EE4308

ANALOG IC DESIGN

Oliver McCarthy

☐ Principal References

RZ: Razavi, Behzad. The Recommended Course Text.

***Design of Analog CMOS integrated Circuits*. McGraw-Hill 2001.**

JM: Johns & Martin.

***Analog Integrated Circuit Design*. Wiley 1997.**

AH: Allen & Holberg.

***CMOS Analog Circuit Design*. 2nd Edition. Oxford Univ Press 2002.**

GM: Gray & Meyer et al.

***Analysis & Design of Analog Integrated Circuits*. 4th Ed. Wiley 2001.**

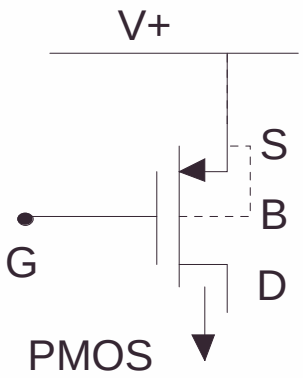
GT: Gregorian & Temes.

***Analog MOS Integrated Circuits for Signal Processing*. Wiley 1986.**

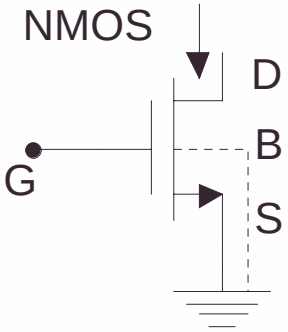
MOSFET DEVICE BASICS

CMOS Structure

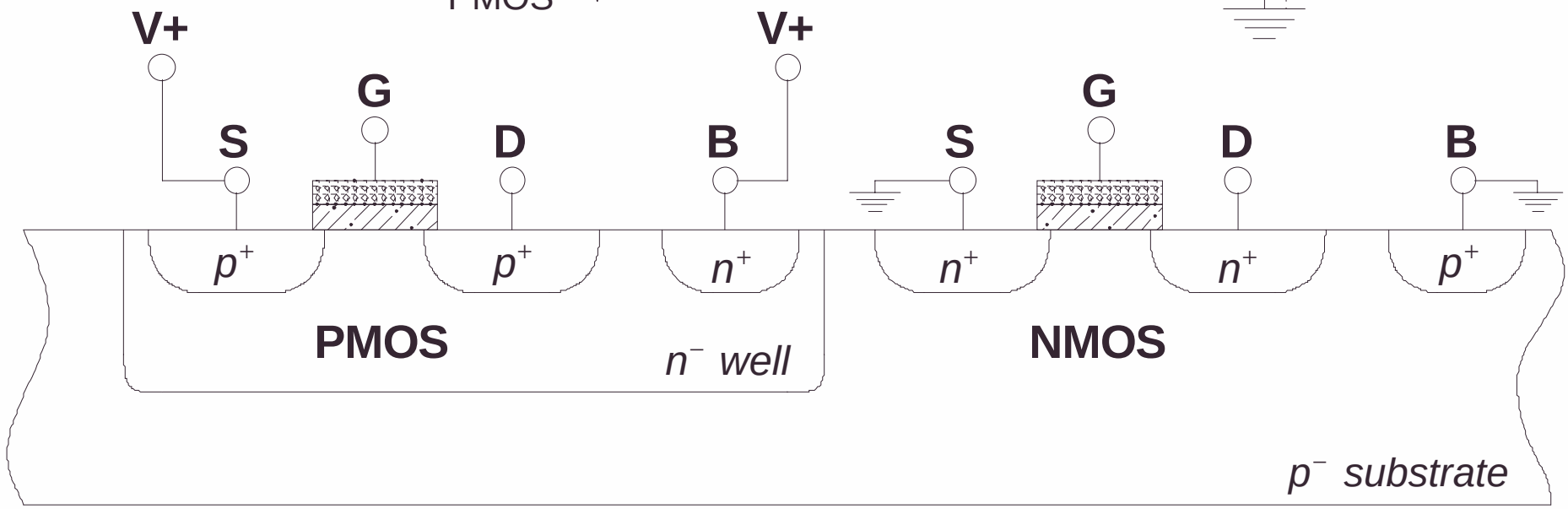
All junctions must be reverse-biased



Tie Bulk to V+ or to S



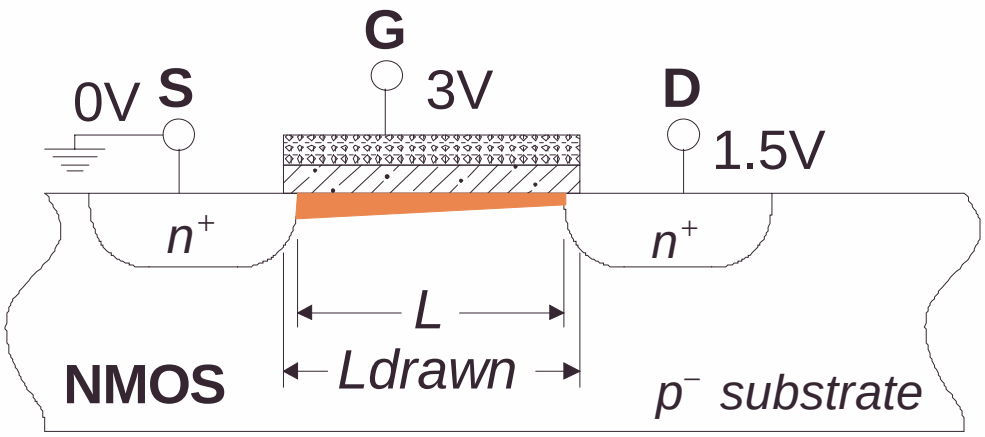
Tie Bulk to GND



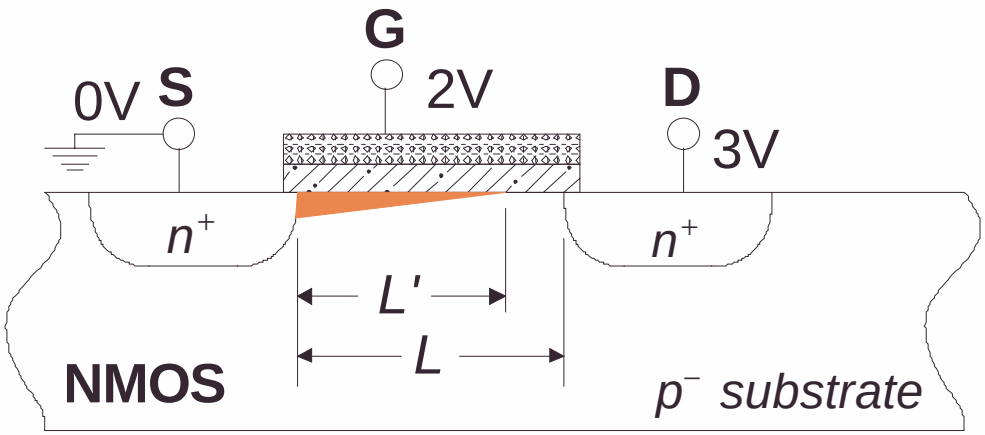
polysilicon

silicon-dioxide, SiO_2

NMOS – operating regions



$_T = 0.8$



Non-saturated (triode) region $V_{DS} < V_{GS} - V_T$

Channel behaves like resistor supported by the *overdrive* voltage (or *effective* voltage V_{eff})

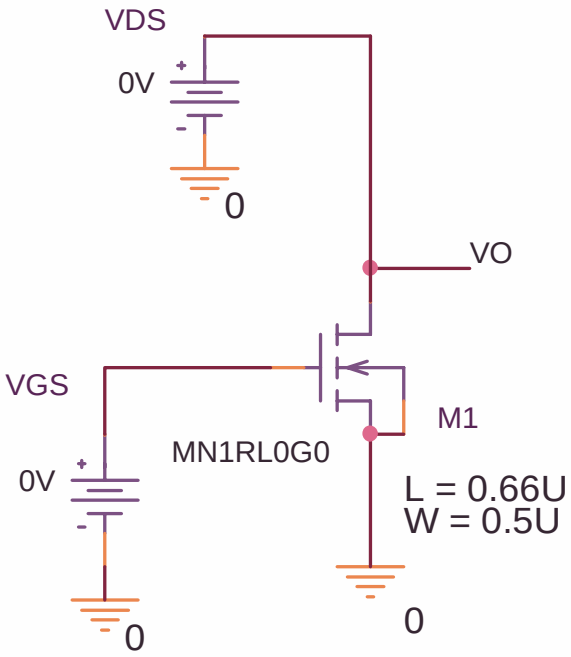
$$_{eff} = V_G - V_T - V_{ch}$$

Current depends on V_{DS} as well as on V_{GS}. Acts like a gate-controlled resistor.

Saturated region $V_{DS} > V_{GS} - V_T$

Channel is pinched-off and V_{DS} now has little effect. Acts like a gate-controlled current sink.

NMOS Current Equations



On PSPICE Diagram:
W = drawn width (U = microns)
L = drawn length (U = microns)

non-sat $V_{DS} < V_{GS} - V_T$

$$I_D = \beta' \frac{W}{L}\left[(V_{GS} - V_T - 0.5 * V_{DS})V_{DS}\right]$$

Ohmic behaviour with a gate-modulated channel resistance

Average value of overdrive (V_{eff})

sat $V_{DS} > V_{GS} - V_T$

Set $V_{DS} = V_{GS} - V_T$ (for top of channel)
Then:

$$I_D = 0.5\,\beta' \frac{W}{L}(V_{GS} - V_T)^2$$

Gate-controlled current sink independent of V_{DS}

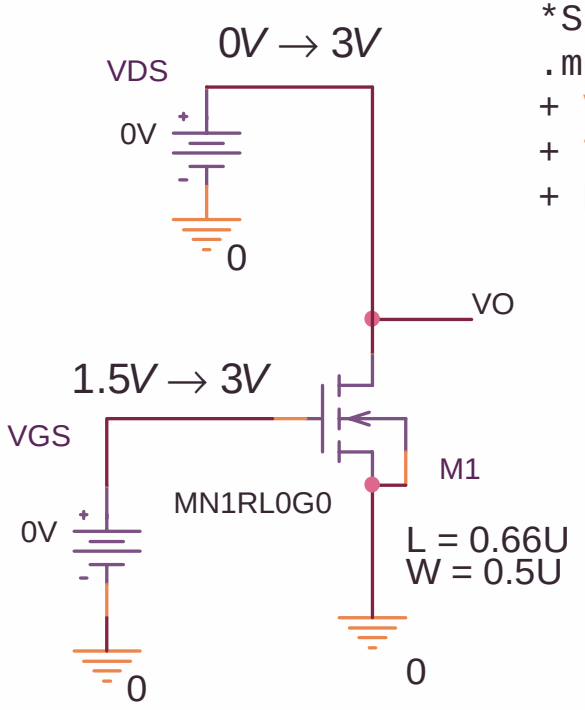
β' is the "gain" in A/V^2 for a square device

$$\beta = \beta' \frac{W}{L}$$

β is the "gain" for a MOSFET of size (***W***,***L***)

The ***L*** of the equations is smaller than PSPICE L by the amount of the two side diffusions:
(i.e. ***L*** = L - 2*LD in SPICE notation).

NMOS SPICE LEVEL 1




```
*SPICE LEVEL1 RAZAVI 0.5U NMOS MODEL WITH LAMBDA = GAMMA = 0
.model MN1RL0G0 NMOS LEVEL=1 LAMBDA=0.0 GAMMA=0.00
+ VTO=0.7 PHI=0.9 NSUB=9E14 LD=0.08E-6 UO=350
+ TOX=9E-9 PB=0.9 CJ=0.56E-3 CJSW=0.35E-11
+ MJ=0.45 MJSW=0.2 CGDO=0.4E-9 JS=1.0E-8
```

non-sat $$I_D = \beta' \frac{W}{L}\left[(V_{GS} - V_T - 0.5 * V_{DS})V_{DS}\right]$$

sat $$I_D = 0.5\,\beta' \frac{W}{L}(V_{GS} - V_T)^2$$

$$\beta' = C_{ox}\,\mu = (k_{ox}\varepsilon_0 / t_{ox})\,\mu \qquad \text{(to be proven later)}$$

$$\frac{W}{L} = \frac{0.5U}{0.66U - 2 \cdot LD} = 1.0$$

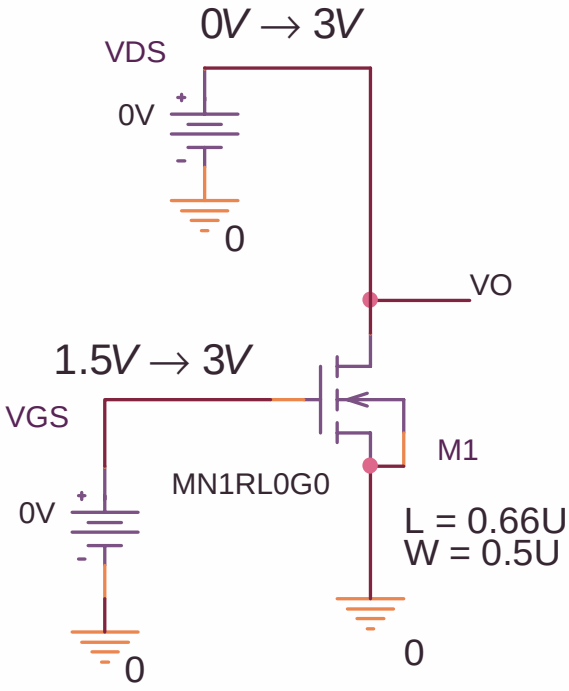
Allowing for side-diffusion *LD*, this is a square MOSFET

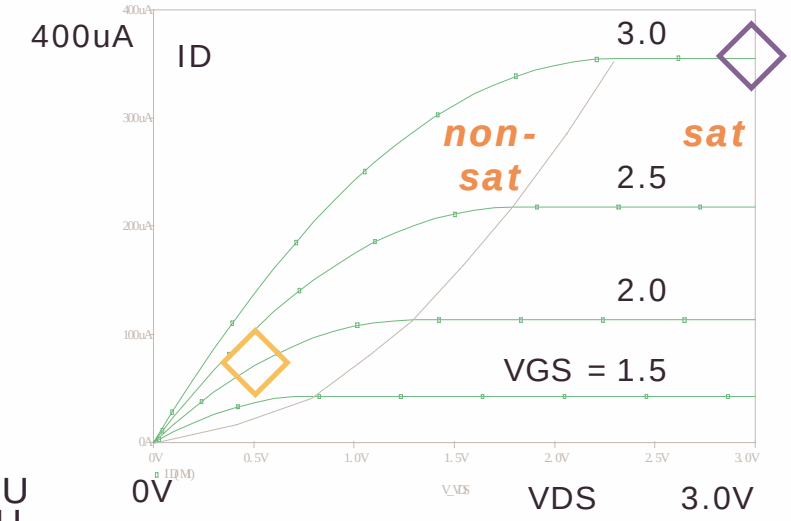
$$\beta' = \frac{3.9 \cdot 8.85 \cdot 10^{-12}}{TOX} \cdot U0 \cdot 10^{-4} = 134\,\mu A/V^2$$

Note: U0 appears in SPICE using cm units (cm^2/V-sec)

We can now apply the current equations ..

