Operational Amplifiers part 2: bread & butter for low supply voltage

May 24, 2019

1 Bread & butter OPAMP for low supply voltage

Reducing the supply voltage below about 2.5V narrows down the common mode range too much for the conventional design used up to 5V. 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.

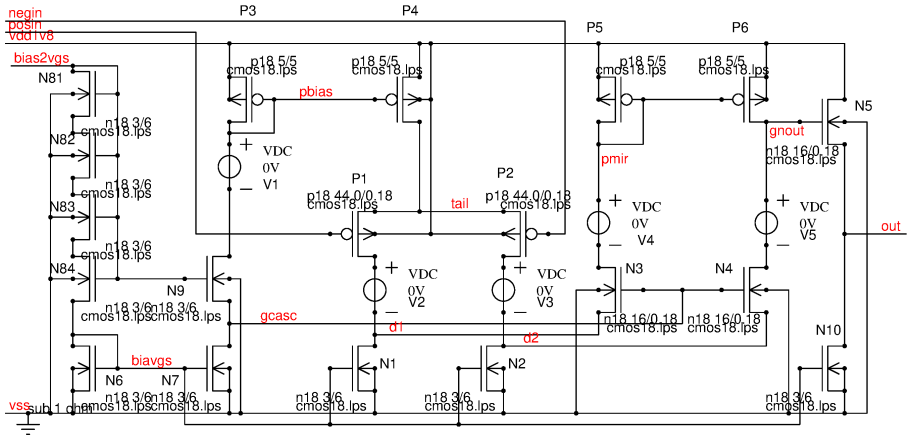

Fig.1: The low voltage bread & butter OPAMP

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 [**?**, **?**]. 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: $V_{ds} < V_{gs} - V_{th} = V_{gseff}$

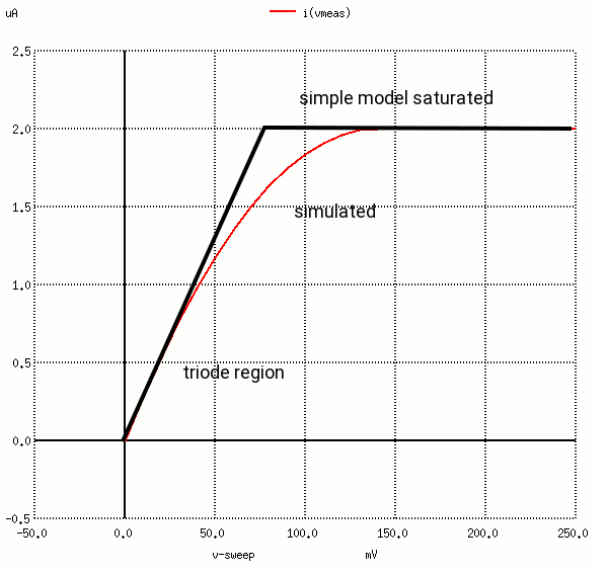

Fig.2: Operating ranges of an NMOS transistor working as a current generator

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.

$$R_{triode} \approx \frac{V_{dssat}}{I_{dsat}} = \frac{V_{gseff}}{I_{dsat}}$$

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

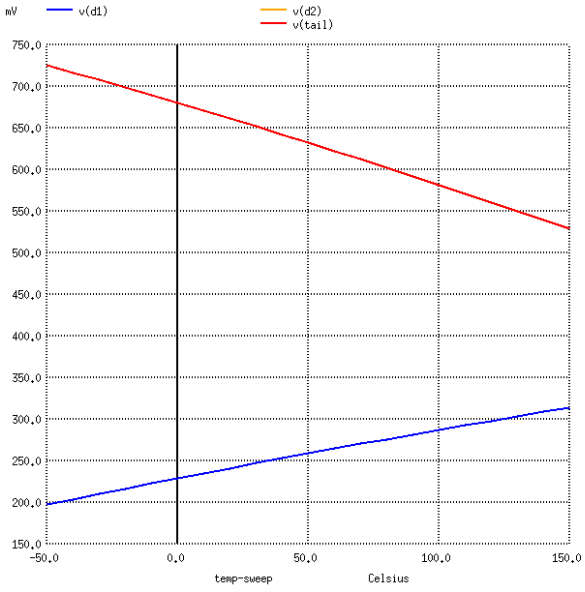

Fig.3: Temperature sweep of the source and drain voltage of the differential amplifier

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.

$$I_{error} = V_{os} * gm = \frac{V_{osproc}}{\sqrt{W * L}} * 2 * \sqrt{\frac{W * K' * I_d}{L}} = 2 * V_{osproc} * \frac{\sqrt{K' * I_d}}{L}$$

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