Operational Amplifiers part 4: rail to rail output

June 10, 2019

The last post described a rail to rail input stage and current adding at the output. So basically this was an OTA (Operational Transconductance Amplifier)

Now we want to an a more powerful rail to rail output stage to the already existing design. The most simple approach is simply adding an inverter as an output stage (and swapping the names of the inputs due to the inversion we added. This approach has the following properties:

- It is simple
- It is fast
- It is easy to compensate
- current consumption is disastrous!

But at least it is a first idea. The current consumption is highest exactly in the middle of the output swing. This usually is the most frequently encountered operating condition.

The frequency compensation is fairly simple. A miller compensation over the last stage. Well. it works, but it is wasting energy like hell.

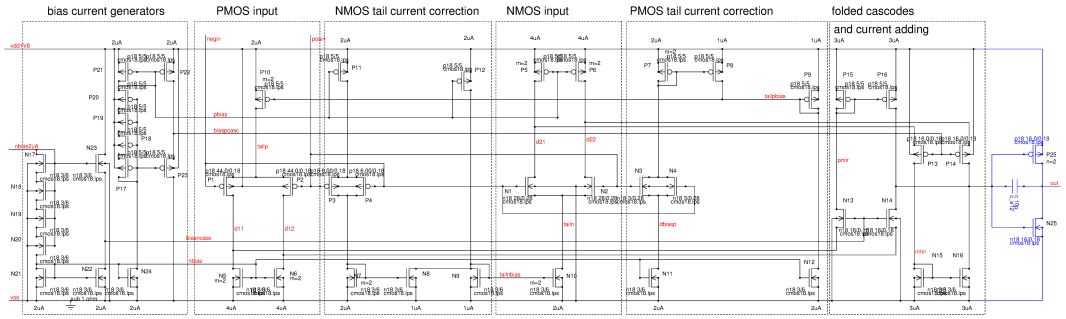

Fig.1: Amplifier with inverter output

1 A translinear output stage to reduce current consumption

What we need is a design that consumes less cross conduction current. One solution is the use of a translinear stage. It has a more or less quadratic DC transfer function. Here comes the concept:

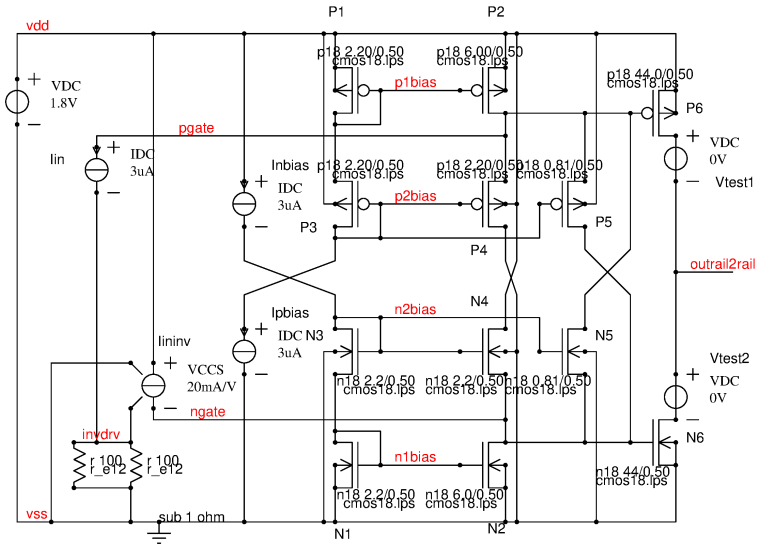

Fig. 2: Rail to rail output using translinear stages

Instead of just driving the output stage with one current (that can have both polarities) like when we are simply using an inverter this stage needs two drive currents. In the balanced operating point both drive currents at net pgate and ngate are equal except for the sign.

To increase the output voltage Iin1 has to be increased and at the same time the current into net ngate (driven by the voltage controlled current source Iininv is decreased. The gate of P6 will be pulled low and at the same time the gate of N6 will be pulled low.

To decrease the output voltage Iin must be decreased and the current in net ngate will be increase. This pulls the gate on N6 up and at the same time P6 turns off.

In the middle of the operational range there still is some cross conduction, but this current is by far lower than the cross conduction current flowing in the inverter shown before.