NMOS Characteristics





Horizontal traces mark the SAT region

VT=0.7 SAT BOUNDARY	
VGS	VDS
1.5	0.8
2.0	1.3
3.0	2.3

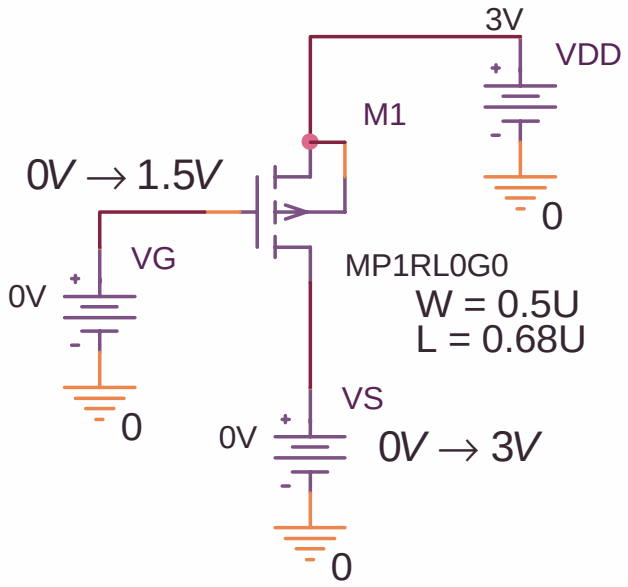
Almost linear through (0uA,0V) marks triode region

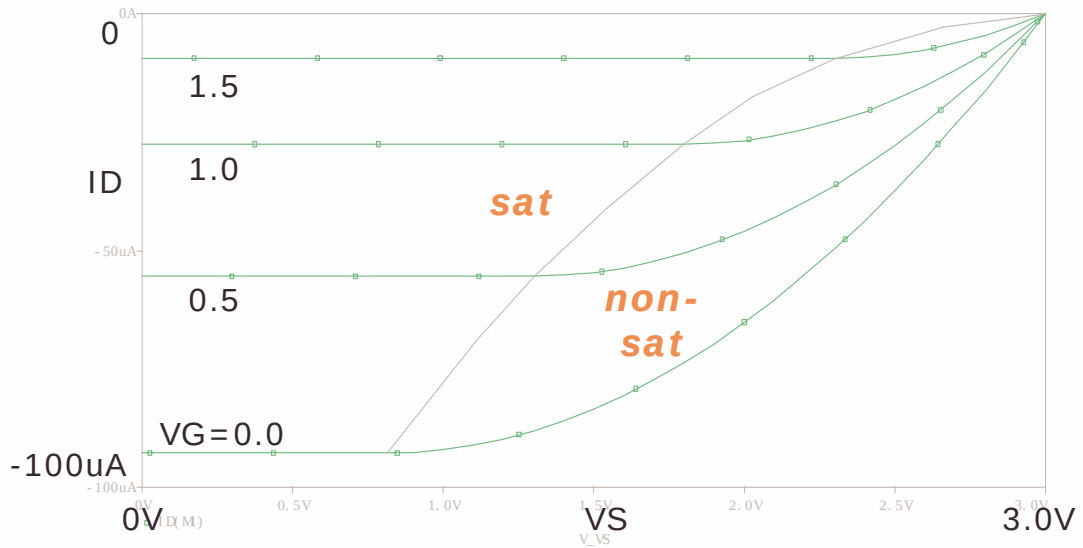
For $V_{DS} = V_{GS} = 3.0V$ $\quad I_D = 0.5 \cdot 134E-6 \cdot 1.0 \cdot (3-0.7)^2 = 354\mu A$

For $V_{DS} = 0.5V, V_{GS} = 2.0V$: $\quad I_D = 134E-6 \cdot 1.0 \cdot (2.0-0.7-0.25) \cdot 0.5 = 70.3\mu A$

But, this model ignores two important effects … described by LAMBDA and GAMMA

PMOS Characteristics





$$\frac{W}{L} = \frac{0.5U}{0.68U - 2 \cdot LD} = 1.0$$

Use same equations but with all polarities reversed !

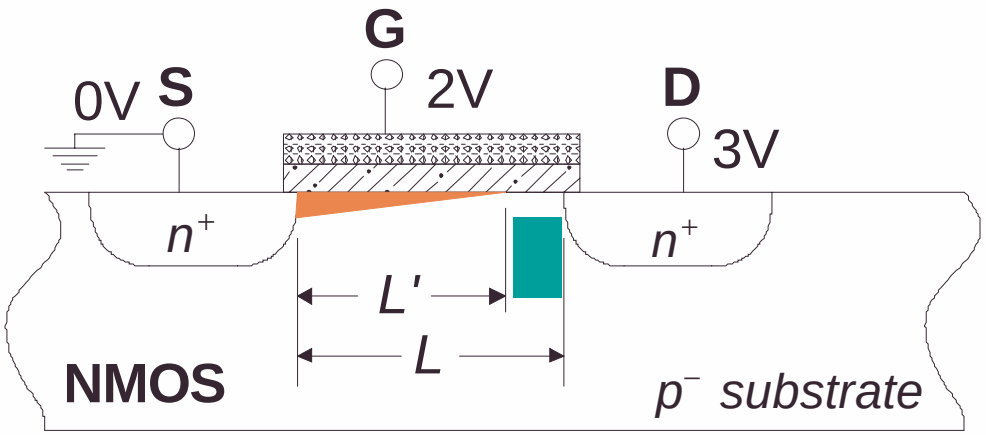
```
*SPICE LEVEL1 RAZAVI 0.5U PMOS MODEL LAMBDA=0 GAMMA=0
.model MP1RL0G0 PMOS LEVEL=1 LAMBDA=0.0 GAMMA=0.0
+ VTO=-0.8 PHI=0.8 NSUB=5E14 LD=0.09E-6 UO=100
+ TOX=9E-9 PB=0.9 CJ=0.94E-3 CJSW=0.32E-11
+ MJ=0.5 MJSW=0.3 CGDO=0.3E-9 JS=0.5E-8
```

Notice the much lower mobility of PMOS devices

VT=-0.8 SAT BOUNDARY	
VG	VS
0	0.8
1	1.8
1.5	2.3

Now try some current calculations for this square PMOS device !

NMOS – the LAMBDA factor



Saturated region $V_{DS} > V_{GS} - V_T$

Channel length is reduced from ***L*** to ***L'***. Its best to use ***L'*** in the SAT-region current equation.

$$I_D = 0.5\,\beta' \frac{W}{L'}(V_{GS} - V_T)^2$$

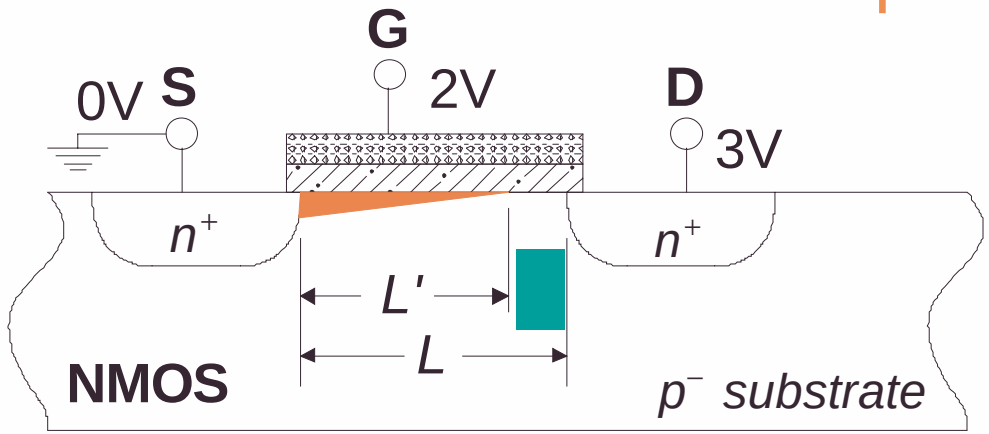
$\Delta L = L - L'$ This ΔL is a depletion region such that ΔL is given by the step-junction approximation ➔, with ΔV as the reverse voltage across ΔL while N_B is the bulk-doping concentration.

$$\Delta L \approx \left[\frac{2k_{si}\varepsilon_0(\psi_0 + \Delta V)}{qN_B}\right]^{1/2}$$

V_{CE} = *channel-end voltage*, and $\Delta V = V_D - V_{CE}$ using ground reference

As V_D is raised, ΔV increases and ***L'*** grows shorter, causing a slight increase in current I_D (whereas previously I_D was constant). A parameter LAMBDA (or λ) is used to account for this effect ..

☐ NMOS – LAMBDA depends on *L*



Saturated region $V_{DS} > V_{GS} - V_T$

$$I_D = 0.5\,\beta' \frac{W}{L'}(V_{GS} - V_T)^2$$

$$\Delta L = L - L'$$

$$\Delta L \approx \left[\frac{2k_{si}\varepsilon_0(\psi_0 + \Delta V)}{qN_B}\right]^{1/2}$$

That is ΔL by the step-junction approximation –
But, to keep things simple, we'll just assume *a linear* proportionality: $\Delta L \propto V_{DS}$

Now, the ***sat*** current I_D increases as:

(where ***k*** is some constant)

$$\frac{L}{L'} = \frac{L}{L - \Delta L} = \frac{1}{1 - \frac{\Delta L}{L}} \approx 1 + \frac{\Delta L}{L} \approx 1 + \frac{k \cdot V_{DS}}{L}$$

Thus, I_D increases linearly with V_{DS}:

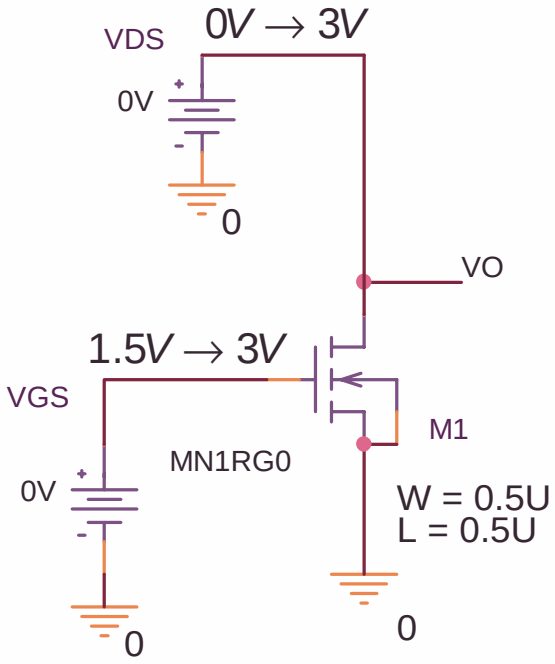
$$I_D \propto (1 + \lambda \cdot V_{DS})$$

where λ (known as LAMBDA in PSPICE) is the constant of proportionality

Notice: λ is inversely proportional to *L*.
If *L* is doubled, then λ is halved.

The LAMBDA value in the SPICE model to follow is for a minimum-length device. If we use a longer MOSFET, we should reduce LAMBDA accordingly.

☐ NMOS LEVEL 1 including LAMBDA (λ)



$$\frac{W}{L} = \frac{0.5U}{0.5U - 2 \cdot LD} = 1.47$$

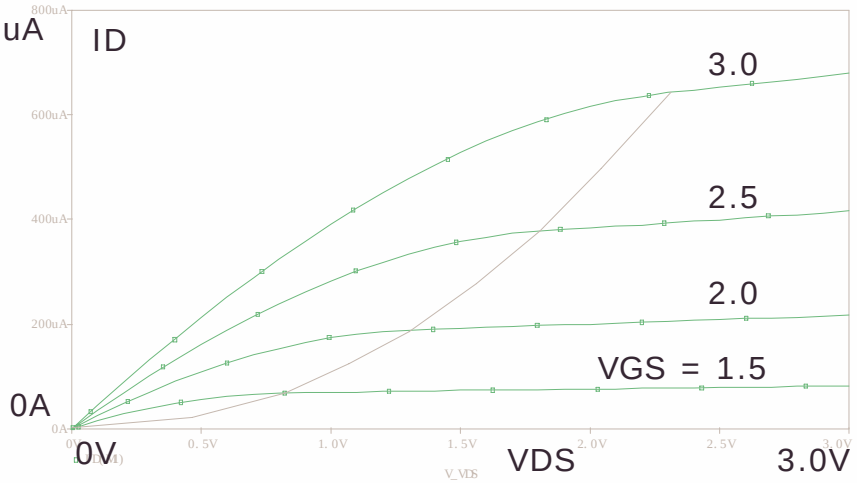
sat $$I_D = 0.5\,\beta' \frac{W}{L}(V_{GS} - V_T)^2 \cdot (1 + \lambda \cdot V_{DS})$$

Now, when V_{DS} is raised, the current increases too.

For continuity of the current plots, we must include λ in the ***non-sat*** equation as well:

non-sat $$I_D = \beta' \frac{W}{L}[(V_{GS} - V_T - 0.5 * V_{DS})V_{DS}] \cdot (1 + \lambda \cdot V_{DS})$$

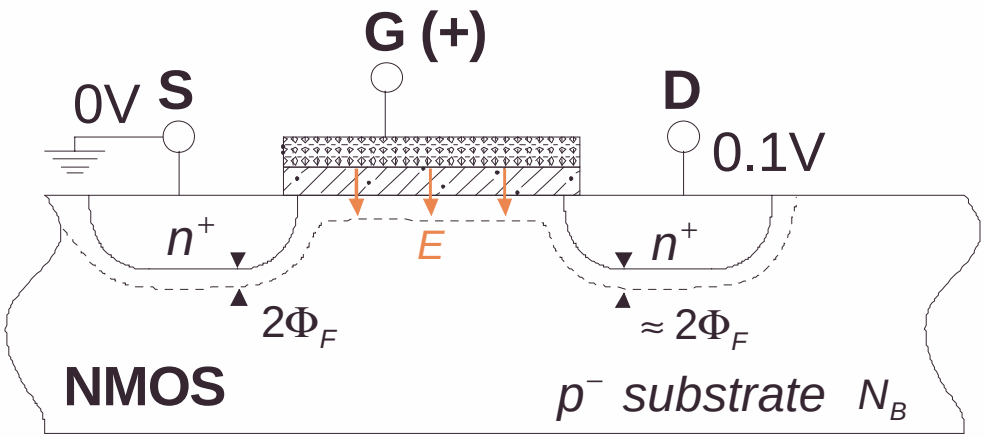
Try some current calculations !

