} = V_0 * exp(\frac{-m_{ch}}{R_{ch} * C * f}) \tag{8.92}$$

In the same way the voltage drop of the pump path at the end of the pumping cycle can be described by

$$V_{pu} = V_0 * exp(\frac{-m_{pu}}{R_{pu} * C * f}) \tag{8.93}$$

$m_{pu}$is the duty cycle of the pump phase. Since in most cases there will be a non overlap time whre all switches are open to avoid cross conduction we have one more condition:

$$m_{ch} + m_{pu} < 1 \tag{8.94}$$

Of course to maximize the energy transfer the sum should be as close to 1 as possible (or permissible for RF emission reasons).

Now we know the voltage swing at capacitor C.

$$\Delta V_c = V_0 - V_{rch} - V_{pu} \tag{8.95}$$

As in the ideal charge pump the short circuit current can be calculated using the voltage swing at the capacitor. Only difference is that the voltage swing at the capacitor now is reduced by the drops remaining at the resistors at the end of the charge cycle and at the end of the pump cycle.

$$I_{sh} = f * C * \Delta V_c \tag{8.96}$$

Again we can calculate the output resistance in the usual way.

$$R_{cp} = \frac{V_0}{I_{sh}} \tag{8.97}$$

Plugging in the equations for the short circuit current we get:

$$R_{cp} = \frac{1}{C * f * (1 - exp(\frac{-m_{ch}}{R_{ch} * C * f}) - exp(\frac{-m_{pu}}{R_{pu} * C * f}))} \tag{8.98}$$

Note: This equation holds a simplification assuming we always charge the capacitor from 0V to $V_0 - V_{rch}$or discharge from V_o to V_{pu}. This is note quite true but the error of the approximation is acceptable as long as at the end of the charge cycles less than 20% of the drop is in the resistors and 80% or more is across the capacitor.

Aproximation for big capacitors (drop over the resistors is dominant): If external capacitors are used the typical component choice is to make the pump capacity so big that there is amlost no change of the volage across the capacitor. Then a different aproximation is needed. Here the assumption is that the voltage across C is constantly V_c and that we have a current ballance.

$$m_{ch} * I_{ch} = m_{pu} * I_{pu} \tag{8.99}$$

The charging current I_{ch}depends on the resistance of the charge path and the difference of $V_0$and V_c.

$$I_{ch} = \frac{V_o - V_c}{R_{ch}} \tag{8.100}$$

The pump current pumping into a short circuit is

$$I_{pu} = \frac{V_c}{R_{pu}} \tag{8.101}$$

Using these equations $V_c$can be calculated:

$$V_c = \frac{m_{ch} * V_0}{m_{pu} * (\frac{R_{ch}}{R_{pu}} + \frac{m_{ch}}{m_{pu}})} \tag{8.102}$$

Dividing with R_{pu}leads to the pulsed short circuit current.

$$I_{pu} = \frac{m_{ch} * V_0}{m_{pu} * R_{ch} + m_{ch} * R_{pu}} \tag{8.103}$$

The average current becomes

$$I_{sh} = m_{pu} * \frac{m_{ch} * V_0}{m_{pu} * R_{ch} + m_{ch} * R_{pu}} \tag{8.104}$$

The output resistance of the charge pump calculates as

$$R_{cp} = \frac{R_{ch}}{m_{ch}} + \frac{R_{pu}}{m_{pu}} \tag{8.105}$$

Note: This equation applies to very big capacitors with a change of the capacitor voltage far below V_0.

8.3.3 Single frequency aproximation

Basic idea is to compose the absolute value of the resistance from the real part neglecting the capacitor and the imaginary part comming from the impedance of the capacitor at the charge pump frequency. This is an aproximation too because we neglect the contribution of higher harmonics but it is a reasonable approach in the range none of the first two aproximations can be used.

$$R_{cp} = \sqrt{Re(R_{cp})^2 + IM(R_{cp})^2} \tag{8.106}$$

As for the high capacity aproximation we take

$$Re(R_{cp}) = \frac{R_{ch}}{m_{ch}} + \frac{R_{pu}}{m_{pu}} \tag{8.107}$$

For the imaginary part we only consider the capacitor. The peak to peak voltage of the sine wave must correspond the signal swing V_0. This means a current reduction of factor $2 * \sqrt{2}$ which equivalent with a multiplication of the imaginary resistance with this factor. Furthermore only 50% of the time a charge transfer takes place. This is taken into account by an other factor 2.

$$Im(R_{cp}) = \frac{2 * \sqrt{2}}{\Pi * f * C} \tag{8.108}$$

This leads to an output resistance of

$$R_{cp} = \sqrt{(\frac{R_{ch}}{m_{ch}} + \frac{R_{pu}}{m_{pu}})^2 + (\frac{2 * \sqrt{2}}{\Pi * f * C})^2} \tag{8.109}$$

This equation strictly applies if the pump is operated with a sine wave. Using it for other signal shapes however still leads to reasonable results.

8.3.4 Comparison of the 3 approximations shown

Approximation 1 is optimized for charge pumps with small capacitors so this approximation is expected to come close to the real behavior at low capacit values.

Approximation 2 is assuming a capacitor so large that the voltage across the pump capacitor can be assumed as constant.

Approximation 3 neglects the duty cycle dependence assuming a sine wave pumping. It is expected to come close to the real behavior of the charge pump for the complete range of capacities but in the extreme regions approximations 1 or 2 are expected to be more accurat. In the following plot a power charge pump operating with charge duty cycles and pump duty cycles of 43% (7% dead time with all switches open) and a path resistance of 7.3 Ohm was calculated with all three approximations. The pump was assumed to operate at 500 kHz. Capacities were sweeped from 1nF to 1uF.

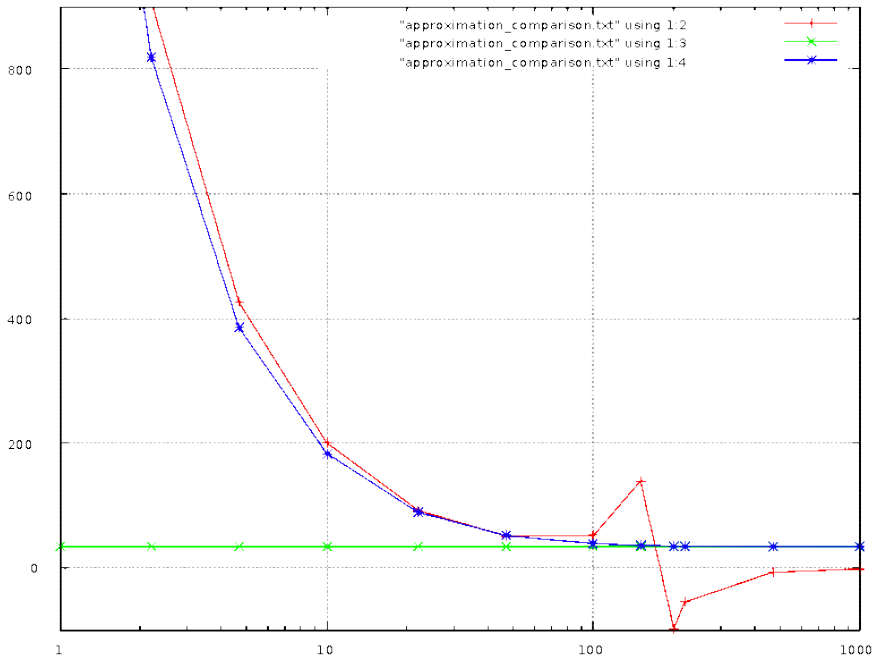


Figure 8.54: Comparison of the 3 approximations for the output resistance of a charge pump vs. pump capacity in nF

The red curve is the approximation for low capacity values.
The green curve shows the approximation for big capacities.
The blue curve shows the approximation using the sine wave concept.

The simple approximation assuming the voltage drop over the resistors can be neglected works nicely for low capacities (red curve). With increasing capacity the resistive losses get more significant and the approximation fails above 100nF.

Neglecting the change of the voltage stored in the capacitors works reasonably well for big capacitors above 100nF (green curve). But of course for low capacities it is definitely wrong.

Simply regarding only one frequency (as a sine wave) and assuming the voltage over the capacitor and the voltage over the resistor have to be added as vectors with 90 degree angle (blue curve) is always somewhat wrong (because it neglects the energy of the higher harmonics) but always close to reality. This approximation fits sufficiently (about +-10% accuracy) well over the whole frequency range and capacity range.

8.3.5 Practical designs of charge pumps

This section describes some basic examples of charge pumps. There are much more complex charge pumps existing or real devices. Often the motivation for the more complex designs ares RF emission or certain technology limitations that need to be circumvented. The simple designs shown are intended to show the concept of charge pumps.

Most simple: A single ended charge pump with bipolar diodes: This is the most simple design of a charge pump. Sometimes this circuit is found on board level implementations as well.

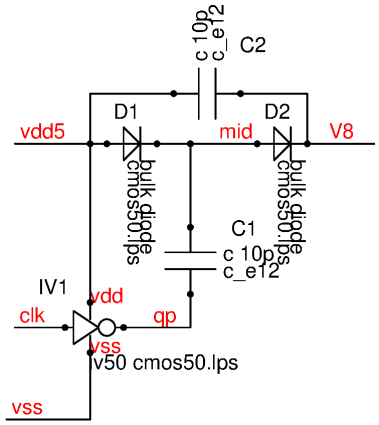


Figure 8.55: single stage charge pump with bipolar rectifiers

The most simple design. But it has it's draw backs:

- The bipolar diodes have a forward voltage in the range of 0.7V. So the achievable output voltage is about $V8 = 2 * vdd5 - 2 * V_f \approx 8.6V$ only.
- The frequency is limited by the reverse recovery time of the bipolar diodes.
- Only one phase is pumping. This leads to a high ripple of the output voltage V8.
- The bipolar diodes may have a parasitic vertical PNP transistor loosing some of the energy to the substrate.

The peak to peak ripple voltage of the single ended charge pump calculates as:

$$V_{ripplepp} = \frac{I_{load}}{C2 * f_{clk}} \tag{8.110}$$

First improvement: push pull to reduce the ripple: Using a charge pump with two inverse signals the ripple can be reduced at the same capacitor real estate as before. The capacitor C1 of the single ended design (10pF) is replaced by the capacitors C1 and C3 (2*4.7pF). The total power transfer is the same. Of course now we need 3 inverters and 4 diodes. But in most technologies these components don't consume much area.

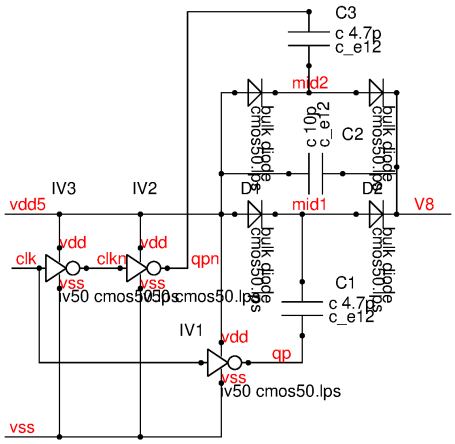


Figure 8.56: push pull charge pump to reduce the ripple voltage at the output

Using the push pull design the ripple voltage calculates as:

$$V_{ripplepp} = \frac{I_{load}}{2 * C2 * f_{clk}} \tag{8.111}$$

2nd Improvement: Using active rectifiers: To reduce the losses the bipolar diodes can be replaced by MOS transistors operating as synchronous rectifiers. This way we can in fact almost double the input voltage vdd5. The output voltage can get close to 10V now.

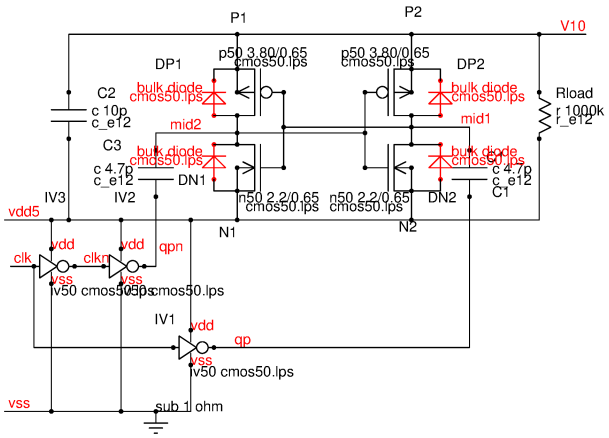


Figure 8.57: charge pump with active rectifiers

The concept is simple:

Phase 1: node qp rises, node qpn falls. N2 turns off, P2 turns on. The charge of C1 gets tranfered into C2.

Phase 2: node qp falls, node qpn rises: N1 turns off, P1 turns on. The charge of C3 gets transfered into C2.

Well, this looks great, but there is a little problem: As long as the voltage difference between V10 and vdd5 is below the threshold of the rectifier transistor this circuit won't work this way. During start up the rectifier transistors remain off. The energy transfer is running via the bipolar bulk diodes! This means the design is starting ith the losses of the bipolar components. The synchronous rectifiers take over as soon as the condition

$$V10 - vdd5 > V_{th}$$

is satisfied. In the example shown this condition is already satisfied after one clock period.

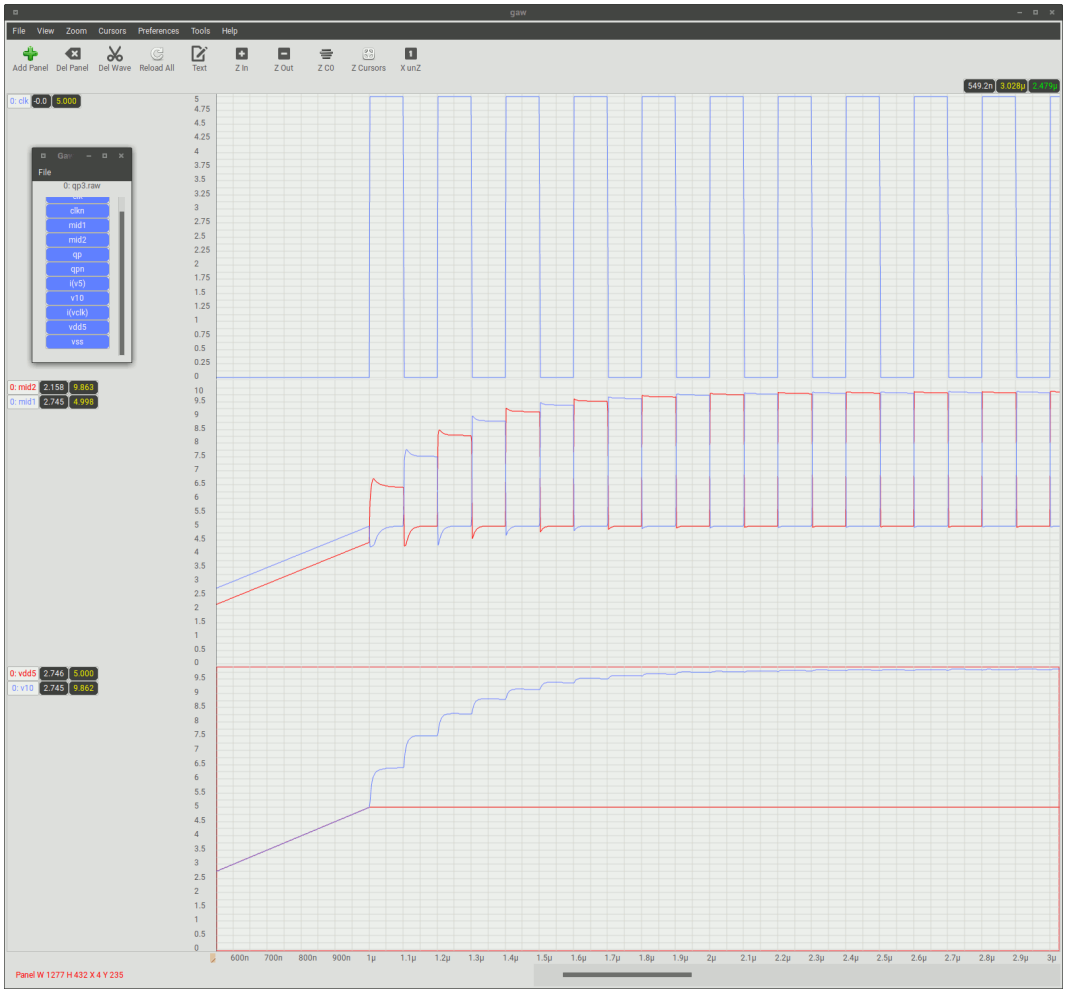


Figure 8.58: Simulation of the charge pump with synchronous rectifier

Beware of the parasitic transistors The circuit shown still looks nice - at least as long as we don't add the substrate PNPs and the vertical NPNs. But here comes trouble!

Every PMOS transistor has a P-drain sitting in an nwell over substrate. This means we have a vertical PNP transistor stealing current!

Every NMOS transistor has an N-drain acting as an emitter. The pwell becomes the base and the epi is the collector. An other component stealing current if the epi is biased from the output V10. This is extremely tempting because sharing the same epi pocket for the NMOS transistors and the PMOS transistors would save space. But placing the NMOS and the PMOS in the same epi pocket is absolutely fatal!

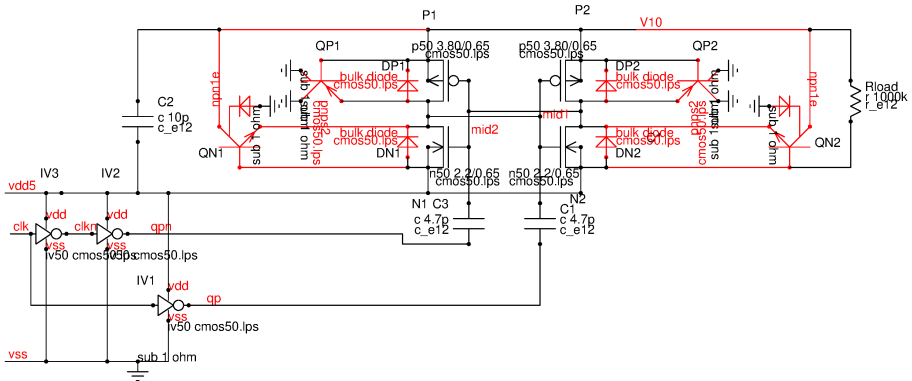


Figure 8.59: Fatal connection of parasitic bipolar transistors

The red colored parasitic bipolar components are active during start up. QP1 and QP2 dump part of the charge down to substrate. The amount of charge lost depends on the current gain B of QP1 and QP2. During start up the average current arriving at the source of P1 and P2 reduces to only:

$$I_{sP1P2} = f_{clk} * (C_3 + C_1) * (V(vdd5) - 2 * V_f)/B_{QN!QN2} \tag{8.112}$$

Very bad, but as long s the load is high resistive enough it still could work. To reduce the gain of the vertical PNP transistors using a burried layer is recommended (provided the technology offers this option).

The real killer is the parasitic NPN QN1 and QN2! Every time current flows through DN1 or DN2 the current is amplified by the current gain of the bipolar transistor. In other words: If the nwell is connected to the output the NPN transistors steal B-times the total charge transported through the diodes. The charge pump will never start!

Bugfix: To get rid of the parasitic NPN the nwell the NMOS transistors are embeded in MUST be connected to vdd5. No other components may be sitting in this nwell tub except N1 and N2. Some DRC tools might complain about possible latch up - ignore it. If nwell and pwell are shorted where should the latch up happen!

The second measure: Disconnect the load until the charge pump has reached Vth.

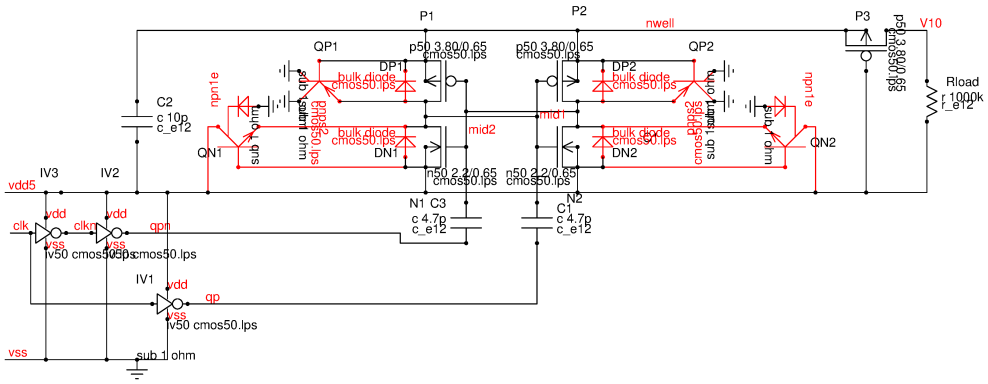


Figure 8.60: Bugfix to reduce the impact of the parasitic transistors

In the shematic above two improvements were done:

1. The nwell that creates the collector of the parasitic NPN transistors is connected to vdd5. Activation of QN1 and QN2 even helps now because the parasitic transistors make the bipolar rectifier lower resistive.

2. The load is disconnected from the charge pump by P3 until the charge pump reaches Vth above vdd5 at net nwell. Equation (8.81) still is valid, but since the load is disconnected there is no more base current into the vertical PNP transistors at start up. C2 simply gets charged with a reduced current during the first cycles of the clock, but it will eventually get charged.

Multi stage chargepumps: It is possible to build multi stage charge pumps. The basic concept is shown as a single ended design. The circuit first was used by Cockcroft and Walton [83] in 1932. In the Cockcroft voltage multiplier each capacitor and each diode only has to withstand Vin. The disadvantage is that the current available at the output decreases with every stage. Since this design is in use since the 1930s the following drawing uses valves acting as diodes.

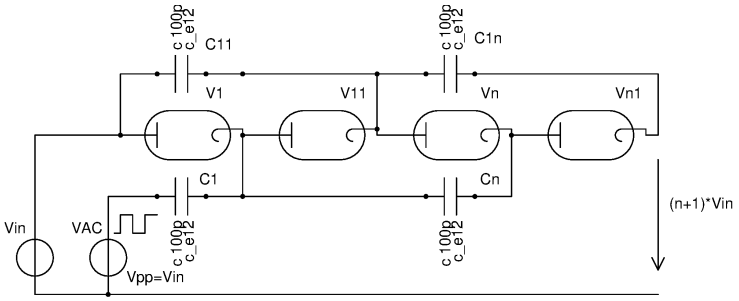


Figure 8.61: Cockcroft voltage multiplier

Dickson Charge pump The Dickson charge pump is a variation of the Cockcroft voltage multiplier that can be used if the capacitors can handle up to n*Vin. This way he capacitors are not in series anymore and the current achievable is higher. If this is really an advantage is debatable because the thinner the oxide the higher the capacity per area. Inversely we can state the higher the voltage capability the thicker the oxide and the lower the capacity per area will be. At the end of the day we probably end up at the same area using the Cockcroft variant with low voltage capacitors or using the Dickson variant with high voltage capacitors. The capacitors Cn and C1n of the Dickson charge pump must be able to withstand n*Vin as a long term DC stress.

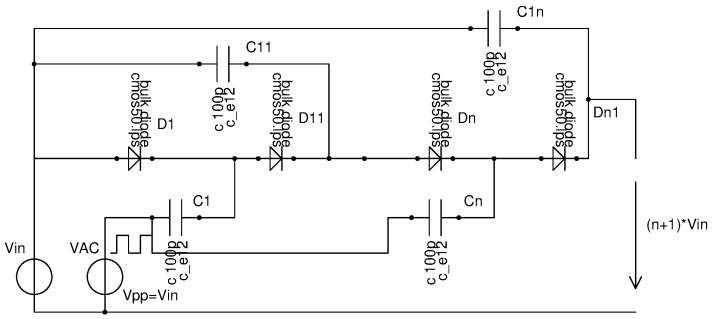


Figure 8.62: Dickson charge pump

In spite of the similarity of the Cockcroft multiplier and the Dickson charge pump the Dickson charge pump was patented in 1980 (US4214174A).

Both, the Dickson approach as well as the Cockcroft approach can of course be used in a push pull configuration as well. In fact the single stage pump with the synchronous MOS rectifiers shown before can be regarded as a single stage push pull Dickson or Cockroft topology.

8.4 Switchmode power supplies

Looking at switchmode power supplies most people immediately think of buck converters and boost converters. In reality the number of possible topologies is much bigger and and the differences of operating principles is much more fundamental. It begins with the question:

Do we want to store energy in the magnetic field or do we want to transfer it to the output right away?

If we store energy for instance in a coil we are looking at the family of flyback converters. The current flowing produces a magnetic flux that must be accommodate by the magnetic material. The amount of energy stored between the charging phase and the discharging phase defines the requirements of the magnetic material involved. Thus the power of the converter is defined by the frequency and the amount of energy the magnetic material can store.

If we are not willing to store energy in the magnetic material but transfer it to the output immediately we are looking at a forward converter. In this topology the cost and weight of the magnetic material can be reduced to a certain extent.

Flyback topologies are very flexible. If a different output voltage is needed we just have to pump a different amount of energy into the magnetic material. In the discharge phase this energy will be transferred to the output increasing the output voltage until the coil discharges. Since the energy will be transferred independently of the output voltage these converters very often are simply built just using a coil, a switch and a diode (or a second switch).

Forward converters are less flexible. The output can only take over the energy provided at the input if input voltage and output voltage match. Usually these converters use transformers. The output voltage is more or less defined by the ratio of the winding numbers. As a consequence forward converters are specialized on applications with a stable input voltage and a stable output voltage. Furthermore the savings made by the core material are absorbed to a certain extent by the cost of having two windings.

Table 41: Comparison of forwrd and flyback converters

	forward converter	flyback converter
advantage	cheaper magnetic material	flexible, simple coil
disadvantage	fixed voltage ratio, several windings	more expensive magnetic material

Summarizing: forward converters are more complicated to design. There is no general cooking recipe. Flyback converters are more simple. Certain topologies almost follow a cook book approach.

8.4.1 Forward converter

The most simple forward converter is a simple transformer power supply. As long as there is no load there is only a little input current flowing. Ideally (neglecting losses) the voltage ratio is defined by the number of windings.

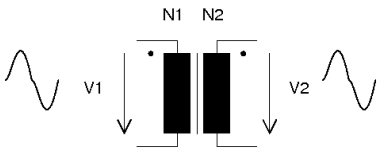


Figure 8.63: Transformer

$$\frac{V1}{V2} = \frac{N1}{N2} \tag{8.113}$$

Since the power must be the same on both sides we also get:

$$\frac{I1}{I2} = \frac{N2}{N1} \tag{8.114}$$

A practical driver circuit for a resistive load could look like this:

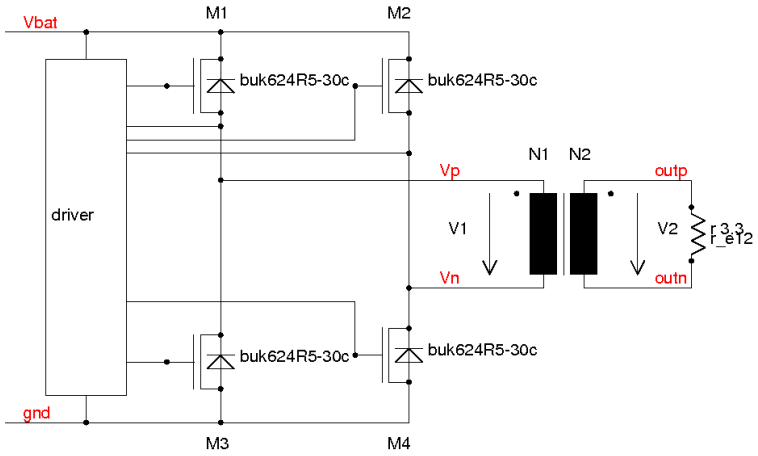


Figure 8.64: Driver of a forward converter in bridge configuration

Care must be taken that the circuit does not build up a DC current flowing in the transformer because this would lead to a loss of efficiency. So the ON-time of M1 and M4 must be (as an average) equal to the ON-time of M2 and M3. The following figure shows an example pulse diagram.

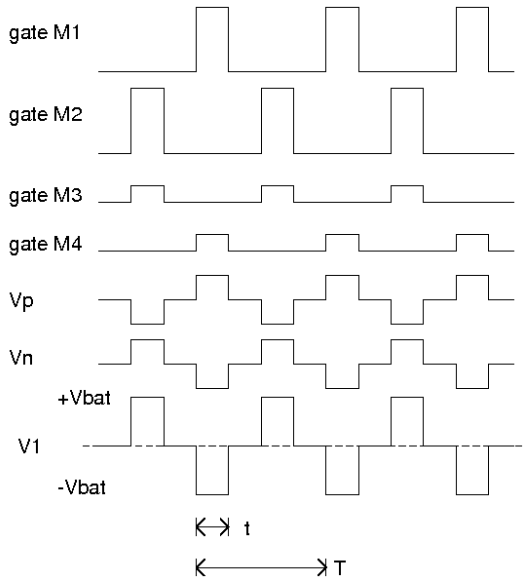


Figure 8.65: Typical pulse diagram of the converter shown above

The duty cycle of the converter calculates as:

$$D = \frac{2 * t}{T} \tag{8.115}$$

The effective voltage of V1 becomes:

$$V1_{eff} = D * Vbat \tag{8.116}$$

The transformer converts this to the effective voltage at the load:

$$V2_{eff} = \frac{N2 * D * Vbat}{N1} \tag{8.117}$$

The pulses may not overlap. So the condition

$$t < 0.5 * T \tag{8.118}$$

must be satisfied. (otherwise we get cross conduction)

This kind of circuit works nicely as long as we are dealing with resistive loads and as long as we are mainly interested in the average power transferred. Typical applications are heating systems.

The signals shown above are a bit too ideal. During the states in which all switches are off we have a more or less floating node system. During that time the voltage is defined by the flyback currents of the transformer and the stray capacities of the transistors. So we must expect ringing there in stead of seeing a step as shown in the theoretical drawings above. To better understand what happens we must have some understanding of a transformer.

Transformer The transformer consists of two windings that are coupled by the magnetic field. The windings are called primary windings (usually the input) and secondary windings (usually the output). If the opposite side is left open we can measure the open load inductances L1 and L2. The relationship to the transformation factor is:

$$N = \frac{N_2}{N_1} \tag{8.119}$$

$$N = \sqrt{\frac{L_2}{L_1}} \tag{8.120}$$

An ideal transformer has a coupling factor of 1. But of course this ideal component can not be manufactures. There always is a little bit of the magnetic field leaving the transformer. The part of the field outside of the transformer can be determined measuring the inductance while the opposite windings are shorted.

$$k = \sqrt{1 - \frac{L_{1short}}{L_{2short}}} \tag{8.121}$$

The real transformer can also be represented by an ideal transformer with a leakage inductance in series. This representation gives some more insight how the circuit would behave in case of short circuits or when the output is clamped.

$$L_{leak} = L_1 * (1 - k^2) \tag{8.122}$$

More sophisticated models of transformers even reflecting saturation of the core etc. can be found in [11] pages 364ff.

Let us assume we are operating our forward converter with a transformer having $L_1 = 160\mu H$ and $L_2 = 40\mu H$ with a coupling factor of $k = 0.99$. We operate the power supply with a frequency of 250kHz and a supply voltage of 12V.

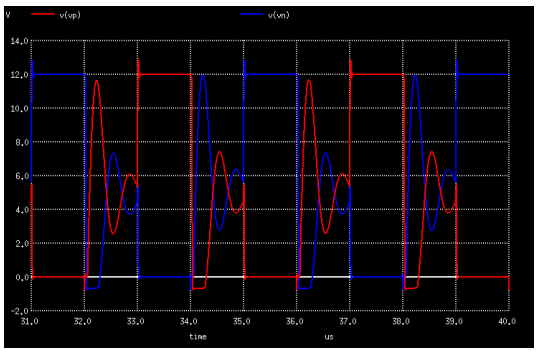


Figure 8.66: Primary voltages of a real converter with ringing in the high impedance states

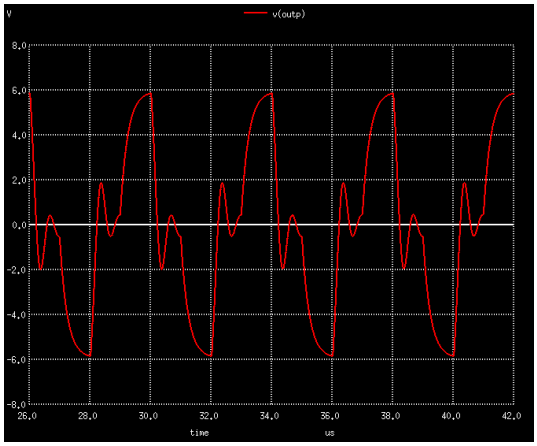


Figure 8.67: Secondary voltage of a real forward converter with resistive load

As soon as we have nonlinear or capacitive loads we need something limiting the current without producing excessive heat. Usually this current limiting device is inductive. The inductance can either be the stray inductance of the transformer itself or an inductance in series with the load.

8.4.2 Drivers for Fluorescent lamps

A fluorescent lamp has a high ignition voltage, but once it has started is has a negative resistance. To limit the current the transformer intentionally is designed with loose coupling ([33] page 69ff). As long as the lamp didn't start the converter provides a high output voltage and the preheating windings provide something like 6V. When the lamp starts the output voltage of the converter decreases due to the loose coupling of the transformer. At the same time the cathode heating drops as well to reduce the stress of the cathodes.

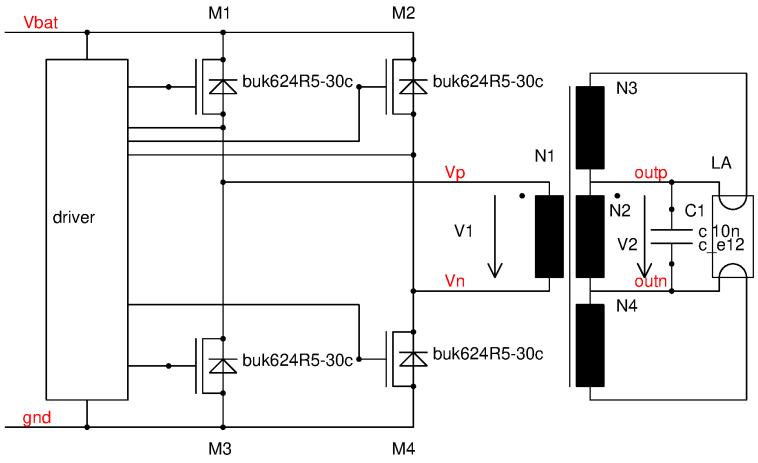


Figure 8.68: Fluorescent lamp driver using a full bridge

The circuit intentionally relies on the loose coupling of windings N1 and the windings N2, N3, N4. The coupling between windings N2 to N4 must be tight to achieve the desired drop of the heating voltage once the lamp has ignited. Capacitor C1 provides a sufficiently high peak current to support ignition of the lamp the first time the ignition voltage is reached. (Without C1 the lamp would start flickering until the complete gas volume is warm enough to ignite. This would wear out the lamp.)

Building a full bridge for such a simple application is more expensive than necessary. Replacing one side of the bridge by an AC coupling reduces cost and even an asymmetrical drive doesn't harm anymore (no more DC component depending on the duty cycle). Ideally the driver circuit should drive the load at the resonant frequency to take benefit of the additional voltage available to faster turn on of the lamp.

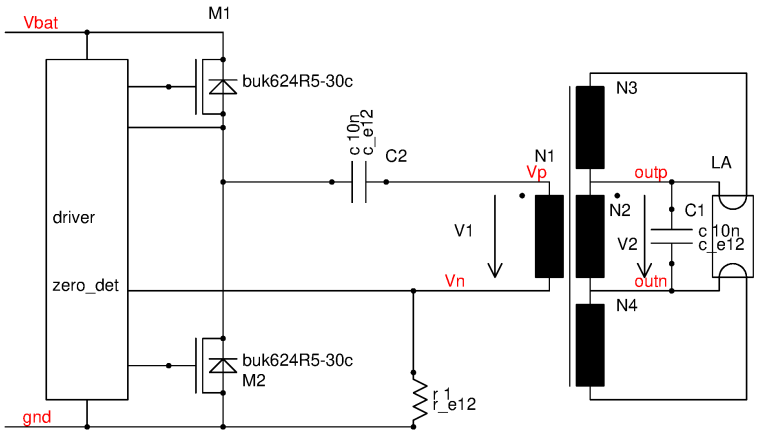


Figure 8.69: Fluorescent lamp driver using a half bridge

If the resonance is hit well the circuit switches at zero current which reduces the switching losses. Switching at zero current allows using slow switches such as IGBTs or even bipolar transistors.

8.4.3 Zero Voltage Switching (ZVS)

With increasing supply voltage the energy stored in the parasitic capacity of the switches starts to dominate the losses. At high supply voltages switching at zero voltage becomes more attractive than switching at zero current. Instead of detecting the zero crossing of the current now the control circuit has to detect when the drop over the power transistors is close to zero volt. In steady state the transistors are always switched on while the voltage drop is close to 0V. Turn off takes place at $V_{ds} \approx 0V$ automatically as long as the capacitors are big enough to prevent excessively fast voltage edges.

To reach the zero voltage switching condition a certain energy must always be present in the resonant tank C2, N1. For this reason the control circuit besides checking for the 0V condition also must monitor the current flowing in the primary side N1 of the transformer. Since the control circuit measures the current in the resonant tank anyway it becomes very attractive to build a current mode regulation. The following figure shows the concept of zero voltage switching.

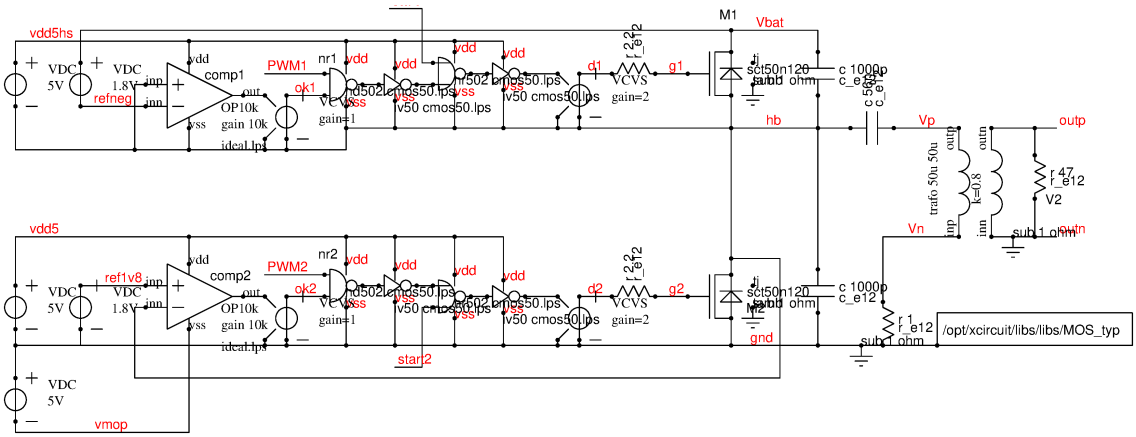


Figure 8.70: Zero Voltage switching half bridge

Comparators comp1 and comp2 monitor the voltage drop over the power transistors. The PWM signals are passed to the driver stage only if the voltage drop is below 1.8V. When the PWM turns off the current flowing in the tank circuit keeps flowing. To describe it in detail lets go through the following steps:

1. M1 is on
2. M1 turns off but M2 can't turn on because comp2 prevents it
3. The current flowing in the tank pulls down node hb. The slew rate is determined by the current and the 2 capacitors in parallel with the power stage.
4. The voltage at hb reaches 1.8V
5. comp2 allows turn on of M2

6. PWM2 turns off and M2 turns off
7. M1 is prevented from turning on by comp1
8. the current through the tank (that has reversed while M1 was on) pulls up node hb
9. hb reaches Vbat-1.8V and comp1 permits turning on M1...

The circuit has three problems:

- At start up the zero voltage condition isn't satisfied. For this reason there are the two signals start1 and start2 to start the circuit. The first pulse is not at zero voltage condition.
- The energy stored in the resonant tank may not completely be passed to the load. Some energy has to remain in the stray inductance. For this reason either the coupling factor of the inductor must be below about 0.8 or we need an auxiliary inductance in series with the transformer.
- The PWM frequency and the duty cycle can only be changed in a certain range that depends on the transformer, the capacitor values and the load resistance. This limits the regulation range. Often the regulation loop needs additional information such as the current flowing in the tank circuit. This usually is measured at node Vn.

The following screen shot shows the signal. Note that g2 turns on after the falling edge of node hb. (cursor position)

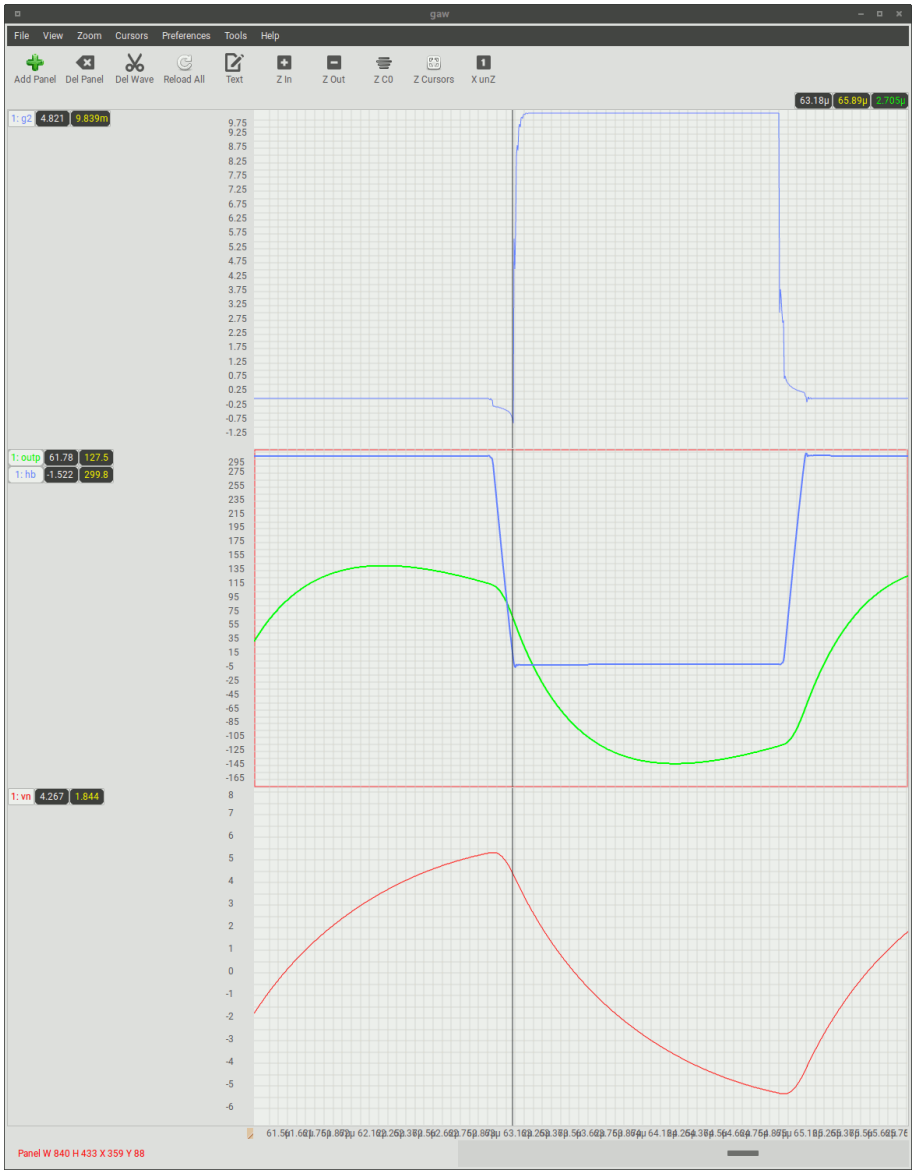


Figure 8.71: Steady state signals of the ZVS converter

At start up the ZVS (zero voltage switching) condition is violated. This is shown in the following plot.

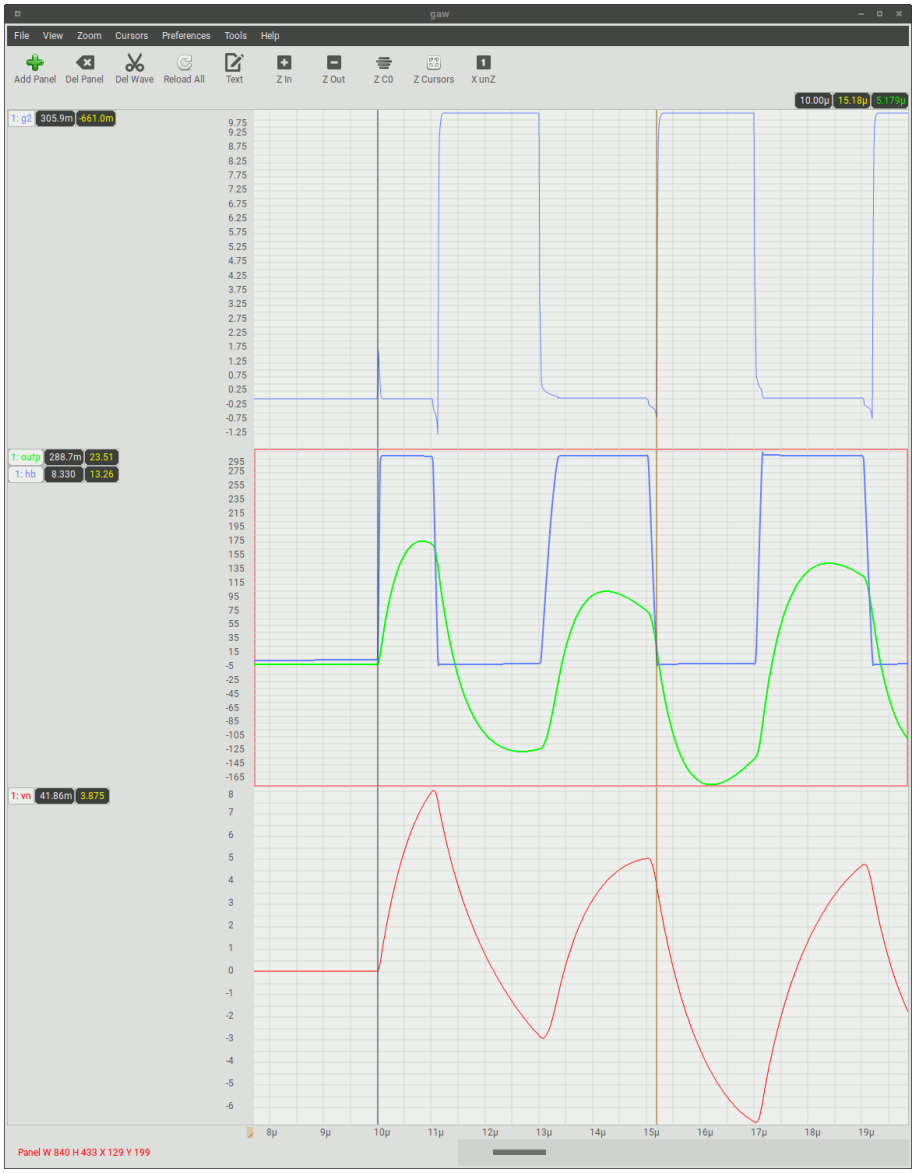


Figure 8.72: Start of the ZVS converter violates the zero voltage condition

Zero voltage switchmode converters can also be built as full bridge designs. This offers additional regulation modes shifting the phase between both half bridges.

Using parallel resonant tanks in stead of serial resonant tanks is possible as well, but may require additional inductors in series with the switches (The parallel resonance is the dual circuit of the serial resonance. Since we are swapping current and voltage we have to switch currents in stead of voltage then.)

8.4.4 Zero current switching (ZCS)

Zero current switching makes sense for power devices that either have slow turn off (e.g. IGBTs that have a current tail associated with the bipolar transistor) or devices that need a current zero crossing to turn off (e.g. Thyristors). Very often zero current switching is realized using quasi resonant approaches. In a quasi resonant converter design more or less sinusoidal current half waves are produced. Turn off takes place when the half wave reaches zero current again. The power is regulated changing the gap between the half waves. This means the frequency changes with the load current.

One big advantage of this discontinuous approach is the simple linear regulator behavior. The energy of each half wave simply is determined by the supply voltage, the inductance and the capacity. The power transferred becomes a linear function of the frequency.

The most important disadvantage is the poor ratio between peak current flowing during the sine half wave (resistive losses in the switches follow I^2) and the average current available in the load.

The following figure shows a little demonstrator for a thyristor quasi resonant converter to explain the concept in detail.

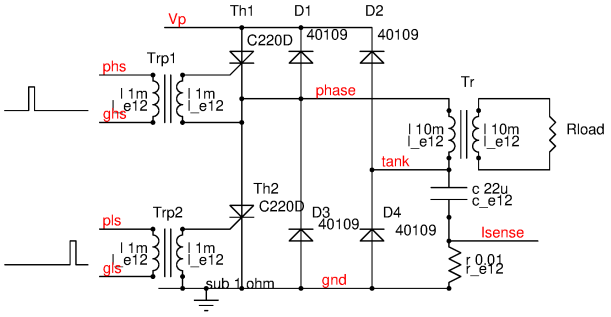


Figure 8.73: quasi resonant thyristor switchmode power supply

As a start condition let's assume node tank is discharged. The first trigger pulse at transformer Trp1 turns on thyristor Th1. Signal phase will be pulled up and current starts to flow through the power transformer Tr. The current increases until node tank reaches Vp. Diode D2 limits the voltage at node tank. Different from a real resonant switchmode power supply the 22uF capacitor can't be charged significantly higher than Vp. After one current half wave the node tank remains at the clipping voltage Vp+Vf and the current returns to zero. Th1 automatically turns off.

When Th1 is off Th2 can be turned on by a pulse at pls. Since the capacitor is charged the current now starts to flow from node tank through transformer Tr to the anode of Th2. Node tank discharges until D4 clamps the voltage at node tank at -Vf. The current decays depending on the back EMF and the inductance of transformer Tr. Again we only get one sinusoidal half wave but not a resonant continuation of the signal.

The following plot shows the signals operating the quasi resonant power supply with open load.

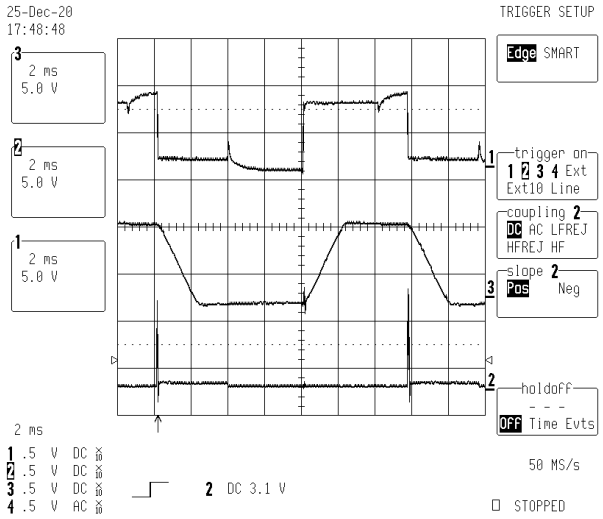


Figure 8.74: signals of the quasi resonant power supply operating with open load

Channel 1 (upper trace) shows the signal at node phase. The thyristors switch at zero current. Channel 2 (bottom trace) shows the trigger signal at the gate of pls. (this signal was used for triggering the oscilloscope). Channel 3 (middle trace) shows the signal at node tank. Here we clearly see that the resonant tank in stead of letting it oscillate freely is clamped at -Vf and Vp+Vf.

The following plot shows the same stage operated at full load. At full load the inductance of the transformer decreases dramatically. The frequency increases by about 2 magnitudes because with load the inductance of the resonant tank almost drops down to the stray inductance of the transformer!

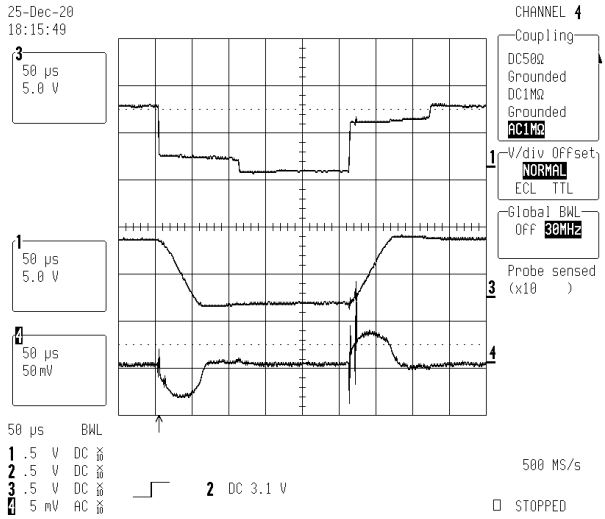


Figure 8.75: signals of the quasi resonant power supply at full load

Channel 1 (upper trace) shows the signal at node phase. Channel 3 (middle trace) shows the clamped node tank. Channel 4 (bottom trace) shows the current measured at node Isense.

The simplicity of the circuit (the thyristor automatically takes care of turning off at the end of each half wave) is tempting. A production quality implementation of the design is extremely difficult for the following reasons:

- If the second thyristor gets triggered while the first one is still conducting the supply Vp directly gets shorted to gnd via both thyristors. The only way to turn off again after a wrong triggering is disconnecting the supply Vp by a fast mechanical switch, a fuse (with special arc extinguisher) or a turn off thyristor.
- Even after zero crossing of the current a certain gap is needed to allow the minority carriers in the thyristors to recombine.
- Disconnecting the load increases the duration of the sine half wave from one cycle to the next. If the regulation circuit doesn't recognize load drop immediately the next pulse will come to early and the supply gets shorted.
- thyristors only allow a limited slew rate at the anode. This leads to additional snubber networks.

Real production quality implementation of a thyristor half bridge supplied from a DC supply is extremely difficult. For this reason thyristor resonant converters are only used in high voltage (above 1kV) , high power applications that justify the high protection effort.

GTOs (Gate Turn Off thyristors) are a possible way out but pulling the holes holes (positive charges) out of the gate requires a strong negative gate drive pulse in the range of several A. This makes the driver circuit significantly more complex.

Below 1kV usually switches that can be turned off by the drive circuit are used (mainly IGBTs because here the current tail justifies zero current switching). Using IGBTs turning off while the current is flowing leads to power dissipation but at least it doesn't lead to immediate destruction.

Table 42: Voltage classes and typical converter topologies

voltage	power device	comment
<1kV	IGBT	false trigger can be intercepted with standard driver
1kV..10kV	GTO	complex emergency off circuit able to drive negative gate current
1kV..10kV	thyristor	emergency turn off by circuit breaker (mechanical breaker of parallel turn off thyristor)
>10kV	thyristor stacks	emergency turn off by circuit breaker (mechanical or parallel turn off thyristor, capacitive voltage balancing required, driver speed must be matched)

The following figure shows the concept of using an emergency turn off thyristor.

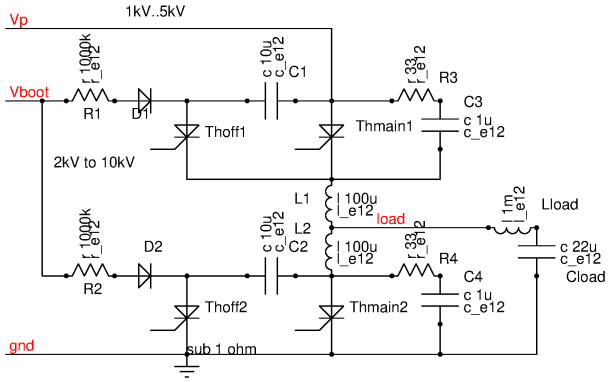


Figure 8.76: half bridge with turn off thyristors

The main switching path uses thyristors Thmain1 and Thmain2. Vboot is the bootstrap supply charging C1 and C2 to always bias the anodes of Thoff1 and Thoff2 with a voltage higher than Vp. Resistors R1 and R2 must be high resistive to keep the DC current through Thoff1 and Thoff below the hold current.

If a short circuit between node load and Vp or gnd takes place L1 and L2 limit the current. At short detection the corresponding turn off thyristor gets triggered and completely drains the current for a time long enough to turn off the main thyristor. When the turn off thyristor turns off again the slew rate at the anodes and the peak voltage must be limited by the snubber networks R3, C3 and R4, C4 to prevent parasitic turn on of the main thyristors due to extreme dV/dt caused by the current still flowing in L1 or L2. (In some cases diodes from Thmain1 cathode to gnd and from Thmain2 to Vp will do the job as well.)

Since the turn off thyristors Thoff1 and Thoff2 don't conduct periodically like the main thyristors the usually can be designed smaller than the main thyristors.

Well, enough of using thyristors. Better stick with IGBTs as long as you can! It makes life MUCH easier.

8.4.5 Flyback converters

Flyback converters intentionally store energy in the magnetic field of the inductor. They don't necessarily rely on a transformer ratio like forward converters. The most simple converters of the flyback family are the classical buck and boost converters. In the case of classical buck and boost converters the input inductor and the output inductor simply are coincident.

In flyback converters at least part of the energy is getting stored in one switch phase and released in the other phase. (Well, we did that in the ZVS and ZCS converter too, but there we only used the energy stored to discharge the parasitic capacities to reduce switching losses but we didn't transfer the stored energy to the output like in the flyback topologies)

Flyback converters are more flexible (wider supply range, wider load range, wider output voltage range independent of a transformer ratio.) and easier to control than forward converters .Very often the current flowing in the inductor can be approximated as a constant current making the math a lot easier. On the other hand the magnetic components are more expensive because the magnetic core must store the energy in the magnetic field without saturating for half of the period of the pulse width modulation.

Flyback converter topologies are very typical for medium loads up to some hundred Watts.

8.4.6 Buck Converter (Step down)

The buck converter is the most familiar switchmode power supply topology used to produce an output voltage that is lower than the input voltage. In a buck converter the coil may not saturate. This means the core of the inductor must store the energy needed during the flyback phase in the magnetic field. The buck converter belongs to the family of the flyback converters.

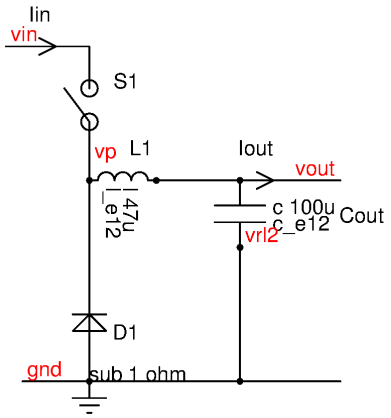


Figure 8.77: Buck converter

Continuous Current Mode operation (CCM): In continuous current mode the current through the inductor never reaches 0. As a simplification the current is assumed to be constant. This simplification makes the calculation of a buck converter a lot easier.

This simplification is justified for high load currents. At low load currents the calculation gets a bit more complex. This is described one paragraph late.

The energy taken from the power supply is:

$$P_{in} = D * V_{in} * I_{in} \tag{8.123}$$

D is the duty cycle of the switch S1. Vin is the supply voltage applied between pin vin and gnd. I_{in} is the pulsed current flowing through the switch (We assume continuous mode with a negligible change of the current trough inductor L1).

The power delivered to the load is:

$$P_{out} = V_{out} * I_{in} \tag{8.124}$$

(During the time S1 is open the current is supplied by the flyback diode D1.) This way duty cycle of the ideal lossless converter calculates as:

$$P_{in} = P_{out} \tag{8.125}$$

$$D = \frac{V_{out}}{V_{in}} \tag{8.126}$$

The DC current consumption (average current flowing into the converter becomes:

$$I_{in_{av}} = D * I_{out} = I_{out} * \frac{V_{out}}{V_{in}} \tag{8.127}$$

Well, of course a duty cycle of more that 100% is not possible. So the output voltage of a buck converter will always be less than the input voltage.

In practical designs we have to take into account the losses in the switches and diodes. Usually the designs are done in a way that resistive (static) losses dominate. So to start we neglect dynamic losses (taking place in the switching slopes)

$$P_{in} = P_{sw} + P_{diode} + P_{out} + P_{coil} \tag{8.128}$$

The losses of the switch calculate as:

$$P_{sw} = D * I * V_{sw} \tag{8.129}$$

In the same way the diode losses calculate as:

$$P_{diode} = (1 - D) * I * V_{diode} \tag{8.130}$$

The resistive losses of the coil calculate as:

$$P_{coil} = I^2 * R_{coil} \tag{8.131}$$

This leads to:

$$D * I * V_{in} = D * I * V_{sw} + (1 - D) * I * V_{diode} + I * V_{out} + I^2 * R_{coil} \tag{8.132}$$

The output voltage of a lossy buck converter becomes:

$$V_{out} = D * (V_{in} - V_{sw}) - (1 - D) * V_{diode} - I * R_{coil} \tag{8.133}$$

Solving the equation for D we get:

$$D = \frac{V_{out} + V_{diode} + I * R_{coil}}{V_{in} - V_{sw} + V_{diode}} \tag{8.134}$$

Power dissipation and efficiency in CCM:

The losses of a switchmode power supply consists of the biasing losses of the driver stage Pbias, the losses of the switch and the inductor and the losses of the rectifier. Additionally there are the losses caused by charging and discharging the switching node.

$$P_{loss} = P_{dyn} + P_{static} \tag{8.135}$$

The static losses are caused by the currents flowing and the voltage drops over the switch and the rectifier.

$$P_{static} = D * I * V_{sw} + (1 - D) * I * V_{diode} \tag{8.136}$$

If the switch is a MOS transistor and the rectifier is a synchronous rectifier using an other MOS transistor as a switch the voltage drops over these devices is proportional to the current. The equation for the losses becomes:

$$P_{static} = I^2 * (R_{coil} + D * R_{sw} + (1 - D) * R_{diode}) \tag{8.137}$$

Note that at low duty cycles the voltage drop over the diode is the dominant source of losses! For low output voltages the use of synchronous rectifiers with very low ON-resistance is recommended.

The dynamic losses depend on many more factors:

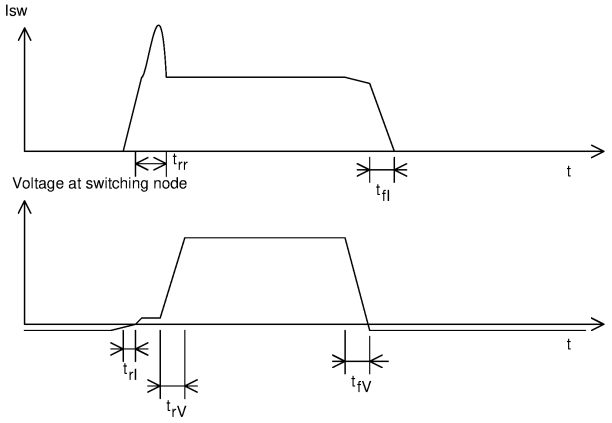


Figure 8.78: Delay times causing dynamic losses

1. The reverse recovery time of the diode. This loss is caused by the diode but the dissipation is inside the switch because usually the switch is the current limiting element. The current I_{on} is the current flowing in the inductor when the switch turns on. $I_{rr}$is the current flowing in the diode during reverse recovery.

   $$P_{rr} \approx f * I * V_{in} * (I_{on} + I_{rr}) \tag{8.138}$$

   Usually the current flowing in the diode before turn off and the reverse recovery current are in the same magnitude. Often the approximation

   $$P_{rr} \approx 2 * f * t_{rr} * V_{in} * I_{on}$$

   is used.

2. The time t_{rI} needed to build up the current in the switch until it reaches the load current .

   $$P_{rI} = \frac{1}{2} * f * t_{rI} * V_{in} * I_{on} \tag{8.139}$$

3. The time t_{rV} needed for the rising edge of the voltage.

   $$P_{rV} = \frac{1}{2} * f * t_{rV} * V_{in} * I_{on} \tag{8.140}$$

4. The time t_{fV} needed for the falling edge of the voltage while the current through the load is I_{off}.

$$P_{fV} = \frac{1}{2} * f * t_{fV} * V_{in} * I_{off} \tag{8.141}$$

5. The time t_{fI} needed to bring the current through the switch down to 0A.

$$P_{fI} = \frac{1}{2} * f * t_{fI} * V_{in} * I_{off} \tag{8.142}$$

6. Charge losses of the capacity of the switching node.

$$P_{csw} = \frac{1}{2} * f * C_{sw} * V_s^2 \tag{8.143}$$

This leads to a sum of the dynamic losses.

$$P_{dyn} = P_{rr} + P_{rI} + P_{rV} + P_{fI} + P_{fV} + P_{csw} \tag{8.144}$$

PWM gain in CCM: The regulation loop in the most simple case consists of a feedback network, an error amplifier, a duty cycle generator, the power stage and the external inductor and filter capacitor. The following figure shows the regulation loop of a straight forward voltage mode regulator.

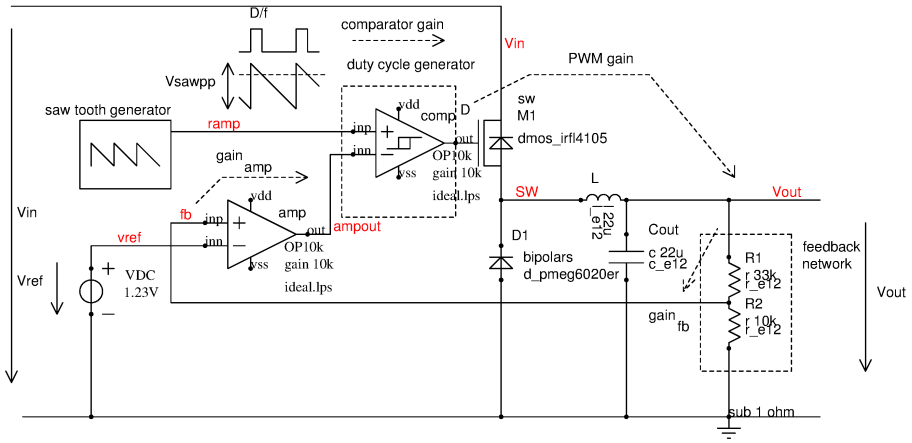


Figure 8.79: Voltage mode switchmode regulator

To understand the regulation loop we need to understand the change of Vout as a function of the duty cycle. So we have to rearrange the equation of the duty cycle and derivate it.

$$gain_{PWM_{CCM}} = \frac{dV_{out}}{dD} = V_{in} - V_{sw} + V_{diode} \tag{8.145}$$

First order the gain of the PWM simply is proportional to the drop of the regulator. The higher the supply the higher the gain!

Comparator gain: The duty cycle D is produced by a saw toot signal and a comparator. The peak to peak amplitude of the saw tooth signal is V_{sawpp} . The saw tooth signal is compared with the output signal of the regulator amplifier. If the output of the regulator amplifier changes by V_{sawpp} the duty cycle D changes from 0 to 100%. Thus the gain of the comparator stage is:

$$gain_{comp} = \frac{1}{V_{sawpp}} \tag{8.146}$$

regulator amplifier gain:

The regulator amplifier gain depends on the design of the regulator amplifier. Let's just assume the DC regulator amplifier of a voltage voltage mode switch mode power supply is a given number of $gain_{amp}$.

Feedback gain: The feedback network attenuates the output voltage of the regulator such that if the output voltage reaches the target the feedback voltage crosses the reference voltage. So the gain of the feedback network is

$$gain_{FB} = \frac{V_{ref}}{V_{out}} \tag{8.147}$$

Loop gain of a voltage mode switchmode power supply: The voltage mode switchmode power supply has a loop gain of

$$gain_{loop_{CCM}} = gain_{FB} * gain_{amp} * \frac{1}{V_{sawpp}} * gain_{PWM_{ccm}} \tag{8.148}$$

$$gain_{loop_{CCM}} = \frac{V_{ref}}{V_{out}} * gain_{amp} * \frac{V_{in} - V_{sw} + V_{diode}}{V_{sawpp}} \tag{8.149}$$

Regulator error: The error of the regulator can be calculated using the same equations as for the closed loop operation of amplifier.

$$Err_{rel} = \frac{1}{1 + \frac{V_{ref}}{V_{out}} * gain_{amp} * \frac{V_{in} - V_{sw} + V_{diode}}{V_{sawpp}}} \tag{8.150}$$

$$Err_{abs} = Err_{rel} * V_{out} = \frac{V_{out}}{1 + \frac{V_{ref}}{V_{out}} * gain_{amp} * \frac{V_{in} - V_{sw} + V_{diode}}{V_{sawpp}}} \tag{8.151}$$

Example: $V_{in} = 12V$, $V_{out} = 5V$, $V_{ref} = 1.23V$, $V_{sawpp} = 1V$, $gain_{amp} = 100$ leads to

$$Err_{abs} = \frac{5V}{1 + \frac{1.23V}{5V} * 100 * \frac{11V}{1V}} = 18.4mV$$

Output voltage ripple: The current flowing through the inductor is not completely constant. In stead the current has a saw tooth shape. Up to now this detail was neglected but to calculate the output ripple voltage of a buck converter we have to take a closer look. In the following it is assumed that resistive drop of the coil can be neglected. (This applies to reasonable converter designs). When the switch is closed the voltage over the inductor is:

$$V_L = V_{in} - V_{out} - V_{sw} \tag{8.152}$$

The voltage V_{in} is the supply voltage of the converter. $V_{out}is$the output voltage. $V_{sw}$is the voltage drop over the switch. The current builds up linearly.

$$\frac{dI_{swon}}{dt} = \frac{V_{in} - V_{out} - V_{sw}}{L} \tag{8.153}$$

In the same way the current decays when the switch is open and current recirculation runs via the rectifier.

$$\frac{dI_{swoff}}{dt} = -\frac{V_{out} + V_{rect}}{L} \tag{8.154}$$

The following figure shows the signals.

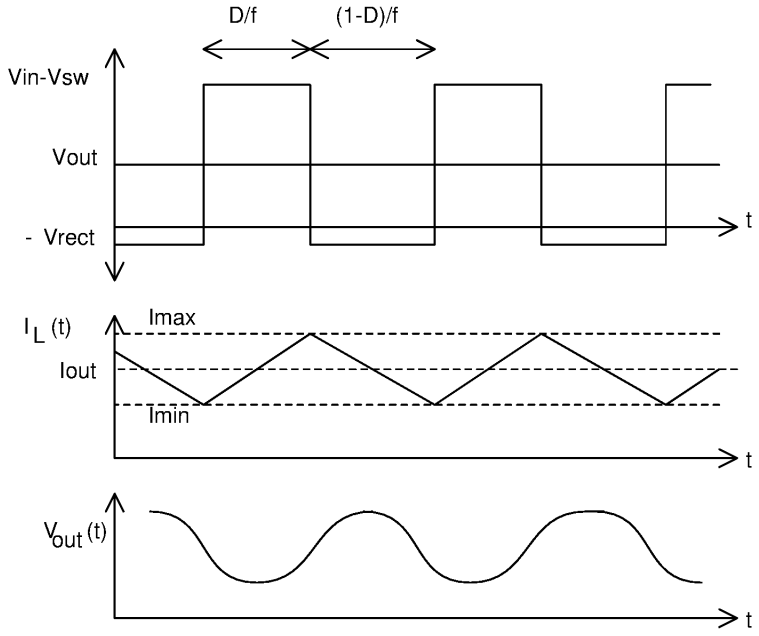


Figure 8.80: Signals of a buck converter in continuous mode

The voltage at the output capacitor (neglecting the ESR) can be calculated integrating over the difference between the current flowing through the coil and the output current.

$$V_{out}(t) = \frac{1}{C_{out}} * \int (I_L(t) - I_{out})dt \tag{8.155}$$

The minimum voltage of the ripple is reached in the middle of the time the switch is on because then the current flowing through the inductor crosses the load current. The maximum of the ripple is reached in the middle of the off tome of the switch. To calculate the peak to peak voltage of the ripple we just have to integrate the closed switch current from D/2f to D/f and the open switch current from 0 to (1-D)/2f. This leads to:

$$V_{ripple_{pp}} = \frac{1}{C_{out}} * (\int_{D/2f}^{D/f} (I_{swon} - I_{out})dt + \int_{0}^{(1-D)/2f} (I_{swoff} - I_{out})dt) \tag{8.156}$$

During switch on time the current difference is

$$I_{swon} - I_{out} = I_{charge1} = \frac{V_{in} - V_{out} - V_{sw}}{L} * t \tag{8.157}$$

with t running from 0 to D/2f. Doing the integration we get:

$$Q_{charge1} = \frac{V_{in} - V_{out} - V_{sw}}{L} * \frac{t^2}{2} \tag{8.158}$$

Since t runs to D/2f (from zero crossing to end of the pulse) the first charge becomes

$$Q_{charge1} = \frac{V_{in} - V_{out} - V_{sw}}{L * 8 * f^2} * D^2 \tag{8.159}$$

During switch off time we get

$$I_{swoff} - I_{out} = I_{charge2} = I_{max} - \frac{V_{out} + V_{rect}}{L} * t \tag{8.160}$$

with t running from 0 to (1-D)/f.

To get rid of the initially unknown I_{max} the integration simply uses swapped limits and starts at 0A (at t=(1-D)/2f) and ends at I_{max}(at t=0). (We don't care about signal shape. We only want to know the peak to peak voltage. So we only need the charge.)

$$Q_{charge2} = \frac{V_{out} + V_{rect}}{L} * \frac{t^2}{2} \tag{8.161}$$

with the limit for t=(1-D)/2f

$$Q_{charge2} = \frac{V_{out} + V_{rect}}{L * 8 * f^2} * (1 - D)^2 \tag{8.162}$$

$$Q_{charge2} = \frac{V_{out} + V_{rect}}{L * 8 * f^2} * (1 - 2D + D^2) \tag{8.163}$$

This leads to a - well more or less - nice expression for the ripple voltage

$$V_{ripple_{pp}} = \frac{1}{C_{out} * L * 8 * f^2} * [D^2 * V_{in} + (1 - 2D) * (V_{out} + V_{rect})] \tag{8.164}$$

If we neglect the drop of the switch and the drop of the rectifier the equations get more simple:

$$D = \frac{V_{out}}{V_{in}} \tag{8.165}$$

$$V_{ripple_{pp}} \approx \frac{V_{out}}{C_{out} * L * 8 * f^2} * (1 - D) \tag{8.166}$$

Imprtant note: This is the ripple voltage over an ideal capacitor without ESR.

Especially in high current application the equivalent series resistance of the capacitor may become dominant. The peak to peak ripple current flowing through the capacitor and the ESR can be calculated:

$$I_{ripple_{pp}} = (V_{out} + V_{diode}) * (1 - D) * \frac{1}{f * L} \tag{8.167}$$

This current is triangular shaped. The effective ripple current is

$$I_{ripple_{eff}} = \frac{I_{ripple_{pp}}}{12} \tag{8.168}$$

Now we can calculate the losses in the ESR

$$P_{ESR} = ESR * I^2_{ripple_{eff}} = ESR * [\frac{(V_{out} + V_{diode}) * (1 - D)}{12 * f * L}]^2 \tag{8.169}$$

The peak to peak ripple voltage over the ESR calculates as

$$V_{rippleESR_{pp}} = ESR * (V_{out} + V_{diode}) * (1 - D) * \frac{1}{f * L} \tag{8.170}$$

Due to the integrating behavior the voltage over the ESR and the ripple voltage over the ESR are shifted in phase by 90°. As a rough approximation the total ripple voltage can be estimated (regarding both voltages as vectors with 90° angle between them)

$$V_{ripple_{total}} \approx \sqrt{V_{ripple_{pp}}^2 + V_{rippleESR_{pp}}^2} \tag{8.171}$$

Discontinuous Current Mode operation (DCM): If the load current is too low (or the inductor has a low inductance) the switchmode power supply operation is in discontinuous current mode. In discontinuous current mode the current through the inductor is triangular shaped. Between the current triangles the current through the inductor becomes 0.

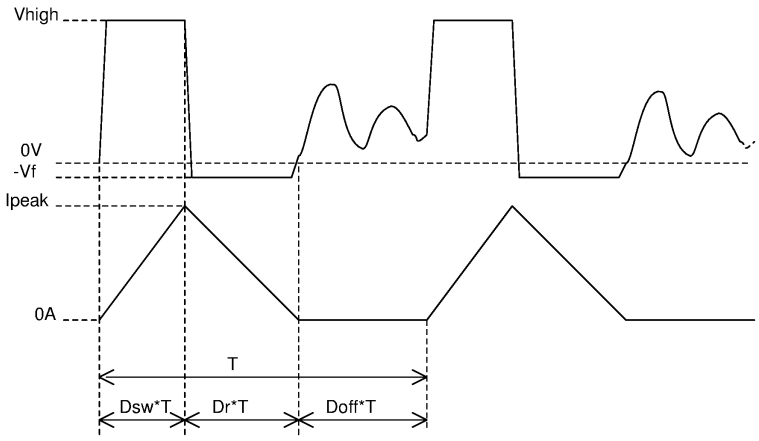


Figure 8.81: Signals of a buck converter operating in discontinuous mode

The period T of the signal is determined by the clock frequency of the switchmode power supply.

$$T = 1/f_{osc} \tag{8.172}$$

The signal consists of 3 phases. During Dsw the switch is on. The voltage at the switch becomes Vhigh. Neglecting the inductor resistance the current through the inductor increases to Ipeak.

When the switch turns off the rectifier takes over the current. During Dr the voltage at the inductor drops to -Vf, the forward voltage of the rectifier. (In discontinuous mode most switchmode power supplies don't use synchronous rectification because then the current can flow the reverse way through the rectifier creating additional losses it the rectifier remains on during Doff)

When the current reaches 0A the rectifier and the switch both are current less. At the switching node we will observe ringing depending on the inductance of the inductor and the parasitic capacity of the switch and the rectifier. After Doff the switch turns on again.

The sum of the duty cycles is 1.

$$D_{sw} + D_r + D_{off} = 1 \tag{8.173}$$

The output current in discontinuous mode calculates as:

$$I_{out} = \frac{I_{peak} * (D_{sw} + D_r)}{2} \tag{8.174}$$

The current slope during Dsw is determined by the difference between the input voltage and the output voltage of the switchmode power supply and by the inductor. So the peak current becomes:

$$I_{peak} = \frac{D_{sw} * (V_{in} - V_{out})}{f_{osc} * L} \tag{8.175}$$

The current decay during Dr depends on the output voltage, the rectifier voltage, the inductance and the clock frequency (we are assuming the voltage drop over the diode is more or less constant because we use a bipolar diode during discontinuous mode).

$$I_{peak} = \frac{D_r * (V_{out} + V_f)}{f_{osc} * L} \tag{8.176}$$

Having two equations for Ipeak we can find the relationship between Dsw and Dr.

$$D_r = D_{sw} * \frac{V_{in} - V_{out}}{V_{out} + V_f} \tag{8.177}$$

Keeping the diode drop in the equation we get:

$$D_{sw} = \sqrt{\frac{2 * L * I_{out} * (V_{out} + V_f)}{V_{in}^2 - V_{out} * (V_{in} + V_f) + V_{in} * V_f}} \tag{8.178}$$

If we neglect the drop of the rectifier (Vf=0) the duty cycle becomes:

$$D_{sw} \approx \sqrt{\frac{2 * L * I_{out} * V_{out} * f_{osc}}{V_{in} * (V_{in} - V_{out})}} \tag{8.179}$$

and the diode

$$D_r \approx \sqrt{\frac{2 * L * I_{out} * (V_{in} - V_{out}) * f_{osc}}{V_{in} * V_{out}}} \tag{8.180}$$

Usually this simple form of the equation is good enough (Later we will use this approximation to estimate the PWM gain because it is by far less complicated than the exact solution). But the lower the output voltage of the SMPS the more significant the impact of the diode forward voltage will be. For low output voltages 3V or less) the forward voltage of the rectifier should not be neglected anymore. So the equations become a bit more complex.

$$D_{sw} = \sqrt{\frac{2 * L * I_{out} * (V_f + V_{out}) * f_{osc}}{(V_{in} + V_f) * (V_{in} - V_{out})}} \tag{8.181}$$

and for the diode

$$D_r = \sqrt{\frac{2 * L * I_{out} * (V_{in} - V_{out}) * f_{osc}}{(V_{in} + V_f) * (V_{out} + V_f)}} \tag{8.182}$$

PWM gain in DCM: As a good enough approximation we will calculate the PWM gain from the approximation. The equation can be reordered:

$$V_{out} \approx \frac{D_{sw}^2 * V_{in}^2}{2 * L * I_{out} * f_{osc} + D_{sw}^2 * V_{in}} \tag{8.183}$$

Deriving Vout leads to the approximation of the PWM gain in discontinuous mode:

$$\frac{dV_{out}}{dD_{sw}} = \frac{4 * D_{sw} * V_{in}^2 * L * I_{out} * f_{osc}}{(2 * L * I_{out} * f_{osc} + D_{sw}^2 * V_{in})^2} \tag{8.184}$$

Not very intuitive. But some things can be seen:

1. At very low duty cycles D_{sw}^2 approaches 0 faster than the rest.

$$\frac{dV_{out}}{dD_{sw}} \rightarrow \frac{D_{sw} * V_{in}^2}{L * I_{out} * f_{osc}} \tag{8.185}$$

2. Since $D_{sw} \sim \sqrt{L * I_{out} * f_{osc}}$ the loop gain follows

$$\frac{dV_{out}}{dD_{sw}} \sim \frac{V_{in}^2}{\sqrt{L * I_{out} * f_{osc}}} \tag{8.186}$$

The lower the load current the higher the lop gain of the regulator in discontinuous mode!

gain of the duty cycle generator: Well, nothing new under the sun. We already know it from the continuous mode:

$$gain_{comp} = \frac{1}{V_{sawpp}} \tag{8.187}$$

amplifier gain: As in the continuous current calculation we regard it as a given $gain_{amp}$.

feedback gain: The feedback gain is determined by the ratio of the feedback divider. This is a function of V_{out} and V_{ref} .

$$gain_{fb} = \frac{V_{ref}}{V_{out}} \tag{8.188}$$

Loop gain in DCM: Multiplying all these numbers we get the loop gain in discontinuous mode:

$$gain_{loop_{DCM}} = \frac{4 * D_{sw} * V_{in}^2 * L * I_{out} * f_{osc}}{(2 * L * I_{out} * f + D_{sw}^2 * V_{in})^2} * \frac{V_{ref}}{V_{sawpp} * V_{out}} * gain_{amp} \tag{8.189}$$

Power dissipation and efficiency in DCM: Since the currents are triangular the losses in discontinuous mode calculate:

$$P_{loss} = 0.5 * C * f_{osc} * (V_{in}^2 - V_{out}^2) + P_{bias} + \frac{1}{3} * D_{sw} * I_{peak}^2 * R_{sw} + 0.5 * D_r * I_{peak} * V_f + \frac{1}{3} * (D_{sw} + D_r) * R_l * I_{peak}^2 \tag{8.190}$$

Due to the gap Doff the duty cycle in discontinuous mode always is smaller than the duty cycle in continuous mode. This allows a simple calculation algorithm using for instance octave: D=min(Dccm, Ddcm) with Ddcm being the result of the discontinuous mode calculation and Dccm being the result of the continuous current mode calculation.

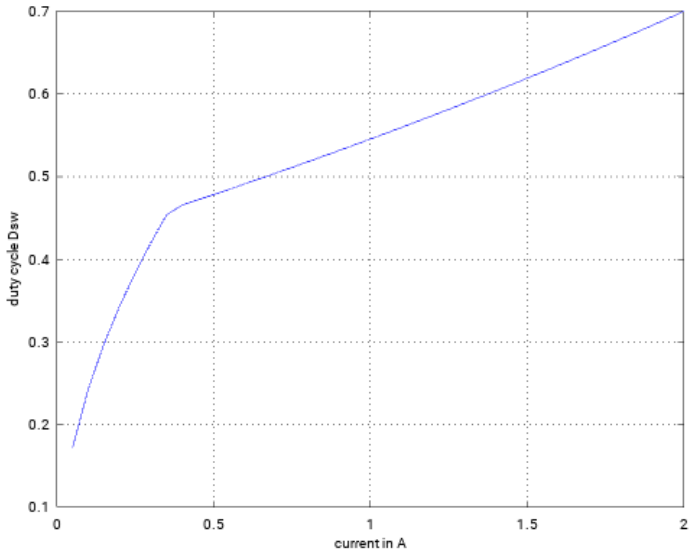


Figure 8.82: Duty cycle of a buck converter from 12V to 5V with only 10uH inductance. At the knee the duty cycle of the discontinuous mode calculation becomes bigger than the result of the continuous mode calculation. So left of the knee we are in DCM while right of the knee we are in CCM

Output voltage ripple in discontinuous current mode: The output capacitor gets charged during the on time of the switch (Dsw) and during the on time of the rectifier (Drect). The integration time however runs from the rising crossing of the inductor current and the load current and the falling crossing of the currents. The following diagram shows the integration to be done. The integration has to determine the dashed area.

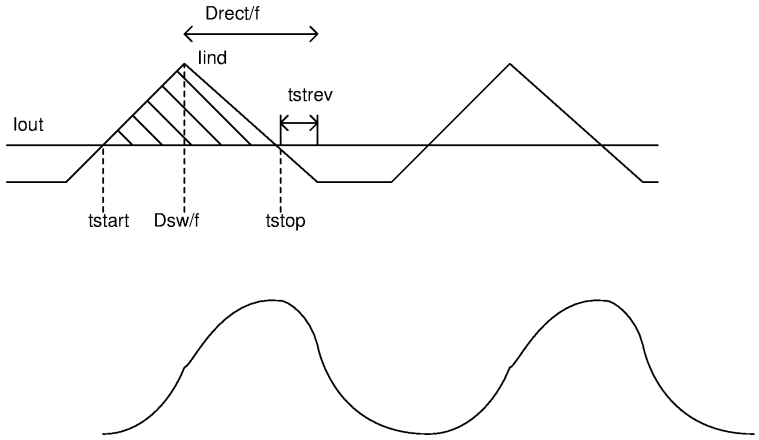


Figure 8.83: Current and ripple voltage in discontinuous mode with the limits to be used for the integration

At the same time the output current is discharging the capacitor. So we have to consider 2 charges. During switch on time the current through the inductor increases with a triangular shaped current.

$$I_L(t) = t * \frac{V_{in} - V_{out} - V_{sw}}{L} \tag{8.191}$$

During this time the capacitor gets discharged by I_{out}. Charging starts when $I_L(t)$ crosses I_{out}. Integration yields:

$$Q(t) = \frac{V_{in} - V_{out} - V_{sw}}{2 * L} * t^2 \tag{8.192}$$

To get the charge this current must be integrated from t_{start} to D_{sw}/f_{osc} . The starting point of the integration is

$$t_{start} = \frac{L * I_{out}}{V_{in} - V_{out} - V_{sw}} \tag{8.193}$$

Since it is a linear increase of the current we can simplify and just integrate a triangle from 0 to $D_{sw}/f - t_{start}$

$$Q_{sw} = \frac{V_{in} - V_{out} - V_{sw}}{2 * L * f_{osc}^2} * D_{sw}^2 - \frac{I_{out} * D_{sw}}{f} + \frac{L * I_{out}^2}{2 * (V_{in} - V_{out} - V_{sw})} \tag{8.194}$$

For the free wheeling part of the signal from D_{sw}/f to t_{stop} it is easier to reverse the borders and integrate a rising current from 0 to $D_{rect}/f - t_{srev}$.

$$t_{srev} = \frac{L * I_{out}}{V_{out} + V_f} \tag{8.195}$$

This way we get a very similar expression for the charge transfer during the free wheeling part of the signal.

$$Q_r = \frac{V_{out} + V_f}{2 * L * f_{osc}^2} * D_r^2 - \frac{I_{out} * D_r}{f} + \frac{L * I_{out}^2}{2 * (V_{out} + V_f)} \tag{8.196}$$

The ripple peak to peak voltage in discontinuous mode becomes

$$V_{ripple_{pp}} = \frac{1}{C_{out}} * [\frac{(V_{in} - V_{out} - V_{sw}) * D_{sw}^2 + (V_{out} + V_f) * D_r^2}{2 * L * f_{osc}^2} - \frac{I_{out}}{f} * (D_{sw} + D_r)$$
$$+ \frac{1}{2} * L * I_{out}^2 * (\frac{1}{V_{in} - V_{out} - V_{sw}} + \frac{1}{V_f + V_{out}})] \tag{8.197}$$

Neglecting the drop voltage of the switch and the rectifier and assuming a very low output current (deep discontinuous mode. The last terms $-\frac{I_{out}}{f} * (D_{sw} + D_r)$ and $L * I_{out}^2 * ...$ are neglected) the whole equation simplifies dramatically!

$$V_{ripple_{pp}} \approx \frac{1}{C_{out} * f_{osc}^2 * 2 * L} * [(V_{in} - V_{out}) * \frac{2 * L * I_{out} * V_{out} * f_{osc}}{V_{in} * (V_{in} - V_{out})} + V_{out} * \frac{2 * L * I_{out} * (V_{in} - V_{out}) * f_{osc}}{V_{in} * V_{out}}]$$

$$V_{ripple_{pp}} \approx \frac{I_{out}}{C_{out} * f_{osc}} \tag{8.198}$$

Well, not really a surprise. This is the case of a very short charging pulse and a very long time with the switch open and the diode being currentless. So this is simply the discharging of the capacitor C_{out} by the load current.

The equation doesn't look intuitive. Let's try a geometric approach. The current is almost triangular. The capacitor gets charged as soon an the current is higher than the output current. We simply find two triangles

charging the capacitor. The height is $I_{peak} - I_{out}$. The base line is $(D_{sw} + D_r) * (I_{peak} - I_{out}) * T/I_{peak}$. This leads to the area of the charging triangle.

$$Q_{charge} = \frac{(I_{peak} - I_{out}) * (D_{sw} + D_r) * (I_{peak} - I_{out})}{2 * f_{osc} * I_{peak}}$$

The ripple voltage now becomes:

$$V_{ripple_{pp}} = \frac{(I_{peak} - I_{out})^2 * (D_{sw} + D_r)}{2 * f_{osc} * I_{peak} * C_{out}} \tag{8.199}$$

Hysteretic buck converter: Since we have no restrictions of the duty cycle range even very simple regulator topologies such as hysteretic buck converters will work. Even more simple circuits have been published in [34, 36]. Probably the concept of hysteretic switchmode power supplies is even older.

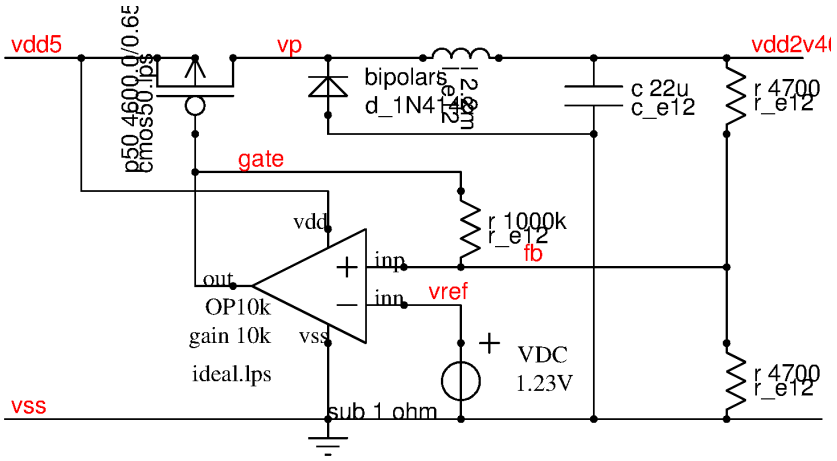


Figure 8.84: Hysteretic buck converter

Of course the output voltage has a voltage ripple depending on the hysteresis of the regulator. In the little example shown above the hysteresis is 5V*4.7K/(2*1000k)=12mV. Thus the output voltage will hoover around 2.46V +-2*Vhyst. The reason for the overshot of 2*Vhyst is the energy stored in the inductor when the schmitt trigger changes state. The oscillation frequency is determined by the inductor, the capacitor, the hysteresis and the load current. It may change dramatically if the load current changes!

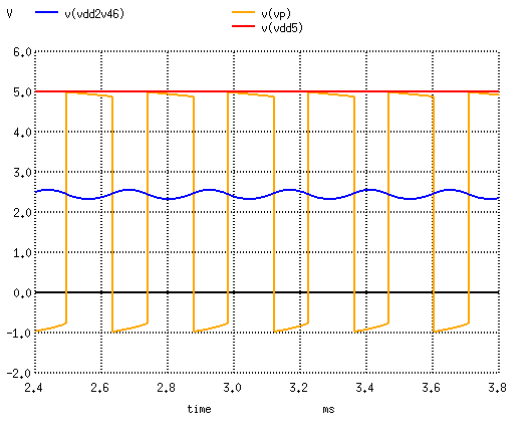


Figure 8.85: Simulation result of the hysteretic regulator operating at 100mA load current and 5V supply voltage

The voltage ripple of the output voltage of a hysteretic converter can not be reduced increasing the output capacity. This will only change the frequency. Furthermore hysteretic converters tend to reduce the frequency at low load leading to a wide range of discontinuous mode operation. Discontinuous mode leads to higher peak currents and reduce efficiency. Nevertheless for very simple requirements such a simple hysteretic regulator may already be sufficient.