To test this behavior the following simulation hold the output in the middle of the swing. The currents flowing through P6 and M6 is measured while the current of the control source Iin is sweeped.

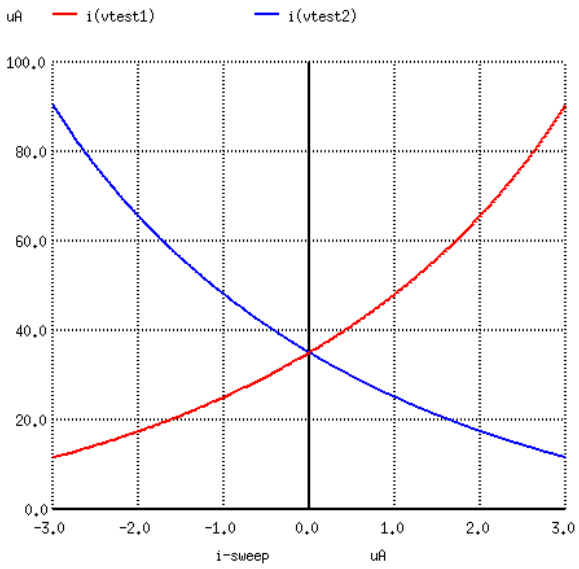

Fig.3: Currents through P6 and N6 sweeping the input current

This figure shows the following:

1. In the middle of the range the currents cross at 36μA. This is the cross conduction when there is no input signal (current=0)

2. At the extremes (drive current=-3μA and drive current $= +3\mu A$) the current in one of the transistors reaches 30uA while the current flowing through the opposite side drop down to 12μA.

3. The output drive capability is the difference between the sourcing and the sinking transistor. It is about 80μA

This means we have created an output stage that has a cross conduction of 36$\mu A$and an output drive capability of $\pm 80\mu A$. A second observation is that the transfer curves are approximately quadratic. The following plot shows the output current flowing into a 0.9V source holding the output while the input currents are sweeped.

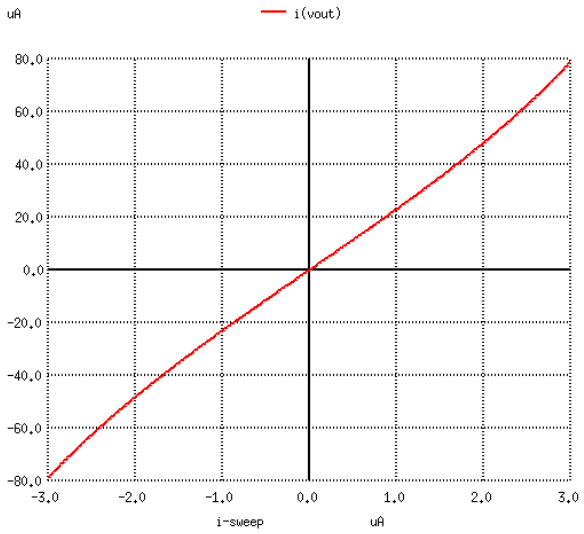

Fig.4: Output current versus input current

For the design this means the frequency compensation has to be designed for the corners with the highest current gain. At the same time the small signal accuracy depends on the low current gain in the middle of the curve (at input current close to 0). On one hand for efficiency reasons it is tempting to minimize the cross conduction compared to the current capability at the extreme ends. on the other hand this leads to a loss of gain and a loss of accuracy operating in the more typical operation point in the middle. In most cases it is required to find a reasonable compromise more or less close to the scaling shown here.

In the first draft output stage the driving current can become positive as well as negative. In the practical design usually we can either sink current (NMOS driver) or source current (PMOS driver). To achieve both polarities we have to substract or add a constant current. The circuit has to be extended a little bit. The blue part of the following circuit is the original folded cascode of the previously used rail to rail input. The output is cut open into two outputs to drive the two inputs of the translinear stage. The translinear stage has been enhanced with a current sink and a current source to make the input signal bipolar ($\pm 3\mu A$) to match it with the requirements of the translinear stage.