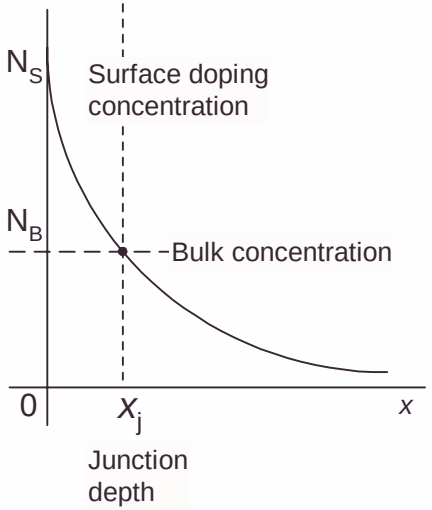



```
*SPICE LEVEL1 RAZAVI 0.5U NMOS MODEL WITH GAMMA=0
.model MN1RG0 NMOS LEVEL=1 LAMBDA=0.1 GAMMA=0.00
+ VTO=0.7 PHI=0.9 NSUB=9E14 LD=0.08E-6 UO=350
+ TOX=9E-9 PB=0.9 CJ=0.56E-3 CJSW=0.35E-11
+ MJ=0.45 MJSW=0.2 CGDO=0.4E-9 JS=1.0E-8
```

THRESHOLD VOLTAGE AND BODY EFFECT

How Thresholding Occurs





$$\Phi_F = \frac{kT}{q} Ln\left(\frac{N_B}{n_i}\right)$$

With $2\Phi_F$ across dep layer, the surface is "as negative" as the Bulk is positive, which conveniently defines V_T.

$V_G < V_T$:

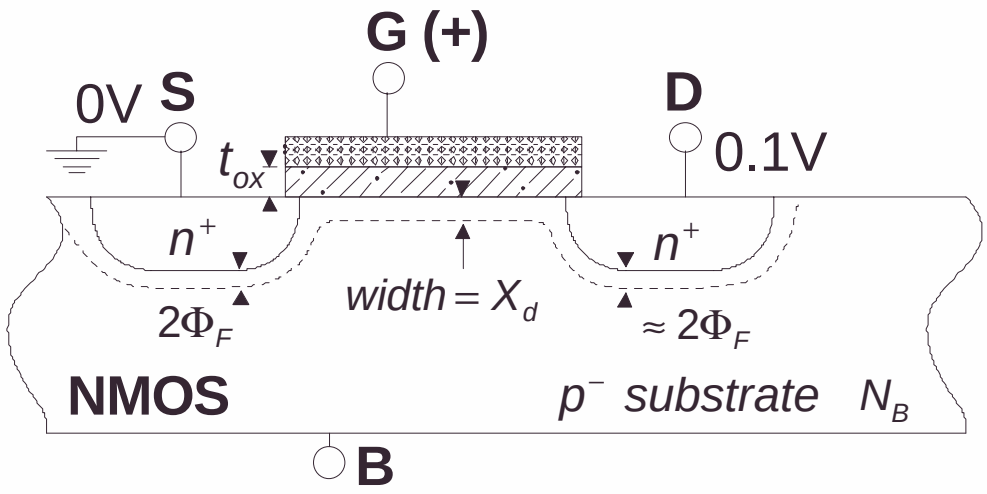
As gate voltage increases, the *E* field pushes holes down creating a widening depletion layer and exposing negative ionic charge to balance the positive Gate charge.

$V_G > V_T$:

Surface reaches equi-potential (with $2\Phi_F$ across dep layer) and electrons now flow in from **S** to form a conducting film or ***channel*** on the silicon surface. As V_G increases further, dep layer does not grow wider, but balancing charge comes from a strengthening of the channel on the surface.

In *SPICE*, $2\Phi_F$ is " PHI ", typically 0.8V.

Components of Threshold Voltage



$$X_d = \sqrt{\frac{2k_{si}\varepsilon_0(Bias)}{qN_B}}$$ Depletion layer width

$$Q_B = qN_B X_d$$ Depletion layer ionic charge in coulombs/m^2.

$$C_{ox} = \frac{k_{ox}\varepsilon_0}{t_{ox}}$$ Gate oxide capacitance in Farads/m^2.

$$_T = 2\Phi_F + \frac{Q_B}{C_{ox}} + \Phi_{MS}$$

across dep layer ***across gate oxide***

Φ_{MS} is a component (some tenths of a volt) due to work-function difference between the gate material and the silicon. It is a fixed value, and need not concern us, because we can independently adjust V_T by ion implantation.

The Q_B term is voltage dependent, a function of the (*Bias*) that determines X_d. The "nominal" threshold is V_{TO}. It occurs when $V_{DB}=V_{SB}=0$ such that *Bias* = $2\Phi_F$ (across X_d) and . . .

then:

$$_{T0} = 2\Phi_F + \frac{Q_{B0}}{C_{ox}} + \Phi_{MS}$$ where $$Q_{B0} = qN_B\sqrt{\frac{2k_{si}\varepsilon_0(2\Phi_F)}{qN_B}} = \sqrt{2k_{si}\varepsilon_0 qN_B(2\Phi_F)}$$

☐ Threshold Voltage Calculations


```
*SPICE LEVEL1 RAZAVI 0.5U NMOS MODEL WITH GAMMA=0
.model MN1RG0 NMOS LEVEL=1 LAMBDA=0.1 GAMMA=0.00
+ VTO=0.7 PHI=0.9 NSUB=9E14 LD=0.08E-6 UO=350
+ TOX=9E-9 PB=0.9 CJ=0.56E-3 CJSW=0.35E-11
+ MJ=0.45 MJSW=0.2 CGDO=0.4E-9 JS=1.0E-8
```

$$_T = 2\Phi_F + \frac{Q_B}{C_{ox}} + \Phi_{MS}$$

+ve for NMOS -ve for PMOS

may be +ve or -ve

Sample calculation **(NMOS) :**

$$Q_{B0} = \sqrt{2k_{si}\varepsilon_0 qN_B(2\Phi_F)} = \sqrt{2\cdot 11.7\cdot 8.85E-12\cdot 1.6E-19\cdot 9E20\cdot 0.9} = 1.64E-4 \quad \text{Coulombs/m}^2.$$

Notice: NB = NSUB = 9E20 in atoms per m^3.

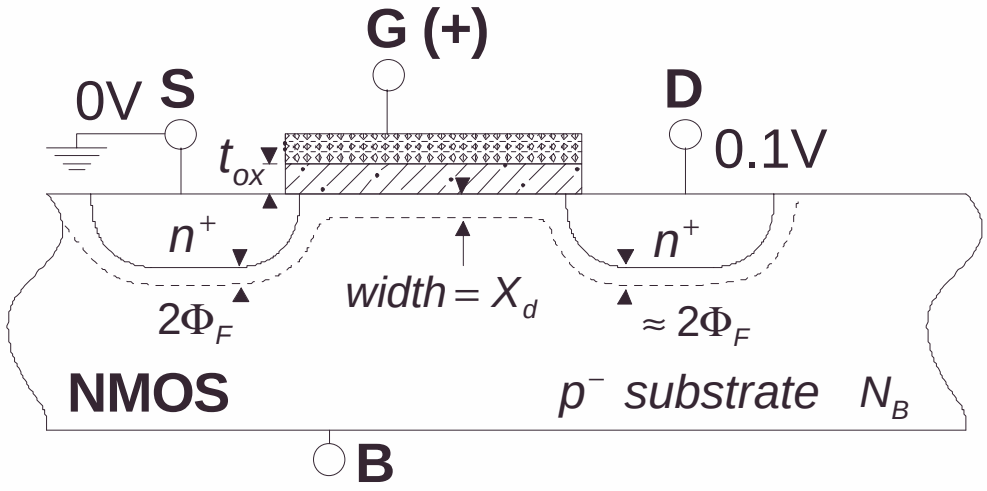
$$C_{ox} = \frac{k_{ox}\varepsilon_0}{t_{ox}} = \frac{3.9\cdot 8.85E-12}{9E-9} = 3.835E-3 \quad \text{Farads/m}^2. \qquad C_{ox} = 3.835 \quad \text{fF/}\mu\text{m}^2.$$

$$2\Phi_F + \frac{Q_{B0}}{C_{ox}} = 0.9 + 0.042 = 0.942$$

For PMOS calculations, both of these components are negative.

When Φ_{MS} is included, NMOS thresholds tend to be too low (say +0.1) and PMOS thresholds tend to be too high (say, -1.8). An ***ion implantation*** step is used during processing to adjust thresholds to around +0.8 (NMOS) and –0.8 (PMOS).

Threshold Adjustment by Ion Implantation



$$Q_B = qN_B X_d$$

N_B is the bulk doping in atoms/cm^3.

$$_T = 2\Phi_F + \frac{Q_B}{C_{ox}} + \Phi_{MS}$$

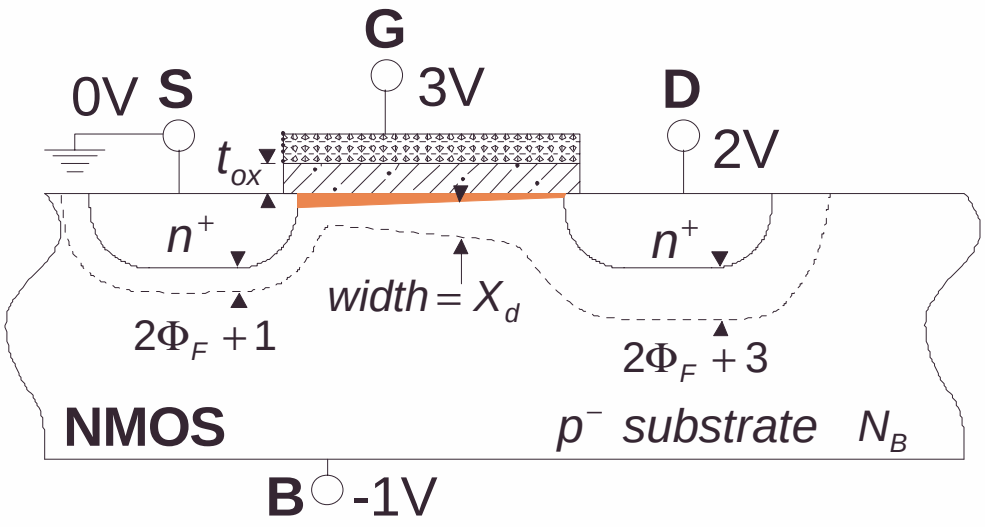
Effective N_B can be increased under the NMOS gate region by using *p*- ion implants. e.g. Boron implants.

Boron implants will raise Q_B and hence the threshold voltage V_T. (whereas *n*-type implants will lower the NMOS threshold). In general, the threshold is shifted by Q/C_{OX} where Q is the implanted charge density (coulombs per m^2).

Threshold adjustment by ion implantation is a routine process step. Implants are also used in ***field*** regions to raise the threshold sufficiently so as to *avoid* channel formation.

Diffusion yields doping profiles with maximum concentration at the surface, but ion implant levels can peak *below* the surface. Ion implantation offers more accurate control of dopants than is possible by diffusion.

☐ Body Effect on Threshold Voltage



For arbitrary bias conditions:

$$X_d = \sqrt{\frac{2k_{si}\varepsilon_0(V_{CB} + 2\Phi_F)}{qN_B}}$$ Depletion layer width

where VCB = Channel-to-Bulk potential
Then:

$$Q_B = qN_B X_d = \sqrt{2k_{si}\varepsilon_0 qN_B(V_{CB} + 2\Phi_F)}$$

whereas: $Q_{B0} = \sqrt{2k_{si}\varepsilon_0 qN_B(2\Phi_F)}$ for zero-bias

The threshold voltage increases to: $_T = V_{T0} + \frac{Q_B - Q_{B0}}{C_{ox}}$ where $C_{ox} = \frac{k_{ox}\varepsilon_0}{t_{ox}}$

Introducing: $\gamma = \frac{t_{ox}}{k_{ox}}\sqrt{\frac{2k_{si}qN_B}{\varepsilon_0}}$ known as " GAMMA " in SPICE (typically 0.3 to 0.8)

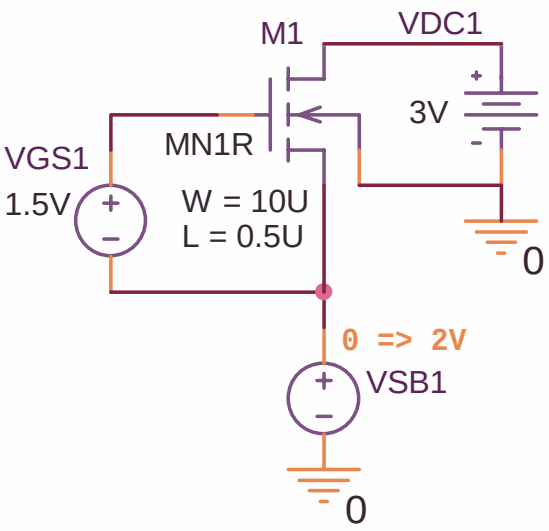
we find ..

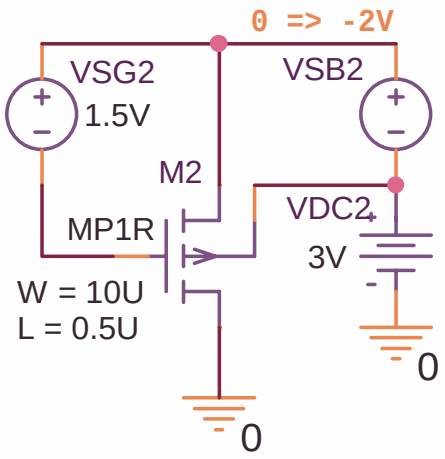
$$_T = V_{TO} + \gamma \cdot \left(\sqrt{V_{CB} + 2\Phi_F} - \sqrt{2\Phi_F}\right)$$

Threshold voltage increases with channel-to-bulk potential. But, for simplicity, we usually use V_{SB} in place of V_{CB}.

The threshold voltage increase is known as "body effect".