Since the output ripple is inherent to the hysteretic regulator a quite nice solution for laboratory equipment is combining a hysteretic regulator acting as a prestabilizer and a linear regulator as a second stage to remove the ripple. In this combination the switchmode regulator reduces the power dissipation but we still get the fast load

regulation of the linear regulator. Efficiency of such an approach however is limited by the drop the linear regulator needs. (About 80% are common)

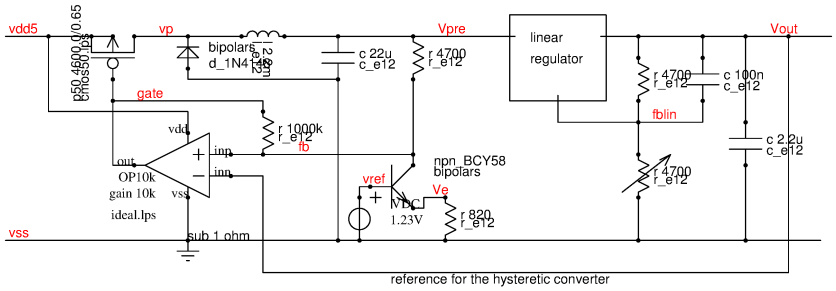


Figure 8.86: Example of an adjustable linear regulator with hysteretic SMPS as a prestabilizer

In the above figure the output voltage of the linear regulator is used as a reference for the hysteretic SMPS. The voltage drop over the resistor in the feedback path must match the dropout voltage required by the linear regulator. It is produced by a constant current sink.

8.4.7 Practical design considerations of a buck converter.

After all the theory let's have a look at the practical design considerations of a buck converter. Typical design targets are:

1. High efficiency, low cooling cost
2. Low cost of the inductor
3. Low cost of the capacitor
4. high accuracy and stability
5. Low output ripple voltage

Often these targets are conflicting!

For demonstration let's have a look at a buck converter with the following parameters in the typical parameters: $I_{bias} = 20mA$, $f_{osc} = 475kHz$, $L = 20\mu H$, $R_L = 0.2\Omega$, $C_{sw} = 100pF$, $R_{onsw} = 0.2\Omega$,$R_{onrect} = 0.2\Omega$, $V_f = 0.8V$ (in discontinuous mode).

Efficiency: Efficiency depends strongly on the load current, the Rdson of the switches, the resistive losses of the inductor and the frequency.

At low loads the dynamic losses (charging and discharging of the parasitic capacity of the switching node) and the biasing losses are dominant. Dynamic losses can be reduced by minimizing the capacity of the switching node. Static losses can be reduced using driver chips with low DC current consumption. Low DC current consumption usually limits the frequency the switchmode powersupply can be operated with. If it can be tolerated that the frequency changes with the load current it may be a good idea to reduce the frequency at low load to improve efficiency. This approach has it's limits when we start to enter deep discontinuous mode because then the peak currents will increase boosting resistive losses again.

At high load currents the resistive losses in the inductor, the switch and the rectifier become dominant. Therefore at high currents using a synchronous rectifier is recommended. The resistive losses of the inductor can be reduced using less windings. This however leads to a lower inductance and higher losses at low load current.

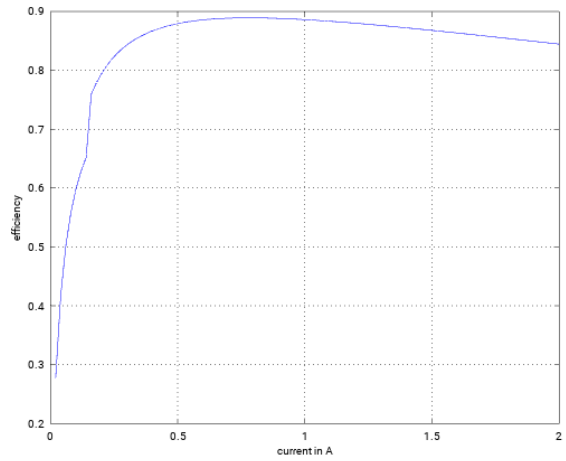


Figure 8.87: efficiency of our example SMPS versus load current

The efficiency is discontinuous at the transition between discontinuous mode and continuous mode at 130mA because the rectifier is an real diode rectifier n discontinuous mode to prevent current from flowing back to ground at low duty cycles. Above 0.7A the efficiency decreases again due to the resistive losses in the switch and the rectifier. Looking at the losses instead of the efficiency the limitations become much more obvious!

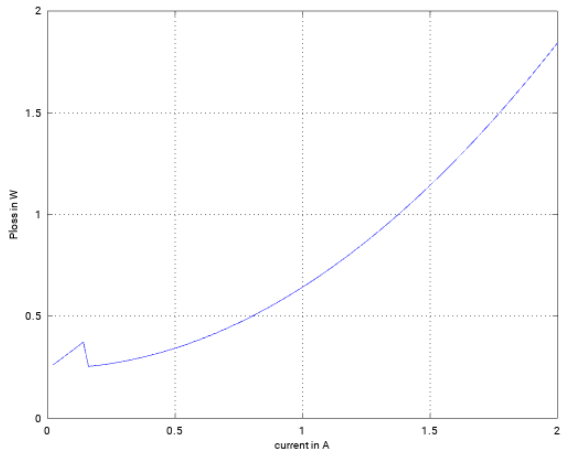


Figure 8.88: Losses of our example SMPS versus load current

The increase of the losses to values higher than about 1.5W (at about 1.7A) is a problem. Typical IC packages have thermal resistances in the range of 40K/W to 100K/W (depending on the package and the mounting on the board)

Low cost of the inductor: Optimizing the inductor for low cost usually means using a lower inductance. Lower inductance leads to higher peak currents and ripple in discontinuous mode. This effect can be compensated increasing the frequency. But an increase of the frequency leads to higher dynamic losses and may require a faster control chip that probably requires more bias current.

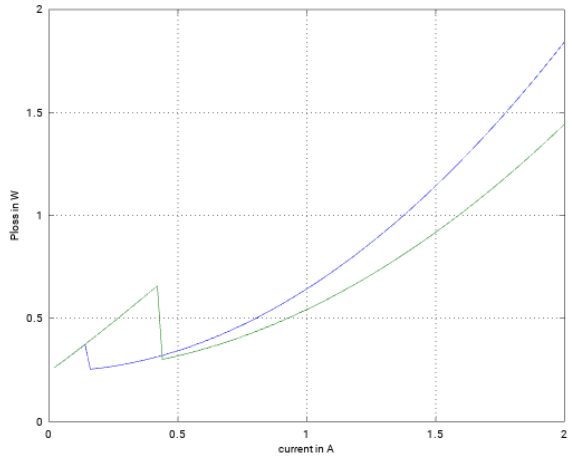


Figure 8.89: comparison of the losses of a switchmode power supply with $7\mu H$ inductor (green) and with $20\mu H$ inductor (blue)

The lower inductance leads to a wider discontinuous mode range and to more losses up to 400mA load current. Above 0.5A the $7\mu H$ inductor has less resistive losses than the $20\mu H$ inductor. The widening of the discontinuous mode range gets worst at high supply voltage because the higher the voltage drop over the inductor the faster the current builds up.

Low cost of the capacitor: Capacitors with high capacity are more expensive than capacitors with low capacity. Further more bulky capacitors require more effort for mechanical fixation on the board. For these reasons switchmode power supply designer are interested in lowering the value of the output capacitor. This however leads to a higher ripple voltage. Looking at out example power supply let's vary the capacitor from $2.2\mu F$ to $47\mu F$ and observe the resulting ripple voltage at a load current of 0.5A.

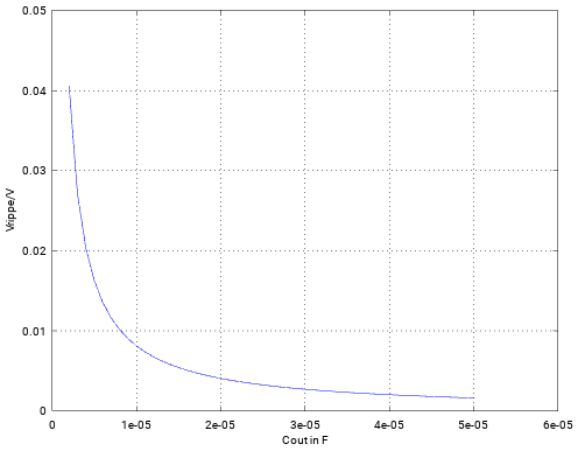


Figure 8.90: Peak to peak ripple voltage of a switchmode power supply versus Cout at 0.5A load current

To reduce the ripple voltage while keeping the capacitor value small switchmode power supply designers try to increase the frequency of the switchmode power supply. For low voltage applications frequencies up to 3MHz are not uncommon anymore. Above about 3MHz the technical problems (dynamic losses, power stage timing - e.g. for synchronous rectifiers, electromagnetic compatibility, parasitic inductance of the blocking capacitors...) become too difficult to handle.

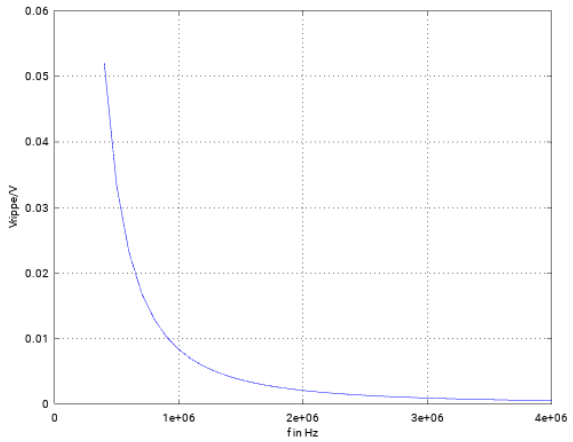


Figure 8.91: Peak to peak ripple voltage versus frequency at $C_{out} = 2.2\mu F$ and 0.5A load current

A nice result. The ripple voltage follows $V_{ripple_{pp}} \sim 1/f^2$. At the same time the dynamic losses -as long as we don't run into cross conduction - follow $P_{loss_{dyn}} \sim f$. The above figure applies as long as the ESR (equivalent series resistance) of the capacitor can be neglected. Most ceramic capacitors have ESR values in the range of $10m\Omega$. Parasitic inductance of the output capacitor is a second problem. The calculations the plots are based on only apply below the resonant frequency of the capacitor. Above the resonant frequency the capacitor starts to have an inductive characteristic! Using SMD components and a well optimized board the achievable resonance (for $2.2\mu F$) is in the range of

$$f_{res} = \frac{1}{2 * \pi * \sqrt{L * C}} = \frac{1}{2 * \pi * \sqrt{2.2\mu F * 3nH}} = 1.96MHz$$

If we want to go to higher frequencies than about 2MHz we will have to use several capacitors in parallel rather than one big capacitor to reduce the inductance.

Building high power converters for tens of Amperes the capacitors become bigger and the resonant frequencies will go down. Therefore high power switchmode power supplies tend to either have lower frequencies or use special layouts to create higher order output filters by the inductance of the traces on the board. (Well, this makes the regulation loop more tricky and design of the buck converter then becomes a real art!)

Transient response on load changes (estimation): The transient response to load changes depends on the regulator speed, the output capacitor, the inductor and the difference between the input voltage and the output voltage. In most cases the speed of the regulation loop is the limiting factor. The fastest possible regulation loop is a hysteretic converter. If the regulator is clocked (for instance because the compartor is clocked) the response delay under worst case conditions becomes:

$$t_{response} = 1/f_{clk}$$

During this time the output capacitor has to carry the change of the load current. The resulting drop (in addition to the normal ripple) is:

$$\Delta V_{out1} = \frac{t_{response} * \Delta I_{out}}{C_{out}} \tag{8.200}$$

In case of an infinitely fast unclocked hysteretic regulator ΔV_{out1} can get very small because the response time approaches 0. Practical design however are limited by the comparator speed and the gate charging time of the power transistor. Even using a fast hysteretic regulator we will barely be able to achieve less than about 100ns.

Example: $C_{out} = 22\mu F$,$\Delta I_{out} = 1.5A$,$t_{response} = 100ns$ already leads to a dip of $\Delta V_{out1} = 6.82mV$. As soon as we have a clocked system this dip caused by latency will get significantly larger!

After this delay the current in the inductor has to build up. It starts at the initial current I_0 and has to increase until it reaches $I_{equi} = I_0 + \Delta I$. When $I_{equi}$is reached the current flow through the inductor exactly matches the new output current after the load step. The time needed for building up the current is:

$$t_{buildup} = \frac{L * \Delta I_{out}}{V_{in} - V_{out} - V_{sw}} \tag{8.201}$$

In most cases the drop over the switch can be neglected.

Example: $\Delta I_{out} = 1.5A$, $L = 1\mu H$, $V_{in} = 2.8V$, $V_{out} = 1.5V$yields $t_{buildup} = 1.1538\mu s$.

Since the current increases almost linearly the additional charge taken out of the capacitor until the current is built up becomes:

$$Q_{buffer} = \frac{\Delta I * t_{buildup}}{2} \tag{8.202}$$

The voltage dip becomes:

$$\Delta V_{out2} = \frac{\Delta I * t_{buildup}}{2 * C_{out}} \tag{8.203}$$

Example: $C_{out} = 22\mu F$, $t_{buildup} = 1.1538\mu s$, $\Delta I = 1.5A$yields $\Delta V_{out2} = 39.334mV$.

In addition to the discharge of the capacitor there also is a voltage drop caused by the ESR of the capacitor.

$$\Delta V_{out3} = R_{ESR} * \Delta I_{out} \tag{8.204}$$

Example: $R_{ESR} = 20m\Omega$, $\Delta I_{out} = 1.5A$yields $\Delta V_{out3} = 30mV$.

The drop over the equivalent series resistance is visible at the beginning of the load change when the load current change has to be carried by the capacitor alone. When the current through the inductor reaches the required value the voltage drop caused by the ESR disappears. A very rough estimation already show some worst case boundaries:

$$\Delta V_{out1} + max(\Delta V_{out2}, \Delta V_{out3}) < \Delta V_{out} < \Delta V_{out1} + \Delta V_{out2} + \Delta V_{out3} \tag{8.205}$$

This somewhat simplistic estimation already shows that the switchmode power supply of out example can't be better than $46.152mV < \Delta V_{out} < 76.152mV$. Since we usually have component tolerances in addition these two corners aren't too far away from what we have to expect.

Transient response to load changes (exact calculation): If we want to know the load response dip with higher accuracy we have to calculate the shape of the dip. The current flowing out of the capacitor is simple:

$$I_{cap}(t) = \Delta I_{out} * \frac{t_{buildup} - t}{t_{buildup}} \tag{8.206}$$

The resistive drop over the ESR becomes

$$\Delta V_{out3}(t) = R_{ESR} * \Delta I_{out} * \frac{t_{buildup} - t}{t_{buildup}} \tag{8.207}$$

with t running from 0 to $t_{buildup}$.

The discharge of the capacitor calculates as:

$$\Delta V_{out2}(t) = \frac{\Delta I_{out}}{C_{out} * t_{buildup}} * (t * t_{buildup} - \frac{1}{2} * t^2) = \frac{\Delta I_{out}}{C_{out} * t_{buildup}} * t * (t_{buildup} - 0.5t)$$

$$\Delta V_{out2}(t) = \frac{\Delta I_{out}}{C_{out} * t_{buildup}} * t * (t_{buildup} - 0.5t) \tag{8.208}$$

Example: Let's assume there is no equivalent series resistance (perfect capacitor). In this case the maximum dip is reached at $t = t_{buildup}$. The dip ΔV_{out2}becomes

$$\Delta V_{our2ideal} = \frac{\Delta I_{out} * t_{buildup}}{2 * C_{out}}$$

but wait, we already had this equation in the simple estimation approach before! Using the numbers of our example it yields 39.334mV. This is the best we can get! There is no way to get better unless we change components or boundary conditions such as the input voltage V_{in}.

Back to the real world that also has an ESR:

Now we have to find the maximum of $\Delta V_{out2}(t) + \Delta V_{out3}(t)$

$$\frac{d}{dt}(R_{esr} * \Delta I_{out} * (t_{buildup} - t) + \frac{\Delta I_{out}}{C_{out}} * t * (t_{buildup} - 0.5t)) = 0$$

$$-R_{ESR} * \Delta I_{out} + \frac{\Delta I_{out}}{C_{out}}(t_{buildup} - t) = 0$$

$$t = t_{buildup} - R_{ESR} * C_{out} \tag{8.209}$$

This is the time the the dip reaches it's maximum. The maximum can be calculated:

$$\Delta V_{out23} = \frac{\Delta I_{out}}{t_{buildup}} * (R_{ESR}^2 * C_{out} + \frac{t_{buildup}^2 - t_{buildup} * R_{ESR} * C_{out} - 0.5(t_{buildup}^2 - 2 * t_{buildup} * R_{ESR} * C_{out} + R_{ESR}^2 * C_{out}^2)}{C_{out}}$$

$$\Delta V_{out23} = \frac{\Delta I_{out}}{2 * t_{buildup}} * (R_{ESR}^2 * C_{out} + \frac{t_{buildup}^2}{C_{out}}) \tag{8.210}$$

Example: Using the same numbers as before we get: $\Delta V_{out23} = 45.055mV$

Adding the drop caused by the assumed 100ns delay we get

$$\Delta V_{out} = \Delta V_{out1} + \Delta V_{out23} = \frac{\Delta I_{out}}{2 * t_{buildup}} * (R_{ESR}^2 * C_{out} + \frac{t_{buildup}^2}{C_{out}}) + \frac{t_{response} * \Delta I_{out}}{C_{out}} \tag{8.211}$$

Coming back to our little example: $\Delta V_{out} = 45.055mV + 6.82mV = 51.875mV$. This fits reasonably well into our first estimation.

8.4.8 RF emission of a buck power supply

Before going into details why are we making a model? Simulation of a switchmode power supply in time domain requires long simulation times (hours to days) because settling time is long. Even worse in system simulation each clock edge of the switchmode power supply excites the complete transistor netlist that can hold millions of transistors. To reduce simulation time a simplified EMC model is suggested.

Most simple: a current consumption model: Buck converters in continuous mode have a trapezoid pulse current consumption. The output current is more or less constant (It is not quite constant because the inductors always have a certain stray capacity).

Boost converters can be regarded as the dual circuit of a buck converter. Here (ideally) the current flowing into the circuit is constant while the output current consists of trapezoid shaped pulses.

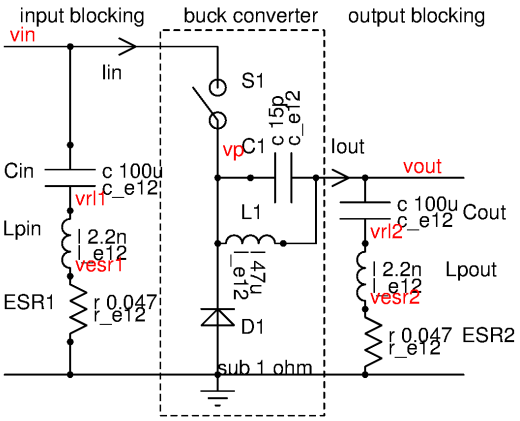


Figure 8.92: Simplified buck converter

In this simplified buck converter the input current flows while switch S1 is closed. When switch S1 opens the input current drops to 0. For analysis let us further simplify that the load current is constant and the supply current flowing from an outside source into pin vin is constant.

Neglecting the stray capacity C1 (We will revisit C1 later) and assuming ideal continuous mode (current through the inductor L1 stays constant) Iout remains constant. While S1 is closed the RF current flows through the loop Cin, Lin, ESR1, S1, L1, Cout, Lpout, ESR2. While S1 is open the RF current flows through the path D1, L1, Cout, Lpout, ESR2.

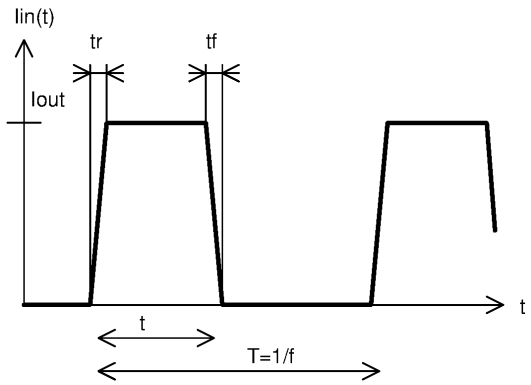


Figure 8.93: Input current of the buck converter in the time domain

The period T of the signal usually is under control of an oscillator. The ON time of the switch calculates as:

$$t_{ON} = T * D = D/f \tag{8.212}$$

The rise (tr) and fall time (tf) of the current depends on the strength of the driver stage of the switch. To minimize losses tr and tf usually are chosen in the range of 1% to 10% of the ON time t (The faster the rise and fall of the current the lower the losses). Looking at RF emission levels we usually are interested in the frequency domain representation of the signal. Since the signal is periodic we have to consider integer multiples of the frequency f of

the oscillator controlling the system. (For EMC we are not interested in the DC current - 0*f - of the converter.) The first frequency of interest is the clock frequency of the switchmode power supply (f1=1*f). Since we are only interested in the amplitude ($|I_n|$), not in the phase we can make things a bit easier shifting the pulse to a position where it ranges from -t/2 to t/2.

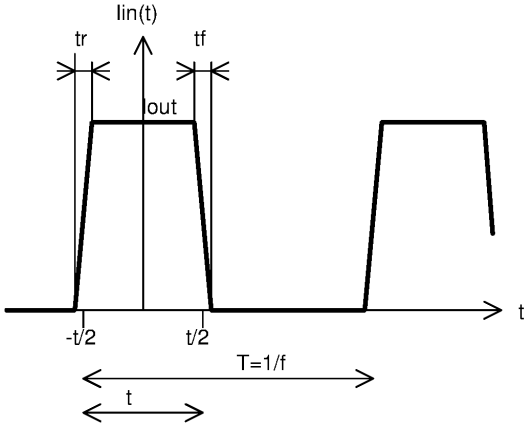


Fig.8.4.4.3: A little trick to simplify the calculations taking advantage of symmetries

Now we only have to consider the cosine waves. Due to the symmetries the sine terms cancel. The Fourier representation for fast edges compared to the harmonic considered (tr=tf<<1/n*f) can be approximated by:

$$I_n = \frac{1}{T} *_{-t/2} \int^{t/2} I_{out} * cos(\omega t) * dt \tag{8.213}$$

$$I_n = |\frac{1}{\pi * f_n * T} * I_{out} * sin(\pi f_n t)| \tag{8.214}$$

In this term $f_n * T$ is the number of the harmonic. (Note that f now is the frequency of the harmonic considered. f1=1/T, f2=2/T....)

The amplitude of a harmonic always drops to 0 when $\pi * f_n * t = k * \pi$ with k being a natural number. In other words whenever a complete wave of the harmonic fits into the pulse the integral over the cosine wave of the harmonic becomes zero.

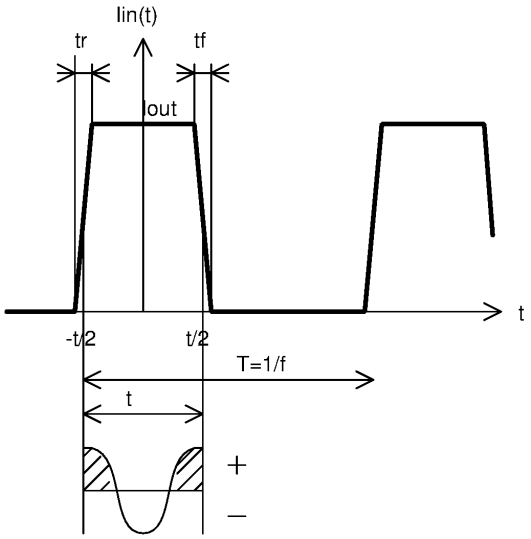


Figure 8.94: Cancellation of the Fourier integral if a full wave fits into the pulse

Example: We have a duty cycle of 20% and an amplitude of the pulse of 1A. This leads to:

$$I_0 = 0.2 * 1A = 200mA$$

$$I_1 = \frac{1A}{\pi * 1} * sin(\pi * \frac{t}{T}) = 187mA$$

$$I_2 = \frac{1A}{\pi * 2} * sin(\pi * \frac{t}{T/2}) = 149mA$$

$$I_3 = \frac{1A}{\pi * 3} * sin(\pi * \frac{t}{T/3}) = 100.9mA$$

$$I_4 = 58.779mA$$

$$I_5 = \frac{1A}{\pi * 5} * sin(\pi * 5 * 0.2) = 0$$

$$I_6 = \frac{1A}{\pi * 6} * sin(\pi * 6 * 0.2) = -31.183mA$$

$$I_7 = -43.247mA$$
$$I_8 = -37.841mA$$
$$I_9 = -20.789mA$$
$$I_{10} = 0$$

For an infinitely steep edge of the pulse (tr, tf ->0) the amplitudes absolute values observed decay with 1 over the number of the harmonic. At certain frequencies the harmonics disappear due to the sinus function in the equation. This sin(x)/x like function can be approximated by an envelope with a -6dB/octave (-20dB/decade) roll off starting at $1/(t * \pi)$.

A simple way to calculate such coefficients is offered by octave:

```
n=1:1:1000
In=Iout*sin(pi*D*n)./(pi*n)
In=rot90(rot90(rot90(In)))
```

The result is one column of numbers starting at f1 (first frequency and ending at harmonic number 1000. Well, a logarithmic presentation is more nice. We have to get rid of the negative numbers before calculation of the logarithm.

```
Indb=20*log10(abs(In))
save Indb.txt Indb
```

Now we can plot Indb scaled in dB(Ampere) using gnuplot which offers more flexibility than the octave built in plotting tool. The first cut off frequency in fact is at about $f_{g1} * 5/\pi$ as expected for a 20% duty cycle.

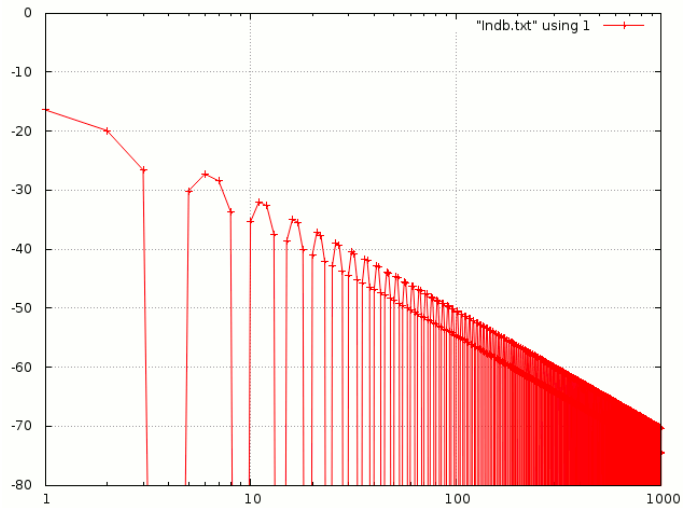


Figure 8.95: Current spectrum in dB(Ampere) of a 20% duty cycle, 1A switchmode power supply assuming tr, tf ->0. Y-axis in dBA, X-axis is the number of the harmonic (1 corresponds 500kHz, 1000 corresponds 500MHz)

Exceeding a frequency of $1/(tr * \pi)$ the roll off of the envelope becomes -12dB/octave (-40dB/decade) [31]. The analytical equations become very inconvenient to solve. Therefore in the following either simplified approximations or a pure numerical discrete Fourier transformation is used. For the brute force numerical approach the following netlist can be used describing a pulsed current consumption of 1A peak with 5ns rise and fall time and a pulse duration of 400ns (50%) with a period of 2us (So it corresponds our ideal rectangular pulse except for the rise time). Since the period is $2\mu s$ the fundamental frequency f1 is 500kHz.

```
imodel vin pgnd pulse 0 1 0 5n 5n 395n 2u
R1 pgnd 0 1
R2 vin pgnd 1
.end
```

Before running the fast Fourier transformation (FFT) we need equidistant sample points. Then we can run the FFT and plot the result. The following lines show the most important lines of interactive nutmeg code for ngspice.

```
ngspice 655 -> linearize
ngspice 658 -> fft v(vin)
ngspice 687 -> plot db(v(vin)) vs log(frequency)
```

Since the resistor is 1 ohm 1V corresponds 1A. The result of the FFT is shown below:

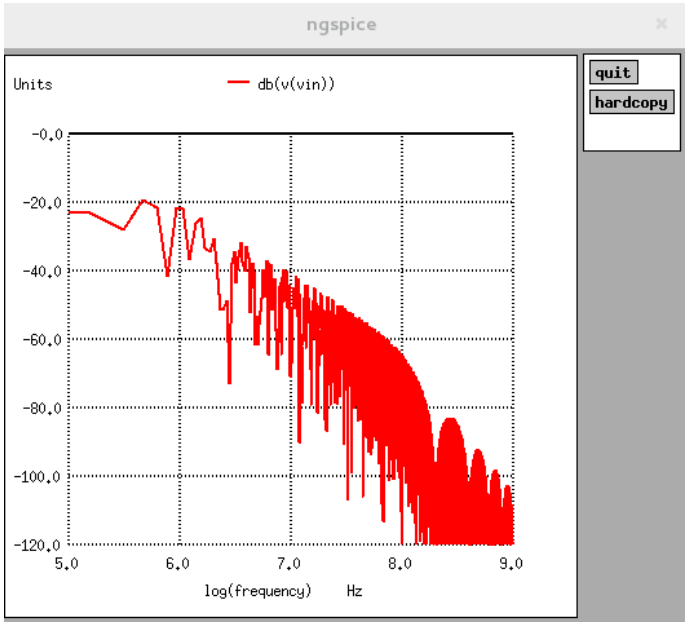


Figure 8.96: Numerical solution using SPICE with 5ns rise and fall time. Y-axis is in dBV (corresponding dBA). X-axis scale is log10(f in Hz)

From about 1MHz to 50MHz the results of the calculation and the FFT look quite similar. Above 50MHz the FFT shows the roll off caused by the 5ns rise and fall time. The effect of the rise and fall time can be included into the estimation by adding a first order low pass filter. Since regarding emission we are not interested in the phase we just consider the amplitude transfer function.

$$\frac{V_{out}}{V_{in}} = \frac{1}{\sqrt{1 + (f_n * \pi * t_{rf})^2}} \tag{8.215}$$

The term $\pi * t_{rf}$ represents the second cut off frequency provided by the rise and fall time of the trapezoid signal.

$$f_{g2} = \frac{1}{\pi * t_{rf}} \tag{8.216}$$

Plugging this second cut off frequency into the octave expressions we get:

```
n=1:1:1000
Iout=1
D=0.2
fg2=2000/(pi*5)
T2=1/fg2
Inlp=Iout*sin(pi*D*n)./(pi*n.*sqrt(1+n.*n.*T2*T2))
Inlp=rot90(rot90(rot90(Inlp)))
Inlpdb=20*log10(abs(Inlp))
save Inlpdb.txt Inlpdb
```

So Inlp is the low pass filtered spectrum of the ideal rectangular signal that we can use as an approximation of the trapezoid signal. Comparing the spectrum of the rectangular signal and the approximated trapezoid signal we see the following:

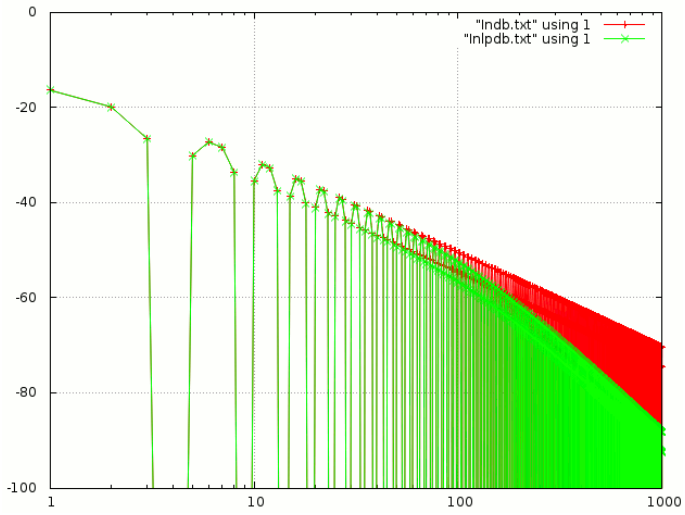


Figure 8.97: Current spectrum in dB(Ampere) of a 20% duty cycle, 1A switchmode power supply assuming tr, tf ->0 (red) compared to tr=tf=5ns. Y-axis in dBA, X-axis is the number of the harmonic (1 corresponds 500kHz, 1000 corresponds 500MHz)

For the analysis of the emission level we are only interested in the peaks. So we can draw an envelope around the green spectrum to represent the emission level. First cut off frequency is

$$f_{g1} = \frac{1}{\pi * t} = \frac{1}{\pi * T * D} \tag{8.217}$$

Since the fundamental f1 is very close to the DC value f0 we can proximate the signal at f_{g1}:

$$I_{db} = 20 * log10(D * I_{peak})$$

This is the current flowing for frequencies below f_{g1}scaled in dBAmpere. Above $f_{g1}$the level rolls off with -20dB/decade.

The second cut off frequency is

$$f_{g2} = \frac{1}{\pi * t_{rf}}$$

Above f_{g2} the current spectrum envelope rolls off with -40dB/decade. To check how well this envelope fits it is drawn into the spectra we just had a look at.

$$I_{db} = 20 * log10(0.2) = -14dB$$

$$f_{g1} = \frac{1}{\pi * 400ns} = 800kHz$$

This is approximately the second harmonic looking at a fundamental frequency of 500kHz.

$$f_{g3} = \frac{1}{\pi * 5ns} = 64MHz$$

corresponding the 128th harmonic in our example of a 500kHz power supply. At the corners we expect:

$$f_{g1} = -14dBA$$

$$f_{g2} : -52dBA$$

$$500MHz : -94dBA$$

$$1GHz : -100dBA$$

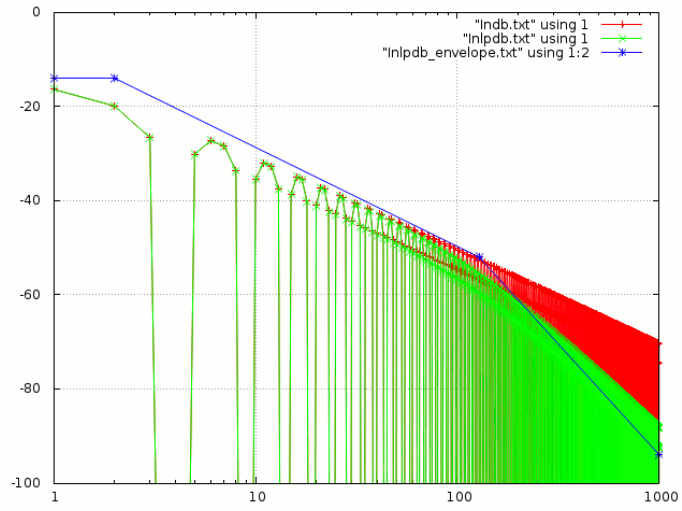


Figure 8.98: Calculated spectra and envelope approximation of the spectrum of the trapezoid signal with rise and fall time 5ns (X-axis: Number of harmonic)

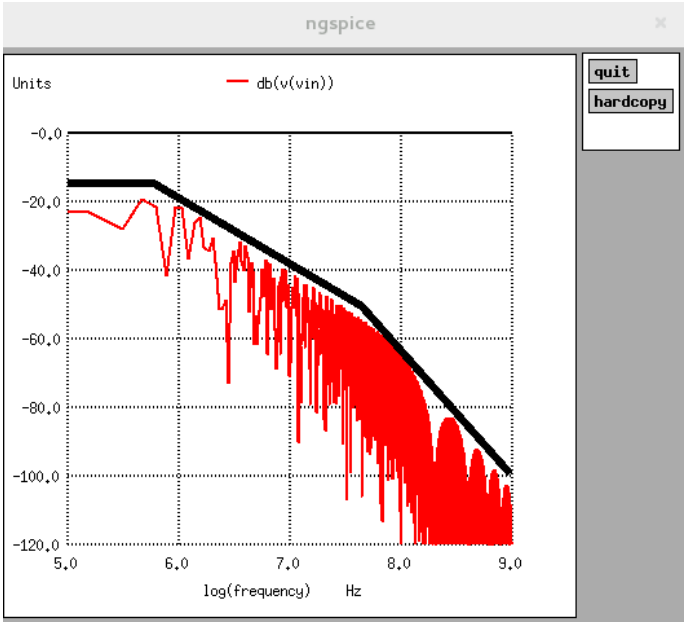


Figure 8.99: Spice simulated spectrum and envelope approximation

The envelope approximation of the spectrum of the current flowing into a switchmode power supply leads to a reasonable estimation of the worst case RF emission.

Application of the emission model Next step is to find out how this spectrum excites the components connected to the power switch. Here comes an example assuming the switchmode power supply is blocked by 3 capacitors:

- a 100nF capacitor with 1nH parasitic inductance and an ESR of $22m\Omega$
- a $2.2\mu F$ capacitor with 10nH parasitic inductance and an ESR of 0.1Ω
- a $22\mu F$ capacitor with 10nH parasitic inductance and an ESR of 1Ω

The wires between the capacitors are assumed to have 4.7nH and an impedance of 56Ω.

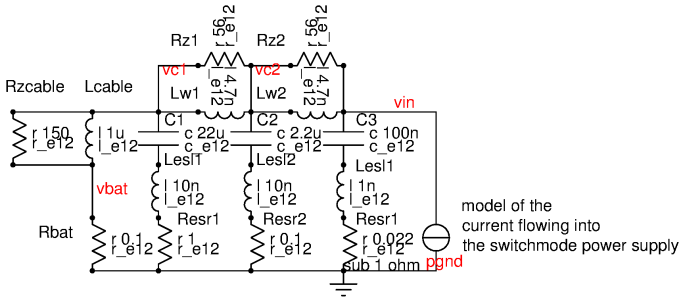


Figure 8.100: Usage of the emission model

The AC model for the current source in SPICE code looks like this:

```
vemc vbb pgnd ac 1
* low pass 800kHz
rlp1 vbb vlp1 1k
clp1 vlp1 pgnd 198p
* low pass 64MHz
rlp2 vlp1 vlp2 10k
clp2 vlp2 pgnd 0.248p
* conversion to current
g1 vin pgnd vlp2 pgnd 0.2
```

Since we want to have a result scaled in $dB\mu V$ we let the nutmeg graphics back end do some calculation.

```
plot vdb(vc1)+120
```

The resulting envelope of the spectrum at node vc1 looks like this:

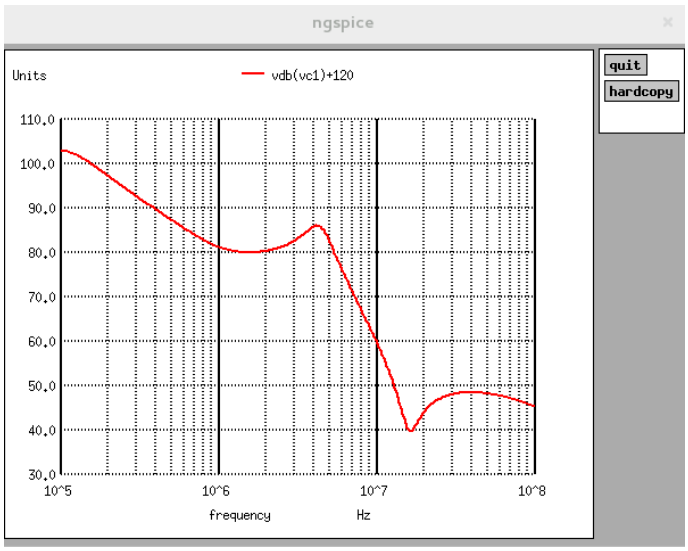


Figure 8.101: Envelope of the expected RF emission of a buck converter using an ideal current consumption model

Limitations of the the current consumption model: A model based on current flows is very intuitive and serves well for a first understanding of the EMC behavior of a switched mode power supply. But the current consumption model does not include current spikes flowing into the capacity of the rectifier and into the stray capacity of the coil. This simple model underestimates emission at high frequency. Furthermore it does not give any answers what happens at the output of the buck converter.

Alternative model using a voltage source: To verify the influence of the diode side of the switch in stead of a current model a voltage model is needed. The switch is replaced by an AC voltage source representing the envelope of the spectrum of the voltage across the switch. In series with the voltage source a resistor is placed representing the Rdson of the switch. This kind of model is fairly accurate for the closed switch but inaccurate for the open switch (An open switch is high resistive interrupting resonant LC loops while a voltage source always is low resistive. This leads to an overestimation of bond wire resonances.)

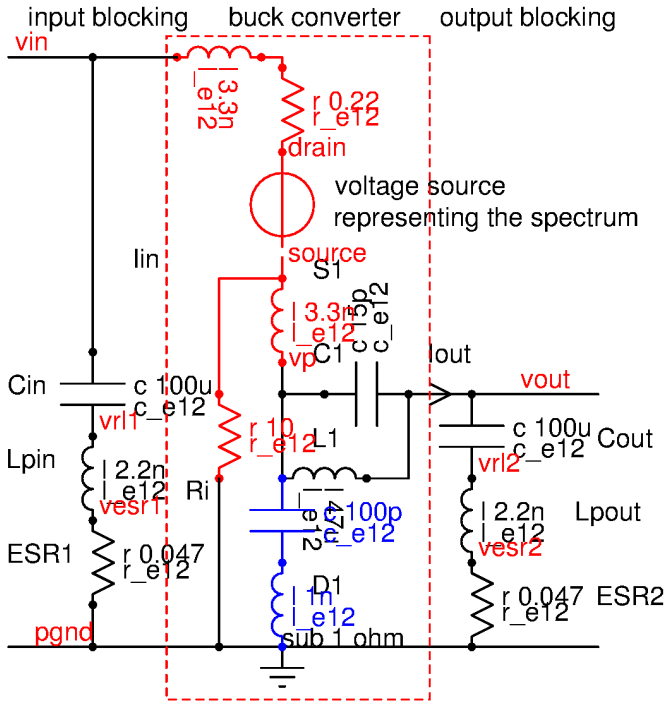


Figure 8.102: EMC Model of the buck converter using a voltage source

The red components represent the switch and the bond wires. The blue components represent the capacity and the parasitic inductance of the rectifier. The current through the voltage source now includes the current flowing into the parasitic capacities. But the rectangular pulse that is created by the DC current through the coil being part of the time delivered by the switch, part of the time delivered by the diode gets lost. This part of the spectrum must artificially be created by the resistor Ri.

As before the shape of time domain signal of the VOLTAGE across the switch must be converted into an envelope of the spectrum. Different from the first model there are two outputs now: vin and vout.

Of course we can remove the representation of the diode (the blue components) and the coil (L1, C1) to disconnect vout. In this case we have fallen back to (almost, there still are bond wires) the same as the first model - provided we have chosen Ri such that we get the same currents.

To try it out we chose Ri=10 Ohm and the fundamental amplitude of the voltage source 2V. The spice code exciting the circuit is shown below:

```
vemc vbb pgnd ac 1
* low pass 800kHz
rlp1 vbb vlp1 1k
clp1 vlp1 pgnd 198p
* low pass 64MHz
rlp2 vlp1 vlp2 10k
clp2 vlp2 pgnd 0.248p
eswitch drain source vlp2 pgnd 2
R1 source pgnd 10

* connection to the pins
L1 int14 vin 3.3n
L2 source vp 3.3n
R2 int14 drain 0.22
```

Again we use a 1V broad band reference source (Vemc) that is low pass filtered to create the cut of frequencies of the trapezoid signal. In stead of converting it into a current here a voltage is generated by eswitch. The current flow into the coil is provided by R1. The second part of the netlist connects the switch model to the pins. (Well, since SPICE doesn't like floating nets some nodes are tied to something defined using 1M resistors.)

Application of the second EMC model: To better compare the different models the first simulations are done without the coil and the diode. The input blocking is chosen identical.

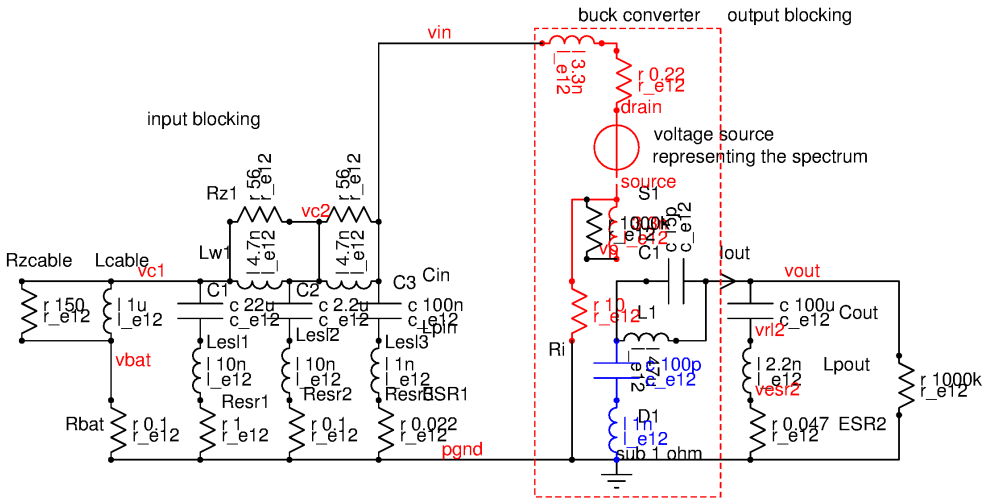


Figure 8.103: The same simulation as before but now using the voltage source representation of the switch

Note that the load (model of the diode and the coil model) is disconnected. Since we used the same supply blocking we expect the same result as before.

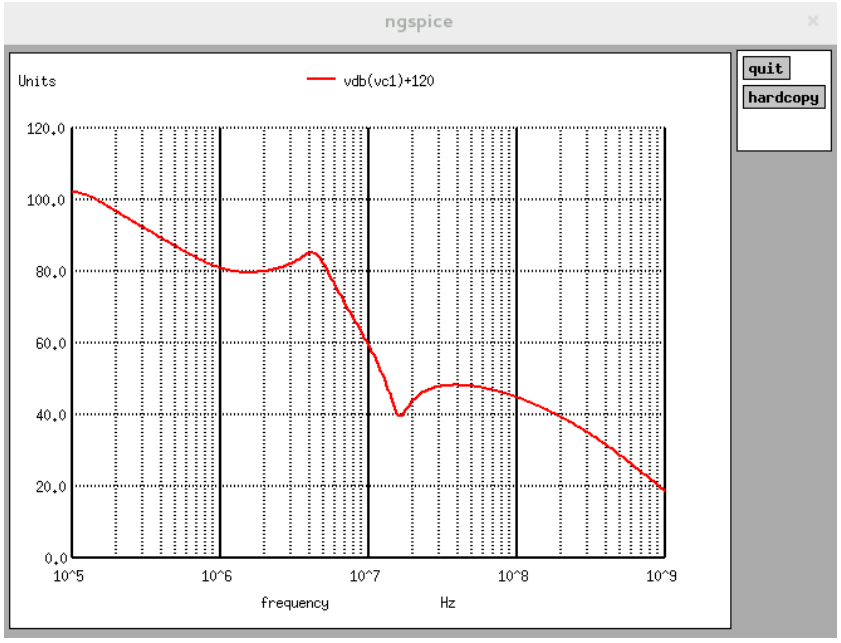


Figure 8.104: Result of the simulation with the load disconnected

The result corresponds expectations (except that now the simulated frequency range is 100kHz to 1GHz). Next step the load gets connected.

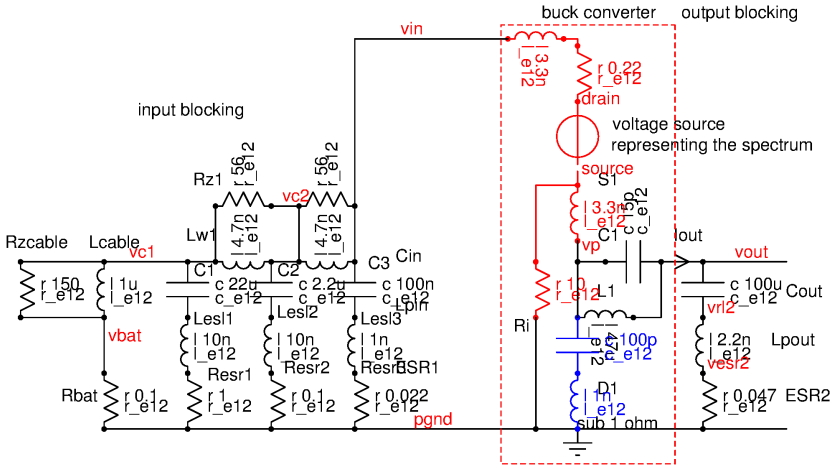


Figure 8.105: Now the load is connected

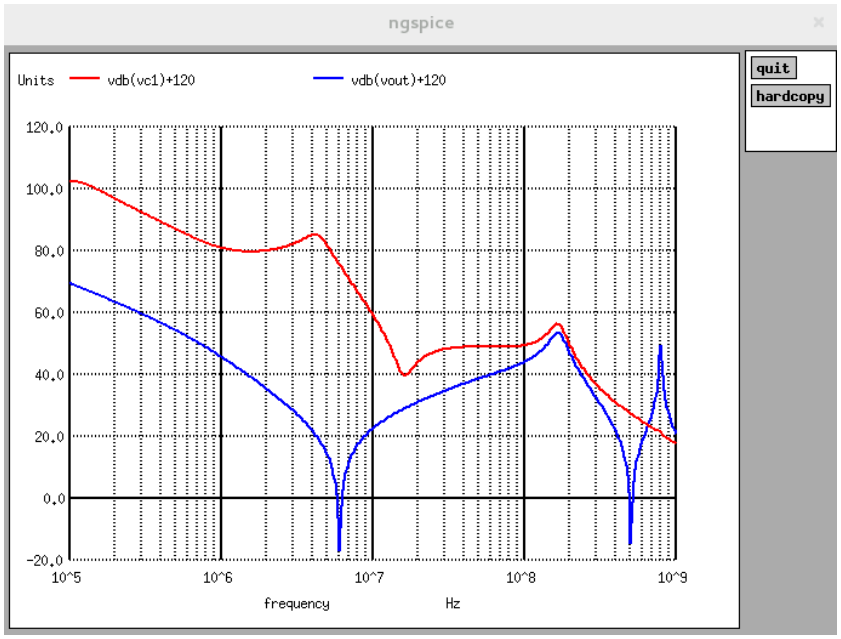


Figure 8.106: The voltage source model shows the resonance of the load together with the bond wires (at about 160MHz). Furthermore there also is a result for the emission at vout

Comparison with measurements: If RF emission is measured using the standardized setup there usually is an attenuator between the pin under test and the spectrum analyzer. This attenuator has an attenuation of about 15dB. To compare the simulated results with the measured ones these 15 dB must be substracted from the simulation results (or added to the measured results).

8.4.9 Boost Converter

Boost converters like buck converters temporarily store energy in the magnetic field of the coil. The most simple start of calculations is an ideal loss less boost converter. The output voltage is higher (or equal, but never less) than the input voltage (minus the diode forward voltage).

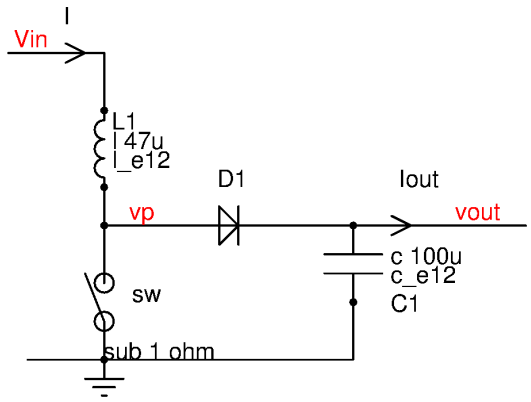


Figure 8.107: Most simple boost converter

In continuous mode the current in the coil can be assumed to be (almost) constant. The current only flow to the output while the switch is off (and the diode is on). Thus the output current becomes:

$$I_{out} = I * (1 - D) \tag{8.218}$$

Assuming an ideal converter without losses we find:

$$P_{out} = P_{in} \tag{8.219}$$

Replacing the power by voltage and current:

$$V_{out} * I_{out} = V_{in} * I \tag{8.220}$$

Or:

$$\frac{V_{out}}{V_{in}} = \frac{I}{I_{out}} = \frac{1}{1 - D} \tag{8.221}$$

In boost converters the energy transfer to the output only takes place while the switch is OPEN. For this reason the duty cycle D must be limited to a value of less than 100% (here the boost converter differs from a buck converter!). Ideally for a voltage doubler we need D=50%. Since there always are losses in the system practical designs allow a higher duty cycle. For voltage doublers to triplers duty cycles of $D_{max} = 75\%$ are a common choice.

In practical designs we have 2 sources of static losses: The voltage drop over the switch V_{sw} and the voltage drop over the rectifier V_{diode}. The power dissipation of the switch becomes:

$$P_{sw} = V_{sw} * I * D \tag{8.222}$$

Similarly the power dissipation of the diode is:

$$P_{diode} = V_{diode} * I * (1 - D) \tag{8.223}$$

The resistive losses of the coil calculate as:

$$P_{coil} = I^2 * R_{coil} \tag{8.224}$$

Now we can do the calculation for the boost converter including static losses.

$$P_{in} = P_{coil} + P_{sw} + P_{diode} + P_{out} \tag{8.225}$$

$$V_{in} * I = I^2 * R_{coil} + V_{sw} * I * D + V_{diode} * I * (1 - D) + V_{out} * I * (1 - D) \tag{8.226}$$

Solving this equation for D we get:

$$D = \frac{V_{out} + V_{diode} + I * R_{coil} - V_{in}}{V_{out} + V_{diode} - V_{sw}} \tag{8.227}$$

As long as the voltage drops over the diode, the switch and the inductor resistance are small compared to V_{in} and V_{out} the duty cycle only changes very little. With increasing load current the duty cycle moves up to compensate the ohmic losses. In a reasonable design this change of the duty cycle is in the range of some % to some 10%.

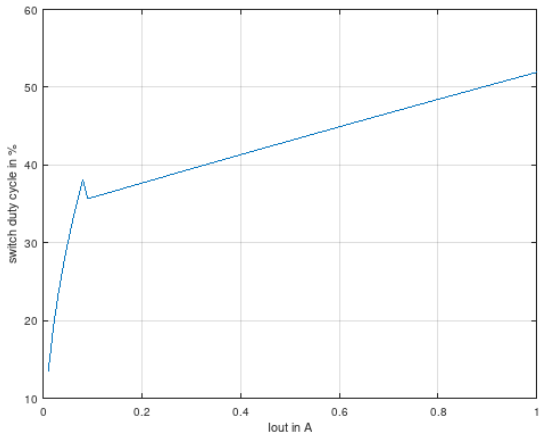


Figure 8.108: Change of the duty cycle of a boost converter from 3.3V to 5V

Below 80mA the boost converter operates in discontinuous mode. This is why the duty cycle changes from 15% to 36% in the low current range.

The efficiency calculates as:

$$\eta = 1 - \frac{V_{sw} * D + V_{diode} * (1 - D) + I * R_{coil}}{V_{in}} \tag{8.228}$$

Using bipolar diodes at low currents the forward voltage of the diode determines most of the losses. Often boost converters use active diodes (MOS transistors operated in reverse mode) as soon as continuous mode is reached. This increases the efficiency dramatically for low load currents. If the load further increases resistive drops start to become the most important contributor of losses.

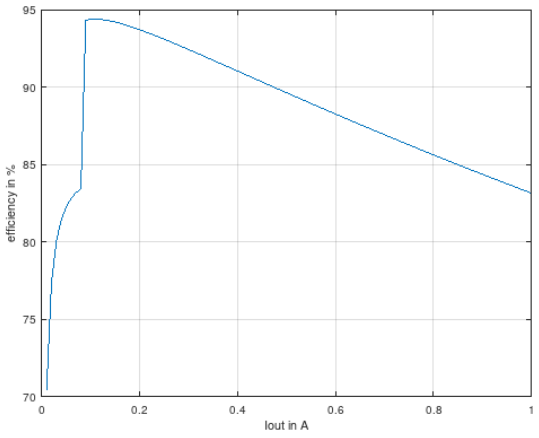


Figure 8.109: A typical integrated 1A boost converter efficiency operating at 5V output voltage (Vin=3.3V)

The converter shown above works in discontinuous mode until about 80mA. The drop of the bipolar diode limits the efficiency to about 80% to 85% there. At 80mA the converter reaches continuous mode and the rectifier is operated as a synchronous switch with a low resistance of 0.8Ω. This reduces the losses significantly and the boost converter reaches a peak efficiency of almost 95%. Further increase of the load current increases the resistive losses in the switch (0.6Ω) and the inductor (0.2Ω). Due to these losses the efficiency rolls back to about 83% at 1A load current.

This is quite a typical example of a small boost converter using integrated switches and rectifiers. If higher load currents are to be supported using external transistors is the more common approach because building transistors with an ON-resistance in the $m\Omega$-range on chip becomes too expensive.

From cooling point of view the total losses are most important. While the efficiency decreases linearly the current at the same time increases. So the losses of the converter follow an almost quadratic curve as soon as we are in continuous mode.

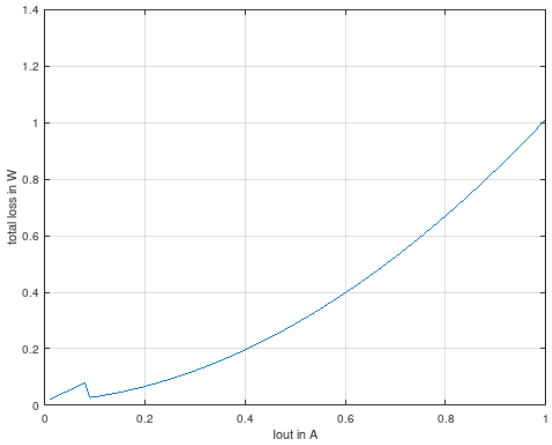


Figure 8.110: Losses of a 5V, 1A boost converter

Losses above 1W usually can't be tolerated on integrated circuits unless they are directly attached to a heat sink. Reducing the losses only is possible reducing the ON-resistance of the switches.

In many boost converters D is limited to a certain range. The maximum possible output voltage (in continuous mode) becomes:

$$V_{out_{max}} = \frac{V_{in} - I * R_{coil}}{(1 - D_{max})} - \frac{D_{max}}{(1 - D_{max})} * V_{sw} - V_{diode} \tag{8.229}$$

Often it is of interest which is the lowest supply voltage that can deliver a certain output voltage operating at the maximum duty cycle.

$$V_{in_{min}} = I * R_{coil} + V_{sw} * D_{max} + V_{diode} * (1 - D_{max}) + V_{out} * (1 - D_{max}) \tag{8.230}$$

Example:

We want to build a boost converter with an output voltage of 8V and a maximum duty cycle of 75%. As a peak current we allow 1.6A. The coil has resistance of 0.3 Ohm. The switch has a voltage drop of 0.8V (Power transistor and reverse polarity protection diode together). The diode is assumed to have a drop of 0.6V. The minimum supply voltage needed to provide 8V output voltage becomes:

$$V_{in_{min}} = 1.6A * 0.3\Omega + 0.8V * 0.75 + 0.6V * 0.25 + 8V * 0.25 = 4.18V$$

Note that at high duty cycles the losses in the switch and the coil become dominant. To produce high output voltages using transformers in stead of simple coils will dramatically reduce the duty cycle D and improve efficiency. Losses in the external components (resistance of the coil, ESR of the capacitors) have not yet been considered up to now for the sake of simplicity (Adding resistive losses we get higher order equations). Since the current in the coil is high these losses in some cases should not be neglected when these losses get close to the losses inside the semiconductors!

Looking at the example calculated above we find an efficiency of only 47.8% mainly due to the losses of the switch (1.28W) and the coil (0.77W).

Boost converter operation at current limitation: Operating at the current limitation the duty cycle differs from the duty cycle target coming from the voltage regulation loop. The current limiter turns off the power transistor before the regulation loop. Thus D decreases and 1-D increases. Let us assume we know the current limit $I_{max}$and the load current I_{out} we want to provide. When the supply voltage $V_{in_{min}}$ reaches exactly the point the output voltage reaches it target at the load current I_{out}we have the following condition:

$$I_{out} = (1 - D_{lim}) * I_{max} \tag{8.231}$$

$D_{lim}$is the duty cycle created by the current limitation of the switch. So we can calculate the the current limitation duty cycle.

$$D_{lim} = 1 - \frac{I_{out}}{I_{max}} \tag{8.232}$$

Again we can calculate the resulting voltages from the power flowing into the system, the power flowing out of the system and the losses in the switches, diodes, inductors etc. (Well, for low currents the losses in the capacitors are neglected up to now because the ESR is assumed to be so low that the impact is very little. This applies for some amps. As soon as we are looking at systems operating at several 10A we have to include them as well.)

$$P_{in} = P_{coil} + P_{sw} + P_{diode} + P_{out} \tag{8.233}$$

$$V_{in_{tran}} * I_{max} = I_{max}^2 * R_{coil} + D_{lim} * I_{max} * V_{sw} + (1 - D_{lim}) * I_{max} * V_{diode} + V_{out} * I_{load} \quad (8.234)$$

Now we are interested which is the supply voltage needed to provide a correct output voltage at the duty cycle forced by the current limitation. Reaching this voltage the current limitation will release and the voltage regulation loop takes over. Worst case the voltage regulation then tries to reduce (1-D) and the output voltage decreases again until the current regulation limitation increases (1-D) again). We may get a system oscillating between two modes at this transition. Not nice, but a working system. Let us plug in the equation for D_{lim} to see where this transition point is.

$$V_{in_{tran}} * I_{max} = I_{max}^2 * R_{coil} + (1 - \frac{I_{out}}{I_{max}}) * I_{max} * V_{sw} + \frac{I_{out}}{I_{max}} * I_{max} * V_{diode} + V_{out} * I_{load} \quad (8.235)$$

$$V_{in_{tran}} = I_{max} * R_{coil} + (1 - \frac{I_{out}}{I_{in}}) * V_{sw} + \frac{I_{out}}{I_{in}} * V_{diode} + V_{out} * \frac{I_{load}}{I_{max}} \quad (8.236)$$

Example: Let us assume we have a current limitation of 1.3A. The drop over the switch is 0.8V. The voltage drop over the diode is 0.6V. With this converter we want to provide an output voltage of 8V at a load current of 0.4A. The coil has a resistance of 0.3 Ohm.

The current limitation forces a duty cycle of

$$D_{lim} = 1 - \frac{0.4A}{1.3A} = 69.2\%$$

With this duty cycle the required input voltage becomes

$$V_{in_{tran}} = 1.3A * 0.3\Omega + 0.308 * 0.8V + 0.692 * 0.6V + 8V * \frac{0.4A}{1.3A} = 3.513V$$

The efficiency operating at 1.3A is about 70%.

But hey, wasn't this exactly the example we calculated before using a 75% duty cycle? Yes! We just have two possible operating points the design can freely move in between. Current limited operation mode that will even reach the target voltage at a reduced current by reducing the duty cycle and a fixed maximum duty cycle mode with D=75%. So we get the following behavior:

1. Below $V_{in_{tran}} = 3.513V$the booster tries to reach the target voltage operating in current limit mode with D<75%. This will even work if the current limit is lower than 1.6A (1.3A in the example).
2. Between $V_{in_{tran}} = 3.513V$ and $V_{in_{min}} = 4.18V$ the booster may oscillate between current limit mode at peak currents of 1.3A and voltage control mode. In voltage controlled mode the maximum duty cycle is limited to 75% allowing a peak current of 1.6A. The output voltage oscillates around the target voltage. The duty cycle oscillates between about 69% and 75%
3. Above 4.18V the booster goes into voltage regulation mode. Here two kinds of operation are possible:

- Hysteretic mode at maximum duty cycle. This leads to periodic bursts.
- Linear regulation that regulates the duty cycle exactly to the value needed.

Efficiency of the linear regulation (which means the duty cycle is regulated in a steady way without skipping pulses) is significantly higher than hysteretic mode (70% compared to 48% in our example).

Boost converter operating in discontinuous mode At low load current the the rectifier can either be operated as a synchronous switch or as a diode that really turns off when the current crosses 0A. Operating the switch in a synchronous mode simply means the energy is getting pulled back from the output when the current becomes negative. In this case we end up having more or less the same equations as in synchronous mode. This operating mode with the energy just swapping back and forth however leads to more resistive losses than turning off the rectifier when the current crosses the 0A line.

In extreme cases a boost converter with synchronous rectifier can is fact start to become a buck converter starting to supply the input from the energy stored in the output capacitor. Disconnecting the supply will increase the duty cycle to the maximum D_{max} and the input voltage will be supplied from the output at

$$V_{indisconnected} = (1 - D_{max}) * V_{cout}$$

In this equation the voltage V_{cout} is the voltage stored in the output capacitor.

This reverse operation can quickly discharge the output. Often this kind of operation is undesired and the converter intentionally is operated using the bipolar diode only in disconnected mode. In the following it is assumed the rectifier is a real diode that turns off at the zero crossing of the current.

Similar to the buck converter we can distinguish three phase of operation. During $D_{sw}$the switch is on and the current in the coil increases. When the switch opens the current is passed to the rectifier. The duty cycle the rectifier is conducting is called D_r. Once the current reaches 0A the rectifier turn off and we have a phase where the switch and the rectifier are both off. This is called D_{off}. We have already seen the equation investigating the discontinuous mode of the buck converter.

$$D_{sw} + D_r + D_{off} = 1$$

The output is supplied from the coil only during D_r. Assuming the current is triangular (This assumption is valid as long as the inductance doesn't change too much with the current) we get

$$I_{out} = I_{peak} * 0.5 * D_r \tag{8.237}$$

The peak current flowing is determined by the on time of the switch, the supply voltage minus the switch drop and the inductance. Usually the voltage drop over the switch can be neglected at low current. With $T = 1/f_{osc}$ the peak current becomes approximately

$$I_{peak} = \frac{V_{in} * D_{sw}}{f_{osc} * L} \tag{8.238}$$

When the switch opens the current in the inductor decays depending on the output voltage, the inductance and the drop V_f of the diode.

$$D_r * T = \frac{I_{peak} * L}{V_{out} + V_f - V_{in}}$$

$$D_r = \frac{f_{osc} * I_{peak} * L}{V_{out} + V_f - V_{in}} \tag{8.239}$$

The duty cycle of the diode can also be expressed by the duty cycle of the switch.

$$D_r = D_{sw} * \frac{V_{in}}{V_{out} + V_f - V_{in}} \tag{8.240}$$

The resulting output current of the switchmode power supply becomes

$$I_{out} = 0.5 * D_r * I_{peak} \tag{8.241}$$

Or reordered

$$I_{peak} = \frac{2 * I_{out}}{D_r} \tag{8.242}$$

Combining both equations we now have for the peak current I_{peak}we get

$$I_{out} = \frac{V_{in} * D_{sw} * D_r}{2 * f_{osc} * L}$$

$$I_{out} = \frac{V_{in} * D_{sw}^2 * \frac{V_{in}}{V_{out}+V_f-V_{in}}}{2 * f_{osc} * L} \tag{8.243}$$

This equation is important if a minimum duty cycle is required to supply a bootstrap circuit. Solving for the duty cycle required we get:

$$D_{sw} = \frac{1}{V_{in}} * \sqrt{I_{out} * 2 * f_{osc} * L * (V_{out} + V_f - V_{in})} \tag{8.244}$$

This equation is valid as long as the following condition is satisfied:

$$D_{sw} + D_r < 1$$

(D_{off}may not become negative) Expressing D_r by D_{sw}this condition becomes:

$$D_{sw} + \frac{f_{osc} * I_{peak} * L}{V_{out} + V_f - V_{in}} < 1$$

Replacing I_{peak}:

$$D_{sw} < \frac{1}{1 + \frac{V_{in}}{V_{out}+V_f-V_{in}}} = \frac{V_{out} + V_f - V_{in}}{V_{out} + V_f} \tag{8.245}$$

The losses in discontinuous mode can be calculated as follows:

Assuming the rectifier drop is more or less independent of the current (a reasonable assumption for a bipolar diode working at low current density) the rectifier losses become

$$P_r = V_f * I_{out} \tag{8.246}$$

The switch losses depend on the peak current, the ON-resistance and the duty cycle. The power dissipated in the switch is

$$P_{sw}(t) = I^2(t) * R_{on}$$

With $I(t) = t * V_{in}/L$ the energy dissipated in the switch over one clock period is

$$W_{sw} = \frac{V_{in}^2}{L^2} * R_{on} \int t^2 dt$$

with t running from 0 to D_{sw}/f_{osc}.

$$W_{sw} = \frac{V_{in}^2 * R_{on} * D_{sw}^3}{3 * L^2 * f_{osc}^3}$$

The average power dissipation over one period is

$$P_{sw} = \frac{V_{in}^2 * R_{on} * D_{sw}^3}{3 * L^2 * f_{osc}^2} \tag{8.247}$$

The resistive losses of the inductor can be calculated in the same way. The only difference is that the current in the inductor flows during the on time of the switch and the time the diode is conducting.

$$P_{Rind} = \frac{V_{in}^2 * R_{ind}}{3 * L^2 * f_{osc}^2} * (D_{sw}^3 + D_r^3) \tag{8.248}$$

Replacing the rectifier duty cycle by the switch duty cycle we can rewrite the equation

$$P_{Rind} = \frac{V_{in}^2 * R_{ind}}{3 * L^2 * f_{osc}^2} * (1 + \frac{V_{in}^3}{(V_{out} + V_f)^3}) * D_{sw} \tag{8.249}$$

Output ripple voltage: The output ripple can be calculated using a geometrical approach. The current charging the output capacitor triangular. The height of the triangular simply is $I_{peak} - I_{out}$. The base line of the triangular is $T * D_r * I_{peak}/I_{out}$. The charge increasing the voltage in the capacitor becomes:

$$Q = \frac{(I_{peak} - I_{out}) * D_r * T * (I_{peak} - I_{out})}{2}$$

The resulting peak to peak ripple voltage becomes

$$V_{ripple_{pp}} = \frac{(I_{peak} - I_{out})^2 * D_r}{2 * f_{osc} * C_{out}} \tag{8.250}$$

Current limiting behavior of a boost converter: There are two operating modes that must be distinguished looking at the current limiting behavior of a boost converter:

- Vout<Vin
- Vout>Vin

A classical boost converter using bipolar diodes usually can't be protected against short circuits. As soon as the output voltage drops below the input voltage the current only is limited by the resistance of the rectifier and the inductor.

$$I_{outlim} = \frac{V_{in} - V_{out} - V_f}{1 + R_{ind} + R_r} \tag{8.251}$$

This equation applies when the following condition is fulfilled:

$$V_{out} < V_{in} - V_f - I_{outlim} * (R_{ind} + R_r)$$

In this mode the switch will immediately detect an overcurrent and turn off at the smallest possible duty cycle.

If the condition is not fulfilled the overloaded boost converter will operate at the maximum permitted peak current I_{lim}. The regulation loop will increase the duty cycle until the peak current reaches the highest permitted value. The duty cycle of the switch usually is limited as well in a boost converter. The maximum duty cycle is called D_{lim}. Usually at overload the boost converter operates in continuous mode (unless we have a very low input voltage compared to the output voltage. But this is a very unusual case). The resulting output current becomes

$$I_{outlim} = (1 - D_{lim}) * I_{lim} \tag{8.252}$$

Since the current in the inductor keeps increasing this operation only takes place for a short time. In the successing pulses the switch will reach the limitation current I_{lim}before the maximum duty cycle is reached. The current limitation will reduce the duty cycle again until a balance between building up current (while the switch is on) and current decrease (while the switch is off) is reached operating at the limitation current.

Minimum load of a boost converter: Most boost converters support a certain range of duty cycles.

The maximum duty cycle must be limited to prevent excessive current flowing through the inductor just because the regulation loop is stuck at maximum power.

The minimum duty cycle often is limited because the driver stage of the rectifier is supplied by a bootstrap circuit. If the duty cycle would drop to zero this supply gets lost.

As long as the rectifier works in synchronous mode (switch instead of a bipolar diode) the energy can be taken back again. This means the boost converter can carry excess energy stored in the output capacitor back to the input. In other words the converter can operate as a step down converter transporting energy from the output back into the input. In this operating mode even without any load the boost converter will provide the correct output voltage.

Things change as soon as the rectifier is a bipolar diode. If the output voltage becomes higher than the regulation target the energy will not swap back because the diode blocks. All the energy stored in the inductor MUST be carried to the output. Without load the output voltage theoretically will become infinite. In practical designs there off course are some losses or some device at the output will break down or leak away the energy in some other way (diode recovery delay can play an important role).

In some applications it is required to protect the input from energy swapping back because the input side for instance uses 5V components while the output voltage is much higher. If the supply gets disconnected energy swapping back from the output can lead to destruction of the circuits connected to the low voltage input side. Such designs include a reverse current detection of the synchronous rectifier or a detection that the regulation loop hits the minimum duty cycle. If such an exceptional operating mode is reached the synchronous rectifier will turn off (to block the reverse current) and only the parasitic diode will act as a bipolar rectifier.

Usually this happens in discontinuous mode. The minimum turns on time provides a peak current through the inductor of

$$I_{peak_{min}} = \frac{V_{in} * D_{min}}{f_{osc} * L} \tag{8.253}$$

If we have no clue of the duty cycle but we know the minimum torn on time the formula can be rewritten

$$I_{peak_{min}} = \frac{V_{in} * t_{on_{min}}}{L}$$

As soon as the switch turns off the current will decay again. Neglecting the resisitive losses the decay time is

$$t_{decay_{min}} = \frac{I_{peak_{min}} * L}{V_{out} + V_f - V_{in}} \tag{8.254}$$

Since the current is triangular shaped the average current can be calculated

$$I_{out_{min}} = \frac{I_{peak_{min}} * t_{decay_{min}} * f_{osc}}{2}$$

This can be rewritten

$$I_{out_{min}} = \frac{I^2_{peak_{min}} * L * f_{osc}}{2 * (V_{out} + V_f - V_{in})}$$

Replacing the peak current we get

$$I_{out_{min}} = \frac{V^2_{in} * D^2_{min}}{f_{osc} * L * 2 * (V_{out} + V_f - V_{in})} \tag{8.255}$$

If we want to use the minimum ON time instead the formula becomes

$$I_{out_{min}} = \frac{V^2_{in} * t^2_{on_{min}} * f_{osc}}{L * 2 * (V_{out} + V_f - V_{in})} \tag{8.256}$$

The following plot shows an example of the minimum load current required for a 1.88MHz boost converter with a minimum on time of 18ns. The input voltage is sweeped while the output voltage is forced to remain at 5.1V. The inductor was assumed to have a minimum inductance of $2.5\mu H$.

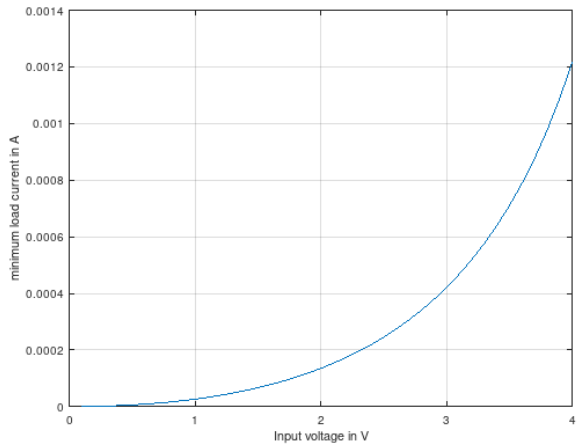


Figure 8.111: Minimum load current of a boost converter with bipolar diode

The closer the input voltage gets to the output voltage the higher the current provided by switch operating at minimum on time gets. At a certain point the current in the inductor doesn't decay anymore. The converter changes to continuous mode. Since the current won't decay anymore the current will increase with every switch cycle. The current is limited only by the current limitation of the switch. (horizontal plateau in the following plot)

Further increasing the input voltage (above Vout+Vf) leads to a continuous current flow through the inductor that is only limited by the parasitic resistance of the inductor. Characteristics in this range are independent of the boost converter control. (This is the range with the linear current increase).

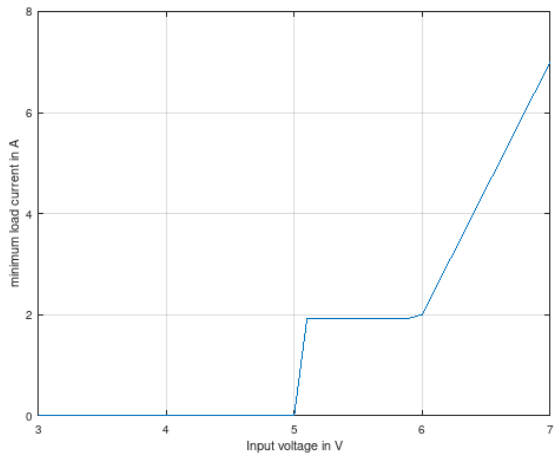


Figure 8.112: When the input voltage exceeds the target output voltage the boost converter changes to CCM and then becomes bypassed by the inductor and the resistance of the inductor

8.4.10 Regulation loops

There are two basic concepts how to regulate a switchmode power supply:

1. voltage mode regulation
2. current mode regulation

Each of these concepts has it's pros and cons.

Voltage mode regulation: Voltage mode regulation is the most straight forward concept. It simply measures the output voltage of the switchmode power supply and compares it with a reference voltage. To create a PWM the output of the regulator amplifier is compared with a saw tooth signal.

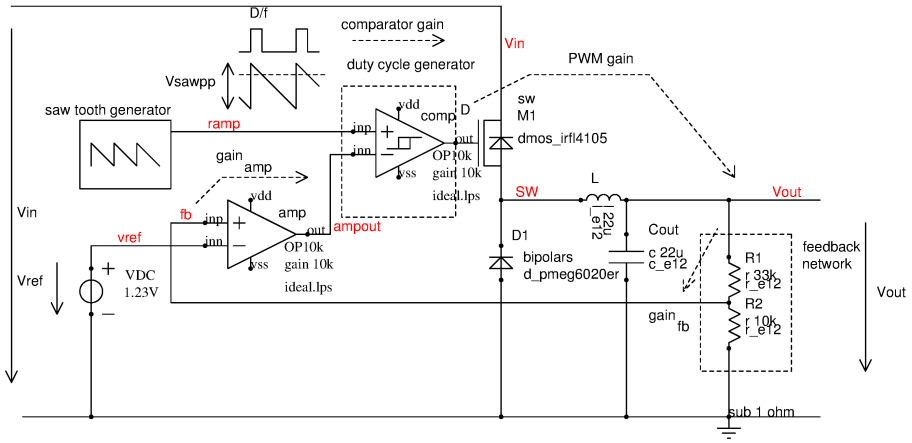


Figure 8.113: voltage mode regulation loop

The voltage mode regulation loop is fairly simple. The problems are hidden in the poles of the loop. Usually the output capacitor Cout together with the load resistance determines the dominant pole. The second pole is determined by the inductor L together with the capacitor Cout. Too keep this loop stable the loop gain must fall below 0dB before reaching the second pole. This either requires a very big output capacitor or the gain of the regulator amplifier must be limited to very low values. Limiting the gain of the regulator amplifier leads to a poor accuracy of the complete switchmode power supply.

Supply rejection is a second problem of the voltage controlled switchmode power supply. A change of the supply voltage Vin immediately changes the current flowing through L. The output voltage increases within a few clock cycles. Since the regulation loop has a low gain and a low cut off frequency the time needed to correct the duty cycle is long and the power supply rejection is poor.

Voltage feed forward: One possible work around is to measure the supply voltage Vin and add a correction voltage to the output of the regulator amplifier.

Current mode regulation: Current mode regulation attempts to keep the current under control in a fast regulation loop. In a current controlled regulator we have two regulation loops making the complete system more complex. The output of this voltage loop sets a target current used for the current control loop. The fast current loop compares the current flowing in the inductor with the target current provided by the voltage loop. Since the current loop can be made fast, changes of the supply voltage will not significantly change the average current through inductor L. This leads to a much better supply rejection than a pure voltage control can provide.

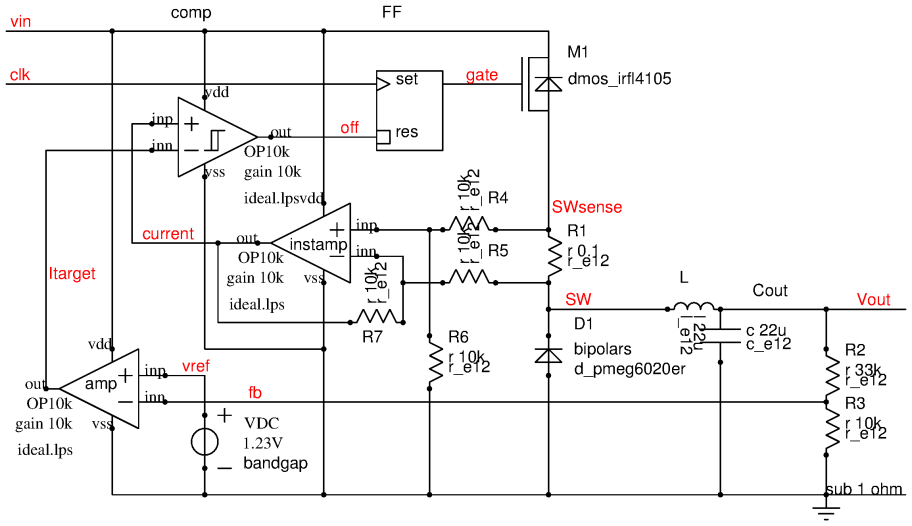


Figure 8.114: Concept of a current controlled switchmode power supply

In the conceptual current mode control the amplifier instamp measures the current flowing through R1. The current measured is compared with the target current Itarget by comparator comp. The clock signal clk periodically turns on M1 via set-reset flip flop FF. The transistor turns off again as soon as the measured current exceeds the target current.

This first implementation however has a stability problem. If the difference between the input voltage and the output voltage becomes too low the current builds up very slowly but after turn off of M1 the current drops much

faster. This leads to operation below the clock frequency. The following plot shows an example of such a subharmonic operation.

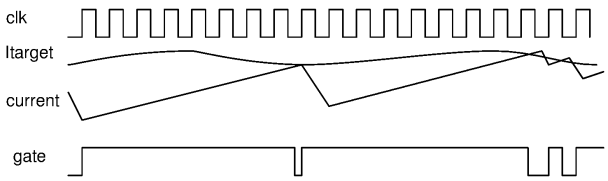


Figure 8.115: Subharmonic operation without slope compensation

To prevent this kind of subharmonic operation an additional saw tooth signal called slope compensation is added to the target current.

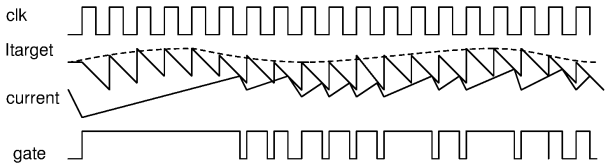


Figure 8.116: Slope compensation reduces subharmonic operation

The modification of the regulation loop is the additional saw tooth input comp and the adder adding this compensation signal to the target current.

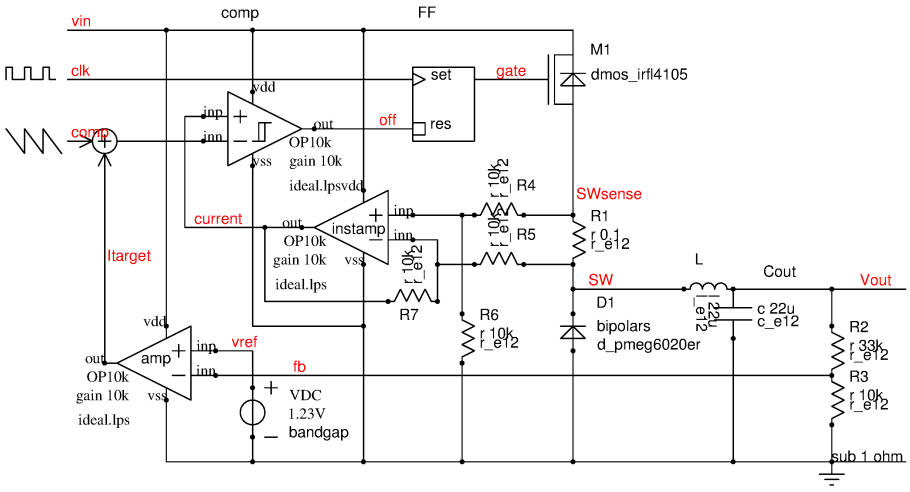


Figure 8.117: Current mode regulation with slope compensation

Up to now we considered switchmode power supplies as linear systems except for the PWM itself. With logic getting cheaper and cheaper it is tempting to add more and more features using digital controls.

Slope compensation using a digital slope generator: This is typically the first step to replace linear functions by a digital function. A linear slope compensation requires an analog saw tooth generator. The same function can be implemented using a higher clock frequency and a counter. The counter drives a DAC. The DAC approximates the ramp by a staircase function. This approach offers several advantages:

1. If the high frequency clock (some 10 MHz) is already available the area of the saw tooth generator can be saved.
2. A crystal controlled clock offers higher precision than an RC saw tooth generator
3. The staircase function can intentionally be tailored non linear to better accommodate discontinuous mode as well

Replacing the linear slope compensation by a staircase function however will lead to discrete timings of the PWM!

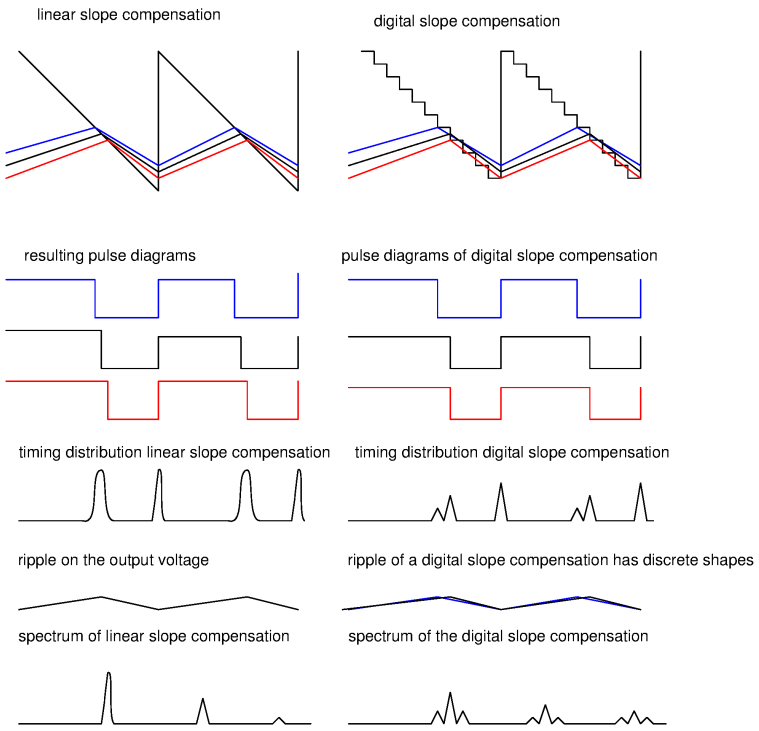


Figure 8.118: comparison of the timings of a digital slope compensation and a linear slope compensation

The linear slope compensation leads to an intersection of two linear functions. Each of these functions has a certain spread caused by noise. As a result the timing of both edges are more or less gaussian.

A digital slope compensation leads to discrete timings. The timing distribution at least at the falling edge has multiple peaks. As a consequence the PWM will have a periodic pattern of changes of the ON-time. This leads to a more or less periodic pattern in the output voltage of the switchmode power supply. This periodic pattern leads to side bands of the harmonics of the PWM frequency. The period however is a function of load current, Cout and the gain of the regulation loop.

8.5 Reset and voltage monitoring circuits

Almost all supply systems, no matter if they rely on a linear regulator or a switchmode regulator, need a reset generator. The reset generator starts the logic in a well controlled initial state. Furthermore it deactivates all circuits during reset to prevent them from malfunction or even destructive behavior.

8.5.1 Reference of a reset generator

The first problem of a reset generator is that it must be in a defined (reset) state even when the reference generators aren't started yet. This means, a comparator simply comparing the output voltage of a regulator with the reference voltage isn't sufficient. The reset generator must have an inherent reference. The reset generator first compares the system bandgap with this internal reference. Once the system reference is up and running the more precise bandgap can be compared with the regulator output. The following drawing shows the concept.

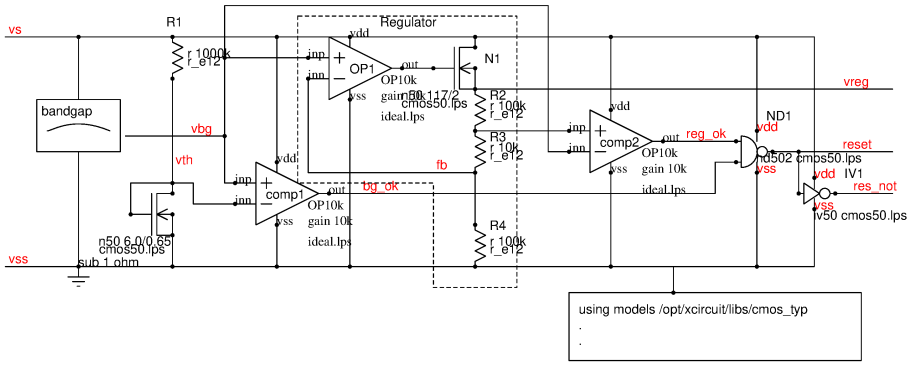


Figure 8.119: Concept of a reset generator

comp1 compares the bandgap with an NMOS threshold. The signal bg_ok indicates the bandgap has reached more than one MOS threshold. comp2 compares the bandgap voltage with the voltage of a tap of the feedback

divider of the regulator. If the regulator is working this tap is expected to have a voltage of about 1.1*vbg. The output bg_ok and reg_ok are ANDed by the NAND gate ND1 and the inverter IV1. The result is the signal res_not. While in reset state signal res_not is 0V.

Usually a signal res_not is preferred because providing 0V still is possible if the supply of the complete circuit is close to 0V. Supporting a signal reset that is logic 1 while the circuit is unsupplied usually causes some conflicts!

8.5.2 Functional safety

Some people claim functional safety (ISO26262) requires the reset generator to have it's own bandgap just to be sure that if the regulator bandgap fails the reset circuit still works. The argument has several flaws:

1. If the same bandgap is used twice there is no guarantee that only one of them fails. (If the bandgap fails in a certain process corner this will happen in all identical bandgaps!)
2. At RF injection it is common that more than 1 bandgap fails.
3. Two bandgaps don't protect against comparator failure.

If we really want to have redundancy this means using 2 different bandgap topologies, two comparators with two different topologies and sending two reset signals to the logic.

Well, and if we have all this: what happens if one reset input of one flip flop at a critical signal doesn't work? Even assuming the logic AND the reset is redundant, somewhere there will be a levelshift driving an actuator that still can fail...

I rather stick with robust and well encapsulated design upfront than taking everything to the limit and then duplicating things to make them robust again.

Encapsulated design: Encapsulated design means that protection and security functions must be as close to the function to be protected as possible. The lower the number of components involved the lower the chance of failure. This is no guarantee that there is no failure. But it is a reduction of likelihood.

Example of encapsulated design: A short circuit protection function should be in a short closed loop with the power stage. No handing over any protection to the logic or even worse to software! The encapsulation of power stage together with the protection has the following benefits:

1. It still works even if the logic gets a reset due to ground bounce at the short
2. If level shifts are failing (unsupplied, transient disturb ...) the encapsulated protection still works
3. Missing supplies at the logic side don't harm protection
4. The design is still protected even if the software didn't load and boot correctly

The only path to the logic is a status flag that an error has occurred.

Testability of encapsulated design: Encapsulated design often leads to local logic in the supply domain different from the logic supply. A classical scan chain test isn't possible. In most cases encapsulated protections must be tested by functional tests.

Design time of encapsulated designs: Encapsulated designs are bare hardware. There is no chance of re-configuring the chip by a software path. Quick and dirty design style and then patching in the NVM (non volatile memory) isn't possible. As a consequence bugs in encapsulated design always require a redesign. This is a price we have to accept building safe designs.

8.5.3 Comparator bias

The bias current of the comparators must be present as soon as the supply voltage vs reaches about one threshold. Taking the bias current from the bandgap isn't working. So the comparators must either be designed self supplying or in a way that we automatically get a correct reset signal and a correct res_not signal if the comparator bias isn't available.

In the L4949 and L4938E voltage regulator series this was achieved by a resistor bias driving the output stage as soon as the supply reaches Vbe.

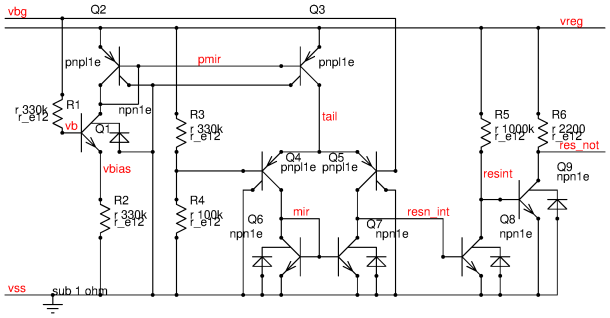


Figure 8.120: Simplified reset circuit of the L4938E

The comparator Q4, Q5 is unbiased until the bandgap voltage exceeds Vbe and at the same time the regulator output vreg reaches about 1V. As long as this condition isn't satisfied the node resn_int is floating. Since Q8 needs a certain base current to turn on the resistor R5 will drive the base of Q9. Q9 turns on and pulls down the output res_not.