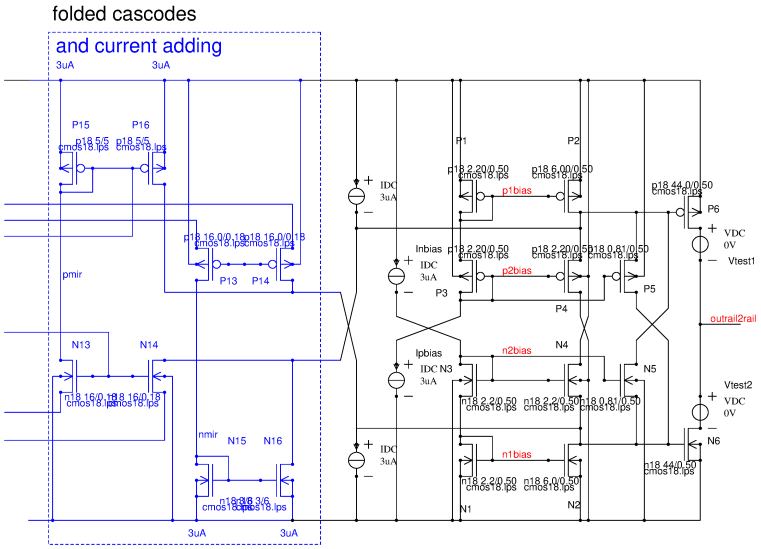

Fig.5: connecting the rail to rail input stage folded cascodes to the translinear output stage

2 Appendix: Calculation of the translinear stage

To calculate the translinear stage let's just look at one side of it. To make it a complete push pull stage a complementary PMOS translinear stage has to be added.

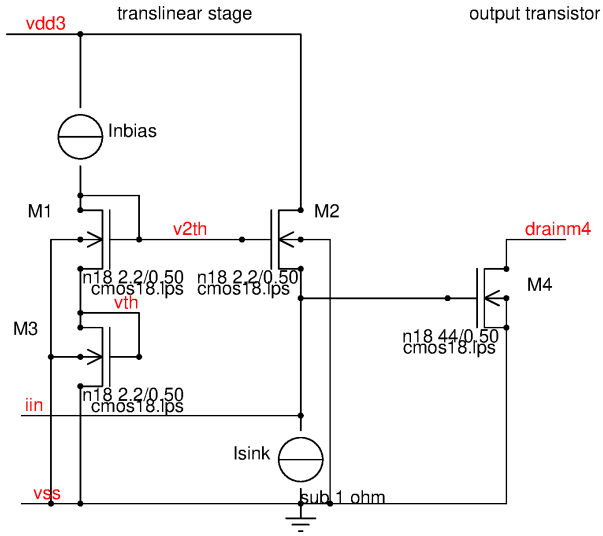

Fig. 7: Example of one half of a rail to rail output stage

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:

$$I_{drainm4} = I_{nbias} * \frac{W_{M4} * L_{M3}}{W_{M3} * L_{M4}} \tag{1}$$

Example: Isink=Inbias=3μA, W/L of M4 is 88, W/L of M3 is 4.4 leads to an operating point of M4 of

$$I_{drainm4} = 3\mu A * \frac{44}{2.2} = 60\mu A$$

If a current is applied at pin iin the small signal gain becomes:

$$gain = \frac{gm_{M4}}{gm_{M2}} = \frac{W_{M4} * L_{M2}}{W_{M2} * L_{M4}} \tag{2}$$

In our example the small signal current gain is 20. This gain however is not 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 while 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:

$$V_{gseffm4} = V_{gsm4} - V_{th} = \sqrt{\frac{I_{bias}}{gm}} * (\sqrt{\frac{L_3}{W_3}} + \sqrt{\frac{L_1}{W_1}}) - \sqrt{\frac{I_{sink} * L_2}{gm * W_2}} \tag{3}$$

To make life a bit easier let us substitute the constant part of the equation (The voltage at the gate of M1 and M2).

$$V_{gseffOP0} = \sqrt{\frac{I_{bias}}{gm}} * (\sqrt{\frac{L_3}{W_3}} + \sqrt{\frac{L_1}{W_1}}) \tag{4}$$

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

$$I_{drainm4} = \frac{gm * W_4}{L_4} * (V_{gseffOP0} - \sqrt{\frac{I_{sink} * L_2}{gm * W_2}})^2 \tag{5}$$

or if we multiply it out:

$$I_{drainm4} = \frac{W_4}{L_4} * (I_{bias} * (\frac{L_1}{W_1} + 2\sqrt{\frac{L_1 * L_3}{W_1 * W_3}} + \frac{L_3}{W_3}) - 2\sqrt{\frac{I_{bias} * I_{sink} * L_2}{W_2}} * (\sqrt{\frac{L_1}{W_1}} + \sqrt{\frac{L_3}{W_3}}) + I_{sink} * \frac{L_2}{W_2}) \tag{6}$$