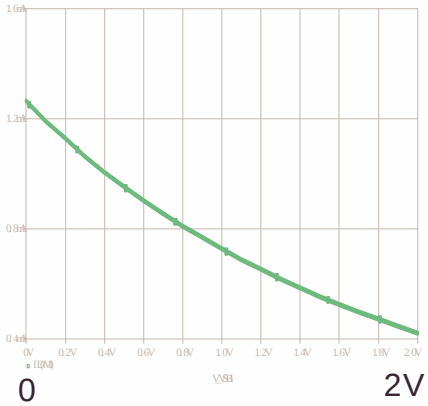
Body Effect Illustrations



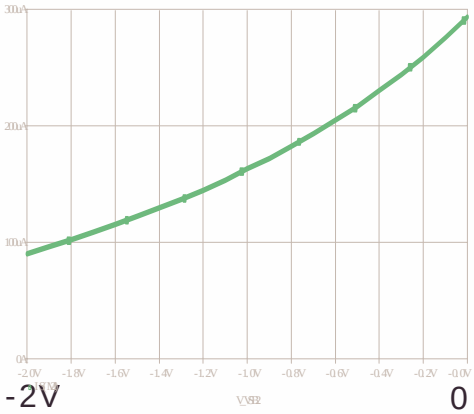



```
*SPICE LEVEL1 RAZAVI 0.5U PMOS MODEL
.model MP1R PMOS LEVEL=1 VTO=-0.8 GAMMA=0.4 PHI=0.8
+ NSUB=5E14 LD=0.09E-6 UO=100 LAMBDA=0.0 TOX=9E-9

*SPICE LEVEL1 RAZAVI 0.5U NMOS MODEL
.model MN1R NMOS LEVEL=1 VTO=0.7 GAMMA=0.45 PHI=0.9
+ NSUB=9E14 LD=0.08E-6 UO=350 LAMBDA=0.0 TOX=9E-9
```



Current falls from 1.26mA (0V) to 0.42mA (2V)



Current falls from 293µA (0V) to 90µA (-2V)

These tests use constant VGS values while VSB is varied over a 2V range. Notice the large current changes: they are due to a $|V_T|$ that increases with |VSB|.

We've set LAMBDA=0 so as to see the body effect on its own (without having LAMBDA effect as well).

Check the values reported here against the body-effect on threshold equation. Remember to use (L-2LD).

☐ MOSFET Circuit Plots : Examples

The examples noted here use the SPICE LEVEL 1 models from Razavi page 37.

```
*SPICE LEVEL1 RAZAVI 0.5U PMOS MODEL
.model MP1R PMOS LEVEL=1 VTO=-0.8 GAMMA=0.4 PHI=0.8
+ NSUB=5E14 LD=0.09E-6 UO=100 LAMBDA=0.0 TOX=9E-9

*SPICE LEVEL1 RAZAVI 0.5U NMOS MODEL
.model MN1R NMOS LEVEL=1 VTO=0.7 GAMMA=0.45 PHI=0.9
+ NSUB=9E14 LD=0.08E-6 UO=350 LAMBDA=0.0 TOX=9E-9
```

The examples Razavi_ch2q5 and Razavi_ch2q7 are based on exercises Q5 and Q7 from Razavi Chapter 2. They are provided as Mathcad files with pdf prints and with PSPICE simulations in folders of the same name.

IC AND MOS CAPACITANCES

☐ On-chip Capacitance Calculations

Between parallel tracks:

$$C = \varepsilon_r \varepsilon_0 \frac{(W + 0.8D)(L + 0.8D)}{D}$$

Visible length L, track separation D, depth W into screen.

Track over plane:

$$C = \varepsilon_r \varepsilon_0 \frac{(W + 1.6D)(L + 1.6D)}{D}$$

Track between planes:

$$C = \varepsilon_r \varepsilon_0 \frac{2(W + 0.9D)(L + 0.9D)}{D}$$

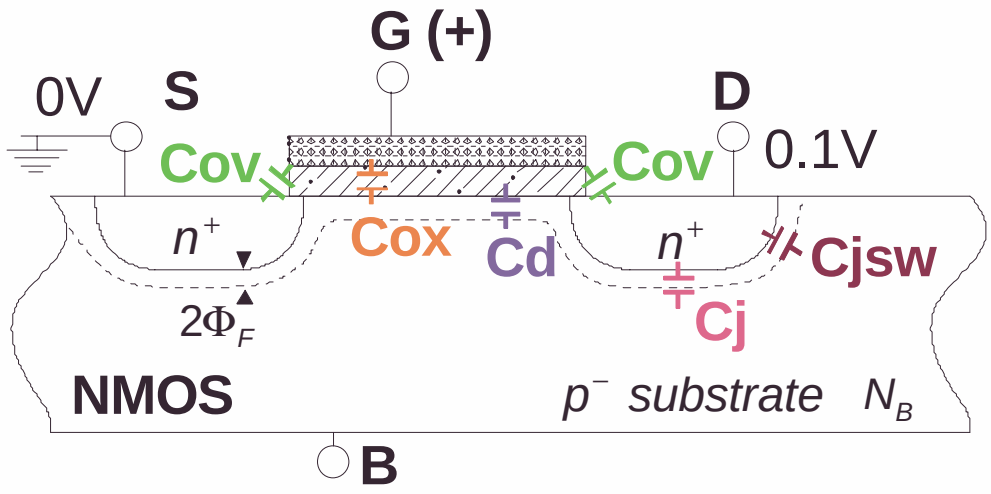
Diffused layers:

Assuming $N_B = 1E15\ /\mathrm{cm}^3$, and $(\psi_0 - V) = 1V$ (V = applied *forward* bias)

$$C_j = \sqrt{0.5 k_{si} \varepsilon_0 q N_B / (\psi_0 - V)} = \sqrt{0.5 \cdot 11.7 \cdot 8.85E-12 \cdot 1.6E-19 \cdot 1E21} = 0.091E-3 \quad (\text{in F/m}^2)$$

Rule of thumb: $C_j = 0.1\,\mathrm{fF}/\mu\mathrm{m}^2$ .. increasing as root of $N_B/10^{15}$.

☐ CMOS Feature Capacitances



The various capacitances of a MOS device are featured here. But we must convert these to *terminal* capacitances for circuit calculations.

When we do this, the roles of C_{OX} and of C_d will vary, depending on the bias conditions ..

$C_{ox} = \dfrac{k_{ox}\varepsilon_0}{t_{ox}}$ Gate oxide capacitance (F/m²).

$C_{ov} = LD \cdot C_{ox}$ overlap capacitances (F/m of W)

$C_d = \sqrt{0.5 k_{si}\varepsilon_0 q N_B / (2\Phi_F)}$ depl cap (F/m²) for zero bias

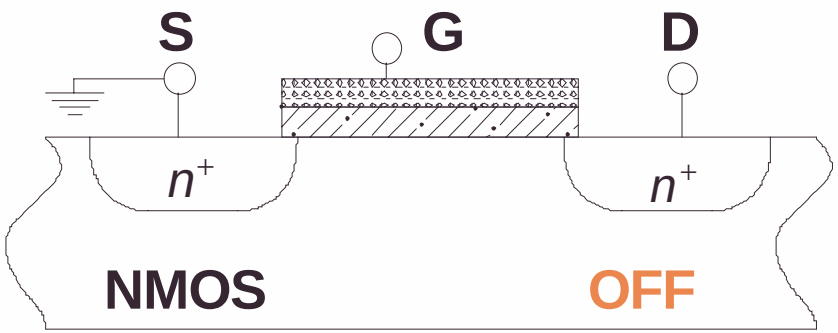
CJ = Zero-bias F/m² (underface)

$CJSW$ = Zero-bias F/m (sidewall)

with voltage dependence ..

$$C_U = \frac{CJ}{\left(1+\dfrac{V_R}{PB}\right)^{MJ}} \qquad C_{sw} = \frac{CJSW}{\left(1+\dfrac{V_R}{PB}\right)^{MJSW}}$$

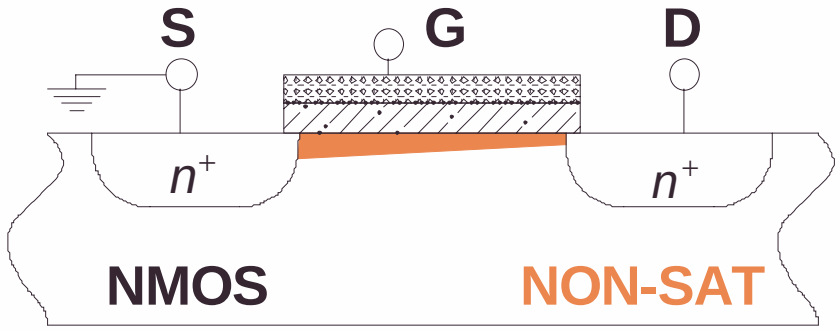
CMOS Terminal Capacitances



$$C_{GS} = W \cdot C_{OV} \ *$$

$$C_{GD} = W \cdot C_{OV} \ *$$

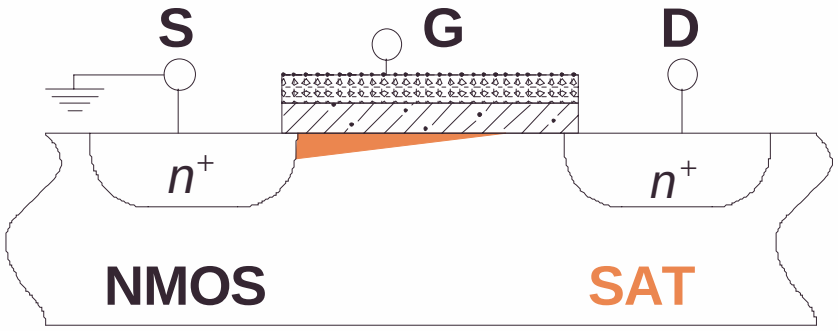
$$C_{GB} = WL \cdot \frac{C_{OX} C_d}{C_{OX} + C_d}$$



$$C_{GS} = W \cdot C_{OV} + \tfrac{1}{2} C_{GATE}$$

$$C_{GD} = W \cdot C_{OV} + \tfrac{1}{2} C_{GATE}$$

$$C_{GB} \approx 0$$

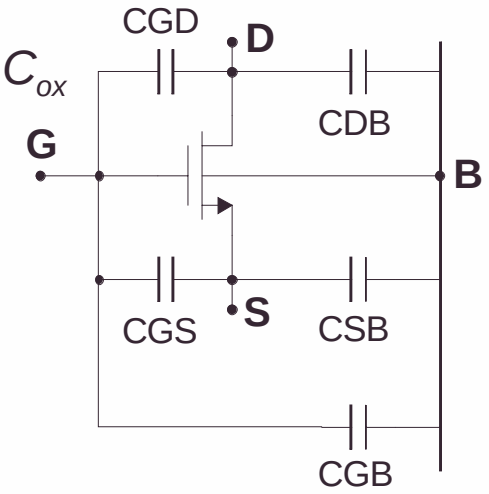


$$C_{GS} = W \cdot C_{OV} + (2/3) C_{GATE}$$

$$C_{GD} = W \cdot C_{OV} \ *$$

$$C_{GB} \approx 0$$

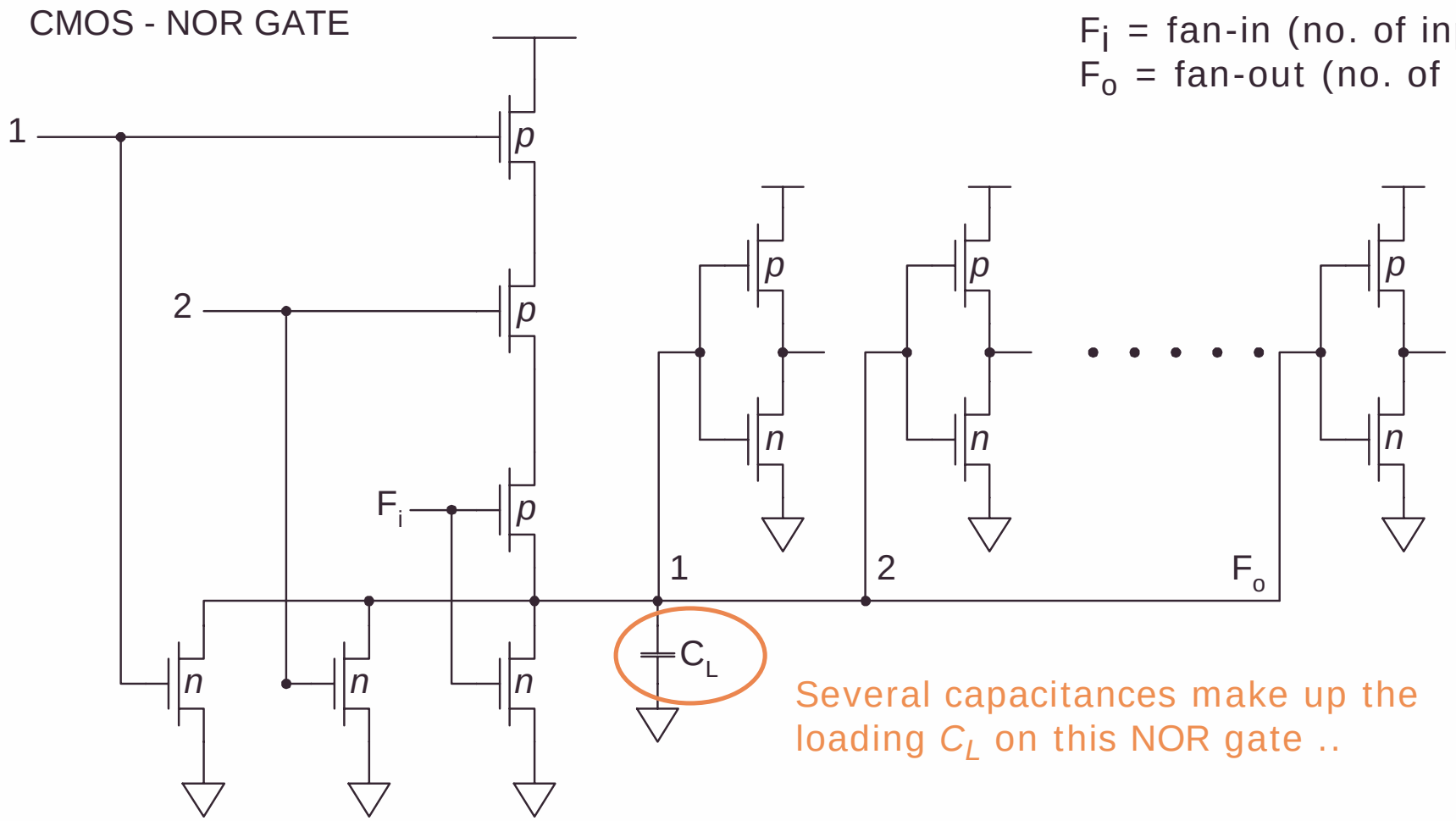
$$C_{GATE} = W \cdot L \cdot C_{ox}$$



C_{DB} and C_{SB} are mostly the voltage-dependent diffusion capacitances, but may be augmented by channel protrusions.