Comparing this design with the concept shown before we can identify the pieces. Transistor Q1 corresponds the comparator comp1. It compares the bandgap voltage with Vbe.

The pair Q4, Q5 and current mirror Q6, Q7 correspond comp2 of the concept.

The only difference is: Instead of relying that the comparator works down to low voltage the resistor R5 is used to determine the state of the circuit while the comparators aren't working correctly.

8.5.4 Autonomous undervoltage detection circuits

We have seen one of the biggest problems is how to detect a critical supply state while the reference voltage and the bias current generators are not yet running. In the paragraph above there already were some precautions shown using resistors as a firs bias generator. This can be made a concept. One possible implementation is a ring current generator. Using bipolar transistors we finally end up at a bandgap like behavior. The circuit can be regarded as a double bandgap consisting of an NPN bandgap below the dashed symmetry line and a PNP bandgap above the dashed line. As expected the threshold is close to $2 * V_{bg} \approx 2.46V$.

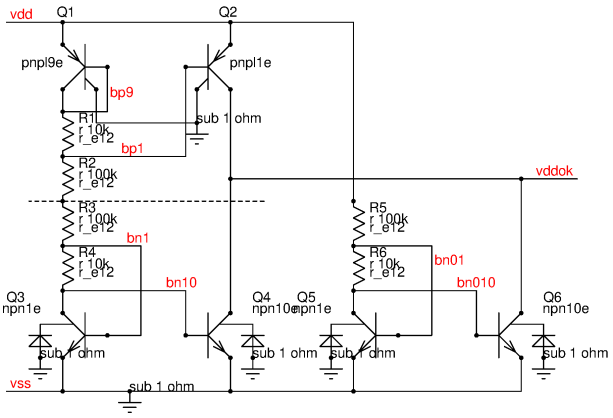


Figure 8.121: Bandgap comparator used as a undervoltage detection circuit

The main circuit consists of Q1 to Q4 and R1 to R4. Since we are stacking two diodes (Q1 and Q3) the node vddok would remain undefined as long as vdd is below 2*Vbe. To define the node vddok to be low as soon as the supply reaches Vbe the additional circuit Q5, Q6, R5, R6 is required. This define-circuit enforces a low voltage at low supplies between Vbe and 2*Vbe. At about 1V Q6 turns off and the behavior of the circuit is only determined by the double bandgap Q1 to Q4.

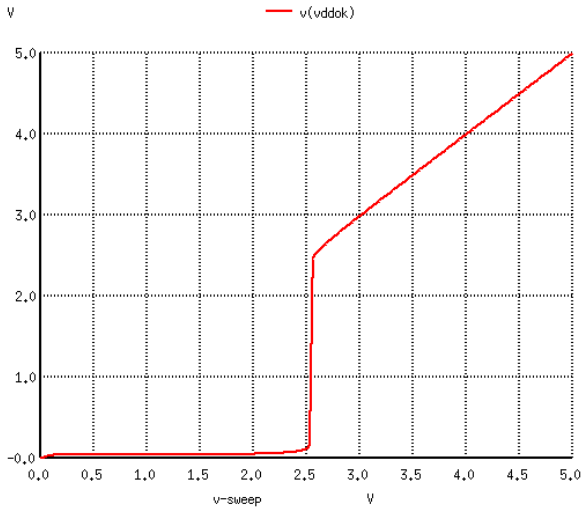


Figure 8.122: Simulation of the undervoltage detection circuit (DC sweep of vdd)

The same concept can be implemented using MOS transistors. Using MOS transistors the trip point can significantly differ from 2.46V because the threshold can be modified by he process engineer changing the doping of the bulk. As usual the MOS implementation like weak inversion bandgaps too has a wider spread than the bipolar counter part. Analytic calculation only gives a first starting point. At the end the circuit must be fine tuned using the technology models of the MOS transistors. Since the spread is determined by the spread of the threshold the tolerance of the MOS implementation must be determined by Monte Carlo simulation and corner simulation.

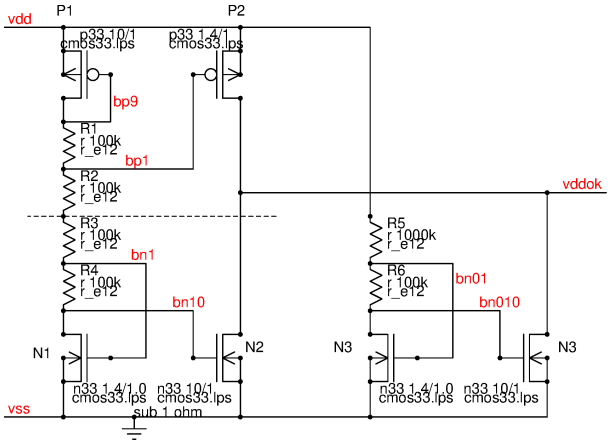


Figure 8.123: Same undervoltage detection concept using MOS transistors instead of bipolar transistors

The simulation of a DC sweep is shown below.

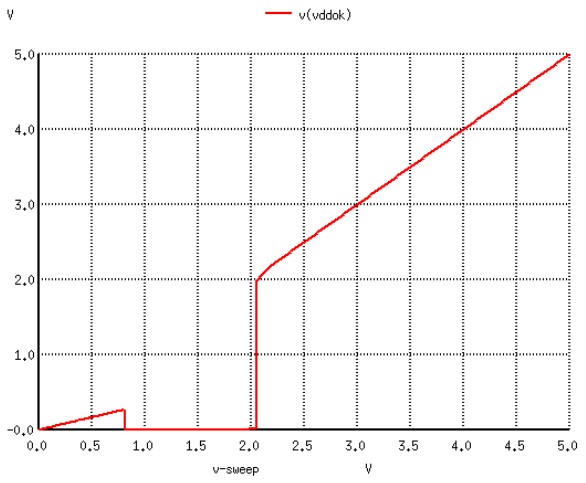


Figure 8.124: DC sweep of the MOS undervoltage detection circuit.

Typically the circuit will use the same type of transistors as the logic gates supplied from vdd. Below the threshold voltage the node vddok becomes floating. As long as the NMOS transistors used in the logic and in the undervoltage detection are of the same type the logic won't work there anymore.

Some technologies offer the same transistors with halo implant for the logic and without halo implant for analog applications. The transistors without halo implant have a lower threshold voltage. In this case it may make sense to use transistors with halo implant for N1 and N2 to match the threshold of the logic gates. The hold circuit N3, N4 can be designed using the transistors without halo implant to achieve a non floating low state down to below Vth of the logic NMOS transistors.

Often the circuit is intentionally designed not to have a trip point at double bandgap but at a voltage slightly higher than the sum of the thresholds of the PMOS transistor and the NMOS transistor. In this case resistors R2 and R3 may have fairly low values (In the bipolar version they were designed 10 times higher resistive then R1 and R4 to achieve a PTAT voltage of about 1.2V at room temperature) or may even be taken out on purpose.

8.5.5 Minimum reset time

During reset the following things must happen:

1. The reset signal must be assigned.
2. The bias current generators must be activated.
3. Logic supply must be powered up.
4. The clock generator must be started (if it isn't already running)
5. Proper operation of the clock must be monitored.
6. Reset signal must be synchronized with the clock inside the logic
7. The logic reads the configuration registers and gets initialized.
8. In case of a microcontroler the program counter is set to the start address.
9. All memory pages must be mapped to an initial position.
10. logic brings all I/O cells and power amplifiers into a safe initial state.

This sequence requires a certain time. For this reason once a reset generator triggers the reset, the reset must be held for a certain time. Leaving the reset state before all the initial configuration is done may lead to unpredictable system states. The classical analog approach used in many pure analog regulators with reset looks like this:

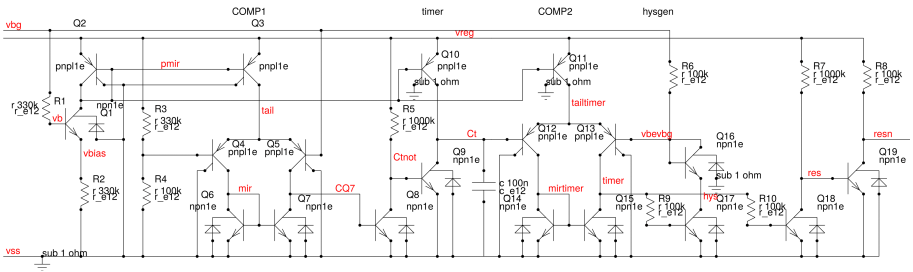


Figure 8.125: Reset of the L4938 including the RC timer

Since a synchronous logic depends on a clock to propagate the reset to all registers the reset in most cases turns on the clock generator no matter in which state the system is. (It must override STOP mode!)

The clock generation module often is part of the reset circuit. The following figure shows a typical automotive system on a chip.

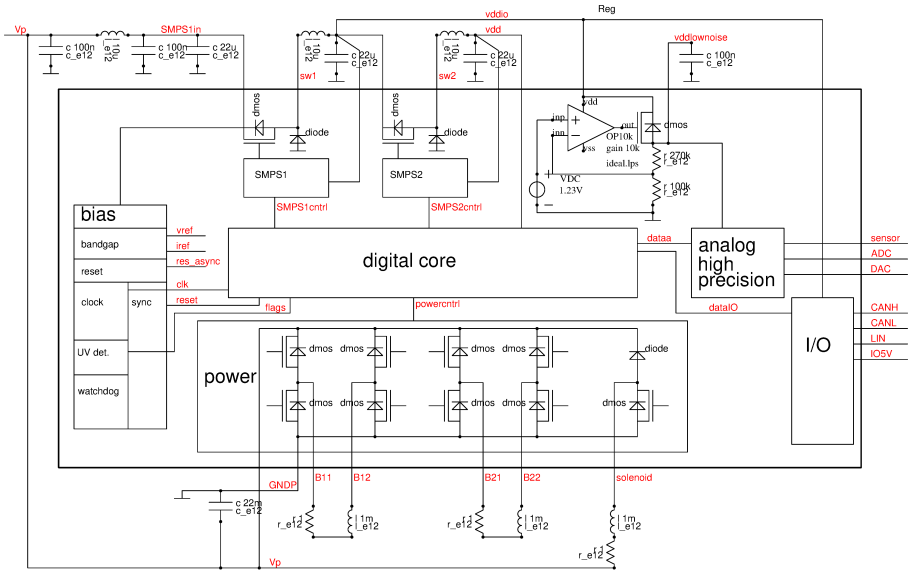


Figure 8.126: Example of a complete system chip

The reset is derived from the voltage vdd that is supplying the logic. The bias block does not only provide the reset but also provides the clock and synchronizes the reset and the flags informing the logic of the state of all the other supplies and the clock.

Before handing over the signals to the logic all the asynchronous events must be synchronized with the clock. For this reason the clock must be started at reset in case it is not yet running. (So a reset event will always end a STOP mode.) The following plot shows the signal threads and the conceptual timing diagram.

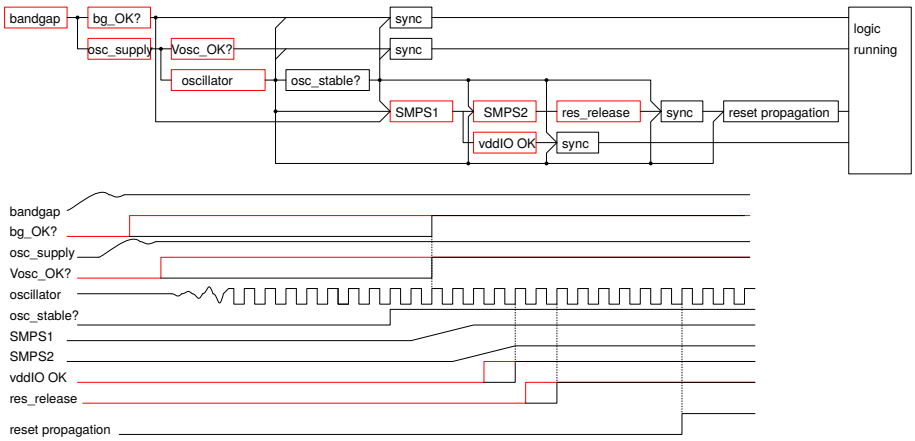


Figure 8.127: Startup threads and simplified pulse diagram

The asynchronous events are drawn in red color. Each asynchronous event has a synchronized copy drawn in black color on the same trace.

At power up first the bandgap and the bias generators are started. Once the bandgap is running signal bg_OK goes HIGH. Since the clock is not yet running there is no synchronization for a long time.

The asynchronous bg_OK starts the supply of the oscillator. The oscillator is monitored for stable operation. After a certain number of cycles the signal osc_stable changes state. This signal can be designed to be inherently synchronous (it is derived from the clock anyway).

When the oscillator is running the synchronization of the signals will start.

Reset release takes place as soon as the supply vdd (the output of SMPS2) reaches it's target value.

Inside the logic depending on the logic design it may take several clock cycles to propagate the rising edge of the reset signal through all the flip flops. This is represented by the block reset_propagation.

As soon as the reset is released and propagated through the complete logic the booting process begins. During booting the logic can perform the following tasks:

1. read I/Os that configure modes (test mode, debug mode, normal mode .... Usually hard coded in the logic design)

2. read default memory mapping table, move memory pages to the addresses of the map (there may be different tables for different modes!)

3. start executing software at initial address (usually this is #h0000. Typically the first command is a JMP to the real position of the software in the memory. The initial JMP allows having multiple software versions for development in the NVM. Switching between the versions can be done modifiying the JMP address.)

4. If required modify memory map

Under normal circumstances (normal mode) the non volatile memory (NVM) is mapped to address #h0000. The boot process executes the code of the NVM.

In debug mode often a ROM code is mapped to #h0000. This ROM code holds a piece of software that establishes a communication via the debug pin. Once this communication is running often the registers are mapped to #h0000. This permits reading and writing the registers in the debug mode. Debug mode often gives access to analog test modes too.

Placing the port registers of port A and port B at addresses #h0000 and #0001 in debug mode is a very elegant solution for software development. This way port A can be used as a command (for instance JMP) and port B can be used as a target address of the test code to be executed.

Some test mode switches the clock to an external pin, disables all analog circuits and enables the scan chain. This permits running a digital scan test.

This complex boot process is typical for a microcontroler type logic.

8.6 Stepper Motor Drivers

Fully integrated stepper motor drivers became common in the 1980s. First designs used bipolar transistors and could handle some hundred mA per coil. The limitations were mainly of thermal nature. With improvements of the packages (exposed dice pad for better cooling) higher currents became feasible.

A stepper motor usually consists of a rotor with a permanent magnet usually having multiple poles and a stator that can be controlled by magnetizing it electrically.

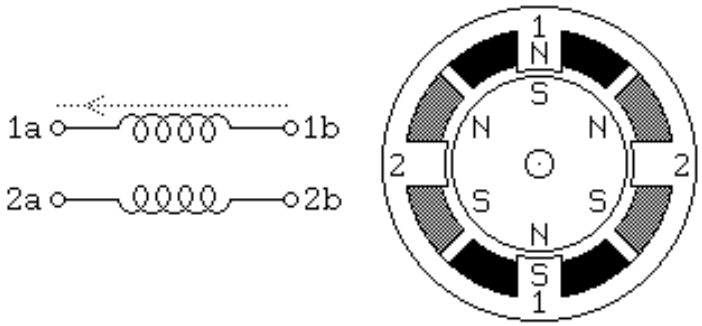


Figure 8.128: stepper motor

In the motor shown the coil 1 (pins 1a and 1b) is magnetized. The rotor will move into a position where the south pole of the rotor is facing the north pole of the stator and the north pole of the rotor is facing the south pole of the stator. This is exactly the position drawn above.

If winding 2 is energized the rotor will move into a position where the rotor poles are facing the stator poles of winding 2 accordingly. The following figure shows a typical full step current control scheme where always only one of the two windings is excited at a time..

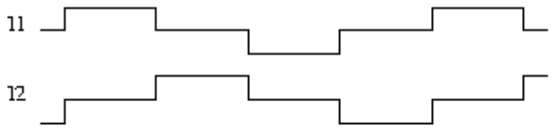


Figure 8.129: currents in full step mode 1

If a higher torque of the motor is desired it makes sense to use both windings simultaneously. The rotor poles in this case will always move to a position in the middle between the stator poles. Each winding permanently is excited with a current of +I or -I but it will never be current-less.

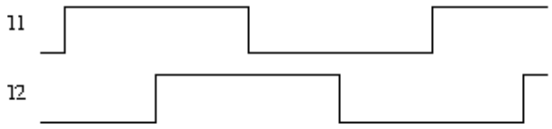


Figure 8.130: currents in full step mode 2

Since the rotor moves to a position between the poles the torque typically only increases by a factor $\sqrt{2}$ instead of a factor 2.

Full step mode makes the rotor move in rough steps. The movement is not continuous and may excite mechanical resonances because the rotating mass of the rotor and the force of the magnetic field act as a "spring and mass" mechanical system. With every step the rotor runs over its ideal position and then swings back.

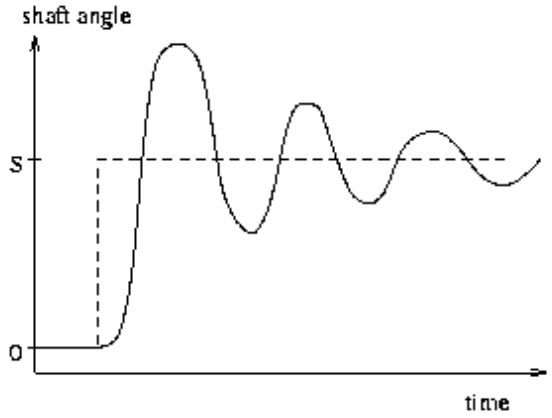


Figure 8.131: movement of the rotor after a step

The dashed line shows the angle of the magnetic field. The solid line shows the angle of the rotor following the field. The torque making the rotor move is the product of the cosine of the angle ϕ between the rotor and the magnetic field B and the pole area A ([51] page 430).

$$M = cos(\phi) * B * A \tag{8.257}$$

For smooth movement of the rotor the field should rotate at a constant rotation speed and a constant magnetic field strength and the rotor should follow at a constant angle. This means the currents in the windings should follow a sine function and a cosine function.

Well, this is an ideal assumption requiring a constant air gap between rotor and stator which isn't feasible. So some roughness of movement and excitation of resonances will always remain.

Typically the sine function and the cosine function are approximated with half step operation or even finer granularity of the steps.

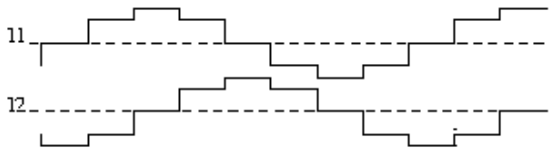


Figure 8.132: currents used for half step operation

Mechanical resonances can reduce the torque of a stepper motor dramatically when the next step exactly hits the moment where the rotor just bounces back. The frequency of the resonance depends on the spring constant which is proportional to the magnetic flux and the rotating mass. Friction and other mechanical losses are acting as a damping of the resonant system. One trick to circumvent such resonances it to switch between stepping schemes to just before a resonance is hit. A typical example is switching between full step mode 1 and full step mode 2 to change the magnetic flux (spring constant) when a resonance is hit.

This way of circumventing torque loss caused by resonances requires an exact knowledge of the complete mechanical system (motor, rotor mass or inertia, damping, mechanical load, stiffness of the shaft between motor and mechanical load).

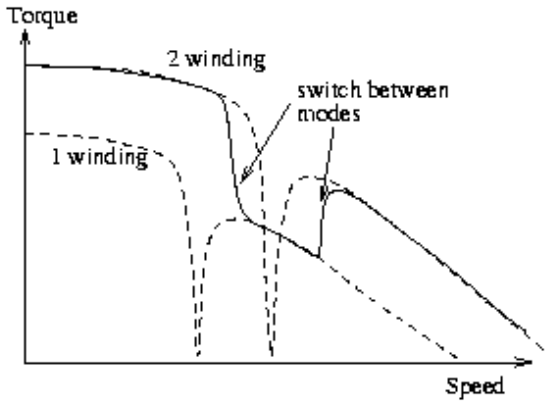


Figure 8.133: switching modes to circumvent resonances

Measuring the response of the mechanical system (torsion measurement of the shaft, observation of back EMF of the motor etc.) may help but lead to very complex digital regulation loops. The limitation of such regulation loops are production tolerances (spread of air gaps and field strength B, Errors of current regulation) on one hand and the sheer amount of data (stream of ADC results of torsion measurements, digital noise filtering of back EMF measurements) coming from the sensors that has to be processed in real time. For small to mid size systems over-engineering (choosing a motor and a driver with enough torque margin to overcome resonances) often is cheaper than spending the effort of complex digital regulation loops.

Using simple analog regulations to compensate the drop of the torque has been tried, but since the 'spring constant' involved depends on the current flowing in the windings the non linear behavior can hardly be handled by simple linear analog circuits consisting of OPAMPs and some RC filters.

Step loss Worst case one or more steps are lost because the mechanical system doesn't follow the rotating field anymore. The most reliable way of detecting (and correcting) step loss is to use angular sensors connected to the shaft. These can be HAL sensors or optical sensors.

Such sensors add cost to the system and it has to be considered what is cheaper: over-engineering the system and possibly running at lower speed of adding the sensors.

Initialization Since the exact position of a motor at power on is unknown if no sensors are present it is common practice to run the system into a mechanical block for initialization. Typically a mechanical load is moved into one direction using more steps than the total range of movement would require. After having executed more steps than the system allows we can assume that the mechanical block is reached and the position is known (block position). A typical example is a car headlight adjust that runs to the mechanical block each time the car ignition is turned on. Just observe your car headlights when turning on the engine to the initialization in action.

8.6.1 Unipolar drivers

In the beginning the power transistors were the dominating cost driver. To minimize the cost of the power transistors the windings of the motor were tapped in the middle. Only half of the volume of the motor windings was used this way but it saved half of the power transistors compared to a full bridge. The design using a taped winding is called a unipolar stepper motor driver.

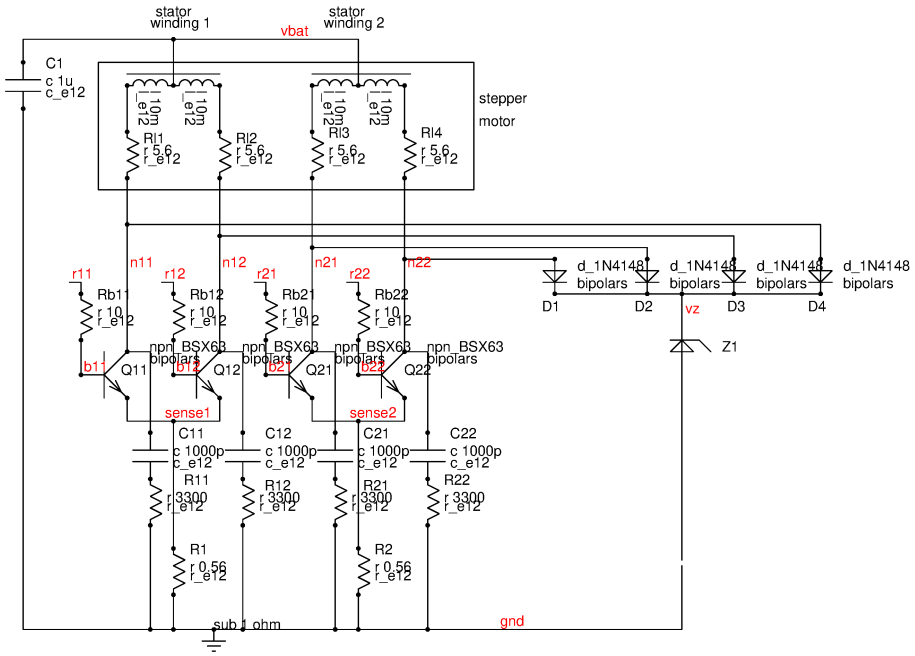


Figure 8.134: unipolar stepper motor driver

The unipolar stepper motor driver works with only 4 power transistors but to protect the transistors against the flyback voltages of the stray inductance of the motor windings the clamping diode Z1 and the diodes D1 to D4 are required. The clamping voltage of Z1 must be chosen at least twice as high as the supply voltage vbat. The windings are mutually coupled. Usually the coupling factor is about 0.9 or tighter. In addition the rotating rotor couples a back EMF into the stator windings.

If the clamping voltage of Z1 is too low the power dissipation increases rapidly and efficiency of the unipolar driver drops.

The break down voltage of the power transistors Q11 to Q22 must be higher than the clamping voltage of Z1 plus the forward voltage of the diodes. In addition often snubber networks are required C11 to C22 and R11 to R22) because while the power transistors are off the dangling ends of the windings will ring. The ringing wires of the stepper motor are sources of significant RF emissions if the snubbers are removed.

Resistors R1 and R2 are the current sense resistors. These resistors are usually connected to a comparator input controlling a chopper to regulate the average current through the windings.

The base currents must be limited either by resistors (Rb11..Rb22) or by a current limitation of the driver stages driving the base of Q11 to Q22.

Summarizing we can say the unipolar stepper motor driver minimizes the number of active devices but we have to pay a high price in terms of poor winding usage and passive external components required for EMC reasons.

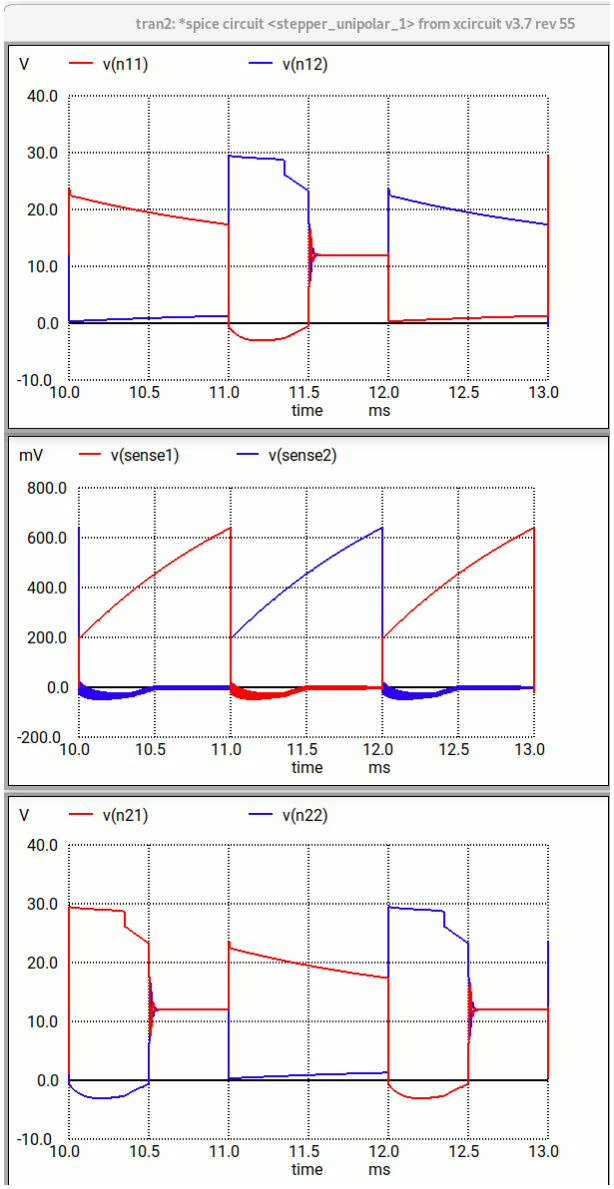


Figure 8.135: signals of the unipolar stepper motor driver working at high step rate

The high peaks (close to 30V) are the flyback signals of the stray inductance after turn off. The tilted roofs at about 20V are the back EMF while the other side of the winding is switched on.

8.6.2 Bipolar Stepper Motor Driver

Today bipolar stepper motor drivers using full bridges are the standard approach. Bipolar stepper motor drivers exploit the full volume of the windings. Ideally a bipolar stepper motor driver can generate double the torque of a unipolar stepper motor driver. In addition the voltage stress of the power transistors is reduces because there is no back EMF lifting the drain- or collector-voltage to double supply. The first generation of bipolar stepper motor drivers where implemented using bipolar transistors. These chips went into production in the mid 1980s. Typical examples are TEA3717 or TCA3727.

The low side transistors are chopping to regulate the current. While high side power transistor Q101 is on the transistor Q211 acts as a chopper. While high side transistor Q201 is on low side transistor Q111 acts as a chopper.

To reduce the drop over the high side power transistor at least during a certain time the predrivers of the high side (Q103, Q103, Q202, Q203) always use the highest voltage available. When the chopper turns off the flyback voltage gets one diode forward voltage higher than vbat and Q103 and Q203 respectively take over the base drive of the active high side stage. This little trick reduces the power dissipation of the high side stage by about 30%.

The PNP transistors Q102, Q103, Q202, Q203 have two collector rings. The inner one is the load collector. The outer ring is the saturation sense. When the PNP transistors approach saturation the outer ring takes over the base current. This way the base will not be flooded with holes ant turn off time of the PNP transistors remains in a reasonable range.

A similar trick is used for the low side. Q111 and Q211 consist of the power transistor itself and one transistor where the emitter is connected to the input of the darlington stage. When the NPN power transistor approaches

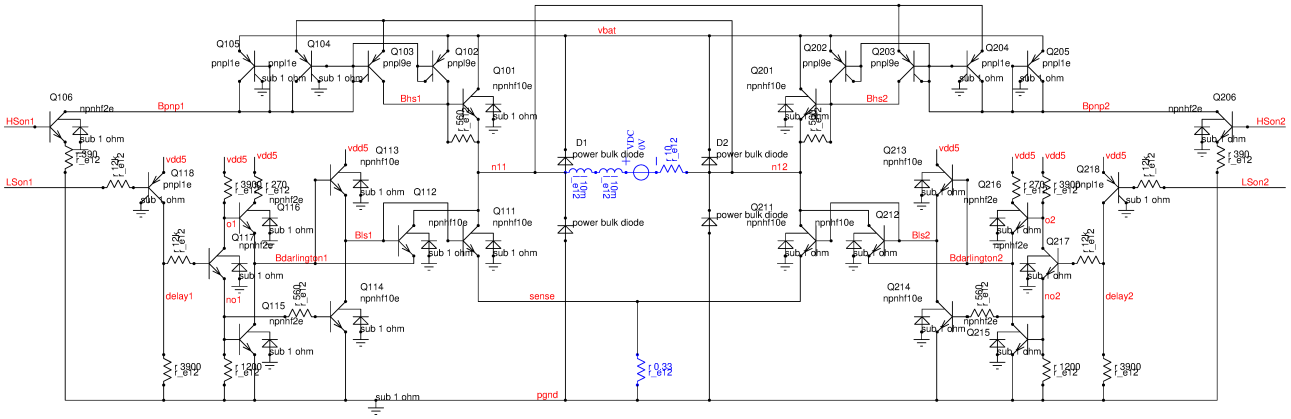


Figure 8.136: Simplified circuit of the TCA3727 power bridge

saturation the emitter connected to the base of the darlington input (nets Bls1 and Bls2) starts to work as an NPN transistor's collector (NPN operating in reverse mode) reducing the drive of the darlington stages. This prevents hard saturation of the NPN transistors and turn off is possible within some 10ns. (Without this trick turn off would take several 100ns increasing slope losses significantly.)

Q118 and Q218 are delay lines using the turn off delay of the PNP transistors. This way turn on of the low side is delayed to avoid cross conduction at phase reversal. (turn off delay of Q118 must be longer than turn off delay of Q102, Q103). Q118 and Q218 intentionally have not second ring to reduce turn off time. These two transistors intentionally are the slowest transistors in the whole bridge circuit.

The following plot shows the function of the bridge driving a stepper motor winding with 20mH and 10 Ohm resistance.

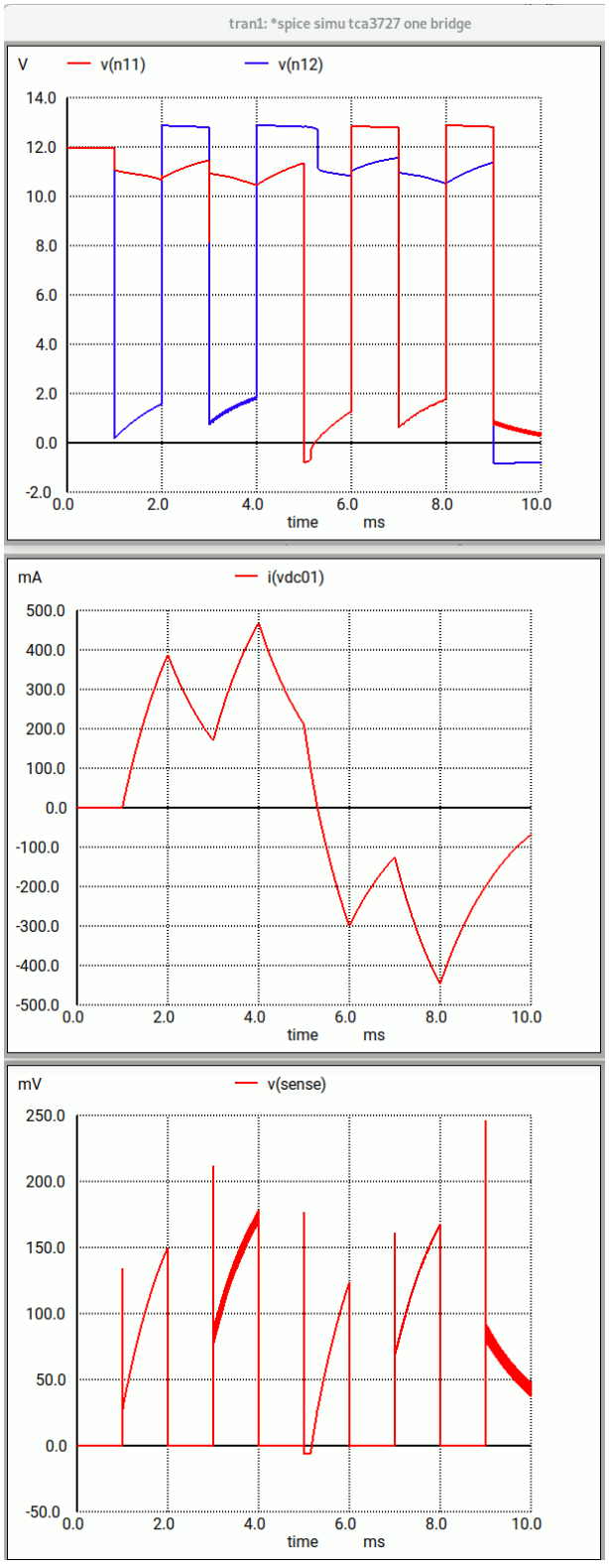


Figure 8.137: simulation of the TCA3721 power bridge

Note that there are two different current decay time constants. As long as a only the chopping low side turns off but the high side remains on the current flows through the high side and the free wheeling diode. The current decay is slow. When the polarity changes (chopper turns off AND high side turns off the back EMF has to overcome the drop of the winding resistance plus the battery voltage. The current decay becomes much steeper. During this fast current decay energy is flowing back from the inductance of the winding into the supply vbat (negative current consumption! The system temporarily acts as a generator feeding the power supply.) This energy has to be absorbed by the power supply or by blocking capacitors between vbat and pgnd. If the power supply doesn't allow reverse current (no operation in the second quadrant) the blocking capacitor will be charged at phase reversal. Following the worst case assumption that there are no losses due to the resistance of the winding and that there is now capability

of the supply system to swallow the energy the flyback energy can be approximated:

$$\frac{I^2 * L}{2} = V_{bat} * \frac{\Delta V_{bat} * C}{2}$$

Which can be reordered to

$$C = \frac{I^2 * L}{V_{bat} * \Delta V_{bat}} \tag{8.258}$$

With ΔV_{bat} being the permitted change of the supply voltage due to the back EMF of the winding. C is the capacity required for systems with a reverse supply protection diode between the power supply and the stepper motor driver.

Example: I=1A, $L = 20mH$,$V_{bat} = 12V$, $\Delta V_{bat} = 2V$ leads to $C = 417\mu F$.

8.6.3 MOS power transistor bridges

In the 1990 the RF emission of stepper motor drivers increasingly became a selling argument. In the 1980s the application engineers still were willing to solve EMC requirements by external filters between the hard switching power bridge with rise and fall times in the 10..100ns range and the cable to the motor that acted as an antenna. In the 1990 customer specifications requesting a defined dV/dt became common to reduce the effort of external blocking. Slower rise and fall times however increased the slope losses of the design. This was compensated by the design of power packages with exposed dice pad featuring lower thermal impedance. Improved board technologies (thicker metal layers of $105\mu m$instead of the standard $35\mu m$and in extreme cases embedded metal cores acting as heat spreaders) further helped to handle power stage losses of several W.

The capacitors used to control the dV/dt should be insensitive to the voltage. Miller capacities of power transistors didn't work well. So the standard design for dV/dt controlled stages became regulation loops using dedicated high voltage capacitors.

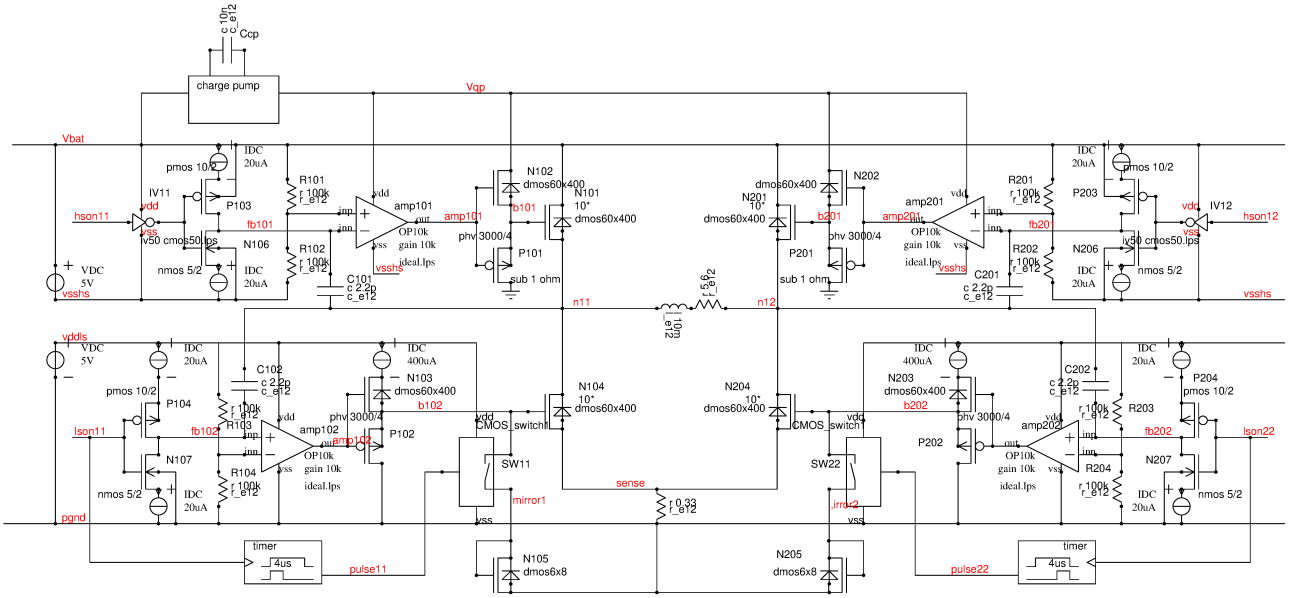


Figure 8.138: Concept of the power stage of L9935

To keep the high voltage capacitors small the slope generator stage uses fairly low current limiters in series with P103, N106, P104, N107.

The load of the stage ranges power bridge ranges from inductive to capacitive (stray capacity of the windings, cable capacity..). For this reason the amplifiers amp101 to amp202 are designed for high speed rather than precision. The open loop gain of the high side stage is limited by the amplifier alone. At the low side stage (amp102, follower stage N103, P102, power transistor N104 the situation is much more difficult because there are two stages producing a voltage gain: amp102 and power transistor N104 together with the impedance of the winding of the motor. Even worse the quadratic characteristic of the power transistor changes the gain dramatically with the load current. To keep the loop stable the solution is to operate N104 as a current mirror with N105 acting as a mirror diode. When the slope has ended we want N104 to be operated with the highest possible gate overdrive. For this reason after $4\mu s$the timer opens the switch SW11. This creates a little step in the signal but it was the only solution found to get the loop stable.

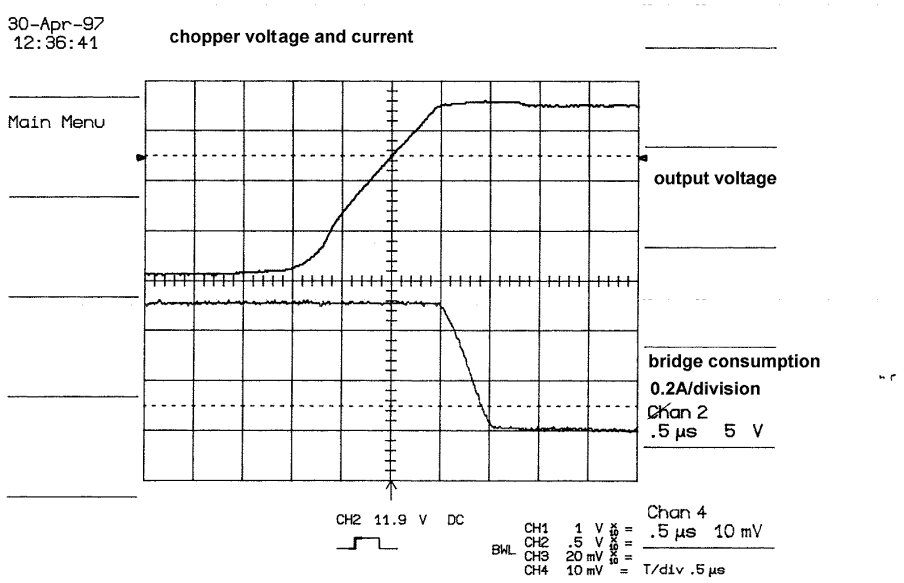


Figure 8.139: switching slope rising edge of L9935

The oscilloscope screenshot of the rising edge of the chopper signal shows that in spite of the effort spent the slopes are not ideal:

- The start of the rising edge of the voltage still has a ringing component.
- The dV/dt is regulated but the current still has a faster edge than the voltage.

Checking the spectrum at the bridge output the charge pump becomes visible. In the time domain picture shown above the charge pump is invisible because $55dB\mu V$ correspond only about 1.5mV peak to peak.The high emission level is caused by the high frequency of the charge pump (4MHz) and the high current consumed by the high side drivers and the amplifiers (0.5mA). This high emission could have been circumvented using a little boost converter rather than a charge pump. But customers rejected this idea because of the cost of the inductor.

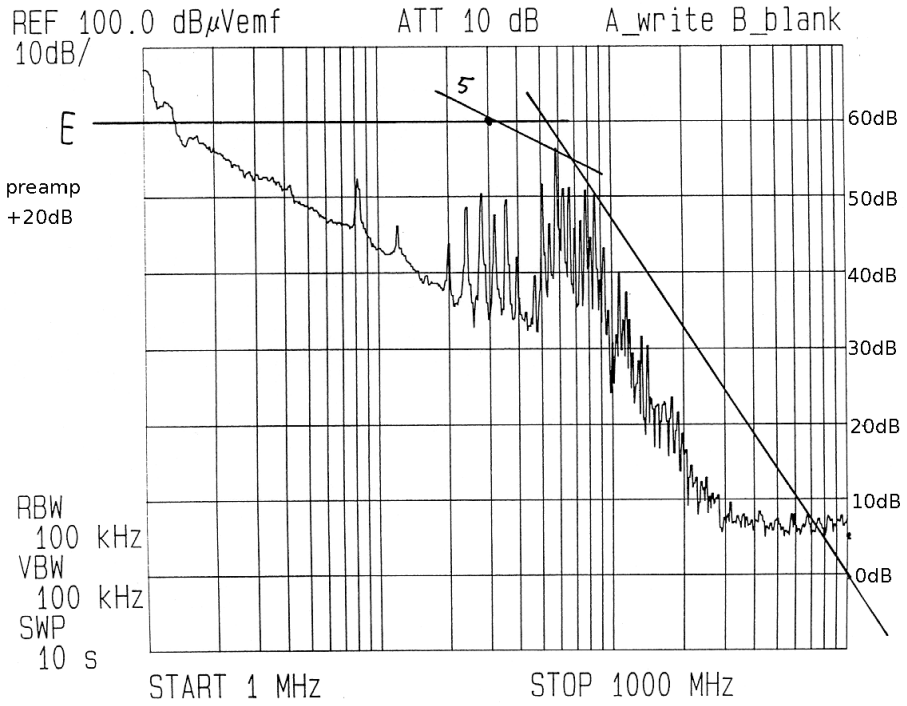


Figure 8.140: L9935 output conducted emission in $dB\mu V$

Since a 2 phase pump was used the base frequency is barely visible. The first significant line is the 2nd harmonic at 8MHz. At about 60MHz the resonance of the supply bond wires and the well capacities on the chip are boosting the harmonics of the charge pump to about $55dB\mu V$. Changing the package this resonance can change significantly.

At the end the RF emission improvement anticipated from the slower switching edges could not be realized due to the charge pump needed to supply the high side driver. The charge pump became the limitation of the improvement.

8.6.4 Cross conduction and break before make

There is a risk of creating a current shoot through when the high side N101 and the low side N104 are turned on simultaneously. The critical situation can occurre if the polarity of the bridge is changed and at the same time the low side transistor that is chopping to regulate the current still is on.

In a hard switching design without dV/dt regulation this shoot through can be accepted because the duration is only in the range of some 10ns. The energy is limited and thermal destruction of the transistors won't happen.

In a slow switching design with dV/dt limitation cross conduction can take place for several microseconds if no countermeasures are taken. A first draft break before make circuit measures the gate voltage of the low side power transistor and prevents turning on the high side before the gate of the low side is discharged. In addition there is a second protection measuring the gate voltage of the high side and blocking turn on of the low side while the high side still is on. The concept is shown in the following figure.

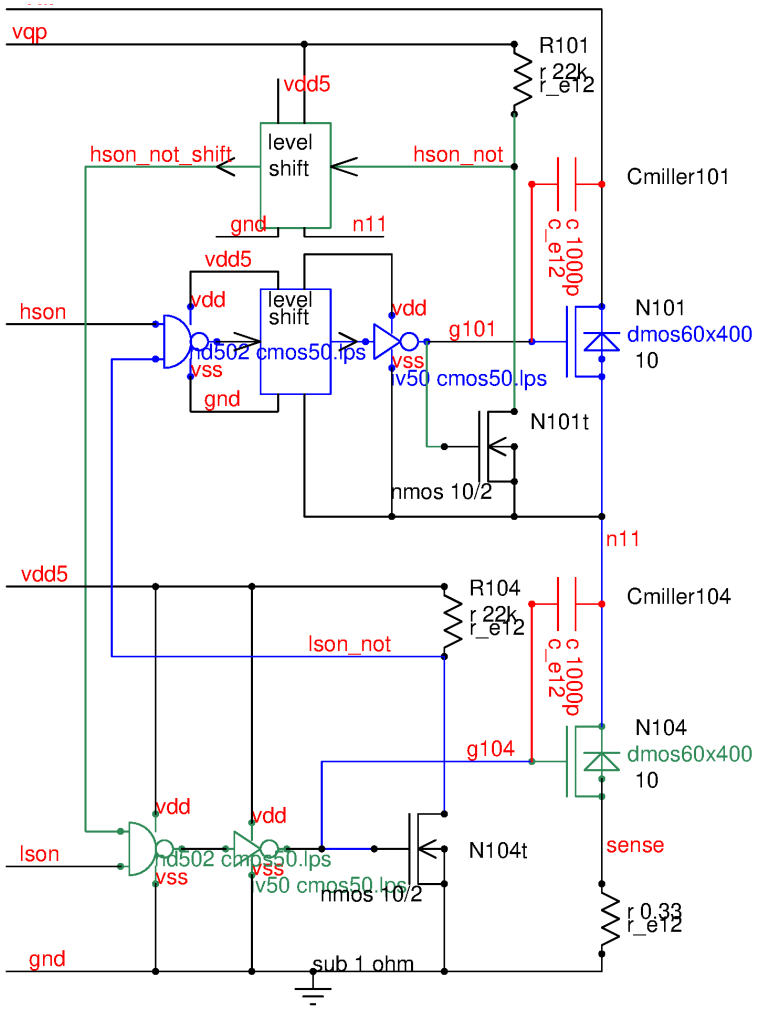


Figure 8.141: cross conduction protection with oscillation risk

The first draft circuit shown has two unintentional closed loops that can (and will!) oscillate. These loops are drawn in blue and green color.

Both power transistors have a big miller capacity (drawn in red).

Loop 1 (green loop): Let's assume we turn off N101 and turn on N104. As soon as the gate voltage of N101 falls below the threshold of N101t hson_not will go up and the already existing logic 1 can propagate to the gate of N104. N104 pulls down node n11 and Cmiller101 charges the gate of N101 and N101t again. The signal hson_not goes down blocking the turn on signal of N104 again.

Now n11 goes up again and Cmiller101 turns off N101 and N101t. We are back at the beginning of the cycle. N104 is allowed to turn on again. This leads to a nice oscillation in the range of some MHz.

Loop 2 (blue loop): Same same but different. This time we assume we want to turn on N101 while N104 is still on. lson_not blocks turn on of N101 until the gate voltage of N104 drops below the threshold of N104t. As soon as N104t turns off lson_not goes up allowing N101 to turn on. The rising edge on n11 changes the gate voltage of N104 and N104t via the miller capacity Cmiller104. This turns on N104 and N104t again.

Now lson_not goes down and turns off N101 again and the cycle starts again.

These two conditions are only hit at very special timings of the control signals hson and lson. In addition the load (inductive load, not shown) and the current flowing in the load plays a role to start the oscillation. On L9935A this condition was hit in about one out of 1000 steps. During simulation this condition never was met, but in practical application a loss of steps was observed and on the wafer prober the loop could be found scrolling through thousands of turn on events with a deep memory oscilloscope (You can't trigger on signals as long as you don't know what you are really looking for! There is no replacement for waveform memory except MORE memory and fast scrolling!) The circuit had to be improved. A turn on permission once committed may never be revoked to prevent oscillation. Additional logic had to be added.

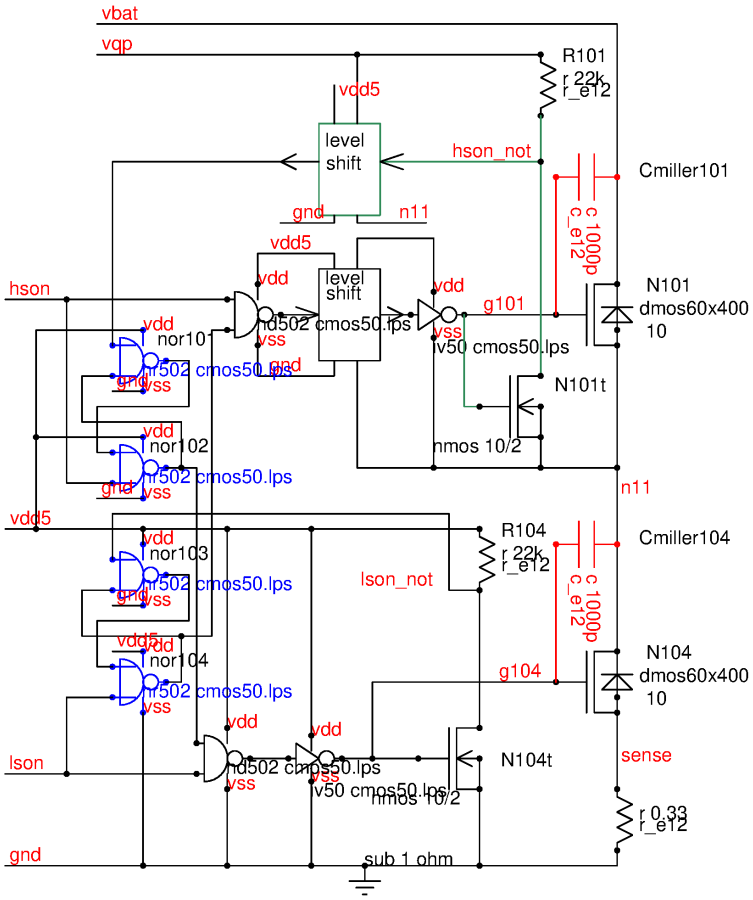


Figure 8.142: cross conduction protection with latches to prevent oscillation

The circuit added are the latches nor101, nor102 and nor103, nor104. As soon as the gate is discharged the information will be stored in the latches. Taking back this information isn't possible until the corresponding stage is intentionally turned on again. This way revoking a turn on permission of the opposite transistor by charges flowing through the miller capacities is prevented.

8.6.5 floorplan of stepper motor driver power stages

Each time the polarity of the bridge changes the low side diodes and the substrate diodes of the low side transistors carry the full load current. This excites a lateral NPN transistor. For this reason the low side transistors (N104, N204) must be kept far away from the sensitive analog circuits.

A lateral NPN draining current from the high side power transistor with the drain connected to the supply $V_{bat}$is much less critical. Placing the high side power transistors between the aggressive low side and the sensitive analog circuits is a classical floorplan for power bridges.

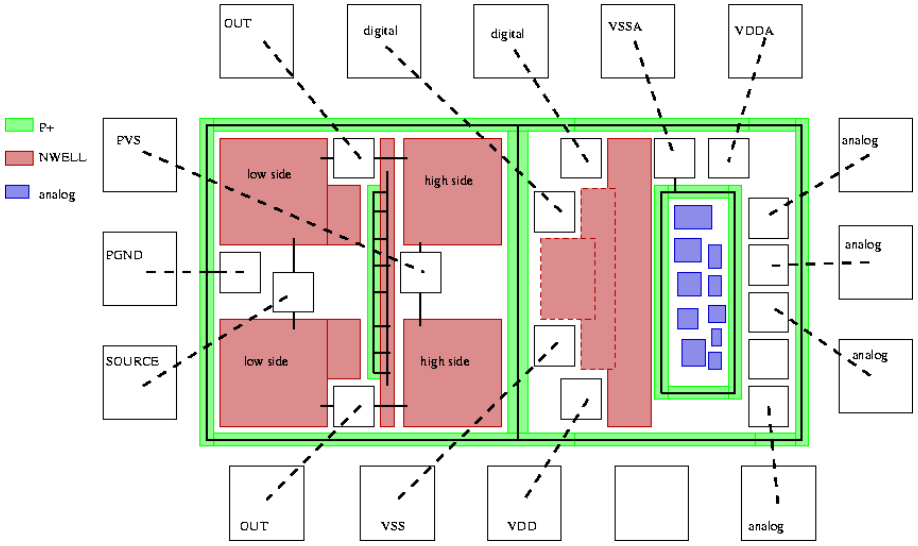


Figure 8.143: floorplan of an H-bridge designed for inductive loads

The low side is on the left side of the floorplan. During phase reversal the tubs of the low side transistors act as emitters of a lateral NPN. Between the low side and the high side the substrate is intentionally not connected to ground but to an nwell acting as a first collector of the lateral NPN. When the low side is injecting electrons into the substrate the nwell between the low side and the high side is pulled negative. Since the nwell is connected to the substrate this creates a drift field in the substrate and most of the electrons will move to the left side to the edge seal of the chip.

Electrons passing the nwell and the floating substrate connection have to travel a significant distance under the high side. The a big part of the electrons either recombine with holes or will move upward and will be removed by the drain regions of the high side acting as a second collector.

Logic can be reasonably well protected against electrons in the substrate provided the nwell (isolation) of the logic is connected tightly to the logic supply. So the logic can act as a second recombination zone for minority carriers in the substrate.

Electron reaching the analog functions on the right side of the chip will modify the currents flowing in all components bringing their own epi regions (usually high voltage transistors and bipolar transistors). The higher the current density chosen for the analog circuits and the higher the distance from the low side power transistors the less sensitive the analog functions will be.

Restricting as much of the analog circuits to low voltage CMOS allows shielding most of the analog circuits by the nwell that can be tightly connected to a supply.

If multiple power bridges are placed on one chip the floorplan shown can be double and flipped horizontally. The we exactly end at the floorplan of the L9935.

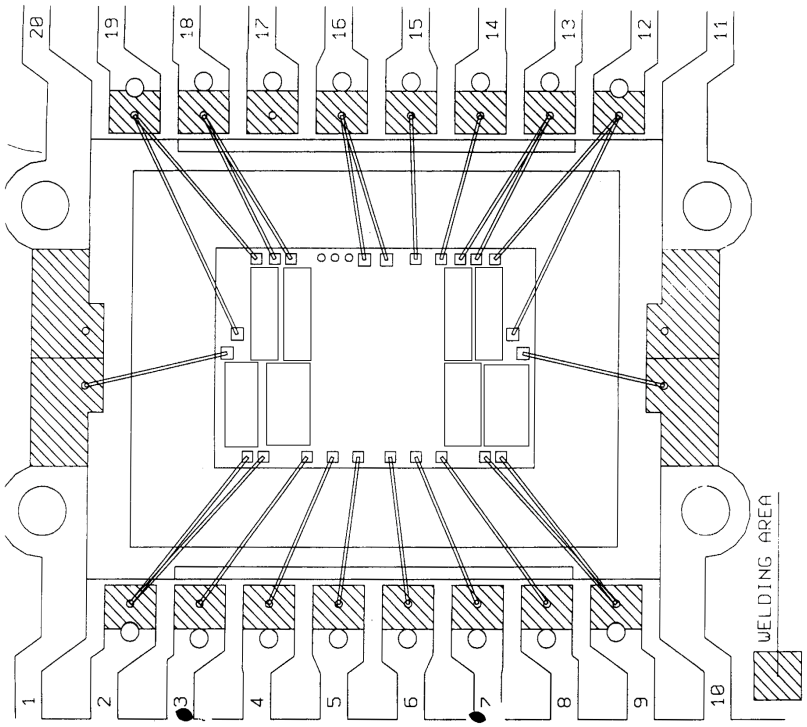


Figure 8.144: Floorplan of L9935

On the L9935 the low side transistors intentionally are at the left and the right edge of the chip. The high side transistors are more in the middle acting as a spacer and as a collector extracting electrons injected into the substrate. The sensitive analog circuits and the logic are in the middle of the chip as far as possible away from the low side transistors.

8.6.6 Regulation of the current

To minimize the losses and the power dissipation of the chip the regulation uses a chopper. The chopper regulation loop measures the voltage drop over the sense resistor between node sense and pgnd. One regulation loop per winding is required. Since most low cost stepper motors only have two windings we need two regulation loops. For high performance applications there are some manufacturers producing motors with more than 2 windings too (I have seen up to 5 windings). In this case we need more than two current regulation loops and the control sequences get significantly more complex. The benefit of more winding is that the motor can be controlled with higher accuracy and runs smoother and with less significant resonances and torque dips (at least the motor manufacturers claim this is the case.)

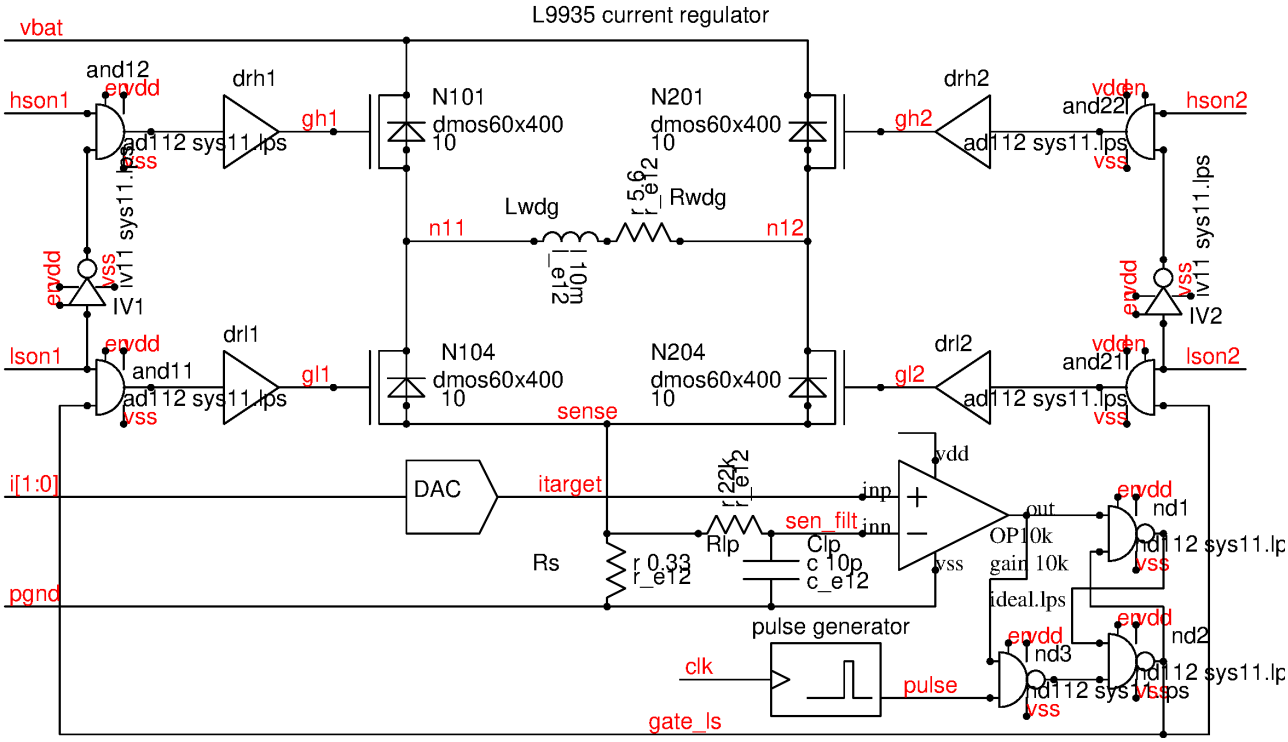


Figure 8.145: current regulation loop using a fixed clock frequency

The current regulation resets the latch nd1, nd2 each time the the target current is exceeded. The gates and11, and21 turn off the low side stage.

The clock signal periodically turns on the low side driver again. NAND gate nd3 prevents a simultaneous set ad reset of the latch nd1, nd2. The reset coming from the comparator is dominant over the set signal of the clock generator and the pulse generator.

The DAC sets the target current to allow half step or quarter step mode operation.

and12 and and22 prevent simultaneous activation of the low side and the high side of the same half bridge.

On drawback of the periodic turn on is that under certain conditions pulse skipping may take place. Pulse skipping leads to operation at a divided clock frequency. This may lead to audible operation although the clock frequency is higher than 20kHz.

An alternative control scheme is the fixed turn off time operation shown in the following schematic. Fixed turn off time operation avoids pulse skipping but the frequency varies with the inductance of the windings, the back EMF induced by the rotating rotor and the supply voltage vbat. Fixed turn off time operation is preferred for single H-bridge drivers such as the TEA3717 because is saves the effort of building a clock generator. On the other hand the monoshot (pulse generator) requires a capacitor in the range of some nF. So the more bridges are built using a fixed off time concept the more external capacitors are needed.

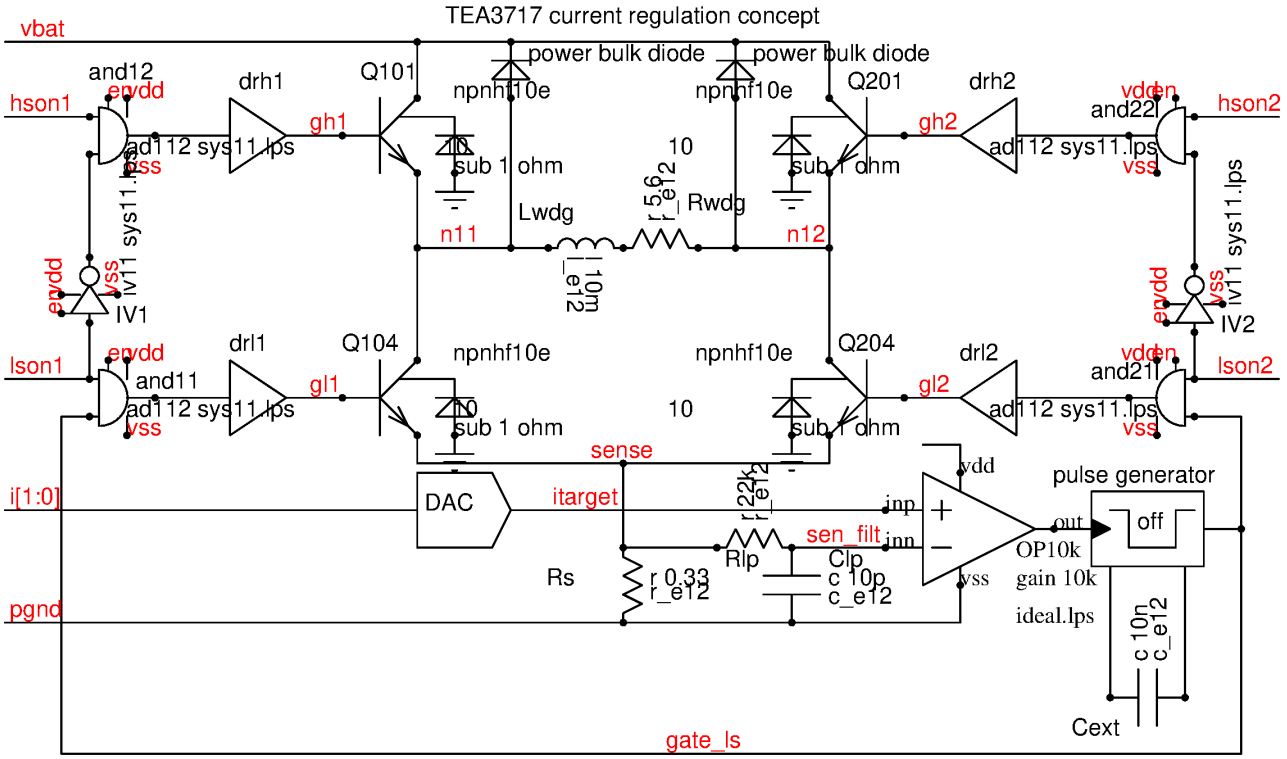


Figure 8.146: current regulation loop with fixed turn off

An other variant of the circuit is to reduce the pulse skipping using a slope compensation concepts as in current controlled switchmode power supplies. It helps, but it doesn't work as good as in switchmode power supplies because the back EMF of the rotating rotor changes the effective inductance much more than any load changes in switchmode power supplies. Furthermore the average current flowing in the winding changes with the ON-time of the low side. This means with a slope compensation we pay the reduction of pulse skipping with a loss of accuracy. Probably it is due to this reduction of accuracy that (up to my knowledge) nobody is using the slope compensation to regulate the current in stepper motor drivers.

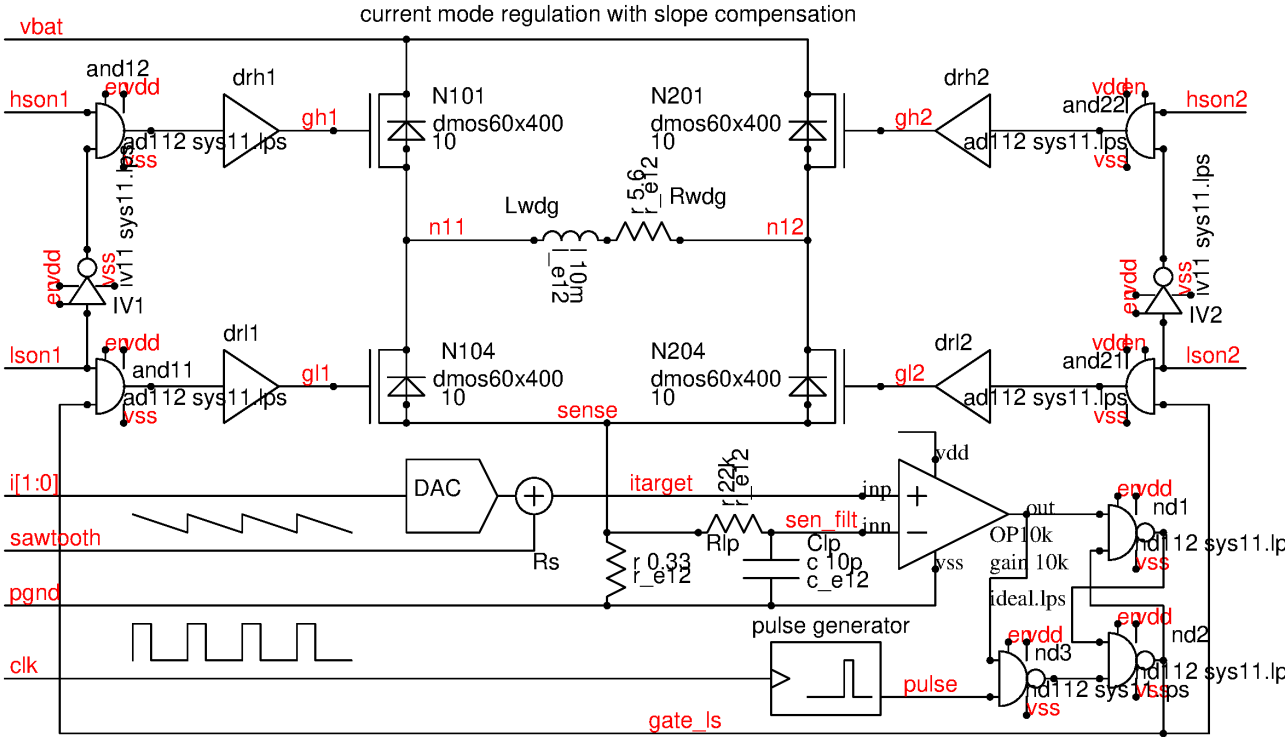


Figure 8.147: clocked current regulation with slope compensation

The loss of accuracy is proportional to the amplitude of the sawtooth signal used for the slope compensation.

Table 43: properties of current regulation methods for stepper motor drivers

method	pro	contra	other remarks
fixed frequency without slope compensation	defined operating frequency	pulse skipping is likely	cheap for multi channel
fixed time off	no pulse skipping	variable frequency	one timer per bridge
fixed frequency with slope compensation	reduced pulse skipping, defined frequency.	current changes with duty cycle	cheap for multi channel

Offset chopping The idea of offset chopping is to run the choppers of both full bridges out of phase to reduce peak currents. Ideally only one full bridge consumes current at a time while the other one is in freewheeling mode. Practically this only works part of the time when the duty cycles of both choppers are below 50%. The most simple way of implementing offset chopping is to trigger one full bridge with the positive edge of the clock and the other full bridge with the negative edge of the clock. The following figure shows offset chopping and current consumption of the L9935 driver.

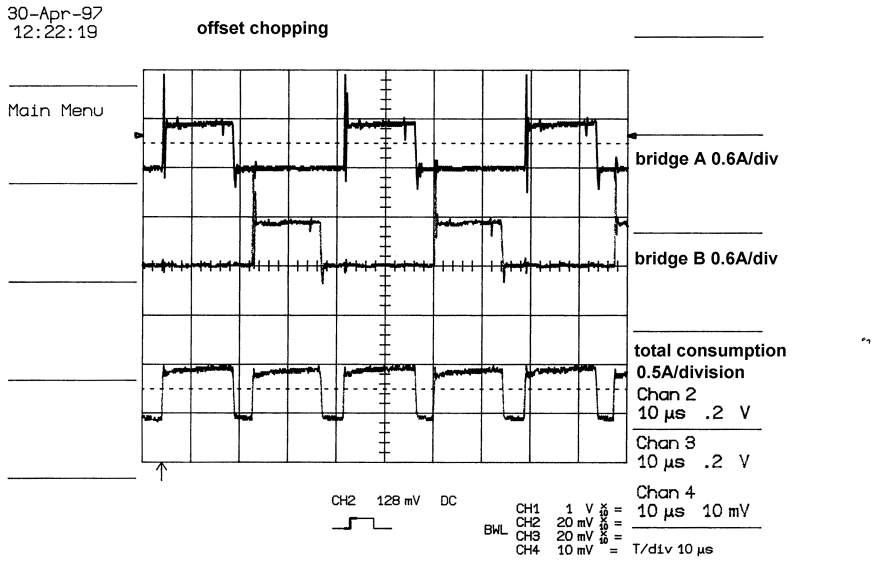


Figure 8.148: offset chopping and current consumption

Offset chopping only works in clock triggered designs. It doesn't work in constant off time systems because there the frequency is variable and both full bridges are shifting the phase versus the other full bridge.

half step mode and micro step mode Full step mode operates the windings always at full current. Whenever there is a current direction change the back EMF of the winding has to work against the supply voltage. This leads to a fast current decay and the current in the windings follows the command with only a small delay.

In half step mode there are times the current is supposed to decrease but since the polarity is unchanged the current only shows a slow decay. From full step to half step the high side remains on and the chopper simply drops to the lowest possible duty cycle as long as the current is too high. When the chopper transistor is off the current flows in a local loop consisting of the high side transistor (which is in on state), the winding and the resistance of the winding and the flyback diode feeding the current back into vbat. The only voltage drops in the loop are the drop over the high side, the forward voltage of the flyback diode on the opposite side (other half bridge) and the I-R drop over the winding resistance.

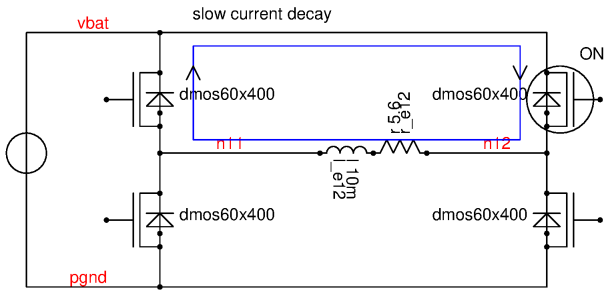


Figure 8.149: current flow during slow current decay

The change of the current flowing in the loop calculates as:

$$\frac{dI}{dt} = -\frac{I^2 R_{on} + I^2 R_{wdg} + V_f + EMF}{L} \tag{8.259}$$

EMF is the electromagnetic force induced into the winding by the rotating rotor. The faster the motor rotates the more energy is taken out of the field and the faster the current decays.

From half step down to 0A both sides of the bridge can be switched to high impedance. The current flows through the flyback diodes against vbat.

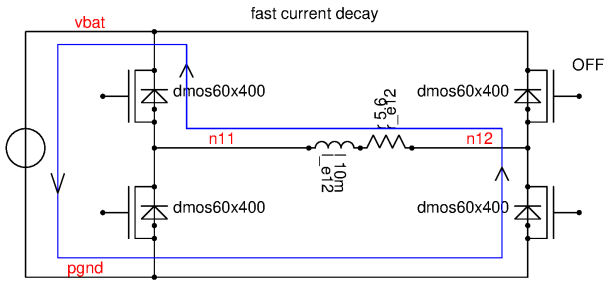


Figure 8.150: current flow during fast current decay

When the bridge is in high impedance mode the current decay calculates as:

$$\frac{dI}{dt} = -\frac{2 * V_f + V_{bat} + EMF}{L} \tag{8.260}$$

The higher the supply voltage and the faster the motor rotates the faster the current decays. Usually the drop over the diodes is negligible compared with the other contributors.

When the bridge turns on again the current increase is defined by the load inductance, the resistance, the drops over the power transistors and the supply voltage.

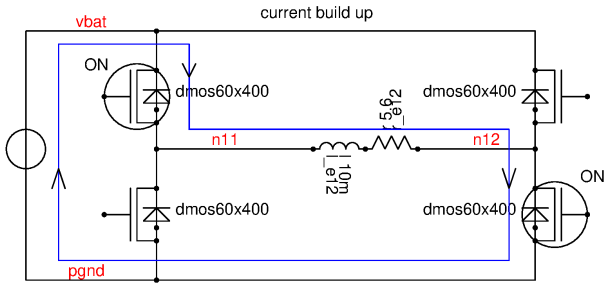


Figure 8.151: current flow when bridge is on

The current increase can be calculated

$$\frac{dI}{dt} = \frac{V_{bat} - I^2 R - 2 * I^2 R_{on} - EMF}{L} \tag{8.261}$$

The current increases the faster the higher the supply is and the slower the motor rotates. In a well designed system the drops over the power transistors usually can be neglected.

In real application the behavior is a bit less ideal because the rotor has a variable air gap to the stator. So EMF not only changes with the speed of the rotor but also with the change of the air gap and with the angle between the field and the rotor position. In reality the current change fluctuates around some kind of average value for current build up, fast decay and slow decay.

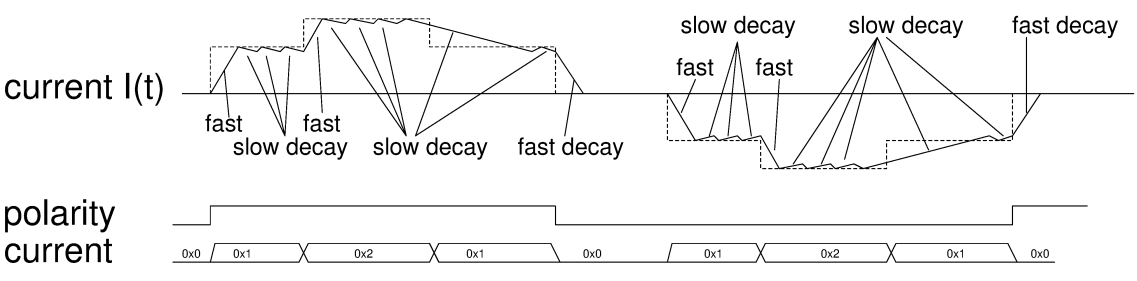


Figure 8.152: fast and slow decay and current programming

In the plot the dashed signal is the ideal target current. The current really flowing in the inductive load is drawn as a solid line. The highest deviation between the target current and the current really flowing is found when the current is reduced from +100% to about + 50% and from -100% to about -50%.

One possibility to make the current better follow a decrease of the current is to switch the bridge into a high impedance state for a limited time each time the current is to be reduced.

A second possibility is to introduce a second chopping mode for fast decay of the current. In this fast decay chopping mode the high side stage and the low side stage are chopping simultaneously. This forces the flyback current to flow via the bulk diodes instead of free wheeling it via the high side transistor. This kind of fast decay chopping mode produces more losses than the normal chopping mode. So a fast decay chopping mode should only be active for a limited time each time a current reduction is intended without changing the polarity.

8.7 Galvanic isolation circuits

For high operating voltages often different parts of a circuit have to be isolated galcanically from each other. A typical application of such galvanic isolation circuits are driver stages for primary side chopping switchmode power supplies or drivers for high power electrical motors (in the range of up to 1000V). There is a big variety of implementations of such systems existing on the market.

1. transformer coupling
2. capacitive coupling
3. opto couplers
4. piezzo couplers

Transformer coupling is a common method using external transformers. Integration of transformers on chip is a fairly new approach because on chip the inductors available are only in the range of fractions of nH to some nH. First on chip transformer coupled circuits exist since about the end of the 1990s.

Capacitive coupling on chip is a more frequent approach. The most simple form is a capacitive driven latch. (See level shift circuits). Since the initial state of such a simple latch in undetermined more sophisticated solutions use a carrier that is detected on the receiving side.

Opto couplers are the first implementation of galvanic decoupling in a single IC package. First opto couplers are dating back to the 1970s. The disadvantage of opto couplers is the need of exotic materials for the LED (for instance GaAs) and a more complex packaging process. A second issue is the aging of the LED.

Piezzo couplers in theory can transport a lot of energy. This may one day remove the requirement of providing an extra power supply for the receiving side. Although first concepts have already been discussed around 2000 I'm not aware of any commercial use of such a system. Piezzo coupled systems however are not new. In the 1960s it was common to build the 64us delay lines of PAL color TV receivers with an acoustic delay line. So piezzo coupled systems are not completely new.

In the following only transformer coupling is discussed.

8.7.1 Transformer coupled systems

Transformer coupled systems were in common use for valve amplifiers. With the advent of semiconductors transformers more and more disappeared. The most simple method is a pulse transformer directly driving a power transistor. This method only works in a certain frequency range but not for DC. For a switchmode power supply this however can be sufficient.

To transfer signals down to DC the receiving part of the circuit must have some kind of storing element. Usually this is a latch or a flip flop. The output of the flip flop can drive a power transistors down to DC.

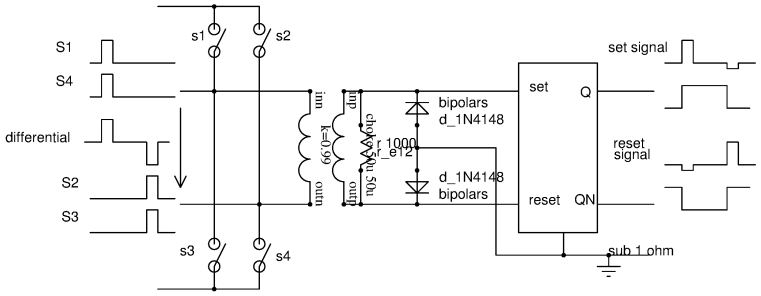


Figure 8.153: Using a latch as a storing element

One of the problems of the circuit is that the initial state of the latch after turning on the secondary side power supply is unpredictable. A second problem is that short pulses coupled into the circuit (for instance due to a big

power transistor switching) may get latched as well. The simple circuit shown here is not robust against RF and pulse distortion. Some improvement can be achieved placing a filter between the transformer and the inputs of the latch.

In case of AC operated systems the storing element can be the power thyristor (also called SCR or silicon controlled rectifier) or triac itself. In this case the thyristor or triac is turned on by a pulse and turns off at the zero crossing of the supply current.

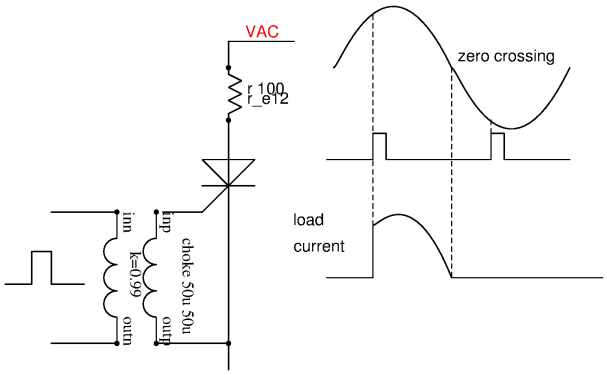


Figure 8.154: Driving a thyristor with a transformer

Using a thyristor as a storing element is elegant in the sense that the thyristor usually isn't too fast and requires gate currents in the range of some 10mA to trigger. So false triggering due to short distortions is less likely than using a fast flip flop. Additionally turning on the power supply the initial state is OFF. (At least as long as dV/dt of the supply voltage doesn't exceed it's critical value that leads to capacitive turn on. The maximum dV/dt usually can be found in the specification of the thyristor). The low speed of thyristors is both, an advantage for robustness and a disadvantage for dynamic switching losses. Therefore thyristor switching is limited to few kHz only.

To make systems more robust in stead of directly transmitting the control pulse a carrier can be used. The information whether to turn on of off the power device is in the modulation. Carrier based systems using a narrow bandwidth where the system is sensitive to distortion (So pulse distortions having a wide bandwidth but only have very little energy in the sensitive frequency range.) can be very robust. In addition check sums or parity bits can be added to the modulation.

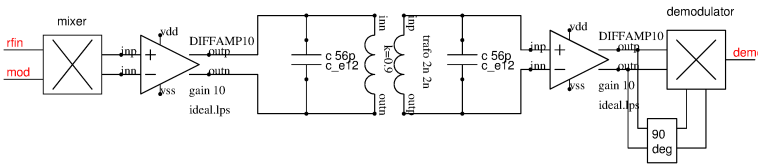


Figure 8.155: Information transfer with transformer and carrier

Such systems using carrier frequencies in the range of 500MHz to 2GHz can be integrated on silicon. Usually the transmitter side is on one chip and the receiving side is a second chip sitting in the same package. Size of the transformer is about 30u diameter, 4 to 5 windings. The isolation between the windings depends on the voltage to be isolated and the quality of the oxide. For thick oxides usually CVD (chemical vapor deposition) is used. CVD provides thick oxides but the quality is less than the quality of a dry grown gate oxides. Typically CVD oxides can handle 0.1V/nm or $100V/\mu m$. A 1000V isolation thus requires about $10\mu m$ of isolation.

To transfer a significant amount of power bigger transformers are required to accommodate a bigger part of the magnetic field. Most of the area simply is required to bring down the wire resistance using wide wires.

Designing a transformer on chip always is a careful optimization of number of windings, trace width to reduce the resistance and frequency of operation. This can be done with field solvers like fasthenry. Creating the netlist for fasthenry manually is too error prone. In stead a perl script can be used to produce the netlist.

Fasthenry provides a complex impedance. The main inductance and the stray inductance can be determined operating the simulated transformer in open mode and in shorted mode. For simplicity here the most important equations of section 4.5.2 are repeated:

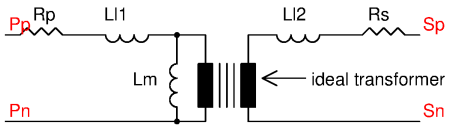


Figure 8.156: equivalent circuit of a transformer

The coupling can be determined comparing the inductance open and shorted.

$$K_{ind} = \sqrt{1 - \frac{L_{short}}{L_{open}}} \tag{8.262}$$

$$L_{l1} = (1 - K_{ind}) * L_{open} \tag{8.263}$$

$$L_{l2} = (1 - K_{ind}) * \frac{L_{open}}{N^2} \tag{8.264}$$

On chip usually the transformer is 1:1 to use the same layout on the primary side and on the secondary side.

$$L_m = K_{ind} * L_{open} \tag{8.265}$$

Running a fasthenry simulation the substrate must be included as a resistive ground plane because depending on the resistivity of the silicon under the transformer there can be significant eddy current losses.

Here comes a little example

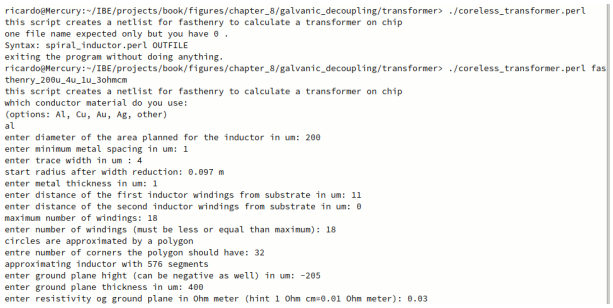


Fig.8.5.1.5: Screenshot of the inputs of a little perl script creating the fasthenry netlist

The results of fasthenry at 1GHz for open and shorted operation are shown below.

```
Open operation
Impedance matrix for frequency = 1e+09 1 x 1
        42.5971       +158.028j

shorted operation
Impedance matrix for frequency = 1e+09 1 x 1
        64.2723       +76.5142j
```

The resulting parameters can be calculated:

$$R_p = 42.5971\Omega$$

$$K_{ind} = \sqrt{1 - \frac{76.51}{158.03}} = 0.72$$

$$L_{OPEN} = \frac{Z_{ind}}{2 * \Pi * f} = 25.14nH$$

$$L_{l1} = (1 - K_{ind}) * L_{open} = 7.04nH$$

$$L_m = K_{ind} * L_{open} = 18.1nH$$

The secondary side simply uses the symmetry:

$$R_s = R_p = 42.5971\Omega$$

$$L_{l2} = L_{l1} = 7.04nH$$

A nice little transformer usable for about 10mA to 20mA, 2V at 1GHz or higher.

8.8 Near field communication

Near field communication can be distinguished in two classes. Passive near field communication uses the energy provided by the field to supply the transponder chip (Typical example: Immobilizer systems with a communication distance of just a few cm). Active near field communication has a power supply for each partner in the network (Typical example: Keyless entry with communication distances up to some meters).

8.8.1 Passive near field communication

In a passive near field communication system the transmitter (in most cases the master) provides a powerful magnetic field usually in the frequency range of 100kHz to 150kHz. The coupling between the two partners (master and slave) can be regarded as a resonant transformer coupling. The partner system (slave) draws it's supply energy out of the magnetic field. Once the supply capacitors of the slave is charged the system goes through a reset and then starts operating. Communication from the slave to the master in most cases is done by damping the field with a data pattern. Data transfer from master to slave is done modulating the transmitter energy. The resulting communication process looks like this:

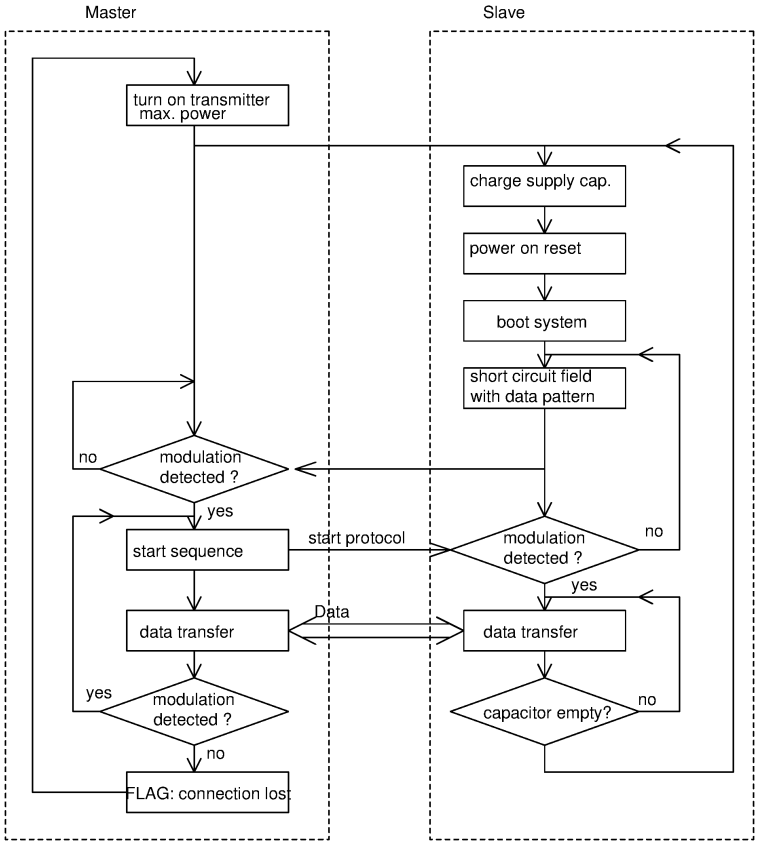


Figure 8.157: Simplified software flow of master and slave in a near field communication system

The most simple circuit for such an application is shown below. The master drives a resonant tank. The inductors Lm and Ls are designed to have a big stray field. They are loosely coupled. The power amplifier PA drives the inductor Lm. The data to be sent is modulated either changing the gain of PA or the duty cycle. The drive signal at mode power can be a sine wave or a rectangular signal. The resonant tank C1, Lm, C2 forces a sinusoidal current even if the driver provides rectangular pulses.

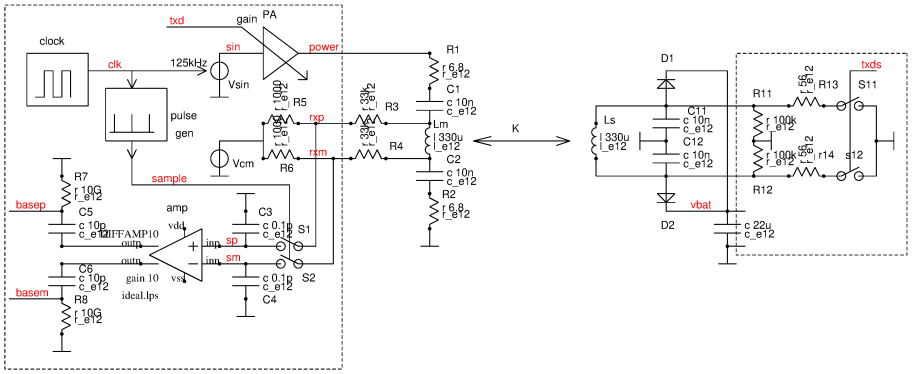


Figure 8.158: Near field communication system

The slave picks up the magnetic field. Diodes D1 and D2 rectify the signal and charge the supply of the slave system. To transfer data from the slave to the master the master drives with a constant power. The slave periodically damps the system closing switches S11 and S12. This changes the amplitude on the driving side.

The master during receive mode samples the sine wave in the resonant tank C1, C2, Lm. The switches S1 and S2 are synchronized with carrier. If the sampling takes place only once per period the sampling already demodulates the signal.

Due to the loose coupling K between the master and the slave the modulation index at the pins rxp and rxm is extremely low. It can be as low as 10^{-5}or -100dBc. This low modulation index sets extreme requirements for the amplitude stability of the transmitter and for the clock stability and the clock jitter.

In practical systems the driver PA offers peak to peak voltages in the range of some 10V. Due to the resonance ($Q \approx 5..20$) the voltage across Lm can reach more than 100V. To protect the receiver the signal must be attenuated by R3, R4, R5, R6. At nodes rxp and rxm the carrier typically has 1V to 2V peak to peak and the modulation is in the range of $50\mu V$ to some mV. Typically the amplifier amp consists of several stages with AC coupling. The AC coupling is represented by C5, C6 and R7, R8. In practical implementations resistors R7 and R8 are realized using switched capacitor techniques.