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.

$$gain = \frac{dI_{drainm4}}{dI_{sink}} = \frac{W_4}{L_4} * (\frac{L_2}{W_2} - V_{gseffOP0} * \sqrt{\frac{L_2 * gm}{W_2 * I_{sink}}}) \tag{7}$$

$$gain = \frac{W_4}{L_4} * (\frac{L_2}{W_2} - \sqrt{\frac{I_{bias} * L_2}{I_{sink} * W_2}} * (\sqrt{\frac{L_3}{W_3}} + \sqrt{\frac{L_1}{W_1}})) \tag{8}$$

For better visibility what happens we can rewrite (8):

$$gain = \frac{W_4}{L_4} * \frac{L_2}{W_2} * (1 - (\sqrt{\frac{L_3}{W_3}} + \sqrt{\frac{L_1}{W_1}}) * \sqrt{\frac{I_{bias} * W_2}{I_{sink} * L_2}}) \tag{9}$$

Note that in equation 8 and 9 the technology dependent gm has disappeared. The only remaining parameters left in the equation are the aspect ratios of the transistors and the currents. We can use the same calculation for the NMOS and the PMOS side. The only difference is the scaling. To achieve the same currents the PMOS transistors usually are all adjusted to about double or three times the W/L than the corresponding NMOS transistors. This way the gate overdrive voltages of the NMOS transistors and the PMOS transistors become very similar.

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 using equal aspect ratios W/L for M1, M2 and M3 the gain is roughly proportional

$$gain \sim 1 - 2 * \sqrt{\frac{I_{bias}}{I_{sink}}}$$

Reducing $I_{sink}$to 0 leads to an infinite gain. To get a better feeling of these complex looking equations let us have a look at the example of above:

$$L_1 = L_2 = L_3 = L_4 = 0.5\mu m$$

$$W_1 = W_2 = W_3 = 2.2\mu m$$

$$W_4 = 44\mu m$$

$$I_{bias} = I_{sink} = 3\mu A$$

This leads to the following results:

$$I_{drainm4} = 88 * (3\mu A * \frac{2}{2.2} - 2 * 3\mu A * \sqrt{\frac{0.5}{2.2}} * 2 * \sqrt{\frac{0.5}{2.2}} + 3\mu A * \frac{0.5}{2.2}) = 60\mu A$$

$$gain = 88 * (\frac{0.5}{2.2} - \sqrt{\frac{0.5}{2.2}} * 2 * \sqrt{\frac{0.5}{2.2}}) = -\frac{44}{2.2} = -20$$

(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 and very non linear!

2. Looking at equation (5) we see the current $I_{drainm4}$ disappears when I_sink gets too high (equation (5) crosses the zero. Negative values don't make sense for $I_{drainm4}$)

Well, the calculations done above assumed strong inversion. Case number 2 in reality means the M4 operates in weak inversion. So the calculation done doesn't apply anymore. Instead of really reaching at zero drain current we reach at a weak inversion current.

Reasonable values for I_{sink} are between:

$$0 < I_{sink} < 4 * I_{bias}$$

2.1 Practical Design Methods

The calculation involves many transistors. Deviations from the idealized quadratic behavior (Manual calculation for instance neglected the early effect) have an impact on the design. Typically the manual calculation is used to get a starting point under ideal conditions. As a second step the real circuit has to be fitted taking into account second order effects. This fitting is done with simulation. Normally optimum transistor sizes of the the fitted design shouldn't deviate too much from the manual calculation (some % to some 10%). If the deviations between manual calculation and Spice simulation are higher the design should carefully be rechecked.

To drive the stage the sink current is modulated. It must be verified that the current I_{sink} remains within the usable range. Leaving this range in overdrive conditions can lead to unstable behavior of the output stage and must be prevented by limiting the input swing of the drive current.

Current mirrors with big ratios and high currents are sensitive to voltage drops over metal paths. Careful layout checking or back annotation of the metal resistances is recommended.