* .. *or higher due to fringing fields*

CMOS Gate Loading Calculation



$$C_L = F_i C_{DSn} + (F_i + 1)C_{GDn} + C_{GDp} + C_{DSp} + F_o(C_{GSn} + 2C_{GDn} + C_{GSp} + 2C_{GDp}) + C_{stray}$$

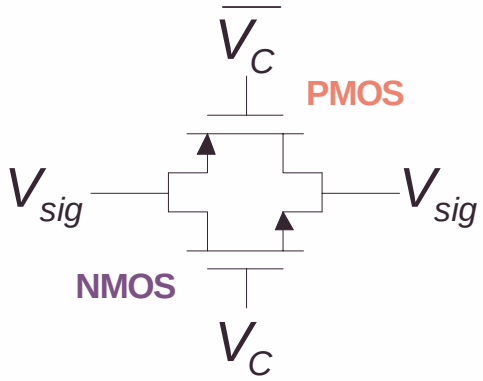
MOS LARGE SIGNAL CIRCUIT BASICS

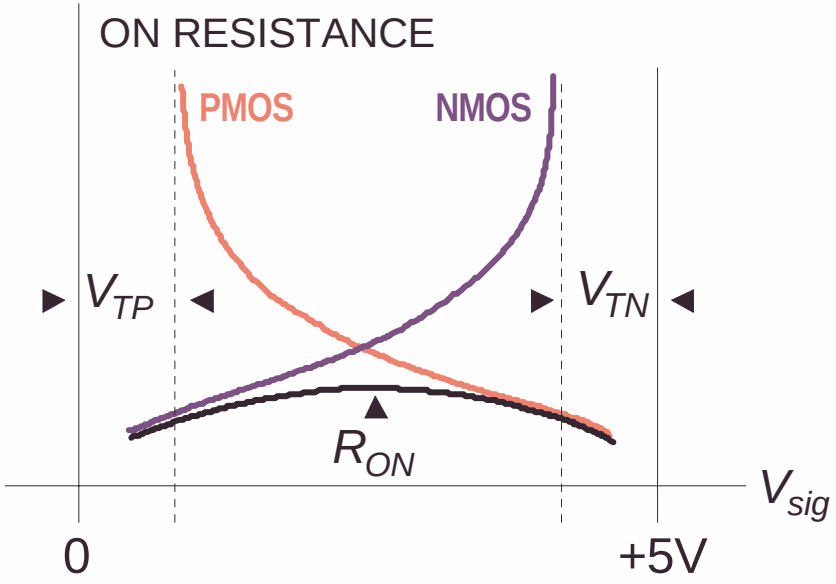
☐ MOSFET "ON" Resistance, R_{ON}

non-sat: $I_D = \beta[(V_{GS} - V_T - 0.5 * V_{DS})V_{DS}]$ where .. $\beta = \beta' \frac{W}{L}$

To function as an ON switch, V_{DS} is small and then .. $I_D \approx \beta[(V_{GS} - V_T)V_{DS}]$ ➔ $\frac{V_{DS}}{I_D} \approx \frac{1}{\beta(V_{GS} - V_T)} \equiv R_{ON}$

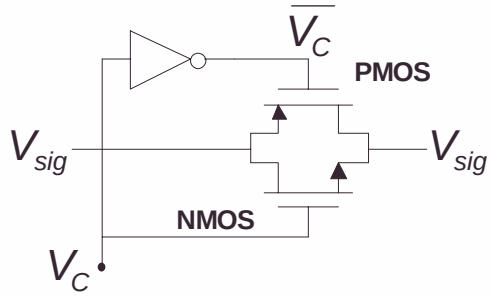
But, R_{ON} varies with signal voltage. To help avoid this, we often connect an NMOS and a PMOS in parallel ➔





In this way, we achieve low R_{ON} ➔ over the full range of V_{sig}.

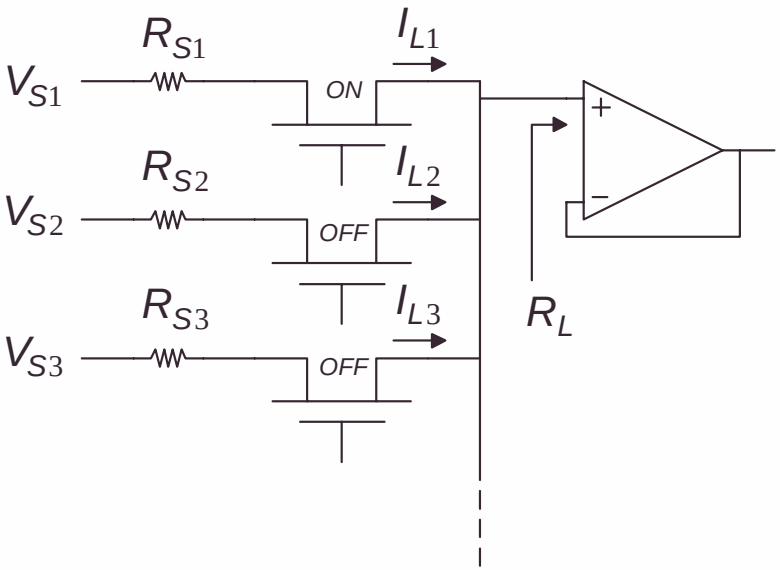
A logic inverter must be included ➔



☐ Multiplexers using MOS switches

MOSFETS make good switches with no inherent offset (unlike bipolars).

Switches with large (W/L) will give suitably low ON resistance ..



← *an analogue signal multiplexer*

$$\frac{V_L}{V_{S1}} = \frac{R_L}{R_{S1} + R_{ON} + R_L}$$ ideally = 1.0

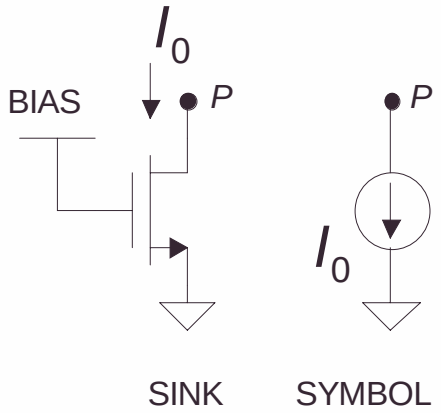
a high load impedance R_L reduces the need for low R_{ON}.

When several switches are in parallel, leakage current may cause an offset error E_L , especially if R_{S1} is high, because ..

$$E_L = (I_{L2} + I_{L3} + ..) \cdot (R_L \,||\, (R_{S1} + R_{ON}))$$

Current Sinks and Sources

Simple Current Sink



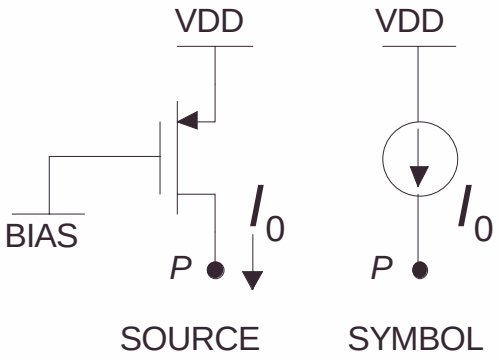
Current entering point *P* is (ideally) independent of the voltage on *P*.

In practice, there is a limit on V_P *below* which current begins to fall.

Even within the valid range, current *increases* slightly as V_P increases.

Sinks require NMOS transistors

Simple Current Source



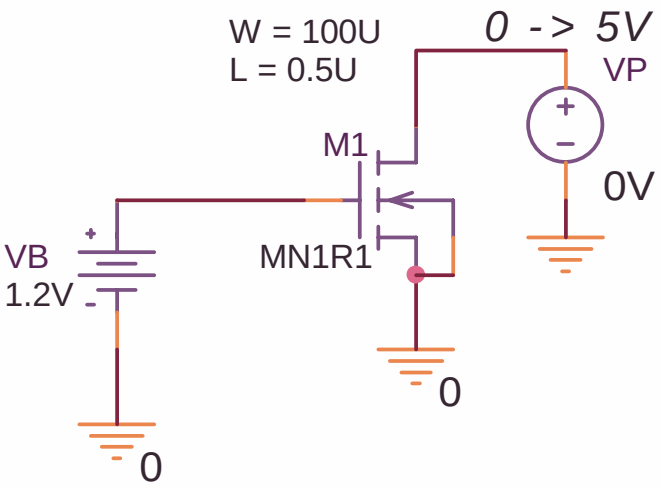
Current exiting point *P* is (ideally) independent of the voltage on *P*.

In practice, there is a limit on V_P *above* which current begins to fall.

Even within the valid range, current *decreases* slightly as V_P increases.

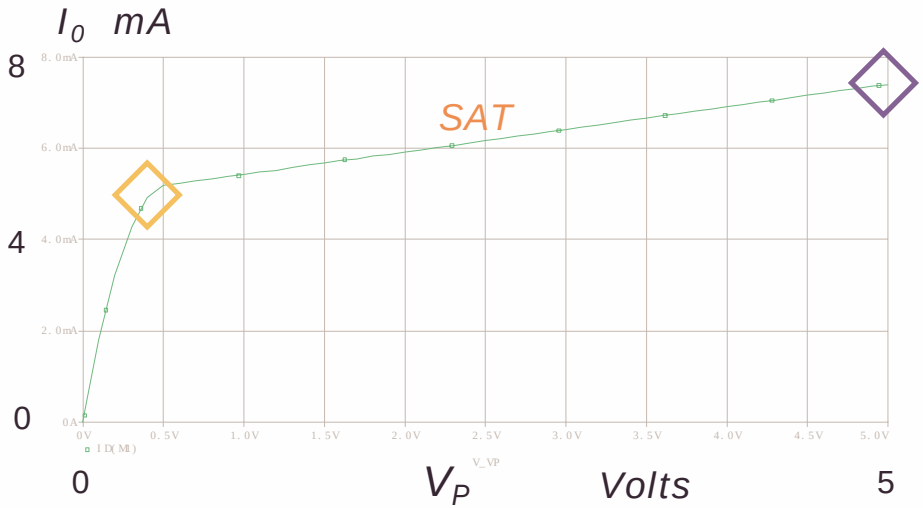
Sources require PMOS transistors

Current Sink Example




```
*SPICE LEVEL1 RAZAVI 0.5U NMOS MODEL MODS 1 GAMMA
.model MN1R1 NMOS LEVEL=1 VTO=0.7 GAMMA=0.00 PHI=0.9
+ NSUB=9E14 LD=0.08E-6 UO=350 LAMBDA=0.1
+ TOX=9E-9 PB=0.9 CJ=0.56E-3 CJSW=0.35E-11
+ MJ=0.45 MJSW=0.2 CGDO=0.4E-9 JS=1.0E-8
```

Measured SAT slope = 2.02kΩ



Compliance: Vmin = 1.2 – V_T = 0.5

Calculations:

$$\beta' = C_{OX} \cdot \mu = \frac{k_{ox}\varepsilon_0}{t_{ox}} \cdot \mu = \frac{3.9 \cdot 8.85E-12}{9E-9} \cdot 350E-4 = 134\mu A/V^2$$

$$I_0 = \frac{\beta'}{2}\frac{W}{L-2\cdot LD}(V_B - V_{T0})^2 = \frac{134E-6}{2}\frac{100}{0.5-0.16}(1.2-0.7)^2 = 4.9mA$$

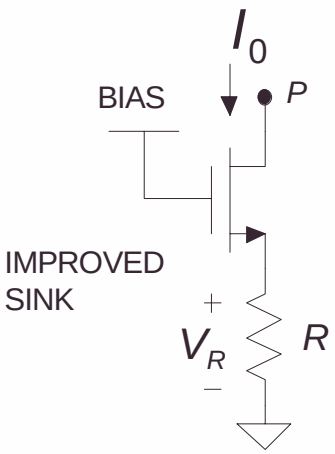
$$I_{max} = I_0 \cdot (1+\lambda \cdot V_{DS}) = 4.9mA \cdot (1+0.1\cdot 5) = 7.35mA$$

Ideally, the *SAT* slope should be *horizontal*.

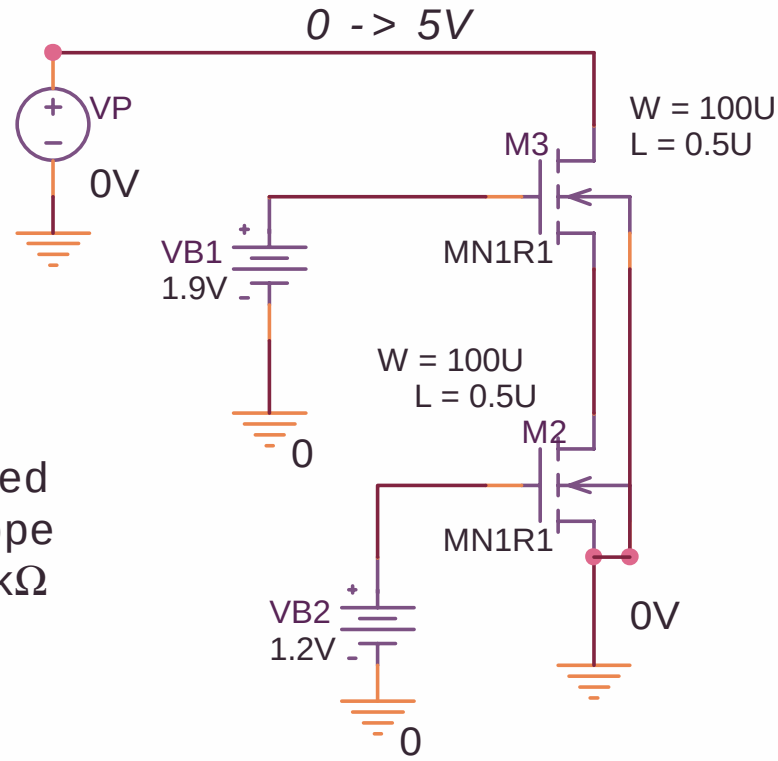
λ is the culprit. But there are ways to make things better ..

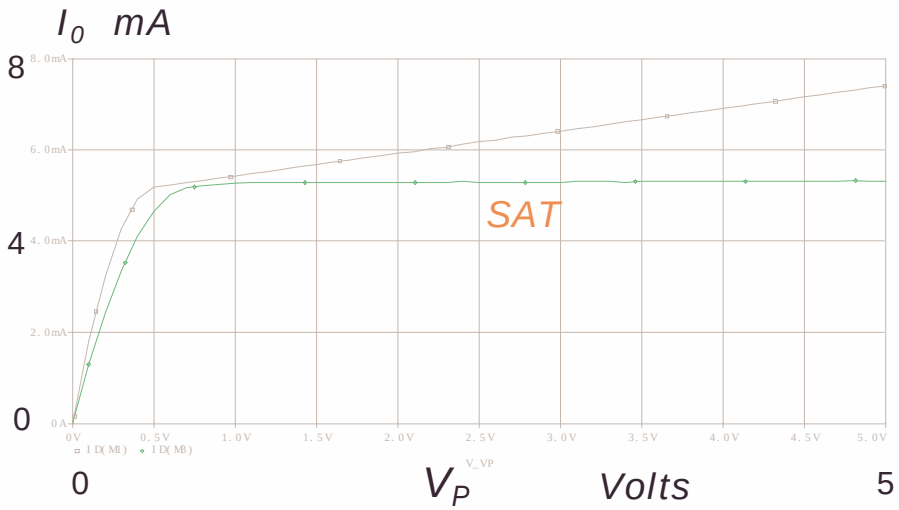
Improved Current Sink

A resistor *R* in the Source lead ➔ gives a much improved sink. That's because the drop across *R* reduces V_{GS} to restrict current growth.