The following figure shows the operation with well tuned inductive antennas. The power amplifier drives the resonant tank in an asymmetrical way. This way the switching noise is mainly in one half wave of the signal while the sampling is done during the undistorted half wave.

8.9 Digital analog converter (DAC)

A digital to analog converter converts a digital signal into an analog signal. In most case the digital input is binary codes and the output is a more or less linear representation of the binary input as a voltage or current.

Besides being the interface of the logic to the physical outside world of the chip DACs are used an building blocks inside analog to digital converters (ADCs).

DACs can be designed as current sources or voltage sources or as charge pumps. In a charge pump DAC the digital input is the number of pumping events (cycles) of the charge pump.

Theoretically the dual circuit would be an inductor and the output would be current - but I have never seen that kind of implementation intentionally operated as a DAC. A switchmode current supply can be regarded as an inductor DAC if it is operated in burst mode controlling the number of pulses. If somebody plays around with superconductors and Josephson circuits it may in fact make sense!

In the following we rather stick with more classical implementations that really will be encountered on normal chips.

8.9.1 Error types of DACs and ADCs

DACs and ADCs have a quantisized digital side and a non quantisized analog side. At an ADC (Analog o digital converter) the input side is linear and the output is quantisized. At a DAC (digital to analog converter) the input is quantisized and the output is linear in the sense that the next processing stage has a linear input. Of course converting from discrete digital values into an analog signal leads to discrete analog values again. Nevertheless if we want to create a linear ramp the closest approximation we can do is a staircase function.

Quantization error: The difference between the linear ramp and the discrete values can be regarded as a quantization error. This applies for both, DACs and ADCs. The following figure shows the concept:

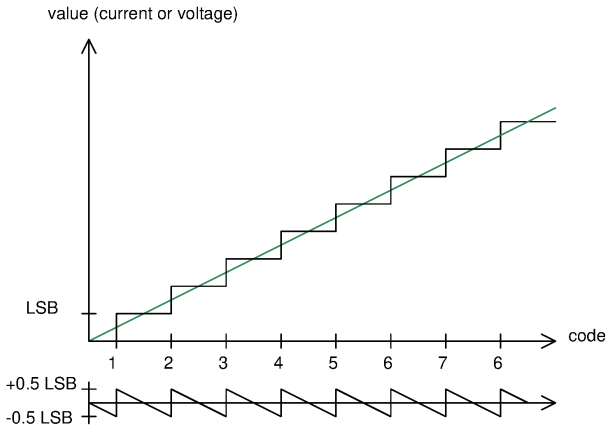


Figure 8.159: quantization error of a discrete system

The system is almost perfect. The set wise discrete curve follows the linear green one in the best possible way. The height of the steps are called the resolution of the discrete system. Often it is abbreviated as LSB (least significant bit). The most interesting point is the deviation between the linear function and the approximation by the step function. The bottom curve shows the deviation V(quant)-V(lin). In the perfect example the quantization error always remain between:

$$-V(LSB)/2 \leq V_{err} \leq V(LSB)/2 \tag{8.266}$$

Knowing the number of bits N and the full range signal $V_{full}$the step height and the maximum quantization error of this ideal system can be calculated.

$$V(LSB) = V_{full}/2^N \tag{8.267}$$

The maximum quantization error becomes

$$V_{errmax} = 2^{-(N+1)} * V_{full} \tag{8.268}$$

In signal processing it is common to describe an error signal as an effective voltage. So we have to calculate the effective voltage of the saw-tooth signal of the bottom curve. Just to remember:

$$V_{eff}^2 = \frac{1}{T} * \int_0^T V(t)^2 dt$$

The integration time is at minimum one period T. If the signal is not such a nice periodic one the integration has to run over n periods. Then we get:

$$V_{eff}^2 = \frac{1}{n*T} * \int_0^{n*T} V(t)^2 dt$$

For one period the triangular signal can simply be described as:

$$V(t) = \frac{V(LSB) - 2*t*V(LSB)/T}{2} = \frac{V(LSB)}{2} * (1 - 2t/T)$$

The effective voltage becomes:

$$V_{eff}^2 = \frac{V(LSB)^2}{4*T} * \int_0^T (1 - 2t/T)^2 dt$$

$$V_{eff}^2 = \frac{V(LSB)^2}{4*T} * (T - \frac{4T}{2} + \frac{4T}{3}) = \frac{V(LSB)^2}{12}$$

The effective quantization error voltage becomes (it is called quantization noise, so we give it a new name V_{qn})

$$V_{qn} = \frac{V(LSB)}{\sqrt{12}} = \frac{V(LSB)}{2*\sqrt{3}} \approx \frac{V(LSB)}{3.464} \tag{8.269}$$

Things are getting interesting if we try to represent a sine wave with our converter. The biggest sine wave we can represent is:

$$V_{pp} = 2*\sqrt{2}*V_{effmax} = V_{full} = 2^N * V(LSB)$$

transforming this equation leads to

$$V_{effmax} = \frac{2^N * V(LSB)}{2*\sqrt{2}} \tag{8.270}$$

Comparing the biggest sine wave we can represent and the quantization noise leads to a signal to noise ratio. Since this is in a linear scale it is called SN_{qlin}.

$$SN_{qlin} = \frac{V_{effmax}}{V_{qn}} = \sqrt{\frac{3}{2}} * 2^N \tag{8.271}$$

Using a logarithmic scale in dB is a bit more convenient.

$$SN_{qdB} = 20 * log10(SN_{qlin}) = 20 * \frac{log2(\sqrt{\frac{3}{2}} * 2^N)}{log2(10)} = 20 * \frac{N + log2(\sqrt{3/2})}{log2(10)} = (6.0207 * N + 1.7609)dB \tag{8.272}$$

Well, most people just simplify it to

$$SN_{qdB} \approx (6 * N + 1.76)dB \tag{8.273}$$

This is the best signal to noise (in dB) you can get using an N bit quantization of a sine wave.

Gain error: If the reference is wrong the steepness of the staircase function deviates from the linear function it is intended to approximate. Usually the gain error is expressed using the minimum and the maximum values of the (target) linear function ($V_{linmin}$and V_{linmax}) and the minimum and the maximum values of the staircase quantisized function ($V_{qmin}$and V_{qmax}).

$$Err_{gain} = 100 * \frac{(V_{linmax} - V_{linmin}) - (V_{qmax} - V_{qmin})}{V_{fullrange}}\% \tag{8.274}$$

In a DAC the expression $(V_{qmax} - V_{qmin})$ follows the reference voltage . If the reference is too high the staircase function gets too steep. If the reference is too low the staircase function gets too flat.

Differential non linearity (DNL) The differential non linearity is the deviation of single steps from the ideal value. The following figure shows an example.

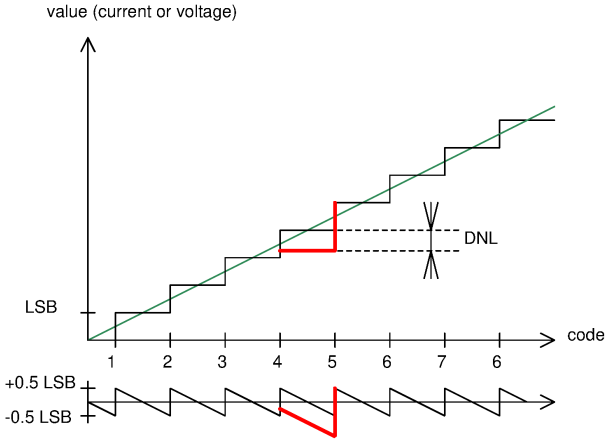


Figure 8.160: Differential non linearity of the red curve

From code 4 to code 5 the real curve of the DAC drawn in red deviates from the ideal curve drawn in black. This difference is called the differential non linearity. If the differential non linearity gets bigger than one LSB the code of the converter gets non monotonous. This effect sometimes is also called 'missing code'. In a regulation loop an ADC or a DAC with non monotonous behavior can lead to oscillation of the loop.

Integral non linearity (INL) Integral non linearity is found if errors caused by several differential non linearities accumulate.

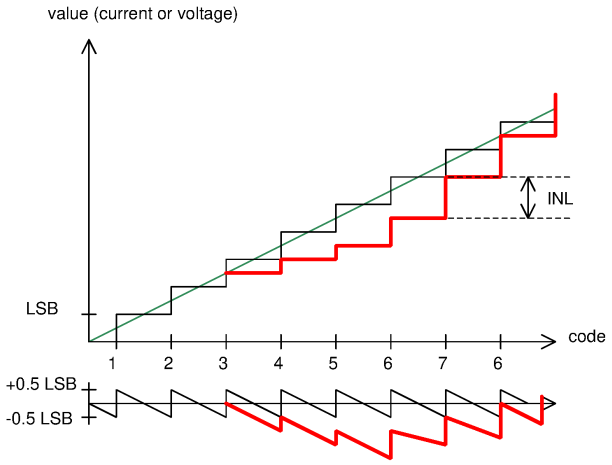


Figure 8.161: Integral non linearity of the red curve

The integral non linearity is the maximum deviation of the real converter (drawn in red) from the ideal converter (drawn in black).

8.9.2 DAC using resistors

A DAC built using resistor is the most simple possibility to build a DAC ([77] p.752). Well, unfortunately it usually is the one the performs worst on ICs because it requires very low resistive switches. Furthermore resistors covering several magnitudes of different values are area consuming. So simple resistor DACs are usually restricted to few bits (4 bit still may be reasonable, but this is about the limit)

The differential non linearity (DNL) mainly depends on the matching of the resistors of the MSB (R31 and R32) and on the Rdson of P3 and N3.

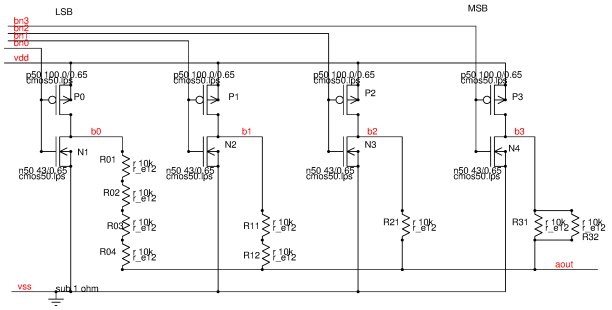


Figure 8.162: resistor DAC

- Highest output voltage is VDD (bits bn0 to bn3 all are 0).
- Lowest output voltage is VSS (bits bn0 to bn3 are 1).
- The output impedance is constant.
- The signal can be scaled by adding a load resistor.
- Load current on the reference voltage changes with code!

Since we have 16 possible values the LSB becomes (V(VDD)-V(VSS))/(16-1).

Example: 4 bit DAC, 5V supply leads to LSB=0.333V

As long as the resistance of the switches is low compared to the resistors this simple DAC works quite well. Nevertheless the binary weighted DAC has glitches in the range of the switching delay whenever higher weight bits change.

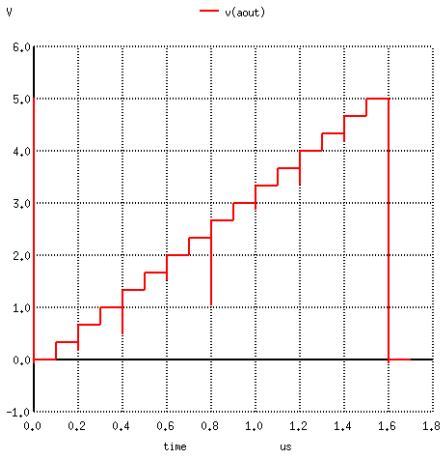


Figure 8.163: performance of an almost ideal binary weighted resistor DAC

In the next simulation the width of the transistors was significantly reduced (PMOS from 100um to 1um, NMOS from 43um to 430nm) . The performance suffers immediately. The worst problems are found when the MSB switches in the middle of the range. There we even find missing code. The characteristic is no more monotonous. In a regulation loop this can lead to oscillation because at the missing bit we invert the feed back. (Well, I agree ideally the transistor width should be scaled with the weight of the bit. But even then we run into trouble because we have to match Rdson of PMOS and NMOS transistors. So in corner simulations we will find issues for small transistors even if we scale transistor width with the bit weight)

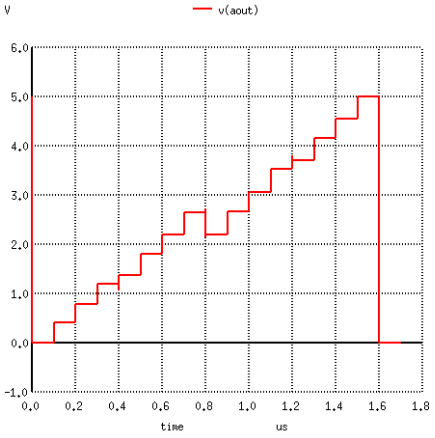


Figure 8.164: Simulation of a binary weighted resistor DAC with too high resistive switches

DAC with tapped resistor ladder: One possible way to avoid the influence of the resistance of the switches is to use a tapped resistor ladder ([77] p.752). In older technologies the cost of the digital decoder was prohibitive to using this approach. With logic shrinking more and more thermometer code DACs become more common. To explain the concept a 4 bit thermometer code resistor DAC is shown.

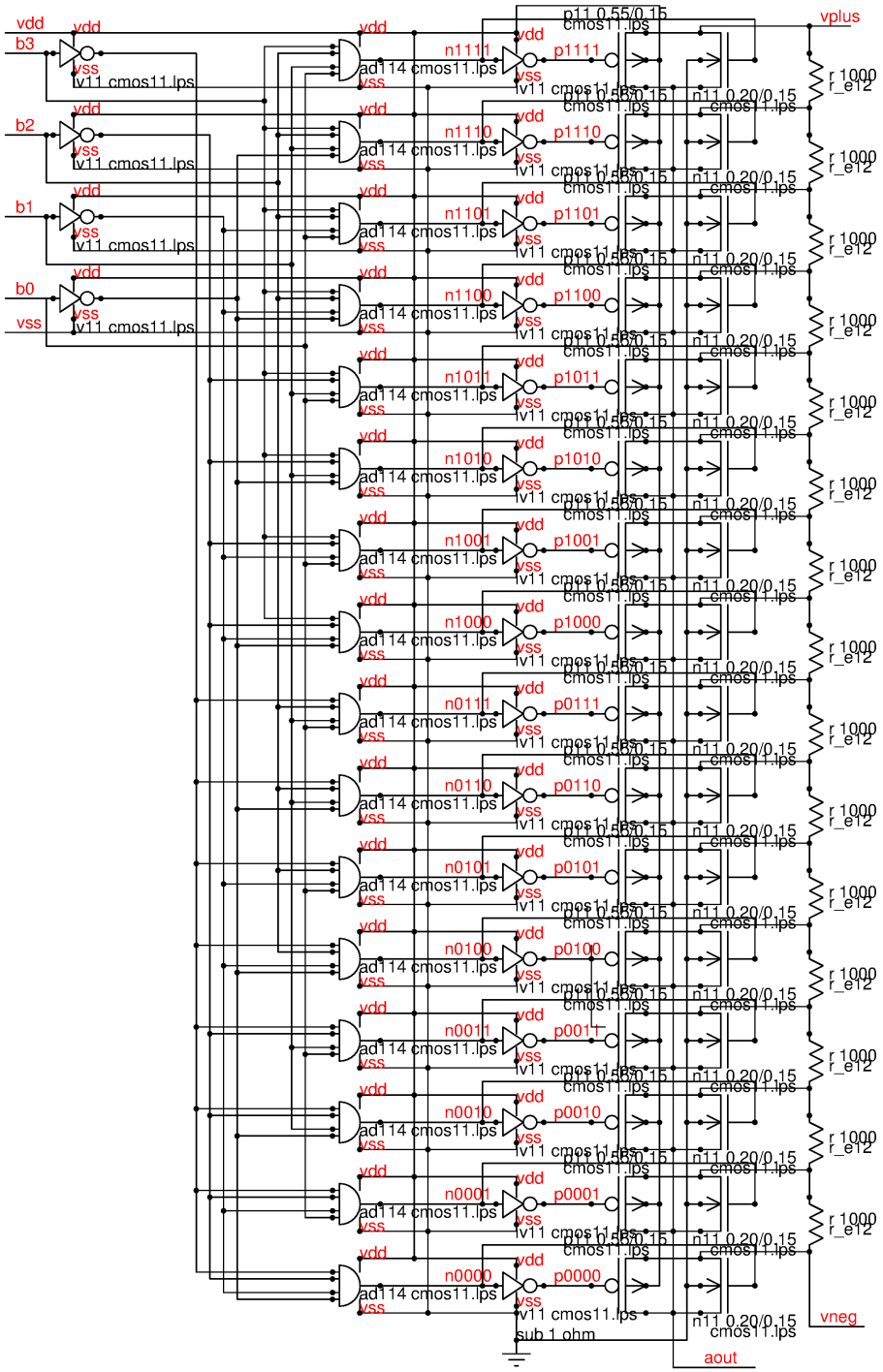


Figure 8.165: 4 bit DAC using a resistor ladder

This kind of DAC has the following properties:

- Constant load current on the reference supply
- Can't drive a resistive load (makes it non linear, integral non linearity INL)
- No missing code, excellent differential non linearity (DNL)
- very regular structure perfect for modern technologies with high logic density
- resistance of the switches doesn't change the conversion result.
- output impedance is code dependent

The output of this DAC can be regarded like a slider of a potentiometer. Changing the slider from K=0 to K=1 varies the output signal from vneg to vplus. Calling the two resistors of the potentiometer R1 and R2 we get:

$$R_{out} = \frac{R_1 * R_2}{R_1 + R_2} \tag{8.275}$$

Introducing K as a DC transfer factor we get:

$$K = \frac{R_2}{R_{ges}} \tag{8.276}$$

with Rges=R1+R2. Thus the output resistance (neglecting the resistance of the switches) becomes:

$$R_{out} = K * (1 - K) * R_{ges} \tag{8.277}$$

Assuming a perfect divider in combination with a load current we can calculate the integral non linearity in the mid position. The voltage error is simply

$$V_{error} = R_{out} * I_{load} \tag{8.278}$$

The integral non linearity (INL) becomes:

$$INL = \frac{V_{error}}{LSB} = \frac{R_{out} * I_{load}}{LSB} \tag{8.279}$$

To show the impact of a resistive load in the following figure the DAC was operated without load (blue) and with a 5K load to the mid range voltage (0.5V, red color). In the range where the output resistance is high the signal of the DAC is compressed. Close to the ends of the range (vplus and vneg) the distorted signal is stretched. In this example the integral non linearity caused by the resistive load reaches +2LSB and -2LSB!

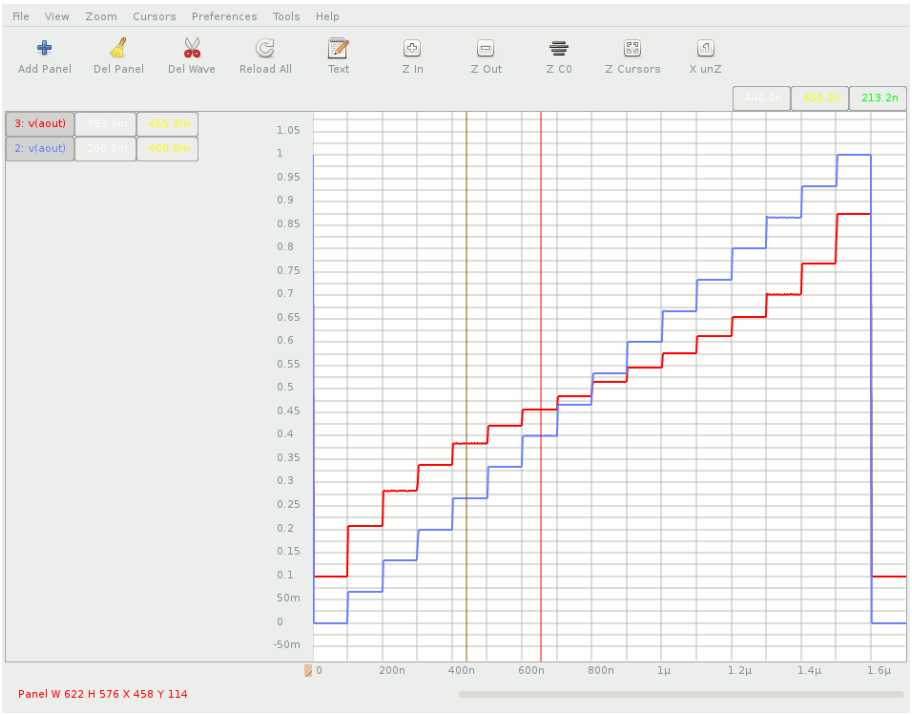


Figure 8.166: Comparison of DAC with resistive load to mid range (0.5V) and without resistive load

This plot of the load dependence of the output signal shows that a resistor ladder DAC must always be operated without a DC load. If a load has to be driven the DAC needs a buffer amplifier connected to signal aout.

The achievable resolution is mainly limited by the area needed for the digital binary to thermometer decoder. Today (2015) DACs up to 10 bit are common using 1.2V logic technologies. Beyond that point the logic effort doubling with each bit becomes too high.

An additional feature is the posibility of creating an intentional non linear DAC by choosing different resistor values in the ladder. This way for instance a logarithmic or exponential curve can be constructed easily. Resistor mismatch (because the resistors don't all use the unit device) won't lead to missing code.

Subranging resistor ladder DAC: The idea of a subranging DAC is to use a resistor ladder for the more significant bits to keep the matching requirements of the analog components in a moderate range and to add a shift stage for the lower significant bits. This way it is possible to limit the effort for the digital thermometer decoder. The concept is shown in the following figure. It is basically the same DAC as before but the blue colored components are added to shift up and down the whole ladder. The shift stage now gets the LSB while the MSB is the most significant bit of the thermometer decoder. This way with modern technologies offering a high logic density it is possible to build 14 bit DACs (10 bit go into the thermometer decoder, the four LSB are used for the shifter). Take care that the ON-resistance of the transistors shifting up and down the ladder is decisive for the accuracy of the DAC.

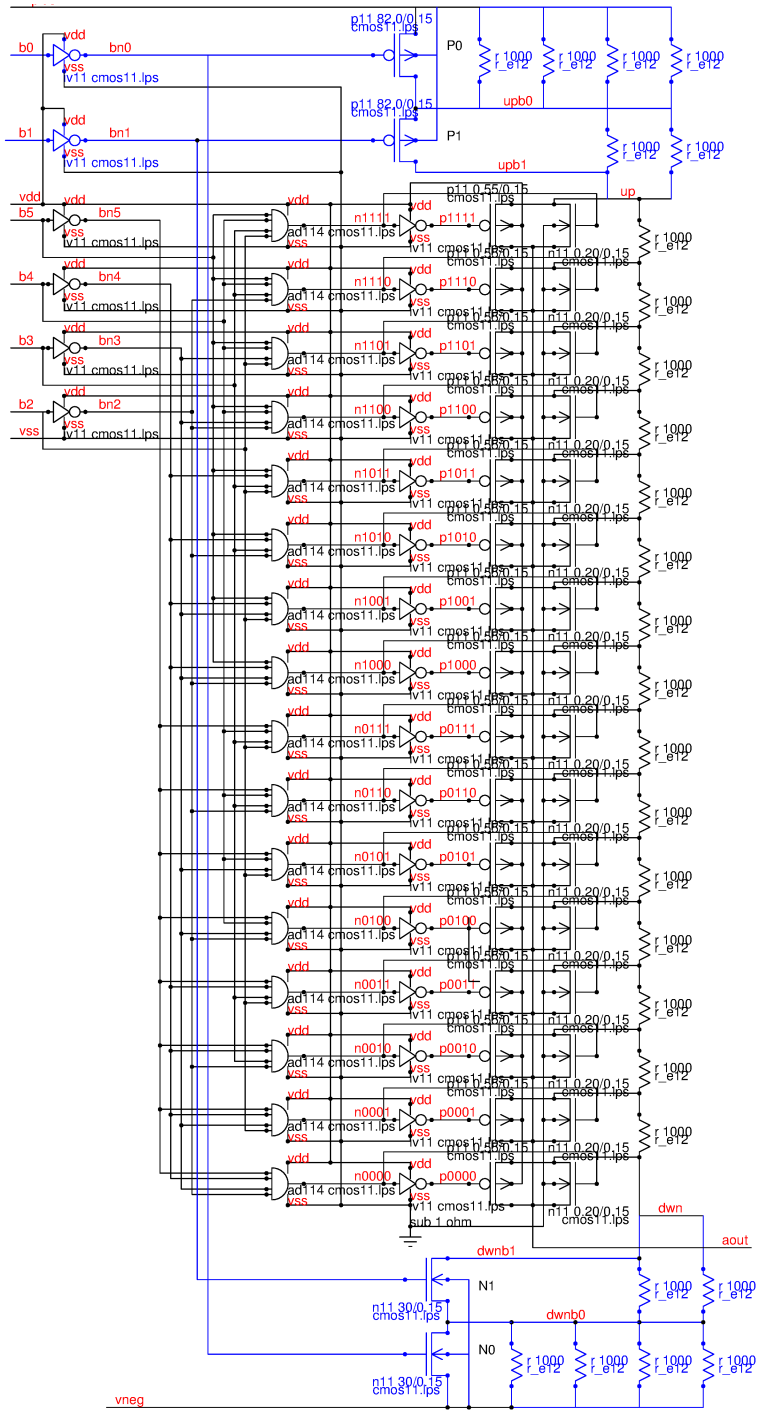


Figure 8.167: Subranging DAC using 4 bit for a thermometer decoder and the two LSB for the subranging shifter

The transfer function of this DAC is shown in the following simulation plot. Besides the analog signal V(aout) the plot also shows the voltages at the top and the bottom of the resistor chain.

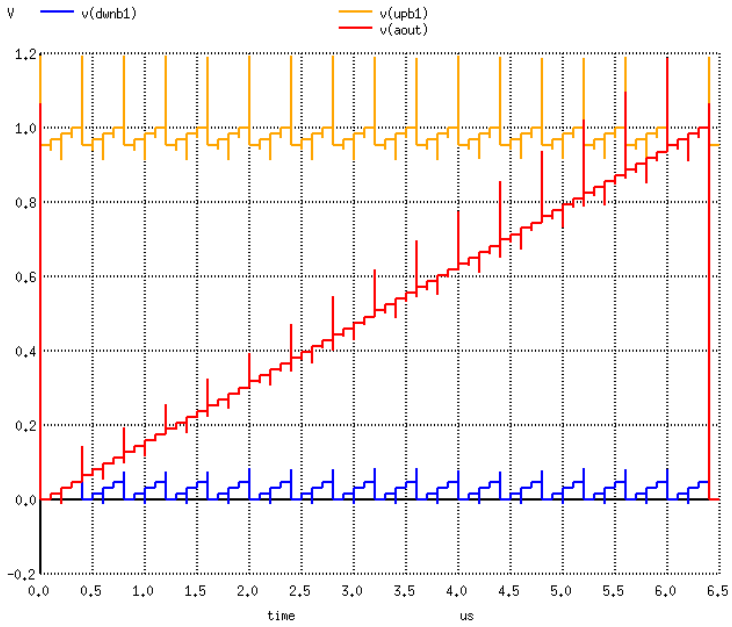


Figure 8.168: Signals of the subranging resistor ladder DAC

R2R DAC: The R2R DAC doesn't require a decoder [32]. So it is well suited for a technology with a lower logic density. The DAC only needs 2 resistor values. In most implementations the two resistors are composed of 1 unit resistor and 2 unit resistors. Basically it is a current splitter. The price for it's simplicity is the need of an operational amplifier. The voltage drop over all switches must be (ideally) equal. Thus the switches must be scaled according to the current they switch. But since all switches can be designed using the same transistor type building R2R converters up to 10 bit works fairly well.

One drawback of the R2R converter is the inverting amplifier. With a positive reference voltage the output of the R2R converter becomes negative!

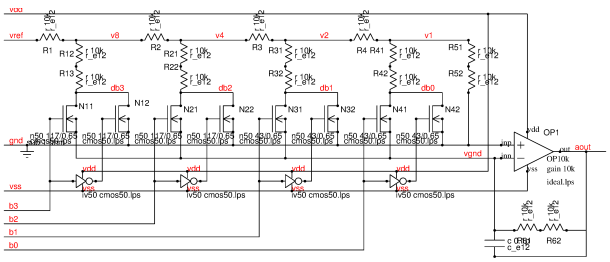


Figure 8.169: R2R DAC

In the example simulated the reference voltage was 1.2V. So the output voltage range of the DAC is 0V to -1.2V. Since the example uses a fairly large 5V technology the feed through of the digital signal via the gate-drain capacities of the switches is well visible.

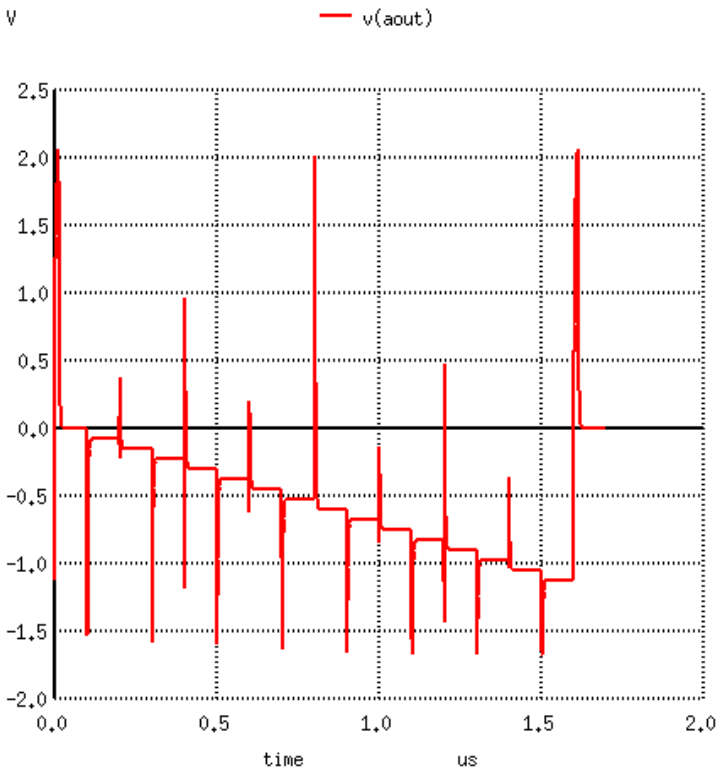


Figure 8.170: The R2R DAC inverts the reference voltage

8.9.3 Current DACs

On integrated circuits very often it is an elegant solution to use current DACs in stead of voltage DACs. The digital input signal is converted into a current instead of converting it into a voltage. The most straight forward approach is building weighted current sources and switching them on and off with the digital input signal. Usually all transistors used in the current DAC are built using the same cell. These cells can be used in series or in parallel to increase or decrease the current. To prevent missing code the spread of the MSB must be less than the value of the LSB. This limits the design of binary weighted current source DACs to about 6 bit.

The most basic concept of a current DAC:

The current is generated by the transistor type that offers the best matching parameters. Usually this is a low voltage device. Using devices with halo implant however is depreciated because halo implant degrades the matching properties. (See chapter 4.5). In most technologies 3.3V transistors are a good choice (typically 7nm gate oxide, no halo implant yet. 1.5V transistors usually have an automatically generated halo implant that is not always obvious looking at the Cadence or GDS layers). Using cascodes to obtain equal Vds for all transistors is mandatory.

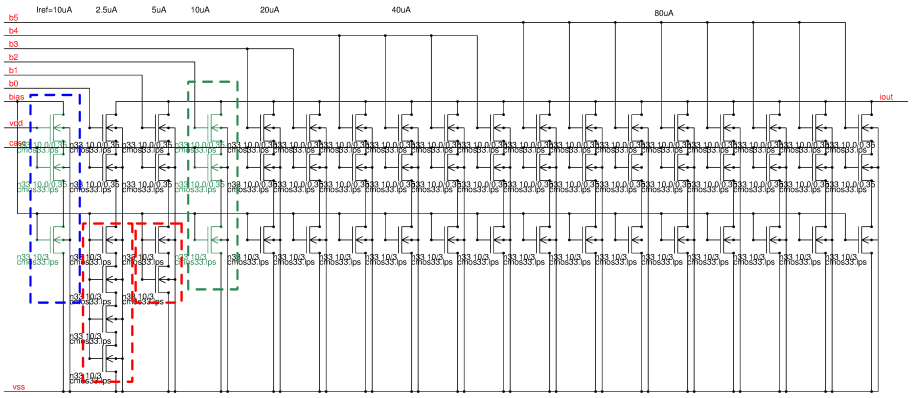


Figure 8.171: The most simple current DAC

The current DAC always uses the same stack of transistors shown in the blue and the green box. This is the unit cell. The stack can be created as a single active region just with 3 gates on it. For this reason the bias generation connects the gate of all current generator transistors to the bias current input. This way there is no need to tap the stack of the reference generator. (If we would do this we would have to place a dummy contact on every stack to

obtain correct matching with the reference in the blue box. This would waste area and add one more polygon that can be disaligned and cause mismatch!)

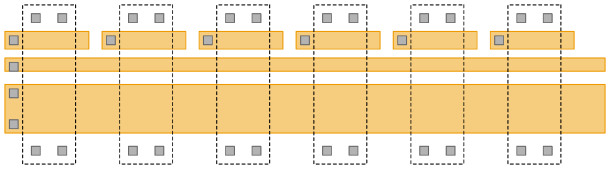


Figure 8.172: Layout concept of the unit cell approach

There are two exceptions of this unit cell approach. These are the least significant bits. Here the symmetry is broken. The error is in the range of some percent of the LSBs. For the LSBs this can be accepted. The transistors that are the most critical usually are the MSB and the single transistor of the 10uA source (green box)

Low resistive ground (VSS) is essential. Even a voltage drop of only some mV along the VSS trace will immediately lead to a deviation of the currents. Distributing the transistors of the higher bits over the array is recommended mainly to compensate possible drops in the VSS line. For very big DACs making the VSS a big metal plate covering the whole DAC with top metal is a good idea to minimize the VSS-drop.

For DC this simple current DAC works acceptably well. The AC performance however is bad!

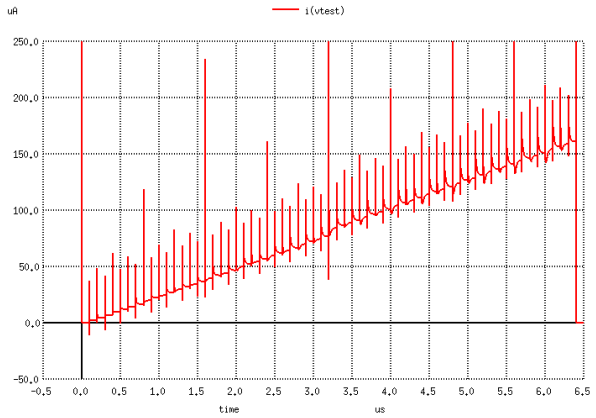


Figure 8.173: Noise superimposed on the output current of the simple DAC

The noise we find on the current signal is coming from charge injection when the drain of the big current generator transistors changes its voltage every time a switch is turned on or off. This change of the drain voltage capacitively couples into the gates of all current generator transistors. (Notice the slow recovery after each switching event. This is the recovery time of the current generator bias node!)

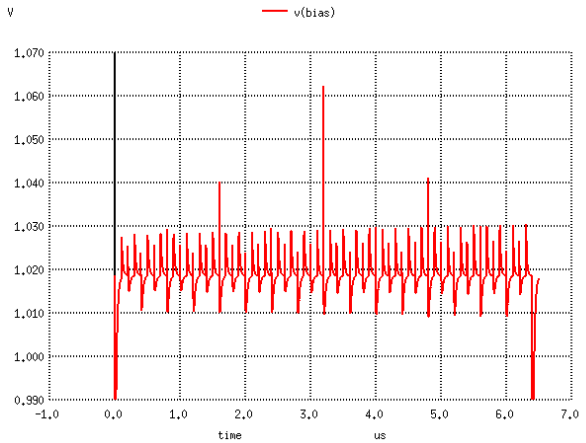


Figure 8.174: Noise coupled into the gate bias of the current generator

To solve this noise problem the drain voltage of all current generator transistors must be kept constant. Furthermore the charge injection from the switch into the load (gate capacity) must be canceled. Since all those parasitic capacities are voltage dependent and have about +-20% of production spread this cancellation never is perfect.

The better solution is to keep the voltage drop over the current generator transistors constant. This means in stead of turning off the current we must switch the current between the load or vdd. (Well, the second node can also be used as an inverting output of the current DAC.)

Current DAC with better AC performance: The basic cell of the improved current DAC looks like a differential amplifier. The current simply is switched to either one of the sides. Additionally every switch is driven with exactly the same driver strength. There still is a little error between the edges driving both switches due to the propagation delay of the inverters.

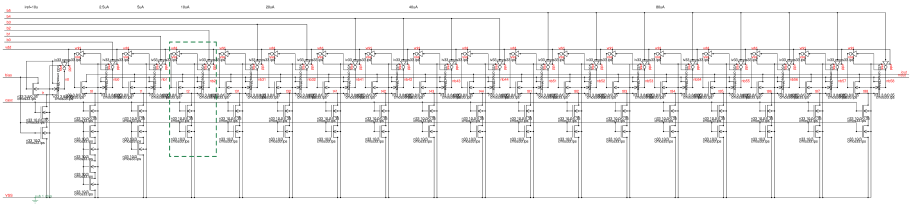


Figure 8.175: Current DAC that doesn't interrupt the currents

The performance is already much better. The price for that improvement is double the current consumption compared to the single ended approach. Since we have two output currents now both are shown in the plot.

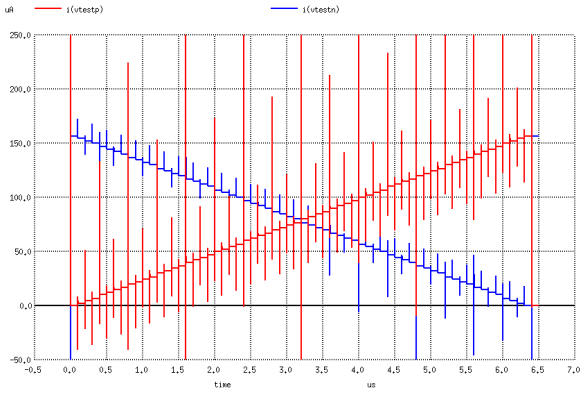


Figure 8.176: Output currents of the improved DAC

The glitches still are there but there is no more slow settling after the glitch. The energy content of the glitches is lower. In fact the remaining glitch (in the following plot the change of the MSB is shown) mainly is caused by the timing error of the inverters driving the switches. Ideally the speed of all bits should be exactly the same. So either the driver stage must match the total size of the switches (If several switches share one driver) or each switch should have it's own driver. If this rules is violated (for instance if the MSB has the same driver for 8 transistors as the LSB has for one transistor the MSB would switch slower than the lower significant bits.), this leads to an increase of the energy in the glitches. In the example design each switch segment has the same driver to avoid this effect.

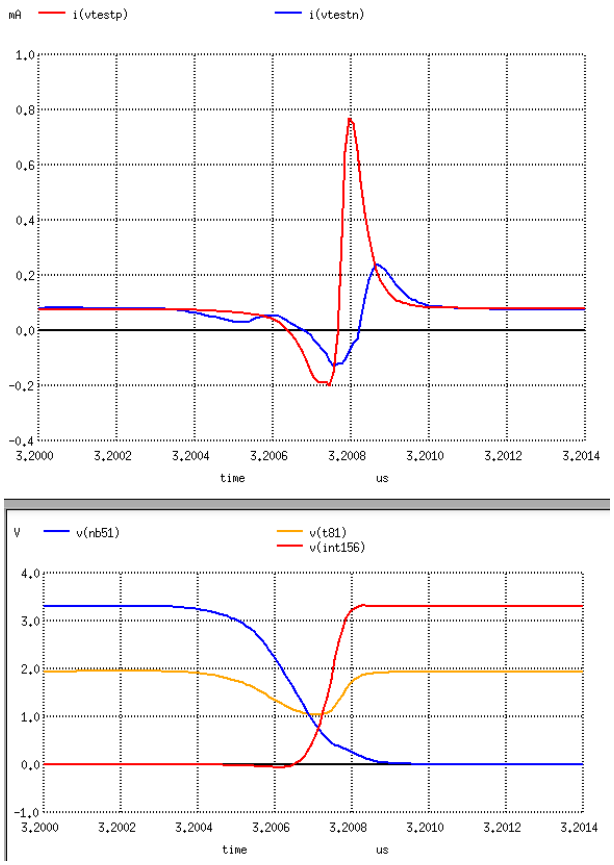


Figure 8.177: Remaining glitch at the change of the MSB and the signals driving the switch together with the voltage at the tail of the switches

For further performance improvement the voltage swing at the gates of the switches must be reduced and the signals should come from a fully differential driver similar to current mode or ECL logic (a voltage swing of differentially no more than 500mV is a good choice). But that costs even more supply current (For high performance DACs the current consumption of the switch drivers in fact can be higher than the current flowing through the switches!

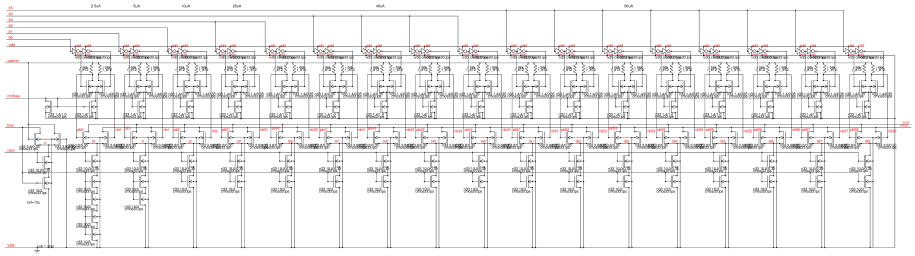


Figure 8.178: Same DAC as before but now with a current mode logic driver

For better visibility here is one unit cell of the DAC with the CML driver stage.

Using a current mode driver and reducing the voltage swing at the gates of the switches reduces the cross talk. If the current per current mode logic driver is chosen to be 5μA the voltage swing becomes 250mV. The supply voltage of the CML driver must be slightly higher than the cascode voltage (about 1.5V is a good choice if the cascode is driven with 1.2V). Now most of the remaining noise is coming from the logic driving the CML stage. The high swing of the inverters couples to the output of the CML stage and produces a fast common mode signal there. This common mode signal reaches the DAC switches and produces the first short but high spike.

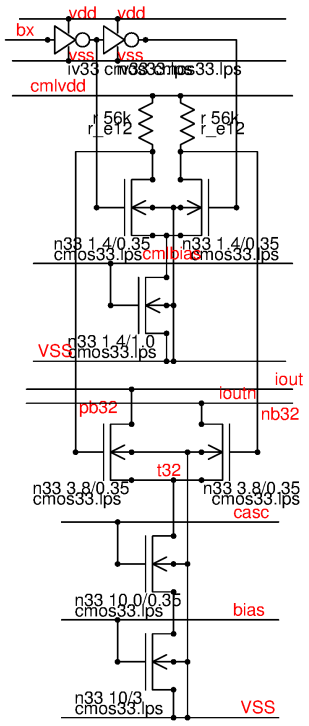


Figure 8.179: Unit cell of the CML driven DAC

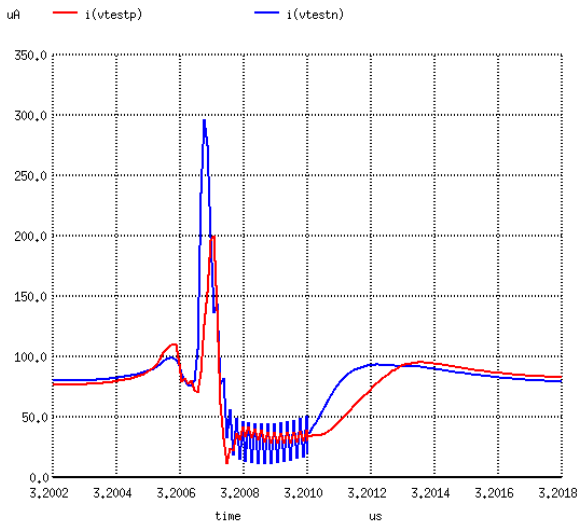


Figure 8.180: Current DAC with CML driver spike at the change of all bits in the middle of the range

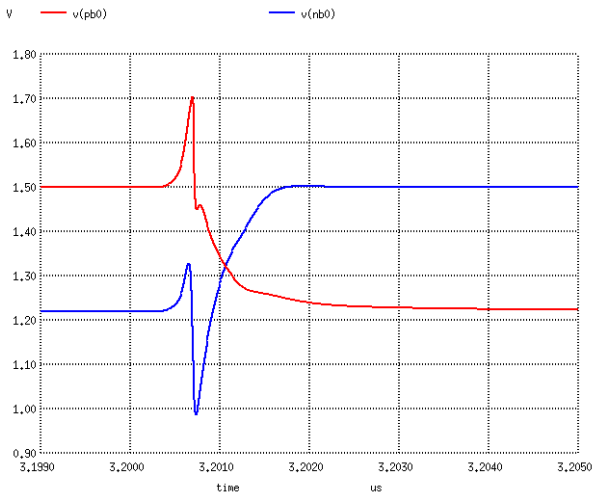


Figure 8.181: Feed through of the logic signal through the CML driver

Best counter measure against this feed through is a reduction of the drive signal at the inputs of the current mode logic and improved timing symmetry (latched drivers similar to clock driver for perfect symmetry).

The last resort to reduce logic signal feed through is to minimize the number of synchronous switching events. So again we end up at thermometer code switching on one stage after the other in stead of using binary weights of current switched together. In stead of a running 1 like in the resistor ladder DAC here we need an increasing number of ones each time the data is incremented.

Wire coupling and layout requirements All the analog traces are low level signals that MUST be protected from noisy signals. The worst aggressor in this circuit are the clock lines that have a high voltage swing (typically 5V or 3.3V) and fast edges (rise and fall times in the range of 100ps). The analog signals and the clock lines may under no circumstances be routed in close proximity. The recommended solution is to route the clock lines perpendicular to the analog lines. This minimizes the parasitic capacities.

In addition every analog line that is crossed by a clock line must also be crossed by an inverted clock line to compensate the clock injection.

If clock lines have to be routed in parallel with analog lines the clk and clk_n line must be designed like a twisted pair to achieve the same parasitic capacity between signal and clk as well as between signal and clk_n.

Current DACs with high number of bits: Designing a 10 bit current DAC with binary weighted current sources would require a 6s matching of the MSB of 0.1%. to prevent missing code. This of course is hopeless. In stead of creating binary weighted sources a unit source is used. Each of the cells must be driven by a thermometer coded control signal.Even if one of the cells deviates by 50% we still don't get missing code. But integral non linearity (INL) in fact suffers in a statistical way if the cell get too small.

In case of a 10 bit DAC this means we get 10 bit input signal and that gets converted into 1023 select lines (one for each cell) coming out of the thermometer decoder. For this reason current source DACs with more than 6 bit can only be implemented at reasonable cost using technologies with high gate density and narrow wiring pitch.

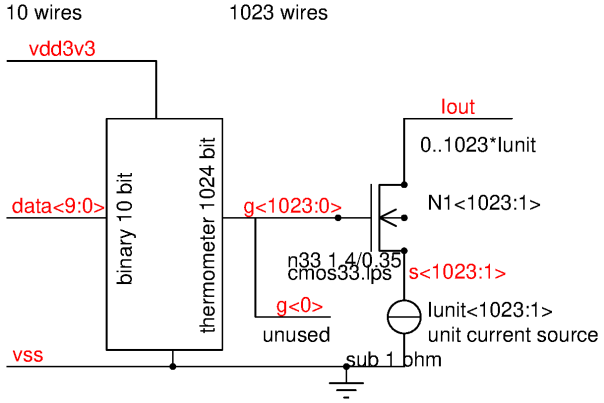


Figure 8.182: 10 bit thermometer coded DAC

Of course nobody wants to draw 1023 wires! This is why everything is described as an array using the notification <1023:0> (1024 wires) or <1023:1> (1023 items). For current 0 no current source and no switch is needed. But

usually the thermometer decoder is a synthesized cell offering a signal g<0> that simply remains unconnected.

Most commercial layout tools and netlisters can handle such notifications. Layout XL will in fact automatically place 1023 cells of transistor N1 and current sink Iunit.

8.9.4 Voltage gradients in the ground wire

Typically the unit current generators are cells of a big current mirror. The voltage drop in the ground metalization chages the gate-ground voltage over the array. In the following a current generator with resistive source degradation is shown to explain what happens. In the following we assume R>>1/gm. This linearization simplifies the calculation. (In reality for full accuracy we must always take into account the sum of R+1/gm instead of R only. If the aspect ratio of the NMOS transistors is big enough the error caused by the linearization can be kept in the 10% range.)

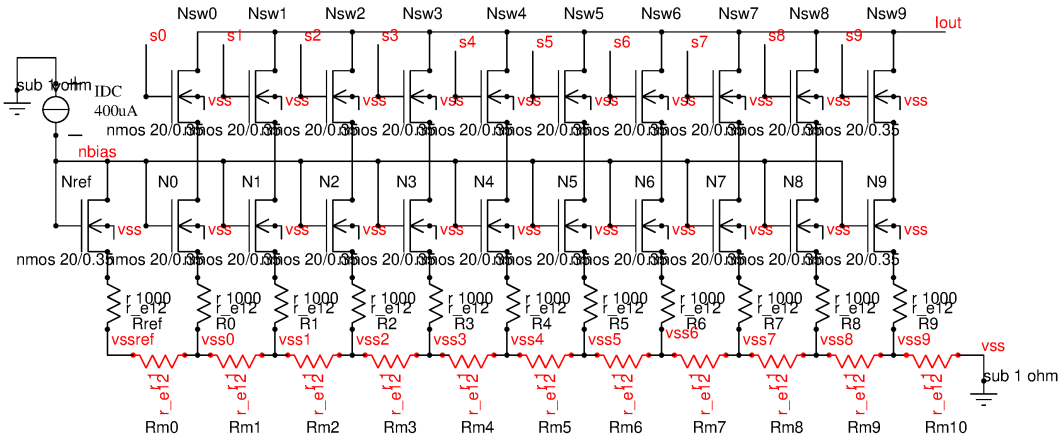


Figure 8.183: thermometer coded current DAC with parasitic metal resistance

In an ideal world the drop over all resistors Rref, R0 to R9 should be 400mV while the metal path resistors Rm0..Rm10 should be negligible. Ideally setting all select pins s0 to s9 the output current should be 4mA.

Including the metal resistors Rm0 to Rm10 the voltage of vssref and vss0..vss9 differs the more we move from left to right. vssref is already 0.4mV higher than vss0. The current through the metal path (drawn in red increases from left to right. Assuming a small deviation we can approximate a linear increase of the error.

$$I(Rmx) = (x+1) * I_{bias} \tag{8.280}$$

The voltage drop over the metal resistors increases accordingly.

$$V(Rmx) = (x+1) * I_{bias} * \frac{L}{W} * r_{met} \tag{8.281}$$

In this equation $r_{met}$is the resistivity of the metal trace (usually in Ohm/#), L is the trace length from one resistor to the next and W is the width of the trace. V(Rmx) is the voltage drop over one metal resistor. To calculate the error of each current source we are interrested in the voltage drop from vssref to the tap vssx.

$$vssref - vssx = \Sigma_{k=1}^{x+1} V(Rmk) \tag{8.282}$$

This equation can be rewritten as:

$$vssref - vssx = \frac{L}{W} * r_{met} * I_{bias} * (1 + 2 + 3 + ... + (x+1)) \tag{8.283}$$

Looks a bit like the sum over a ramp. For $x \to \infty$ we can approximate the sum by an integration. Running from x=0 at vssref to x=Lw at vss9 the current through the resistors builds up linearly.

$$I(x) = \int i * dx \tag{8.284}$$

i is a current per length along the metal trace. Since all the currents flowing into the red drawn metal trace is (almost) constant we can determine i as

$$i = I_{bias}/L = I_{tot}/L_w \tag{8.285}$$

Here $I_{tot}is$the total current flowing and Lw is the total length of the metal trace of vss. The calculation of I(x) simply is:

$$I(x) = x * \frac{I_{tot}}{L_w} \tag{8.286}$$

The voltage drop from vssref to position x calculates

$$V(L_w) = \int_0^{L_w} I(x) * dR$$

with

$$dR = r_{met} * \frac{dx}{W}$$

This leads to

$$V(L_w) = r_{met} * \frac{I_{tot}}{L_w * W} * \int_0^{L_w} x dx = \frac{r_{met} * I_{tot} * L_w}{2 * W} \tag{8.287}$$

If we are not interrested in the end point but some position X in the middle the equation transforms into

$$V(X) = \frac{r_{met} * I_{tot} * X^2}{2 * W * L_w} \tag{8.288}$$

Example: $r_{met} = 0.082\Omega/\#$, $L_w = 100\mu m$, $I_{tot} = 4mA$, $W = 1um$ leads to

Table 44: Example calculation of the influence of metal paths on the I-DAC accuracy

X/μm	V(X)/mV	error/μA	cumulative error/uA	current/mA	relativ cumulative error
10	0.164	0.164	0.164	0.400164	0.041%
20	0.656	0.656	0.820	0.80082	0.1025%
30	1.476	1.476	2.296	1.202296	0.1913%
40	2.624	2.624	4.92	1.60492	0.3875%
50	4.100	4.100	9.02	2.0092	0.451%
60	5.904	5.904	14.924	2.40149	0.622%
70	8.036	8.036	22.96	2.823	0.82%
80	10.496	10.496	33.456	3.233	1.0455%
90	13.284	13.284	46.74	3.647	1.2983%
100	16.4	16.4	63.140	4.0631	1.5785%

The example shows that the metal path resistance matters. The bigger the current DAC gets and the longer the metal trace gets the higher the impact of the metal path. Placing the reference in the middle doesn't change the shape of the curve. It only changes where the 0-reference is. The biggest problem of building current DACs is not the circuit design but the layout implementation!

The higher the currents the higher the impact of the metal resistance gets.

Power stages: If power current DACs are to be designed it is tempting to most layouters to use the metal of the DAC supply or ground to route the supply of other stages through the current DAC. This is absolutely fatal because in addtion to the triangular current distribution of the DAC itself and additional DC current is added (and integrated over the metal path). The current routed through doesn't decrease from stage to stage. So we don't integrate over a triangle but over a rectangle. Current routed through has double the weight of the DAC current! This current further increases the errors. Every mV of drop counts!

Moving the reference to the middle and placing the current sources alternately on both sides of the reference reduces some of the impact on the integral non linearity (INL). But the differential non linearity (DNL) still is determined by the current running through the metalization and the distance between the reference and the current generator the furthest away from the reference.

Metalization between the resistors and the sources of the current generators: In power stages the metal path from Rx to the source of transistor Nx may significantly contribute to the total gain degradation. The ratio between Rx and the resistance of this metal path must be constant over all current generator stages and it must match the reference stage. In a thermometer coded current DAC this often 'comes for free' because all stages use the same cell.

In binary weighted designs the voltage drop in all metal paths must be kept equal! As a consequence the lower significant bits of a binary weighted current DAC require compensation metal resistors to create the same voltage drop over metal in all stages..

Basic design approach:

1. Layout the MSB.
2. Extract the metal path resistance of the MSB (R_{mMSB})

3. Layout the next lower bit and add the dummy trace required to reach $2 * R_{mMSB}$ (factor 2 because now we have half the current)

4. repeat step 3 until you reach the LSB. Design the metal path for

$$R_{currentbit} = \frac{I_{MSB}}{I_{currentbit}} * R_{mMSB}$$

5. If the reference cell uses a different current than the LSB add a metal resistor to the reference cell to achieve the same drop over the metal as in all the bits of the current DAC.

Ideally the metal in the MSB and all the compensation traces should use the same metal mask. If different masks are used (for instance inside the MSB metal 3 and metal 4 are used, but all other stages are using m3 compensation resistors) the temperature coefficient will approximatelly be matched but the missmatch of the different metal layes will be visible!

8.9.5 Gain error of a DAC

The DAC can't be better than the reference. This applies to DACs that provide an output voltage as well as to DACs that provide a current.

The output voltage of a DAC providing a voltage output is caused by the tolerance of the reference voltage. Well designed bandgaps can be trimmed (at one temperature) to about +-0.1%. The problem is the curvature of the bandgap. Usually this is an upside down parabola shaped curve. Typical bandgaps have a parabola that hangs down at cold (-40°C) and hot (150°C) by about 0.2%. Aging of the bandgap and mechanical stress may add an other 0.2%. Unfortunately these are not statistical errors. So the usual adding of the squares as done with statistical data is not applicable!

Current DACs usually are designed as current mirrors. The high number of output mirrors averages out the error of the output transistors. But the error of the reference cell remains. Usually the MOS diodes is significantly smaller than the total area of the output transistors. Thus the reference current defines the size of the MOS diode and the statistical gain error of the whole current DAC.

If the gain error is dominated by the reference diode most of this error can be removed by trimming of the reference current. The temperature coefficient caused by the curvature of the bandgap driving the reference current generator however remains.

8.10 Analog digital converter (ADC)

Analog to digital converters convert an analog input signal into a digital output signal. The achievable resolution runs from 1 bit to about 20bit for high precision measurements. Often the resolution of an ADC is limited by thermal noise and the bandwidth. Reaching 20 bit only works for an extremely low bandwidth. The low bandwidth can be achieved by low pass filtering of the input signal and by averaging many samples to reduce the noise produced inside the ADC.

A second limiting factor can be digital noise produced by the logic on the same chip. Digital noise often has a more or less flat envelope of the spectrum but the energy is sitting in discrete spectral lines. Using synchronous logic there usually is a very strong line at the clock frequency and harmonics of the clock frequency. To start easy let's begin with a 1 bit ADC.

The most simple analog to digital converter is a simple comparator or a schmitt trigger. This can be regarded as a 1 bit ADC. Even this very simple one bit ADC already has some remarkable properties!

8.10.1 Noise in ADCs

Every signal source has a certain resistance. This resistance produces thermal noise. Just to remember:

$$V_{nr} = 2 * \sqrt{R * k * T * BW} \tag{8.289}$$

(BW is the bandwidth taken into consideration)

Resistive input dividers produce noise as well. The higher the divider resistance the more noise we get! The divider noise present at the input of the comparator depends on the impedance the comparator is connected to.

Some ADCs use capacitive dividers. These capacitive dividers don't produce noise themselves but there is a certain staistics in the current charging the capacitor. It can be regarded as a quantization noise of the electrons charging or discharging the capacitor and the statistics of electron movement. A current must be regarded as electrons randomly "hopping" into the capacitor. The current itself is just the average of these electrons., that by far don't arrive in periodic times. Freezing the charge in the capacitor by opening a switch will lead to a random distribution of the number of electrons eventually "sitting" in the capacitor. In other words the sampled charge has a random spread which leads to a noise like random sampled voltage as well

$$V_{nc} = \sqrt{\frac{k * T}{C}} \tag{8.290}$$

$V_{nc}$is the random fluctuation of the sampled voltage every time the switch is opened. The nice feature of capacitive dividers is that the random voltage fluctuation doesn't increase with the bandwidth. This makes capacitive dividers attractive for RF attenuators.

The same applies to the comparator itself.

The differential stage produces noise as well (1/gm can be regarded as a resistor in the source or emitter of the differential pair). If the comparator has no hysteresis the noise can randomly make the comparator flip between the two states. The threshold simply can be regarded as the voltage at which the comparator produces 50% logic ones and 50% logic zeros.

The energy of the noise (if it is white noise) is proportional to the bandwidth. The effective noise voltage follows the square root of the bandwidth.

To make an ADC work well the noise voltage (of the divider(s) and the comparator(s) inputs) must be significantly lower than the value of the LSB.

8.10.2 The most basic: a 1 bit ADC

The 1 bit ADC has two output values: 1 and 0. The logic 1 usually is assigned to 50% full scale and higher and the logic zero is assigned to levels below 50% full scale. Neglecting offset and noise the comparator will switch at exactly Vref/2.

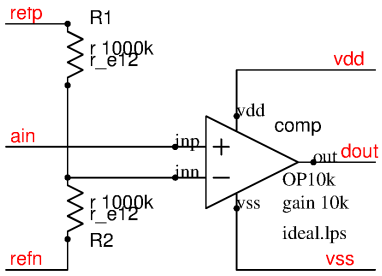


Figure 8.184: The most simple 1 bit ADC

This simple example already shows a very important concept: Reference ground (refn) and reference voltage (refp) should never be shared with any traces carrying current!

The impedance at the negative input of the comparator is 500K. The resulting resistor noise becomes about $90nV/\sqrt{Hz}$. Sounds harmless, but if the comparator has a response time in the range of 0.5ns we end up with a bandwidth of about 1GHz. So for this bandwidth $90nV/\sqrt{Hz}$ immediately produce an effective noise voltage of 27mV. Since the amplitudes of the noise are statistically distributed the ADC as it is shown here will not work well anymore for an LSB of less than about 120mV.

To reduce the noise of the divider either lower resistor values must be chosen or a capacitive divider in parallel must be added.

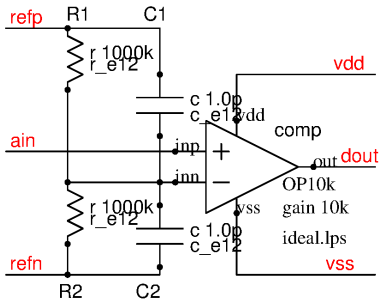


Figure 8.185: One more time but with better noise performance

Now there are two resistors acting as a noise source and 2 capacitors reducing the noise bandwidth to 140kHz. The effective noise voltage becomes 1.5mV. Additionally this noise is bandwidth limited. For applications ranging from DC to 140 kHz we can use the divider for an LSB of about 8mV now.

If the ADC is used for pure AC signals in the GHZ range the noise from the resistor divider is even lower (it rolls off with -20dB/decade starting at 140kHz). But of course the source driving nets refp and refn must be able to drive the load of 0.5pF.

Layout recommendations: On chip there is no ideal resistor and no ideal capacitor. Every component has a stray capacity to whatever is underneath or above it. It is extremely important to take parasitic capacitive coupling into consideration.

Resistors must be placed over silent wells. Ideally these should be connected to refn and refh (So the parasitic capacities are in parallel to C1 and C2 and have the same ratio as C1 and C2.)

Capacitors C1 and C2 ideally are designed as coaxial structures with the inner layer connected to the tap of the divider and the outer layer connected to refn and refl.

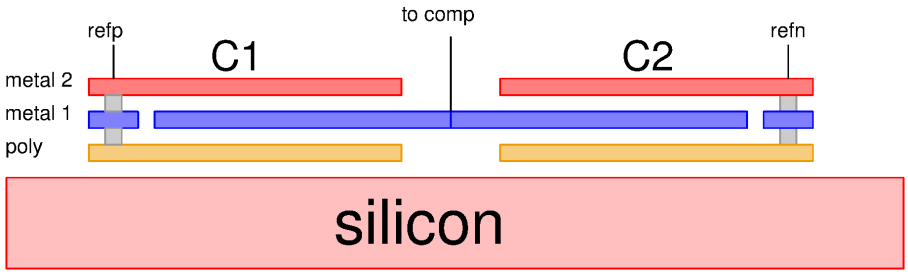


Figure 8.186: Two coaxial capacitors used to avoid parasitic coupling of substrate noise and noise from wires passing above the capacitors

If more than 2 layers of metal are available the bottom layer can use metal1, the inner node metal2 and the top layer metal3.

Using classical fringe capacitors for ADCs is not recommended because each finger has stray capacities to outside of the capacitor.

One bit ADC with capacitive input: Capacitors can be integrated with very good matching properties. Therefore it is a good idea to use capacitors in stead of switches as a voltage divider. The complete system becomes a sampling system. This allows cancellation of the offset of the amplifier.

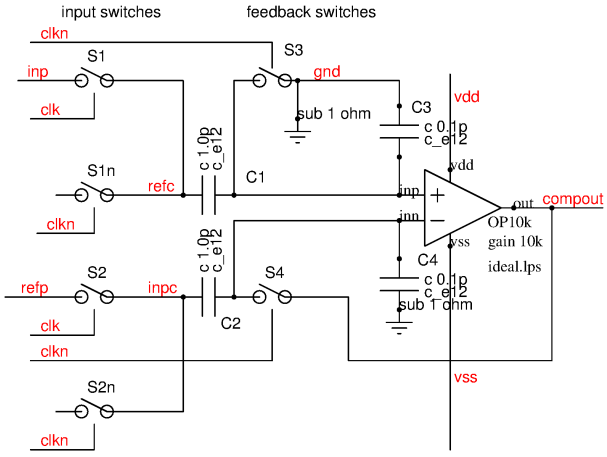


Figure 8.187: Sampling capacitor 1 bit ADC

In the preparation phase clkn is logic 1 closing switches S1n, S2n, S3 and S4. Assuming the amplifier doesn't have an offset S4 could be connected to gnd as well. Connecting S4 to the output of the amplifier in stead of gnd offers offset cancellation as an additional feature of the clocked system. Connecting C2 to the output of the amplifier charges C2 with exactly the offset voltage.

In the sampling phase S1n, S2n, C3 and S4 are opened. Closing C1 and C2 the amplifier compares V(inp) with V(refp). If V(inp) is higher than v(refp) the output of the amplifier becomes vdd. Vice versa if v(inp) is less than V(refp) the output switches to vss.

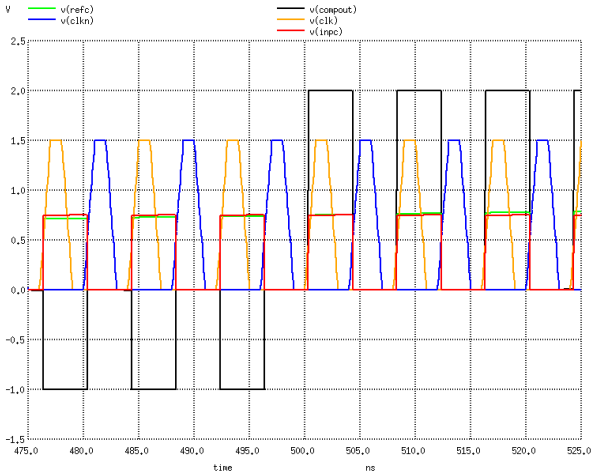


Figure 8.188: Signals of the sampling 1 bit ADC when V(inp) crosses v(refp)

Capacitive coupled comparators can also be used for adding and subtracting signals. This way the sampling comparator can also be used to measure true differential signals. The negative nodes of the reference voltage and the input signals simply are sampled at reverse polarity. Reverse polarity is achieved charging the input with the voltage of the positive node of the voltages and at the same time removing the charge of the negative node. This approach works well as long as the signal doesn't change too much between the preparation phase (clkn=1) and the measurement phase (clk=1).

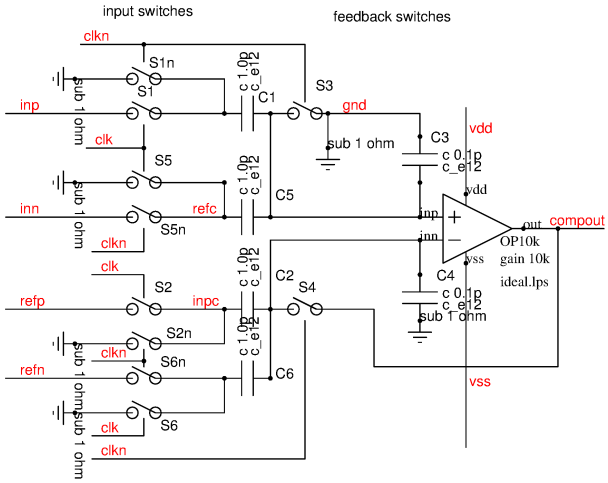


Figure 8.189: Extension of the switched capacitor comparator to fully differential measurement

The input signal at the amplifier inputs now has half the amplitude as in case of the single ended measurement. On the other hand the capacity is doubled now. So regarding capacitive noise the loss of accuracy is 3dB.

Quantisation noise: The quantization noise is a consequence of the quantization error. To calculate the quantization noise the effective error voltage must be compared with the effective voltage of a sine wave converted with an ADC using the full range [46].

The error voltage is the difference between the output of the ADC and the input of the ADC. In case of the 1 bit ADC the error is a saw tooth signal with only one period.

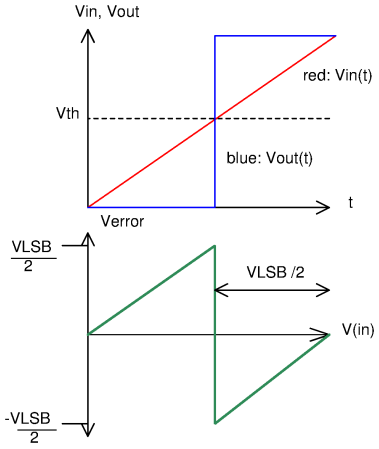


Figure 8.190: Quantization error of a 1 bit ADC

To calculate the effective quantization error voltage the square if the error voltage must be integrated over the whole signal range.

$$V_{erroreff}^2 = \frac{1}{V_{inmax}} * \int_0^{V_{inmax}} V_{error}^2 * dV_{in} \tag{8.291}$$

since the signal is symmetric with respect to the trip point of the comparator the equation can be simplified integrating from 0 to 0.5*VLSB.

$$V_{erroreff}^2 = \frac{2}{V_{LSB}} * \frac{1}{3} * (V_{LSB}/2)^3$$

$$V_{erroreff} = \frac{V_{LSB}}{2 * \sqrt{3}} \tag{8.292}$$

To calculate the signal to (digital) noise ratio of an ADC we assume we convert a sine wave exploiting the full range of the ADC without clipping. To prevent clipping the effective voltage of the sine wave must be

$$V_{sineeff} = \frac{V_{fullrange}}{2 * \sqrt{2}} \tag{8.293}$$

Since the full range of an ADC with n bits is

$$V_{fullrange} = 2^n * V_{LSB} \tag{8.294}$$

The signal to noise (SNR) of the ADC becomes (in dB):

$$SNR = 20 * log10(\sqrt{3/2} * 2^n) \tag{8.295}$$

This can be written in a more convenient way:

$$SNR = n * 6.02dB + 1.76dB \tag{8.296}$$

Example 1: A 1 bit ADC has a SNR of 7.78dB. (So a 1 bit ADC has a total distortion of 16.6%)
Example 2: A 8 bit ADC has a SNR of 49.9dB

8.10.3 FLASH ADC

A flash ADC is a converter with many comparators working in parallel to achieve the highest possible speed. It can e regarded as a straight forward extension of the 1 bit ADC presented before.

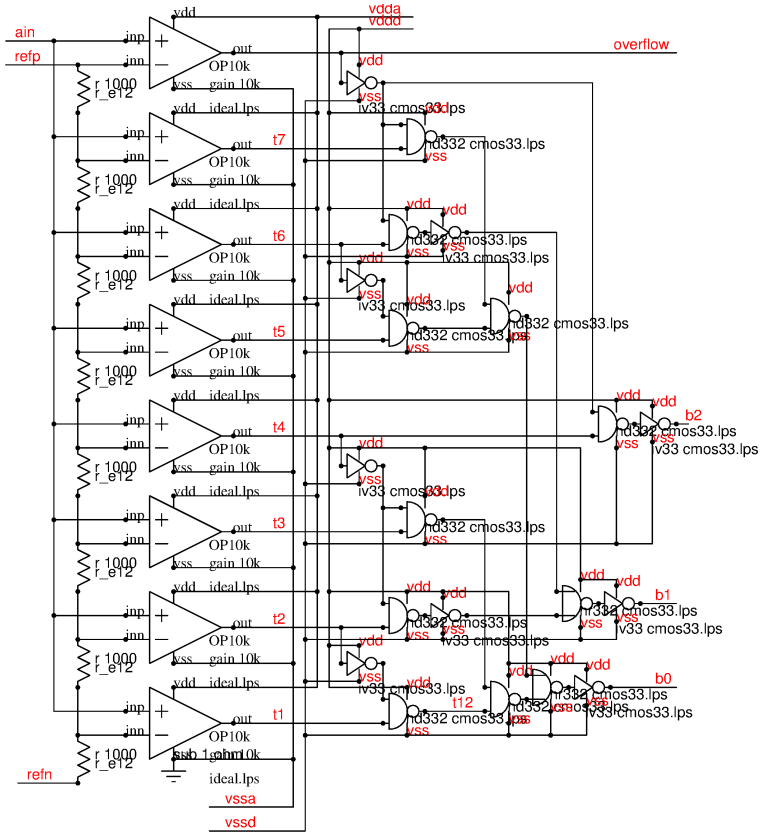


Figure 8.191: Core of a 3 bit flash ADC

The simple 3 bit flash ADC core shown above already demonstrates the limitations of the flash converter concept:

1. The number of analog comparators is very high leading to high chip real estate, high power consumption and low yield
2. The decoder logic explodes with the number of bits of the ADC

So flash ADCs are restricted to applications with low resolution (usually 5 to 9 bit) and high speed. Classical applications are oscilloscopes and RADAR signal conversion.

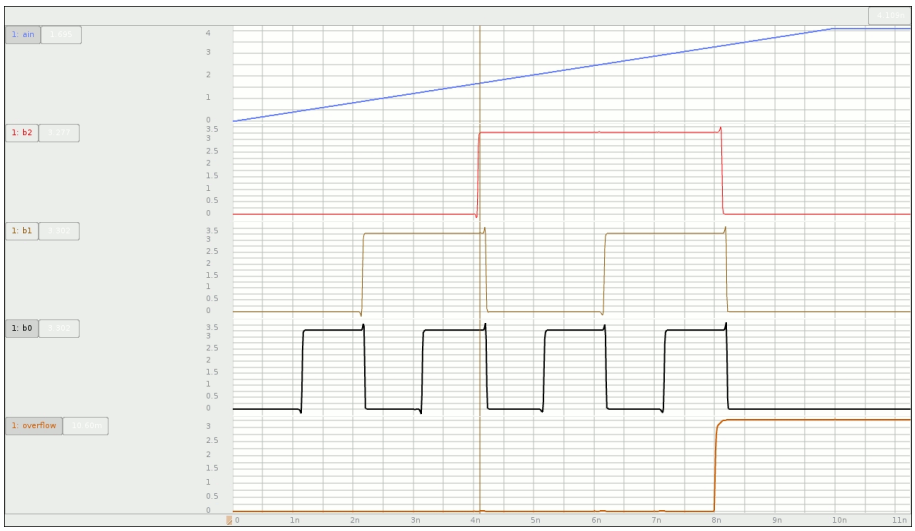


Figure 8.192: Simulation of the 3bit flash ADC

The simulation already shows one issue of the flash converter. Due to the different propagation delays of the parallel paths the conversion is not glitch free (at the cursor position there is a glitch with data b111 in the middle of the range). The converter must always be used in combination with a sample & hold input stage and flip flops that sample the output data while the signal is frozen. Even this doesn't guarantee the converter always works correctly. The best solution is to convert into Gray code (So if one bit is wrong the digitized signal only changes by one LSB), then latch the result and convert to binary code after the first latch.

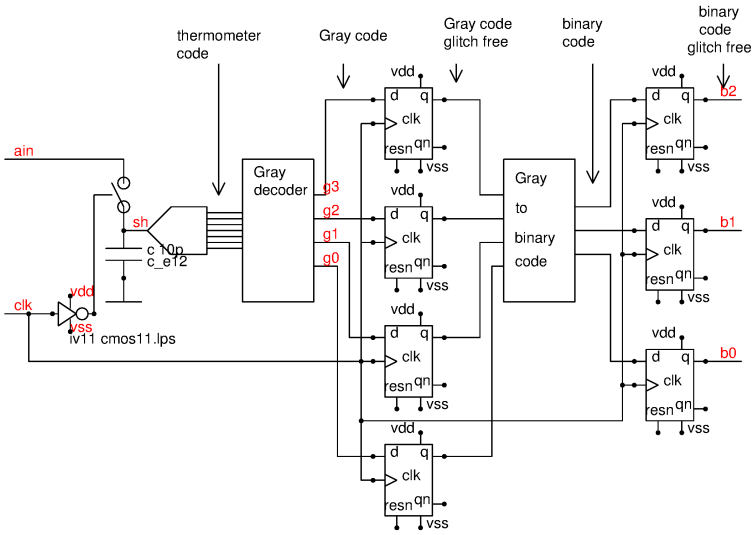


Figure 8.193: Glitch free implementation of a flash ADC with intermediate conversion to Gray code

Now the ADC is free of glitches (well, not quite, but the Gray code limits the transitional error to +-1 LSB) but we added a propagation delay of 2 clock cycles plus one clock cycle latency of the sample & hold stage.

Requirements of the comparators: The requirements of the comparators can be derived from the value of the LSB and the required speed of the ADC.

8.10.4 Successive Approximation Register (SAR) converters

To reduce the number of comparators the ADC is constructed as a comparator that compares the output of a DAC with the voltage to be measured. The result of the comparison is an indication which signal is higher. If the input signal is higher than the signal provided by the DAC the digital input of the DAC needs to be further increased. If analog input id lower than the value provided by the DAC the digital input vector must be decreased. This way the digital value at the DAC approaches the correct conversion result.

Ideally the conversion starts with the most significant bit and ends with the least significant bit. This way the conversion takes at maximum 2N cycles but it can be done with only one comparator.

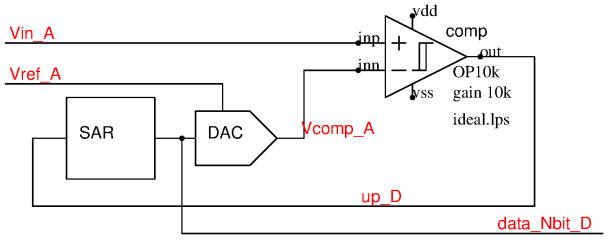


Figure 8.194: SAR ADC

The comparator simply compares the analog input signal Vin_A with the signal Vcomp_A coming from the DAC. The result of this comparison simply it an indication whether the input value of the DAC must be increased (up_D=1) or decreased (up_D=0) to match the signal Vcomp_A and Vin_A. In most cases the signal Vcomp_A simply calculates as something like:

$$Vcomp_A = Vref_A * gain * \frac{data_{NbitD}}{2^N - 1} \tag{8.297}$$

The gain depends on the design of the DAC. In many cases it simply is 1. When Vcomp_A crosses the input signal up_D will toggle.

The most simple way to implement the logic sitting in the SAR block is just building an up/down counter. But this is very slow because if Vcomp_A and Vin_A differ significantly Vcomp_A only approaches Vin_A is LSB teps. Worst case the conversion using an up/down counter takes 2^Nclock cycles.

The much more elegant and faster implementation is a successive approximation register. The conversion starts with the most significant bit and ends with the last significant bit. It can be written like a software loop:

Algorithm 3 SAR algorithm

```
/* start condition all bits 0 */
for (i=N; i>=0; i--)
  bit[i] = 0;
/* check for underflow because it comes for free */
if (up_D == 1)
  {
  underflow=1;
  return underflow;
  }
/* SAR algorithm */
for (i=N; i>=0; i--)
  {
  bit[i] = 1;
  if (up_D == 1)
    bit[i] = 0;
  }
/* end of SAR. now vector bit holds the conversion result */
```

The core of the SAR algorithm starts with the most significant bit (MSB) and tests if the signal Vin_A is below or above the 0.5*Vref_A. Depending on the result the MSB is kept or taken back. Then the same test is done with the next lower significant bit to test for the quarter of the range. This is repeated until the least significant bit (LSB) is reached. Depending on implementation (how much time it takes for the comparator to provide the result up_D) the SAR algorithm requires between 2*N and 4*N cycles. (4*N assumes the comparator requires 2 clock cycles).

Typical implementation use a clocked comparator that requires 1 clock cycle. Therefore most SAR converters require 3*N clock cycles for the conversion.

8.10.5 ADCs as a load

ADCs are intended to convert an analog signal into a digital signal with the lowest possible error. The precision of this conversion is influenced by the precision of the voltage at the input of the ADC. The ADC acts as a load for the driving stage. The driving stage ideally should be an ideal voltage source. In real applications the input of the ADC either is driven by a voltage divider (attenuator) or a buffer amplifier.

The input of an ADC can be resistive (for instance resistor coupled flash converter) or capacitive (capacitor coupled clocked converters).

As long as the resistive behavior dominates the input resistance of the converter and the output resistance of the driving stage act as a voltage divider with a ration very close to 1 but not exactly 1. This leads to a reduction of the input signal or in other words a gain error. Let's assume the ADC has an input impedance of $R_{adc} = 1M\Omega$ and the source has an output impedance of $R_{source} = 9k\Omega$. The resulting gain error calculates as:

$$Err_{gain} = 100\% * [R_{adc}/(R_{adc} + R_{source}) - 1] \tag{8.298}$$

In the example the gain error is -0.892%. This kind of gain error can be compensated, as long as the resistances are constant.

Leakage currents CMOS transistors used as switches at the input of an ADC (switches of the ADC itself and switches used in analog multiplexers) have leakage currents that change with temperature. Together with the source impedance these currents lead to offset errors. For technologies with 3nm and higher gate oxide thickness the gate leakage usually can be neglected. The domination leakage current is produced by the bipolar junctions at the drain and the source of the transistors (bipolar diode leakage).

Charge coupled ADCs acting as a load Charge coupled ADCs have to be chaged with an accuracy of one LSB at each conversion cycle. The voltage swing present at the capacitor usually starts at 0V and goes up to Vin. The error decreases with the time given to settle.

$$Err_{RC} = exp(-t/RC)$$

The settling time required is

$$t = -RC * ln(V_{fullscale}/V_{LSB})$$

Replacing the ratio $V_{fullscale}/V_{LSB}$by the number of bits of resolution in bits N we can calculate the settling time required for one appoximation step.

$$t = -R * C * ln(2^N) \tag{8.299}$$

Assuming we use a successive approximation (SAR) vonverter the total time required for the conversion becomes

$$t_{SAR} = 2 * R * C_{ADC} * N * ln(2^N) \tag{8.300}$$

This equation assumes that the input voltage is close to 0 and every approximation has to be taken back again leading to the factor 2.

The Example: C=0.1pF, N=10bit, R=10k leads to $t_{SAR} = 10K * 0.1p * 10 * ln(1024) = 139.4ns$.

This equation applies as long as no wire capacities between the divider and the ADC are involved. In practical applications often analog multiplexers are used and the voltage divider has to drive a lot of wire capacities. The wire capacity has to be charged to 1LSB accuracy before the first conversion starts. So we get an additional analog settling time.

$$t_{wire} = R * C_{wire} * ln(2^N) \tag{8.301}$$

To prevent w wrong firs conversion the wire capacity must be charged longer than t_{wire}before the first conversion starts.

The time needed for settling the complete system consisting og the divider with output reistance R, the wires with capacity $C_{wire}$and the SAR ADC with capacity C_{SAR}becomes

$$t_{system} = t_{SAR} + t_{wire} = R * ln(2^N) * (C_{wire} + 2 * N * C_{ADC}) \tag{8.302}$$

As an example let's assume we use the same ADC together with a multiplecer and a wire capacity of 5pF:

$$t_{system} = 0.485\mu s$$

$$t_{wire} = 0.346\mu s$$

Modulators and Demodulators

There are many parameters of a carrier that can be modulated:

1. amplitude
2. frequency
3. phase (well, can be regarded as frequency modulation as well)
4. any combinations of the above leading to phase modulation and single side band

Besides these very fundamental ones there are dozens of other such as QAM (quadrature Amplitude Modulation), APSK (Amplitude and Phase Shift Keying), to just name some of them. But they all boil down to a combination of changing anplitude and phase at the same time to get more (usually discrete) symbols into a signal space with the dimensions phase and amplitude. This means most of these complex modulation schemes are just a superposition of amplitude modulation and phase modulation on the same carrier. If you understand AM and FM and the meaning of phase changes the rest is no rocket science.

8.10.6 Amplitude modulation

The most traditional way of building a modulator is the amplitude modulator (AM). The easiest way of building an amplitude modulator is a gain adjustable amplifier. It can be implemented single ended because usually the base band is removed in a filter behind the modulator. Not very elegant and a lot of distortion - but this primitive design already does the job.

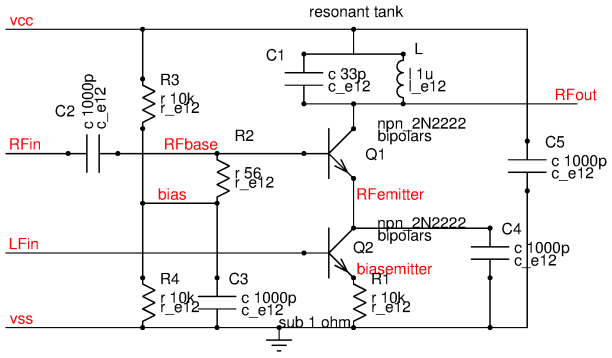


Figure 8.195: classical AM modulator

This very old circuit was used for AM modulation with low effort (low transistor count). The bias current of Q1 is determined by the voltage at LFin and the resistor R1. Changing the current flow through Q1 the gm of transistor

Q1 changes. This changes the gain of the amplifier consisting of Q1 and the resonant tank C1, L. The resonant tank removes the base band from the signal present at RFout. Only signals with about 27MHz will pass this filter.

Minimizing the number of transistors while accepting resonant tanks was a very typical approach of the 1960s when transistors were much more expensive than inductive components.

Assuming the RF carrier is a pure sine wave and the base band signal is a constant operating point (represented by the "1") and a sine wave with amplitude m the resulting signal becomes:

$$Vout = sin(\omega_c * t) * (1 + m * sin(\omega_m * t) \tag{8.303}$$

Since the current can't be reversed by Q2 the amplitude of the modulation must always be less or equal than 1. Solving the above equation yields:

$$Vout = sin(\omega_c * t) + 0.5 * m * [cos((\omega_c - \omega_m) * t) + cos((\omega_c + \omega_m) * t)] \tag{8.304}$$

This means the output signal consists of the carrier with amplitude 1 plus two side bands at $\omega_c - \omega_m$ and at $\omega_c + \omega_m$ both with an amplitude of 0.5*m.

The modulated information is in the energy of the side bands. This is why the modulation index m should be chosen as close to 1 as possible. Making m=1 however leads to distortions (in the modulator as well as in the demodulator). To keep distortions low most AM radio transmitters keep m in the range of 30%. This is a reasonable compromise between efficiency and distortions.

The amplitude of the side bands compared to the carrier often is expressed in dB. Usually the reference to the carrier is notified by a little "c" naming it dBc. An AM transmitter with m=100% has -6dBc. (The side bands have half the amplitude of the carrier)

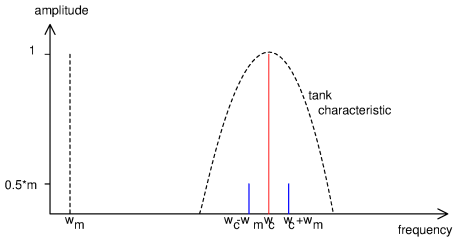


Figure 8.196: Spectrum of an AM transmitter

Only the two blue lines transport information. The red line requires energy but doesn't transport any information. The dashed line will not get through the filter. (it is present in the collector current but there is no voltage swing)

8.10.7 AM demodulation

Demodulation of an amplitude modulated carries is just as easy as modulating it. A rectifier will do the job. The demodulator usually consists of an input filter to remove all signals except for the carrier (and the side bands) you re interested in. After that there is a rectifier and a low pass filter letting the demodulated signal pass, but not the carrier.

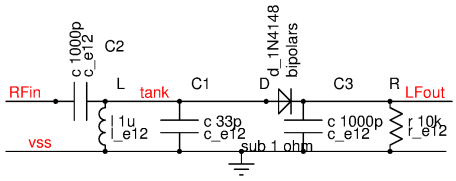


Figure 8.197: AM demodulator with diode

This is the most classic way of demodulating an AM signal. To achieve a good sensitivity a diode with a low forward voltage is preferred. This is why for AM demodulators germanium diodes have been in use for a long time.

8.10.8 AM modulation with suppressed carrier

To enhance efficiency and AM modulation without the carrier offers a lot of advantages. instead of using a single ended modulator we need a modulator that performs a multiplication with positive as well as negative sign. The equation is easy:

$$Vout = sin(\omega_c * t) * sin(\omega_m * t) = cos((\omega_c - \omega_m) * t) + cos((\omega_c + \omega_m) * t) \tag{8.305}$$

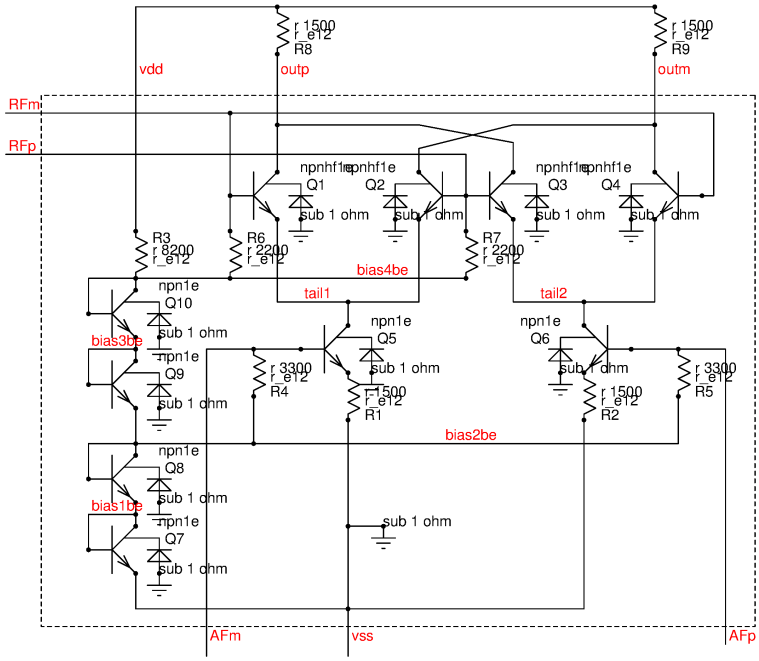


Figure 8.198: The SO42 as an example of a Gilbert cell (the chip is inside the dashed rectangle)

Now the carrier is gone. The circuit required to perform a real multiplication is a bit more complex. We need a Gilbert cell. A typical example is the SO42 chip [64]. The Gilbert cell relies on good matching of the transistors. For this reason it doesn't make much sense to implement a Gilbert cell using discrete devices.

The input signals must be applied fully differential now. The transistors Q7 to Q10 are just bias generators. Q5 and Q6 create the bias currents of the two differential stages Q1, Q2 and Q3, Q4. The output currents of the differential stages are added with inverted signs. If there is no differential voltage applied between AFp and AFm both differential amplifiers have the same gain and the signals at the output cancel. The carrier is rejected this way.

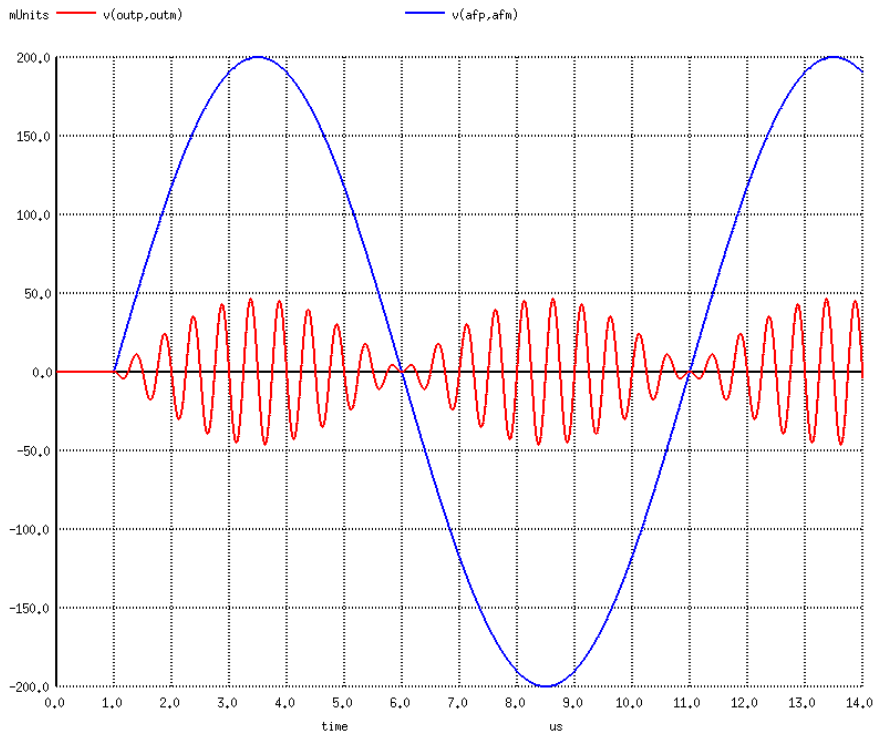


Figure 8.199: Simulation of the SO42 Gilbert cell

From t=0 to t=1us the differential input voltage of the inputs AFp and AFm is 0V. So there is no carrier visible at the output. Starting at 1us the input has a sinusoidal signal. Note the phase jumps of the output signal at the zero crossings of the modulation.

An FFT of the signal (command "fft V(outp)") shows that the signal in fact has the two side bands but the

carrier (that would have been at 2MHz) is suppressed.

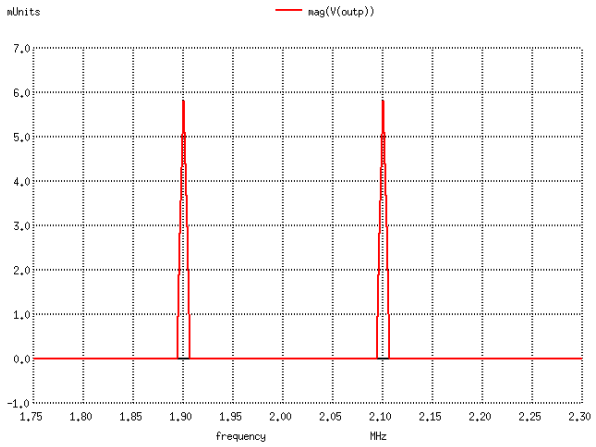


Figure 8.200: Spectrum of the signal v(outp)

This kind of modulation also is called double side band, no carrier (DSB). Since there is no more carrier in the signal the energy efficiency is much higher than using classical AM modulation.

By the way did you notice the phase shift of 180° when the modulation signal crosses the zero? You can regard this as a AM combined with phase shift modulation.

The same circuit can be implemented as a MOS circuit. To make it work nicely linear the MOS transistors must be operated in weak inversion (we need the exponential characteristic for a linear multiplication). This makes a MOS implementation slow. If we can accept signal compression the MOS Gilbert cell can be operated in strong inversion as well. In strong inversion MOS Gilbert cells can be much faster than most bipolar technologies. MOS Gilbert cells in strong inversion make sense if digital signals have to processed (here the compression of the parabola characteristic doesn't harm).

If the modulation signal is pure rectangular the Gilbert cell can be replaced by 4 switches. This approach is frequently done in pure digital systems. This example also shows that a double side band modulation with suppressed carrier with rectangular modulation is the same as a phase shift modulation with two discrete values: 0 and Π (0° and 180°).

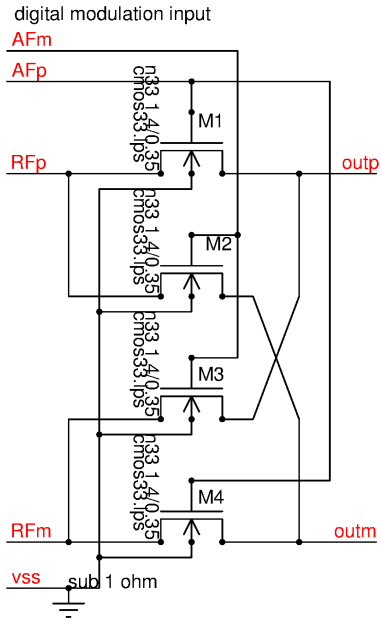


Figure 8.201: A very simple version of a DSB modulator for pure rectangular signals

This version of the modulator only switches the phase. The RF doesn't have any kind of an envelope as before. The only information remaining is the phase jump.

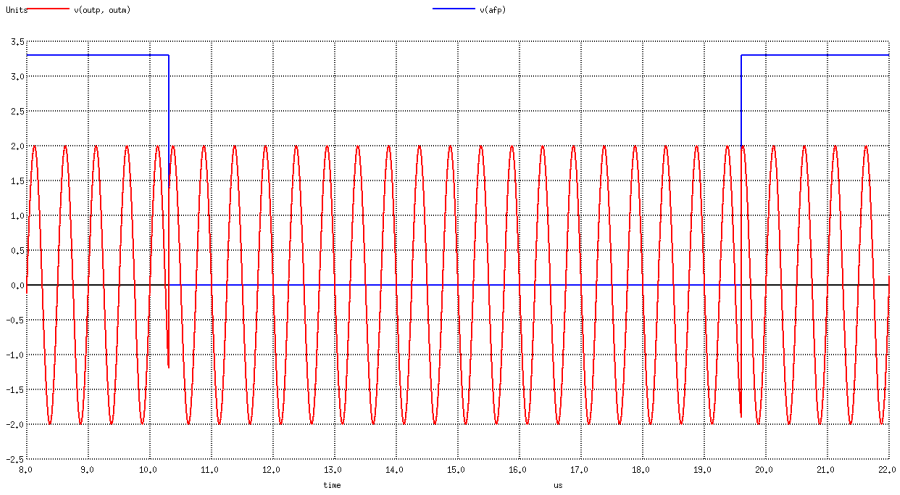


Figure 8.202: Modulation input and differential output signal of the digital modulator

The spectrum again has a suppressed carrier (there is no line at 2MHz) and the side bands of the rectangular modulation signal.

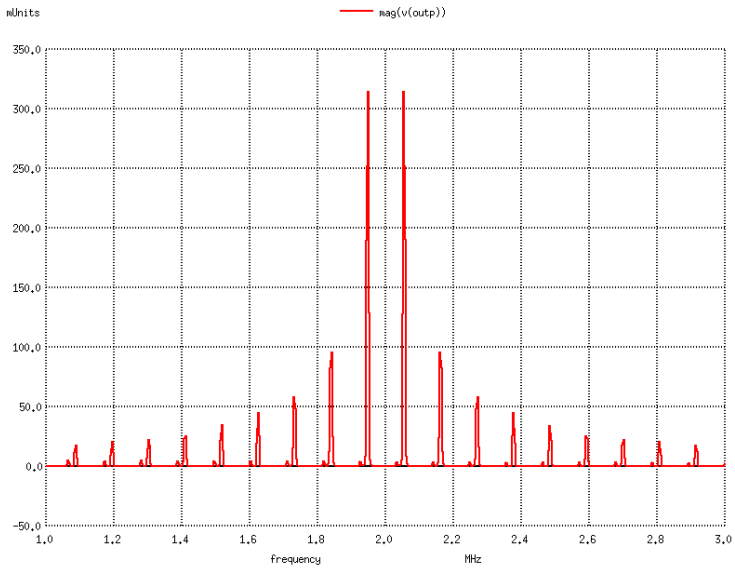


Figure 8.203: Spectrum of a DSB signal with rectangular modulation

8.10.9 Demodulation of a DSB signal with suppressed carrier

The example of the digital modulation shows that a classical AM demodulator detecting the envelope of the signal won't work anymore. The demodulator must detect the phase changes rather than the amplitude. The best way of doing this is the multiplication of the signal with the carrier. The demodulator becomes the same circuit as the modulator.

$$Vout = sin(\omega_c * t) * [cos((\omega_c - \omega_m) * t) + cos(\omega_c + \omega_m) * t)] \tag{8.306}$$

$$Vout = \frac{1}{2}[sin((\omega_c - \omega_c + \omega_m) * t) + sin((\omega_c + \omega_c - \omega_m) * t) + sin((\omega_c - \omega_c - \omega_m) * t) + sin((\omega_c + \omega_c + \omega_m) * t)]$$

$$Vout = \frac{1}{2}[sin(\omega_m t) + sin(-\omega_m t) + sin((2 * \omega_c + \omega_m)t) + sin((2 * \omega_c - \omega_m)t)] \tag{8.307}$$

The term $sin(\omega_m * t)$ is exactly the modulation we want to get back. The negative frequency component doesn't have a physical meaning.

To test the setup the Gilbert cell is placed twice in the test circuit. The carrier is applied at the inputs RFp and RFm of the first Gilbert cell and at the inputs BFp and BFm of the second gilbert cell.

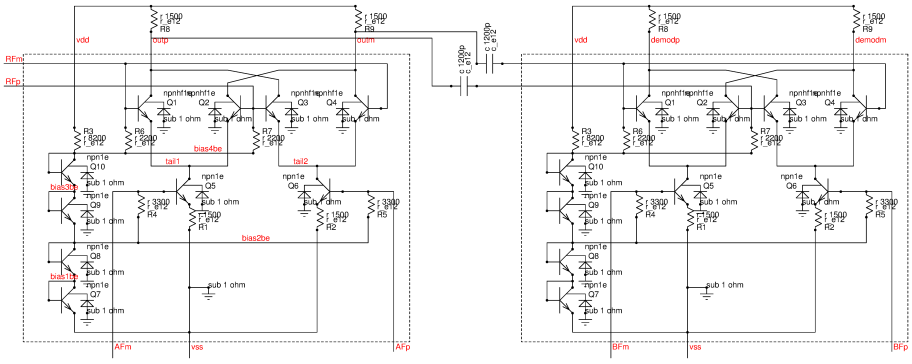


Figure 8.204: Modulation and demodulation using the same Gilbert cell

The stimuli file looks as follows:

```
*SPICE circuit <two_SO42> from XCircuit v3.7 rev 55
Vs vdd vss dc 12
Vaf AFp AFm sin 0 0.2 1e5 1u 0
Vrf RFp RFm sin 0 20m 2e6 0 0
Vbf BFp BFm sin 0 0.2 2e6 0 0
```

The simulation result is shown below:

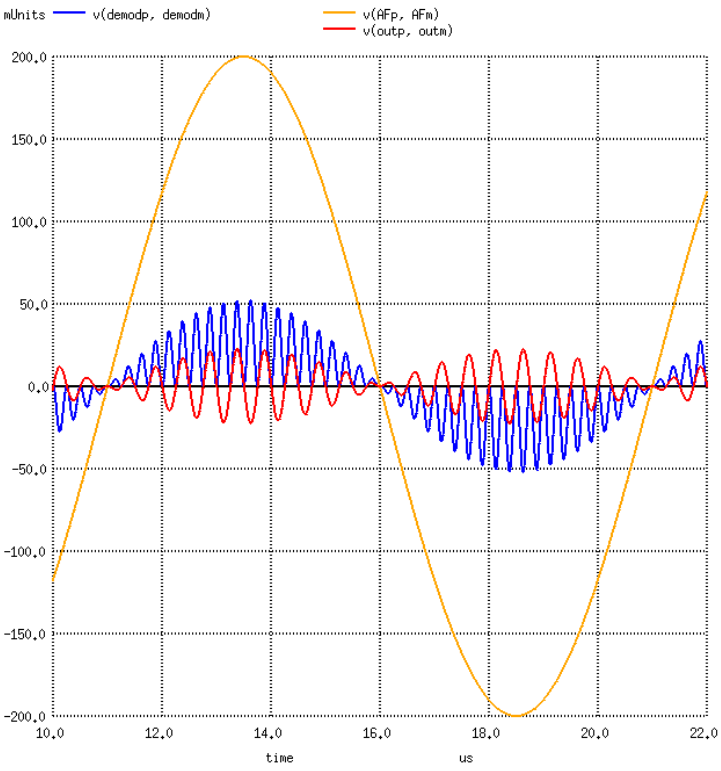


Figure 8.205: Simulation of the two Gilbert cells leads to the blue colored signal at the second stage

The output signal of the second stage holds the spectrum expected from the calculation holding the frequencies ω_m, $2 * \omega_c - \omega_m$, $2 * \omega_c + \omega_m$.

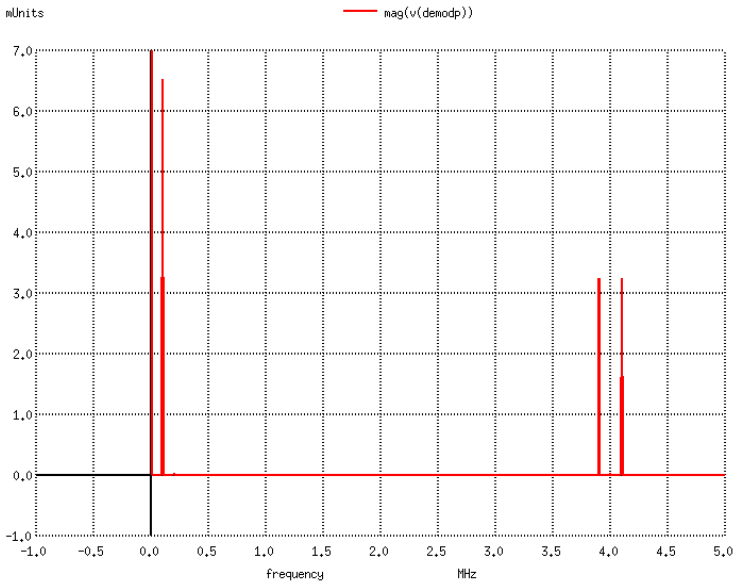


Figure 8.206: Spectrum of the output signal of the second Gilbert cell

Since the carrier isn't available the carrier must be regenerated for instance using a PLL (phase locked loop). If the second Gilbert cell is operated with a slightly different frequency than the carrier the base band signal will split in two lines above and below the original base band signal. The deviation is +- the deviation of the original carrier and the beat frequency used in the second cell. The following example shows stimulus operating the two Gilbert cells with a deviation of 1kHz.

```
*SPICE circuit <two_SO42> from XCircuit v3.7 rev 55
Vs vdd vss dc 12
Vaf AFp AFm sin 0 0.2 1e5 1u 0
Vrf RFp RFm sin 0 20m 2e6 0 0 Vbf
BFp BFm sin 0 0.2 2.001e6 0 0
```

The resulting split of the base band is 99kHz and 101kHz instead of having one line at 100kHz.

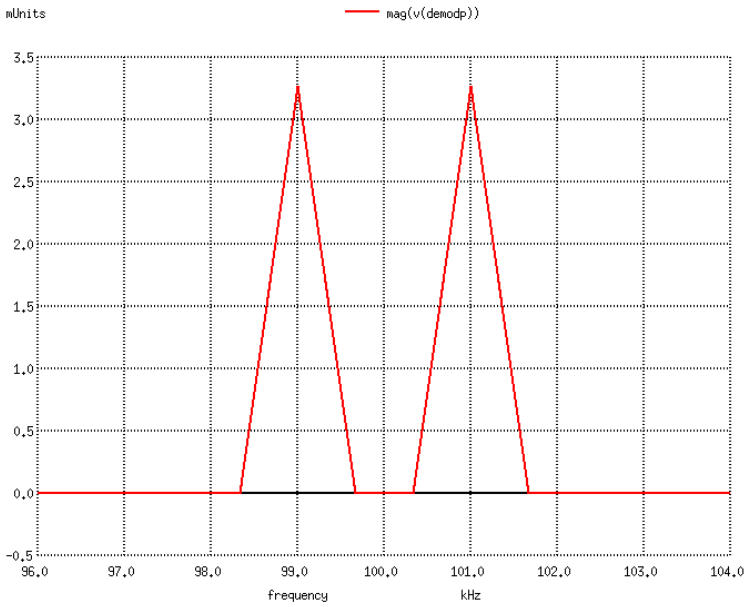


Figure 8.207: Split of the base band spectral lines due to an offset of the oscillator frequencies of 1kHz

8.10.10 Frequency modulation (FM)

Frequency modulation (FM) modulates the frequency to transport information. Usually the oscillator tuning is changed to perform the modulation. The following schematic is one possible example how to build a frequency modulated oscillator.

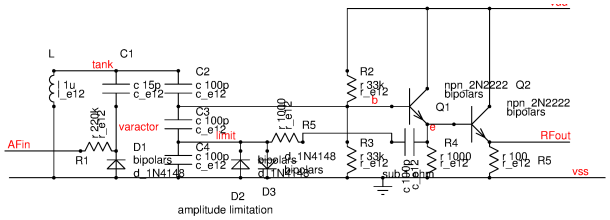


Figure 8.208: Oscillator frequency modulation by the junction capacity of D1

To achieve a stable oscillation without clipping by the transistors the feedback path has a limiter consisting of D2, D3 and R5. frequency modulation is done tuning the junction capacity of D via R1 and modulation input AFin. Transistor Q2 is just an impedance converter to make the oscillator insensitive to load changes.

To simulate the oscillator the following stimulus was used:

```
vs vdd vss pulse 0 12 0 1n 1n 10m
vmod AFin vss sin 6 5 10k 0 0
```

The side bands of an FM are a Bessel function. We will not only find the 10kHz of the modulation but also multiples of the side bands with 10kHz spacing.

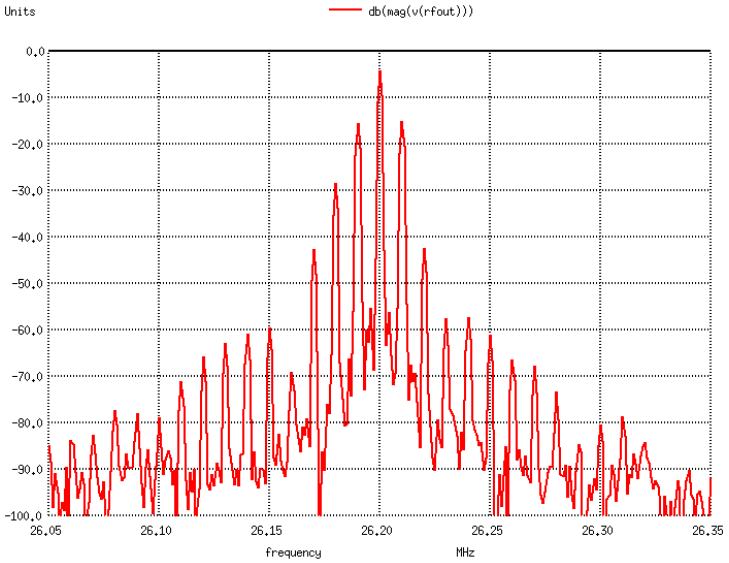


Figure 8.209: Spectrum of a frequency modulation with 10kHz modulation frequency

8.11 Chopper stabilized amplifier

Chopper amplifiers are in use at least since 1949 [36, 47] . Even chopper amplifiers using magnetic field plates as chopping elements are already described in 1967 [36]! In the beginning chopping was used to eliminate the drift of valve-amplifiers. Transistor amplifiers have a significantly lower drift and statistical offset, but as soon as offsets below 1mV are required either chopping techniques or auto zero techniques become mandatory.

Chopper amplifiers are used for two reasons:

1. Remove amplifier offsets from the signal chain
2. Fold 1/f noise out of the signal bandwidth

The input signal is converted into an AC signal by the input mixer. The amplifier amplifies both, it's own noise and offset as well as the input signal that was folded around the chopper frequency. The second mixer at the chopper amplifier output folds the noise and the offset to around the carrier and it's harmonics and the input signal back from left and right of the chopper frequency to DC and baseband.

Typically a chopper amplifier uses a switch modulator instead of a Gilbert cell. (Theoretically a Gilbert cell can be used as well, but since a Gilbert cell depends on matching it has offset too and there is no more benefit.)

Frequencies close to the chopper frequency can be folded into the use-bandwidth. This is something that is not desired. A chopper amplifier MUST have an analog anti aliasing filter at the input and at the output with a bandwidth significantly less than half of the chopper frequency.

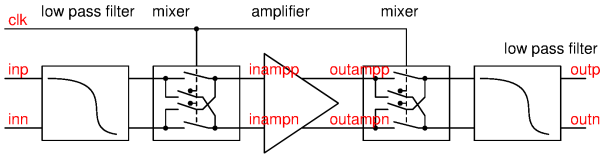


Figure 8.210: Concept of a chopper amplifier

Besides this fully differential approach single ended (asymmetrical) switching can be found as well. But asymmetric switching only works for fairly low performance requirements. For this reason asymmetrical designs are no more used today. A second reason for no more using asymmetrical modulators is that a perfectly symmetrical modulator also removes the clock. The multiplication of clock and signal without any DC component at the clock input (perfect 50% duty cycle) is a dual side band modulation with suppressed carrier

The following figure shows a simplified spectral representation of the signal.

After the first mixer the signal is folded around the clock. (the clock is a rectangular signal. So we have harmonics at (1+2N)fclk (N=0,1,2,3..) that are neglected in the simplified drawing). The clock frequency itself will only be present when the base band signal holds a DC component.

At the amplifier output before the second mixer we will find the amplified side bands (green), The amplified white noise (red), the 1/f noise of the amplifier (red, dashed fill) and the amplified offset of the amplifier at f=0.

After the second mixer the signal and the white noise are folded back to the base band. (green and red filled). The offset and the 1/f noise are folded to the clock frequency (offset) and into the side bands around the clock frequency.

The low pass filter at the output removes the 1/f noise and the offset that is folded around the clock frequency. Only the amplified signal (with green fill) and the amplified white noise (red fill) are present at V(outp, outn). Since the 1/f noise usually is much bigger than the resistive noise for frequencies below 10kHz this trick to get rid of the 1/f noise is quite versatile.

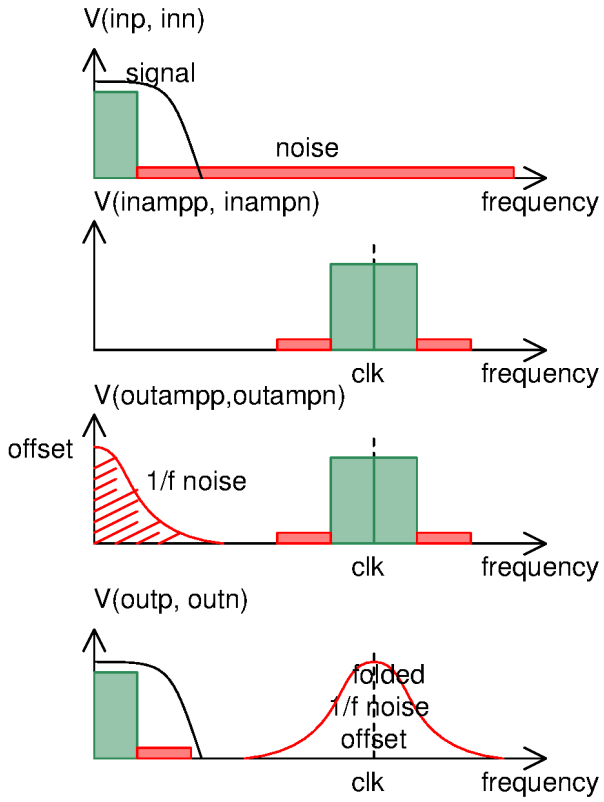


Figure 8.211: Spectral distribution of the signals at different positions of the signal chain

The advantages of chopping are paid with requirements for the amplifier and the switches:

1. The amplifier bandwidth must be significantly higher than the chopping frequency. (Theoretical minimum 3 times the base band bandwidth, typically 10 times the base band bandwidth!)
2. Due to the high bandwidth the amplifier has a high current consumption.
3. The switches must be designed for low clock feed through.
4. To cancel remaining clock feed through the slew rates of the amplifier for rising and falling edge must match

5. The duty cycle must be as close to 50% as possible

The remaining offset error caused by duty cycle deviations is:

$$V_{oschopper} = V_{osop} * D - V_{osop} * (1 - D)$$
$$V_{oschopper} = V_{osop} * (2 * D - 1) \tag{8.308}$$

Clock jitter will lead to duty cycle fluctuations. This converts amplifier offset into noise.

Probably the most important argument for building a chopper amplifier instead of an auto zero amplifier is the folding of the 1/f noise. The synchronous demodulator at the output of the amplifier folds the 1/f noise into the side bands of the chopping frequency. The following analog low pass filter finally removes the noise from the signal. The beauty of noise folding has a price. The amplifier must almost perfectly settle during half of a clock period. For precision applications the bandwidth of the amplifier should be:

$$BW_{amp} > 10 * f_{clk}$$

The consequence of the high bandwidth requirement is a high current consumption of the amplifier.

8.12 Auto Zero Amplifier

Different from a chopper stabilized amplifier the signal is not shifted to an other frequency band . The signal remains in the base band. Only the offset of the amplifier is removed. But the 1/f noise remains in the signal chain.

One of the advantages of auto zero amplifiers is the lower bandwidth required compared to a chopping amplifier. The second advantage is the low frequency the switches are operated with. This leads to less clock feed through (in terms of average clock power).

The offset correction can operate incremental (This applies especially for the in the loop offset correction). The more correction steps the better the offset correction gets. This means the charging of the capacitor storing the correction voltage can be made slow (even switched capacitor solutions to pump discrete portions of charge into the capacitor can easily be implemented). This way the storage capacitor can intentionally be made the dominant pole of the regulation loop.

On the negative side is the auto zero signal, that is somewhere in the middle of the use bandwidth. So the switching noise is visible in the signal. The way this switching is visible depends on the topology used. It depends on the application which kind of distortion can be accepted.

There are several approaches of building auto zero amplifiers:

1. Zeroing while not in use
2. Ping-Pong design
3. In the loop auto zero

Zeroing while not in use is the cheapest approach. The amplifier is not permanently used. While the amplifier is not used the inputs are shorted and nulled. There is a variant of the concept: 2 amplifiers that are periodically swapped (ping pong). The one that is swapped out of the signal path is getting nulled. "Zeroing while not in use" can be done in open loop applications too (This technique is frequently used for comparators of $\Delta\sigma$ ADCs).

In the loop auto zero requires two amplifiers for nulling and for measuring. The total effort is like a ping pong topology. In the loop auto zero provides nicely continuous signals. It only works in systems with a feedback (OPAMPs in closed loop application). It doesn't work for open loop systems such as comparators.

Storage of the nulling: The adjustment must be stored in one or another way. The most classical analog approach is capacitive storage of the correction value. The capacitive storage must be refreshed regularly because of leakage. Since leakage is temperature dependent the refresh cycles must be more frequent for high temperature applications. Regarding simulation tools and simulation time analog storage is faster to design.

With digital processing getting cheaper and cheaper digital storage is getting more common. The advantage is that digital storage doesn't suffer from leakage. Repetition rates can be kept lower. The price is that and ADC (analog to digital converter) is needed. Design of a system with digital storage requires significantly more computation power for the simulation (either a high number of digital transistors to be simulated with the analog part or mixed signal simulation will be required.)

For ease of understanding let's start with analog storage.

8.12.1 Zeroing while not in use

In this "poor man's auto zero amplifier" the main amplifier is only used part of the time. While the main amplifier is not used the inputs are disconnected by switches sw1 and sw2. sw0 short circuits the inputs of the amplifier. sw3 connects the amplifier measuring the output voltage to the adjust input adj. The correction amplifier will tune the adj node until the output voltage reaches 0V.

When the amplifier is returned to operational mode the switches sw0 and sw3 open. The adjust voltage is capacitively stored. switches sw1 and sw2 connect the amplifier to inp and inn again.

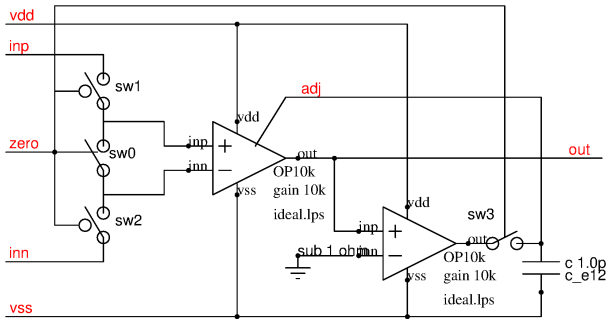


Figure 8.212: Auto zero circuit performing the auto zero while the amplifier is not operated

Since the nulling amplifier compares the already amplified signal with a reference the gain of the nulling amplifier can be kept low. This leads to a very simple design of the nulling amplifier. The main silicon real estate is in the main amplifier while the nulling amplifier often can be neglected.

This approach has limited use for applications in which the signal is only needed at certain times. Typical examples are current sense amplifiers that are only needed while the power transistor is on.

Signal folding of the simple auto zero amplifier: The simple auto zero amplifier multiplies the signal with the gating clock. Since there is no polarity change the equation looks like this:

$$V_{out} = k * D * 0.5 * (1 + rect(t))$$

K is an amplification factor. In most applications K will be determined by a resistor feedback network like for other OPAMPs in closed loop operation as well. The factor D is the duty cycle. rect(t) is a rectangular function that is +1 for 50% of the time and -1 for the other 50% of the time. In the time domain the signals look like this:

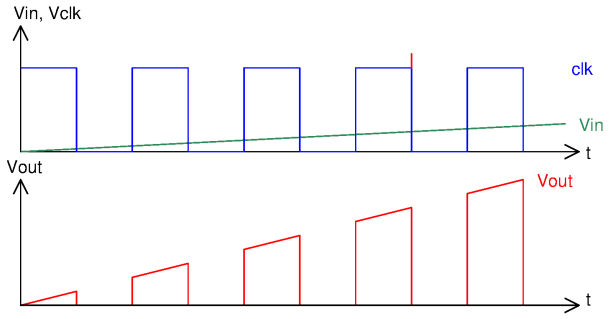


Figure 8.213: Time domain signals of the most simple auto zero amplifier

In the multiplication done in the time domain leads to a folding in the frequency domain. The colors in the frequency domain are the same as the colors used in the time domain:

- Input signal: green
- clock signal: blue
- output signal: red

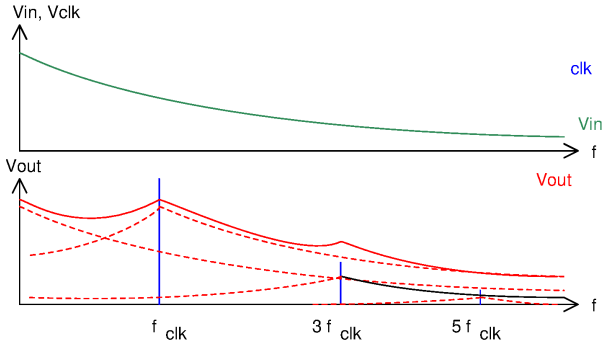


Figure 8.214: Frequency domain representation of the signals

The clock becomes visible as discrete frequencies at fclk and odd multiples of fck. The original spectrum of the input signal gets superimposed by the spectrum of the input signal folded around the clock frequency and it's harmonics.

Things get a bit more usable if we interpolate the signal when the cock is zero. Basically all we have to do is store the last value seen at the falling edge of the clock until we get a new usable value at the next rising edge. A ample and hold circuit can do this job in the analog world. In the digital world this can be implemented using an ADC and a register.

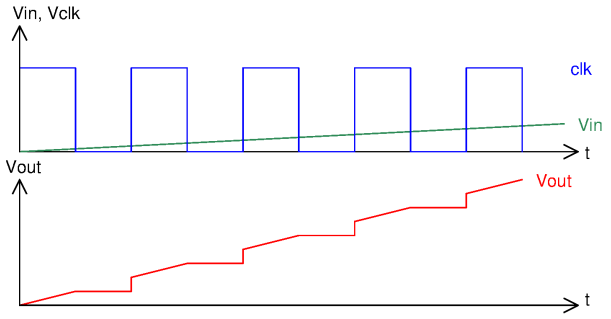


Figure 8.215: Time domain signal interpolating while the clock is 0

The question is, what are we doing in the frequency domain? The interpolation rejects the clock and all it's harmonic frequencies (well, not perfectly. There is a triangular error signal left over, similar to ADC quantization noise). For the spectrum it means we have added a notch filter at N*fclk with N=1,3,5,7.... The resulting output spectrum is very similar to the input signal except for the notches in the frequency domain.

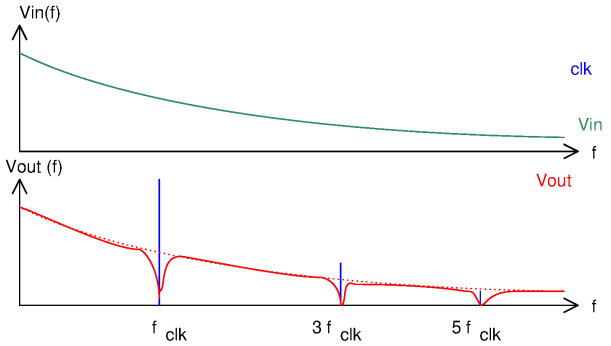


Figure 8.216: Spectrum of the simple auto zero amplifier if the signal change is ignored while the clock is 0

For many applications retrieving the original spectrum with the exception of the notches is good enough. This is especially true if the information during the time the clock is 0 isn't needed. (For instance applications that are simply not operating while the clock is 0.)

8.12.2 Ping-pong auto zero amplifier

In open loop systems without any "inactive time" a ping pong design may make sense. In a ping-pong system there are two amplifiers that are periodically changing places. One is acting as an amplifier while the other one is nulled. This double the effort but the system is permanently available.

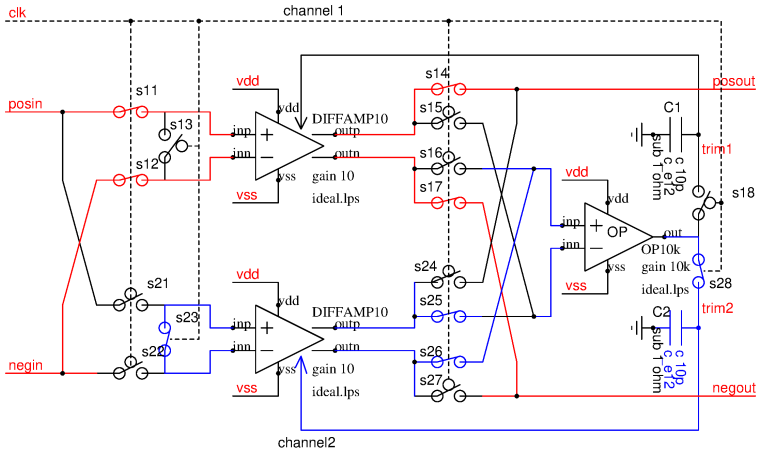


Figure 8.217: Ping Pong auto zero amplifier

In the figure above the operating path is drawn in red color and the nulling path is shown in blue color.

One of the problems of the ping pong design is the gain matching of the two channels. Even the smallest gain mismatch of the two amplifier becomes visible as a change of the output signal with the clock frequency. This disturbance increases with the input signal. For this reason the spectrum at the output of a ping pong amplifier usually holds the clock frequency and it's harmonics. As long as mainly the offset is of interest but not (mostly rectangular) clock signal superimposed on the output this can be accepted. Usually the clock is suppressed in the range of 40dB to 60dB corresponding a gain mismatch of 0.1% to 1% between the two channels.

A second effect that cant be fully avoided is the slewing of the amplifier back into the operating point when the two amplifiers are swapped. The amplifier coming from the nulling starts with an output voltage close to zero. Assuming we want to amplify a sine wave the signal looks similar to the following figure.

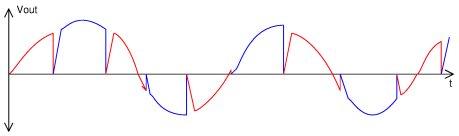


Figure 8.218: Output signal of a ping pong auto zero amplifier

I admit I overemphasized the transitions between the two amplifiers. The amplifier is operating correctly most of the time. Furthermore in mixed signal applications the output signal of the amplifier usually is sampled (for instance by an ADC). Reasonably the sample time of the ADC will be chosen somewhere in the middle between the swapping events.

In addition the settling time can be reduced dramatically if the frequency compensation is not swapped. This way the frequency compensation capacitor more or less hands over the last operating point from the active amplifier (that now enters the nulling procedure) to the amplifier taking over the signal (that was nulled before).

The signal can be interpreted as a multiplication of two signals in the time domain. One of them is the sine wave, the other one is a (1-trinagular_pulse).

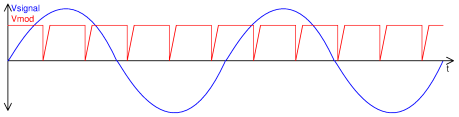


Figure 8.219: Representation of the amplifier output signal by the two functions to be multiplied

These two signals that are multiplied in the time domain get folded in the frequency domain. The constant (DC) part of the red signal folds most of the sine wave directly back into the base band. The short triangular pulses are getting folded with the sine wave. Since the sine wave is symmetrical the frequency of the triangular signal gets suppressed.

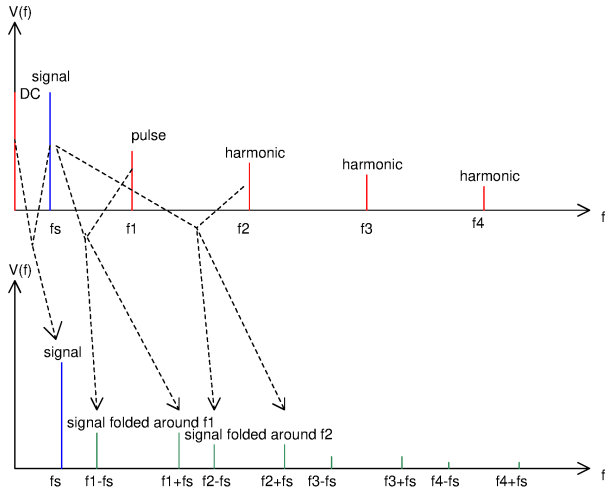


Figure 8.220: Folding of the signal (fs) around the DC component back into the base band and around the triangular pulses into side bands of double the swapping frequency

The unwanted spectral components at $fx \pm fs$ can be minimized reducing the energy of the triangular pulses as far as possible. But in a ping pong design these distortions can't be completely eliminated. In the following figure the simulation of the output of a ping_pong auto zero amplifier (without storing the operating point in the frequency compensation) is shown. In the example the signal was 1kHz and the swapping frequency was 9.8kHz (so there is a disturbance every 51μs)

The choice of swapping period of 102μs was intentional to obtain a signal that doesn't exactly repeat for a long time.

Simulation was pure transient without noise. For this reason there is no noise floor visible (Using "transient noise" the noise floor would be visible at about 10μV instead of 10nV)

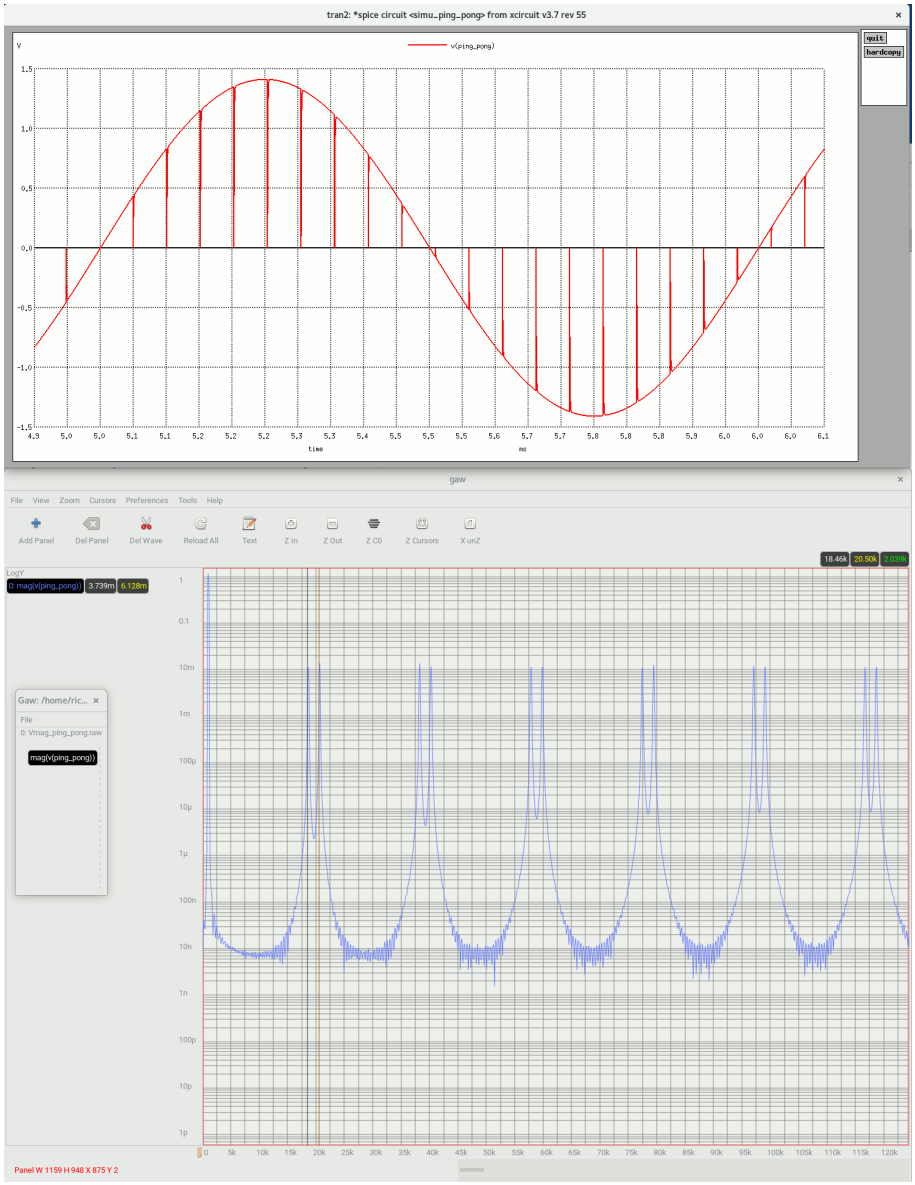


Figure 8.221: Simulation of a ping pong auto zero amplifier and the spectrum obtained

As expected the carrier at 9.8kHz is invisible and due to the folding the double carrier frequency of 19.6kHz is rejected while the two side bands are visible in the spectrum.

Typical applications are detection of fast signals with requirements for low offset but low gain accuracy requirements. One example is the fast detection of short circuits using very low resistive sense resistors or detecting open loads of a power stage. (Often both tasks are performed by the same amplifier. The open load detection requires a low offset voltage while the short detection must be fast.)

Like in the "adjust while not in use" topology the nulling amplifier OP observes the already amplified signal. The requirements of this nulling amplifier usually are low. Nevertheless the effort is almost double now because we need two amplifiers with a high gain.

8.12.3 In the loop auto zero

In the loop auto zero is the most preferred auto zero technique. It offers the advantage that the amplifier is permanently available similar to the ping pong amplifier. The ping pong design has the problem that switching between the channels can produce significant noise due to channel mismatch. In the loop auto zero topologies always use the same amplifier channel for the signal path. The switching noise gets eliminated by concept. (Practical implementations still have some switching noise due to coupling via the gate capacities of the switches. But in most cases this parasitic coupling noise is lower than the noise of a ping pong design.)

One of the first in the loop auto zero amplifiers was the ICL7650 of Intersil [66]. The basic idea is that an ideal operational amplifier always makes the (offset free) input voltage equal 0V. If the amplifier doesn't reach the 0V across it's input it obviously has an offset and need to be corrected. The in the loop auto zero consists of two amplifiers both having a high gain. The nulling amplifier needs a gain of the same magnitude as the main amplifier because it monitors the input instead of the output of the amplifier it has to adjust. For this reason the effort of an

in the loop auto zero amplifier is the same as the effort of a ping-pong design. Open loop operation of an in the loop auto zero isn't possible.

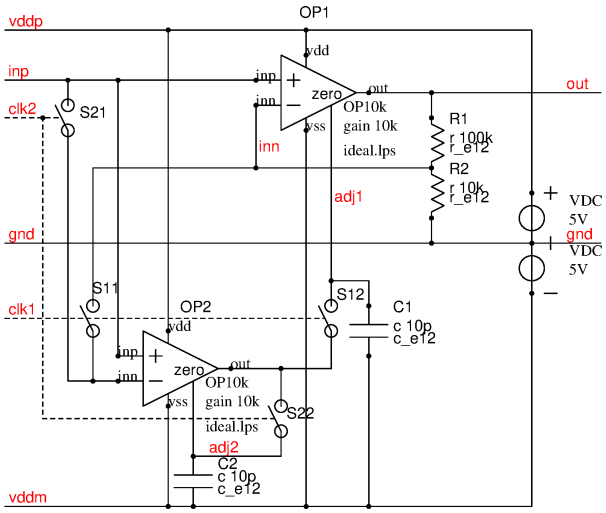


Figure 8.222: Auto zero amplifier with in the loop adjustment

The basic idea of in the loop adjustment is that an ideal OPAMP (with zero offset and infinite gain) will always bring the differential voltage between its inputs inp and inn to 0V. The signal path is running via OP1. OP2 is used for adjustment. The adjustment is done by the following steps:

1. Short circuit the inputs of OP2 closing S21
2. Close S22 to let OP2 correct it's own offset.
3. Open S22 to store correction of OP2 in C2
4. Open S21 and close S11 to measure the input offset of OP1
5. Close S12 to adjust OP1
6. Open S12 to store correction of OP1 in C1
7. Restart at step 1.

8.12.4 Beyond all measures

Theoretically it is possible to use an in the loop auto zero topology inside a copping amplifier. The auto zero amplifier reduces the offset to about $10\mu V$ and the chopper amplifier further reduces the offset to about 100nV and removes the $1/f$ noise. But what is the benefit? As soon as we have to work with signals coming from outside the chip there are many interconnects interfacing different metals at different temperatures. On low power chips the temperature differences can be kept under control to a certain extent. On high power chips the temperature gradients can already become significant.

The following table holds the Seebeck coefficients of some commonly used materials [67] at room temperature versus platinum (Pt) . Ideally each junction exists as a pair. As long as these junctions have the same temperature the Seebeck voltages cancel.

The exact numbers (e.g. of semiconductors) depend on doping and the exact alloys used. Furthermore the Seebeck constants depend on temperature. So these numbers are approximate values for room temperature only.

Table 45: Seebeck coefficients of the most important contacts found in ICs

material	usage/position	Seebeck const.
Si	gate contact, source contact	$\approx 440\mu V/K$
Ge	emitter contacts	$\approx 300\mu V/K$
NiCr	Nichrome	$25\mu V/K$
Fe	Pin of ICs	$19\mu V/K$
W	Tungsten vias	$7.5\mu V/K$
Au	bond wires	$6.5\mu V/K$
Ag	RF inductors	$6.5\mu V/K$
Cu	wires, bond wires	$6.5\mu V/K$
Pb	lead solder	$4.0\mu V/K$
Al	pad, wires on chip	$3.5\mu V/K$
C	Carbon resistors	$3.0\mu V/K$
Pt	reference	0
Ni	Nickel	$-15\mu V/K$
	Constantan resistors	$-35\mu V/K$
Bi	Bismuth solder	$-72\mu V/K$

Especially the solder joints on the board are a source of errors. Two solder joints with temperature difference of only 1K will already lead to a measurement error of some $10\mu V$. In other words designing an OPAMP that can be connected to a IC pin for offset voltages below $10\mu V$ doesn't make sense because the Seebeck voltages on the board will contribute errors that are much higher than those of the amplifier itself!

On chip it may be a different situation. On low power chips the temperature gradients are low. Amplifiers optimized for reading HAL sensors placed on the same chip as the amplifier are in fact designed to reach offset voltages of $1\sigma = 100nV$ or even better. However as soon as power stages are on the same chip the temperature gradients are higher. Combining HAL sensors with power stages on the same chip will easily degrade performance by one magnitude.

8.13 Input and output cells (IO cell)

These are the cells interfacing the IC with the outside world. They have to protect the inside parts of the IC against ESD and other kinds of overloads. At the same time the I/O cells must be capable of driving wire capacities and and resistive loads. Typically the ESD protection is part of the I/O cell.

8.13.1 Standard logic IO cells

These are cells designed for logic level. What exactly is 'logic level' depends on the application and technology. In most cases 'logic level' runs from about 1.2V to 5V. The maximum Vgs is consistent with the supply voltage.

The standard I/O cells have the following states:

Table 46: Output states of a standard logic IO cell

state	description	usage
high Z	high resistive. Cell may stoill operate as an input	used for digital BUS applications
H	logic 1. Output is pulled up to vddx	digital output
L	logic 0, Output is piulled down to vssx	digital output
weak H	high resistive but pullu up if left floating	used for digital BUS applications
weak L	high resistive but pulled down if left floating	used for digital BUS applications
analog	output is used for an analog function	analog test bus, low voltage analog

Sometimes the digital outputs have selectable drive strength as well. The drive strength is used depending on the application. Driving short wires is possible using a low drive strength. Using low drive strength reduces RF emission. If the digital I/O has to drive a tester the strong drive strength is used because most testers are connected via coax cables that can have up to 200pF capacity. During testing RF emission of an I/O usually doesn't matter.

In the I/O cell shown above the red drawn parts of the schematic are ESD protections. The green parts (N2, P2, R2, R3) are weak pull up and weak pull down structures. Pull up and pull down structures with some $10k\Omega$ or more are mainly designed to define the voltage while the output is passive. This prevents cross conduction in receiving structures.

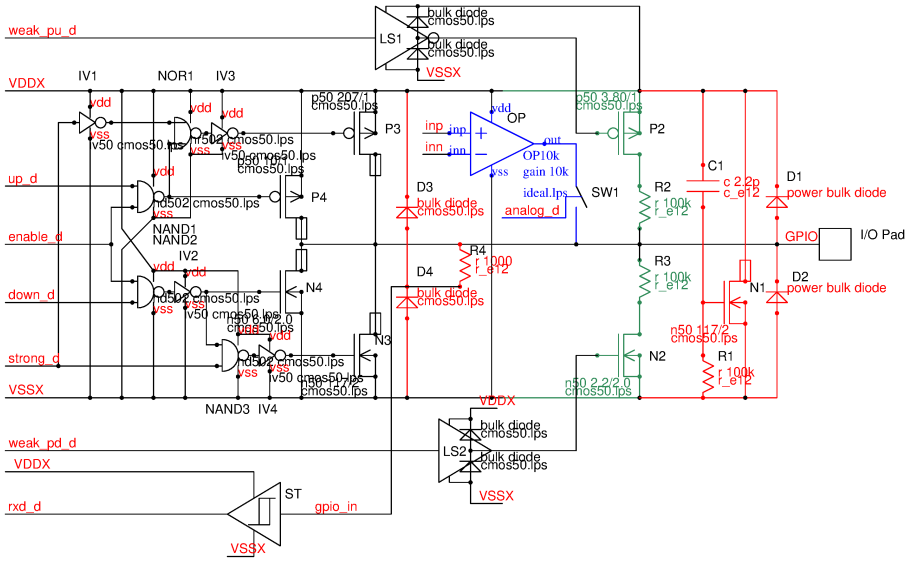


Figure 8.223: A typical 5V I/O cell

The blue colored part is an analog operational amplifier. The switch SW1 simply is a pass gate disconnecting the amplifier if the cell is used as a digital input or output. The pass gate SW1 must be designed using non salicided drain extensions because it is exposed to ESD pulses.

The digital I/O functions are the weak driver N4 and P4 and the strong driver N3 and P3. Since N3, N4, P4, P4 are exposed to ESD pulses the drains have a low doped non salicided drain extension. The digital receiver ST is protected by a secondary ESD protection R4, D3, D4. Signal rxd_d is not only an input. It can also be used to monitor if the output follows the control signals or if it is stuck at 0 or stuck at 1.

Floating inputs may need special attention. If the input voltage is at some mid level the digital schmitt trigger draws cross conduction current. For low current consumption requirements it is recommended to give ST an enable input (not shown).

8.13.2 Digital IO cells for extended voltage ranges

High voltage outputs: Classical digital IO cells are designed for 5V or 3.3V signal swing. In some cases a wider voltage range is needed. These IOs are generally called high voltage IOs. Very often these high voltage IOs are open collector or open drain outputs.

High voltage inputs: These high voltage outputs usually are combined with dedicated high voltage protected inputs. The easiest way of implementing a high voltage input is to build a low voltage input combined with an overvoltage clamp as shown in the following figure.

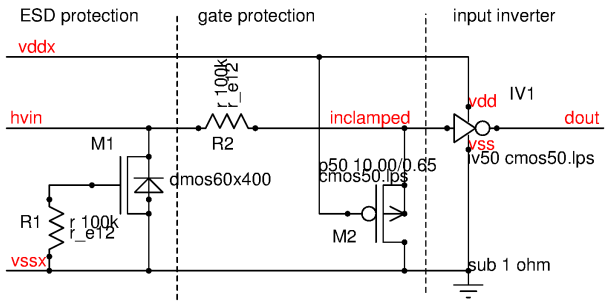


Figure 8.224: High voltage digital input

This input consists of 3 parts. M1 acts as a grounded gate NMOS ESD protection. It is needed to protect the resistor R1 and the oxide under R1 against overvoltage. Usually M1 is a fairly big transistor designed to adsorbe the enery of the ESD pulse. It's size is determined by the thermal capacity of the active area and the ESD energy it is designed for.

R1 together with M2 clamps the signal to protect the gates of the inverter IV1.

The inverter usually is the smallest part of the IO cell.

The input characteristic of this kind of input is determined by the bulk diode of M1, The clamp voltage of M2 (vddx+vth) and the resistor value. Below -Vf there is a low resistive path to vssx. Between -Vf and vddx+vth the input is high resistive. Above vddx+vth the input resistance is determined by R2.

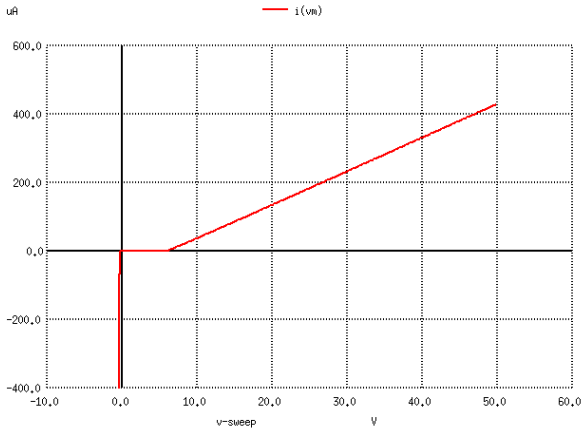


Figure 8.225: DC input characteristic of the simple high voltage input

Since high voltage inputs often connect to wires exposed to RF injection the RF behavior is of interrest. The bulk diode of M1 will act as an RF rectivier shifting up the average voltage at net hvin if the amplitude is big enough. Since the trip point of IV1 is about 2.5V the critical amplitude where rectification starts is about 3V. In a 50 Ohm system this roughly corresponds 26dBm. The range the input is rugged against RF injection can be widened replacing the simple grounded gate ESD protection by anti serial zener protections.

The input of IV1 is protected by the capacity of the clamp M2 and the resistor R2. If required the protection can be improved adding a capacitor in parallel with M2.

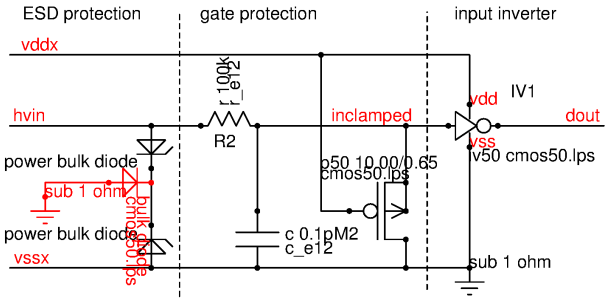


Figure 8.226: Enhanced high voltage input with ESD protection allowing the input to swing negative

The anti serial zener diode ESD protection can be built placing the (rec colored) parasitic substrate diode in the mid tap. This way the input can swing negtative. The input characteristic changes accordingly.

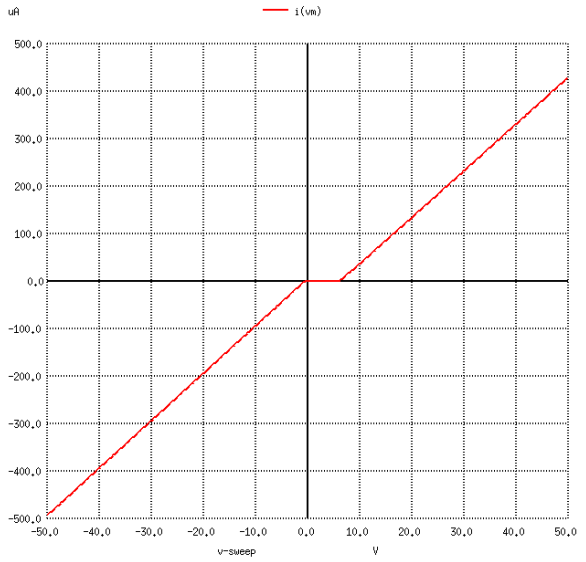


Figure 8.227: DC input characteristic of the optimized HV input permitting negative voltages and reducing RF rectification at the ESD protection

RF reaching the input of the inverter is reduced by the resistor and the capacitor acting as a low pass filter. This

however reduces the speed of the input too.

A frequently found method building high voltage inputs is using PNP transistors. In most technologies PNP transistors use the low doped epitaxy a a base. So the base emitter break down voltage and the base collector break down voltage is in the range of several 10V. The advantage of the PNP input is the lower input current up to the break down voltage of the base. (Usally there is a zener gate clamp limiting the voltage because a base break down damages the transistor!)

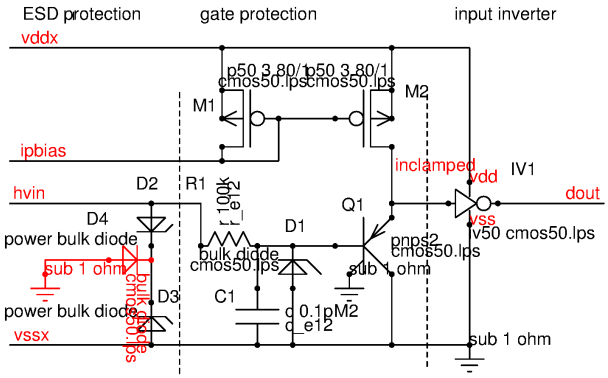


Figure 8.228: Using the high base break down voltage for a high voltage input stage

The DC characteristics are mainly determined by the bias current and the gain of the PNP transistor. The input current (neglecting leakage of ESD protections etc.) calculates as:

$$I_{in} = \frac{I_{pbias}}{B} \tag{8.309}$$

This leads nicely low input currents in the nA range to uA range.

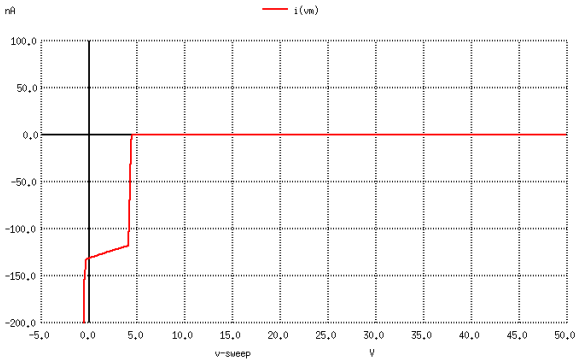


Figure 8.229: DC input characteristic of a PNP input stage

Between about -0.5V and 4.5V the input current is determined by the PNP transistor gain. Above vddx the transistor is in off state and the input current depends on the leakages of the ESD protections.

The most important disadvantage of the PNP input stage is the rectification of RF. The negative half waves of RF reaching the base of Q1 will rapidly discharge net inclamped. During the positive half wave the bipolar transistor turns off and the emitter capacity of Q1 and the drain capacity of M2 will only be pulled up by the bias current. This leads to a negative peak rectification. Therefore the RF filter R1, C1 is much more important to protect the stage against RF injection than in the resistor input shown before.

A second problem of this kind of input stage is the sensitivity to substrate noise. The base of a (non isolated) PNP transistor as well as the cathode of most zener diodes have a significant substrate capacity. If the substrate bounces (compared to the ground of the circuit) the substrate bounce becomes visible at the input (base of the PNP transistor) and gets rectified as well.

An improved version of the circuit might use an isolated vertical PNP transistor and isolated zener diodes. This modification just swaps the problem from substrate noise to supply noise injection. In case of stacking isolated zener diodes and non isolated zener diodes (for cost reasons. isolated zener diodes require more space) the substrate noise still remains visible as long as the drop over the isolated zener diodes is low. (At low blocking voltage the capacity Cbe of the transistor used as a zener still is considerably high and the noise present at the anode still gets coupled to the cathode.)

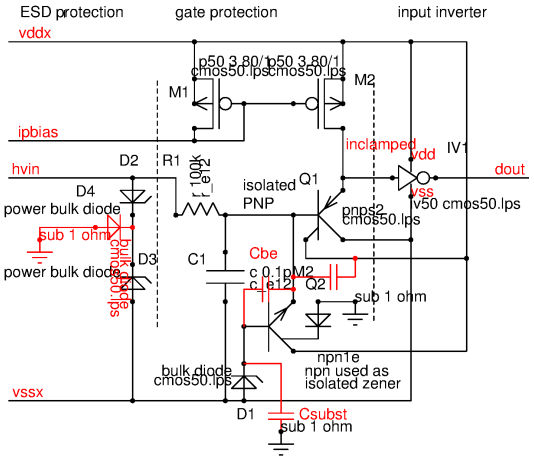


Figure 8.230: Stack of zeners still doesn't protect against substrate bounce

The benefit of replacing the standard components by isolated components in most cases is limited. Having a sufficiently high capacity C1 is a must to protect this kind of input from RF injection.

Modern semiconductor processes often don't permit using the vertical PNP transistor anymore. This doesn't mean there is no vertical PNP anymore. It usually just means it was not characterized and therefore it is not allowed to use it.... In these processes often fhe following high voltage input stage is used.

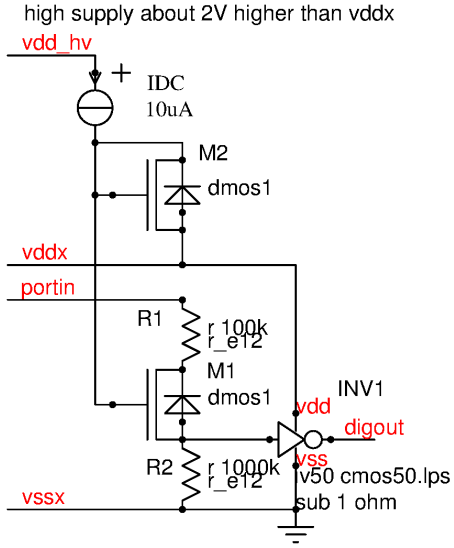


Figure 8.231: High voltage input with DMOS protection

The DMOS transistor limits the voltage swing at the input of the inverter. The port needs an additional supply that is about 2V higher than vddx for the gate of M1. Without this additional supply the swing at the input of the inverter is limited to less than vddx and the inverter will consume cross conduction current or will not switch propperly.

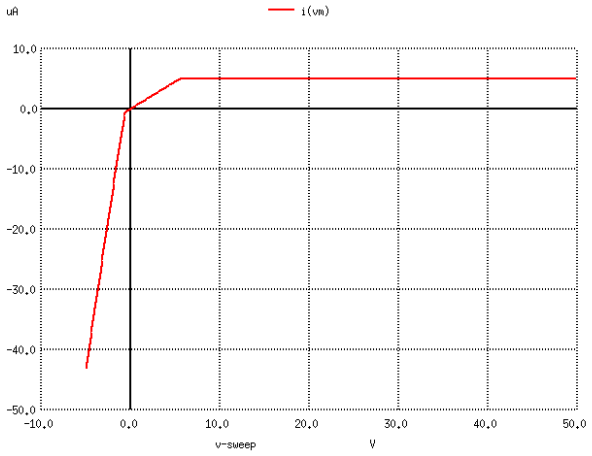


Figure 8.232: DC input characteristic of the high voltage input with dmos protection

The slope below -0.6V is defined by the substrate diode of M1 together with resistor R1.

One drawback of this protection is that in most technologies the drain of M1 is capacitively coupled either to substrate (drain=epi).

In some technologies isolated DMOS transistors exist. In this case the source of M1 usually is capacitively coubled to the supply of the nwell the transistor is sitting in.

In oxide isolated technologies (ABCD or similar) it depends on the details of the design of M1 where the parasitic capacities of M1 are connected to (N-tub of P-tub inside the oxide isolation ring?).

8.13.3 LIN

LIN is the abreviation for Local Interconnect Network. It is used for slow bus systems in the automotive world with data rates up to 20Kbit/s. The bus system had a predecessor: The ISO9141 bus. ISO9141 was a simple test bus intended as a maintenance access to the car electronics in the garage. The following table compares the features of typical implementations of LIN and ISO9141.

Table 47: Comparison of LIN and ISO9141 interfaces

parameter	LIN	ISO9141
supply voltage	6V..18V	6V..40V
slew rate	limited	no limit (higher RF emission)
data rate	<20Kbit/s	mostly <100Kbit/s
reverse polarity proof	-16V	-16V
internal pull up	yes (20K)	possible, but not required
standard protocol	yes	no

Especially the higher RF emission of ISO9141 does not permit using the ISO9141 outside of the garage.

First LIN implementations used the concept of ISO9141 but with slower switching edges. These designs were built with just a few components because the main objective of LIN was low cost. A typical implementation of the early days of LIN (about 1997) is shown in the following figure:

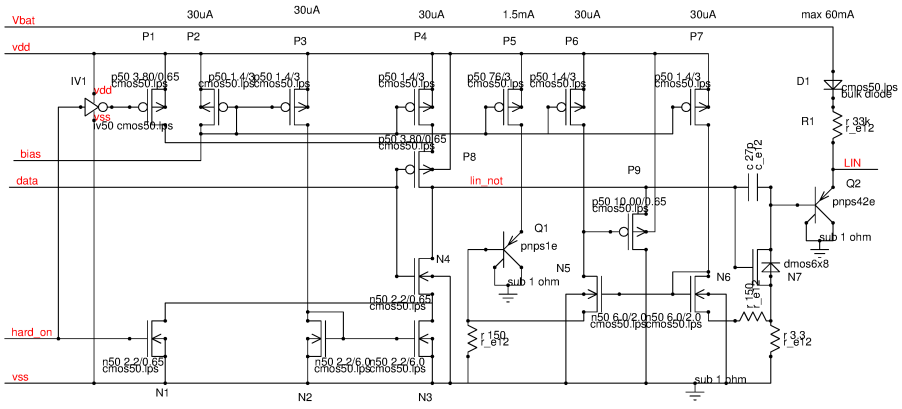


Figure 8.233: LIN transmitter stage (simplified) of HC12GA32 and HC12G60 processors

During the switching slopes the driver current flows via N4, N3 and P4, P8. After a certain time the signal hard_on becomes a logic 1 and the gate driver pulls the gate of N7 hard to vss or vdd. (Via N1 and P1).

Since a standard CMOS process doesn't provide a lossless diode substrate PNP transistors had to be used. So the collector current could not be measured. Instead the base current of a reference transistor was measured and compared to the base current of the output transistor. This comparison was used for the short circuit protection of the output stage.

This first generation of LIN drivers had a couple of weaknesses:

1. With increasing temperature the gain of the substrate PNP transistor increases significantly. The signal used to measure the current becomes smaller and smaller with increasing temperature. So the accuracy of the current limit becomes very poor at high temperature.

2. RF reaching the drains of P8 and N4 will not be shorted to vss or vdd while N1 and P1 are off. So RF injection gets rectified at the bulk diodes of N4 and P8. Typically a voltage of about vdd/2 will be established at the signal lin_not turning on the output stage in stead of producing a smooth slope.

3. Before the output transistor switches on or off the current generators N3 or P4 have to charge or discharge the gate of the power transistor. This leads to additional delays in the driver stage. To achieve the required data rates the slew rate must be designed faster than ideally required.

4. The slew rates depend on the spread of the miller capacity and the feedback capacity and the current sources. This further narrows down the time available for slewing.

5. The miller capacity (Cgd) is voltage dependent. So the edges become non linear.

To overcome these weaknesses the second generation of LIN transmitters was built differently (about 2002). As a first step manufacturers started to use process lines with much lower gain B of the PNP transistors. This was mainly achieved introducing a burried (N+) layer in the base of the PNP transistors. In extreme cases the substrate PNP transistors became more like a diode (B=0.05 or less). This way the tolerances of the current limitation could be reduced significantly.

To solve item 2 the driver stage was replaced by a low impedance class AB buffer or a transliear loop. This low impedance driver can better absorb RF injected via the miller capacity of the output stage.

As a consequence the slope shaping via the miller capacity did not work anymore. One way arround it was to use current mirror output stages and to rely on the correct pull up value to be used in the application. The low impedance driver also solved the item 3 of the list above.

Throughout the years the LIN specification was tightened more and more. Today LIN drivers are very complex designs even including regulation loops and digital curve shaping.

Regulation loops in an interface driver leads to a high risk of RF sensitivity. To protect the regulation loops on chip RF filtering is needed.

At the same time digital curve shaping leads to RF emission lines at the clock frequency (and multiples of it) caused by the step wise approximation of the target curve shape. LIN transceivers today have moved far away from the concept of a simple low cost design. Many of them today are extremely tweaked to the process they are using. Porting a LIN IP from one semiconductor process to an other has become a big effort now.

8.13.4 Differential signal interfaces overview

In noisy envirnment it is common practice to use differential interfaces. The data to be transfered is present on two wires using opposit polarity. Distortions comming from off chip sources are present on both lines as a common mode signal while the payload is a differential mode signal. Similar concepts are used in high speed logic such as ECL (emitter coupled logic, used in bipolar technologies.) or more modern in CML (current mode logic, usually implemented in MOS technologies).

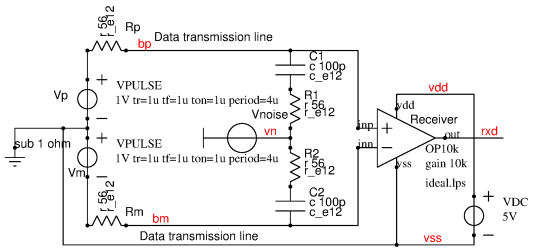


Figure 8.234: Concept of a differential signal interface

In the figure above the concept of a differential data transfer is shown. The transmitter consists of the voltage sources Vp and Vm producing a ditterential mode data signal on wires bp and bm. Rp and Rm represent the output impedance of the driver stage. Vnoise couples a common mode noise on the wires (via C1, C2, R1, R2). So the

receiver sees vp+vnoise at the positive input and vm+vnoise at the negative input. In spite of the noise the receiver reading the differential signal only will still be able to read the data.

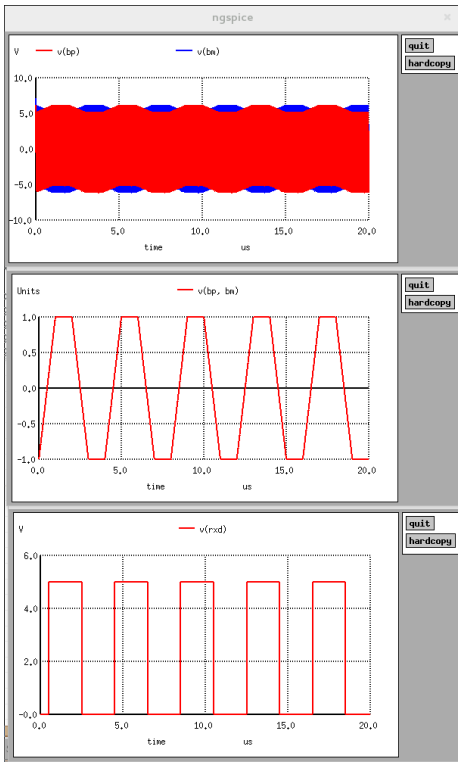


Figure 8.235: Signals on the bus lines, differential signal v(vp)-v(vm) and the received signal at pin rxd

Emission minimization: Besides the requirement of being immune to noise coupled into the system most differential data transmission systems additionally have the requirement of low RF emission (EME). Especially in automotive environment the emission requirements are extremely strict because it has become common to use antennas in the wind shield of the car. Typical coupling between cables in the dashboard and the windshield antenna are in the range of -25dB. Since the sensitivity of a typical FM radio is in the range of 1μV (0dBμV) most car vendors require bus systems with common mode signals on the wire in the range of less than 22dBμV in the FM band. Since about 2015 some car manufacturers request about 10dBμV in the FM band. This requirement mainly applies to CAN, flexray and ethernet PHYs for automotive applications.

Typical matchings achievable on integrated circuits for the voltage sources Vp and Vm is in the range of 0.5% at the nominal data rate. Due to dynamic effects (mismatch of delay times) this matching decreases with increasing frequency. Thus higher harmonics are suppressed less. As on consequence most differential data transfer systems have common mode chokes to suppress the common mode signal provided by the transmitter.

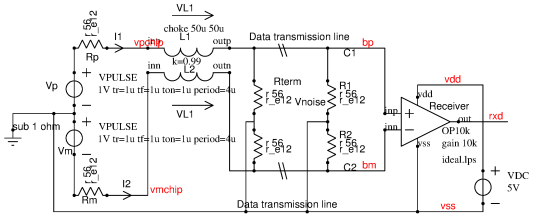


Figure 8.236: Emission reduction using a common mode choke

Common mode signals present at the pins Vpchip and Vmchip are suppressed by the common mode choke. Typical chokes available on the market for CAN and flexray applications have coupling factors of abpout K=0.995. Introducing a choke the bus at vp and vm MUST be terminated with a resistor. This resistor is chosen equal with the impedance of the bus lines (usually these are cables of several meters of length). The cable must be terminated at both ends. At least one of the terminations must be tapped and connected to a DC voltage (otherwise we get an undefined common mode DC input voltage overdriving the receiver!)

Usually the manufacturers of the chokes provide data of the inductance of L1 and L2 and the coupling. The inductive voltage drop at inductor L1 is:

$$V_{L1} = \frac{dI_1}{dt} * L_1 - \frac{dI_2}{dt} * L_2 * K \tag{8.310}$$

in the same way the voltage drop at L2 calculates as:

$$V_{L2} = \frac{dI_2}{dt} * L_1 - \frac{dI_1}{dt} * L_1 * K \tag{8.311}$$

Assuming both currents are equal

$$I_1 = I_2 = I$$

we find:

$$V_{L1} = \frac{dI}{dt} * (L_1 - K * L_2) \tag{8.312}$$

$$V_{L2} = \frac{dI}{dt} * (L_2 - K * L_1) \tag{8.313}$$

The resulting differential mode inductances become:

$$L_{d1} = L_1 - K * L_2 \tag{8.314}$$

$$L_{d2} = L_2 - K * L_1 \tag{8.315}$$

The bandwidth limitation of the choke is:

$$f_{gdiff} = \frac{R_{term}}{2 * \pi * (L_{d1} + L_{d2})} \tag{8.316}$$

Example: Rterm=56Ω, K=0.99, $L_1 = L_2 = 50\mu H$

$$L_{d1} = L_{d2} = 50\mu H * (1 - 0.99) = 0.5\mu H$$

$$f_{gdiff} = \frac{56\Omega}{2 * \pi * 1\mu H} = 8.9MHz$$

Common mode chokes only work well as long as L1 and L2 are equal. Mismatch of L1 and L2 will lead to differing voltages VL1 and VL2. This means a poor choke will convert differential mode signals into common mode. Whether this differential to common mode conversion becomes visible on the chip side or on the bus side of the choke depends on the source impedance of the bus driver. In case of a driver significantly lower resistive than the termination resistors on the bus side the common mode produced by a choke with missmatch will be visible on the bus side.

If the driver is a current source (high impedance) the common mode produced by a choke with poor matching will show up on the chip side.

Unfortunatelly the impedance of a bus driver is frequency dependent (pin capacities!).

We will investigate these non linearities discussing CAN and flexray systems in the following sections.

RF Sensitivity reduction using a common mode choke: The common mode chocke also rejects RF comming from the bus. So besides being a good emission filter a common mode chocke also can improve the sustem ruggedness against RF injected from outside. The AC transfer function depends on the common mode inductance, the pin capacity of the bus driver to be protected and the impedance of the common mode noise source. The setup typically looks as shown in the following figure.

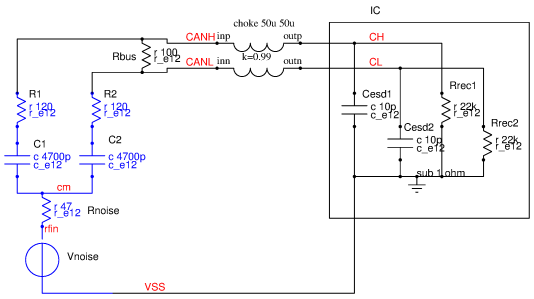


Figure 8.237: Typical immunity test setup with common mode choke used for CAN transceivers and flexray transceivers

The blue components belong to the test RF generator and the coupling network required for the test. These components are not present in the application of the IC. In application the RF can get picked up by the wires acting as an antenna. Assuming the wires CANH and CANL are exposed to the same electromagnetic field the RF coupled into the system during the test is applied as a common mode signal with equal values for R1 and R2 as well as C1 and C2.

To verify the AC transfer function the source Vnoise usually is simply set to 0dBV. The AC transfer function to CH and CL is shown in the following plot:

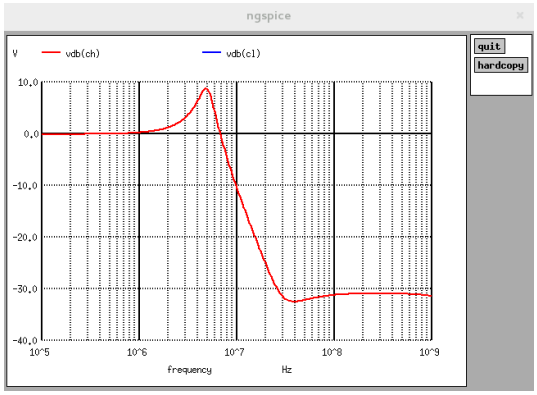


Figure 8.238: AC transfer function from the RF source to nets CH and CL

The transfer function has a resonance at 5MHz. This resonance is caused by the common mode inductance of the choke together with the pin capacities of the IC. The resonant frequency calculates as:

$$f_{res} = \frac{1}{2 * \pi * \sqrt{L * (Cesd1 + Cesd2)}} \tag{8.317}$$

The square root in the denominator holds the sum of the capacities of the chip. The reason is the tight coupling of the windings of the choke that transforms the capacities ot the other side. So both capacities becomes visible on both sides. If the coupling of the windings of the choke is reduced to K=0.1 the resonance shifts up to about 7.4MHz.

At high frequencies the attenuation of the common mode choke is limited by the parasitic stray capacity (300fF) bypassing the inductance. This limits the common mode rejection to -30dB.

The resonant peak at 5MHz can cecome a problem for the bus receiver IC if the amplitude exceeds the common mode range of the receiver. Therefore most CAN systems are most RF sensitive close to the resonant frequency of the common mode choke together with the pin capacities.

A second limiting factor is the symmetry of the choke. Usually the two inductors will not match perfectly well. This mismatch converts part of the common mode signal into a differential mode signal that will be detected by the bus receiver. In the following example the two common mode inductances are 51μH and 49μH in stead of both inductors being 50μH. As a consequence about 4% of the common mode signal is getting converted into a differential mode signal from about 5MHz to about 100MHz. Above 100MHz the stray capacities (that are matched again) dominate over the effect of the choke.

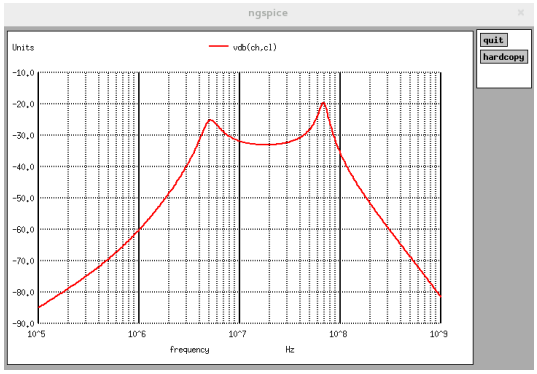


Figure 8.239: Conversion of a common mode signal into a differential mode signal by an asymmetrical choke having 4% mismatch between the windings

On a first glance a common mode to differential mode conversion with -25dB may look negligible. But in most automotive systems we have to expect common mode disturbance in the range of +-30V. At these levels a conversion with -25dB leads to differential signals on the other side of the choke that will be detected by the bus receiver! These very fundamental simulation based on ideal components already show that for building a good differential bus system not only the IC must be designed almost perfectly symmetrical. The symmetry requirement also applies to the common mode choke and the impedances on both bus wires! It even may become advisable to omit using a poor choke if the IC offers better symmetry than the common mode choke.

Emission reduction using digital wave shaping: Modern technologies offer increasing digitap performance that can be used for digital wave shaping of the BUS signals. The basic concept is to use a DAC (digital to analog converter) to produce a signal shape with smoothened edges. The target of the rounded edges is to make the spectrum of the signal decay faster than if a simple linear edge is used.

The price of this approach is the emission of a sampling frequency. The signal comming from the DAC approximates the desired signal by a staircase function. Assuming an unlimited fast bus driver stage the sawtooth like error would propagate to the BUS wires. Real implementations use the limited speed of the bus driver stage as an anti aliasing filter.

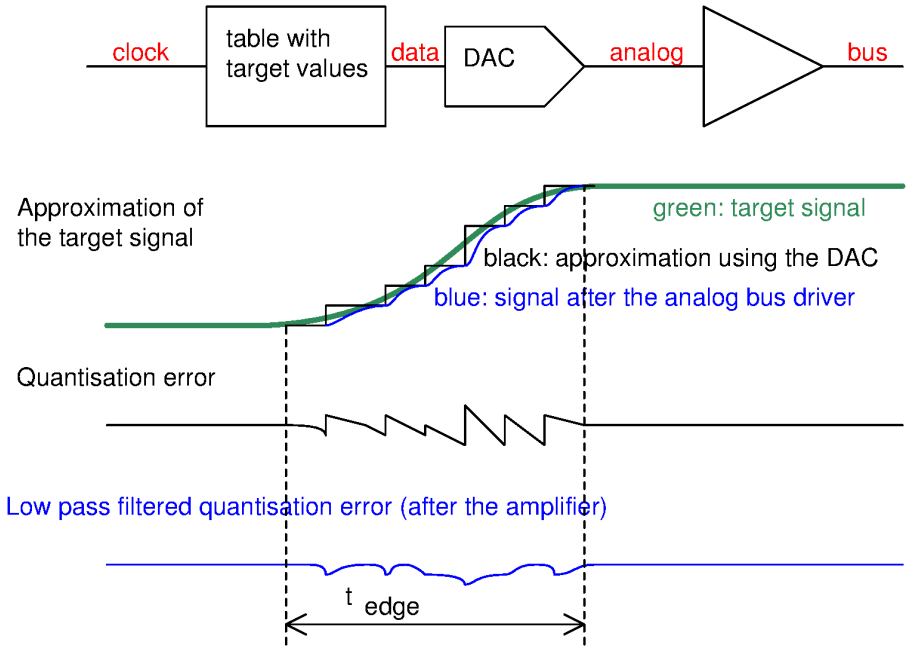


Figure 8.240: Using a DAC to create a rounded edge

The clock rate should be chosen as high as possible to move the undesireable carrier far away from the intentional bus signal. Often gate propagation delays are used to create a time quantisation is the ns range or even less. The sampling points usually are equidistant but the step size usually depends on the speed of the edge. So quantisation errors become a maximum in the fast part of the edge while at HIGH or LOW state there is no quantisation noise.

The amplitude of the sampling frequency follows the amplitude of the saw tooth signal. The amplitude of the n-th harmonic of the saw tooth signal calculates as:

$$A_n = A_{saw} * \frac{2}{n * \pi} \tag{8.318}$$

Since the saw tooth signal only is present during the rise and fall of the signal but not while the signal is constantly high or low the average quantisation noise at the n-th harmonic of the sampling frequency becomes:

$$A_{neff} = A_n * \frac{t_{edge}}{t_{bit}} = A_{saw} * \frac{t_{edge} * 2}{t_{bit} * n * \pi} \tag{8.319}$$

This signal is present at the input of the analog amplifier. Most analog amplifiers can be regarded as a first order or second order low pass filter attenuating the sampling noise.

There is one very interresting case to consider:

If the slew rate of the amplifier exactly matches the speed of the edge the quantisation noise during the fast part of the transistion disappears! (In this case the amplifier creates a constant delay but the blue curve simply is running in parallel with the blue target signal.) This can be used to minimize the quantisation noise.

8.13.5 CAN

There are two different flavours of CAN bus systems. Low speed CAN is used up to 125kbit/s. High speed CAN is used up to 1Mbit/s. The basic concept of both transmitters is the same. The transmitter consists of a switch to ground and a second switch to 5V. The termination of both systems differs. Low speed CAN is terminated to the supplies at every node with a termination that does not necessarily correspond the bus line impedance. High speed CAN has a termination resistor at the ends of the wires. The termination of a high speed CAN corresponds the wire impedance.

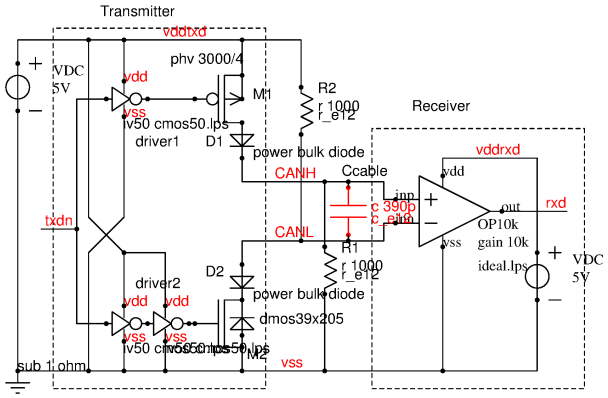


Figure 8.241: Basic concept of a low speed CAN. Ccable represents the capacity of the cable

The CAN bus has two possible states: If M1 and M2 are ON the CAN bus is in the so called dominant state. CANH is pulled to about 4V while CANL is pulled to about 1V. If M1 and M2 are turned off CANH is pulled to ground by R1 and CANL is pulled to vddtxd by R2. This is the so called recessive state. The following figure shows the resulting signals on the bus wires.

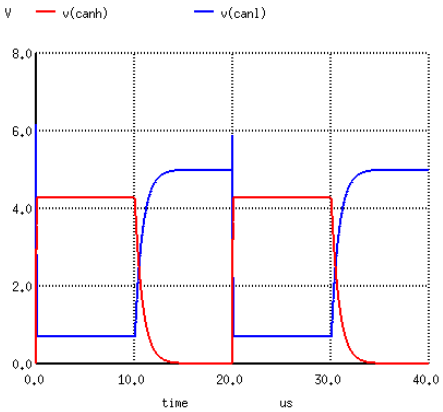


Figure 8.242: Eye diagram of a low speed CAN

High speed CAN uses the same driver concept but the termination is from CANH to CANL at both ends of the cable.

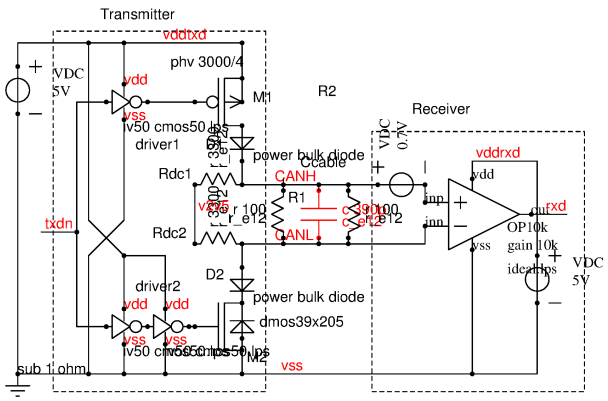


Figure 8.243: Basic concept of a low speed CAN. Ccable represents the cable capacity

The receiver of a high speed CAN must have a designed offset. Usually this offset is in the range of 0.7V+- some hundred mV of hysteresis. In recessive state the bus is floating. To give the bus a defined DC operating point Rdc1 and Rdc2 act as weak pulls to a mid voltage of typically 2.5V during recessive state.

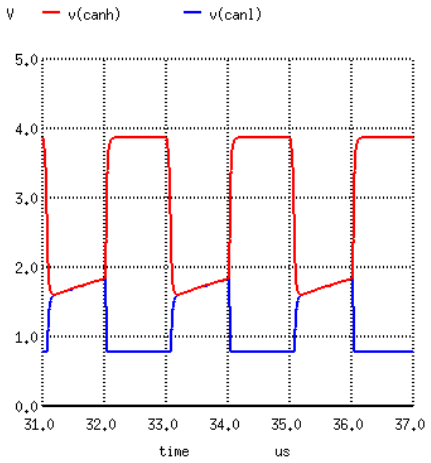


Figure 8.244: Eye diagram of a poor designed high speed CAN

The eye diagram of the high speed CAN shown above already shows one of the most difficult design issues of the high speed CAN concept. During the recessive state the bus is almost floating. As a consequence the transistor turning off later (In this case the DMOS transistor) pulls the DC level of the bus away from the intended common mode level of 2.5V. The strength of the switches and the timing must be balanced extremely well. The simplified example shown here would not fulfill the emission requirements of today's CAN transceivers. It only is used to illustrate the basic concept.

More sophisticated concepts try to better match timings as well as currents, impedances and slopes of the CANH driver and the CANL driver. A very common concept is using the slow high voltage transistors as protections only. For switching the smaller and faster low voltage transistors are used. The timing is controlled by the delay line that can be trimmed in stead of the slow high voltage components now. To prevent clock feeed through the switched (time discrete!) drive signals are low pass filtered by Rh1..Rh4 and Rl1..Rl4 together with the gate capacities of the switching transistors. The time constant chosen typically is 3 times the granularity of the delay line.

One possible approach to match currents is to use current generators referring to the same reference current. To match the slopes the power transistor is segmented and each segment is controlled by a tunable delay line.

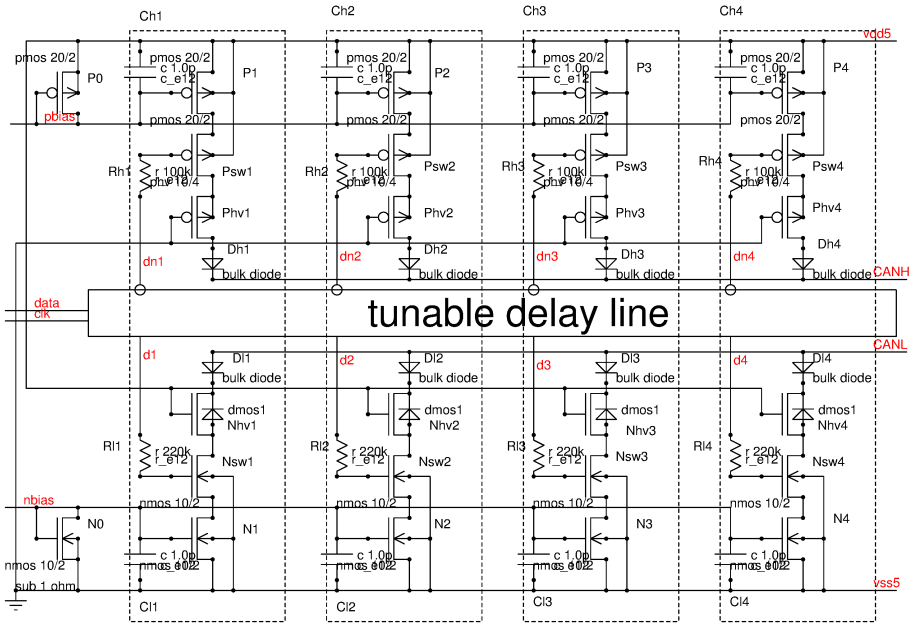


Figure 8.245: Concept of a CAN transmitter with segmented driver, current generator and fast switching low voltage transistors

Practical implementations often use up to 20 segments or more to shape the switching edges. Here only 4 segments are shown to illustrate the concept.

Practical considerations: To shape the switching slopes with rounded edges (Faster roll off of the harmonics) the channel differ in size (current). As a conequence the delay line outputs have to drive different load capacities. Additionally to the different transistor sized mainly at small stages (range where the edge rounding is done) the wiring

capacity has a significant influence. To solve these influences of load capacities on the delay time the outputs of the delay line should be buffered.

Delay lines driving the NMOS side and the PMOS side of the transmitter must track each other. Using synchronization latches similar to clock trees is strongly recommended.

To make the bias generators rugged against RF injected into the nets CANH and CANL the gates of the bias generators are blocked with Ch1..Ch2 and Cl1..Cl4. Typically these capacitors are at leas one magnitude bigger than the miller capacities of the current generator transistors.

In the following plot the eye opening of the bus is shown under two operating conditions. The upper trace shows nominal operation at a common mode voltage of 2.5V. The lower trace shows the operation at a common mode voltage of 10V.

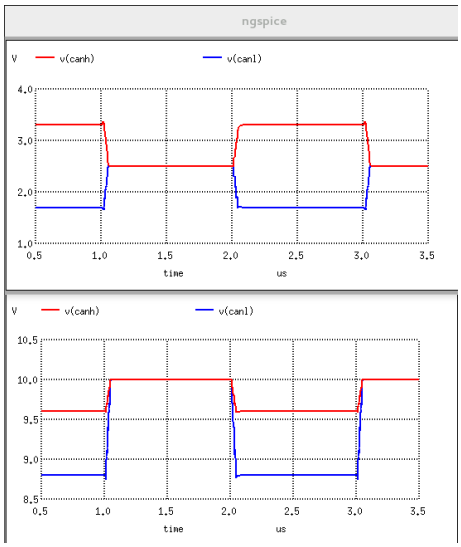


Figure 8.246: Operation of the CAN transmitter at a common mode voltage of 2.5V and 10V

Obviously the eye opening of the current source driven CAN transmitter changes with the common mode voltage. The reason simply is that pulling the common mode voltage beyond the supply rails either turns of the high side diodes (Dh1..Dh4) or the low side diodes (Dl1..Dl4).