In practice, *R* is replaced by a SAT MOSFET (M2). Bias must ensure that M2 and M3 are both SAT.





Measured SAT slope = 69.5kΩ

But, the range of compliance is reduced : (Vmin = 1.9 – V_T = 1.2 Volts).

A Simple Design Procedure

SAT Eqn: $I_D = 0.5\,\beta\,(V_{GS} - V_T)^2 \cdot (1 + \lambda \cdot V_{DS})$

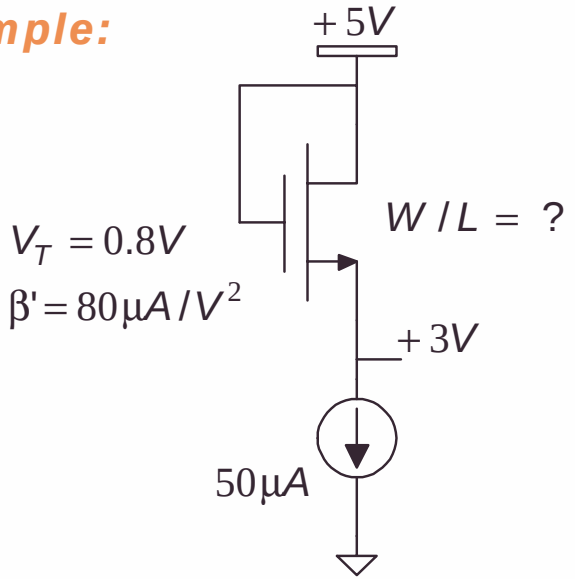
Re-arranging: $_{GS} = V_T + \sqrt{\dfrac{2I_D}{\beta(1+\lambda V_{DS})}} = V_T + V_e$

V_e is the "**effective voltage**" or the "**overdrive**"

$_e = \sqrt{\dfrac{2I_D}{\beta(1+\lambda V_{DS})}} \approx \sqrt{\dfrac{2I_D}{\beta}}$ (when V_{DS} is small)

$_e \approx \sqrt{\dfrac{2I_D}{\beta'(W/L)}}$ ➔ $\dfrac{W}{L} \approx \dfrac{2I_D}{\beta' V_e^{\,2}}$

Example:



Determine (W/L) which sets bias at +3V.

$_{GS} = 5 - 3 = 2V$

$_e = V_{GS} - V_T = 2 - 0.8 = 1.2V$

$\dfrac{W}{L} \approx \dfrac{2I_D}{\beta' V_e^{\,2}} = \dfrac{2 \cdot 50E-6}{80E-6 \cdot 1.2^2} = 0.87$

The idea of ***overdrive voltage*** V_e is a widely used concept when setting up a circuit.

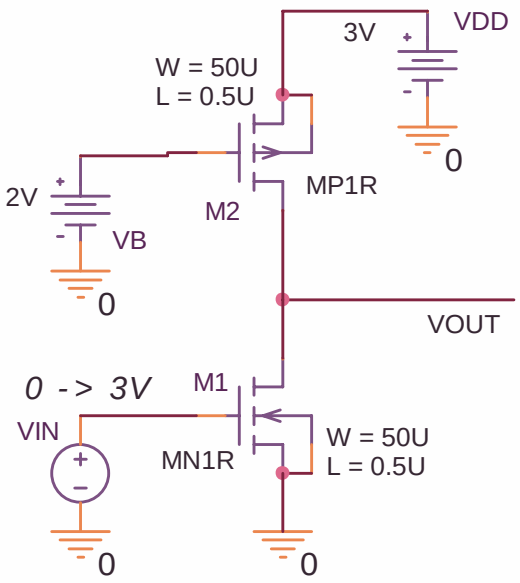
We should allow for *LD* when we make sizing decisions.

Common Source Amp

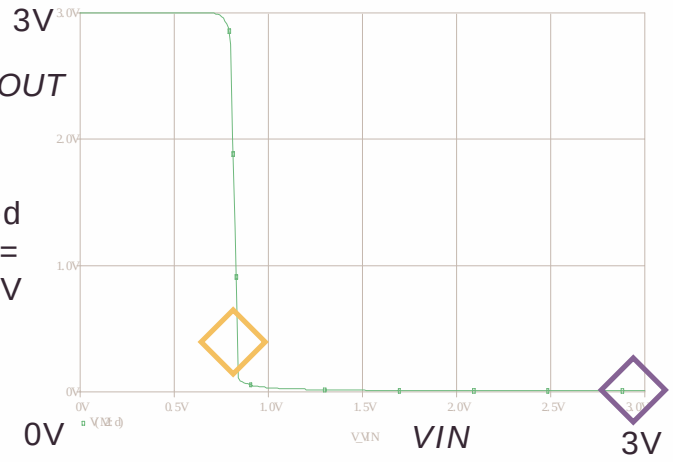
PMOS Load

```
*SPICE LEVEL1 RAZAVI 0.5U PMOS MODEL
.model MP1R PMOS LEVEL=1 VTO=-0.8 GAMMA=0.4 PHI=0.8
+ NSUB=5E14 LD=0.09E-6 UO=100 LAMBDA=0.2 TOX=9E-9

*SPICE LEVEL1 RAZAVI 0.5U NMOS MODEL
.model MN1R NMOS LEVEL=1 VTO=0.7 GAMMA=0.45 PHI=0.9
+ NSUB=9E14 LD=0.08E-6 UO=350 LAMBDA=0.1 TOX=9E-9
```



An NMOS driver with a current-source PMOS load.



Measured slope = 68 V/V

When biased in transition region, the ac gain (the *slope* of V_O/V_I) is about –68.

Gain is limited by the λ of M1 and of M2.

Calculations:

$$I_{max} = 0.5\beta_P' \frac{W}{L-2LD}(V_{SG}-V_T)^2(1+\lambda V_{SD}) = 0.5\cdot 38.3E-6\frac{50}{0.32}(1-0.8)^2(1+0.2\cdot 3) = 191\mu A$$

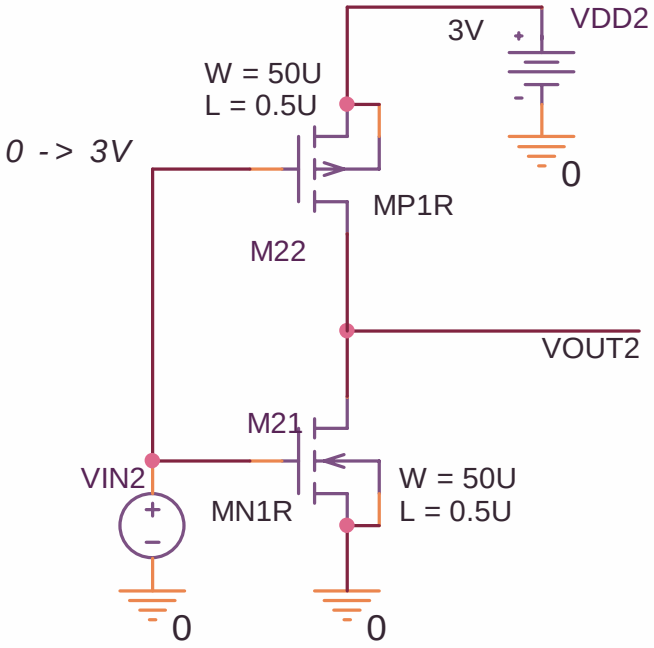
$$R_{ON} = \frac{L-2LD}{\beta'_N W(V_{GS}-V_T)} = \frac{0.34}{134E-6\cdot 50\cdot(3-0.7)} = 22.0\ \Omega$$

$$_{min} = I_{max}R_{ON} = 4.2mV \qquad _{INX} = V_T + \sqrt{\frac{2I_{max}}{\beta}} = 0.7 + \sqrt{\frac{2\cdot 191E-6}{134E-6\cdot(50/0.34)}} = 0.839V$$

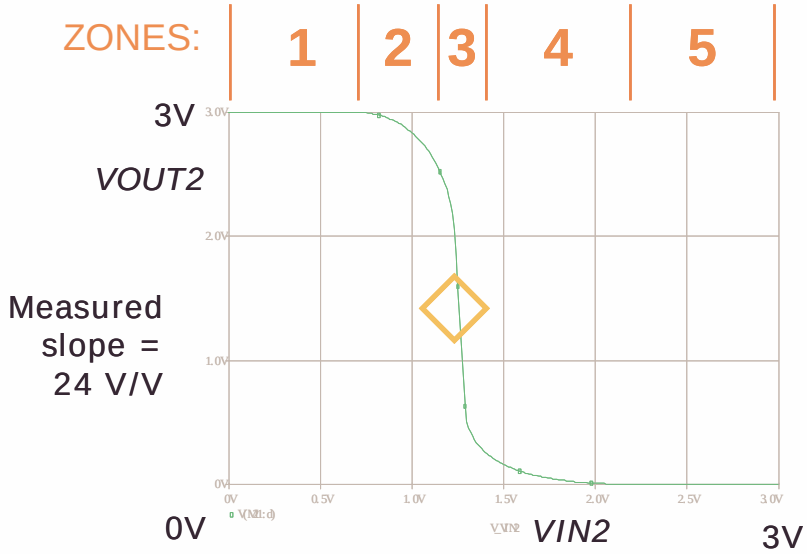
☐ CMOS Inverter / Amp

```
*SPICE LEVEL1 RAZAVI 0.5U PMOS MODEL
.model MP1R PMOS LEVEL=1 VTO=-0.8 GAMMA=0.4 PHI=0.8
+ NSUB=5E14 LD=0.09E-6 UO=100 LAMBDA=0.2 TOX=9E-9

*SPICE LEVEL1 RAZAVI 0.5U NMOS MODEL
.model MN1R NMOS LEVEL=1 VTO=0.7 GAMMA=0.45 PHI=0.9
+ NSUB=9E14 LD=0.08E-6 UO=350 LAMBDA=0.1 TOX=9E-9
```



ZONE	NFET	PFET
1	OFF	NS
2	SAT	NS
3	SAT	SAT
4	NS	SAT
5	NS	OFF



This is the connection for a simple CMOS logic inverter. It has amplifier properties in the transition region where $V_{IN} = V_X$:

$$\frac{\beta_N}{2}(V_{IN} - V_{TN})^2 = \frac{\beta_P}{2}(V_{DD} - V_{IN} - |V_{TP}|)^2$$

Solve for $V_{IN} = V_X$:

$$_{IN} = V_X = \frac{V_{DD} - |V_{TP}| + V_{TN}\sqrt{\beta_N / \beta_P}}{1 + \sqrt{\beta_N / \beta_P}}$$

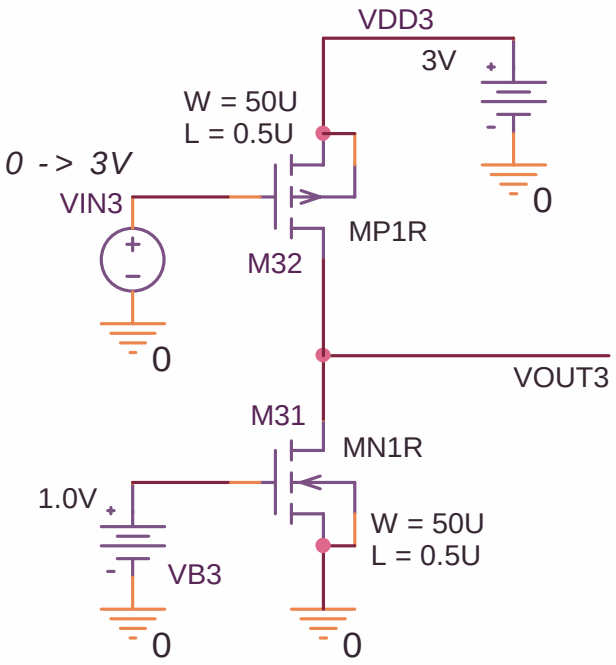
V_X becomes $VDD/2$ when $\beta_N = \beta_P$. Requires larger PFET. Why ?

Common Source Amp

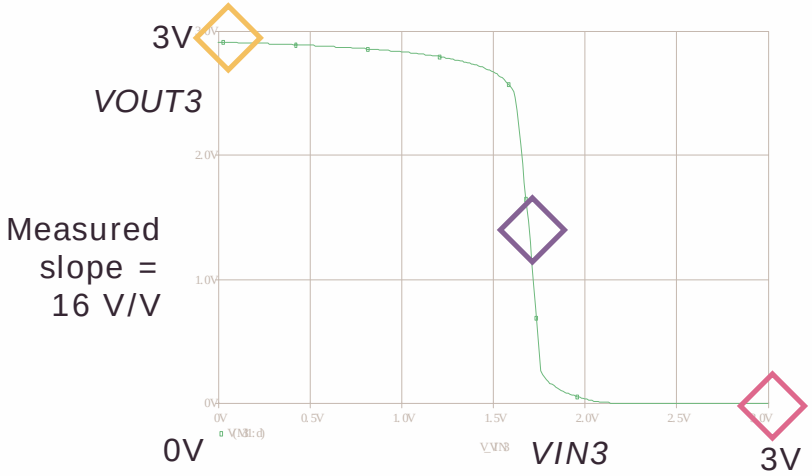
NMOS Load

```
*SPICE LEVEL1 RAZAVI 0.5U PMOS MODEL
.model MP1R PMOS LEVEL=1 VTO=-0.8 GAMMA=0.4 PHI=0.8
+ NSUB=5E14 LD=0.09E-6 UO=100 LAMBDA=0.2 TOX=9E-9

*SPICE LEVEL1 RAZAVI 0.5U NMOS MODEL
.model MN1R NMOS LEVEL=1 VTO=0.7 GAMMA=0.45 PHI=0.9
+ NSUB=9E14 LD=0.08E-6 UO=350 LAMBDA=0.1 TOX=9E-9
```



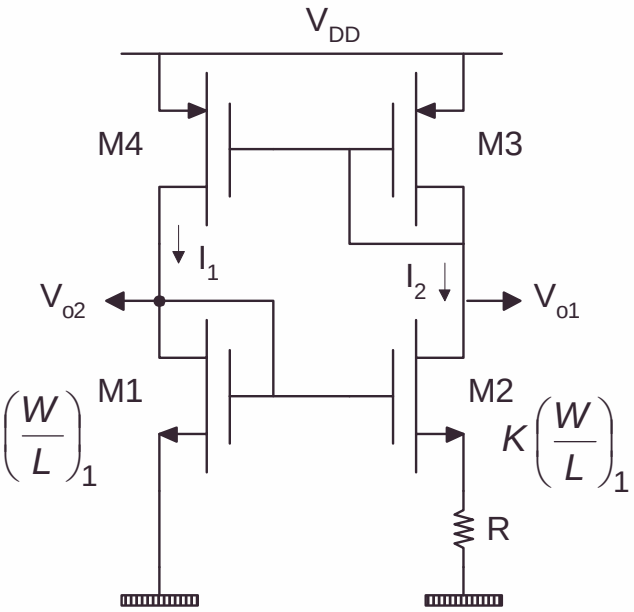
A PMOS driver with a current-sink NMOS load.