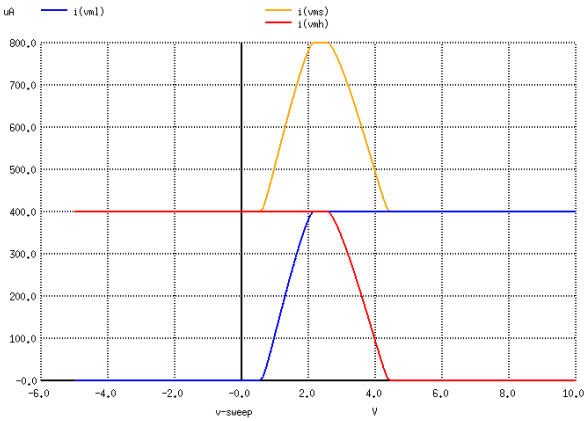


Figure 8.247: Currents found sweeping the common mode voltage from -5V to 10V

The sum of both currents (i(vms) is the current opening the eye. This current reaches a maximum at a common mode voltage of 2.5V. Ideally i(vms) should be constant from -40V to 40V for a CAN driver stage.

To counteract this effect the currents must be made voltage dependent! There are several solutions to achieve an increase of the currents moving the common mode voltage away from the mid point. Ideally the sum of the currents flowing in N1..N4 and P1..P4 should be constant sweeping the common mode voltage.

1. The current generators (N1..N4 and P1..P4) could be operated in triode region intentionally. Designing such a current mirror that performs well throughout the whole temperature range and production spread is extremely cumbersome and technology dependent. Since I have not yet seen a successful implementation of this concept we will not follow this idea any further.

2. Detection when the diodes turn off. This can be done monitoring the voltage at the current generators. If one of the current generator leaves saturation the opposit side will be driven with an increasing current to

compensate the loss. This however is an active regulation loop influenced by the commn mode disturbance. Such a regulation loop opens the door for disturbance by RF.

3. Using resistors in stead of current generators. There the problem is that the high voltage transistors Nhv and Phv have threshold voltages that do not match. So the stage becomes asymmetrical.

4. Use a hybrid of the current source driven stages and the resistor driven stages. The resistor provides the increase of current moving away from the 2.5V common mode voltage while the current generators can be used to regulate the symmetry. The regulation loop in this case is independent of the output and not exposed to direct RF injection.

CAN specific RF emission problems: CAN transceivers have some very specific EMC problems that are not common in other differential bus systems. These result from the transition into a high impedance changing from dominant state into recessive state. The CAN transceiver has a certain output capacity in the range of 10pF to 30pF. If the CAN transmitter turns off (change to recessive state) this capacity in combination with an inductive load will ring. Most users of CAN drivers assume that this ringing will be suppressed by the common mode choke. Unfortunately the contrary is true. Ideally the common mode choke should be symmetrical. Both sides should have the same common mode inductance. Chokes found on the market (e.g. EPCOS chockes) show asymmetries of the inductances of about 0.3%.

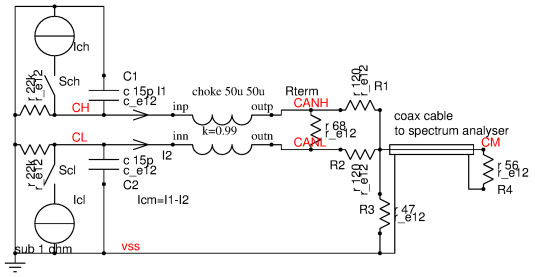


Figure 8.248: Simplified test setup for direct coupled RF emsission (DC decoupling capacitors were omitted). (The two 22K resistors represent the impedance of the receiver attenuator)

During dominant stae the (ideal) CAN transmitter forces equal currents in both coils of the choke. In case of differing inductances $L_1$and L_2 this builds up a magnetic field in the choke. (In a perfect choke the field of the CANH and CANL current would perfectly cancel). Turning off the tight coupling of the coils makes the choke act as a current transformer. The ratio of the currents in both winding becomes:

$$\frac{I_1}{-I_2} = \sqrt{\frac{L_2}{L_1}} \tag{8.320}$$

Things are getting more interresting expressing $I_1$by I_2 and summing the two currents. $I_{cm}$is the common mode current flowing into C1 and C2 after opening the switches.

$$I_{cm} = I_2 * (\sqrt{\frac{L_2}{L_1}} - 1) \tag{8.321}$$

Immediatelly after turning off the current $I_2$is almost equal to the current that has been forced by Ich and Icl during dominant state (assuming the mismatch of the choke is in the range of a few percent only). Thus the common mode current provided by the choke is simbly:

$$I_{cm} \approx I_{dom} * (\sqrt{\frac{L_2}{L_1}} - 1) \tag{8.322}$$

The initial first peak of the resulting ringing at the spectrum analyser thus becomes:

$$V_{cmpeak} = 25\Omega * I_{dom} * (\sqrt{\frac{L_2}{L_1}} - 1) \tag{8.323}$$

The resulting frequency of the common mode ringing of the CAN transmitter is:

$$f_{cmres} = \frac{1}{2 * \pi * \sqrt{L_{cm} * (C_1 + C_2)}} \tag{8.324}$$

Besides the common mode ringing we also will observe a differential mode ringing of the CAN transmitter at:

$$f_{dmres} = \frac{1}{2 * \pi * \sqrt{L_{cm} * (1 - K) * \frac{C_1 * C_2}{C_1 + C_2}}} \tag{8.325}$$

For better visibility of the common mode ringing in simulation K can be chosen K=1. (Well, this component doesn't exist but in simulation we can calculate with such an ideal thing).

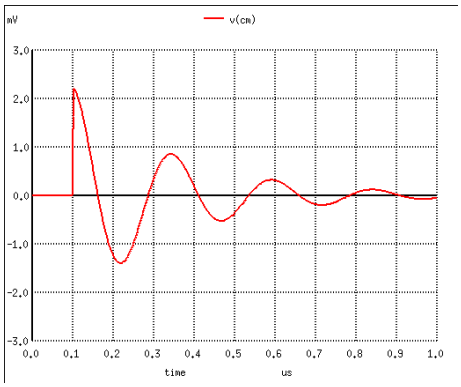


Figure 8.249: Simulated choke ringing assuming K=1, $L_1 = 50.15\mu H$ $L_2 = 49.85\mu H$

To get a more realistic picture of what happens let us set K=0.99 and rerun the simulation.

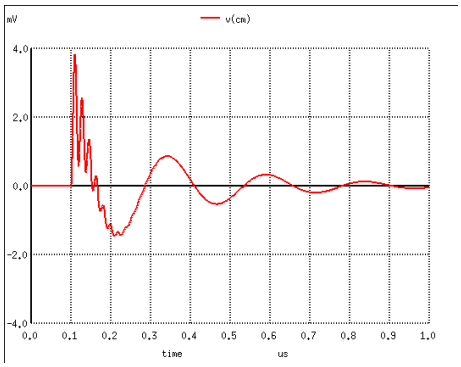


Figure 8.250: Same simulation as before but with K=0.99 and resulting differentiol mode ringing

The differential mode ringing becomes clearer looking at nodes CH and CL.

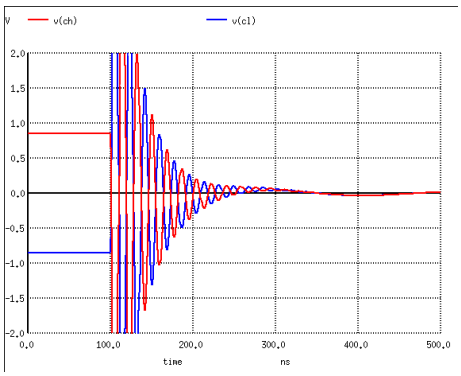


Figure 8.251: Ringing on the IC side of the choke

The above figure shows that in the beginning there mainly is a differential mode ringing. The differential mode ringing gets converted into a common mode ringing on the other side of the choke by the asymmetry of the choke. After decay of the differential mode ringing (after about 300ns) the slower common mode ringing dominates.

Since these simulations used a model of the CAN transmitter representing an ideal transmitter without any mismatch these simulations show that this ringing is caused by the choke only!

8.13.6 Flexray

The flexray specification defines the differential output voltage, the differential load impedance and the timings of a flexray bus system. It does not make anry recommendations if a flexray transceiver is built using current sources ot voltage sources. Transceivers can be connected to the bus anywhere along the bus.

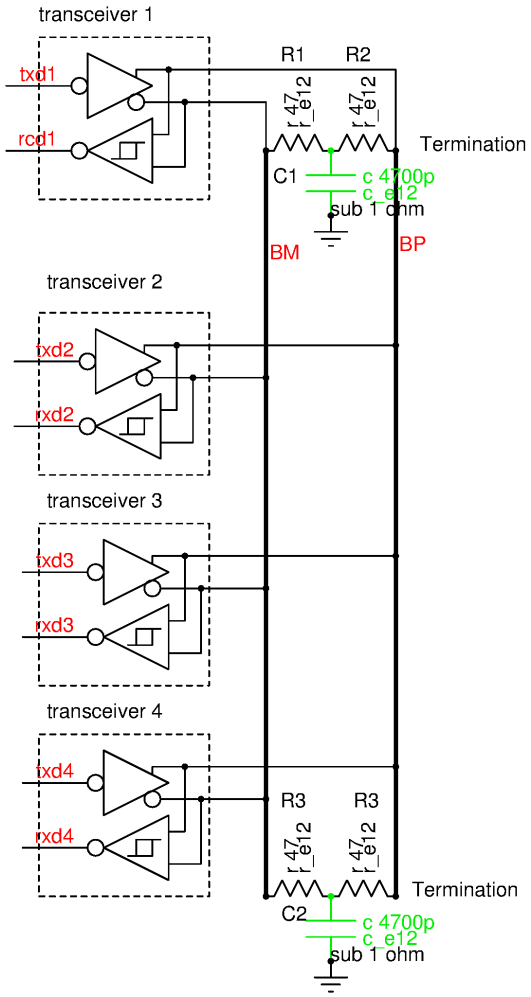


Figure 8.252: Concept of a flexray system

The wires BP and BM are the positive node and the negative node of the bus. R1 to R4 terminate the the bus line. The differential impedance is 100Ω to 110Ω. So R1 to R4 can range from 50Ω to 55Ω.

For RF immunity and RF emission the system assumption is that BP and BM both carry RF but at opposit polarity. The common mode termination C1 and C2 is optional. Thus the DC operating point of the bus wires is defined by the input attenuator of the receiver comparator.

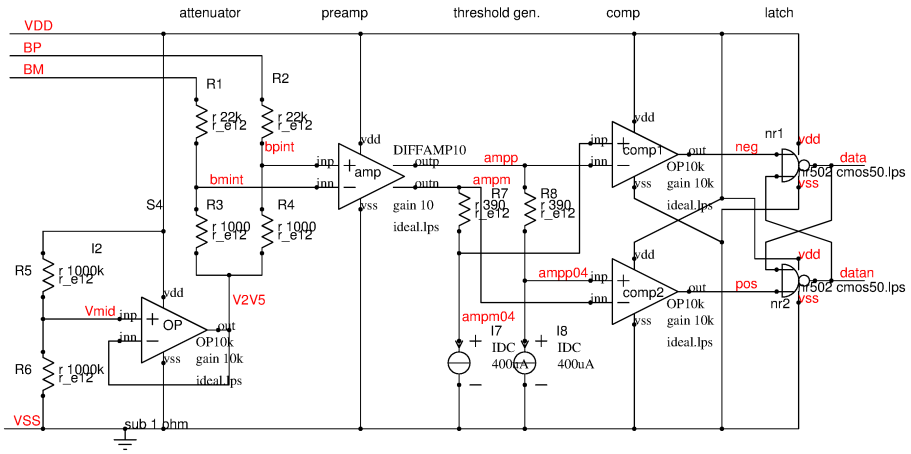


Figure 8.253: Concept of a flexray receiver

R1 to R4 pull the DC voltage of the bus to 2.5V. To make the foot of the attenuator low resistive the attenuator is connected to the output of OP that buffers the internal 2.5V reference Vmid. Vmid must be a low impedance node for the complete frequency band from 0 to about 1GHz. So additional capacitive buffering is recommended. The 2.5V buffer is one of the most critical parts for the EMC performance of a flexray receiver.

In some cases the amplifier buffering the node V2V5 even is designed to compensate the common mode signal (See chapter instrumentation amplifiers). This trick requires extreme caution because approaching the cut off frequency of

the amplifier the phase shift turns the compensation into a common mode boost! The common mode boost mostly happening between 200MHz and 400MHz often leads to an EMC problem of the receiver at 200MHz to 400MHz.

Current generators I1 and I2 together with R1 and R2 define the hysteresis of the receiver.

The first transmitter designs used current sources together with switches to drive the bus wires BP and BM. Since the specification of the flexray system requires the bus to operate even at DC offset voltages of +-15V and does not permit reverse supply of VDD from the bus lines, diodes have to be added to the transmitter stage. The transmitter consists of multiple stages that are switched with a delay line to achieve defined switching slopes.

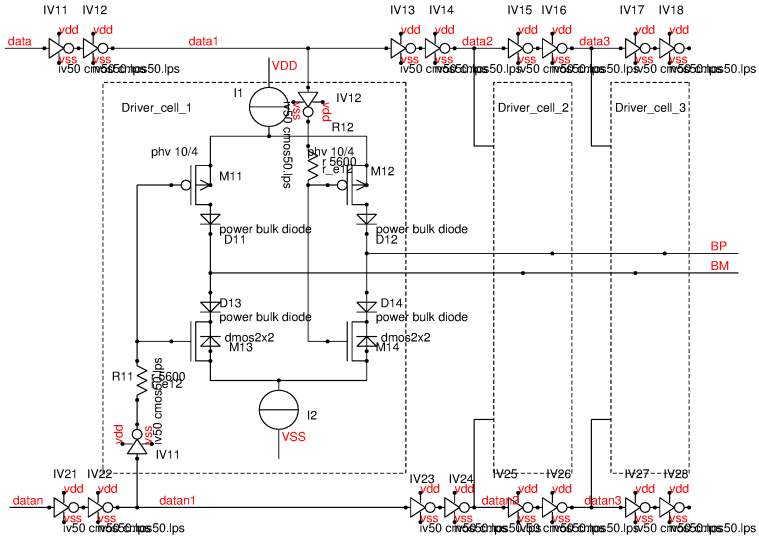


Figure 8.254: Concept of a flexray transmitter

to minimize common mode signals on the bus the driver signals data and datan must have exactly opposit phase. Even delay errors in the range of 100ps are already visible as a common mode signal on the bus. Current sources I1 and I2 also must be matched to prevent common mode signals. If bipolar diodes are used as shown here they may not have losses to substrate (parasitic vertical PNPs) because these losses lead to different currents pulling up the bus line or pulling down the bus line.

The resistors R11 and R12 have two motivations: They smoothen the switching of the segment of the flexray transmitter and they reduce feed through of fast logic edges comming from the delay line. (even the input edge of IV11 and IV12 can capacitively propagate to the output of IV11, IV12 although IV11 and IV12 are designed with long channel transistors!) Building inverters IV11 and IV12 with resistors in the supply paths is by far less performant than placing the resistors between IV11, IV12 and the gates of the power transistors M11 to M14. If resistors are to area consuming a starved current inverter type 2 should be considered (current starving transistor in the middle of the inverter).

The gates of M11, M13 and M12, M14 intentionally are connected and not driven from seperate inverters. The reason is that current sources I1 and I2 are not tolerant versus current interruptions (These lead to fast changes of the voltage across the current source and modulates the currents via the miller capacities. In stead there intentionally is a make before break timing in flexray drivers.)

If seperate drive stages for te PMOS side and the NMOS side are needed for wave shaping reasons care must be taken that the make before break condition is not violated.

If the timing for a make before break operation can't be guaranteed the current generators can be kept in their desired operating range by a tail clamp (source follower holding the tail at a certain minimum voltage).

To keep the eye opening even at high comon mode voltages on the bus the current generators can be operated in triode region on purpose. (So if the common mode of the bus is pulled to ground the increase of the current on the PMOS side partly compensates tha the NMOS side can not pull down anymore. If the common mode of the bus is pulled to 5V the current on the NMOS side increases and keeps the eye open)

Tail capacity of the sources of the switches: Even with make before break timing the sources of M13, M14 and M11, M12 move. This leads to currents flowing into the drain capacity of the current sources I1 and I2. To minimize the capacitive currents these tail nodes must be layouted extremely compact. One possible solution is to use low voltage MOS transistors as switches and stack them with high voltage cascodes. (In this case the cascodes are biased at the 5V supply rail and at the 0V ground rail. So the cascodes are only voltage limiters to protect the low voltage transistors)

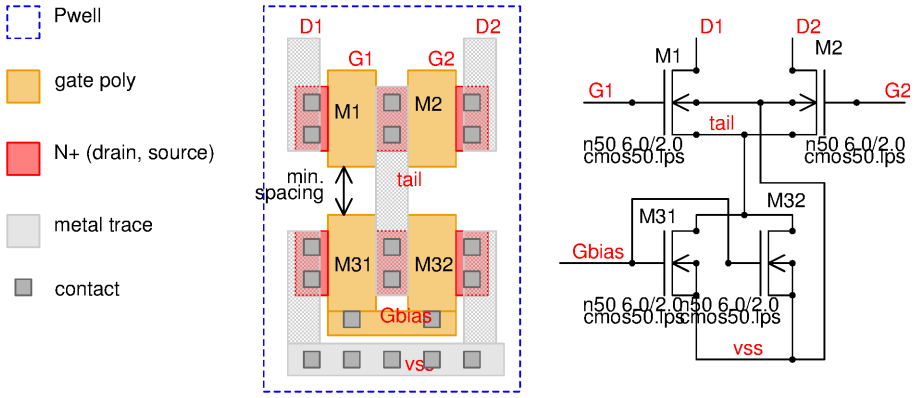


Figure 8.255: Compact layout to minimize the tail capacity of the current switch

RF emission considerations: The wires BP and BM are assumed to radiate RF with the same characteristics. To measure the RF emissions of a flexray system usually the same setup is used as for CAN systems.

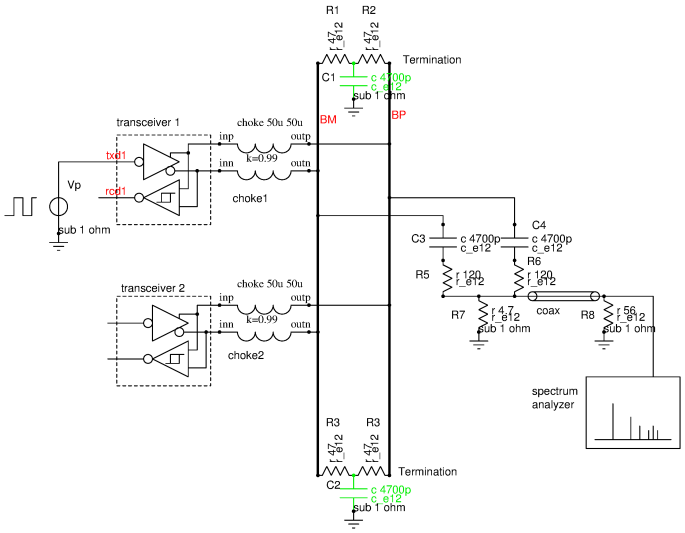


Figure 8.256: Flexray common mode emission test setup

This setup uses exactly the same components as recommended for the application of a flexray transceiver. The bus wires are replaced by symmetrical strip lines on a test board with a length in the range of some cm. The divider R5, R6, R7 and the transformed input impedance R8 of the spectrum analyzer provides a signal proportional to the common mode signal of the bus system. Typically one transmitter is operating while the second transceiver is in receive mode. The second transceiver is used to monitor the signal on the bus lines. Ideally if everything is perfectly symmetrical the expectation is to find no more common mode signal at the spectrum analyser. Depending on the asymmetries we can make the following observations:

Case 1: The transmitter is ideally symmetrical but the chokes have a sysmmetry error: Since the signals at the chip are assumed to be perfect we will not observe any common mode signal operating this hypothetical ideal chip without a choke. Operating the ideal transmitter with the non ideal choke we will observe that the non ideal choke converts a part of the differential signal into a common mode signal. The spectrum analyzer will measure the attenuated spectrum of the differential signal.

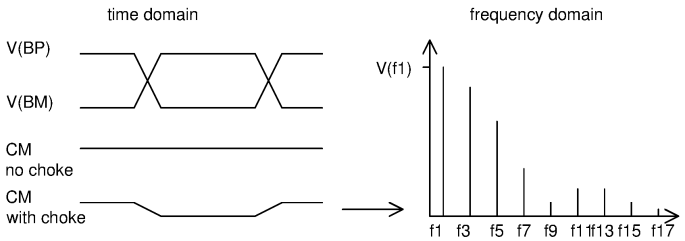


Figure 8.257: RF emission of an ideal transmitter with an asymmetrical choke

Since the common mode signal is a replica of the differential mode signal the spectrum holds the odd harmonics of the 5MHz rectangular pulses. The amplitude of the fundamental (f1=5MHz) is defined by the choke asymmetry and the attenuation of the test network R5..R8.

Case 2: the transmitter has an amplitude error: If one of the channels BM or BP has a different amplitude than the other a common mode signal can be observed at the output of the chip. In case of a current controlled transmitter the choke is less effective. It will attenuate the signal by about 5dB to 10dB only because the current source will simply force a higher voltage at the IC pins until the current will flow. The signal on the bus typically is a low pass filtered copy of the error. The error has the same period as the payload signal. If the pulse train used to excite the system has a duty cycle of exactly 50% the resulting spectrum consists of odd harmonics only. At the chip the spectrum rolls off like the spectrum of the differential mode signal. On the bus lines the roll off starts at the cut off frequency of the choke's common mode inductance in combination with the bus common mode impedance.

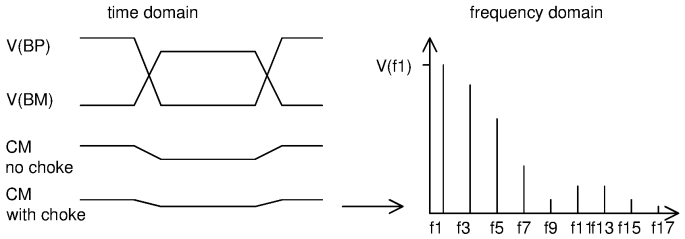


Figure 8.258: RF emission of a transmitter with an amplitude error

Most common reason for amplitude errors of a flexray transmitter is a non ideal current source in combination with deviating resistances of the switches.

Looking into practical applications the matching of the resistors terminating the bus lines or the impedance of the cables (different coupling of BP to ground and BM to ground) must be taken into consideration as well. If the cable or the termination resistors are the limiting factor even a perfect chip will not solve the problem and we MUST rely on the choke to filter the common mode signal.

Case 3: the two flexray channels have a different delay: If the delay of the channels BP and BM differs the resulting time domain signal and the frequency domain signal looks like the following plot.

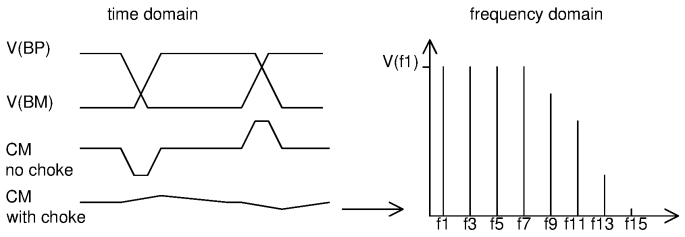


Figure 8.259: RF emission of a flexray transceiver with a different propagation delay of BP and BP

The common mode signal has the same period as the payload signal. As a consequence the odd harmonics are present. Since the common mode pulses have a short duration the spectrum has a high bandwidth. It start to roll off at about

$$f_{g1} = \frac{1}{\pi * t} \tag{8.326}$$

with t being the duration of the pulse. Since the pulses are short the total power of the lines is fairly low. So this contributor to the RF emission usually is visible above several 10 MHz.

The pulses occurre at the moment the bus is switching. During that short time the bus nodes are fairly low resistive due to the miller capacity of the switches and due to the capacities of the ESD protections. A high resistive choke will not lead to a significantly higher voltage on the chip side. So these pulses usually are nicely attenuated by the common mode choke with an attenuation of about 30dB to 40dB.

Case 4: The signals at BM and BP differ in duty cycle In this case the observed common mode signal has double the repetition rate than the bus signal itself. This leads to a spectrum mainly consisting of the even harmonics. The common mode pulses have a similar duration as those produced by duty cycle errors. Thus the observed spectral lines have a low power but a high bandwidth. Furthermore these pulses are driven low resistively for the same reasons as the duty cycle error pulses. So the common mode choke is an effective filter.

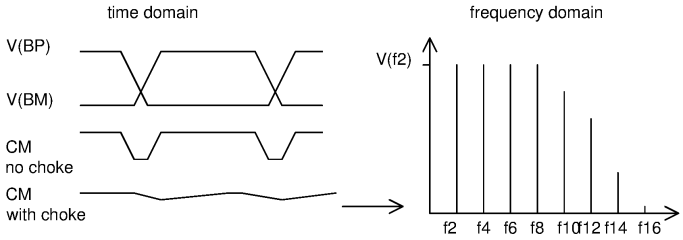


Figure 8.260: RF emission of a flexray transmitter with a duty cycle error

So after all this theory how does a real spectrum look? It can be composed with the information we have! First the spectrum without choke:

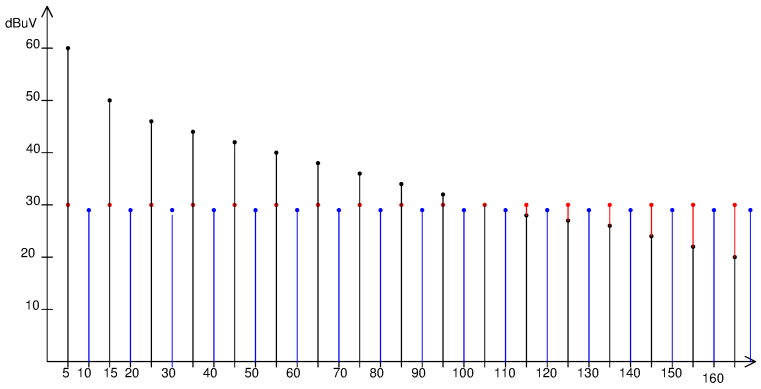


Figure 8.261: Spectrum composed of the amplitude error, delay error and the duty cycle error.

In the spectrum shown above the amplitude error is black. Since the signal at BP and BM is more or less trapezoid in decays with abozr -20dB/decade to about 40MHz and with -40dB/decade above 40MHz.

The contribution of the daly difference between BP and BP is in red color. It is almost flat until 160MHz. Since it has a period of 200ns (5MHz) it holds odd harmonics.

The contribution of the duty cycle error has a period of 100ns (10MHz) which leads to even harmonics. These are colored in blue.

Introducing the common mode choke we expect a reduction of the emission according to the common mode suppression of the choke. The resulting spectrum using an ideal choke with 25dB suppression above 30MHz is shown below. It looks quite nicely and as if the system fulfills the usual requirements.

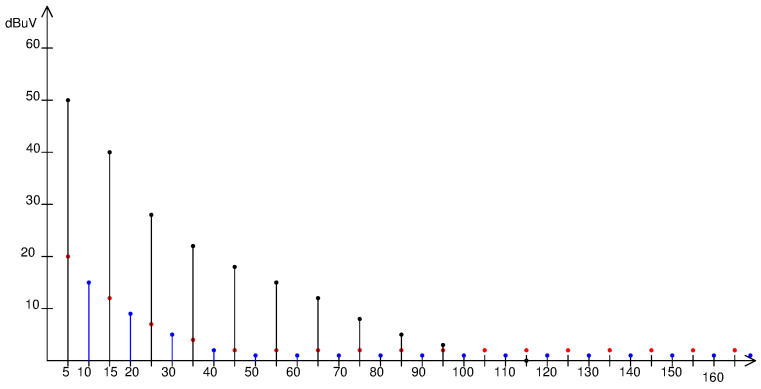


Figure 8.262: Theoretical spectrum assuming we use a symmetrical choke with about 25dB common mode rejection above 25MHz

Unfortunatelly the choke also has a matching error between the two coils This matching error converts some of the differential mode signal into common mode. The differential mode signal is about 120dBμV at 5MHz and rolling off with -20dB/decade up to about 40MHz. The typical attenuator used for the common mode measuremnt has -15dB. Assuming an asymmetry of the choke of 1% we expect a differential mode to common mode conversion with -40dB. Thus at 5MHz the converted differential mode becomes visible with 65dBμV. The converted differential mode almost only has odd harmonics. In the following figure the converted differential mode is represented by the green circles.

https://de.wikipedia.org/wiki/Thermoelektrische_Spannungsreihe

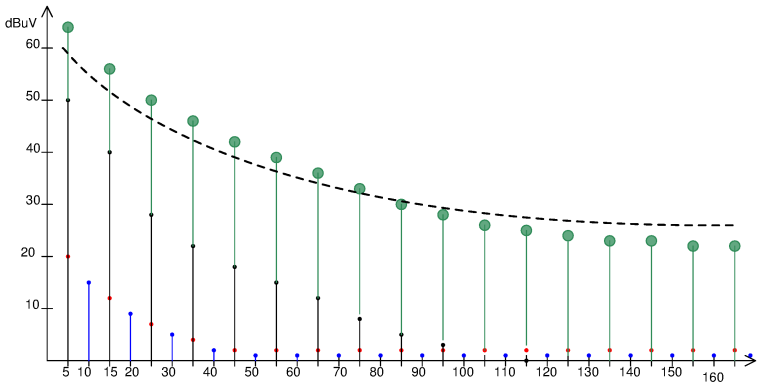


Figure 8.263: The green circles represent the effect of the differential mode to common mode conversion of a choke with 1% mismatch

In the figure shown above the dashed line represents the envelope of the spectrum without choke we had before. This creates the impression that the choke only rejects the even harmonics. The truth is that the choke rejects both, odd and even harmonics, but also superseeds the rejection by the common mode to differential mode conversion of the choke itself!

In other words a choke with more than 0.5% mismatch will lead to RF emissions higher than the G5 requirements even if the transmitter were perfect!

8.13.7 USB

8.13.8 Ethernet PHYs

8.13.9 Test IOs

8.14 ESD protection

ESD protection is not just a single circuit. Usually ESD protection is a complete concept including the pin inductance and the circuit to be protected. Designing an ESD protection first requires the answer to the following questions:

1. What is the fastest destruction mechanisme the ESD protection has to protect against?
2. What is the pin inductance limiting the current?
3. Where does the current flow?
4. Is a multi stage protection consisting of primary protection and a secondary protection possible?

8.14.1 Destruction mechanisms to protect against

The following table lists typical destruction mechanisms found in integrated circuit design and the time constants to be considered.

Table 48: Comparison of ESD destruction mechanisms

problem	mechanism	response (typ)	critical voltage, current	cumulative
gate protection	oxide rupture	voltage dependent	$V_{break} \approx \frac{1V*t_{ox}}{nm}$	yes
Base-Emitter break down	hot carrier	some μs to ms	some V	yes
drain bulk break down	hot carrier	some μs to ms	some V	yes
junction break down	thermal	100ns to 10μs	$V_{break} \approx 25..50V/\mu m$	no
poly resistors	grain melt	100ns		yes
diffused resistors	contact melt	some μs		no
metal resistors	thermal	200ns	0.200A*100ns $/\mu m^2$	no
latch up	thermal	1..10μs	time to trigger	no

For cumulative destruction mechanisms the total stress time must be summed up. Non cumulative events don't have any storage effect.

Example: If a detruction is thermal we either reach melt down or the device can stand the same stress multiple times provided between the pulses there is enough time to coold down again.

Gate protection: Gate break down depends on oxide thickness, homogenity and quality. Thermal oxides have (gate oxides) have been investigated thoroughly and are described well in literature. CVD oxides (chemical vapour deposition) found between metal layers have lower break down field strength and are more process dependent.

For gate oxides Reza Moazzami, Jack Lee, Ih-Chin Chen and Chenming Hu publisched [19]the equation

$$t_{br} = t_0 * e^{\frac{G*X_{eff}}{V_{ox}}} \tag{8.327}$$

with the parameters $t_0 = 10^{-11}s$, $G = 35V/nm$ (the original publication states 350MV/cm), $X_{eff} = t_{ox}$ at the thinnest spot. This means if there is a defect reducing the gate oxide locally the remaining thickness exactly at this defect must be taken as X_{eff} . Applying this equation to an ESD pulse of 200ns (worst case assumption the ESD pulse is approximated by a rectangular function - clearly too pessimistic, but pragmetic - we get something like:

Example: $X_{eff} = 15nm$, leads to the following time to break down:

Table 49: Gate stress found in a 15nm oxide of a 5V transistor

V_{ox}	E in V/nm	t_{br}/s	remark
5	0.33	$4 * 10^{34}$	normal operation
10	0.66	$6.3 * 10^{11}$	typical gate stress test
15	1.0	15380	ESD, Flash memories
20	1.33	2.51	too risky
25	1.66	0.013	

This table shows that during ESD we can tolerate quite a bit more stress. Exposing a 15nm gate to 10V (instead of 5V, which is specified for long term operation) still looks non fatal.

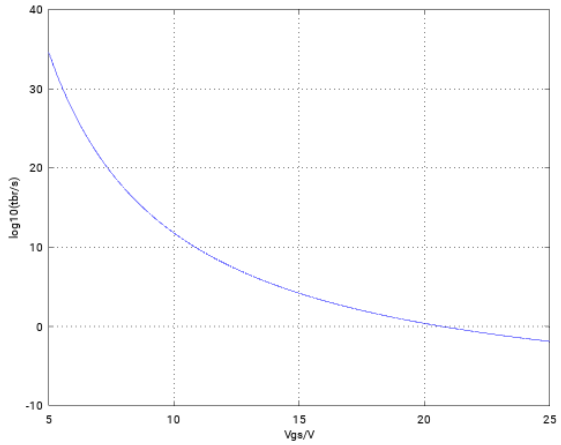


Figure 8.264: log10(tbr/sec) versus Vgs of a 15nm gate oxide

Hot carrier protection: Hot carriers are found if electrons inside the silicon are accelerated such that they can cause lattice defects or get injected into the oxide. Hot carriers typically are produced at zener break down. In a normal diode this doesn't harm too much. The resulting lattice defects reduce minority carrier recombination time. The diode gets a bit more resistive over time.

In the base of a bipolar transistor lattice defects are fatal. The current gain of a transistor with a base damaged by hot carriers degrades immediately! Bipolar transistors connected to inputs need a protection that under no circumstances allows reaching a base emitter break down. Either the base protection has such a low clamping voltage that there is a margin of about factor 2 between the base break down or the base is protected by an anti parallel diode.

Hot carriers inside a MOS transistor usually inject charges into the gate oxide. The damage is caused by a break down of the drain-bulk junction. These charges change the threshold voltage of the transistor. In some cases the change of the threshold can be reversed by hot storage (200°C or more for 24h is a nice test). Hot carrier injection into CMOS transistors can lead to parameter changes of ICs that disapear again after some days or weeks. If not anealed (high temperature storage) hot carrier injection is a cumulative effect.

Junction break down protection: In normal diodes hot carriers are less harmful. Here we rather find thermal destruction when high current and high voltage are present at the same time. Thermal destruction typically starts at about 450°C.

Grain melting protection: Polysilicon resistors consist of many silicon crystals. The resistance strongly depends on the grain boundaries. The more boundaries a current has to cross the higher the resistance. If these grain boundaries melt two grains merge. This leads to a reduction of the resistance. Since poly silicon resistors are embedded between oxide layers there is very little thermal capacity swallowing the energy. Typically at every ESD event some grains will merge. The process is unidirectional (merged grains won't separate anymore). The ESD protection has to clamp the voltage before the current density inside the resistor reaches a destructive value.

Diffused resistor protection: Diffused resistors have more thermal capacity and a lower thermal resistance. Therefore diffused resistors usually can absorb about one magnitude more energy than a poly silicon resistor using the same area.

Metal traces protection: For short pulses in the 100ns range metal traces can handle up to $200mA/\mu m^2$. Since metal layers in modern technologies often are in the range of $0.3\mu m$ thick a single layer metal trace should at least be about $30\mu m$ wide. Stacking several layers of metal reduces the required width.

Latch up protection: ESD events can trigger latch up. Usually the bipolar parasitic transistors have a high base width. Therefore the act as integrators with time constants in the μs range. If latch up tests (carried out with DC signals) show a latch up robustness of 200mA or more it is very unlikely that a short chip level ESD pulse leading to a current of 2A for 100ns (corresponding about $0.2\mu As$) triggers a latch up. Directly applying the board level gun test to a chip is a different pair of shoes because ESD gun tests can lead to peak current up to 25A! (corresponding $2.5\mu As$).

8.14.2 ESD models

There are several ESD models used for testing ICs. The most common one is the human body ESD model (HMB). The tester mimics the typical impedance of a person touching an IC. The resistance of the skin is asumed to be $1.5k\Omega$. This resistance limits the current. The capacity is estimated to be about 100pF. This model was first used in MIL-STD-883E [60]. Later revisions have slightly enhanced the model adding some parasitic inductance and some pF bridging the resistor. But the core of the model remains more or less the same since 1989.

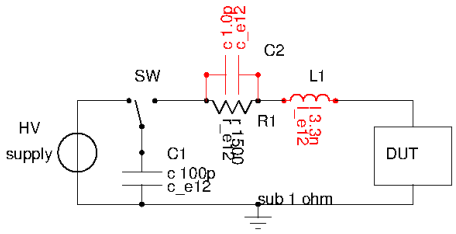


Figure 8.265: human body ESD model used to test integrated circuits

In the mode shown above L1 and C2 represent parasitic capacities found in a typical tester. C2 and L1 vary depending on the test equipment manufacturer. Idealy L1 should be an ideal short and C2 shouldn't exist.

The ESD gun test is a variant of the human body model. Normaly the gun test is applied to boards rather than to naked ICs. The ESD gun test model is some kind of a worst case assumption what might happen to complete systems in rough environment. Since this test usually is applied to complete boards additional protections present on the board will absorb a considerable part of the energy. There is one exception: BUS drivers directly connected to cables leaving the board usually are tested with the ESD gun test too! The test equipment is grounded with a long metal cable (about 2m long!). The inductance of this cable ($2\mu H$) leads to several periods of ringing. The resistance of the test source differs from from the normal human body model: 330Ω. Mainly due to the long cable the model of an ESD gun test loos as follows:

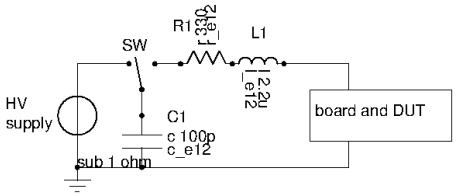


Figure 8.266: ESD gun test model

An important difference between the ESD human body model and the ESD gun test model is the value of the resistor R1. The peak current flowing in an ESD gun test is 5 times higher than the current flowing in the classical human body model!

The ESD machine model (MM) mimics automatic production equipment. Handlers or robots are expected to be grounded and the arms of robots and handlers are assumed to be well conducting (metal arms). So the voltage is lower than using the human body model. The only current limitation in many cases is the pin inductance of the IC package and the inductance defined by the shape of the conducting path. This leads to a model mainly consisting of a capacitor and an inductor. The resistor only is there for damping the resulting resonance. AEC-Q100-003 specifies the following ESD test circuit using a 200pF capacitor, an inductance of $1\mu H$ and a resistance of less than 1Ω. The biggest problem of a MM ESD tester is to build a switch of less than 1Ω and a switching speed in the ns range. Therefore the reproducability of MM ESD tests is unsatisfactory.

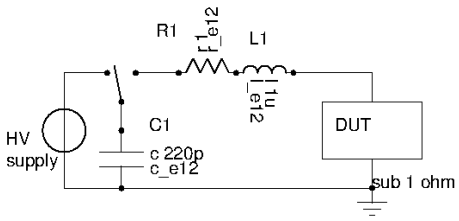


Figure 8.267: ESD machine model test circuit

The charged device ESD model (CDM) mimics a charged IC dropping on a conductive ground. The complete chip gets charged and then it is touched with a probe tip. This leads to a spark discharge with almost unlimited dV/dt. The peak currents flowing for some hundred ps to some ns can exceed tens of Amperes.

CDM pulses are too fast for the classic ESD protections to react. The energy reaches the circuit and the only thing that can be done is to limit the currents adding intentional resistors (even if these aren't required by the circuit) on the chip. Since CDM uses an air discharge reproducability of a charged device ESD test is poor. Even worse the time constants of the discharge strongly depend on the package of the chip and the mechanical geometry of the test probes.

There are different standards how to test CDM ESD events. Most standards accomodate to the technical limits of certain test set ups. (AEC-Q100-011, ESD STM5.3.1, JEDEC JESD22-C101-A).

The following circuit shows a - more or less - representative equivalent circuit.

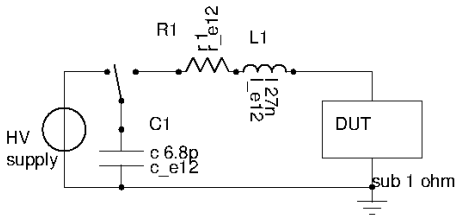


Figure 8.268: CDM ESD model

The circuit here is an approximation to the "average" of the different equivalent circuits of different testers. Simulation using this model should show a considerable margin to achieve a good probability of passing a real CDM test on different testers.

TLP, Transmission line pulse ESD tests are a method to analyse what happens inside the chip applying an ESD pulse. For a TLP test a transmission line is getting charged with a known current. When the current is flowing a switch is being opened. The resulting - ideally rectangular - pulse can reach an amplitude of:

$$V_{TLP} = I * 2Z$$

Z is the impedance of the transmission line. If the ESD pulse triggers a break down the voltage will be lower than the open line voltage. The reflected wave can be analyzed using a network analyzer (using inverse FFT). Even without using reverse FFT a lot can already be seen looking at the time domain signal (using a very fast oscilloscope).

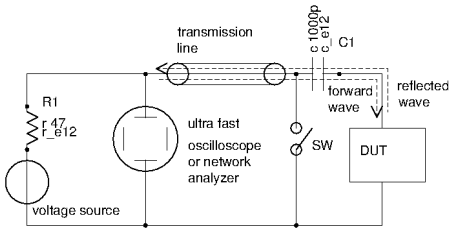


Figure 8.269: TLP test setup

TLP ESD testing is not a tool for qualification and mass testing of components. It is more a tool used for circuit analysis.

8.14.3 ESD protection circuits

The cost of ESD protection depends on the energy dissipated in the protection. The energy is converted into heat. The temperature limit for silicon is about 400 deg. C. Above this temperature the contacts start to melt and the liquid metal fill be drawn into the silicon by the electric field. This created conductive filaments short circuiting the pin.

Since ESD pulses have a duration of only a few ns to some hundred ns the heat will not flow but stay in the dissipating volume.

The power dissipation of an ESD protection simply calculates:

$$P_{esd} = V_{clamp} * I$$

For the heat calculation the total energy matters.

$$Q_{esd} = \int V_{clamp}(t) * I(t) * dt$$

asuming the clamp voltage is more or less constant the energy becomes:

$$Q_{esd} = V_{clamp} * \int I(t)dt$$

Assuming an RC discharge network (like human body) and a clamp voltage that can be neglected compared to the ESD voltage the current trough the ESD protection is very close to:

$$I_{esd}(t) = I_{peak} * exp(-\frac{t}{RC}) \tag{8.328}$$

and integration over the current becomes:

$$\int I_{esd}(t) = \frac{V_{esd}}{R} * RC * (0 - (-1)) = V_{esd} * C$$

(Note: The equation holds a simplification. The peak current is calculated as V_{esd}/R . This neglects the voltage drop over the ESD protection structure. This simplification only is justified as long as $V_{clamp} << V_{esd}$.)

The thermal limit for the energy is

$$Q_{esd} = C_{th} * \Delta T = W * L * D * \rho_{si} * C_{thsi} * \Delta T$$

W is the width of the dissipating junction, L the length and D is the thickness of the conducting channel.

This way the required dissipating volume can be estimated:

$$W * L * D = Vol = V_{clamp} * V_{esd} * C * \frac{1}{\Delta T * C_{thsi} * \rho_{si}} \tag{8.329}$$

Since R, C and V_{esd} are defined by the ESD standards and the thermal properties of silicon cant't be changed by design the only way to create a small protection is to reduce the voltage drop over the protection. The following table lists the energy dissipated using a MIL STD 883 human body ESD model.

Table 50: Energy of ESD pulses

Vesd	2kV	4kV	8kV	16kV
$V_{clamp} = 1V$	$0.2\mu J$	$0.4\mu J$	$0.8\mu J$	$1.6\mu J$
$V_{clamp} = 5V$	$1\mu J$	$2\mu J$	$4\mu J$	$8\mu J$
$V_{clamp} = 10V$	$2\mu J$	$4\mu J$	$8\mu J$	$16\mu J$
$V_{clamp} = 40V$	$8\mu J$	$16\mu J$	$32\mu J$	$64\mu J$
$V_{clamp} = 80V$	$16\mu J$	$32\mu J$	$64\mu J$	$128\mu J$

The energy to be absorbed by the ESD protection directly leads to the sil1con volume needed for the junction or the channel dissipating the energy. In the following table a change of temperature of the silicon of 350K is assumed.

Table 51: Area needed to adsorb the ESD energy

Vesd	2kV	4kV	8kV	16kV
$Vol@1V$	$3269\mu m^3$	$6538\mu m^3$	$13076\mu m^3$	$26152\mu m^3$
$Vol@5V$	$16345\mu m^3$	$32690\mu m^3$	$65380\mu m^3$	$130760\mu m^3$
$Vol@10V$	$32690\mu m^3$	$65380\mu m^3$	$130760\mu m^3$	$261520\mu m^3$
$Vol@40V$	$130760\mu m^3$	$261520\mu m^3$	$523040\mu m^3$	$1046080\mu m^3$
$Vol@80V$	$261520\mu m^3$	$523040\mu m^3$	$1046080\mu m^3$	$2092160\mu m^3$

For a simple bulk diode the length of the depletion zone can be designed to be about $3..4\mu m$. (This usually is done adding some resistive path between the contacts and the PN junction). Pushing the junction away from the surface the "hight of the conductive path too can be designed to be about $3\mu m$. So the required width of an ESD diode becomes about $300\mu m$ for 2KV HMP ESD protections.

A 5V clamp can be designed (adding some drain extension regions and source extension regions) to have a channel length of about $5\mu m$ (from drain contact to source contact). So a typical 2KV ESD clamp for 5V needs a width of about $800\mu m$.

Building clamps with 40V clamping voltage in most cases is done using DMOS transistors. Adding some resistive drain extensions and resistive source extensions the distance from drain contact to source contact can be made as long as about $10um$. Using lateral transistors (LDMOS) the "thickness of the conducting channel will not become significantly more than $4\mu m$. So a 40V clamp can be expected to need a width of $3000\mu m$ for 2kV HMB ESD.

For an 80V clamp this size can be expected to double leading to about $W = 6000\mu m$ for a 2V HMB ESD protection.

The numbers for the 40V clamp and the 80V clamp are on the optimistic side! Practical designs often suffer from inhomogenous current distributions. So widths of about twice to 4 times the numbers calculated here are common to have some margin for non ideal effects.

If the on chip ESD protection has to absorbe the energy of a gun test (for instance CAN, LIN, Flexray transceivers) the size of the ESD protection will go up by factor 5 due to the lower resistance (330Ω instead of 1.5kΩ !)

For cost reduction using thyristor like structures is tempting. These structures trigger at a high voltage (for example 80V) and then snap back to a lower voltage (for example 10V) to reduce the power dissipation of the protection. Snap back protections however are fatal if there are ESD events during operation and some external capacitors are conneted to the pin (for instance voltage regulator pins, supply pins etc.)

These area and cost considerations directly lead to the typical ESD protection using diodes and one central clamp. The drop over the diodes is low and the resulting silicon real estate for the diodes is low. The area needed for the central clamp scales with the clamping voltage but it only is required once in the network.

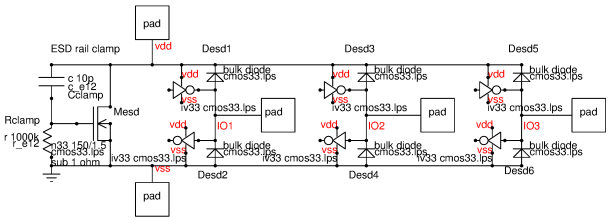


Figure 8.270: classical ESD protection of logic ports

In the example shown above 3 IO pins are sharing one clamp structure. The voltage drop over the clamp Mesd is in the range of several volts while the drop over the diodes in in the range of 1V.So the diodes can be made much smaller than the ESD clamp transistor Mesd. The more IOs are sharing the same clamp the lower the impact of the ESD protection on the size of an individual port. Typical logic chips have about 8 to 32 IOs sharing one clamp transistor. This of course requires a low resistive path from the diodes to the clamp. Typical design rules require that the metal path from each of the diodes to the clamp has less than 2Ω for a 2kV HMB ESD protection.

9 Random access memories and registers

Random access memories (RAM) usually are big building blocks consisting of thousands or millions of cells. These building blocks usually come together with the address decoders and the read and write amplifiers.

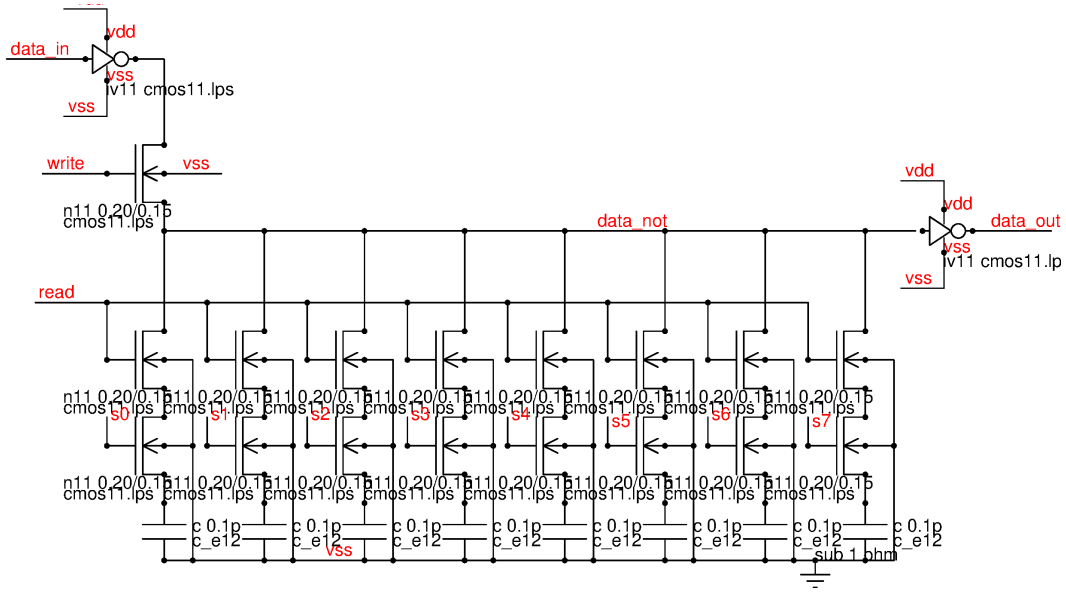
There are two basic flavours of RAM cells:

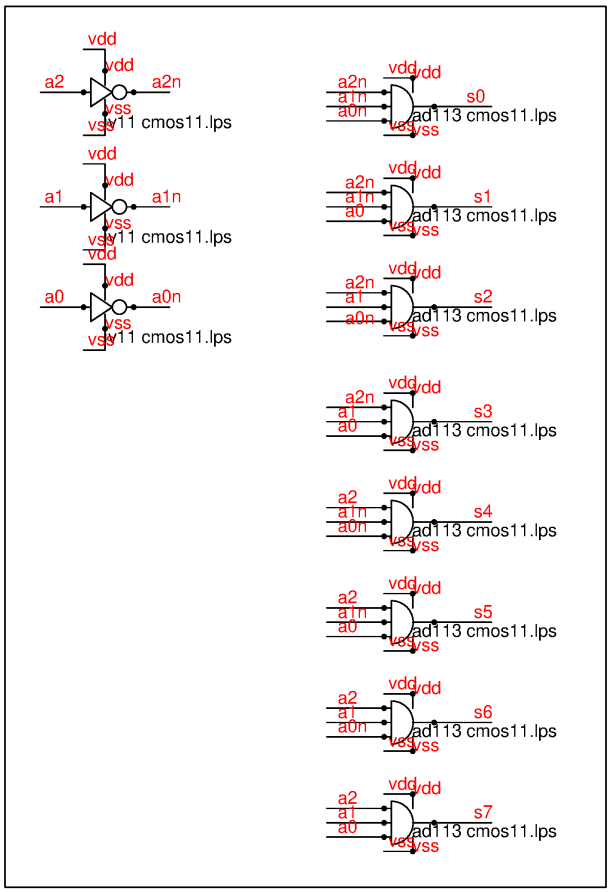
1. Dynamic RAM cell
2. Static RAM cell

The dynamic RAM cell is a capacitor together with a read transistor. Since the capacitor slowly looses charge the dynamic RAM cell needs to be refreshed periodically.

The static RAM cell is a latch together with write and read line switches. The latch holds it's state and doesnt need periodic refresh.

registers usually are composed of standard flip flops. They are not as optimized for a specific process as memory cells. Often registers are used as fast buffers between a slow RAM and a fast CPU.





Decoder

54 transistors

Figure 9.1: 8 bit dynamic RAM

Now we have a total errort of 56 transistors inside the decoder +16 components in the memory cells +5 transistors in the amplifiers. We end up with 9.6 transistors per bit. Still not convincing.

Things get more interresting if we arrange 8*8 bits sharing the same decoder. In other words we write complete bytes in stead of single bits. With the same decoder we can address 8 byte=64 bit. The total effort becomes 56 transistors +8*(5+8*2) = 3.5 transistors per bit.

Writing 16 bits (2 bytes) with the same address decoder further reduces the effort per bit. The bigger we make the array the cheaper it gets! One of the challenges of memory design is to maximize the number of bits and at the

same time find the smallest possible address to select line decoder. However there is one limit. The effort for the memory cell always will remain 2 components (capacitor and 2 NMOS transistors sharing the same active area). So including decoder the cost will always remain above 2 components per bit.

9.0.1 Data line capacity

Up to now we assumed the storage capacity is magnitudes bigger than the stray capacity of the data line. A simplification that will no longer be true with increasing memory size. Worst case the data line is fully discharged and the memory capacitor is charged. After closing the switch the voltage at the read amplifier will be:

$$V_{data} = vdd * \frac{C}{C + C_{line}} \tag{9.1}$$

Using a simple inverter as a read amplifier we need about 70% of vdd to reliably achieve a 0 (at data_out). With the simple design shown above we can calculate the maximum permissible stray capacity of the data line.

$$C_{line} < 0.429 * C$$

One way out is to simply precharge the data line to exactly the trip point of the inverter used as a read amplifier.

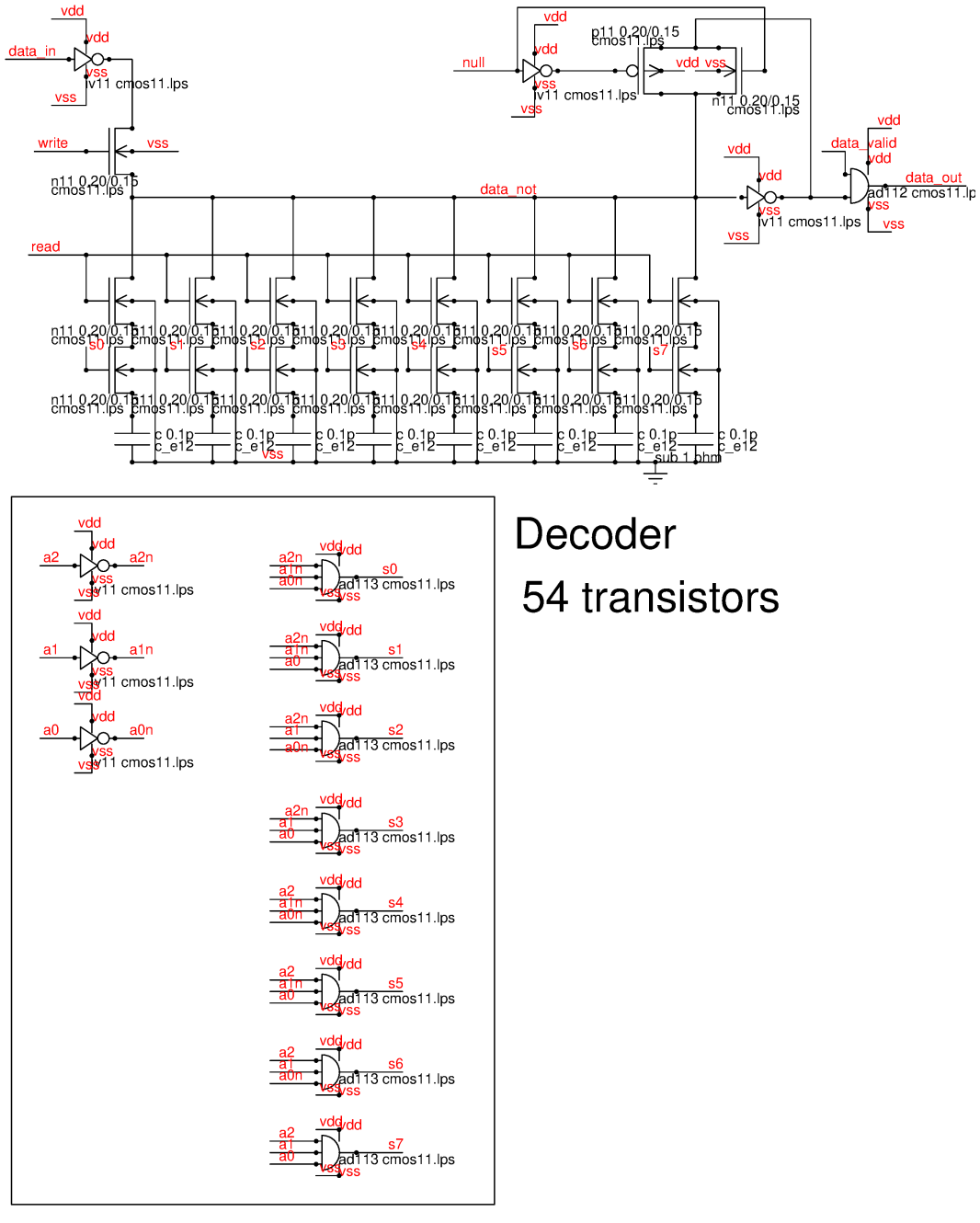


Figure 9.2: 8 bit dynamic RAM with auto zero read amplifier

Basically this is the same design as before with only one anhancement: The 'null' signal. This signal is logic 1 right before a read access is done. It connects the input and the output of the read inverter. The inverter will settle

exectly in the middle between HIGH and LOW. Of course we don't want such an undefined state to propagate into the logic. For this reason the signal of the amplifier may only be propagated to the logic while no nulling is taking place. During nulling the signal 'data_out' is forced to 0 by a LOW at signal 'data_valid'. The required ratio between the storage capacitor and the wire capacity now is determined by the gain of the inverter used as a read amplifier. The required signal at the input of the read amplifier becomes:

$$\Delta V_{in} > \frac{vdd}{2 * gain} \tag{9.2}$$

Since we are starting at about vdd/2 the voltage change reading the capacitor is

$$\Delta V = \frac{vdd}{2} * \frac{C}{C + C_{line}} \tag{9.3}$$

Combining both equations we get the requirement:

$$C_{line} < C * (gain - 1) \tag{9.4}$$

Depending on the early voltage of the transistors (channel length!) and the transconductance a single inverter usualy has a voltage gain in the range of 3..10. If higher gains are needed using 3 inverters in stead of one inverter is an option. On the other hand the more inverters are used the more critical the stability of the inverter chain during nulling will get. (each inverter adds a pole to the loop during nulling! An odd number of inverters higher than 1 can lead to ring oscillation. This must be treated exactly like the stability of an operational amplifier.)

An alternative to using multiple inverters is using a single inverter with increased channel length. The increased channel length increases the early voltage of the transistors and the output impedance of the amplifier stage. (In some technologies the halo implant destroys the analog performance of the transistors. If transistors without halo implant are available consider using the transistors without halo for better voltage gain of the inverter acting as a read amplifier.) So the gain increases without the risk of ring oscillation. The draw back of long channel inverters is the reduction of speed of the read amplifier.

9.0.2 Refreshing

The dynamic RAM storing data in a capacitor requires a periodic refresh because leakage currents may charge or discharge the capacitors. To refresh the data each bit must be read periodically and rewritten again.

Leakage can be caused by:

1. Junction leakage (thermal carrier generation)
2. Oxide tunneling in the capacitors
3. particles (in the capacitor oxide as well as in the switches)
4. short channel leakage
5. photogeneration of carriers
6. carrier generation by radioactive radiation

Thermal carrier generation Thermal carrier generation either leads to an increase of the refresh rate with increasing junction temperature and/or a limitation of the application temperature of dynamic RAMs

Oxide tunneling Oxide tunneling prevents scaling dynamic RAMs for oxides below 10nm. Since modern semiconductor technologies use oxide thicknesses of about 2nm the oxide required for the capacitors can't be scaled with technology. While the minimum transistor size used decreases the storage capacitor cant be shrinked anymore. As soon as the size of the capacitor becomes bigger than 6 minimum transistors dynamic RAMs become unattractive because the static cell is smaller.

Particles Particles in the oxide are statistically distributed. The bigger a memory block the higher the risk that there is a cell that suffers from particle induced leakage. This makes the complete memory block unusable. For this reason dynamic RAMs often are organized in several blocks plus spare blocks in stead of using one single block. If one of these blocks has a particle damage one of the spare blocks can be mapped into the same memory location to replace the damaged block.

This remapping can be done after production at wafer sort or even dynamically during operation (typically during the memory self test at device booting).

Particles cause a weak spot in the oxide. Often the weak spot develops an increasing leakage with operation time of the chip and temperature of the chip. This means the cell is tested good after production but fails after a certain operating time. As long as there are enough spare blocks available this can be recoverd. Once the number of damaged and aged blocks exceeds the number of spare blocks the emory will fail.

Short channel leakage To prevent short channel leakage the switches between the capacitors and the data line must have a certain minimum length (about $0.7\mu m$) . If shorter channel lengths are used halo implants that loacally increase the threshold of the transistors (at the ends of the channel) will help. But even with halo implant a minimum length of about $0.2\mu m$ has to be used. This is a second reason why dynamic RAM cells loose attractivity for technologies with channel lengths of 200nm and below.

Photogeneration Visible and ultraviolet light will generate minority carriers in the switches. The switches become leaky and the memory looses data. For this reason dynamic RAMs must be protected against light.

Radioactive radiation Radioactive radiation can directly discharge the cells producing minority carriers in the switches (like optical light) and it can damage the oxide and the junctions themseves.

Minory carrier generation will immediatelly discharge the memory. After the impact of radiation the memory may be usable again.

Hard radiation having enough energy to damage the oxide or the lattice of the semiconductor will destroy the complete memory. There is no more chance to recover it. For this reason dynamic RAMs are not recommended for space use and for airborne applications (even operating dynamic RAMs at higher elevation of some thousand meters will already lead to a significant reduction of operating life time!). The same applies to application in radiactive environment (nuclear power plants, uranium mining etc.)

9.1 Static RAM cell

The static RAM cell is a latch together with read and write channel switches. So the minimum cell requires 6 transistors. Since transistors have gotten smaller and smaller throughout the years the area advantage of dynamic cells has diminished and using static RAMs has become common.

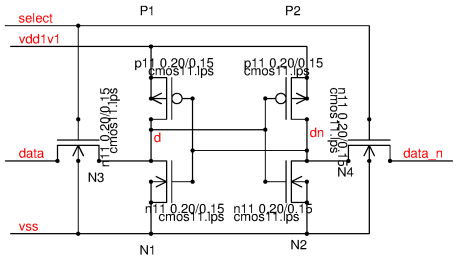


Figure 9.3: 6 transistor RAM cell

The latch storing the data consists of N1, N2, P1, P2. The transistors N3 and N4 are needed to connect the cell to the data lines. To write the cell one of the data lines must be forced HIG while the opposit data line is forced LOW. To overwrite the cell the HIGH level must be above the threshold of the NMOS transistor while the LOW level must be below the supply rail vdd1v1 minus the threshold of the PMOS transistor. Since the transistors are small these thresholds have a significant spread. Making things worse the number of cells can be in the millions. So designing for 6 to 8 sigma is mandatory to achieve sufficient yield. The required cell overdrive for secure programming becomes:

$$V_{overdrive} > 6 * \sqrt{4} * \frac{1mV\mu m * t_{ox}}{\sqrt{W * L} * nm}$$

Let us plug in numbers to see what happens:

$$V_{overdrive} = 12 * \frac{1mV\mu m * 3nm}{\sqrt{0.2 * 0.15}um * nm} = 208mV$$

This is an offset that can be positive as well as negative. Even worse: The estimation of the offset assumed a mature process. During process ramp up offset voltage often is less under control and the estimation shown above is too optimistic. So the minimum supply voltage (assuming a mature process) becomes about:

$$vdd1v1_{min} = V_{thn} + V_{thp} + 2 * V_{overdrive}$$

A supply voltage of 1.1V may already become narrow. Our RAM performance and the required supply voltage significantly depends on the offset of the transistors of the cell!

To not loose drive voltage the select transistors must be driven with about 0.7V higher levels than vdd1v1.

Using low Vt transistors may help to reduce the supply voltage. Unfortunatelly low Vt transistors have more weak inversion leakage. So this is surely not the preferred way out. Eventually the minimum size of the memory cell is limited by the matching properties of the process.

To allow using smaller transistors it is possible to increase the supply voltage temporary while the memory is written. This helps to reduce the cell size but increases the stress level. The transistors will age at every write access. To build small memory cells the process must be tailored exactly to the memory requirements and the life time requirements of the application.

9.2 Registers

Registers often are used for just some bit to some thousand bit. In most cases it isn't worth the effort to tailor each transistor to the absolute minimum required for it's specific task. In stead standard logic transistors from the PDK are used.

10 Non volatile memories

10.1 Zener zap

Zener zapping was a classical trimming process used in bipolar technologies. The initial state of a zener zap is open (open means there is a zener diode in the circuit. Take care of the polarity of the diode!). To make a connection the zener diode is exposed to a current pulse of about 400mA for about 1ms. In the beginning the power dissipation of the zener diode is about 20V*0.4A=8W (The zener voltage is about 7V. The rest of the drop is caused by the resistance). The zener diode reaches a temperature high enough to melt the metalization. The aluminum diffuses into the junction (zener zaps must be designed with some excess metal at the n+ side. If there isn't enough metal the trace simply gets disrupted). As soon as the aluminum filament connects both sides of the zener structure the voltage drops to about 1V..2V reducing the power dissipation.

After zapping a typical zener zap has about 5Ω to 15Ω.

The biggest draw back of zener zapping is the high current required. Each zener zap must be connected to 2 probe pads to provide the current. These 2 pads mainly determin the area needed for each trim bit.

Using tungsten plugs as contacts zener zapping will not work anymore because the tungsten doesn't melt and blocks the aluminum from diffusing into the junction of the diode.

The high power dissipation of the zener zapping process may damage the passivation over the zener diode. If this happens the zapped diodes are exposed to corrosion. If the diode and the zapping process ared engineered properly (no rupture of the passivation) zener zapping will work reliably.

10.2 Laser trimming

10.2.1 Poly silicon fuse

10.2.2 Thin film resistor trimming

10.3 Memresistor

10.4 EEPROM

An EEPROM is an electrical erasable and programmable read only memory. Well, not quite state of the art any more. Todays EEPROMS can be written and erased multiple times.

This was already the case in the 1990s. But at that time there was no production experience of how many times the process of writing and erasing can be repeated without damaging the cell. For this reason the in the beginning these memories were specified as "write once, read multiple times" or in other words as a read only memory (ROM).

With increasing production experience writing the memory multiple times was permitted.

Each write and erase cycle stresses the gate oxide of the memory cell. Accumulation of surface states prevents erasing the cell after a certain number of cycles. Gate damages cause by the high gate voltages during write and erase lead to a leakage that discharges the gate storing the data.

10.4.1 SLC single level cell

The single level cell (SLC) is the most classic form of an EEPROM cell. The cell is used in two states: charged and discharged. One cell can only store one bit. This kind of cell has the highest charge per bit but the lowest memory density. This makes the SLC the most expensive variant of memory cells.

10.4.2 MLC multi level cell

The multi level cell (MLC) uses multiple levels of the shift of the threshold of the storage transistor to code 2 bit in one cell. Distinguishing these levels requires more accurate read amplifiers and even slight discharges of the cell change the data. MLCs are cheaper than SLCs but the data retention suffers.

10.4.3 TLC tripple level cell

In a tripple level cell (TLC) the single cell is used to store 3 bit. This means the range of thresholds of the storage transistor represents 8 different analog values. The TLC requires even higher read amplifier accuracy and is even more sensitive to gate discharge than the MLC.

10.4.4 Pseudo-SLC (pSLC)

A pseude single level cell (pSLC) is a multi level cell that has a read out amplifier that can be programmed to either splitt the range in 4 or 8 levels or only in two levels like in a SLC. Using a MLC in SLC mode leads to excellent data retention comparable with a classic SLC.

Table 52: Properties of multi level storage cells

parameter	SLC	pSLC	MLC, TLC
bits per cell	1	1	2-3
no. of R&W cycles	60k-100k	20k-30k	typ 3k
Data retention	10-15 years	1-10 years	1-10 years
reliability	high	medium to high	low to medium

10.5 Cross point memory

Cross point memories are the latest evolution of non volatile memories. It is based on nano tubes [44, 43].

11 Back on the Top Level

After all the basics about the standard building blocks and the technology it is time to revisit the top level. There are hundreds of important considierations! A bad top level will destroy the performance even of the best analog design.

11.1 The package

All signals present inside the integrated circuit are disconnected from board ground or ideal supplies. There is pin inductance everwhere! Usually the application refers to the ground plane on the board. At 100MHz or higher this ground plane doesn't have much in comon with the ground inside the ICpackage or with the supplies present on the chip!

How substrate is connected to board ground depends on the substrate resistivity an on the inductance of the path to the ground plane of the board.

11.2 ESD considerations

Under normal circumstances all ESD protections are located in the pad ring.

11.3 Electromagnetic emission considerations

Even pins that are expected not to be very active can produce a significant RF emission (EME , Electromagnetic Emission). Usually there is an RF generator hidden somewhere in the chip that couples to the pins expected to be silent via parasitc paths. One of the most commom RF propagation paths is the ground system of the chip. This ground system is connected to the board via the impedance of the bon wires. At high frequency the bond wire is mainly an inductance. So RF ground currents may produce unwanted voltages (with reference to board ground almost everywhere.

11.3.1 EME System point of view

From the system point of view EME (Electromagnetic emission) is an RF radiation property. It disturbs other systems (e.g. radio receivers etc) by radiated waves. The system designer first of all is interrested in the measurement of radiated RF. Radiated RF includes the properties of the transmitter (frequency, power, source impedance) as well as the properties of the antenna (antenna length, directivity, impedance of the antenna). Everything attached to the electronic system such as cables or even other conducting material that could act as anatenne is part of the system level emission measurement.

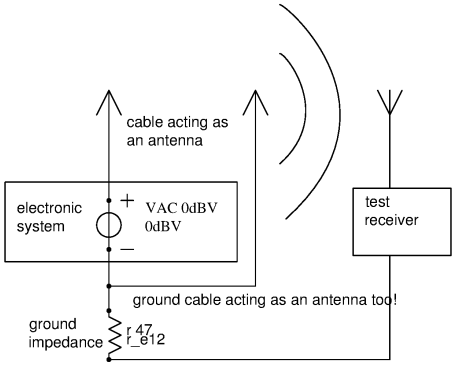


Figure 11.1: electromagnetic emission from system point of view

The test receiver represents the radio systems that may not be distorted. The transfer function from the RF source sitting inside the electronic system to the test receiver depends on the antenna properties of the radiating antenna as well as the ground wire (that carries the same signal but with shifted phase and amplitude). Whether the signals of the different antennas add or cancel depends on the antenna geometries and on the complex ground impedance.

Since the properties of such multi antenna systems are fairly unpredictable the system designers tend to use standardized setups representing some kind of an average worst case application.

For large applications (complete cars, planes, ships...) the complete application is placed in an anechoic chamber and the receiver is made ground-independent using a dipole antenna. The polarization of the dipole antenna is getting rotated to cover all possible orientations of the wave and the receiver is mover around the system.

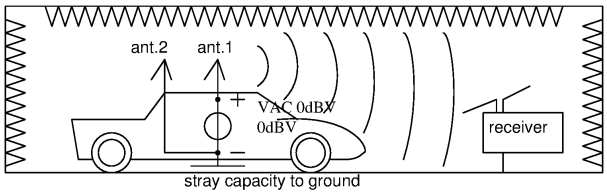


Figure 11.2: A typical test setup for complete systems regarding radiation of all wires together

A test setup using an anechoic chamber is expensive. Therefore before doing this kind of test simplified setups are used to characterize the electronic subsystems included in the car. This way a much smaller (and cheaper) anechoic chanber can be used.

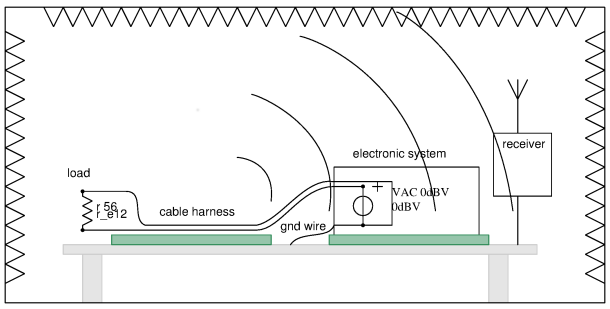


Figure 11.3: Subsystem test with standarized setup

The subsystem is placed over a metal table. The insulation layer typically is about 5cm thick similar to a typical spacing between a board and the chassis of a car. In the same way the cable harness to the load(s) is running over an isolation. The electronic system and the cables act as antennas. The waves reaching the walls of the anechoic chamber will be absorbed. The return current has to flow back into the system via the conducting walls of the chamber, the metal table and the gnd wire. Typically the gnd wire is about 20cm long. So the board ground and the table will differ!

The test receiver uses the metal table as it's ground reference. This setup has important consequences:

- Referring to the metal table as receiver ground the board ground carries an RF signal
- Every wire connected to board ground will radiate (with reference to the metal table)
- At low frequency (some MHz) the return current flowing through the gnd wire corresponds the sum of the currents flowing in the antennas

- At high frequency most of the return current will flow via the stray capacity between the board and the metal table
- At a certain frequency there is a parallel resonance of the inductance of the gnd wire and the board's stray capacity

The fact that the return current flowing in the ground wire represents the sum of the antenna currents can be used to further simplify the setup. This leads to the 1 Ohm method often used to characterize the RF emission of logic ICs. At the characterization of logic ICs the antenna (connected to I/O pins) are replaced by typical load impedances.

11.3.2 Logic acting as an RF source

The current flowing in the supply rails of the logic consists of very short spikes. These can almost be regarded as dirac pulses. (True up to about 500MHz). The logic can be blocked on chip to make the spectrum of the current consumption roll off. This local blocking only solves the problems partially. The following figure shows an example of a locally blocked logic in a chip with substrate connected to an exposed dice pad. In this simplified circuit the inductance of the exposed dice pad is assumed to be 0. (Real numbers are in the range of some 10nH depending of the package).

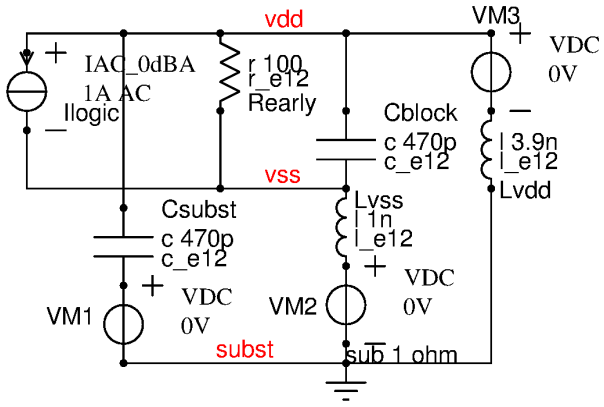


Figure 11.4: Example of a logic acting as a noise source

The current generator Ilogic represents the current consumption of the logic. Since we are mainly interrested in the resonances of the system it is an AC current source.

Real life CMOS circuits are not ideal current sources. So Rearly represents the impedance of all the logic transistors involved. It is determined by the current consumption of the logic and the early voltage of the logic transistors.

The logic nwell is assumed to have a capacity to substrate subst represented by Csubst. Since we assume having an exposed dice pad the substrate capacitor is directly connected to the board ground via VM1 (net subst is connected to the ground sysmbol). VM1 is a 0V DC source used for measuring the current in simulation.

Cblock is the on chip blocking capacitor. It acts as an RF short (or at least low impedance path) between vdd and vss. vss is assumed to connect to a ground pad that has a bond wire down to the exposed dice pad (Lvss, VM2 for measuring the current)

The logic additionally has an external blocking connected to vdd via bon wire Lvdd. (Lvdd represents the sum of all inductances from vdd via the bond wire to the external capacitor, the trace inductance and the parasitic inductance of the external capacitor). Te external capacitor is assumed to have an capacity that is magnitudes higher than the on chip blocking capacity. For this reasen (we are not interrested in DC) the capacitor is siplified by a short from vdd to subst.

As long as no intentional capacitors are placed the capacity between the nwell (tied to vdd) and the pwell (tied to vss) has the same magnitude as the substrate capacity. We find two resonant tanks:

1. Lvdd together with Cblock and Csubst in parallel. The approximate frequency calculates:

$$f_1 = \frac{1}{2 * \pi * \sqrt{L_{vdd} * (C_{subst} + C_{block})}} \tag{11.1}$$

   example: Lvdd=4nH, Csubst=500pF, Cblock=500pF leads to f1=80MHz

2. Lvss together with Csubst and Cblock in series. The approximate frequency calculates:

$$f_2 = \frac{1}{2 * \pi * \sqrt{L_{vss} * \frac{C_{subst} * C_{block}}{C_{subst} + C_{block}}}} \tag{11.2}$$

   example: Lvss=1nH, Csubst=500pF, Cblock=500pf leads to f2=318MHz

The current flowing through Lvss will lead to ground bounce on the chip. From RF emission point of view it is recommended to supply all preriphal drivers (IOs, bus drivers, analog signals...) from a different ground than vss and from a different supply than vdd. Most microcontrollers have multiple supply pairs. vdd and vss supply the logic core while vddx and vssx supply the IOs.

Sometimes the ground nets are all tied together for ESD reasons. In this case the periphal drivers are connected to vss as well. In this case it is important to keep the noise level at vss as low as possible and to shift the resonant frequencies into regions where they will not disturb.

The bounce at vss is determined by the inductance Lvss, ther resonant frequency and the current flowing though this inductance. The resonant current depends on the quality of the resonant tank and the exciting current Iac.

$$I_{res} = I_{ac} * Q \tag{11.3}$$

The calculation of the resonant tank quality depends on the type of resonance (Serial resonance with R in series with the tank or paralell resonace with R in parallel with the tank. At f1 the resistor Rearly is in parallel with the resonant tank (at least as soon as $C_{block} >> C_{substt}$). The quality is approximatelly

$$Q_{f1} = R * \sqrt{\frac{C_{block}}{L_{vdd}}} \tag{11.4}$$

At f2 the resistor is in series with the capacitor and the inductor. In this case the quality becomes about

$$Q_{f2} = \frac{1}{R * \sqrt{\frac{C_{substt}}{L_{vss}}}} \tag{11.5}$$

This is an extremely important result. Increasing the on chip blocking capacitor increases the quality of the resonant tank at the low frequency f1. The current through the external capacitor at the resonance f1 increases proportional to $I_{res} \sim \sqrt{C_{block}}$. At the same time the frequency of the first resonance decreases with $f_1 \sim 1/\sqrt{C_{block}}$. Additional to the increase of the resonant current at f_1 the increase of the on chip blocking capacitor forces more of the resonant current into the bond wire Lvss while at the same time the substrate current flowing through C_{substt} gets reduced.

$$I_{vss_{f1}} = I_{res} * \frac{C_{block}}{C_{block} + C_{substt}} \tag{11.6}$$

Furthermore the quality at f_2 for $C_{block} >> C_{substt}$ is independent of of C_{block} . The current provided by the source however gets divided between the resonant tank and the blocking capacitor C_{block} . This current division reduces the amplitude of the second resonance in spite of keeping the quality of the resonant tank stable.

Since these results up to now apply to approximations making C_{block} bigger than C_{substt} a more accurrat result requires solving the problem nummerically.

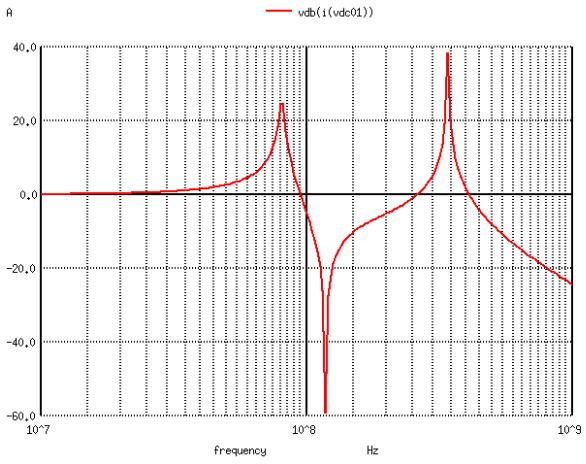


Figure 11.5: Simulation of the transfer function with 470pF blocking capacity

In the above figure the system was excited with 1A (0dB(A)). So we are looking at a transfer function of the excitation current to the bounce of vss with 22dB at the first resonant frequency.

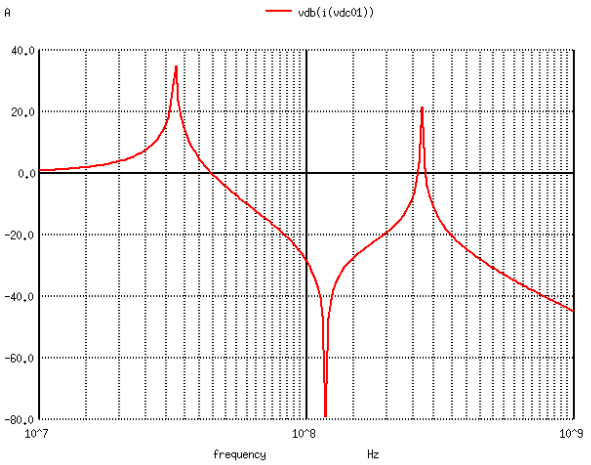


Figure 11.6: Increasing the blocking capacitor to 4.7nF

The increase of C_{block} in simulation leads to an increase of the amplitude at net vss at the first resonace as expected from manual calculation. At tthe same time the current division at the second resonace lowers the amplitude at about 320MHz.

Conclusion: Increasing the on chip blocking capacitor reduces the resonant frequency but increases the noise level on vss by changing the current distribution between C_{block} and C_{subst}.

11.4 Electromagnetic sensitivity considerations

Ideally RF of significant amplitude ($> V_t$) should never ever reach a junction or a gate of a transistor where it can be rectified. In fact some companies specialized on precision measurement equipment have application note that explicitly tell the user that any out of operating band RF reaching an amplifier input WILL BE RECTIFIED! ([47]p 7.122ff). So the first conclusion is: Keep RF away from analog circuits.

Ideally RF injected into an electronic system should be eliminated before it reaches an active component (This does not only apply to integrated circuits. It also applies to discrete transistors!)

Practically complete blocking of RF isn't possible because even on board level nobody has succeeded creating ideal capacitors or inductors. In automotive applications RF levels expected to be present on wires directly connected to cables are in the range of 1W (or 30dBm). Pins connected to traces on a board (so the antenna is only a few cm long) are expected to be able to handle about 100mW (20dBm). Local wires just connecting the adjacent chip are expected to be able to handle 10mW (10dBm). These levels always refer to the chip operating it it's typical application circuit and blocked in the application typical way. As a consequence the EMS (electromagnetic susceptivity) test circuit always consists of a model of the antenna (usually represented by an standard RF source with 50 Ohm), the blocking including ail parasitic inductances and capacities, the board (ground impedance) and the chip and package package (bond wire inductance).

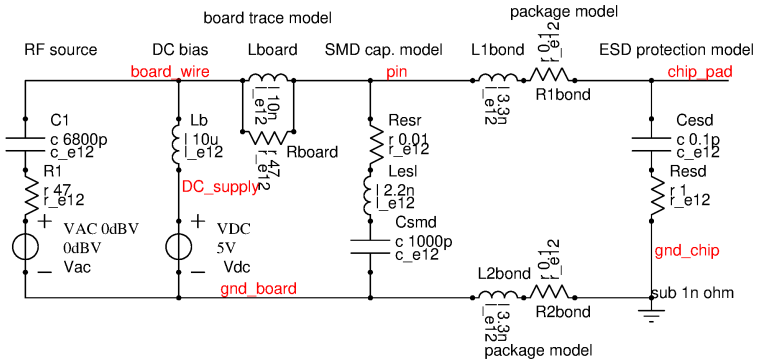


Figure 11.7: Example of an RF injection test circuit

The path from net "board_wire" to net "chip_pad" already provides an significant attenuation. Removing the external circuit (mainly the blocking capacitor Csmd and the wire Lboard) would lead to a completely different result. Typically the properties of the board and the components applied externally play a more important role than the chip itself.

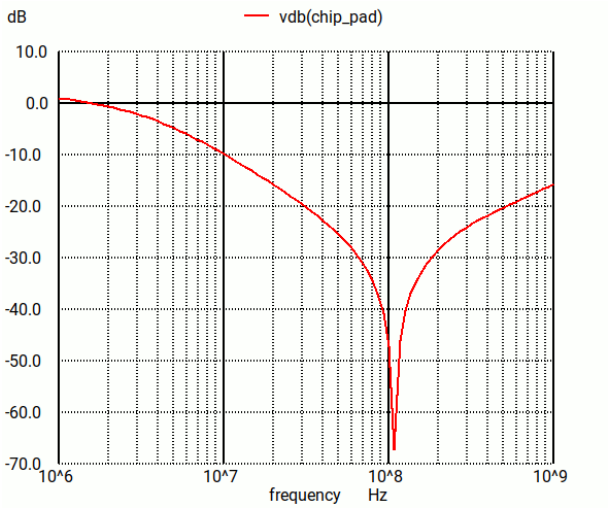


Figure 11.8: AC transfer function to the pad

As soon as more than one pin or multiple ground pins are involved cross coupling effects begin to have an impact on the electromagnetic susceptivity. The sensitivity of circuits inside the chip is strongly influenced by the voltage drop of the ground system. Often blocks that have nothing in common are observed to influence each other.

11.4.1 Low resistive substrate and exposed dice pad

An exposed dice pad provides an extremely low inductive path to the ground plane on the board. In most cases exposed dice pads are used for thermal reasons. To achieve a low thermal impedance he dice pad is soldered to the board ground. If the chip is connected to the exposed dice pad by soft solder the bottom side of the substrate of the chip is directly connected to board ground. The vertical path through the dice pad and the chip can be estimated as a parasitic inductance in the range of 10..40pH depending on the size of the chip. The magntic field of the vertically flowing RF current produces eddy currents in the ground plane. Thus the inductance even is damped.

If the back side of the chip is glued in stead of soft soldered the interface between the substrate and the dice pad forms a capacitor. The dielectric material is a stack silicon oxide (usually only some nm thick) and the resin used. In case conductive glue is used the dielectric consists of the oxide only. Dices in the range of 20 to 30 mm^2can easily have back side capacities of some 10nF to some 100nF if conductive glue is used.

Due to mechanical handling the oxide on the bottom side of the chip usually has scratches. These scratches lead to a parallel resistoance of the back side capacitor in the range of some Ohm to some 10 Ohm.

The vertical path from the bottom side of the substrate to the top side of the substrate is in the range of some milli Ohms. It can be regarded as a condutive plate![39].

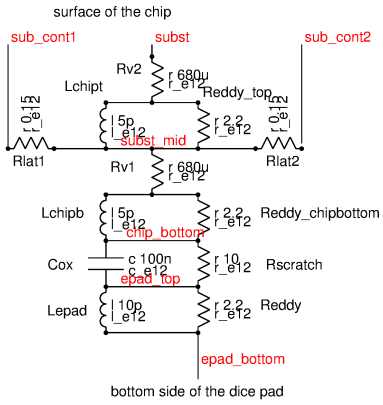


Figure 11.9: Substrate model of a chip with 1mOhm*cm substrate glued to an exposed dice pad with conductive glue

The absolute impedance of the path from subst to epad_bottom varies from 1.5mOhm to about 4mOhm in the frequency range of 10MHz to 1GHz. This is magnitudes lower resistive than the connection to the substrate contacts on the surface of the chip!

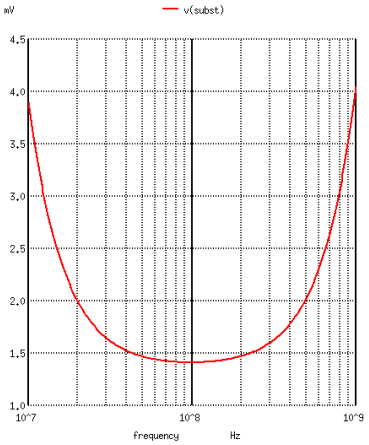


Figure 11.10: Absolute impedance from subst to epad_bottom (System excited with an AC source of magnitude 1)

So for 10MHz to 1GHz the substrate can be approximated as an almost short to board ground. At the same time the ground network can be excited by RF coupled into an other pin of the chip. Here is an example in which RF is injected into an IO pin and part of the RF power flows into the ground of the bandgap.

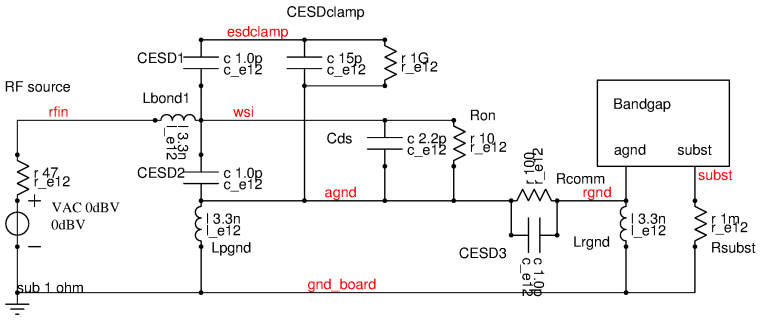


Figure 11.11: Example of a coupling path exciting a different ground domain via the ESD protection between the different ground domains

By this unintentional coupling path RF currents will flow into agnd and will cause a significant voltage drop at Lrgnd. The bandgap will see a differential signal between rgnd and subst that can reach several hundred mV.

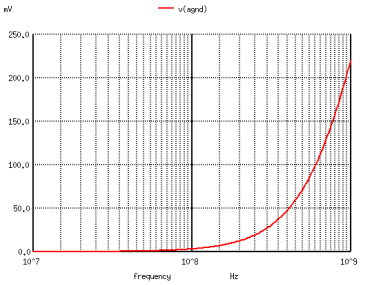


Figure 11.12: RF signal at the ground node of the bandgap

This is a general problem of chips with low resistive substrate connected to an exposed dice pad. All circuits used in chips wth low resistive sbstrate MUST be designed to survive some hundred mV of substrate noise (difference between substrate and cell ground).

12 Testing of intergrated circuits

To test integrated circuits a very wide range of requirements has to be considered. First samples usually have to be tested in the laboratory. Here the main focus is on getting access to almost all parameters but there are no test time constraints. Test conditions often are sweeped to find out margins between the actual silicon and the targets defined in the specification. During qualification such lab tests can be used to evaluate aging effects of the chips.

In production the focus is on testing speed and test coverage. Different from lab testing very often go/nogo tests just checking for the limits are prefered for speed reasons. Devices failing one test will be rejected immediatelly without any attempt to find correlations between the failing test and other tests.

In the field built in tests are often used to detect failure of a safety critical device before any further damage is done. A failing built in test is notified to the microcontroller. Usually this failure gets documented in log files that can be read out at maintenance or trigger a warning even if the system still works.

12.1 Testing in the laboratory

Laboratory or bench testing of an integrated circuit is completely different from testing it on the tester. During bench testing sophisticated high speed equipment such as fast oscilloscopes is available. On the tester the PMUs (power measurement units) usually have settling times in the ms range.

Lab testing often is very close to the typical application of a device. Parasitic board components (for instance parasitic inductance of traces on the board) usually are under reasonable control while on a tester there is a significant spread of parasitic inductances moving from one tester to an other. Functional analog tests at speed usually are much easier to be done at board evaluation than on a tester.

Bench tests usually don't run automatically. Getting statistical data from bench tests usually isn't possible while testing high numbers of chips at wafer sort provides statistical data almost for free.

Test to destruction usually isn't appreciated on a tester because if the tester gets damaged this can be extremely expensive to repair. For this reason tests that might lead to destruction preferably are done during lab. evaluation. In production test to destruction isn't possible at all because this could lead to predamaged parts that later fail at the customer.

12.1.1 Sample preparation and modification methods

Sometimes there is a need to test modified parts. These can be modifications implemented to verify the behavior of a planned redesign under certain boundary conditions that can't be achieved in simulation. These can also be modifications done to make a design work at all or work in certain failure modes.

Samples can be modified in many different ways:

Opening the device: Ceramic packages often can be opened easily using a carpet cutter and a hammer. The cutter is applied at notches where the ceramic package was glued together or at the metal top. To get enough peak force to open the package a careful knock with the hammer helps.

Molded samples can be etched open with acid. The choice of the right mixture is difficult because usually you don't want to destroy the metalization at the pad openings (avoid phosphoric acid at pad openings). For low complexity designs opening from the top to access the metal is common. For modern technologies access from the top side is limited because of the metal filling used in modern processes. Sometimes back side opening has to be used then.

Laser passivation opening: When the mold is removed the passivation usually still hinders access to the wires to be probed. Passivation can in most cases be removed well using a green laser. The green laser is adsorbed well by SiN passivation. The passivation gets vaporized by the thermal energy (use pulses in the range of 5J).

Laser metal cutting: Metal cutting with a laser works best with red lasers. The metal trace gets vaporized.

One of the problems is that some of the vaporized metal settles again on the chip. A laser cut for this reason isn't a real 'open'. To improve this situation remove the passivation on both sides of the opening, place two probes and apply a voltage of about 1V...5V (depending on what the process allows) and allow a current flow of up to 100mA to remove the condensated metal.

All modifications using laser cuts are limited in accuracy due to the wave length of the laser and the precision of the spot.

Focused Ion Beam (FIB): Focused ion beam modifications can be done with by far better precision than laser cuts. To modify technologies with feature sizes below $1\mu m$ focused ion beam (FIB) is the much better choice. Focused ion beam manipulation allows removing material as well as deposition of conductive material (connecting two traces). Metal deposition normally is high resistive ($k\Omega$ range) because the deposited films are very thin.

focused ion beam can also be used to locally remove oxides or passivations to give access to wires buried deep in a low metal level.

FIB manipulation produces charges on the silicon. Using FIB on analog amplifiers can lead to offsets (ions injected into gate oxides). For logic design this doesn't matter too much. For analog amplifiers this ma be fatal. Sometimes hot storage after FIB helps (but this can destroy the content of NVMs!) to recover offsets created by ions in the gate oxides.

A focused ion beam modification usually requires the precise coordinates where the cut needs to be done. The operator works more or less blind. E.g. on planarized chips visibility of structures in an electron microscope (this is part of the FIB equipment) is poor to zero. Typically the operator has an overlay of the layout of the chip and the view of the camera of the electron microscope for orientation on the chip.

Back side preparation: focused ion beam allows drilling holes trough the chip. It is possible to drill holes through hundreds of μm of silicon to access a metal 1 trace from the back side.

Optical manipulation: In some cases behavior of a circuit can be manipulated illuminating transistors connecting to high resistive nets with a 5mW HeNe laser or a diode laser. The laser generates hole-electron pairs in the illuminated area and makes it conductive. One typical application of this method is to find floating nets in a logic (current consumption changes when a critical transistor was hit by the laser light).

The following table gives an overview of the methods

Table 53: Optical manipulation techniques

problem	suggested method	equipment	other remarks
ceramic sample opening	mechanical	carpet cutter, hammer	
plastic sample opening	acid		corrosive, poison
passivation opening	laser	green laser for $\emptyset \geq 1\mu m$	Eye protection
passivation opening	focused ion beam	for $\emptyset \leqq 1\mu m$	
wire cut	laser	red laser	leaves a metal film
	focused ion beam		better precision
rewiring	focused ion beam		some $k\Omega$
drilling holes	focused ion beam		up to 1mm
floating node search	laser	Obirch	
manipulating bits of an NVM	focused ion beam		first check test modes

12.1.2 Functional tests

Functional tests verify the function of a chip using operating conditions identical or at least very similar to the real application. Functional tests are typical for lab evaluation. Usually a device is operated in the lab (on bench) in a typical (or sometimes extreme, possibly out of specification) application. The target is to mimic the operating conditions of the real application as far as possible to verify the performance from customer point of view. (This may even include connecting electromechanical systems to sensors and actuator chips.)

Functional tests may disclose conceptual problems even if the chip is in spec. In other words the chip was specified wrong or insufficiently.

Functional tests may pass even if the test limits are violated. This may mean the application is more error tolerant than anticipated during device specification (possibly overspecified for this specific application).

Functional testing on a tester only is possible up to a certain extent. A simple amplifier or comparator usually isn't a problem. (Unless you want to test offsets in the μV range. Typical testers have quantization in the 1mV range). As soon as high power and/or high frequencies are involved functional tests on a tester can quickly become very tricky or impossible. (Most testers only have PMUs - power measurement units - for about 100mA with analog settling times in the ms range). Adding application circuitry (external components) to the test board is possible, but before running the test the tester should verify that the external components still are present and in spec (board self test). The more external components are placed the more complex calibration and self testing of the board will get.

EMC tests on a tester are almost impossible.

12.1.3 Reliability and life time tests

During device qualification a lot of life time and reliability testing is done. Let's first sort out some of the abbreviations used:

Table 54: List of reliability test abbreviations

abbreviation	what it means
BTI	bias temperature instability
EM	Electromigration
HCI	Hot carrier injection
HTGB	High temperature gate bias
HTRB	High temperature reverse bias
IDDQ	drain quiescent current
	Pattern shift
TDDB	Time dependent dielectric break down

The number of failures permitted during a life time test (or sometimes all life time tests together) is described in Failures In Time of FIT.

$$1FIT = \frac{1failure}{10^9 operatinghours} \tag{12.1}$$

In most cases the so called bathtub curve is expected to describe the change of the failure rate over time. In the beginning we have the early failures caused by production damage of weak spots such as weak spots in a gate oxide. At the end of the curve the failure rate is expected to go up again due to wear out. So the FIT number applies to a certail life time of the devices (for instance 100k hours at a certain mission profile describing temperature and stress voltage or current). The number of failed parts N_{fail}allowed testing N devices for t hours calculates as:

$$N_{fail} = \frac{FIT * N * t}{10^9} \tag{12.2}$$

Example: The failure rate to be guaranteed is 50 FIT. During the reliability test 10000 devices are tested for 10000h (this is 416 days!). The permitted number of failures must be below:

$$N_{fail} = \frac{50 * 10^4 * 10^4}{10^9} = 5$$

Doing stress tests for more than 400 days is barely possible. For this reason reliability tests apply test conditions going beyond the normal operation conditions. Typically the temperature can be modified in an easy way. Increasing the temperature reduces the required test time. The reduction of the test time depends on the physical effect to be detected by the test.

BTI: Bias temperature instability is an effect often observed at semiconductors that have dangling bonds in the interface between the semiconductor material and the dielectric. It is typical for SiC power transistors. Depending on the stress history the threshold voltage changes. For most switching applications this isn't much of a problem. It simply means that there needs to be some margin between the gate drive voltage and the threshold of the power transistor to handle changes of the threshold.

BTI testing usually is part of the technology qualification.

EM: EM is the abbreviation for electromigration. Electromigration is strongly temperature dependent. For this reason it usually is tested at elevated temperatures. The speed of electromigration typically doubles all 6K..20K depending on the mobilization energy ϕ of the metal film. For testing the device is exposed to a high DC stress current while the temperature is significantly higher than the maximum rating to accelerate the test.

At about 290°C additional effects start to become significant for aluminum traces. At such high temperature the interpolation using Black's law doesn't work anymore.

$$MTF = \frac{1}{A * J^2 * exp(-\frac{\phi}{k*T})} \tag{12.3}$$

The parameter A depends on the metal film quality and coverage.

MTF is the time to failure in h at which 50% of the tested devices have failed.

The mobilization energy ϕ depends on the chemistry of the metal and the coverage.

k ist the Boltzman constant and T is the temperature in K.

The following table lists some typical values for amuminum traces:

Table 55: Electromigration parameters to calculate the mean time to failure

material	A	unit	ϕ	unit
Al (small crystallite)	0.404	$\frac{m^4}{A^2*h}$	0.5	eV
AL (large crystallite)	290	$\frac{m^4}{A^2*h}$	0.82	eV
AL (large crystallite, covered with SiO2)	268060	$\frac{m^4}{A^2*h}$	1.16	eV

The parameters were extracted from the plots of [20].

A second possibility to accelerate the test is operating the device at higher currents. The MTF follows:

$$MTF \sim \frac{1}{J^2} \tag{12.4}$$

IDDQ: IDDQ testing stresses the bipolar junctions inside a logic as well as the gate oxides. Typically the test pattern of the logic is stopped at a certain point. Current consumption of the logic is measured while there is no dynamic activity. During the test the supply voltage of the logic is increased to create a certain margin between the maximum supply voltage and the test voltage.

After the test the pattern continues and will be stopped again for the next IDDQ test. This procedure will be repeated until every logic gate was tested in HIGH state as well as in LOW state.

Circuits that have a static current consumption must be disabled or disconnected from the supply during IDDQ test.

IDDQ testing usually only requires some ms to seconds. It can be done in production as well as in the reliability lab and at bench evaluation.

HCI: Hot carrier injection typically is encountered when a transistor is operated at high drain source voltage while a high current is flowing. Hot carriers lead to the injection of "hot" electrons into the gate oxide at the drain edge. These negative charges in the oxide increase the threshold of the transistor.

Testing HCI operating the transistor at high voltage and at the same time at high current usually leads to thermal problems. HCI testing in most cases can only be performed in pulsed mode operation.

HTGB: High temperature gate bias tests target on gate oxide weaknesses such as local thinning of the gate oxide due to particles or scratches. For the test the gate leakage is measured before applying a high gate voltage. Then a higher gate voltage than in normal application is applied for some seconds. After the stress the gate leakage is measured again to check for defects.

Since this test only takes some seconds it can be done in the reliability lab as well as during production testing (Screening of gate oxide defects). This gate stress test requires some preparations in the design of the circuit (avoid bulk diodes that prevent testing).

In some cases HTGB tests are used to characterize the long term stability of a chip. In this case the stress voltage is applied for longer time (days to months).

Polysilicon resistors (over thin oxide layers) and capacitors have break down issues as well. This is especially true it the technology tends to form silicon cones growing into the oxide.

Fringe capacities sometimes have metal filaments. These result from non ideal etching of the metalisation. Such filaments tend to grow with increasing electrical field strength (On board level this effect is called whisker creation)

HTRB: High temperature reverse bias tests characterize mainly the break down behavior of junctions in reverse bias condition. (Typical example: the bulk-drain diode of transistors)

IDDQ tests partially cover HTRB. If the reliability of a junction over a longer life time is to be tested the test typically will be carried out at qualification stressing the junction for several days to months.

Pattern shift: Pattern shift is the lateral movement of the wiring of a chip caused by thermo-mechanical stress. Usually this stress is a consequence of the different thermal expansion of the mold and the chip. To test for pattern shift the molded chip is exposed to thermal cycles. After exposing the chip to some hundred cycles the functionality is tested to detect short circuits or opens caused by the movement of wires.

Pattern shift results always relate to a chip in combination with the package and the mold material used.

TDDB: Time dependent dielectric break down focuses on the reliability of dielectrics exposed to a high field strength over a long time. Usually this test is carried out during technology qualification.

All of these reliability tests require test times in the range of hours to thousands of hours. These tests are carried out on samples using dedicated boards. Test equipment is needed to measure parameter drift before stress and after stress.

12.2 Connectivity test

Connectivity testing is the first step of each test program. It verifies the contact of every probe on the pads. Usually the connectivity test probes a parasitic diode (for instance a drain bulk diode). A typical connectivity test simply measures the substrate diode pulling a pin negative (versus ground or VSS).

If the connectivity test fails this can be caused by bad adjustment of the probes or oxidation of the pads.

For engineering samples, that do not yet have to fulfill high quality standards it may be acceptable to lift the probes and bring them down on the pads again to establish a contact.

In production this usually isn't permitted because multiple probe marks on the pads degrade bonding quality later.

12.3 Scan test

Scan test is a pure logic test. Analog functions are disconnected from the logic and a wrapper holds all the interfaces in a defined state to make the scan test run independent of any analog stimuli.

During scan test the flip flops of a logic and all the gates are reordered such that the logic becomes a large shift register. During scan test a pattern is shifted through this register in a way that is tests as many flip flops and logic gates (that now are placed between the flip flops) as possible.

Scan path insertion usually is done by the software used for logic synthesis. Then the test pattern is run by the logic simulator. The simulation provides the expected result (output pattern) of the scan test.

On the tester the same stimulus (input pattern) is applied and the output pattern is recorded. The measured output pattern is compared with the expected pattern coming from the simulation.

12.3.1 Test coverage

The test coverage describes how many of the theoretical errors will be detected. Ideally the test coverage should be 100%. In reality there always are cases that can't be detected. Especially logic that is fragmented into many little pieces (mixed signal chips) requires a lot of multiplexers. At the end of the day we run into the problem that catching the last gate requires a lot of multiplexers that possibly can't be tested themselfes by the scan test thus reducing test coverage again!

12.3.2 Testing at speed

A second problem is testing at speed. Modern logic technologies are much faster than the testers used. Passing the scan test at low speed doesn't guarantee that the chip will work at speed. (A poor contact or via may still work at low speed, but at high speed in combination with wire capacities it may become a fatal low pass.)

Even if the tester supports speed up to several GHz we will face the problem that the bond wires and the inductace of the probe card limit the test bandwidth.

In most cases the scan test is designed to operate at the same clock frequency that is used in normal operation too. Even if this clock frequency is significantly below the frequency the technology is (or should be) capable of.

12.4 Joint Test Action Group (JTAG)

JTAG is a more general, more standardized approach to chip testing than a simple scan chain. The scan chain however can be part of the JTAG test infrastructure and may use the same ports as the JTAG.

In addition to just scanning the logic with automatically generated pattern JTAG in most cases gives a well documented access to registers inside the chip and permits programming the chip. Most of the JTAG test method is documented in IEEE Std. 1149.1.

Since JTAG is mostly used for board level testing there is a standardized test access port (TAP). The TAP is a serial interface similar to an SPI.

12.4.1 JTAG modes

In most implementations JTAG offers two modes:

functional mode: In normal operation the JTAG doesn't disconnect any function of the chip. At maximum it can be used to monitor internal states (optional). But writing to the chip in functional mode is impossible. The dashed connections between the logic and the periphery are closed.

test mode: In test mode the logic is disconnected from the normal I/Os This allows:

1. Test of the logic like a normal scan test. The I/Os are passive during this scan test because the boundary scan cells disconnect the peripheral functions. This way scan test can be part of a JTAG. The dashed connection between the logic and the periphery are opened.

2. Stimulation of the peripheral functions. The JTAG interface has direct access to the signals driving the I/O cells of the chip. This method can be used to test the I/O functions and possibly the load conditions on the board (provided the I/O cell have means to read back the state of the board wires). The dashed connections between the logic and the peripheral functions are at least partly opened.

3. Modification and/or reading of programming registers (ISP - in system programming and ISC - in system configuration). These registers hold control bits such as trim values or mode settings. In some cases ISP may give access to program code memories. Now the dashed bridges are closed.

4. Execution of code coming from the TAP instead of using hard wired microcode or programmed code sitting in a non volatile memory. Dashed bridges are closed.

5. Bypass: For debugging the scan hardware many chips provide a bypass mode connecting TDI (test data in, pin sdi) directly to TDO (test data out, pin sdo).

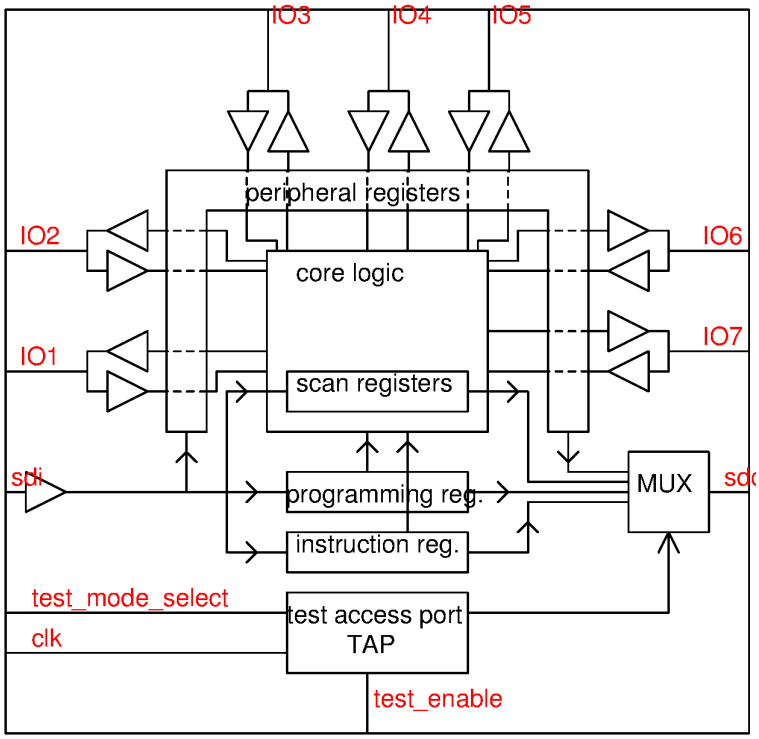


Figure 12.1: Concept of a JTAG interface

Running a JTAG test there always is a stimulus and an expected response. Many chip suppliers offer a so called Boundary Scan Description Language (BSDL) file. This is a standardized language used to describe the boundary scan. BDSL is a subset of VHDL.

In most cases the registers of several devices can be arranged in a daisy chain. So from board point of view it looks like a very long shift register.

12.4.2 JTAG testers

Many testing companies offer JTAG test interfaces that can be driven using SVF-files (Serial Vector Format). This is a kind of assembler language to control a JTAG tester. Here are some example commands:

Table 56: List of JTAG commends

command	description
ENDDR	specifies end state after DR scan
ENDIR	specifies end state after IR scan
FREQUENCY	maximum frequency
HDR	Header Data Register
HIR	Header Instruction Register
PIO	parallel I/O (multiple scan chains)
PIOMAP	Maps PIO pins to logical pins
RUNTEST	runs test for a specified number of clocks
SDR	Scan Data Register
SIR	Scan Instruction Register
STATE	force a specified state
TDR	Trailer Data Register
TIR	Trailer Instruction Register
TRST	Test Reset

12.4.3 JTAG connectors

There is a standardized connector for JTAG. During JTAG test the signals are driven from a tester or observed by a tester. In application mode most of the inputs have to be tied to ground by resistors (soft GND) to prevent unintentional activation of the JTAG test mode.

Table 57: JTAG 20 pin pinout

use	no	no	use
input	1	2	input
input	3	4	soft GND
TDI	5	6	soft GND
TMS	7	8	soft GND
TCK	9	10	hard GND
input	11	12	soft GND
TDO	13	14	soft GND
High	15	16	soft GND
input	17	18	soft GND
input	19	20	hard GND

Some manufacturers however use a reduced connector:

Table 58: JTAG 10 pin pinout

use	no	no	use
TCK	1	2	GND
TDO	3	4	
TMS	5	6	NRST
VCC	7	8	
TDI	9	10	GND

12.5 Built in self test (BIST)

A built in self test is used to test a limited set of functions of a chip either during booting or in special test modes. It relies on a certain minimum functionality such as the availability of a reference voltage. Due to this dependence it can't replace testing of a chip in production. Nevertheless built in self test can reduce the test effort to be done in the production line. Furthermore built in self test can be used to detect chip failure in the field.

Typical examples of built in self test are:

Table 59: Typical built in self tests

test	usage	remark
parity checks	tests the reliability of a memory	
marginal read test	Reads the content of a memory with modified thresholds having reduced margin to detect NVM fails before the real read path fails.	
supply test	internal supplies are compared to a reference voltage	
using ADCs for self test	Internal signals are tested versus expectations using an ADC and an analog MUX	ADC can also be used to test external components
Internal stimulation of functions with test cases	Functional test of signal paths and processing of signals inside a system chip	
Test of external components	During BIST current sources can apply currents to external components and measure voltage drop to test if the application circuit is present and correct.	Often using ADCs to read results.

Most of these tests consists of two parts.

1. During production test the refences, clocks and the test circuits must be tested (usually bandgap and ADCs or DACs)

2. Once the correct function of the references and test circuits is proven other functions of the chip can be tested using the BIST and - if available - the internal DACs and ADCs.

12.6 Analog test bus (ATB)

Modern chips hold a lot of analog functions that are buried deep inside the chip and that are not connected to a pad. Testing these functions only with functional tests can be extremely difficult if not impossible.

Example: A switchmode power supply has a regulation loop controlling the PWM. A functional test of the feedback loop would require sweeping the the feedback signal and monitoring the power transistor. To capture the AC parameters of the feedback amplifier this sweep would have to be done on speed. The nightmare of every product engineer!

Building an analog test bus the following requirements should be considered:

1. Usually you need simultaneous access to an analog signal and a digital signal to test comparators.
2. Testers have a high capacity compared to on chip capacities. Therefore analog signals should be buffered before leaving the chip.
3. Buffers must be fast enough for the fastest analog signal to be tested. (For this reason often the analog test bus is limited to 3.3V or 5V to use fast low voltage components)
4. Buffers (e.g. if designed for speed) have an offset. There must be a possibility to measure the buffer offset.
5. For current measurements and resistor measurements it must be possible to bypass the buffer
6. If differential signals are measured two analog test bus channels are required. (since chip ground is not always coincident with board ground (bond wire impedance) using a two channel analog test bus is strongly recommended)
7. Analog test bus may be useless for testing internal signals during EMC tests. (You don't know whether you measure a failure of the buffer or of the internal function during EMC tests)

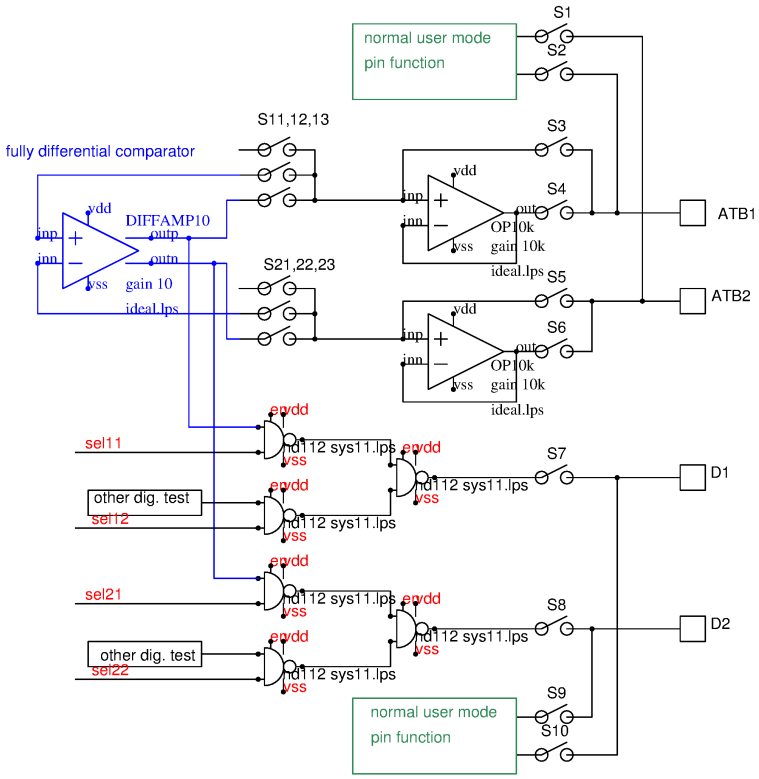


Figure 12.2: Concept of an analog test bus used to test a fully differential comparator

Typically the test starts calibrating the test buffer.

1. connect an internal voltage directly to the analog test bus closing S3, S5.
2. measure internal voltage unbuffered.

3. open S3, S5, close S4, S6 to measure the same voltage buffered.

4. measure buffered voltages.

5. calculate and store offset of the buffers.

Now the offset is known other functions can be measured using the buffers.

The ATB can be used to stimulate internal functions too.

1. To stimulate the input of analog functions open S4, S6 and close S3, S5

2. ramp the voltages on the ATB pins and observe result on the digital outputs D1,D2

If the digital test bus is routed through a synthesized logic all digital signals will have a latency depending on the clock frequency of the logic and the number of sync flip flops used. (Most designs use two sync flip flops both triggering at the same clock edge. So the synchronization delay will be two clock periods and the latency adds one more clock period. The total delay will be between 2 clock periods and 3 clock periods).

For slow functions this latency can be accepted. For high speed functions this error may be unacceptable. In this case the digital test bus must be hand crafted bypassing the synchronous logic.

12.6.1 Design considerations

Design hierarchy: The test switches S11, 12, 13 and S21, 22, 23 should be placed as close to the analog function to be tested as possible. This placement avoids parasitic capacities at the analog function to be tested. To enforce proper layout an analog test bus isn't a compact block but a circuit distributed over many cells.

The test buffer and switches S1..S10 usually is located somewhere in the padring.

The select signals can either be generated in a local decoder inside the analog function to be tested or in the logic. Each approach has it's pros and cons.

- Generating the select signals using local decoders reduces the number of wires in the top level routing. The decoder itself is not included in the SCAN test. The decoder can be inside the 3.3V or 5V supply domain of the analog part to minimize the number of level shifts.

- Placing the decoder in the synthesized logic leads to more top level routing (one wire for each select signal). The decoder will automatically be included in the SCAN test. Now we need a level shift for every select signal. (Example: 4 test addresses lead to 16 select lines!)

Usually the logic uses 1.2V or 1.8V transistors. So every signal running from the logic to the analog part will require a level shift. The level shifts are not part of the SCAN test either. Minimizing the number of level shifts (even if we have to sacrifice a little bit of digital test coverage) is a valid argument too!

12.6.2 A practical example

As case study of test modes let's have a look at the regulation loops of a current mode booster. The simplified regulator is shown in the following figure.

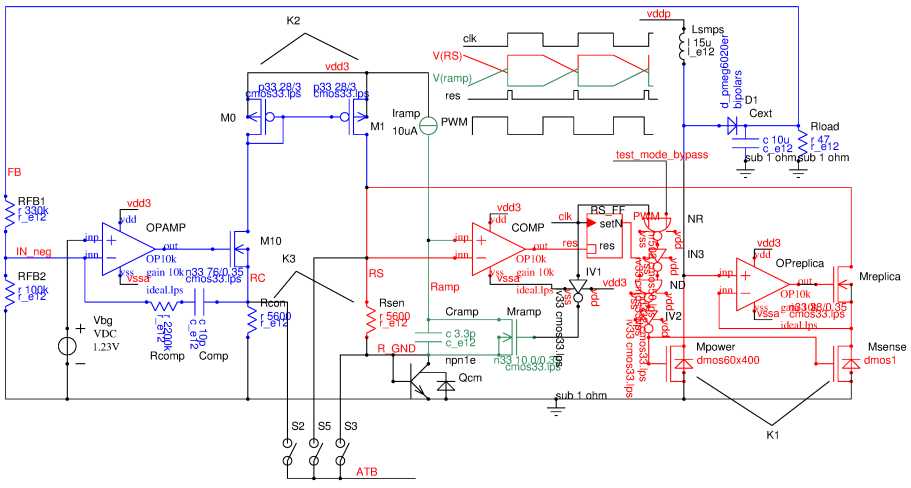


Figure 12.3: A boost converter with current mode regulation

The boost converter has two regulation loops. The fast loop is the current mode loop drawn in red color. The slow loop is the voltage mode loop drawn in blue color. The voltage mode loop measures the output voltage of the

regulator and defines the target current for the current mode loop. The target current is represented by the current provided by the transistor M1.

The current loop measures the current flowing in the power transistor Mpower (using the operation point replica amplifier OPreplica and replica transistor Mreplica). This measured current is subtracted from the target current provided by M1. The voltage across Rsen represents the difference between the target current and the current flowing through the load (inductor Lsmps).

If the inductance Lsmps is very high there wouldn't be any ramp signal at net RS. This would lead to a 2 point regulation oscillating below the clock frequency. To avoid 2 point regulation a ramp is forced on the positive input of the comparator (signal Ramp). This ramp guarantees stable operation independent of the load inductance Lsmps.

The behavior of the regulator in application depends on the loop gain of both regulation loops. Measuring regulator performance in a test environment with fast switching currents in the range of several amperes is very difficult to impossible. Therefore the parameters of the regulation loop are determined by test modes.

Testing the ramp generator: The net Ramp is very sensitive to slight changes of the capacity. Therefore signal Ramp on intention is not connected to any test bus. The ramp generator is tested indirectly observing the PWM duty cycle while switch S5 is closed. Via the analog test bus the voltage at net RS is forced. The gain of the PWM generator (in terms of $dD/dV(RS)$) is measured changing the voltage and observing the change of the duty cycle.

To avoid any influence of the Ron of switch S5 it is suggested to have no current flow in the power transistor and the target current transistor M1. Best way to accomplish this is disconnecting the drain of the power transistor from the load and forcing the feedback voltage V(FB) higher than the regulation target (for instance 5.4V if the regulator is designed for 5V and has +-200mV tolerance).

To test the range of the ramp can be verified forcing V(RS) below the Vbe of Qcm (Then we observe the minimum duty cycle defined by the propagation delays of the circuit).

The maximum duty cycle can be measured forcing V(RS) to vdd3. (This is the duty cycle of the clock signal at net clk).

Test of the signal swing at node RS: To test the signal swing the voltage source at the analog test bus is disconnected. The minimum voltage at RS can be determined forcing the feedback signal higher than the regulation target (in our example of a 5V booster 5.4V).

The maximum voltage at RS can be determined forcing V(FB) below the regulation target (for instance 4.6V or 3.3V). The maximum voltage is stored as V_{RSmax}.

Test of the aspect ratio of Mpower and Msense: The ratio K1 between the power transistor and the sense transistor can be measured sweeping the current through Mpower and measuring the current flowing into Msense. To do this measurement switch S5 is closed and the voltage at the analog test bus is forced to about 1V. Tho avoid any influence of the voltage loop the node FB again is forced above regulation target (5.4V). The current measured at the analog test bus is the sum of the currents flowing through Rsen and the current flowing into Msense. Doing this measurement at two different currents flowing into the power transistor the current through Rsen cancels.

$$K1 = \frac{I_{ATB1} - I_{ATB2}}{I_{power1} - I_{power2}} \tag{12.5}$$

Test of the resistor Rsen: To measure the resistor Rsen we need to know the voltage at net R_GND while the maximum current is flowing through the resistor. S5 is opened, S3 is closed now. As before the feedback is forced below the regulation target (for instance 4.6V or 3.3V). The result of the measurement is stored as V_{RGND} .

To measure the current flowing through Rsen the switch S3 is opened again and S5 is closed. Now the analog test bus ATB is forced to 0V and the current flowing out of the bus is measured. The result is stored as I_{Rsen} . Now the value of the resistor Rsen can be calculated:

$$Rsen = \frac{V_{RSmax} - V_{RGND}}{I_{Rsen}} \tag{12.6}$$

Test of the resistor Rcon: Rcon together with the current mirror M0, M1 and Rsen determines the gain of the voltage loop. To measure Rcon the transistor M10 must be turned off. This can be achieved forcing V(FB) above the regulation target (5.4V in our example). The output of OPAMP drops down to 0V. Now the resistor Rcon can be measured via the analog test bus switch S2.

Test of the maximum current through Rcon: Once the value of Rcon is known the maximum current flowing through Rcon can be calculated from a measurement of the voltage at net RC while the signal V(FB) is forced below the regulation target. The voltage is stored as $V_{RC_{max}}$. The current calculates as:

$$I_{RC_{max}} = V_{RC_{max}}/Rcon \tag{12.7}$$

Calculation of the ratio of the current mirror M0, M1: Knowing the currents through Rsen and Rcon the ratio of the current mirror can be calculated.

$$K2 = I_{Rsen}/I_{RC_{max}} \tag{12.8}$$

Calculation of the gain of the current loop: The gain of the current loops can be regarded as the change of the duty cycle with respect of the change of the current. The gain to node RS simply is

$$gm_{RS} = Rsen/K1 \tag{12.9}$$

Since this gain has the unit of V/A this is a transconductance. At RS the voltage is converted into a duty cycle of the PWM. So the loop gain until the output of the PWM generator becomes:

$$gain_{currentmode} = \frac{dD}{dV(RS)} * \frac{Rsen}{K1} \tag{12.10}$$

Since the duty cycle doesn't have a unit the expression $gain_{currentmode}$has the unit 1/A. The gain of the current loop together with the external components Lsmps, Cext, Rload and the supply voltage vddp determines the stability of the current loop.

Calculation of the voltage loop gain: The gain of the voltage loop is determined by the divider RFB1, RFB2, the gain of OPAMP and the transfer function from Rcon to the PWM.

$$gain_{voltageloop} = \frac{RFB2}{RFB1 + RFB2} * gain_{OPAMP} * \frac{Rsen}{Rcon} * K2 * \frac{dD}{dV(RS)} \tag{12.11}$$

There still are two unknown parameters: the ratio RFB2/(RFB2+RFB1) and the gain of the OPAMP.

The resistor ratio is simply the ratio of the output voltage and the reference voltage.

$$\frac{RFB2}{RFB2 + RFB1} = \frac{V_{bg}}{V_{out}} \tag{12.12}$$

This can be taken from measurements of the bandgap and the regulation loop. (If you don't want to use a regulation loop replica simply observe the voltage at node RC via switch S2 and sweep V(FB) to find the threshold of OPAMP operated as a comparator.)

The gain of the OPAMP is a bit more difficult to determine. In the tests described the OPAMP was simply tested as a comparator applying 5.4V and 4.6V to the node FB. This only is a go -nogo test of OPAMP but does not yet provide parameters. Furthermore the amplifier OPAMP is operated as an integrator. So the gain decays with increasing frequency.

Closing S2 an AC test of the OPAMP is possible establishing a closed loop. This test however is more suitable for lab testing than for testing on a wafer prober. In the following figure the components inside the chip are colored blue while the components on the test board are colored black. The amplifier OP together with resistors R1 to R4 and Vos is an analog adder shifting the output voltage of the amplifier up to the regulator output voltage level. The loop is stimulated by an AC source. The AC transfer function can be measured at the analog test bus ATB.

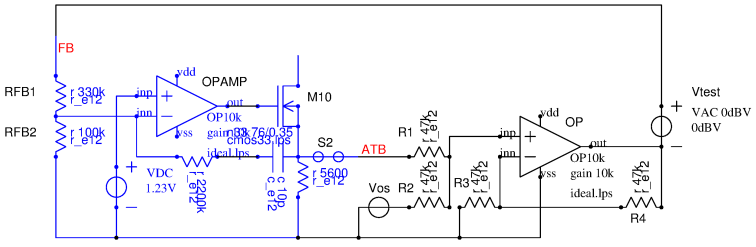


Figure 12.4: Proposal how to measure the parameters of the voltage loop error amplifier using classical lab equipment

This setup probably is difficult to implement in a production test environment. There it might be easier to apply a rectangular pulse at pin FB and to observe the slew rate at the analog test bus. The slew rate test is less accurate but probably good enough to prove the amplifier and the feedback network is working correctly.

12.7 Power transistor test modes

Most testers have a limited current capability in the range of 100mA to 200mA for the analog PMU (power measurement unit). Higher currents can be tested operating several PMUs in parallel. Paralleling several PMUs may lead to regulation loop instability and the frequency compensation has to be tuned to lower cutoff frequencies. For this reason the higher the current the slower the test will get.

As long as possible for test effort reasons most product engineers prefer working with a single PMU per node rather than using several in parallel.

Typical settling times of a single PMU (no parallel operation) is in the range of 1ms to 3ms.

12.7.1 Sense and force

Testing with high currents leads to voltage drops on the wires. For this reason sense and force testing is required. For the design of a chip it means that the pads must either be designed big enough to accommodate several probes or multiple pads have to be made available for the sense probe.

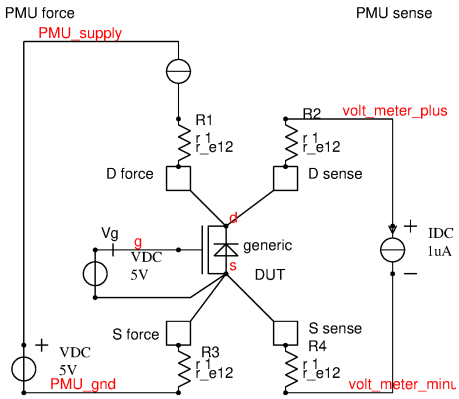


Figure 12.5: Sense and force concept

The PMU is divided in a PMU force (on the left side) and a PMU sense (on the right side). Resistors R1 and R3 carry the high force current. Resistors R2 and R4 are (almost) current-less. So a drop on the resistors R1 and R3 doesn't affect the voltage measured with the voltmeter on the sense side.

Usually a certain minimum current intentionally flows on the sense side as well to identify broken probes or cables. This current usually is chosen so low that the impact on the measurement is negligible.

Ideally the volt meter on the sense side is a fully floating instrument permitting fully differential measurements.

In stead of forcing current and measuring voltage it is also possible to force voltage and measure current. This approach requires a regulation loop measuring the differential voltage at the DUT (device under test) and a controlled voltage source providing the forced voltage plus the drop over the resistors. The current is measured by an amperemeter. (represented by the 0V source in the drain path).

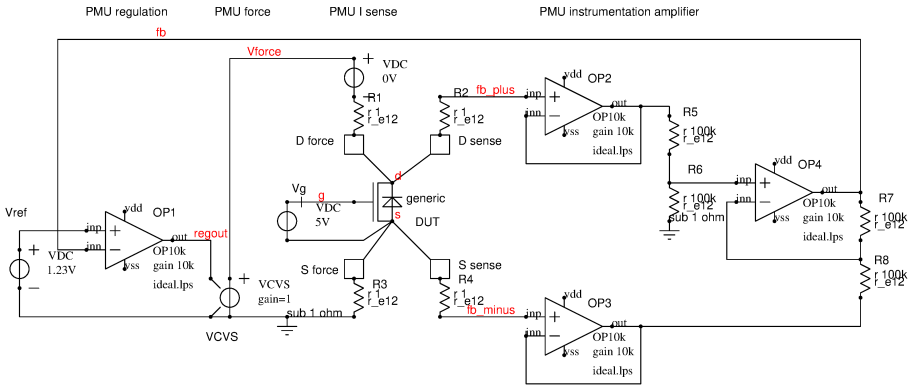


Figure 12.6: forcing voltage measuring current

The circuit forces Vref between the nodes D_sense and S_sense and allows measuring the current at the amperemeter. The accuracy of this loop mainly depends on the matching of the resistors R5 to R8 of the instrumentation amplifier. The dominant pole of the loop usually is implemented in OP1 and must be magnitudes below the first pole of the instrumentation amplifier and the first pole of the power stage represented by the voltage controlled voltage source VCVS.

12.7.2 Quasi differential measurements

Fully differential (floating) volt meters are a sparse resource on most testers. Floating instruments are more complex and more expensive that instruments related to the common tester ground. For this reason it is tempting to replace one fully differential instrument by two single ended volt meters and then calculate the differential voltage.

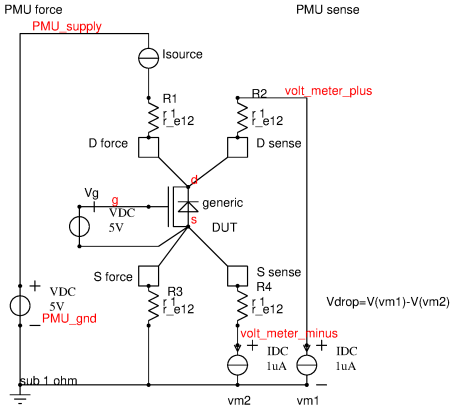


Figure 12.7: quasi differential measurement

The drawback of this approach is that the measurement error is the sum of the absolute errors of the two volt meters vm1 and vm2. At minimum this is twice the quantization error of the two instruments. For low side drivers both volt meters operate in a low range and the absolute errors are acceptable. Using quasi differential measurements for high side drivers can lead to dramatic errors.

In the following example let's assume each volt meter has an accuracy of 0.1%. The supply voltage is intended to be 12V. $Ron = 0.1\Omega$, Isink=0.1A.

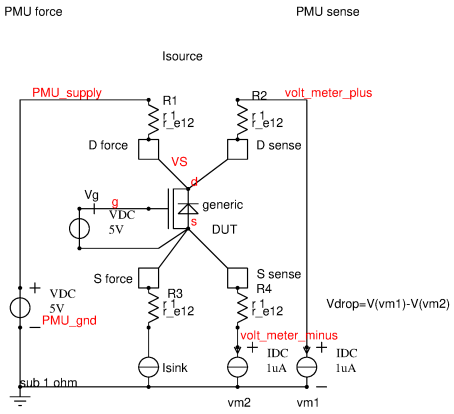


Figure 12.8: quasi differential measurement of a high side driver

To start let's assume the voltage at node is already adjusted correctly.

$$V(PMU_{supply}) = VS + I_{sink} * R_1 = 12.1V$$

The voltage measured by vm1 is $12V \pm 12V * 0.1\%$. So the uncertainty of instrument vm1 is already 12mV.

The voltage measured by vm1 is $12V - R_{on} * I_{sink} \pm (12V - R_{on} * I_{sink}) * 0.1\%$. The uncertainty of vm2 is 11.99mV

The difference $V(vm1) - V(vm2) = 10mV$with an uncertainty of 23.99mV! Using a quasi differential measurements obviously will fail.Testing high side drivers requires real fully differential instruments or connecting the tester's common ground node to R1 instead of R3.

12.7.3 Ron test

Ron tests require a calculation

$$R_{on} = V_{drop}/I_{test}$$

Testing low resistive transistors in the range of fractions of Ohms leads to fairly low drop voltages. At the same time the resolution of the tester (quantization error) is limited. Typical quantization errors of testers are +-0.5mV. The achievable accuracy using a fully differential approach is:

$$Err_{quant} = V_q/(I_{test} * R_{on}) \tag{12.13}$$

Using our example of a 100mA test current and a 0.1Ω transistor the relative error cause by the limited resolution of the volt meter is

$$Err_{quant} = 0.5mV/(0.1A * 0.1\Omega)$$

We can't test more accurate than $\pm 5\%$. The current sink will contribute an additional error depending on the accuracy of the source. Using a quasi differential measurements will at least double the error (for a low side test). Using a quasi differential approach a high side switch test (with tester ground at the negative node of the supply) will yield unusable results because the spread quickly gets bigger than the measured value of Ron. There are 3 ways out of the problem:

1. Increase test current
2. Use a volt meter with finer resolution
3. Increase the resistance of the transistor you want to measure.

Number 1 and 2 are costly (equipment cost, multiple tests with a lot of averaging, longer test time because precision volt meters usually are slower). Number 3 means we test only a fraction of the device and assume the untested part of the device will match. This is called a replica transistor test.

12.7.4 Replica transistor test

Basic idea of testing Ron of a scaled replica transistor is to increase Vdrop while still using a reasonably fast PMU. The Ron of the real power transistor is assumed to correlate according to the aspect ratios of the replica transistor and the power transistor.

$$R_{oninterpolated} = R_{onreplica} * \frac{W_{power} * L_{replica}}{W_{replica} * L_{power}} \tag{12.14}$$

The assumption here is that the replica transistor is tested at the same gate overdrive as the power transistor will have in application. In addition the replica transistor test relies on proper matching of the power transistor and the replica transistor that is intended to mimic it's performance. Transistor matching depends on technology and the transistor type used.

Classical CMOS transistors that have no technology tricks like halo implant etc. usually match well following the Pelgrom approximations. For large replica transistors (n ranging from 10..100) the offset between the power transistor and the replica transistor can be neglected because the gate area is high ($V_{os} \sim 1/\sqrt{W * L}$). If the ratio n becomes large and the replica transistor has a small gate area the mismatch will contribute significantly to the total error.

Transistors with halo implant match reasonably for a high gate overdrive. If they are operated close to the threshold halo transistors don't match well!

DMOS transistors have an effective channel length that is defined by the lateral diffusion of the bulk doping. Often this lateral diffusion is only in the range of fractions of a $\mu m$and matching tends to be poor. Homogeneous field distributions inside the drift zone (under the field plate) add further sources of mismatch. Trench isolations can create mechanical stress additionally affecting the matching in a negative way!

Since the impact of halo implants and the homogeneous fields in DMOS transistors are very difficult to predict the calculations shown only apply to classical CMOS transistors which is the most optimistic case.

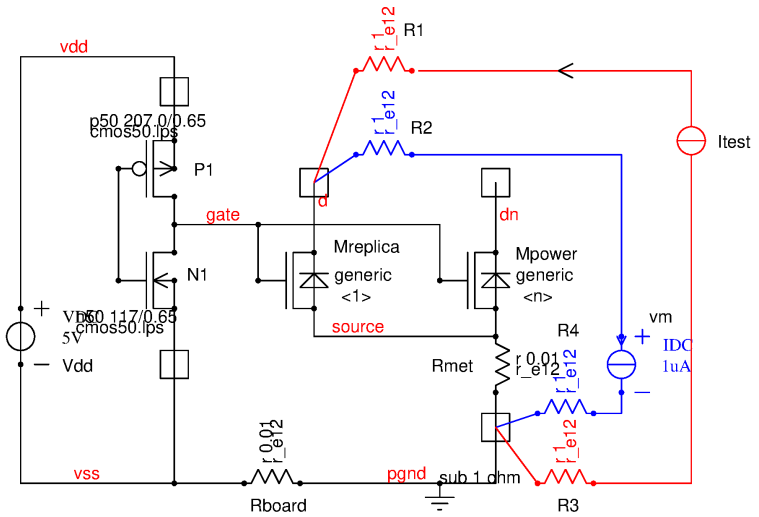


Figure 12.9: Replica transistor testing circuit

In the above shown circuit the power transistor is n-times larger than the replica transistor. Both gates are driven from the same driver stage (P1 and N1). The driver stage has a separate supply pin pair vdd and vss. Between vss and the power ground pgnd there is a connection on board level represented by resistor Rboard. The metal paths inside the power stage (and possibly the bond wire) are represented by Rmet.

During the test the current Itest is flowing through Rmet while in real application a n times higher load current Iload is flowing through Rmet. During test mode the voltage drop over Rmet is:

$$V(R_{mettest}) = I_{test} * R_{met}$$

In application the drop over Rmet becomes:

$$V(R_{metapp}) = I_{load} * R_{met}$$

Since we want to test the replica transistor at the same current density as the power transistor will be operated it is reasonable to chose:

$$n = \frac{I_{load}}{I_{test}}$$

As a consequence the drop over Rmet in application becomes:

$$V(R_{metapp}) = n * I_{test} * R_{met} \tag{12.15}$$

The gate overdrive during application is:

$$V_{gseff} = vdd - V_{th} - V(R_{metapp}) = vdd - V_{th} - n * I_{test} * R_{met} \tag{12.16}$$

While in test mode the gate overdrive is:

$$V_{gseff} = vdd - V_{th} - V(R_{mettest}) = vdd - V_{th} - I_{test} * R_{met} \tag{12.17}$$

Between application and test the gate overdrive changes by:

$$\Delta V = R_{met} * (I_{load} - I_{test}) = R_{met} * (n-1) * I_{test} \tag{12.18}$$

Due to the voltage drop in the source metalization the replica transistor will be tested at a higher gate overdrive than the power transistor will be operated in application! The assumption of an equal gate voltage is broken. Now we have to calculate the error caused by this deviation.

The ON resistance of a CMOS transistor is determined by the aspect ratio W/L, the carrier mobility μ and the gate overdrive V_{gseff}. In addition source spread resistance R_{ssp} and drain spread resistance R_{dsp} will contribute. Drain spread resistance and source spread resistance contribution can be minimized using exactly the same finger design for the main transistor and the replica transistor. Ideally the power transistor should simply be designed placing n modules of the replica transistor (same finger length, same number of fingers, same ratio of inner fingers and edge fingers).

$$R_{on} = \frac{L}{W} * \frac{t_{ox}}{\mu * \varepsilon_{sio2}} * V_{gseff}^{-1}$$

Well, nothing new. We had this equation in section 4.6.1. Since we only have little changes of V_{gseff}we can linearize this equation.

$$\frac{dR_{on}}{dV_{gseff}} = -\frac{L}{W} * \frac{t_{ox}}{\mu * \varepsilon_{sio2}} * \frac{1}{V_{gseff}^2} \tag{12.19}$$

If we insert the difference of the voltage drop over the metal path Rmet between normal application and replica test mode we can calculate the error.

$$\Delta R_{on} = \frac{R_{met} * (n-1) * I_{test}}{(vdd - V_{th})^2} * \frac{L}{W} * \frac{t_{ox}}{\mu * \varepsilon_{sio2}} \tag{12.20}$$

Since we are only interested in the relative error let's divide this expression by the R_{on} we would get in an ideal design with superconducting source metal.

$$err_{rel} = \frac{\Delta R_{on}}{R_{on}} = \frac{R_{met} * (n-1) * I_{test}}{vdd - V_{th}} \tag{12.21}$$

The equation can be rewritten:

$$err_{rel} = \frac{R_{met} * (I_{application} - I_{test})}{vdd - V_{th}} \tag{12.22}$$

Let's calculate an example:
vdd=5V, $R_{met} = 10m\Omega$,n=40 (application current 4A), $I_{test} = 0.1A$, $V_{th} = 1V$:

$$err_{rel} = \frac{0.01\Omega * (40-1) * 0.1A}{4V} = 0.975\%$$

This is a systematic error. The deviation means that due to the change of the gate voltage between the application and the test mode the test mode looks 0.975% lower resistive than the power stage really is.

(Note: $R_{met}$is the sum of the metal path on chip and the bond wire resistance! $10m\Omega$ is an extremely low value that can only be reached using massive copper plates and many bond wires in parallel.)

The worst case error is found when supply vdd is low, the threshold $V_{th}$is high and the metal resistance between the power transistor source and the ground (board ground if driver and power transistors use separate pins!) is high. To show the impact let's recalculate the same circuit using different conditions.

vdd=4.5V, $R_{met} = 15m\Omega$,n=40 (application current 4A), $I_{test} = 0.1A$, $V_{th} = 1.1V$:

$$err_{relworst} = \frac{0.015\Omega * (40-1) * 0.1A}{3.4V} = 1.7206\%$$

The change of conditions shows that the error of the scaled current will have a significant spread even if the transistors are matching perfectly well simply due to the impact of the path from the transistor sources to ground.

For test limits this means there must be a margin between the test of the scaled replica transistor and the target $R_{on}$of the power transistor.

$$R_{onreplicamax} \leqq (1 - err_{relworst}) * \frac{R_{onpowermax}}{K} \tag{12.23}$$

In this equation $R_{onpowermax}$ is the specified maximum on resistance of the power transistor and $R_{onreplicamax}$ is the maximum on resistance accepted as a pass testing the replica transistor.

Since the bond wires are part of R_{met} a change of the bond process impacts the margin we need for the test.

Estimation of worst cases: Typical values for the temperature dependence of threshold voltages are in the range of -2mV/K to -5mV/K depending on technology. Spread of threshold voltages usually is under good control in modern CMOS processes. At room temperature V_{th}should not vary more than ± 0.2. If nothing is known about the technology assuming $V_{th} = 1V$at room temperature (nominal threshold) -3mV/K is a reasonable starting point.

Temperature coefficients for aluminum and copper wires are about 0.0039/K. Technology spread of traces on a chip usually is in the range of $\pm 20\%$ (constant temperature). Production spread of bond wires (diameter of the wire, resistivity spread, wire length) is much lower in the range of $\pm 10\%$ mainly caused by wire length variations in the bond process. Modifications of the package (change of bond wire length) however can have a dramatic impact!

12.7.5 gate stress test

Gate stress testing intends to find weak spots in the gate oxide of large transistors. Such weak spots usually are caused by particles that lead to a locally thinner gate oxide. Gate stress tests usually consist of 3 steps:

1. Apply the nominal gate voltage and measure leakage for reference
2. Apply a stress voltage that is higher than the nominal operating gate voltage but lower than the voltage leading to instantaneous break down (usually between $1.5 * V_{gsnom}$and $2.5 * V_{gsnom}$). This voltage is expected to make the oxide that is thinned locally by the particle break down. Stress time usually is in the range of some ms to some seconds.
3. Repeat the leakage measurement using exactly the same test condition as in step 1.

If step 1 and step 3 show the same result the device is good. If the gate leakage increased in step 3 the part is suspected to have a particle in the gate oxide. Devices with particles in the gate oxide will be binned.

Gate stress test mode may never be accessed in application because it turns on all power transistors connected to this test mode at the same time. This may lead to immediate destruction in the application. There are several ways to prevent unintentional activation of the gate stress test.

- Use a separate pad to that will not be bonded to supply the gate stress voltage (this way gate stress test can only be done during wafer sort but not after packaging. Field returns can't be retested unless the package gets opened).
- Place a diode between the gate stress supply pin and the gates to be tested. In application the gate stress supply pin must be soldered to ground (we rely on reasonable usage of the chip. Field returns can be retested).
- Make gate stress test accessible only with a password that is kept as a company secret to prevent users in the field from accessing this possibly destructive mode. (Field returns can be retested. As long as the password is kept secret it is unlikely the chip can be destroyed in application)

BUT: Any kind of password protection may fail if parasitic transistors can bypass the logic! Bypassing encrypted functions of a chip using parasitic transistors (pulling pins below ground or above supply) is a common attack!

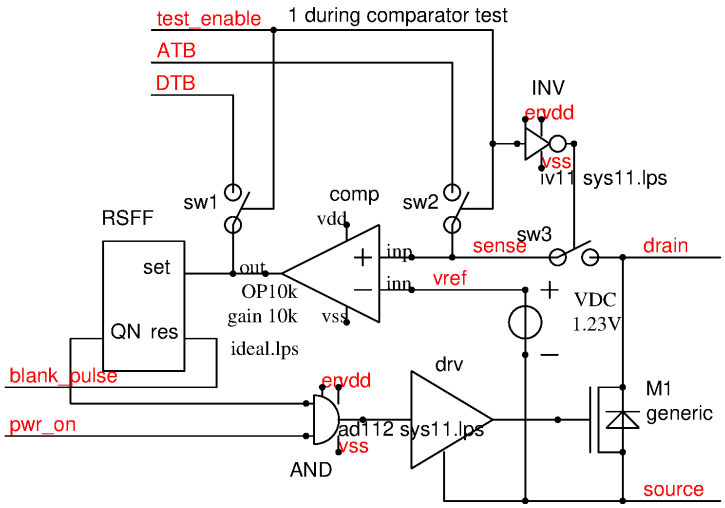


Figure 12.10: Overcurrent shut down circuit with test access

12.7.6 current limit test

Current limit test are the tests running power transistors at the highest possible currents. These tests are intended to find errors in the short circuit protection. Like R_{dson} tests the test current often is reduced using a scaled replica transistor. It must be distinguished between current limitation tests and overcurrent shut down tests.

Current limitations will limit the current but the power transistor remains on. A current limitation will produce a lot of heat on the chip and sooner or later will trigger a thermal shut down. To avoid driving the chip into thermal shut down current limitations usually must be tested with short pulses. Typical time constants are in the range of $100\mu s$ if the temperature sensor is close to the power transistor operating in current limitation. Ideally pulses of less than $10\mu s$should be used for current limitation tests to minimize thermal effects.

An overcurrent turn off test drives the power stage into a situation where the power transistor is turned off. Overcurrent turn off usually doesn't produce excessive heat because the power transistor has two possible states:

- Overcurrent is not reached. Power transistor is on and voltage drop over the transistor is low. Power dissipation still is in an uncritical range
- Overcurrent threshold was exceeded. The power transistor is turned off. There is a high voltage over the transistor but there is no more current flow. Power dissipation is low.

Reducing the currents:

To reduce the currents required the test can be split in two separate tests:

1. Measure the characteristic of the current sensing device.
2. Bypass current sensing device and test the measurement system

The current limitation or shutdown threshold will be calculated from the results of both tests. The calculated result will be influenced by the errors of both tests. The following schematic shows the concept of an overcurrent shut down with test bus access..

In normal application the switch sw3 is closed. The turn off threshold is

$$I_{turnoff} = \frac{vref}{R_{on}}$$

To test the circuit the test mode signal test_enable is set. Switch sw3 opens and sw1 and sw2 close. First the on resistance $R_{on}$is measured (this can be done using a scaled replica of M1 as described before). Then a voltage ramp is applied at the analog test bus ATB and the comparator response is observed at the digital test bus DTB. If the comparator has no offset signal DTB should have a rising edge crossing vref. If the comparator has an offset the rising edge at DTB will take place at a slightly different trip point vtrip. The shut down current is calculated from the measured R_{on} of transistor M1 and the measured trip point vtrip of the comparator.

$$I_{turnoff} = \frac{vtrip}{R_{on}}$$

The calculated turn off current has the sum of the relative errors of both measurements (error of vtrip and error of R_{on}).

12.7.7 void test

12.7.8 bond wire test

Standard contact test for single bond wires: Usually the existence of bond wires is tested first in the whole test flow. The most classical test simply relies on the current flow into the ESD protection or into substrate diodes. If the pin in pulled below ground or above the supply an increase of current flowing is expected at $VSS - V_f$ or above $VDD + V_f$. $V_f$is the forward voltage of the diodes being part of the ESD protection. Typical test currents are in the range of 1mA.

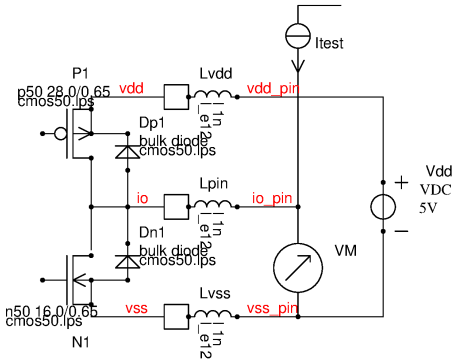


Figure 12.11: Test of bond wires using bulk diodes

If bond wire Lvdd is missing a positive current of Itest will pull node io_pin up until some structure between vdd and vss will reach zener break down.

If bond wire Lpin is missing Itest can pull signal io_pin to almost unlimited voltage levels.

If bond wire Lvss is missing a negative current of Itest will pull node io_pin down until some structure between vdd and vss will reach zener break down.

If all bond wires are present the voltage at the volt meter VM will always remain between $-1V < V(iopin) < vdd + 1V$.

Multiple bond wire test: Especially supply pins may need multiple bond wires to reduce the voltage drop or to carry the peak currents. Testing if all bond wires are present is not always trivial. The difference in resistance if one bond wire is missing is in the range of some $10m\Omega$. To make multiple bond wires testable these can be wired to different pins. The a test current is applied between the supply pins and the voltage drop is measured.

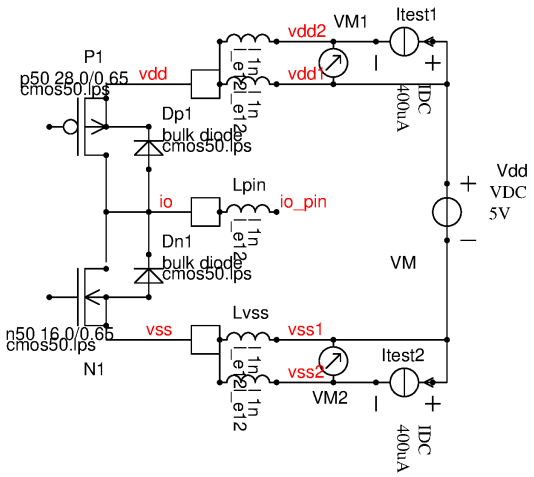


Figure 12.12: Test of multiple bond wires using multiple pins

If all bond wires exist the volt meters VM1 and VM2 are expected to only measure some mV or less. If one of the bond wires is missing the voltage will increase until the limitation of the test current source is reached.

Power transistor bonding test: Power transistors often require multiple bond wires because a single wire isn't able to carry the current. In the following circuit testing the bond wires isn't possible. If one of the bond wires is missing the difference measured is some $10m\Omega$ while the transistor has some $100m\Omega$. The spread of the transistor Rdson is wider than the change of resistance loosing one of the bond wires.

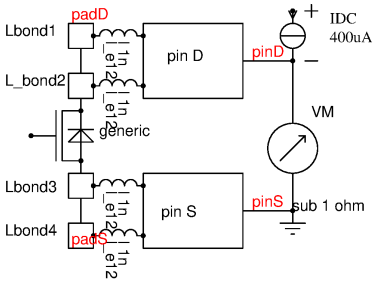


Figure 12.13: untestable double bonds

One way out is to split the transistor in multiple smaller transistors. If one of the bond wires is missing the measurement of the Rdson will yield double the expected good value.

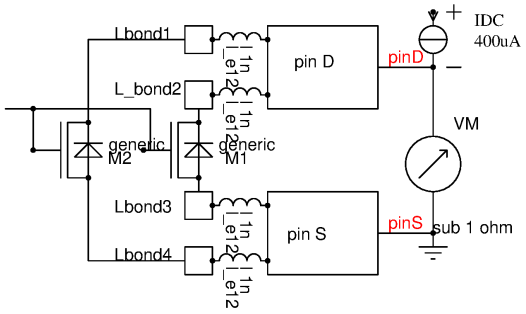


Figure 12.14: testable double bonds

This approach of segmenting the power transistor can be extended to more than 2 devices provided the following condition is satisfied:

$$6 * s(g) < \frac{g}{2 * n} \tag{12.24}$$

In this equation g is the conductivity of the complete power transistor. n is the number of segments. s is the standard deviation of the conductivity. Usually 6s is used to get a low number of false tested 'good' devices. The following plot shows the distributions of good devices and of devices with one missing bond wire.

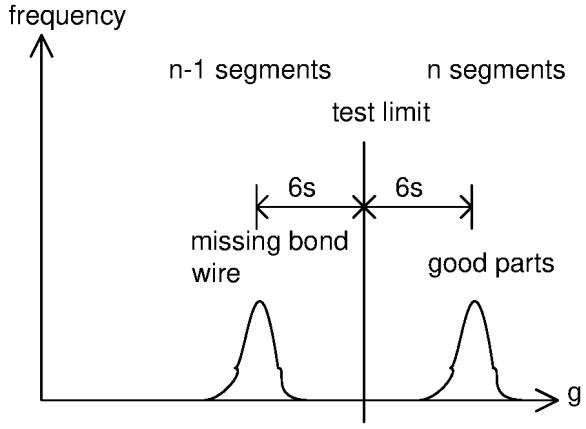


Figure 12.15: distributions of good devices and devices with missing bond wires

Usually this kind of test works well for 2 or sometimes even 3 parallel bond wires. Above 3 transistor segments the distributions start to overlap and good devices and devices with missing bond wires can't be distinguished in a reliable way anymore.

bond wire test using test modes: If the number of bond wires gets too high the only way out is using test modes to drive each segment of the power transistor individually. The following circuit has a normal operating mode with all power transistor segments operating in parallel (signal normal_mode=1) and a test mode (signal normal_mode=0). In test mode the select signals sel_1, sel_2, sel_3 are used to select each path individually. This mode is only used for testing the presence of the bond wires.

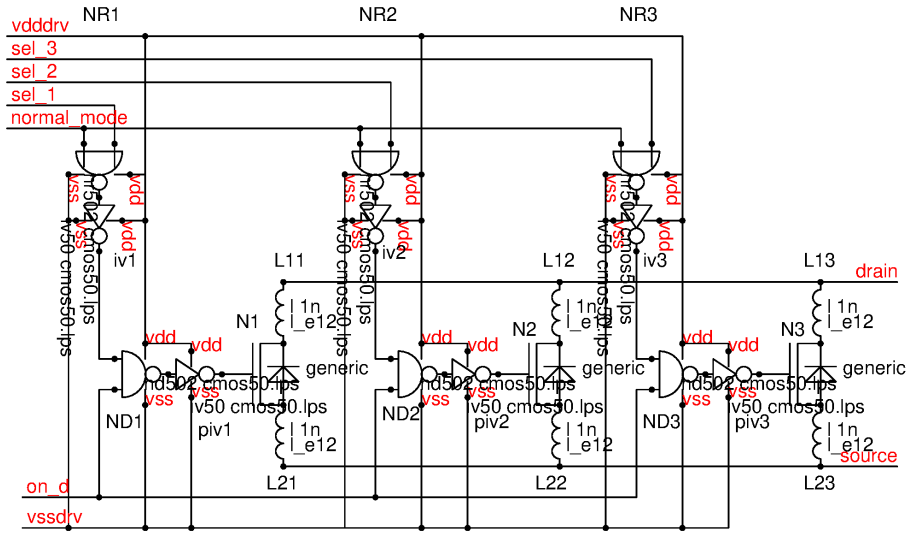


Figure 12.16: Power transistor split in 3 segments and test mode logic

Testing each path individually of course means the driver stage has to be implemented multiple times (piv1 to piv3) as well. The area penalty depends on the complexity of the driver stage. This may become a significant effort if the driver stage has features such as dV/dt regulation etc. ESD protection of multiple separated drivers and power stages may also become quite complex.

12.8 Amplifier testing

Classical phase margin tests cut open the feedback loop.

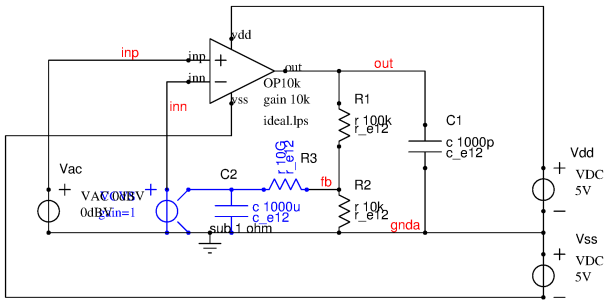


Figure 12.17: Open loop test setup for phase margin and gain margin

The resistor R3 and C2 together with the voltage controlled voltage source (blue components) cuts open the feedback loop. The low pass created by R3 and C2 has a cut off frequency several decades below the cut off frequency of the amplifier. R1 and R2 are the components that belong to the mandatory application circuit. C1 is the worst case capacitive load of the amplifier. If the input of the amplifier is high resistive the voltage controlled voltage source can be omitted.

In many circuits the amplifiers are embedded inside the closed feedback loop. High speed feedback amplifiers with integrated feedback network often don't allow opening the loop because even the pad capacity may already be a significant capacitive load of the feedback network destroying the phase margin. Such amplifiers must be tested inside the feedback loop. The best way doing this is to operate the amplifier with a rectangular signal. The overshoot of the output signal is an indication of the phase margin of the system.

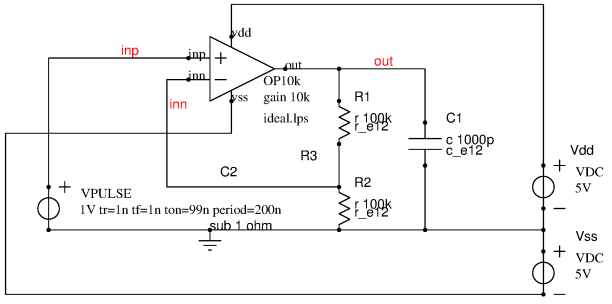


Figure 12.18: Closed loop amplifier test setup

Regarding the amplifier inside the closed loop as a second order system there is a usable correlation between the overshot of a small signal rectangular pulse and the phase margin at unity gain of the loop. The closed loop gain calculates:

$$gain_{closed} = \frac{gain_{amp}}{1 + \frac{R_2}{R_1+R_2} * gain_{amp}} \tag{12.25}$$

The system becomes unstable when the denominator becomes 0. To make life a bit easier the feedback network transfer function can be replaced by β.

$$\beta = \frac{R_2}{R_1 + R_2} = \frac{1}{k} \tag{12.26}$$

$$gain_{amp} = -\frac{1}{\beta} \tag{12.27}$$

In other words the amplifier must shift the phase by 180 degrees in excess of the inversion. Replacing the transfer function of the amplifier by a system of two poles the gain depending on the frequency s becomes:

$$gain(s) = \frac{gain_0}{(1 + \frac{s}{\omega_{p1}}) * (1 + \frac{s}{\omega_{eq}})} \tag{12.28}$$

The second pole ω_{eq} is an equivalent pole approximating a multiple pole system.

$$\frac{1}{\omega_{eq}} = \frac{1}{\omega_{p2}} + \frac{1}{\omega_{p3}} + ... \tag{12.29}$$

For $\omega >> \omega_{p1}$ the equation can be approximated by:

$$gain(s) = \frac{gain_0 * \omega_{p1}}{s * (1 + \frac{s}{\omega_{eq}})} \tag{12.30}$$

The product $gain_0 * \omega_{p1}$is called the gain bandwidth product GBW.

$$GBW = gain_0 * \omega_{p1} \tag{12.31}$$

$$gain(s) = \frac{GBW}{s * (1 + \frac{s}{\omega_{eq}})} \tag{12.32}$$

The closed loop gain becomes:

$$gain_{closed}(s) = \frac{1}{\beta} * \frac{1}{1 + \frac{s}{GBW} + \frac{s^2}{GBW * \beta * \omega_{eq}}} \tag{12.33}$$

$$gain_{closed}(s) = \frac{k}{1 + \frac{s}{Q * \omega_0} + \frac{s^2}{\omega_0^2}} \tag{12.34}$$

with:

$$\omega_0 = \sqrt{\beta * GBW * \omega_{eq}}$$

and

$$Q = \sqrt{\frac{\beta * GBW}{\omega_{eq}}}$$

The loop gain is the feedback factor β multiplied with the gain of the amplifier $gain_{amp}$:

$$\beta * gain_{amp}(s) = \frac{\beta * GBW}{s * (1 + \frac{s}{\omega_{eq}})}$$

The phase margin has to be calculated at the cross over frequency ω_t where the loop gain becomes 1.

$$\beta * gain_{amp}(\omega_t) = 1$$

$s = j\omega_t$ leads to expression:

$$1 = \frac{\beta * GBW}{j\omega_t * (1 + \frac{j\omega_t}{\omega_{eq}})} \tag{12.35}$$

Solving for the amplitude we get

$$\beta^2 * GBW^2 = \omega_t^2 * (1 + \frac{\omega_t^2}{\omega_{eq}^2}) \tag{12.36}$$

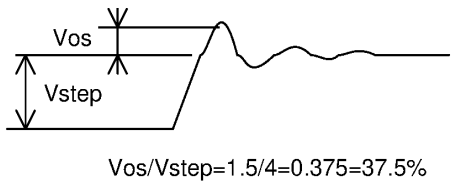


Figure 12.19: Definition of Vos and Vstep

Taking the square root and dividing by ω_{eq} yields:

$$\frac{\beta * GBW}{\omega_{eq}} = \frac{\omega_t}{\omega_{eq}} * \sqrt{1 + \frac{\omega_t^2}{\omega_{eq}^2}} \tag{12.37}$$

This is exactly the square of the quality factor Q. So Q can be expressed as

$$Q = \sqrt{\frac{\omega_t}{\omega_{eq}} * \sqrt{1 + \frac{\omega_t^2}{\omega_{eq}^2}}} \tag{12.38}$$

The resonance quality Q determines the overshot of the system if a step function is applied.

$$os[\%] = 100 * \frac{V_{os}}{V_{step}} = 100 * exp(\frac{-\pi}{\sqrt{4 * Q^2 - 1}}) \tag{12.39}$$

In this equation $V_{os}$is the overshot voltage and $V_{step}$is the height of the step. To test the phase margin in a closed loop configuration only little steps may be used to prevent the amplifiers from running into saturation (clipping of voltage, slew limited operation must be avoided). In most cases the quality factor Q isn't known. So the equation shown above isn't too helpful to determine the phase margin.

The phase margin of the system is:

$$phasemargin = -\frac{\pi}{2} - atan(\frac{\omega_t}{\omega_{eq}}) \tag{12.40}$$

To make life a bit easier here is a little octave script calculating the relationship of phase margin and overshot:

Algorithm 4 octave script used to calculate the relationship between phase margin and overshot

```
% Overshoot as a Function of Phase Margin
% x = wt/weq
x = .27:.01:1.0;
pm = 90-(180/pi)*atan(x);
q=sqrt(x.*sqrt((1+x.^2)));
os=100*exp(-pi./(sqrt(4*q.^2-1)));
plot(pm,os)
title('Overshoot');
xlabel('Phase Margin (degrees)');
ylabel('Percent Overshoot');
grid;
```

The script produces a nice graph of the overshot versus the phase margin. So a closed loop stability test can use the overshoot of the system to estimate the phase margin.

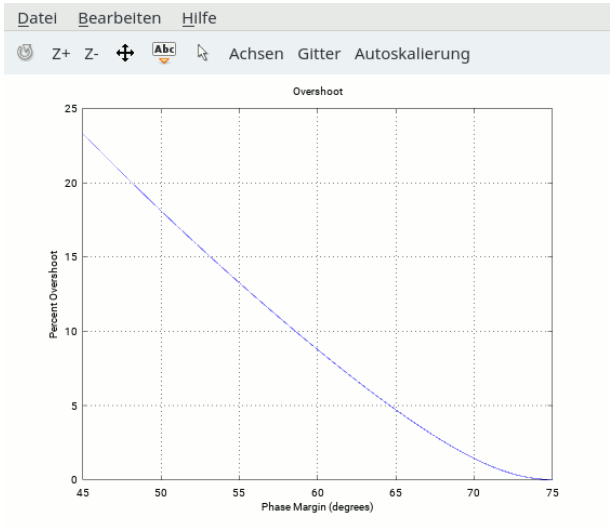


Figure 12.20: Overshoot versus phase margin plot

For most applications a phase margin of 60 degrees is desired. In extreme cases (high capacitive load) 45 degrees may be acceptable in the worst case corner.

12.9 Security rules

Chip manufacturers are liable for their products. For this reason test modes usually are protected.

Test modes like BIST and JTAG give access to internal registers of the chip (instruction registers, parameter registers etc.). Hardware giving access to these modes may under no circumstances be accessible without physical access to the board. JTAG connectors may not be accessible from public interfaces such as Ethernet, USB, WLAN, CAN, LIN. If access to a JTAG connector is possible from outside during user mode the whole system can be attacked!

Using scan test mode for code injection looks a bit more difficult because it rearranges the logic and at the end of the scan test the rearranged logic gets reconnected and in theory all registers should be overwritten. But to be honest I wouldn't trust all registers are really overwritten. Compared to JTAG code injection in scan mode seems less likely but even SCAN interfaces could allow an attack. Keeping SCAN interfaces separate from public interfaces is recommended.

Analog test modes may have features to disable protections (over current protections, over voltage clamps etc.) to simplify testing. Analog test modes may only be activated with interfaces on board level or tester level but not on public interfaces. If access is given from public interfaces (for instance Ethernet, USB, LIN, CAN) this public access can be misused to destroy devices that are expected to be well protected.

Concluding: test access may never ever be possible from public interfaces that allow access from remote. Violating this basic rule opens the door for cyber attacks.

Even communication on public interfaces should run with encryption to prevent unauthorized system access just by taping something like a LIN or CAN wire. If encrypted communication isn't possible the non encrypted bus must be kept local and separated from security critical functions by gate ways acting as a fire wall. (typical example: opening the doors of a car by simply taping the LIN wire of a torn off mirror of the car. Best protection: The mirror LIN may not be connected to the LIN controlling the door locks. Commands from the door lock (for instance a mirror fold command) to the (easy to attack) mirror must be unidirectional. A message from the (possibly attacked) mirror LIN may not block locking the car doors. It must be intercepted by the fire wall function of the gate way!)

Wireless communication must always be regarded as insecure. The wireless interface must be separated from all other functions by a gate way acting as a fire wall. Encryption of wireless communication is a MUST.

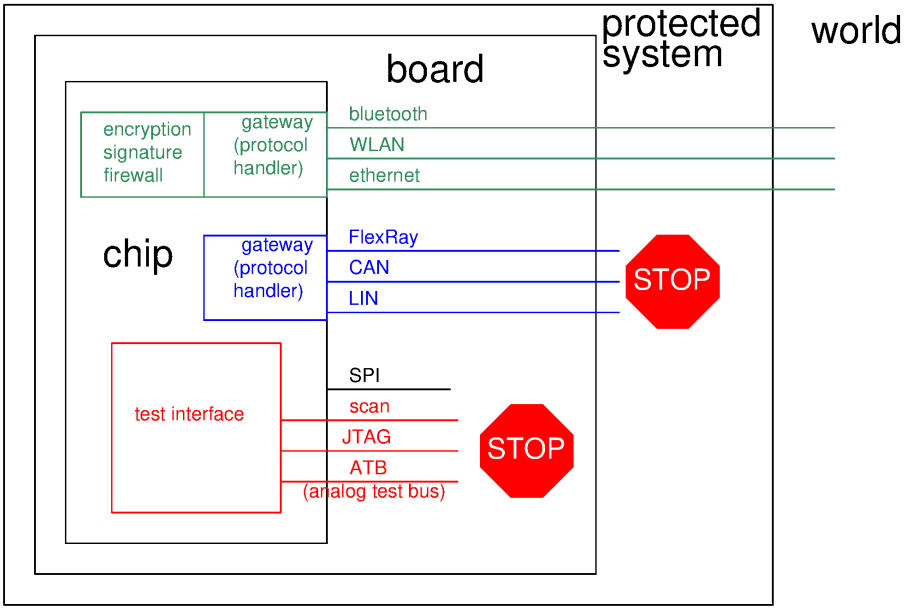


Figure 12.21: security concept of a chip, board and system

Interfaces that give access to test modes end on board level (drawn in red).

Interfaces running without encryption (in some cases the messages are standardized including the OP-codes! So encryption would violate the standard) must have a gate way and must end inside the protected system. System protection could for instance be a locked enclosure that requires a mechanical key for access. These interfaces are drawn in blue color.

Interfaces that can be accessed from the outside world mus have a gate way, a fire wall and use encryption and signatures to protect them from attacks. These interfaces are drawn in green color.

SPI (serial peripheral interface) and classical ports that give access to address bus or data bus inside the chip are less critical than test interfaces. Nevertheless I recommend these interfaces don't leave the board.

12.9.1 Encrypted access to test modes

The access to test modes can be protected by requesting an access key or by using encryption. The key is manufacturer specific and chip specific.

In case of a key authorization only the initialization of the test mode communication requires a key but the communication is running clear text. Once the test mode is initialized the communication can be observed easily.

An encrypted communication requires a key and the communication runs encrypted too. Observing the communication is less easy.

Both using an initialization key as well as using encrypted data exchange can be regarded as an improvement of security. But the test mode interface nevertheless should be kept limited to the board only. It is too easy to steal a key by social engineering of employees designing or testing a chip.

Proving a chip manipulation in the field relying on passwords and encryption only is very difficult.

12.9.2 Protection using write protect bits

An additional protection against unauthorized manipulation of the chip is using a write protection. Changing data on the chip is possible as long as the write protect bit is not set. Usually this is the case during testing and possibly during application development. Once the software is in a stable state the write protect bit is set. Usually this is the case when the project has reached production level.

The write protection bit is part of the protected memory. Once it is set it can't be erased anymore - at least this is the concept.

A possible attack is reading the entire memory and dumping it into an image file. This image can be modified. To erase the write protection bit a global memory erase is used. Typical ways of globally erasing the non volatile memory is exposing the chip to Roentgen or gamma radiation or to excessive heat. Often hot storage at 250°C for 24h is enough to erase an NVM.

Once the memory is erased the modified image can be written into the chip.

Manipulated field returns may be hard to recognize. Somtimes the mold has traces left from hot storage (gray brittle mold).

Table 60: Comparison of test mode security

protection	possible attack	security level
none, test bus shared with public bus	no effort at all	catastrophic
Local test bus ending on board level	manipulation of the board	poor
test bus + password (key)	social engineering of employees knowing the key	medium
test bus + encrypted communication	social engineering of employees knowing the key	medium
test bus + write protection	radiation, hot storage	acceptable
test bus + r/w protection	dedicated methods like electron beam	good
test bus + r/w protection + parity	electron beam + chip knowledge	very good

12.9.3 Read protection and write protection

Combining a read protection and a write protection prevents attacking a chip via global memory erase. The read protection prevents producing the memory image before doing a global erase of the memory. Reading the memory after erasing is worthless. The only thing that still can be done is creating a new firmware from scratch. Individual chip trimming is lost.

The remaining attack is erasing the read protect bit selectively. The effort is high because the chip needs to be opened and a single bit must be erased using an electron beam or similar.

Once the read protect bit is erased the next steps to modify the data are the same as overriding the write protection.

Physical access to the write protect bit leaves traces such as package damage and possibly oxide damages caused by the electron beam used to erase the read protection. After having erased a bit using an electron beam the read protect bit often has a data retention problem. Manipulated field returns can sometimes be recognized by the loss of the read protection bit.

12.9.4 Read and write protection + parity

If the read protect bit is secured by a parity a manipulation of data requires a change of the protect bits and an other bit sitting in the same row. This makes a manipulation of a chip more difficult. This is especially true if multiple parities (row and column) are used.

12.9.5 Central kill

Central kill is a military grade protection. The intention is not to prevent manipulation. The target of a central kill is to make a complete system unusable before it falls into the hands of an enemy. A central kill command at least erases the complete memory. If possible it takes the system into a state that leads to physical destruction. In case of power chips this often is quite easy turning off all protection functions and creating a short circuit (for instance turning on the high side and the low side of a power bridge simultaneously). Ideally a central kill should destroy a chip in a way that even reverse engineering of the destroyed chip is impossible.

Automotive and industrial chips usually are built in a way that the protection functions can't be fully disabled. A central kill of standard chips is difficult unless a dramatic overvoltage is applied.

Crypto chips (bank cards etc.) sometimes have similar functions such as sensors detecting an opening of the package. If the package is opened the kill function of a crypto chip destroys the chip or deletes the memory.

12.9.6 Summary of test mode access protections

As we have seen test modes are a security critical feature. Let's summarize the risks:

12.10 Interpretation of test results

Interpretation of test results is essential to predict production yield. Parameter distributions usually are not really Gaussian. But Gaussian distributions are nice for mathematical prediction of production yield. Therefore usually the results found on the tester are approximated by a Gaussian distribution. This approach can be justified as long as the real distribution is more or less similar to a Gaussian one.

A Gaussian distribution decays continuously the further we go away from the average value. The first step to verify if the approximation by a Gaussian distribution is justified. The real distribution must be checked for heavy ends and multiple maxima.

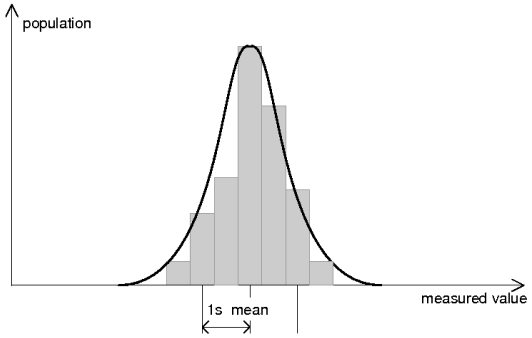


Figure 12.22: This distribution can be approximated with a Gaussian distribution

A typical case of a distribution that should not be approximated by a Gaussian distribution is shown below. It has heavy ends. Using the Gauss approximation to calculate production yield would lead to much too optimistic results.

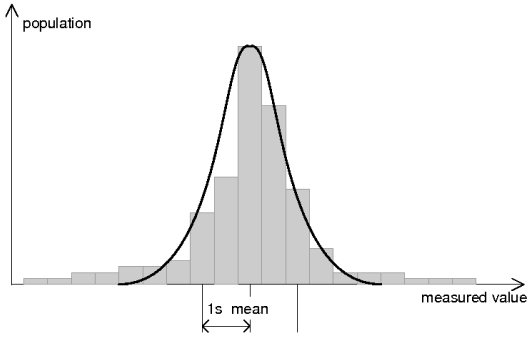


Figure 12.23: Example of a distribution with heavy ends

A distribution of a circuit that does not work reliably has multiple peaks. This kind of distribution is alarming! The two side maxima indicate a severe malfunction or a test problem (tester settling time too short etc.)

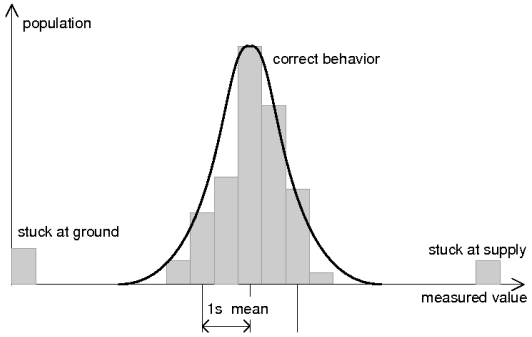


Figure 12.24: Distribution of a circuit failing randomly

Circuits that have been trimmed have distributions with the two original slopes folded in. The population that formerly was outside is moved into the now misshaped distribution. Ideally they should be rectangular. In practice this rectangular shape will not be reached perfectly due to the limited number of trim steps. Trimmed distributions may have multiple peaks within the trimming tolerance.

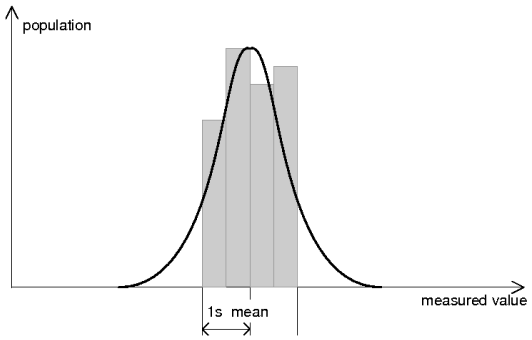


Figure 12.25: Distributions of trimmed parameters are often close to rectangular

Using the Gauss approximation for trimmed parameters often leads to a too pessimistic estimation of production yield.

12.10.1 Calculating with Gaussian distributions

A Gaussian distribution is described by its mean value and by the standard deviation. In literature the mean value usually is abbreviated as μ (very ugly for us. We normally use this letter for magnetic permeability. But let us stick with this naming convention since it is common practice.). The mean value is calculated summing all measured data and dividing it by the number of measurements.

$$\mu = \sum_n x_n/n \tag{12.41}$$

Here n is the number of samples and $x_n$are the results of the measurements.
After determining the mean value the variance van be calculated.

$$Var(x) = \frac{\sum_n (x_n - \mu)^2}{n} \tag{12.42}$$

The variance can be described as a mean square deviation of the samples measured from the mean value. Units such as $V^2$of course are not very nice for imagination. So in stead of using the variance we usually refer to the standard deviation. In mathematical literature it is abbreviated with σ. Engineers are too lazy to always switch fonts. In engineering books it is usually abbreviated with 1s (one sigma). So we get characters we will find again in the test reports...

$$1s = \sigma = \sqrt{Var(x)} = \sqrt{\frac{\sum_n (x_n - \mu)^2}{n}} \tag{12.43}$$

Knowing the mean value and the standard deviation we can write down the density function (please don't ask me where the equation is coming from. There are some mathematical books for further details..)

$$f(x) = \frac{1}{\sigma * \sqrt{2 * \Pi}} * e^{-\frac{1}{2} * (\frac{x-\mu}{\sigma})^2} \tag{12.44}$$

The integration of this equation leads to the error function. Usually it is scaled with

$$t = \frac{x - \mu}{\sqrt{2} * \sigma} \tag{12.45}$$

So what is important for engineers?
The density function f(x) simply describes how many results can be found in which range.
68.27% of the population will be found between -1s and +1s.
95.45% of the population will be found between -2s and +2s.
99.73% of the population will be found between -3s and +3s.
To calculate the percentage of samples within a certain range you can for instance use octave's error function:

```
octave:14> erf(2/1.4142)
ans =  0.95450
```

This line calculates the yield of a Gaussian distribution from -2σ to +2σ.

Things are getting more interesting if we have to deal with many measurements (for some ICs thousand or more measurements are taken!). If we want to know the over all production yield we simply have to multiply the yield of every test carried out.

$$yield = \prod_n yield(test_n) \tag{12.46}$$

To get a good production yield at a high number of tests the loss per test must be as low as possible. This leads to the requirement to design such that the test limits are more than 6s away from the mean value. (There is quite some discussion if we really need a 6s design. But some chip manufacturers during the 1990 published data that process spread eats up about 1s to 1.5s leaving 4.5s of the original 6s design).

Well, in certain cases we can tolerate a lower yield of some parameters:

1. It must be a parameter the customer is willing to pay for
2. It must be very few parameters where we accept less than 6s
3. It must be discussed with manufacturing (Otherwise you get complaints every week. Get a written disclaimer for the specific test!)

Besides mean value and standard deviation test engineers often discuss so called cp and cpk - values. These numbers describe the distance of the test limits (LSL:lower specification limit; USL: upper specifivcation limit) from the standard deviation.

$$c_p = \frac{USL - LSL}{6\sigma} \tag{12.47}$$

And:

$$c_{pk} = \frac{min(\mu - LSL; USL - \mu)}{3\sigma} \tag{12.48}$$

In case the mean value μ is exactly in the middle between the upper specification limit (USL) and the lower specification limit (LSL) the values of $c_p$and c_{pk}become equal. In addition the CPK value describes the distance of the test limits (USL=Upper Specification Limit, LSL=Lower Specification Limit) from the 3 sigma width. A c_{pk} of 1 means the specification limit with the lower margin exactly meets the 3σ value of the distribution. A c_{pk} of 2 means we have the closest specification limit at 6σ of the distribution.

Ideally designing for $c_{pk} \geqq 2$ is the target to have a negligible impact of a parameter on the production yield.

12.11 Searching defects and break downs with optical emission

Certain physical effects lead to emission of photons. Typically this is the case wherever carriers are accelerated or decelerated at high electrical fields. This can be used to find locations on the chip with extreme electrical fields. Typical examples are:

- holes in gate oxides. Here electrons get accelerated passing the hole. Can be used to find gate damages.
- Drain regions with hot electrons (power transistors, NVM write amplifiers)
- zener break down (the field strength is accelerating electrons within a few nm). Can be used to find well break downs. Very bright. (can sometimes even be seen with the bare eye)
- bipolar junctions (well, compared to a zener diode a normal diode is fairly dark). The higher the current density the brighter it gets.
- Activation of ESD protections (extremely bright, but only for some ns)

To detect these optical emissions a very sensitive optical sensor is needed. The device to be searched for defects must be in complete darkness but supplied (and operated) under the microscope. Usually first a reference photo of the unpowered device is taken (dark image). Then the device under test is powered up and a second image is taken. The dark image then is subtracted from the image in operation.

12.12 Test equipment maintenance

12.12.1 Floppy disk replacement

A lot of test equipment from the 1980s and 1990s relies on floppy disks for data exchange. Floppy disk drives are mechanical systems that wear out. Even if seldom in use the rubber belt often used for transmission between the motor and the disk gets brittle and breaks after some years.

The solution is replacing the disk by a floppy emulator. The floppy emulator uses an SDCARD or a USB storage to store the data.

The interface between the floppy and the floppy controller is fairly low level. It consists of the data lines /RDATA, /WDATA, READY and some control signals indicating the position on the disk. So the floppy controller in fact expects

a medium consisting of tracks and sectors. The emulator has to present a data structure of this kind to the floppy controller. The classical solution is to have at least one floppy disk image on the SDCARD or the USB-storage. The floppy emulator makes this disk image available to the floppy controller.

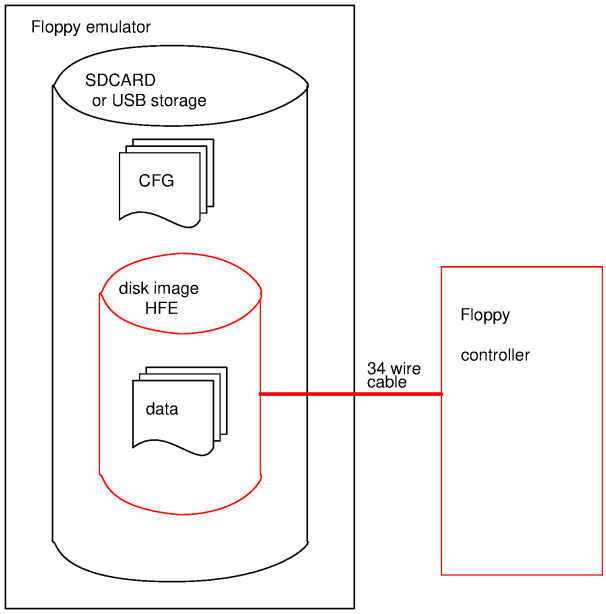


Figure 12.26: Floppy emulation concept

Since the emulator has to present something to the floppy controller that looks exactly like a floppy disk the size of the image corresponds exactly the size of a floppy. (160K, 320K, 720K, 1440K depending on the type of floppy you want to emulate. However FAT12 allows up to 16MB. But tests with bigger images at least failed using the emulation for a LeCroy 9354 scope) Even if the SDCARD or the USB storage is several GB big the floppy image is still limited to the size of the original floppy. The only thing that can be done to take advantage of the storage space is to place several disk images on the SDCARD or the USB storage.

The CFG file is needed to tell the emulator hardware what it should do with the data on the medium.

To make the data files available on a computer a special software to extract the data from the image is required. The required software is available at http://hxc2001.com/download/floppy_drive_emulator/HxCFloppyEmulator_soft.zip . It runs on Windows and on linux using wine (the windows emulator). A typical screenshot of the disk browser looks like this:

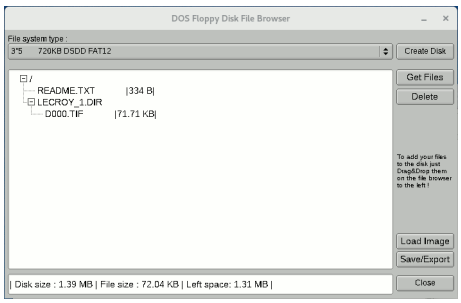


Figure 12.27: Screen shot of the disk browser reading one of the floppy images

Using the export function of the disk browser single files can be extracted from the image. The disk browser also offers the possibility to convert the HFE file into a disk image using the IMG format. The IMG-image then can be mounted as a loop back (something like mount -o loop file.img /mountpoint) on a linux system (probably something of that kind in possible on Windows as well, but I never tried it out).