The slope, and hence the gain, is much lower than for the PMOS Load case. A few factors contribute to this. We will soon do an ac analysis of circuits like this one, and then the difference in gain values will be explained.

We could now attempt calculations for the following:

Supply Independent Reference



Supply-independent reference

The idea is to generate a voltage or a current which is reasonably immune to the supply V_{DD} , to supply ripple and other fluctuations. Here is one example ←.
We first set $(W/L)_3 = (W/L)_4$ to make $I_1 = I_2$.
Then we set $(W/L)_2 = K(W/L)_1$ with $K > 1$.

The loop equation $V_{GS1}=V_{GS2}+I_2R$ becomes a *nonlinear* one which determines the current uniquely, independent of V_{DD}.

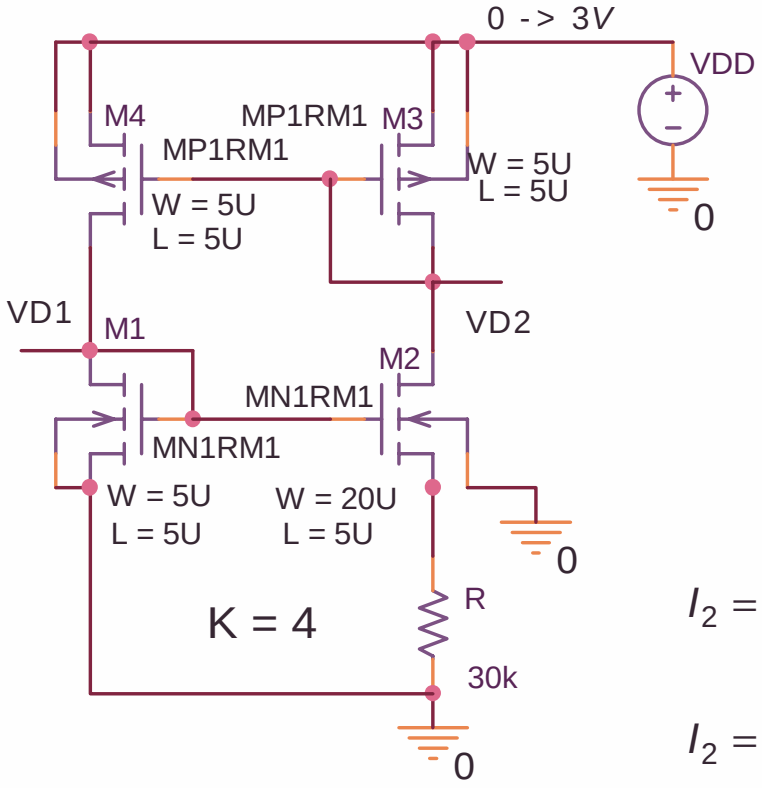
That is: $$V_{TN}+\sqrt{\frac{2I_2}{\beta_N'(W/L)_1}}=V_{TN}+\sqrt{\frac{2I_2}{\beta_N'K(W/L)_1}}+I_2\cdot R$$

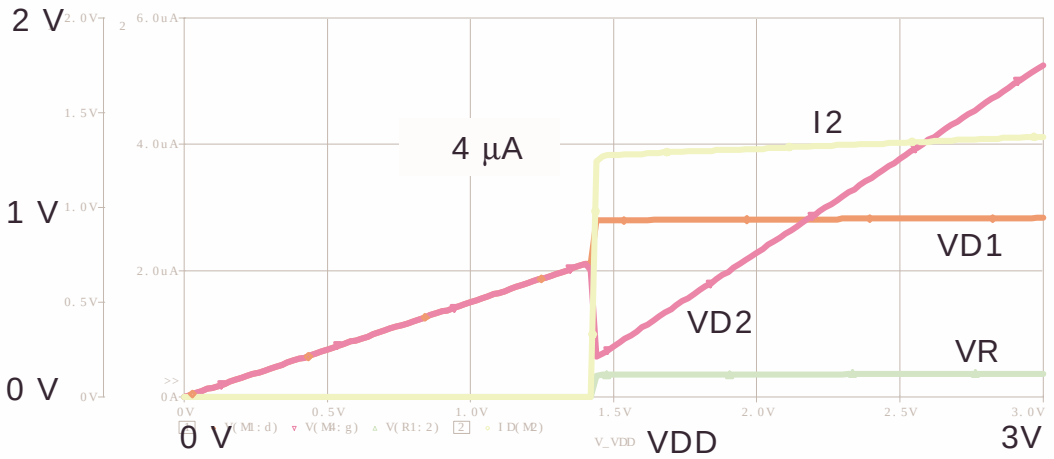
reducing to: $$\sqrt{\frac{2I_2}{\beta_N'(W/L)_1}}\cdot\left(1-\frac{1}{\sqrt{K}}\right)=I_2\cdot R$$

After squaring both sides and dividing by root(I_2): $$I_2=\frac{2}{\beta_N'(W/L)_1\cdot R^2}\cdot\left(1-\frac{1}{\sqrt{K}}\right)^2$$

But notice, the solution $I_1 = I_2 = 0$ is valid also. It can therefore be advisable to have a startup circuit that induces current flow (and switches itself off after power-up).

☐ Supply Independent Reference in *SPICE*





$$I_2 = \frac{2}{\beta_N'(W/L)_1 \cdot R^2}\cdot\left(1-\frac{1}{\sqrt{K}}\right)^2 = \frac{2}{134E-6(1.0)\cdot 900E6}\cdot\left(1-\frac{1}{\sqrt{4}}\right)^2$$

$I_2 = 4.0\,\mu A$ closely matched by the simulation result.

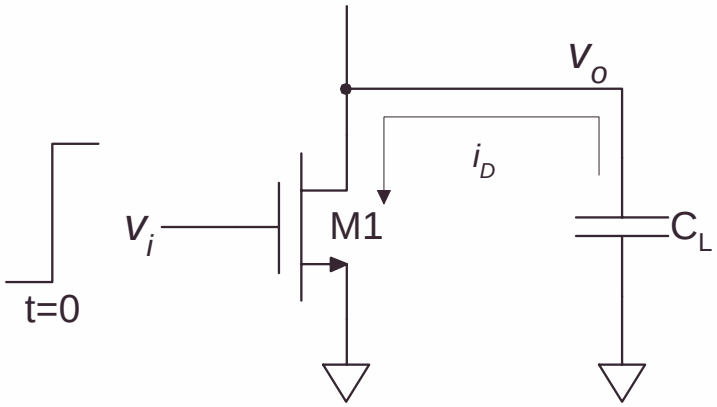
It's important that *R* be sufficiently large !

Use long MOSFETS to keep LAMBDA small.

Current is reduced when GAMMA > 0.

```
*SPICE LEVEL1 RAZAVI 0.5U NMOS MODEL
.model MN1R NMOS LEVEL=1 VTO=0.7 GAMMA=0.0 PHI=0.9
+ NSUB=9E14 LD=0.08E-6 UO=350 LAMBDA=0.01 TOX=9E-9
*SPICE LEVEL1 RAZAVI 0.5U PMOS MODEL
.model MP1R PMOS LEVEL=1 VTO=-0.8 GAMMA=0.0 PHI=0.8
+ NSUB=5E14 LD=0.09E-6 UO=100 LAMBDA=0.02 TOX=9E-9
```

☐ CMOS Inverter Transient Analysis



Here ← we see just the NMOS transistor of a CMOS inverter. When V_i goes high, C_L discharges through M1, and the output V_o falls from its initial value of V_{o1} toward zero. The PMOS is now OFF, so we can ignore it as we analyze the transient ..

M1 is initially saturated, but it goes non-saturated when V_o falls below $V_i - V_T$, where V_i is the input *high* level.

For the saturated discharge: $$\frac{\beta}{2}(V_i - V_T)^2 = -C_L \frac{dV_o}{dt}$$

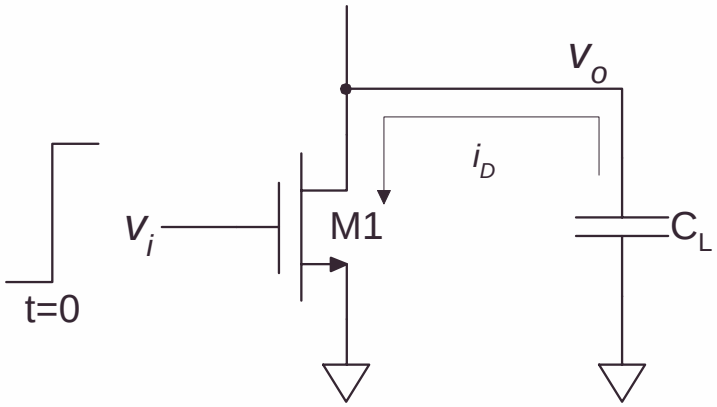
The solution uses a time constant τ : $$\tau = \frac{C_L}{\beta(V_i - V_T)}$$

and it is easy to check that the diff-eqn leads to this result ➔, with duration T_S for the saturated portion of the discharge: $$\frac{T_S}{\tau} = 2\left(\frac{V_{o1}}{V_i - V_T} - 1\right)$$

The subsequent *non-saturated* discharge is governed by: $$\beta\left[(V_i - V_T)V_o - \tfrac{1}{2}V_o^{\,2}\right] = -C_L \frac{dV_o}{dt}$$

This is a more complicated diff-eqn, but can be solved with the help of this indefinite integral: $$\int dx / [x(ax+b)] = (1/b)\cdot Ln[x/(ax+b)]$$

CMOS Inverter Transient Analysis (Cont'd)



For the NON-SAT portion of the transient, down to a voltage level V_{o2}, we get this result ➔

$$\frac{T_{NS}}{\tau} = Ln\left(\frac{2(V_i - V_T)}{V_{o2}} - 1\right)$$

The overall discharge time T_D from initial V_{o1} to a final V_{o2} is the sum of the two components:

$$\frac{T_D}{\tau} = \frac{T_S + T_{NS}}{\tau} = 2\left(\frac{V_{o1}}{V_i - V_T} - 1\right) + Ln\left(\frac{2(V_i - V_T)}{V_{o2}} - 1\right)$$

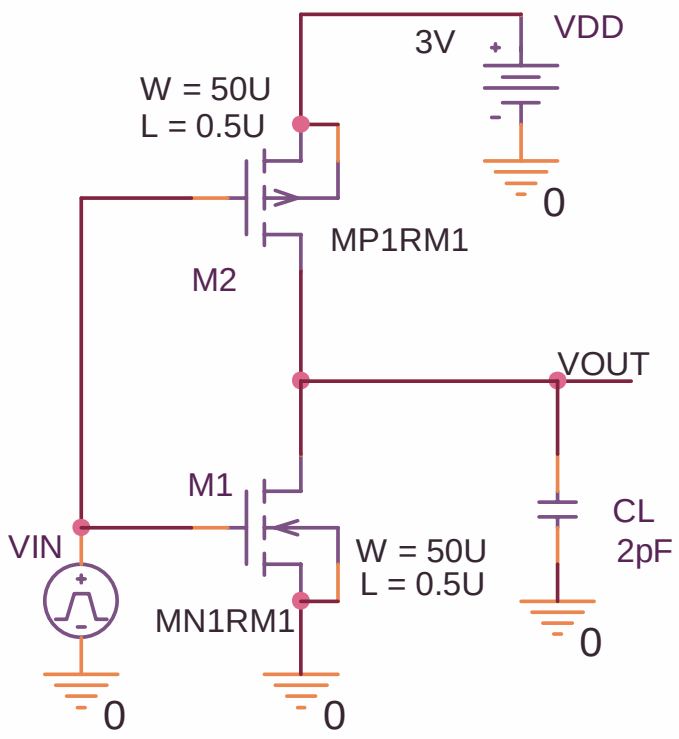
To illustrate for a 3V logic supply:

the discharge time (as a multiple of τ) through the NMOS device (with V_{TN} = 0.7V) from 3V down to 1.5V becomes:

$$\frac{T_D}{\tau} = 2\left(\frac{3}{3-0.7} - 1\right) + Ln\left(\frac{2(3-0.7)}{1.5} - 1\right) = 1.33$$

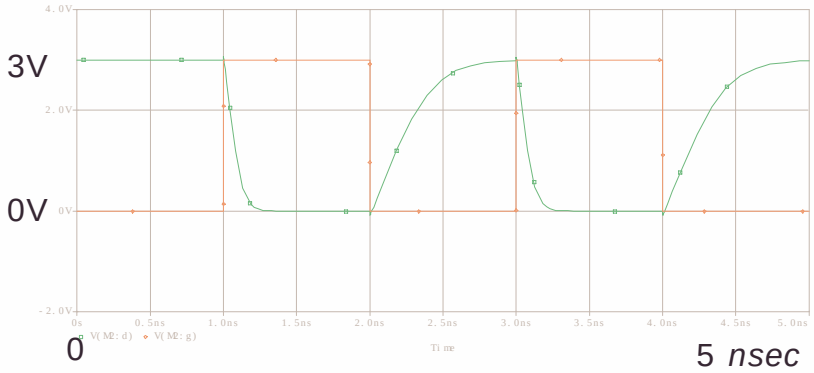
There is no need to do a separate analysis of the charging transient. The symmetry of the CMOS inverter makes it clear that the same analysis applies, using PMOS particulars instead of NMOS. A SPICE LEVEL 1 simulation can be used to check these ideas ..

CMOS Inverter Transients



A SPICE transient analysis

$$\tau = \frac{C_L}{\beta(V_i - V_T)}$$



$$\tau_{NMOS} = \frac{2E-12}{134E-6\{50/(0.5-0.16)\}(3-0.7)} = 0.044\ n\text{sec}$$

$$\tau_{PMOS} = \frac{2E-12}{38.3E-6\{50/(0.5-0.18)\}(3-0.8)} = 0.152\ n\text{sec}$$

$T_{FALL} = 1.33 \cdot 0.044 = 0.058\ n\text{sec}$ (SPICE: 0.065)

$T_{RISE} = 1.33 \cdot 0.152 = 0.202\ n\text{sec}$ (SPICE: 0.233)

```
*SPICE LEVEL1 RAZAVI 0.5U NMOS MODEL MOD1: GAMMA LAMBDA
.model MN1RM1 NMOS LEVEL=1 VTO=0.7 GAMMA=0.00 PHI=0.9
+ NSUB=9E14 LD=0.08E-6 UO=350 LAMBDA=0.0
+ TOX=9E-9 PB=0.9 CJ=0.56E-3 CJSW=0.35E-11
+ MJ=0.45 MJSW=0.2 CGDO=0.4E-9 JS=1.0E-8
```

```
*SPICE LEVEL1 RAZAVI 0.5U PMOS MODEL MOD1: GAMMA LAMBDA
.model MP1RM1 PMOS LEVEL=1 VTO=-0.8 GAMMA=0.0 PHI=0.8
+ NSUB=5E14 LD=0.09E-6 UO=100 LAMBDA=0.0
+ TOX=9E-9 PB=0.9 CJ=0.94E-3 CJSW=0.32E-11
+ MJ=0.5 MJSW=0.3 CGDO=0.3E-9 JS=0.5E-8
```

SPICE times are slightly longer -- but note that we neglected the MOSFET capacitances (about 100 fF).