Besides 34 wire cables there are 26 wire cables as well. For a LeCroy 9354 oscilloscope the following adapter was required:

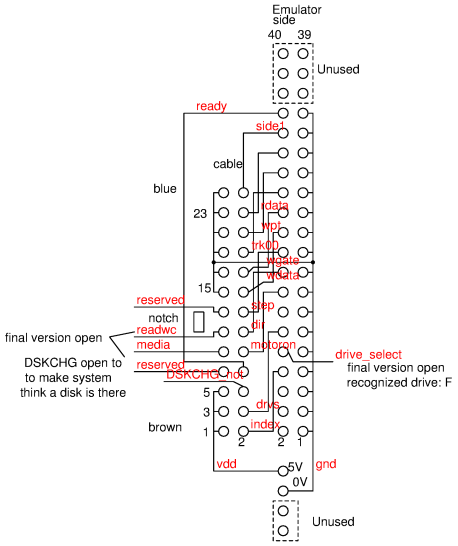


Figure 12.28: 26 wire to 34 wire adapter required for a LeCroy9354 scope

Using the floppy emulation for a HP4145: The HP4145 parameter analyzer uses a different data format called LIF. This format requires a different image than a classical floppy. Since the LIF image is different from a normal floppy writing data to the image of a LIF floppy seems not to work (at least I didn't find a solution) and measurement data must be exchanged via the IEEE interface of the HP4145.

12.12.2 Equipment control using LAN

Many instruments used in the lab can be remote controlled. Old equipment frequently uses the GPIB interface. This is a standard dating back to the 1980s intruduced by Hewlet Packard inc. It can be controled directly from software simply sending and receiving ASCII code. The commands were instrument specific and usually well documented in the instrument's manual. (A typical example is [**?**])

Today (2021) most lab equipment has a LAN connection. So it can directly be connected to the PC. Some of these instruments offer a WEB-browser interface, some of them are designed to operate with equipment specific software. To connect the instrument to the PC the IP address and the port must be known. (Some of these instruments support dynamic IPs using a DNS (domain name server), but for writing software this is making things more complicated). The best thing about using an ethernet LAN connection is that the istrument is galvanically isolated because ethernet uses transformer coupling.

To connect an instrument to a PC via LAN you need:

- IP address
- port used for communication
- List of the SCPI commands of the device

Old equipment often still can be used using interface adapters from LAN to GPIB or USB. (Typical manufacturers are National Instruments or Prologix). These adapter directly transfer te commands and data (ideally) without changing any code.

13 Energy supply

"Don't worry about the waste of nuclear power plants, we get the energy from the mains."

Well, this is a bit short sighted. No mater where the energy is coming from energy always is associated with:

- Costs
- Pollution
- Risks
- Disturbance

So it is worth while looking at the "mains" a little closer.

13.1 Where does the energy come from

What is called the mains is supplied from various sources. These usually are power plants (combinations of different power plants ranging from solar collectors to nuclear power plants with very different properties) combined into a big network with transformer stations and fuses in between. The decision which power plant is used for what depends on the properties of these power plants and the performance of the regulation of the complete network.

Regulation of the network does not only mean technical decisions. This regulation also has to consider economical aspects choosing the cheapest possible mix of energy sources.

With systems becoming more and more intelligent and complex the economy of power supply is more and more becoming a matter of concern even in mobile systems such as cars or cellular phones!

To get a first overview let us have a look at different power sources and their properties.

Table 61: Properties of energy sources

type	typical power	time constant	energy density	typical cost (without transmission)	pollution	start up	shut down
nuclear power plant	1000 MW	hours		2c/kWh *	nuclear waste	months	emergency: seconds normal: months
coal power plant	500MW	hours		5c/kWh	CO_2	hours	hours
oil power plant	200MW	1h		13c/kWh	CO_2	about 1h	minutes to hours
gas power plant	200MW	5..20 minutes		13c/kWh (?)	CO_2	5..20 minutes	minutes to hours
water power plant	200MW	seconds		9c/kWh	none	seconds	seconds
wind power plant	1MW	DC-DC converter required		4c/kWh	none	no control	no control
solar thermal power plant	some 100kW	DC-DC converter required		15c/KWh	none	no control	no control
solar electrical power plant	some kW	DC-DC converter required		15c/KWh	none	no control	no control
car alternator	600W	some ms		98c/kWh	CO_2	seconds	seconds
rechargeable battery	some W	almost ideal voltage source	NiMh: 50..70Wh/kg LiIon: 120..140Wh/kg	20c/kWh	acids, metal oxides	NA	NA
dry battery	some W	almost ideal voltage source		200 Euro/kWh	acids, metal oxides	NA	NA
Super capacitor [82]	some W to kW peaks	buffer for 1s to 10s	up to 5Wh/kg	? (depends on the number of cycles)		NA	NA

*: This number does not iclude the cost of removing the power plant at the end of life!

(Calculation: coal has an energy of about 8KWh/kg. A typical power plant has an efficiency of about 35% to 50%. So from coal we can expect to get about 4kWh/kg of electrical energy. Coal for power plants currently (2013) is about 100 Euro per ton. s a kWh costs about 5c neglecting maintenance and depreciation of the power plant. Adding maintenance and depreciation about 9c/kWh look reasonable

Crude oil provides about 11kWh/kg. Without tax a kg of oil costs about 70c. Assuming the same efficiency as burning coal we get about 13c/kWh.

Table 62: US DoE efficiency requirements for AC-DC power supply, low voltage

Nameplate output power	Efficiency while active	Max. power without load
$P \leqq 1W$	$\eta \geqq 0.517 * P_{out}/W + 0.087$	$P_{in} \leqq 100mW$
$1W \leqq P_{out} \leqq 49W$	$\eta \geqq 0.0834 * P_{in}/W - 0.0014 * P_{out}/W + 0.609$	$P_{in} \leqq 100mW$
$49W \leqq P_{out} \leqq 250W$	$\eta \geqq 0.870$	$P_{in} \leqq 210mW$
$P_{out} \geqq 250W$	$\eta \geqq 0.875$	$P_{in} \leqq 500mW$

Table 63: US DoE efficiency requirements for AC-DC power supply, basic voltage

Nameplate output power	Efficiency while active	Max. power without load
$P \leqq 1W$	$\eta \geqq 0.5 * P_{out}/W + 0.16$	$P_{in} \leqq 100mW$
$1W \leqq P_{out} \leqq 49W$	$\eta \geqq 0.071 * P_{in}/W - 0.0014 * P_{out}/W + 0.67$	$P_{in} \leqq 100mW$
$49W \leqq P_{out} \leqq 250W$	$\eta \geqq 0.880$	$P_{in} \leqq 210mW$
$P_{out} \geqq 250W$	$\eta \geqq 0.875$	$P_{in} \leqq 500mW$

In a car conditions are different. Since the engine is not permanently running at a good operating point it is reasonable to assume about 10% average efficiency. Furthermore we are running the car with a fuel we have to pay tax for. Assuming the alternator has an efficiency of 80% we end up at about 98c/kWh.

Dry batteries offer some Wh for a Euro. Ecpecting about 200 Euro per kWh looks reasonable.

The cost of rechargeable batteries strongly depends on the number of cycles they survive. Well, how about 1000 at the same production cost as a dry battery?)

Depending on the application of chips and the average life time in application it makes sense to consider power consumption as a cost factor.

Example1: Assuming linear automotive voltage regulator (supplied at 98c/kWh) running 3000h (corresponds between 100000km and 200000km of car usage) dissipating 0.5W we end up with a cumulated cost of about 10000h*0.5W*0.98 Euro / 1000Wh = 4.90 Euro. Money enough to replace it by a switchmode power supply!

Example2: Operating this regulator from mains in industrial environment with an average cost of 10c/kWh we end up with energy costs of about 50c. Using a switchmode power supply here becomes financially marginal.

13.1.1 Some efficiency standards to consider

Legislation of the the European Union and the United states requires certain levels of efficiency of power supplies.

Efficiency requirements of the US DoE (status 2021)

Efficiency requirements od the EU for Tier-2 manufacturers

13.1.2 Future trends of energy storage

In the future the energy stored in vehicles can play an important role in the regulation of power grids. A single electrical car is exüected to have an energy reservoir of about 70kWh. Assuming 10 million cars in a country like Germany this means a total energy of $7 * 10^8 kWh$ simply store in the vehicles. Even if the energy supplyers only are allowed to pull back about 10% of this amount the energy reserve to cover the lack of wind- or solar energy this provides $7 * 10^7 kWh$, enough to supply Germany for about 3 to 4 hours without using any other electrical energy source.

Table 64: CoC Tier 2 single voltage AC-DC power supply, low voltage

Nameplate output power	Efficiency while active	10% load efficiency
$0.3W \leqq P_{out} \geqq 1W$	$\eta \geqq 0.517 * P_{out} + 0.091$	$\eta \geqq 0.517 * P_{out}$
$1W \leqq P_{out} \leqq 49W$	$\eta \geqq 0.0834 * P_{in}/W - 0.0011 * P_{out}/W + 0.609$	$\eta \geqq 0.0834 * P_{in}/W - 0.00127 * P_{out}/W + 0.5$
$49W \leqq P_{out} \leqq 250W$	$\eta \geqq 0.88$	$\eta \geqq 0.78$
$P_{out} \geqq 250W$	N/A	N/A

Table 65: CoC Tier 2 single voltage AC-DC power supply, basic voltage

Nameplate output power	Efficiency while active	10% load efficiency
$0.3W \leqq P_{out} \geqq 1W$	$\eta \geqq 0.5 * P_{out} + 0.169$	$\eta \geqq 0.5 * P_{out} + 0.06$
$1W \leqq P_{out} \leqq 49W$	$\eta \geqq 0.071 * P_{in}/W - 0.00115 * P_{out}/W + 0.67$	$\eta \geqq 0.071 * P_{in}/W - 0.00115 * P_{out}/W + 0.57$
$49W \leqq P_{out} \leqq 250W$	$\eta \geqq 0.89$	$\eta \geqq 0.88$
$P_{out} \geqq 250W$	N/A	N/A

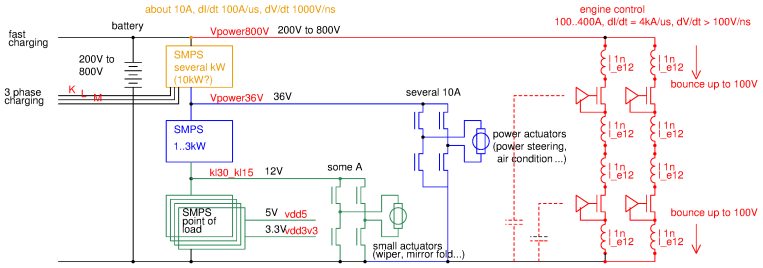


Figure 13.1: Simplified supply concept of an electrical car

The big SMPS between the 800V domain and the 36V domain is likely to also hold the slow charging access (from normal mains) and could provide a patch to convert back from the 800V rail to the mains as well.

13.1.3 Future concepts of supplying mobile applications

Currently (2018) the most common way of charging mobile applications is to simply connect them to the mains either directly (so the converter is on board) or indirectly (example: charging via the USB connector as frequently done with cellular phones. So the conversion to the low voltage domain is is an external switchmode power supply).

For cellular phones or similar devices an alternative solution is inductive charging. The device to be supplyied is holding a coil that picks up the magnetic field of the charging device. In an abstract way the system to be charged can be regarded as the secondary side of a transformer while the charging device is the primary side of the transformer.

Systems using inductive charging since many years are electronical car keys.

The mathematical description of inductive coupling dates back to Michael Faraday, who published his research results 1831.

$$rot(E) = -\frac{dB}{dt} \tag{13.1}$$

In this equation the electrical field E and the magnetic field B are vectors. The integrated version of the Faraday Law looks a bit more handy:

$$\oint Eds = -\int \frac{dB}{dt} dA \tag{13.2}$$

The left part of the equation is the voltage induced in the path surrounding a certain area A. The right part of the equation is the change of the magnetic field B going through the area A. The magnetic field B and the area A both are vectors again. In case of a transformer we can regard the area A as a constant. Only the magnetic field B changes with time.

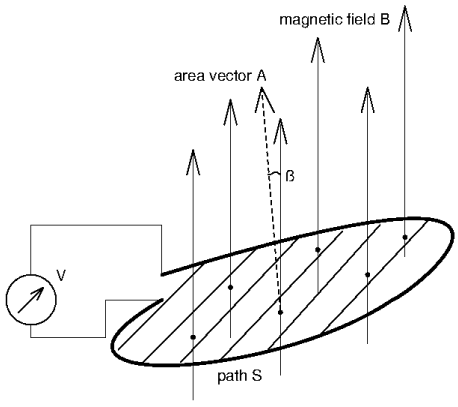


Figure 13.2: magnetic field flowing through a wire winding

In the case of a homogenous magnetic field and a tilt of the area with an angle β the voltage induced becomes:

$$V = -N * A * cos(\beta) * \frac{dB}{dt} \tag{13.3}$$

N is the number of windings of the coil. Ideally the tilt between the area vector A and the magnetic field B is 0. So $cos(\beta)$ becomes 1. l is the length of the path s (length of one winding)

One of the most extreme applications of inductive charging is the concept of inductive charging of an electrical car [54]. The antennas in discussion for 3.6kW charging have an area of about 250mm * 250mm = 0.0.0625m^2. The circumfence is about 1m with a typical number of winding of N=10. Frequencies in discussion are about 100kHz (85kHz to 140kHz).

$$\frac{dB}{dt} = \frac{V}{N * A * cos(\beta)} \tag{13.4}$$

Assuming good allignment of both coils $cos(\beta)$ becomes 1. Since most electric cars are expected to have 400V to 800V batteries we need about 40..80V per winding. So at the receiver coil the magnetic field must satify:

$$\frac{dB}{dt} = \frac{80V}{1 * 0.0625m^2} = 1280T/s$$

Assuming a sine wave shaped current in the primary coil and perfect coupling we get:

$$B(t) = B_{peak} * sin(2 * \pi * 100kHz)$$

$$max(\frac{dB}{dt}) = B_{peak} * 2 * \pi * 100000s^{-1}$$

Since most electric cars are expected to have 400V to 800V batteries we need about 40..80V per winding. So at the receiver coil the magnetic field must satify:

$$B_{peak} = \frac{80V}{2 * \pi * 100000s^{-1} * 0.0625m^2} = 2.04mT$$

In areas accessible by human beings the maximum permitted flux density however is limited by regulations. In most countries the limit is $6.25\mu T$. The wireless energy transfer may only be activated when there is nobody between the primary coil and the secondary coil.

To keep stray fields under control the magnetic field lines must be closed by ferrites. This is expected to keep about 99% of the magnetic field in an area of about $1m^2$.

Inductive cooking for comparison: To get a comparison with already existing technologies let's have a look at inductive cooking.

Measurements setup:
diameter of the cooking plate 2r=24cm, $A = r^2 * \pi = 0.0452m^2$
number of windings N=2
measured voltage (zero to peak) Vp=10V
Measured frequency f=5kHz, T=$200\mu s$
signal shape: rectangular

$$\frac{dB}{dt} = \frac{V_p}{N * A} = \frac{10V}{2 * 0.0452m^2} = 110.6T/s$$

Since the signal is rectangular with a pulse duration of $100\mu s$ we gat a change of the magnetic field:

$$B_{pp} = \frac{dB}{dt} * \frac{T}{2}$$

The peak magnetic filed of a cooking plate thus becomes:

$$B_p = \frac{dB}{dt} * \frac{T}{4} = 110.6T/s * 200\mu s/4 = 5.5mT$$

Let's summarize the comparison:

Table 66: Comparison of magnetic fields of inductive cooking and inductive charging

parameter	inductive cooking (measured)	inductive charging (calculated)
frequency	5kHz (rectangular)	85kHz and 140kHz (proposed)
dB/dt	110.6T/s	1280T/s
peak field B_p	5.5mT	1.42mT to 2.34mT
voltage induced in $1cm^2$	11mV	about 128mV

Any conducting material exposed to the field of an inductive charging device will be heated by eddy currents 10 times faster than if it were sitting on top of an inductive cooking plate! Boards with a ground plane may be destroyed within seconds.

Inductive charging devices must have means to detect any kind on unecpected object in close proximity of the antenna. Any electronic system that might get exposed to the field of an inductive charging device must be designed extremely compact and robust. Cable harnesses may not have any loops that could build up an induced voltage or they must be made intentionally resistive to limit the currents.

Antenna Specifications found: Specifications found in the internet indicate a maximum field of $186 dB\mu V/m$ which corresponds 2000V/m. This number only seems to make sense regarding the coupling as a propagating wave.

$$E = \sqrt{z_{vacuum} * p}$$

The power density in this case calculates as

$$p = \frac{E^2}{Z_{vacuum}} = 10.6 kW/m^2$$

Whether this definition makes sense is debatable because we are looking at a short distance coupling.

System Specifications found: Ford Motor corporation requests an robustness against magnetic fields of 120dBpT ($1\mu T$) at 100kHz. [55]page 52. This means current EMC specifications are factor 2000 (or 66dB) below the magnetic field strength close to a wireless power transfer device (WPT) operating with 2mT at 100kHz. Systems developed according to these specifications will not survive exposure to the fields of wireless charging.

Medical Aspects of wireless charging and inductive cooking The information about permissible fields human beings may be exposed to is scattered into many publication. The "ICNIRP Guidelines for limiting exposure to time-varying electric, magnetic and electromagnetic fields" [56] tries to summarize the available knowledge about medical effects. These guidelines distinguish between 'occupational exposure' and 'general public exposure'. The limits for 'general public exposure' are between factor 3 and factor 5 lower to keep a certain safety marging for the public. Values stated as 'occupational exposure' are considered as absolute limits. The following table displays the calculated fields in comparison with the suggestions of ICNIRP. ICNIRP assumes a maximum field of 2.0/f as a maximum permissible field (Table 6)

Table 67: Wireless charging fields compared to legal requirements

application	calculated	ICNIRP 'occupational'	ICNIRP 'public'	source
wireless charging	1.42mT..2.34mT	$20\mu T$ @100kHz	$6.25\mu T$ @100kHz	[56] table 6 and 7
inductive cooking	5.5mT	$400\mu T$ @5kHz		[56] table 6

This means the fields of wireless power transfer used for charging electrical vehicles is 2 magnitudes higher than the permitted exposures stated by ICNIRP! Exposure of human beings to the fields of a wireless charging antenna must be prevented under any circumstaces.

Even the fields used for inductive cooking are still one magnitude higher than the maximum fields suggested as limits for human exposure. Therefore inductive cooking devices are designed to turn off if there is not pot present, that absorbs the energy.

Currents induced inside a human body The main effect of human exposure to a magnetic field is induction of a current inside the body. To estimate the currents induced the conductivity of the human body must be known. ICNIRP states a value of 0.2 S/m. As a rough guess we can assume the "receiving antenna" as a ring of 30cm diameter (d) corresponding an area of about $700 cm^2$.

$$A = (\frac{d}{2})^2 * 3.1415 = 706 cm^2$$

The ring circumfence is about 94cm.

$$C = d * 3.1414 = 94 cm$$

Assuming a thickness (th) of the ring of 10cm the resulting resistance becomes roughly 600 Ohm.

$$R = C/(0.2A/V * (th/2)^2 * 3.1415) = 0.94m/(0.2A/V * 3.1415 * 0.05^2 * m^2) = 598\Omega$$

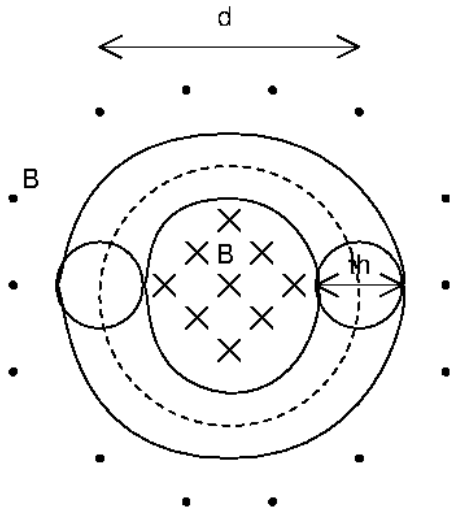


Figure 13.3: Aproximation of the human body acting as an antenna by a torus

Current flowing in a human body caused by the field of an inductive cooking plate thus becomes 13.5mA. Probably not yet fatal.

The current flowing in a human body caused by the field of a wireless power transfer system is about 140mA to 160mA!

Well, this simple calculation assumes a more or les homognous field distribution and neglects that the current density in the torus isn't exactly constant. But this simple calculation clearly shows the magnitude.

13.1.4 Super capacitors

Super capacitors are very similar to electrolytic capacitors. Typically one of th electrodes is porous in the nm scale. This way the elecrode has a very high surface area. The second electrode is some kind of a conductive solution. In some case each pore accomodates a single ion to store the energy.

Super capacitors can be used for energy storage similar to a LiIon battery. The energy per volume about factor 5..20 is less than that of a battery. The number of permitted charge cycles is about factor 10..1000 higher. The sweet spot of using super capacitors is in the range of 1s to 100s.

13.2 Pollution

Pollution is an almost unavoidable consequence of energy production. It is quite difficult to get reliable numbers of the whole process. At least it is fairly easy to calculate the CO_2 footprint of burning chemical fuels. All chemical fuels in use in a large scale are either coal or carbon-hydrates (oil, gasoline, diesel, methane, ethane..). To calculate the CO_2 footprint let's start with some basic chemistry.

13.2.1 Burning coal

The process is:

$$2 * O_2 + C \rightarrow 2 * CO_2 + 32MJ/kg$$

The energy of 32MJ/kg is an average for 100% coal. Brown coal usually provides about 15MJ/kg because more than 50% of it's weight is sediments that don't contribute energy burning the brown coal. Hard coal (Anthracite) consists of more than 70% carbon and less than 30% of sediments. Coke consists of almost 100% of carbon. So this is probably the best reference for further calculations. Coke has an energy content of about 23-31MJ/kg.

Weight of the carbon dioxide produced: carbon has an atomic weight of 12.01. Oxygen has an atomic weight of 15.99 (for simplicity let's use 12 and 16 for the following calculations). The weight ratio of the carbon and the carbon dioxide resulting from burning the carbon calculates as

$$ratio_{CO_2-C} = \frac{12 + 2 * 16}{12} = 3.667 \tag{13.5}$$

This means burning 1kg of coke produces about 3.667kg of carbon dioxide. Other types of coal produce less carbon dioxide but at the same time have a lower energy contents. Here is a little comparison assuming brown coal holds 50% of carbon and 50% of sediments. Hard coal is assumed to consist of 80% carbon and 20% of sediments. These numbers are somewhat arbitrary because every coal mine has a somewhat different distribution of carbon and sediments in the coal. But I think they are reasonable averages. The following table lists the higher heating value

(HHV) assuming the process of burning coal doesn't produce any water that has to be vaporized (so HHV and LHV are identical, lower heating value is the HHV minus the vaporization energy of the water)

Table 68: Comparison of carbon dioxide emissions burning coal

coal type	% of carbon	HHV	Kg CO_2 per kg fuel	kg CO_2per MJ
coke	100	32.5MJ/kg	3.667	0.1128
hard coal	80	26MJ/kg	2.934	0.1128
brown coal	50	16.2MJ/kg	1.833	0.1128

Remark: https://en.wikipedia.org/wiki/Heat_of_combustion states 15MJ/kg for brown coal.

The amount of CO_2 per MJ of energy is constant because in either case we are always burning the same material: carbon. The difference between the different kinds of coal is simply the mineralic left over - in other words the amount of ash.

13.2.2 Burning Natural gas

Natural gas mainly consists of methane, ethane and butane. all are gases. Methane has the chemical formula CH_4. Ethane has two carbon atoms and the chemical formula is C_2H_6 . propane has 3 carbon atoms and 8 hydrogen atoms with the chemical formula C_3H_8. The ratio of the three main components varies from gas well to gas well.

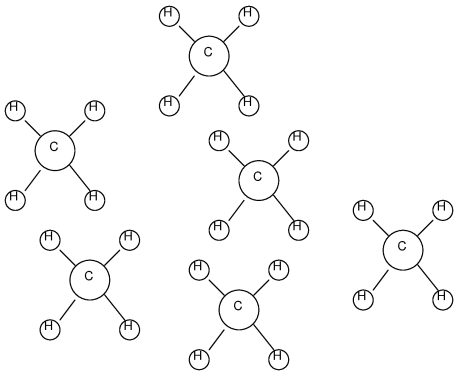


Figure 13.4: 6 molecules of methane

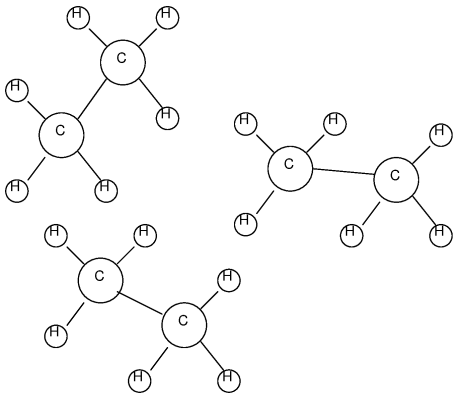


Figure 13.5: 3 molecules of ethane

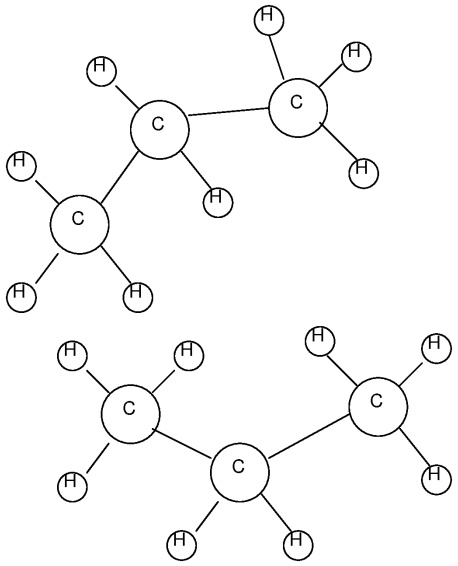


Figure 13.6: 2 molecules of propane

From one figure to the next the number of carbon atoms remains the same, but the number of hydrogen atoms decreases from 24 (methane) to 16 (propane) the bigger the molecules get. We end up with $C_nH_{(2n+2)}$ if we keep increasing the number of carbon atoms.

This however doesn't harm the calculation too much as long as we refer to the weight of the gas. To see the differences let's always have a look at 6 carbon atoms and at the weight of the resulting carbon dioxide after burning.

Table 69: Comparison of carbon dioxide emission burning gas

gas	atomic weight	after burning	atomic weight of CO_2	ratio
6*CH_4	96	6*CO_2+ 12*H_2O	6*44=264	2.75
3*C_2H_6	90	6*CO_2+ 6*H_2O	6*44=264	2.933
2*C_3H_8	88	6*CO_2+ 4*H_2O	6*44=264	3

The ratio of the weight of the gas and the weight of the carbon dioxide doesn't change too much with the content of methane, ethane and propane. It only moves between 2.75 and 3.0. As a reasonable average we can assume burning 1kg of natural gas produces about 2.9kg of carbon dioxide.

To estimate the technical value of a fuel it is important to know the heating value. To produce mechanical energy the Carnot process usually can't be run down to temperatures where the water condensates. For this reason the lower heating value is listed. (Source: https://en.wikipedia.org/wiki/Heat_of_combustion)

Table 70: Energy versus carbon dioxide emission

fuel	LHV	Kg CO_2 per kg fuel	kg CO_2per MJ
methane	50MJ/kg	2.75	0.055
ethane	47.6MJ/kg	2.933	0.061
propane	46.3MJ/kg	3	0.065

The most important observation is, that we get more energy per kg of fuel. This is due to the contribution of combusting hydrogen additionally to the carbon. Since hydrogen is much lighter than carbon the weight doesn't change much. The emission of CO_2 per MJ is the lowest for methane because there the contribution of burning hydrogen to the total amount of energy obtained is the biggest. The longer the molecules get the more heat is coming from burning the carbon and thus the CO_2 emission per MJ increases.

13.2.3 Burning liquid fuels

Liquid fuels are hydrocarbons just like natural gas. The molecules are bigger however and the hydrocarbon gets liquid at room temperature. As an average we can assume to have slightly more than 2 atoms of hydrogen per atom of carbon. We get approximately the weight ratio before oxidation and after oxidation:

Table 71: carbon dioxide emission burning classical fuels

fuel	atomic weight	after burning	atomic weight of CO_2	ratio
$C_nH_{(2n+2)}$	n*12+2n+2	n*$CO_2 + (n+1)H_2O$	n*44	close to 44/14=3.143

The longer the molecules get the more carbon relative to the hydrogen is combusted. But we always remain below the level of carbon dioxide emission of burning pure coal.

Table 72: Comparison of carbon dioxide emission using benzene and diesel

fuel	LHV	specific weight	Kg CO_2 per kg fuel	kg CO_2per MJ
benzene	41.8MJ/kg		3.385	0.081
diesel	44.8MJ/kg		3.157	0.0704

(Note 1 : reaction of benzene is assumed $2 * C_6H_6 + 9 * O_2 \rightarrow 12 * CO_2 + 6 * H_2O$and a weight ratio of CO_2 versus fuel weight of $(6 * (12 + 2 * 16)/2 * (6 * 12 + 6 * 1) = 3.3846$.

Note 2: Diesel:86.1% of the fuel mass is carbon. source: https://en.wikipedia.org/wiki/Diesel_fuel.)

13.2.4 Using electric energy

The amount of carbon dioxide produced if we use electric energy depends on the source of the electricity. Electricity coming from burning fossil fuels produces carbon dioxide. Electric energy coming from renewables produces much less carbon dioxide (to be exact producing the concrete used to build the water power plant or the wind energy tower does produce carbon dioxide too. But this is at least one magnitude less than producing energy from burning fossil fuels.)

To estimate the relationship between electric energy and carbon dioxide we need to know the percentage of electricity coming from fossil fuels. For many countries there are fairly reliable statistics available. In addition we need some information about the efficiency of the thermal power plants and the transportation losses of the electricity.

2015 the world wide the percentage of solid fuels (coal), liquid fuels (crude oil, gasoline, diesel) and natural gas to produce electricity was 66.3% (solid fuels 39.3%, petroleum 4.1%, gas 22.9%) [65, page 16]. The US energy information administration publishes data about the efficiency of thermal power plant. It is rated a bit strange for European citizens however. So here is an important conversion:

$$1BTU = 0.2931Wh \tag{13.6}$$

Data of average heat rates for 2017 published (https://www.eia.gov/electricity/annual/html/epa_08_02.html) in BTU/per kWh:

Table 73: average heat rates published 2017

BTU per kWh	coal	petroleum	gas
steam generator	10043	10199	10353
gas turbine		13491	11176
internal combustion		10301	9120

With these numbers the efficiency of a power plant can be calculated

$$\eta = \frac{1000Wh}{heatrate * 0.2931Wh} \tag{13.7}$$

This leads to the following table:

Table 74: Average efficiency of power plants using heat rates published 2017

η	coal	petroleum	gas
steam generator	33.9%	33.5%	32.96%
gas turbine		25.29%	30.53%
internal combustion		33.12%	37.4%

The different columns indicate the fuel used while the lines relate to different methods of converting the mechanical energy into a electricity.

These numbers do not yet include transformation losses and wire losses. Including these losses a typical electrical grid is expected to have a wheel to mains efficiency of about 30%. To produce 1kWh (at the mains of the customer) fuel with an energy contents of about.

$$P_{fossil1kwh} = 0.663 * 1kWh/0.3 = 2.21kWh = 7.965MJ$$

With the world wide energy mix (solid fuels 39.3%, petroleum 4.1%, gas 22.9%) these 7.965MJ correspond about 0.21kg of fossil fuel or 0.65kg of carbon dioxide.

$$1kWh \sim 0.65kgCO_2$$

This number is a world wide average calculated using the statistical data of 2015 and 2018. In countries with a higher percentage of renewable energy this number is slowly getting better. (For example Sweden only produces about 28% of it's electricity from fossil fuels. So for Sweden 1kWh electric energy corresponds about 0.3kg of CO_2 . For comparison: Germany produces 53.3% of it's electricity from fossil fuels)

Example: running an electric car with an energy consumption of 20kWh/100km leads to a carbon dioxide emission of 120g/km using the world energy mix of 2015! Using the energy mix of Sweden the emission of the same car drops to 55g/km. Using the energy mix of Germany the emission becomes 105g/km.

13.2.5 Cost of pullution

Today (2019) the cost of polluting the environment with carbon dioxide is barely taken into account. With the acceleration of global warning this will change. First numbers are popping up since scientists try to remove carbon dioxide from exhaust gas and even from the atmosphere.

One approach reported is to press the carbon dioxide into basaltic minerals. There the carbon dioxide reacts with the basalt and gets absorbed. First test carried out in Iceland indicated cost of removing 1 ton of carbon dioxide in the range of 600.- Euro.

13.2.6 Conclusion of the comparison

Replacing coal by natural gas or liquid fuels reduces the emission of carbon dioxide by about 30% to 40%. Differences between natural gas and gasoline or diesel are in the 10% range. The advantage of using diesel instead of gasoline is only due to the higher compression of the diesel engine that leads to a slightly higher efficiency.

Simply multiplying the weight of the fuel used by factor 3 to get the total amount of CO_2 produced at combustion is already a good educated guess with about +-20% accuracy.

There is one exception of the rule: brown coal containing only 50% of carbon while the rest is dead weight of sediments. The heat value of brown is so low that the carbon dioxide per MJ again is in the range of 0.113 Kg/MJ just like all other kinds of coal.

Replacing fossil fuels by electricity only makes sense if the electricity used is mainly produced using renewable sources.

No matter which fossil fuel we use the differences for our carbon footprint is just some 10%. On the long run we must resign from burning any fossil fuels.

13.3 Energy distribution

This section deals with the distribution of electrical energy. However with the advent of local storage (for instance in car batteries) and with the advent of fuel cells the boundary between electrical energy transport and other medias of energy is getting more and more fuzzy.

The classical way of distributing electrical energy is using 3 phase high voltage lines. Typically there is a high voltage side with voltages ranging from 110kV to 1MV for long distances. (hundreds of kilometers), a mid level for local distribution (usually 5kV to 110kV) for distances up to some 10 Km. The local supply usually is done with 400V/230V. (190V/110V in the USA). Since it is a 3 phase system the load can either be connected between two phases or from one phase to ground. The three phases can be represented by vectors (U, V, W). The phase angle between the vectors is 120^0. The voltage seen by a load between the phases is represented by the dashed lines.

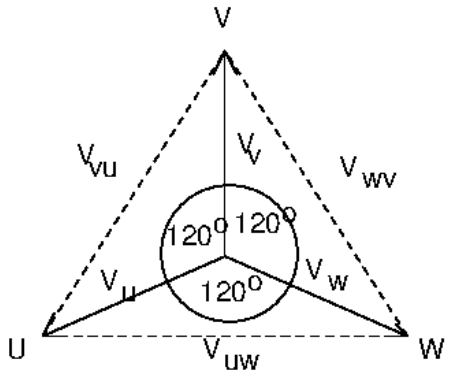


Figure 13.7: Vector diagram of a 3 phase system

The relationship between the phase (to ground) voltage and the phase to phase voltage is:

$$V_{VU} = 2 * sin(60^0) * V_{phase-ground} = \sqrt{3} * V_{phase-ground} \tag{13.8}$$

Transformers built for 3 phase systems can share a common core. If the load of the 3 phases differs too much a magnetic return path not going through the other phases is required.

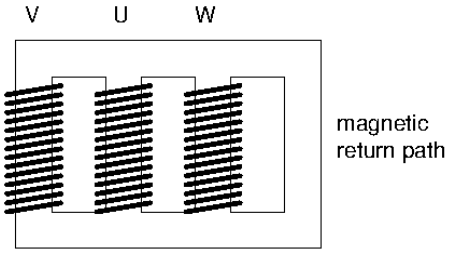


Figure 13.8: 3 phase transformer permitting an unbalanced system

The magnetic return only has to carry the difference of the 3 coils. Well managed distribution systems are balanced good enough to use transformers without the additional magnetic return path. This saves significant hardware cost.

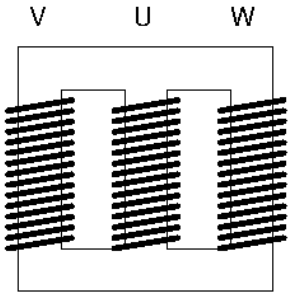


Figure 13.9: 3 phase transformer for balanced systems

This cost reduction using 3 leg transformers and balancing the distribution net is the reason why energy suppliers are so eager to compensate their loads (If there is out of phase current the magnetic return path is required again). If the net is balanced each set of windings can be regarded as a single transformer. Since the coild can be connected between the phases (triangle) as well as in a star or Y connection various configurations with different phases are possible.

The 3 phase distribution system was introduced long before it became common to use complex numbers for electrotechnical calculations. The phase shifts are always multiples of 30^0. So the inventors of the 3 phase system numbered the possible phase shifts of the transformers like the hours of the clock starting at "noon" (first reported 3 phase generator I found in literature dates back to 1887, invented by Friedrich August Haselwander). Understanding this scheme is easy looking at the vector diagram again. For this purpose let's assume we have a triangular input configuration and a star output configuration.

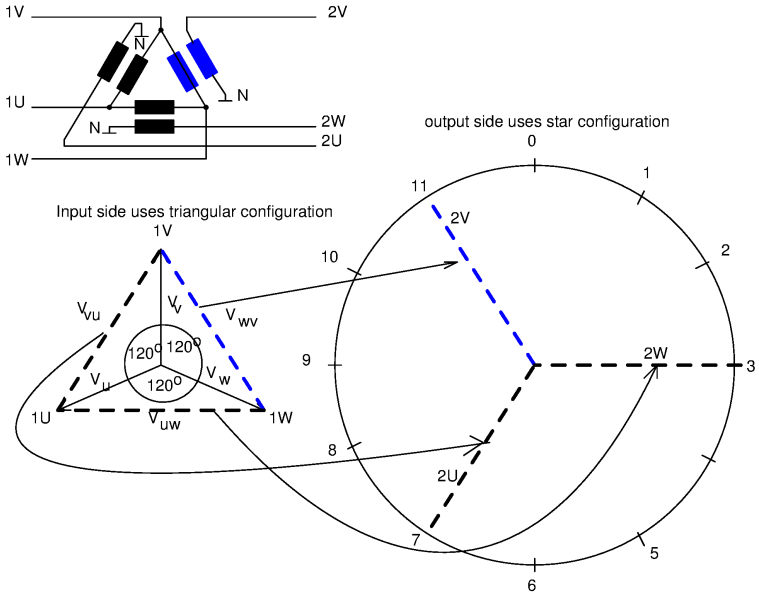


Figure 13.10: Example of a Dyn11 transformer

The reference phase (blue) is sifted left by 30^0 to "$11^o clock$".The input is a triangle marked with a D. The output is a star marked with a Y. if the ground node (N) is made available a small n is added to the name. This leads to the name of the transformer Dyn11.

Note that changing from triangle to star connection the voltage increases by factor $\sqrt{3}$ because the phase to phase voltage on the input side becomes the phase to ground voltage on the output side. Therefore Dy or Yd configurations always have a winding ratio different from 1 (well, unless somebody really wants a change of the voltage by $\sqrt{3}$ or $1/\sqrt{3}$ which is a bit uncommon.)

Reversing the polarities of the wires on one side of the transformer would make it a Dyn5, Rotating the nodes shifts the phase by $n * 120^0$ making it a Dyn0 or a Dyn7.

Reversing only one or two but not all 3 windings will destroy the 3 phase rotating system (it breaks the rotation symmetry!). This can even lead to destruction of the transformer or the generators driving it!

13.3.1 Branche currents and wire currents

The relationship between the currents of the branches (strings) and the wires depends on the configuration (traingle or star). In case of a star configuration the pin current and the current through the transformer are identical.

In case of a triangle the current flowing in the wire splits into two currents flowing into transformers. Since the transformer currents in a triangle configuration are shifted in phase by $120^0$the resulting pin current (assuming symmetrical, real load) becomes:

$$I_{pin\nabla} = \frac{I_{string}}{\sqrt{3}}$$

(This can also be derived by the power of the system. Since in a triangle configuration the string voltage is higher then the voltage from one pin to ground the current must be reduced by exactly this ration to satisfy the same power flow.)

13.4 Management of an energy distribution system

Since the cost per kWh ranges from 2 cent (depreciated nuclear power plant) to about 50c (solar energy) the producers of electrical energy are interrested in providing as much energy as possible from low cost sources (nuclear and coal).

Gas and oil are accepted for its fast regulation response.

Availability of solar energy and wind energy are difficult to predict. Usually weather forecasts combined with statistical models play an important role here.

So a regulation of a complete energy system has a cost weight for every energy source as well as a description of the regulation response of the sources. The regulation algorithme tries to minimize the production cost to be accepted to provide the energy within the tolerances of the network (target voltage, target frequency of the AC current).

13.5 Disturbances

On the way from the power plant(s) to the systems consuming the energy there are various possibilities to disturb the system. Here are some examples but the list is far from being complete!

- Thousands of kilometers of wires transporting the energy act as antennas picking up the magnetic field of the sun. Solar storms can induce thousands of volts!
- Nuclear explosions in the high atmosphere can produce strong surges due to sudden movements of ionized gas.
- Fast changes of the load of the energy distribution system will create transients on the lines.
- Emergency turn off of parts of the system can create severe pulses.
- Storms can break or short circuit wires
- Lightning strike at almost any position in the system.
- Wires will pick up signals from radio transmitters.
- In the future wireless power transfer systems may be strong sources of magnetic interference (up to $\frac{dB}{dt} = 500T/s$ at 100kHz!)
- Charged objects touching in close proximity to electronic components can lead to damage of sensitive circuits.
- Turbulent flows of plasma can create significant fields

There are two ways of determining these disturbances:

1. Knowing the characteristics (impedances) of the (intentional) contributors and the coupling of the disturbance some scenarios can be calculated or simulated.
2. Measuring transients and creating statistical distributions of the measurement reasonable test signals can be defined. (choosing energy levels high enough to cover most cases but not too high to avoid the costs of over engineering.)

Definig disturbance levels that are too low means lower protection cost but the risk of a complete system failure increases. Vice versa minimizing the risks defining high levels of energy to be handled by the protections can increase the costs dramatically.

Usually a combination of both methods is required. There are different levels of disturbance defined at every location of the energy distribution network. In the following some disturbances are described going from the highest enery level to the low levels.

13.5.1 Solar magnetic storms

What is called a solar magnetic storm is caused by exceptionally strong flares on the sun. These flares are accopmanied by strong fluctuations of the solar magnetic field. But what is more important are the ionzed particles traveling from the sun to earth. As long as these storms of charged particles are weak or moderate these particles are deflected by the magnetic field of the earth. If these storms get too strong the magnetic field of the earth gets deformed. Now these paricles can hit earth.

In most cases earth's magnetic field will not fully collaps and most of the particles hit earth close to the poles. Only in extreme cases these particles can even come down close to equator.

Strong showers of ionized particles will charge earth on one side leading to enourmous ground currents. So the ground potentials of the ends of long transmission lines in the energy grit will start to differ by thousands of volts creating excessive direct currents in the enegy supply grit. Transformes are not made for these DC currents and will go into saturation or may be destroyed directly by the DC current itself.

One example wes the solar storm hitting the energy supply of Canada 1989. The DC current built up within about 1 minute!

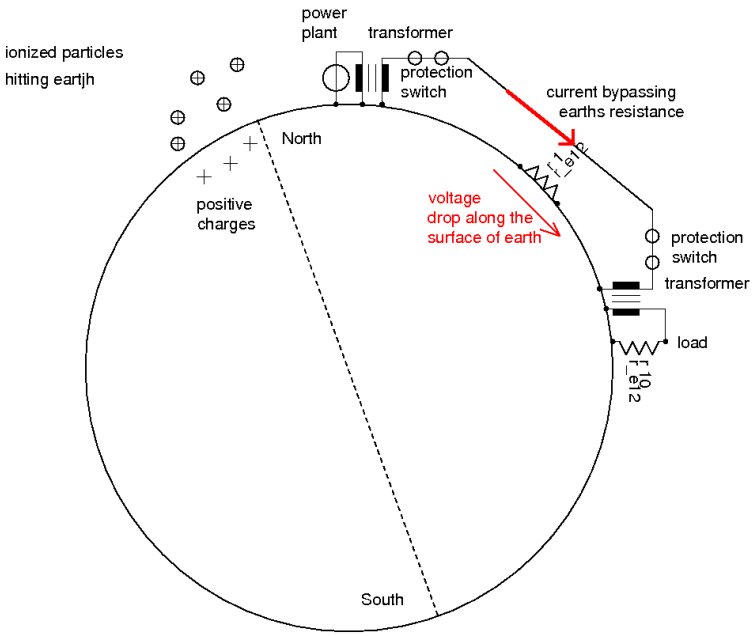


Figure 13.11: DC currents caused by a solar storm

In case of a strong solar storm the priority is to protect the transformers even if the energy supply of big regions of the earth fails. For this reason every transformer station must have circuit breakers (protection switches) able to open even if thousands of ameres are already flowing. The circuit breakers must be on both ends of the transmission line to be sure the protectin still works even if there is a connection of the transmission line to ground between the transformer stations (fallen down cables etc.). *The protection of the hardware must have absolute priority* because destroying thousands of transformers just because we did not want to interrupt the power supply would lead to damages that take years to repair. So in case of strong solar storms black outs of big areas must be accepted.

Warning time: The protons are travelling from the sun to the earth at a typical speed of 800km/s. From observing the flares to the arrival of the protons there is a delay of about two days. Normally this is time enough to intentionally break the power lines in a well controled way..

13.5.2 Nuclear Electromagnetic pulse

A scenario similar to a solar storm can be produced by a nuclear explosion in the upper athmoshere (about 40km to 200km above ground). A nuclear explosion first strips the electrons from the gas molecules of the athmosphere and pushes them down to the ground providing a strong negative pulse.

The ions are a bit heavier and slower than the electrons. These too are pushed to the ground but they arrive later producing a second half wave now with a positive polarity. The energy of a nuclear electromagnetic pulse (NEMP) in close proximity to the explosion (some ten to some hundred kilometers) can (locally) even be higher than the energy of a solar storm. The pulse typically is shorter than that produced by a solar storm (risingedge in the range of microseconds). So mechanical circuit breakers will reakct too late. Any electronics connected to a cable acting as an antenna needs extensive protection to survive such an event. Typical protections consist of multiple spark gaps (reaction time in the range of some 10ns), resistive attenuators, varistors, further attenuators and on chip ESD protection. These protections short circuit the pulse to ground until circuit breakes will react.

Preferably high reliability electronic systems that are intended to survive an NEMP should be mounted in a shielded box.

Alternatively the electronic systems could be built using valves. These are more rugged than semiconductors due to their size and resulting thermal capacity.

13.5.3 Direct lightning strike

Every cable exposed to the world outside of a building can be struck by lightning. The energy of lightnings is statistically distributed. Numbers published in literature [48]are:

Table 75: Energy of lightning strike

lightning pol.	typ. current	typ. charge	rel. frequency	highest currents observed
positive stroke	25kA	25 As	91%	400kA
negative stroke	250kA	250As	9%	

Distribution of currents is non gaussian!

13.5.4 Indirect lightning strike

Indirect lightning strike means the lightning hits something in proximity of the electronic system. The path of the lightning acts as an antenna coupling energy into the electronic system by the electromagnetic field. A typical lightning about 100m away can produce up to 100V/m. So an antenna of 1m length can deliver 100V with its typical impedance. (about 30Ω to 100Ω). 1m of cable attached to an integrated circuit wth a typical 2kV ESD protection must already be expected to damage the chip in case of an indirect lightning strike closer than 100m. (These internal protections can handle about 1.5A for 50ns if they are designed for 2kV human body model)

13.5.5 Spark discharges during operation and RF injection during operation

There are dozens of models of spark discharges during operation of an electronic system. Which model is applicable depends strongly on the application. Usually spark dicharges do not directly hit the IC.

The same applies to RF injection into a system. RF can be picked up by any wire connected to a system that reaches some 10% of the ave length of the RF considered.

There are external first level protections on the board to take the main part of the energy. These protections however are limited by theirv clamping resistance and inductance. Thus part of the energy will surpass the first level protection and arrive at the IC. The ammount of energy an IC can absorbe without damage mainly is limited by the thermal capacity of the chip internal structures. So the energy is directly related to an area and the minimum cost of the protection can be estimated knowing the die cost per area. Practical structures usually become a bit bigger than the ideally estimated area because some house keepin add ons like wide metal traces and current symmetrizing ballast resistors are needed.

A typical protection concept used for automotive systems is shown below. On the left side the signals connected to cables have fairly high levels of disturbances. As an average each wide band (100kHz to 1GHz) protection can be assumed to attenuate disturbances by 10dB (of course at mid band of the protection the attenuation is higher, but it is very difficult to for instance create a choke the works nicely from 100kHz to 1GHz. Parasitic bypasses will reduce the effectiveness of the protection outside of the range it is optimized for.)

Only very few semiconductors can directly be connected to cables (mainly power transistors and special interface ICs). Most semiconductors need at least one level of protection. Sensitive devices such as high performance analog signal pocessing chips or microcontrollers will hardly handle more than 10dBm without degradation of their parameters!

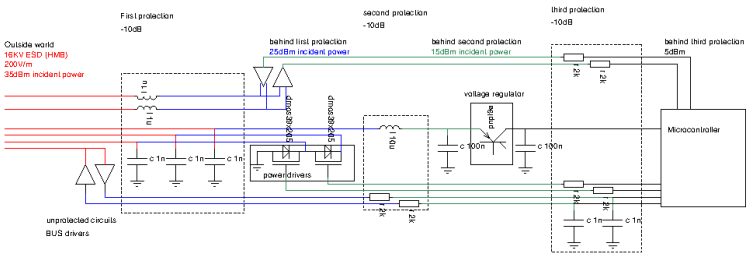


Figure 13.12: A basic board concept used to protect the semiconductors

Other fields of applications may need significantly more (for instance power grid electronics) or less (for instance consumer electronics in a well encapsulated system) protections.

Test standards usually assume worst case conditions that in practice are not encountered. The wide guard bands are requested because the coupling to a disturbance depends on so many parameters (stray capacities, stray inductances, antenna gain, frequency, distance of a source of disturbance...) that almost only empirical statistical data can be used. To achieve a high reliability of electronic systems the test limits are at 4s to 6s of the spead found in real applications.

13.5.6 Plasma induced interference

Moving plasma can produce high electric and magnetic fields. If the plasma touches conducting structures the currents can be quite high. Plasmas can be found more frequently than most people think. The following table lists some, but surely not all, plasma sources and resulting disturbances. A plasma can be regarded as a conductor. Local plasmas between two wires will act as a low resistive connection. (A welding arc has only some ohms!)

Table 78: CISPR classA conducted RF emission limits

fequency range	quasi peak	average
0.15MHz to 0.5MHz	$79dB\mu V$	$66dB\mu V$
0.5MHz to 30MHz	$73dB\mu V$	$60dB\mu V$

Table 79: CISPR class B conducted RF emission limits

fequency range	quasi peak	average
0.15MHz to 0.5MHz	$66 - 56dB\mu V$	$56 - 16dB\mu V$
0.5MHz to 5MHz	$73dB\mu V$	$60dB\mu V$
5MHz to 30MHz	$60dB\mu V$	$50dB\mu V$

Table 76: Electrical stress caused by plasma

occurrence	effect observed	currents	voltages	other remarks
electric welding	DC currents	up to 1kA	some hundred V	
plasma etching	DC currents	some A	some 10V	
exhaust of a rocket	no wireless comunication	?	?	
air plasma around a space ship	no wireless comunication	?	?	entering atmosphere

13.5.7 Fields of an inductive charging device

Currently there is not standardization of frequencies and antennas (2018). Frequencies in discussion are 70kHz and 140kHz. Antennas in discussion have an area of about $0.25m^2$. The voltage per winding needed on the receiver side is about 40..80V. This leads to an approximate estimation of the magnetic fields in close proximity of the inductive charging device:

Table 77: Fields caused by inductive charging devices

parameter	inductive charging (calculated)
frequency	70kHz and 140kHz (proposed)
dB/dt	1280T/s
peak field B_p	2.04mT
voltage induced in $1cm^2$	128mV

13.6 CISPR RF emission limits

CISPR only allows limited RF emission of electronic equipment. This applies to out of band emission (If you build a radio transmitter you WANT it to radiate - but of course only inside a certain bandwidth)

14 Physical properties of material used in semi conductor processe

14.1 Most important physical constants

constant	symbol	value	unit	remark
electron charge	e, q	$1.60217656535 * 10^{-19}$	C	1C=1As
Bolzmann constant	k	$8.617332478 * 10^{-5}$	eV/K	$V_t = \frac{k*T}{e} = 86.1733\mu V/K * T$
Dielectric constant	ϵ_0	$8.85418781762 * 10^{-12}$	As/Vm	$\frac{1}{\mu_0 * c_0^2}$
Stefan's constant	σ	$5.6703 * 10^{-8}$	W/m^2K^4	heat radiation

Table 80: CISPR class A 10-Meter radiated emission limits

frequency	field strength
30-88MHz	$39dB\mu V/m$
88-216MHz	$43.5dB\mu V/m$
216-960MHz	$46.5dB\mu V/m$
above 960MHz	$49.5dB\mu V/m$

Table 81: CISPR class B 3-Meter radiated emission limits

frequency	field strength
30-88MHz	$40dB\mu V/m$
88-216MHz	$43.5dB\mu V/m$
216-960MHz	$46.5dB\mu V/m$
above 960MHz	$54.0dB\mu V/m$

14.2 Mechanical parameters

14.2.1 Density (specific weight)

Material	symbol	value	unit	remark
C (diamond)	ρ_{cu}	$3.4 * 10^3$	$kg * m^{-3}$	
Copper	ρ_{cu}	$8.96 * 10^3$	$kg * m^{-3}$	
Iron	ρ_{fe}	$7.4 * 10^3$	$kg * m^{-3}$	
GaAs	ρ_{GaAs}	$1.86 * 10^3$	$kg * m^{-3}$	
GaN	ρ_{GaN}	$6.15 * 10^3$	$kg * m^{-3}$	
GaP	ρ_{GaP}	$4.13 * 10^3$	$kg * m^{-3}$	
Germanium	ρ_{Ge}	$5.32 * 10^3$	$kg * m^{-3}$	
InN	ρ_{InN}	$6.58 * 10^3$	$kg * m^{-3}$	
InP	ρ_{InP}	$4.81 * 10^3$	$kg * m^{-3}$	
Silicon	ρ_{si}	$2.3 * 10^3$	$kg * m^{-3}$	
SiC	ρ_{SiC}	$3.1 * 10^3$	$kg * m^{-3}$	
SiO2	ρ_{SiO2}	$2.65 * 10^3$	$kg * m^{-3}$	

14.2.2 Elasticy of materials used in semiconductor manufacturing

Mechanical stress can affect matching of devices. To achieve good matching all transistors or resistors to be matched should be exposed to the same mechanical stress. Therefore the material close to matching structures ideally should have the same elasticity (Youngs modulus) as silicon.

Material	Symbol	value	unit	remark
Aluminium		70	kN/mm^2	used for wires
C (diamond)				possible future semiconductor
Copper		100..130	kN/mm^2	used for wires
Silicon		130..190	kN/mm^2	depends on crystal orientation
SiO2		40..90	kN/mm^2	amorphous
SiO2		88	kN/mm^2	crystal (silica)

14.3 Thermal parameters

14.3.1 Thermal conductivity of various materials

Material	symbol	value	unit	remark
Iron	λ_{fe}	74	W/m K	at 300K
Copper	λ_{Co}	385	W/mK	
Diamond	λ_{C}	3320 (?)	W/mK	at 300K
GaAs	λ_{GaAs}	52	W/m K	at 300K
GaN	λ_{GaN}	230	W/m K	at 300K
Germanium	λ_{Ge}	58	W/m K	at 300K
InN	λ_{InN}	45	W/m K	
InP	λ_{InP}	68	W/m K	
SiC	λ_{SiC}	490	W/m K	valid for 6H-SiC, 4H-SiC, 3C-SiC, [26]
SiC	λ_{SiC}	250	W/m K	at 500K
Silicon	λ_{si}	150	W/m K	at 300K
Silicon	λ_{si}	110	W/m K	at 450K
Silicon oxide (SiO2)	λ_{sio_2}	0.8..1.2	W/m K	depends on oxidation process used
Silver	λ_{Ag}	406	W/mK	

14.3.2 Thermal capacity of various materials

Material	symbol	value	unit	remark
Copper	Cthcu	385	Ws/kgK	
C (diamond)	CthC	509	Ws/kgK	
Iron	Cthfe	452	Ws/kgK	
GaAs	Cthgaas	345	Ws/kgK	
GaN	Cthgan	500	Ws/kgK	
GaP	Cthgap	430	Ws/kgK	
Germanium	Cthge	310	Ws/kgK	
InN	Cthinn	320	Ws/kgK	
InP	Cthinp	310	Ws/kgK	
Silicon	Cthsi	760	Ws/kgK	
SiC	Cthsic	950	Ws/kgK	
SiO2	Cthsio2	733	Ws/kgK	

14.3.3 Thermal voltages (Seebeck coefficients)

Seebeck koefficients were taken from Wikipedia [67]

material	usage/position	thermo voltage
Si	gate contact, source contact	$\approx 440\mu V/K$
Ge	emitter contacts	$\approx 300\mu V/K$
NiCr	Nichrome	$25\mu V/K$
Fe	Pin of ICs	$19\mu V/K$
W	Tungsten vias	$7.5\mu V/K$
Au	bond wires	$6.5\mu V/K$
Ag	RF inuctors	$6.5\mu V/K$
Cu	wires, bond wires	$6.5\mu V/K$
Pb	lead solder	$4.0\mu V/K$
Al	pad, wires on chip	$3.5\mu V/K$
C	Carbon resistors	$3.0\mu V/K$
Pt	reference	0
Ni	Nickel	$-15\mu V/K$
	Constantan resistors	$-35\mu V/K$
Bi	Bismuth solder	$-72\mu V/K$

14.4 Electrical parameters

14.4.1 Most important units

unit	symbol	meaning	conversion 1	conversion 2	remark
ampere	A	electrical current			
Farad	F	capacity	$\frac{As}{V}$		
Henry	H	inductance	$\frac{Vs}{A}$		
Joule	J	energy	Ws	Nm	
Tesla	T	magnetic flux density	$\frac{Vs}{m^2}$	$\frac{N}{Am}$	magnetic flux B
			$\frac{A}{m}$		magnetic field intensity H
volt	V	electrical voltage			
watt	W	power	$V * A$	$\frac{Nm}{s}$	
BTU		british thermal unit			$1BTU = 0.2931Wh$

14.4.2 Electrical resistivity of wiring material

Material	symbol	value	unit	temp. coefficient	remark
Aluminium	r_{Al}	28.2	$n\Omega * m$	0.0039/K	wires on chip, bond wires
Copper	r_{cu}	16.78	$n\Omega * m$	0.0039/K	power wires, bond wires
Gold	r_{au}	24.4	$n\Omega * m$	0.0034/K	bond wires
Iron	r_{fe}	100.0	$n\Omega * m$	0.005/K	Die pad, springs
Silver	r_{ag}	15.9	$n\Omega * m$	0.0038/K	Platings used for RF
Tungsten	r_w	56.0	$n\Omega * m$	0.0045/K	Vias and contacts

14.4.3 Carrier mobilities

Material	type	symbol	value	unit
C (diamond)	electrons	μ_{nC}	2200	cm^2/Vs
C (diamond)	holes	μ_{pC}	1600	cm^2/Vs
Germanium	electrons	μ_{nGe}	4500	cm^2/Vs
Silicon	electrons	μ_{nsi}	600-1100	cm^2/Vs
Silicon	holes	μ_{psi}	250	cm^2/Vs
GaAs	electrons	μ_{nGaAs}	6000-8500	cm^2/Vs
GaN	electrons	μ_{nGaNs}	440 at the surface	cm^2/Vs
GaN	electrons	μ_{nGaNb}	10 vertically in the bulk	cm^2/Vs
6H-SiC	electrons	μ_{n6Hsic}	370	cm^2/Vs
6H-SiC	holes	μ_{p6Hsic}	90	cm^2/Vs
4H-SiC	electrons	μ_{n4Hsic}	800..1000 [61]	cm^2/Vs
4H-SiC	holes	μ_{p4Hsic}	115	cm^2/Vs
3C-SiC	electrons	μ_{n3Csic}	750	cm^2/Vs
3C-SiC	holes	μ_{p3Csic}	40	cm^2/Vs

14.4.4 Saturation velocity

Material	type	symbol	value	unit
C (diamond)	electrons	V_{sate}	$27 * 10^4$	m/sec
C (diamond)	holes	V_{sath}	$10 * 10^4$	m/sec
4H-SiC	electrons	V_{sate}	$20 * 10^4$	m/sec
Silicon	electrons	V_{sate}	$10 * 10^4$	m/sec

14.4.5 Typical resistivities found in semiconductor materials

Material	symbol	value range	unit	usage
Low doped substrate		3 to 20	Ohm cm	components in the substrate
High doped substrate		5 to 15	m Ohm cm	latch up hardening. Requires an epitaxy

14.4.6 Dielectric constants

relative dielectric constants.

Material	symbol	value	unit	remark
diamond (C)	ε_{rC}	5.7		
Silicon oxide (SiO2)	ε_{rsio_2}	3.9		gate capacity, wiring capacity
Silicon Nitride (Si_3N_4)	ε_{rsiN}	7.5		gate capacity, capacitors
depleted silicon	ε_{rsi}	12		junction capacity
depleted Gremanium	ε_{rge}	16		
Silicon Carbide	ε_{rsic}	9.7		[26]
Gallium Nitride (GaN)	ε_{rGaN}	10		different sources state values ranging from 6..10
PCB (poly carbonate board)	ε_{PCB}	4.7		printed circuit, interposer

absolute dielectric constants.

Material	symbol	value	unit	remark
Silicon oxide (SiO2)	ε_{sio_2}	0.34	pF/cm	$\varepsilon_0 * \varepsilon_r$
depleted silicon	ε_{si}	1	pF/cm	$\varepsilon_0 * \varepsilon_r$

14.4.7 Typical break down field strengths

Material	Symbol	value	unit	remark
diamond	C	1000	$V/\mu m$	
Silicon	Si	30-60	$V/\mu m$	[26, 28]
Germanium	Ge	10	$V/\mu m$	
Gallium Arsenide	GaAs	60	$V/\mu m$	[26]
Gallium Nitride	GaN	330	$V/\mu m$	
Silicon Carbide	6H-SiC	320	$V/\mu m$	[26]
Silicon Carbide	4H-SiC	250..300	$V/\mu m$	[26][61]
Silicon Carbide	3C-SiC	150	$V/\mu m$	[26]
Mold		200	$V/\mu m$	at this field charges start to creep in the mold

14.4.8 Table of Permeabilities

Important conversion:

$$1J = 1Ws = 1VAs = 1kg * 1m/s^2 * 1m = 1Nm$$

Material	Symbol	value	unit	remark
vacuum	μ_0	$1.2566 * 10^{-6}$	$NA^{-2} = Vs/Am$	$\mu_0 = 4 * \pi * 10^{-7} N/A^2$

14.4.9 Bandgap energies and bandgap voltages usually found at 300K

Material	symbol	bandgap ideal	bandgap voltage typically at 300K	remark
diamond	C	5.45eV		
Gallium Arsenide	GaAs	1.424eV		
Gallium Nitride	GaN	3.40eV		
Germanium	Ge	0.67eV		
Indium Phosphide	InP	1.344eV		
Silicon	Si	1.11eV	1.23V	
Silicon Carbide	6H-SiC	3.05eV		[26]
Silicon Carbide	4H-SiC	3.23eV	3.23V	[26][61]
Silicon Carbide	3C-SiC	2.36eV		[26]

15 Appendix: Mathematical rules

This chapter lists a short summary of mathematic relations used

15.1 Quadratic equations

The equation $ax^2 + bx + c = 0$ has the solutions

$$x_{1/2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{15.1}$$

15.2 Trigononetric rules

Sine and cosined add as:

$$sin^2(x) + cos^2(x) = 1 \tag{15.2}$$

15.3 Hyperbolic functions

Most important are sinh and cosh. The definitions are:

$$sinh(x) = \frac{1}{2} * (e^x - e^{-x}) \tag{15.3}$$

$$cosh(x) = \frac{1}{2} * (e^x + e^{-x}) \tag{15.4}$$

15.4 Polynom differentiation

Polynoms differentiate as:

$$\frac{d(x^n)}{dx} = n * x^{(n-1)} \tag{15.5}$$

15.5 Substitution in differentiations

This is usefull if the argument of a function again is a function:

$$\frac{d}{dx}u(v(x) = \frac{d}{dv}u(v) * \frac{dv}{dx} \tag{15.6}$$

Here comes an example:

$$\frac{d}{dx}ln(sin(x)) = \frac{d}{dv}u(v) * \frac{d}{dx}v(x)$$

with $u = ln(v)$ and $v = sin(x)$. This leads to:

$$\frac{d}{dx}ln(sin(x)) = \frac{1}{u} * cos(x) = \frac{cos(x)}{sin(x)}$$

15.6 Logarithmic differentiation and integration

This may help in some specific cases of differentiation or integration. The basic rule is:

$$\int \frac{f(x)'}{f(x)} = ln(f(x)) \tag{15.7}$$

To check, if this method can be used have a look at the nominator. Is the nominator the differeciated denominator?
Example of usage: $f(x) = x^{x+1}$
First step: we simply differentiate the equation and then plug in the function we want to solve.

$$\frac{f(x)'}{f(x)} = \frac{d}{dx}ln(f(x)) = \frac{d}{dx}ln(x^{x+1}) \tag{15.8}$$

$$f(x)' = f(x) * \frac{d}{dx}((x+1) * ln(x))$$

$$f(x)' = f(x) * (ln(x) + (\frac{x+1}{x})$$

$$f(x)' = x^x * (x * ln(x) + x + 1)$$

15.7 differentiation product rule

in the following equations the following abreviation is done:

$$\frac{df(x)}{dx} = f(x)' = f'$$

A product of functions differentiates as

$$(f_1 * f_2)' = f_1' * f_2 + f_1 * f_2' \tag{15.9}$$

15.8 differential quotient rule

A quotient of two functions can be differentiated using the following rule:

$$(\frac{f_1}{f_2})' = \frac{f_1' * f_2 - f_1 * f_2'}{f_2^2} \tag{15.10}$$

15.9 Logarithmic differentiation

Logarithmic differentiation used the relationship

$$\frac{f(x)'}{f(x)} = \frac{dln(f(x))}{dx} \tag{15.11}$$

Looks a bit strange. So let's have a look at an example:

$$f(x) = x^x$$

$$\frac{df(x)/dx}{x^x} = \frac{d}{dx}(ln(x^x)) = \frac{d}{dx}(x * ln(x)) = ln(x^x) + x * \frac{1}{x}$$

$$dx^x/dx = x^x * (ln(x^x) + 1)$$

This complicated approach applies not only to exotic functions such as x^x but also to more simple functions. However nobody will use this complicated approach for things like polynomial functions. (But it works there too.)

15.10 Basic integration rules

Integrating polynoms is the easiest. It can be deen directly from the corresponding differentiation rule:

$$\int x^n dx = \frac{1}{n+1} * x^{n+1} \tag{15.12}$$

15.11 Integration using substitution

This is simply the inversion of the differentiation with substitution:

$$\int u(v(x))dx = \int udv \tag{15.13}$$

With $dv = v' * dx$. This can be rewritten:

$$\int u(v(x))dx = \int u(v(x)) * v(x)' * dx = \int udv \tag{15.14}$$

Lets make an example:

$$\int sin^2(x) * cos(x)dx$$

Substitution: $u = v^2$, $v = sin(x)$, $v' = cos(x)$ leads to:

$$\int sin^2(x) * cos(x)dx = \int v^2 dv = \frac{1}{3}v^3 = \frac{1}{3}sin^3(x)$$

An interresting case of this substituton is v'=constant. Then both sides may be divided by v'. Here is an example

$$\int sin(kx)dx = \int sin(v)dv$$

with v=k*x and v'=k we get:

$$\int sin(kx) * k * dx = \int sinvdv$$

$$k * \int sin(kx)dx = cos(v)$$

$$\int sin(kx)dx = \frac{1}{k} * cos(kx)$$

Important: Taking v'=k from one ide to the other only is permitted as long as k is a constant!

15.12 Using complex numbers

complex numbers consist of 2 components. The real part and the imaginary part. They can be regarded as vectors sitting in a complex plain. Adding and substracting complex numbers is easy using cartesian coordinates. It is like adding or substracting vectors. The following picture show some examples of complex numbers.

$$A = x_a + y_a j$$

$$B = x_b + y_b j$$

$$C = A + B = x_a + x_b + y_a j + y_b j \tag{15.15}$$

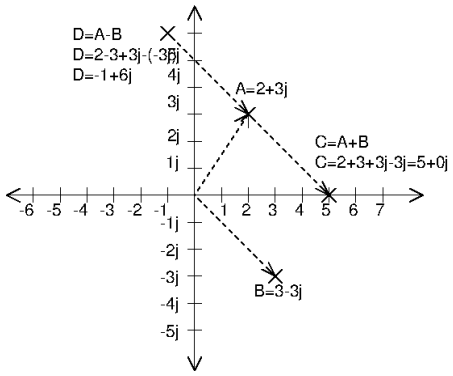


Figure 15.1: Graphical representation of adding complex numbers

The two parts of a complex number A are called the real part Re(A) and the imaginary part of Im(A)

$$A = 2 + 3j$$

$$Re(A) = 2$$

$$Im(A) = 3$$

The length of the 'vector' is called the absolute of the complex number. It can be calculated using the Pythagoras equation.

$$abs^2(A) = Re^2(A) + Im^2(A)$$

Going back to out example we get:

$$abs^2(A) = 2^2 + 3^2 = 13$$

$$abs(A) = \sqrt{13}$$

15.12.1 Multiplication of complex numbers

Multiplication of complex numbers is a bit like multiplication of polynoms. There however is one special thing to consider:

$$j * j = -1$$

The square of the imaginary part is -1 instead of 1!

Thus a multiplication in cartesian coordinates looks like this:

$$E = A * B = (2 + 3j) * (3 - 3i) = 15 + 3j$$

Not very comfortable! Let's see what happened in circular coordinates.

The absolute values are:

$$abs(A) = \sqrt{13} = 3.6056$$

$$abs(B) = \sqrt{18} = 4.2426$$

$$abs(E) = \sqrt{234} = 15.297$$

and surprise:

$$abs(A) * abs(B) = 15.297$$

This means the 'vector' legths multiply.

$$abs(A * B) = abs(A) * abs(B)$$

Now lets have a look at the angles:

$$arctan(\frac{Im(A)}{Re(A)}) = 0.98279$$

$$arctan(\frac{Im(B)}{Re(B)}) = -0.78540$$

$$arctan(\frac{Im(E)}{Re(E)}) = 0.19740$$

Which means, the multiplication adds the angles:

$$arctan(\frac{Im(A)}{Re(A)}) + arctan(\frac{Im(B)}{Re(B)}) = 0.98279 - 0.78540 = 0.19739$$

Conclusion: Representing the complex numbers by circular coordinates abs(A) and the angle is simplifying multiplication.

$$E = A * B$$

$$abs(E) = abs(A * B) = abs(A) * abs(B) \tag{15.16}$$

$$arctan(\frac{Im(E)}{Re(E)}) = arctan(\frac{Im(A)}{Re(A)}) * arctan(\frac{Im(B)}{Re(B)}) \tag{15.17}$$

15.12.2 Using complex numbers to describe reactances

A reactance is a path in which current and voltage are out of phase. The components creating "out of phase" currents are capacitors and inductors. Assuming a sine wave voltage across the component we get:

Capacitor:

$$V_c(t) = V_p * sin(\omega t)$$

$$I_c(t) = C * \frac{dV_c(t)}{dt} = C * \omega * cos(\omega t) \tag{15.18}$$

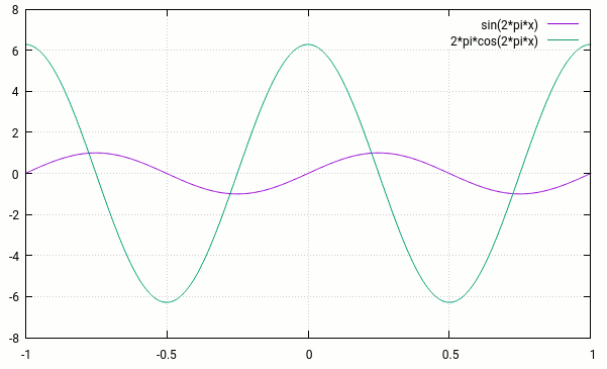


Figure 15.2: Sine wave and dx/dt at $\omega = 2 * \pi$

Inductor:

$$V_l(t) = V_p * sin(\omega t)$$

$$I_l(t) = \frac{1}{L} \int V_l(t) dt = -\frac{V_p}{L * \omega} * cos(\omega t) \tag{15.19}$$

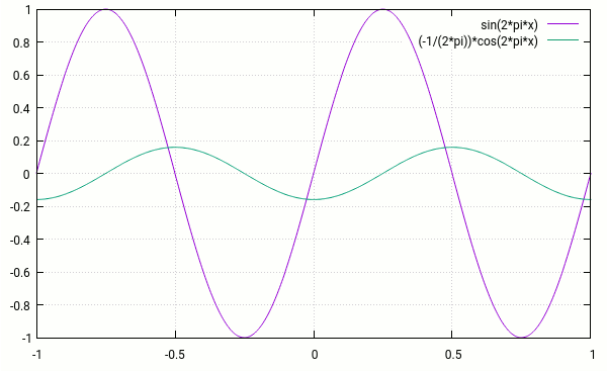


Figure 15.3: Sine wave and $\int sin(\omega x) dx$ at $\omega = 2 * \pi$

The voltage is a sine wave and the currents are cosine waves of either polarity. This is something that can be described using complex numbers as well. We simply have to assigne the sine to the real axis and the cosine to the imaginary axis. (The multiplication with j simply shifts the phase by 90 degrees).

$$cos(\omega t) \rightarrow j * sin(\omega t)$$

$$-cos(\omega t) \rightarrow \frac{1}{j} * sin(\omega t)$$

(remember j*j=-1, -1/j=j, 1/j=-j)

Next step we replace $V_p * sin(\omega t)$ by the voltage V. This transforms the irrents of the above equations to:

$$I_c = j\omega C * V \tag{15.20}$$

$$I_l = \frac{1}{j\omega L} * V = \frac{-j}{\omega L} * V \tag{15.21}$$

The phase (relative to the voltage) now is described by the j in the nummerator or in the denominator.

To given an example of the usage let's assume we have a series of a resistor, a capacitor and an inductor.

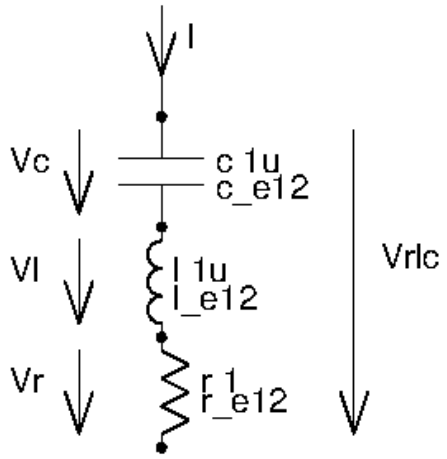


Figure 15.4: circuit to be calculated

Using the complex notation we get the following equations:

$$V_r = I * R$$
$$V_l = j * \omega * L$$
$$V_c = \frac{1}{j * \omega * C} = \frac{-j}{\omega * C}$$

This leads to:

$$V_{rlc} = I * (R + j(\omega L - \frac{1}{\omega C})) \tag{15.22}$$

Sweeping the frequency the real part remains constant while the imaginary par changes from a negative value (at low frequency) to a positive value (for high frequencies). This can nicely be demonstrated using a little octave script.

```
>> R=1
R =  1
>> L=1e-6
L =     1.0000e-06
>> C=1e-6
C =     1.0000e-06
>>w=1e5:1e5:1e7
>> Z=R+j*(w*L-1./(w*C))
>> plot (Z)
```

This leads to the following plot:

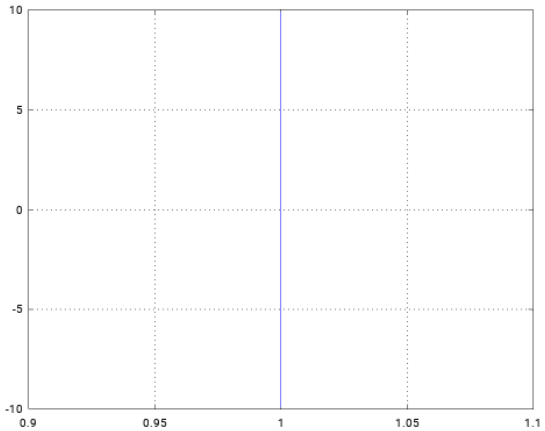


Figure 15.5: complex impedance sweeping the frequency. The horizontal axis is the real part of the impedance. The vertical axis is the imaginary part of the impedance

The frequency at which the real part of the impedance crosses the zero ($\omega L = \frac{1}{\omega C}$) is called the serial resonance of the circuit.

15.13 Laplace transformation

Laplace transformation is a mathematical method replace differential equations by multiplication and division.To do this a time domain equatiion has to be converted into a complex frequency domain equation. In the frequency comain an integration is a simple phase shift of the sine wave by 90 degres and a multiplication of the amplitiude by $1/\omega$. Vice versa a differentiation is a shift by -90 degres and a multiplication of the amplitude with factor ω .

The most simpe case is the differetiation of a sine wave.

Time domain:

$$\frac{d}{dt}sin(\omega t) = \omega * cos(\omega t) = \omega * sin(\omega t + \frac{\pi}{2})$$

The more general approach is to use a complex description of the signal:

$$\frac{d}{dt}e^{j\omega t} = j\omega e^{j\omega t}$$

And:

$$\int sin(\omega t)dt = -\frac{1}{\omega}cos(\omega t) = \frac{1}{\omega}sin(\omega t - \frac{\pi}{2})$$

$$\int e^{j\omega t}dt = \frac{1}{j\omega}e^{j\omega t} = -\frac{j}{\omega t}e^{j\omega t}$$

Frequency domain: Conversion to frequency domain:

$$X(s) =_0 \int^{\infty} x(t) * e^{-st}dt \tag{15.23}$$

with $s = j\omega + \sigma$.

Conversion back:

$$x(t) = \frac{1}{2\pi j} *_{\sigma - j\infty} \int^{\sigma + j\infty} X(s) * e^{st}ds \tag{15.24}$$

Not yet intuitive. So let's just use it.

Time domain:

$$C_1y(t) + C_2y'(t) + C_3y''(t) = K_1x(t) + K_2x'(t) + K_3x''(t)$$

Frequency domain uses multiplication with s instead of differentiation:

$$C_1 * y(s) + C_2 * s * y(s) + C_3 * s^2 + y(s) = K_1 * x(s) + K_2 * s * x(s) + K_3 * s^2 * x(s)$$

In frequency domain we can immediatelly write down the transferfunction:

$$G(s) = \frac{y(s)}{x(s)} = \frac{K_1 + K_2 * s + K_3 * s^2}{C_1 + C_2 * s + C_3 * s^2}$$

15.13.1 Stability of a transfer function:

Instability means there is a 0 in the denominator with positive real part while the nummerator isn't 0. So we simply have to check the denominator for zeros. Thus we just have to find the poles of the function. (Zero in the denominator).

Zeros of the denominator with negative real part are damped. So the function decays with time.

A Zero of the denominator with real part equal 0 means we have a marginally stable system. As long as it isn't excited it remains as it is. As soon as the system gets excited the oscillation continues.

A zero of the denominator with positive real part is instable. The oscillation increases with time.

15.13.2 Calculation of the response to an excitation in the frequency domain

The transfer function G(s) can be excited with an input function x(s). The resulting response at the output calculates as:

$$y(s) = x(s) * G(s)$$

To use this tool we need the correspondence of time domain functions and frequency domain functions.

Table 82: correspondence between time domain and frequency domain

signal	f(t)	f(s)
dirac pulse	$\delta_0(t)$	1
step	$\delta_1(t)$	$\frac{1}{s}$
ramp	$y(t) = t$	$\frac{1}{s^2}$
parabula order n	$y(t) = t^n$	$\frac{1}{s^{n+1}}$
decaying exponential	$y(t) = e^{-\delta t}$	$\frac{1}{s+\delta}$
sinwave, cosine	$y(t) = sin(\omega t)$	$\frac{1}{s^2+\omega^2}$
damped sine	$y(t) = e^{-\delta t} * sin(\omega t)$	$\frac{1}{(s-\delta)^2+\omega^2}$

15.13.3 Limits of the Laplace transformation

To calculate the time domain signal of a function we only know the Laplace transformed is easy for $t \to 0$ and for $t \to \infty$.

$$x(t \to +0) = lim_{s\to\infty}\{s * X(s)\} \tag{15.25}$$

$$x(t \to \infty) = lim_{s\to 0}\{s * X(s)\} \tag{15.26}$$

List of Figures

List of Algorithms

List of Tables

Index

F

G

H

X

Y

Z

References

[1] http://www.staticfreesoft.com/jmanual/

[2] http://opencircuitdesign.com/xcircuit/

[3] P. van Zant, "Microchip Fabrication", McGraw-Hill, 2000

[4] Paul R. Gray, Robert G. Meyer, "Analysis and Design of Analog Integrated Circuits", John Wiley & Sons, 1984

[5] Michael Quirk, Julian Serda, "Semiconductor Manufacturing Technology (Instructor's Manual), http://smt-book.com/instructor_guide.pdf

[6] Albert Z. H. Wang, "On-chip ESD Protection for Integrated Circuits", Kluwer, 2002

[7] Randall L. Geiger, Phillip E. Allen, Noel R. Strader, "VLSI Design Techniques for Analog and Digital Circuits", McGraw Hill, 1990

[8] A. von Weiss, H. Kleinwaechter, "Uebersicht ueber die theoretische Elektrotechnik", Akademische Verlagsgesellschaft Geest & Portig KG Leipzig, 1956

[9] C. Bond, "Problems and Solutions in Mathematics, Physics and Applied Sciences"

[10] Willy M.C. Sansen, "Analog Design Essentials", Springer, 2006

[11] Christophe P. Basso, "Switch mode power supplies", McGraw-Hill, 2008

[12] Jan Sonsky, "Method of manufacturing a semiconductor device with an isolated region and a device manufactured by the method", US Patent US2008/0217653 A1, Sept. 11, 2008

[13] A. Paul Brokaw, "A simple Three-Terminal IC Bandgap Reference", IEEE Journal of solid state circuits, VOL. sc9, no. 6, December 1974

[14] Stephen K. Michalich, Thomas S. W. Wong, "CMOS schmitt trigger and oscillator", US Patent US4295062, Oct. 13 1981

[15] Mansour Izadinia, Tamas Szepesi, "Precision oscillator circuit", US Patent US4904960, Feb. 27 1990

[16] Bronstein, "Taschenbuch der Mathematik"

[17] Michael B. Steer, "SPICE User's guide and reference", m.b.steer@ieee.org, 2007

[18] Paolo Nenzi, Holger Vogt, "Ngspice user manual version 24", http://ngspice.sourceforge.net, 2012

[19] Reza Moazzami, Jack Lee, Ih-Chin Chen, Chenming Hu, "Projecting the minimum acceptable oxide thickness for time-dependent dielectric breakdown", Department of electrical engineering and computer sciences University of California, Berkeley, CA 94720, 1988

[20] James R. Black, "Electromigration - A Brief Survey and some Recent Results", IEEE Transaction of Electron Devices, Vol. ED16, NO. 4, April 1969

[21] R.L. de Orio, H. Ceric, S. Selbherr,"Physically based models of electromigration: From Black's equation to TCAD models", Microelectronics reliability 50 (2010) 775-789

[22] Boon-Khim Liew, Nathan W. Cheung, Chenming Hu, "Projecting Interconnect Electromigration Life Time for Arbitrary Current Waveforms", IEEE Transactions on Electron Devices, Vol 37, No. 5, May 1990

[23] Jee Yong Kim, "Investigation of the Mechanism of Interface Electromigration in Copper Thin Films", December 2006, University of Texas at Arlington

[24] Marcel Pelgrom, "Low Power Analog IC Design", Lecture held at EPFL Premises, Lausanne, Switzerland, July 1-5, 2013

[25] Ki-Ju Baek, Kee-Yeol Na, Jeong-Hyeon Park, Yeong-Seuk Kim, "Suppression Techniques of Subthreshold Hump Effect for High-Voltage MOSFET", Journal of Semiconductor Technology and Science, Vol 13, No 5, October 2013

[26] Indra Mani Sharma, "Analysis of Depletion Region Width and Breakdown Voltages of 6H-SiC DIMOSFET Using Linearly Graded Profile in the Drift Region ", Thapar University Patiala - 147004, Punjab, India, July 2011.

[27] M. Kamon, L.M. Silveira, C. Smithhisler, J. White, "FastHenry USER'S GUIDE", Research Laboratory of Electronics Department of Electrical Engineering and Computer Science, Massachusets Institute of Technology, Cambridge, MA 02139 U.S.A., Nov. 11 1998

[28] R. Mueller, “Grundlagen der Halbleiter Elektronik”, Springer, 1987

[29] ”Linear Integrated Circuits Data Book”, Motorola Semiconductor Products Inc., 1971

[30] ”Operating and service manual 214A pulse generator”, Hewlet Packard Company, Sept 1964

[31] ”Introduction to Electromagnetic Compatibility”, R. Paul Clayton, Wiley & Sons, 2006

[32] ”Designing Analog Chips”, Hans Camenzind, February 2005

[33] ”Schaltungen mit Halbleiterbauelementen Band 2”, Erich Gelder, Walter Hirschmann, Siemens AG, 1965

[34] ”Schaltungen mit Halbleiterbauelementen Band 4”, Erich Gelder, Walter Hirschmann, Siemens AG, 1970

[35] ”AN-140 CMOS schmitt trigger - a uniquely versatile design component”, National semiconductors, 1977

[36] ”Schaltungen mit Halbleiterbauelementen Band 3”, Erich Gelder, Walter Hirschmann, Siemens AG, 1967

[37] ”Analysis and Breakdown Voltage and On Resistance of Super-Junction Power MOSFET CoolMOS Using Theory of Novel Voltage Sustaining Layer”, p.N. Kondekar, C.D. Parikh, M.B. Patil, Indian Institute of Technology, Bombay, 2008

[38] ”Power Electronics: Converter, Applications and Design”, Mohan, Undeland, Robbins, John Wiley & Sons, 1989

[39] “Simulation Techniques and Solutions for Mixed Signal Coupling in Integrated Circuits”, Verghese, Schmerbeck, Allstot, Kluwer, 1995

[40] http://www.learnemc.com/tutorials/Time_Frequency/Time_Frequency_notes.html

[41] http://www.wcalc.sourceforge.net

[42] http://www.fastfieldsolvers.com/products.htm#fasthenry2

[43] ”Electrostatically telescoping nanotube nonvolatile memory device”, Jeong Won Kang, Qing Jiang, Department of Mechanical Emgineering, University of California, Riverside, CA 92521, USA, 2007

[44] ”Nonphotolitographic Nanoscale Memory Density Prospects”, Andre DeHon, Seth Copen Goldstein, Philip J. Kuekes, Paatrick Lincoln, IEEE transactions of nanotechnology, Vol. 4, No. 2, March 2005

[45] “Infineon’s 1200V SiC JFET - The New Way of Efficient and Reliable High Voltage Switching”, Wolfgang Bergner, Fanny Bjoerk, Daniel Doms, Gerald Deboy, Infineon Technologies Austria, 2015

[46] ”Linear and Interface Circuit Applications”, Volume 3, D.E.Pippenger, E.J. Tobaben, Texas Instuments, 1988

[47] “OP AMP Applications”, Walter G Jung, Analog Devices Inc. 2002

[48] ”Atmospheric Electrostatics”, Lars Wahlin, Colutron Research Corporation, 1989 (ISBN 0 471 91202 6 (Wiley))

[49] http://www.murata.com/en-global/tool/netlist/mlcc?intcid5=com_xxx_xxx_cmn_nv_xxx

[50] ”Audio Power Amplifier Design”, Douglas Self, Focal Press, 2013

[51] ”Taschenbuch der Physik”, Horst Kuchling, Verlag Harri Deutsch, 1981

[52] ”Nonuniform Transmission Lines”, Paolo Menti, Drie Vande Ginste, Daniel de Zutter, Gent University, 1999

[53] ”Frequency Response of Thin Film Chip Resistors”, Vishay INC, 2009

[54] ”Smart Systems for Safe, Clean and Automated Vehicles”, Faical Turki, Andre’ Körner, Paul Vahle GmbH and Hella KG, 2014

[55] ”Electromagnetic Compatibility Specification for Electrical/Electronic Components and Subsystems”, Ford Motor Corporation, 2016

[56] ”ICNIRP Guidelines for limiting exposure to time-varying electric, magnetic and electromagnetic fields”

[57] ”Laser diodes, an introduction”, Matthias Pospiech, Sha Liu, University of Hannover, 2004

[58] ”Handbook of optics, chapter 13, semiconductor lasers”, Derrry, Figueroa, Hong, Boeing Defense & Space Group Seattle Washington

[59] “Engineering Electromagnetism”, A. J. Baden Fuller, Wiley & Sons, 1993

[60] ”On-Chip ESD protections for integrated circuits”, Albert Z. H. Wang, Springer, 2013

[61] “AN2017-46_CoolSiC 1200V SiC MOSFET”, Infineon AG, 2018

[62] “A Dual Pixel CMOS CCD Image Sensor”, Shinji Uya, Jun Hasegawa, Jin Murayama, Testuno Yamada, Fujifilm Microdevices Co. Ltd VLSI Department, 1999

[63] “Silicon Analog Components”, Badih El-Kareh, Lou N. Hutter, Springer, 2015

[64] “ICs fuer die Unterhaltungselektronik”, Siemens AG, 1986

[65] “EU energy in figures statistical pocket book 2017” European Union, 2017

[66] “Intersil SE Spezial-Electronic”, Spezial-Electronic, 1976

[67] “Thermoelektrische Spannungsreihe”, https://de.wikipedia.org/wiki/Thermoelektrische_Spannungsreihe, snapshot June 19 2019

[68] “Dioden”, Siemens AG, 1980

[69] “Bauelemete der Halbleiter Elektronik (Halbleiter Elektronik 2)”, R. Mueller, Springer 1987

[70] “Metal-dielectric band alignment and its implications for metal gate complementary metal-oxide-semiconductor technology”, Yee-chia Yeo, Tsu-Jae King, Chenming Hu, Journal of applied physics volume 92 No.12 2002

[71] “Product information CF190 - PSI5 receiver”, Robert BOSCH GmbH, 11/2010

[72] “AS5172A/AS5172B AMS data sheet”, Austria Micro Systems, 2017

[73] “Temperature Acceleration of Time-Dependent Dielectric Breakdown”, Reza Mouzzami, Jack Lee, Chenming Hu, IEEE Transactions on Electron Devices, Vol.36, No.11, November 1989

[74] “Smart power ICs”, Bruno Murari, Franco Bertotti, Guiovanni A. Vignola, Springer 2002

[75] “Electrical Performance of Packages”, Application note AN-1205, Nat. Semiconductors 2003

[76] “FAST Applications Handbook”, National Semiconductor, 1987

[77] “Halbleiter Schaltungstechnik”, Tietze, Schenk, Springer, 1993

[78] “Praktikum Automatisierungstechnik (Praktikumsskript)”, Prof. Dr. G. Schmidt, TUM 1985

[79] William R. Blood Jr. “MECL system design handbook”, Motorola Semiconductor Products Inc., 1971

[80] Maricau, E. ; Gielen, G. “Analog IC Reliability in Nanometer CMOS”, Springer, 2013

[81] s. Mahapatra, M.A. Alam, P. Bharath Kumar, T.R. Dalei, D. Varghese, D. Saha, “Negative Bias Temperature Instability in CMOS Devices”, preprint, 2020

[82] Mustapha Jammal, “Stand und Technik der Anwendung von Superkondensatoren”,Grin Verlag, 2009

[83] J.D. Cockcroft, E.T.S. Walton, “Experiments with High Velocity Positive Ions”, 1932

[84] National Semiconductor, “Linear Databook”, 1982

[85] Lianghui Ding, Kehong Chen, Falong Huang, Feng Yang and Liang Qian, “Modeling and Evaluation of Piezzoelectric Transducer (PZT)-Based Through-Metal Energy and Data Transfer”, Sensors 2020, 20, 3304;doi:10.3390/s20113304