When λ is included (0.1 NMOS, 0.2 PMOS), the transients go faster: 0.053 FALL and 0.161 RISE)

MOS SMALL SIGNAL CIRCUIT BASICS

☐ Small-Signal Parameters : R_{ON}

Small-signal parameters describe transistor behaviour for purposes of small excursions around an operating point. For amplifier purposes, a MOSFET should be saturated (SAT), but other MOSFETs in the circuit may be non-saturated (NON-SAT), and we need small-signal models for both.

NON-SAT OPERATION (operation as a SWITCH):

A non-saturated condition generally involves a low V_{DS}, allowing some simplification of the current equation:

$$I_D = \beta(V_{GS} - V_T - \tfrac{1}{2}V_{DS})V_{DS} \approx \beta(V_{GS} - V_T)V_{DS}$$

$$\beta = \beta' \left(\frac{W}{L}\right)$$

Then: $dI_D / dV_{DS} \approx \beta(V_{GS} - V_T)$ and $R_{ON} \approx 1/[\beta(V_{GS} - V_T)]$

R_{ON} is the ON-RESISTANCE measured in Ohms. It provides a simple characterisation which is generally adequate when V_{DS} is small. It is both a large-signal and a small-signal device parameter, and no other parameter is needed at low frequency.

SAT OPERATION (operation as an AMPLIFIER):

Here we need a more detailed current equation, as follows:

$$I_{DS} = \frac{\beta}{2}(V_{GS} - V_T)^2 \cdot (1 + \lambda V_{DS}) \quad \text{where} \quad {}_T = V_{T0} + \gamma \cdot \left(\sqrt{V_{SB} + 2\Phi_F} - \sqrt{2\Phi_F}\right)$$

☐ Small-Signal Parameters : g_m , g_{mb}

Combining these equations:

$$I_D = \frac{\beta}{2}\left(V_{GS} - V_{T0} - \gamma\sqrt{V_{SB} + 2\Phi_F} + \gamma\sqrt{2\Phi_F}\right)^2 \cdot (1 + \lambda V_{DS})$$

The current now depends on V_{GS}, on V_{DS} , and also on V_{SB}. These three dependencies lead to three separate AC parameters, as follows.

Transconductance: $g_m = dI_D / dV_{GS}$:

Differentiation yields: $g_m = dI_D / dV_{GS} = \beta\left(V_{GS} - V_{T0} - \gamma\sqrt{V_{SB} + 2\Phi_F} + \gamma\sqrt{2\Phi_F}\right) \cdot (1 + \lambda V_{DS})$

which we can write more simply as: $g_m = \sqrt{2\beta I_{D0}(1 + \lambda V_{DS0})}$ I_{D0} and V_{DS0} are bias values at the DC operating point.

g_m is the primary current-control parameter. MOSFET g_m values are typically much lower than corresponding bipolar values { $g_m = I_{C0}/(kT/q)$ }.

Back-Gate Transconductance: $g_{mb} = dI_D / dV_{BS}$:

$$g_{mb} = dI_D / dV_{BS} = \beta\left(V_{GS} - V_{T0} - \gamma\sqrt{V_{SB} + 2\Phi_F} + \gamma\sqrt{2\Phi_F}\right) \cdot (1 + \lambda V_{DS}) \cdot \frac{\gamma}{2\sqrt{V_{SB} + 2\Phi_F}}$$

☐ Small-Signal Parameters : g_{mb} , δ , g_d

which we can write more simply as: $$g_{mb} = \frac{g_m \cdot \gamma/2}{\sqrt{V_{SB0} + 2\Phi_F}} = \delta \cdot g_m \qquad \text{with} \qquad \delta = \frac{\gamma/2}{\sqrt{V_{SB0} + 2\Phi_F}}$$

(This same δ can be shown to approximate V_T as $\{V_{TS} + \delta.V_{CS}\}$ where V_{CS} is the channel-to-source potential at any point in the channel).

The typical δ value is about 0.3. Clearly, this "back-gate" transconductance is not a small one. But g_{mb} is active only when the Source voltage has an AC component.

Drain Transconductance: $g_d = dI_D/dV_{DS}$:

$$g_d = dI_D/dV_{DS} = \frac{\beta}{2}(V_{GS} - V_T)^2 \cdot \lambda = \frac{\lambda I_{D0}}{(1 + \lambda V_{DS0})}$$

I_{D0} and V_{DS0} are bias values at the DC operating point.

Also: $$r_d = 1/g_d = \frac{1 + \lambda V_{DS0}}{\lambda I_{D0}}$$ r_d is the AC drain output resistance

Approximate parameters for a quick estimate:

$$g_m = \sqrt{2\beta I_{D0}} \qquad g_{mb} = \delta \cdot g_m \qquad g_d = \lambda I_{D0}$$

☐ Small-Signal Parameters : using V_e

Some useful expressions for g_m :

Recall that: $V_{GS} \approx V_T + \sqrt{2I_D / \beta} = V_T + V_e$ where V_e is the "**effective voltage**" or the "**overdrive**"

Then: $g_m \approx \sqrt{2\beta I_{D0}} = \sqrt{\beta^2 \cdot \frac{2I_D}{\beta}} = \beta\sqrt{\frac{2I_D}{\beta}} = \beta V_e$ .. worth noting

We get an alternative expression, involving the bias current I_D, by observing that: $I_D = \frac{\beta}{2}(V_{GS} - V_T)^2 = \frac{\beta}{2}V_e^2$

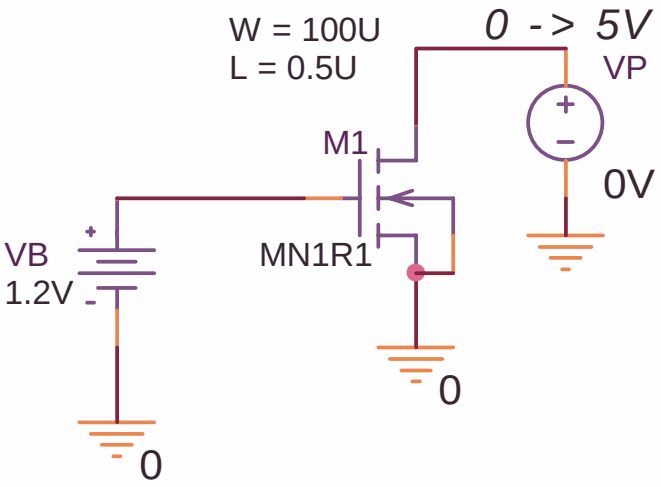
It then follows that: $g_m = \frac{2I_D}{V_e}$

To summarize:

$$g_m = \sqrt{2\beta I_{D0}} \qquad g_m = \beta V_e \qquad g_m = \frac{2I_D}{V_e}$$

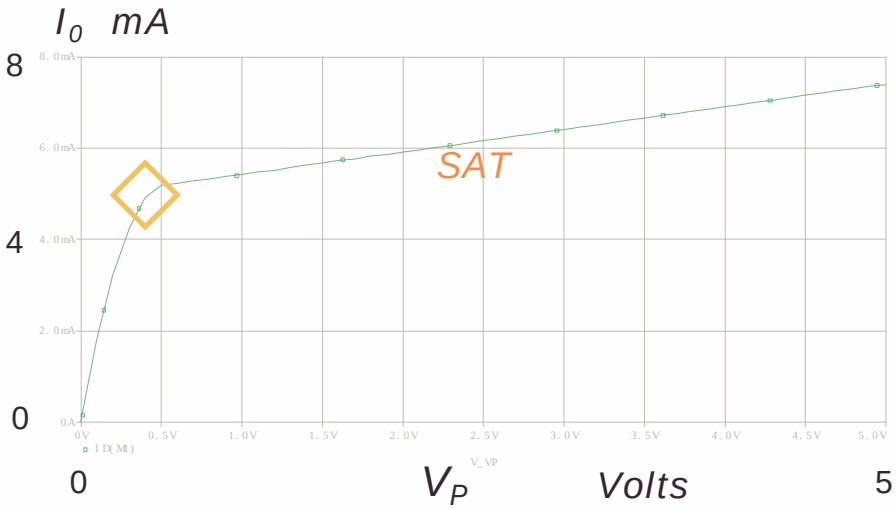
These represent different insights, the use of which depends on the constraints imposed by the problem at hand

Current Sink Example Revisited



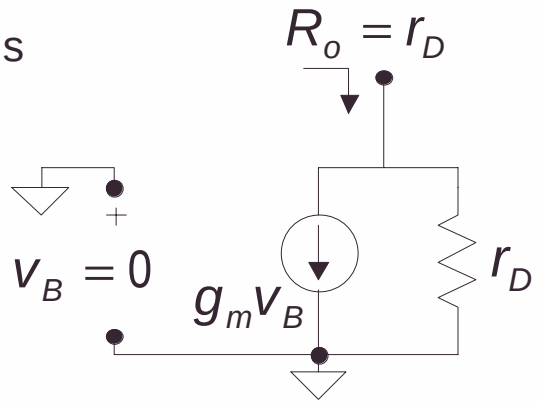

```
*SPICE LEVEL1 RAZAVI 0.5U NMOS MODEL MODS 1 GAMMA
.model MN1R1 NMOS LEVEL=1 VTO=0.7 GAMMA=0.00 PHI=0.9
+ NSUB=9E14 LD=0.08E-6 UO=350 LAMBDA=0.1
+ TOX=9E-9 PB=0.9 CJ=0.56E-3 CJSW=0.35E-11
+ MJ=0.45 MJSW=0.2 CGDO=0.4E-9 JS=1.0E-8
```

Measured SAT slope = 2.02kΩ



Small-signal Model

Current sink is zero because v_B (ac) = 0

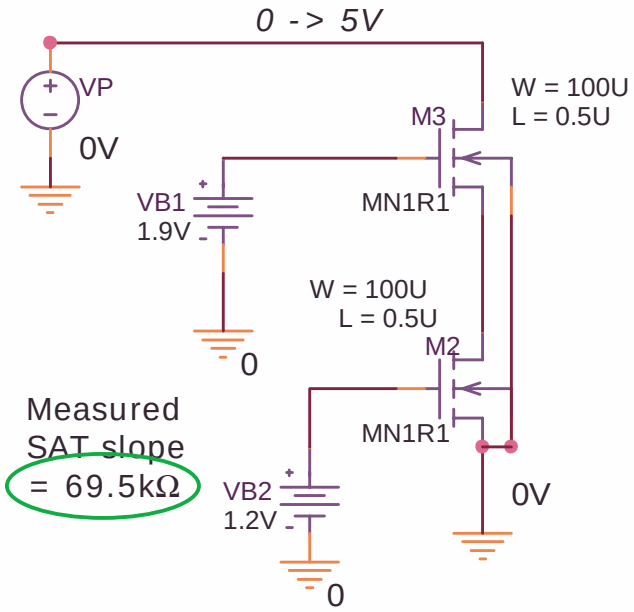


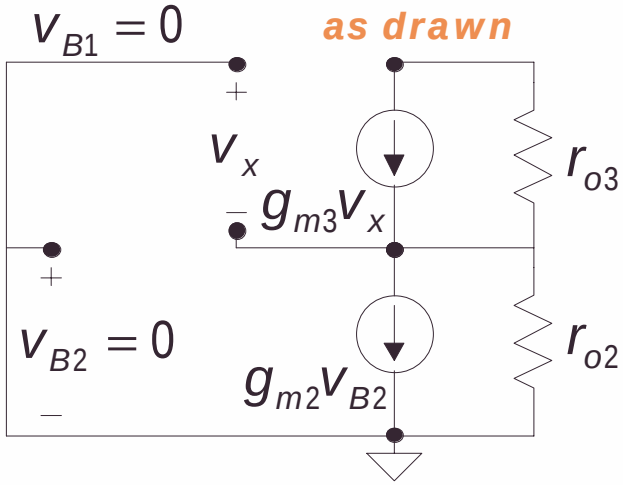
Calculations: $I_0 = 4.9mA$

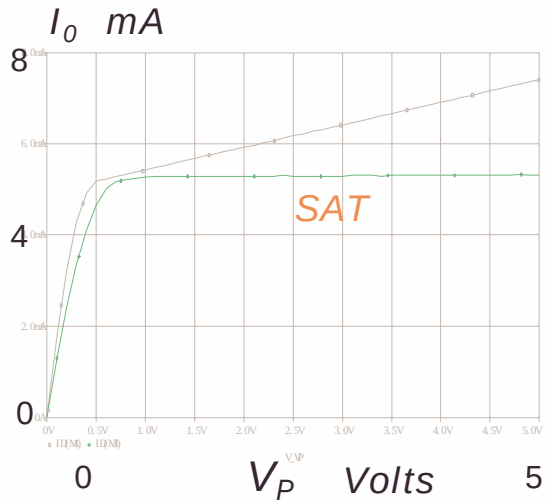
$$g_d = \lambda_N \cdot I_O = 0.1 \cdot 4.9mA = 0.49mA/V \qquad R_O = \frac{1}{g_d} = 2.04k\Omega$$

The computed small-signal R_O matches the slope from the large-signal plot – as it should do !

Improved Current Sink Revisited







Two equations to solve:

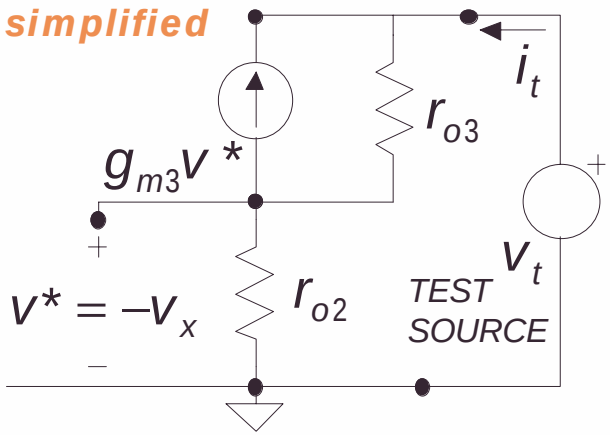
$$v^* = i_t r_{o2}$$

$$v_t = i_t r_{o2} + (i_t + g_{m3} v^*) r_{o3}$$

Solving for $R_O = v_t / i_t$:

$$R_O = r_{o2} + r_{o3} + g_{m3} r_{o3} r_{o2}$$

$$R_O \approx g_{m3} r_{o3} r_{o2}$$



The *TEST SOURCE* V_t is for calculation of output impedance R_O.

Quick Check :

$I_B \approx 5.3mA$ from graph

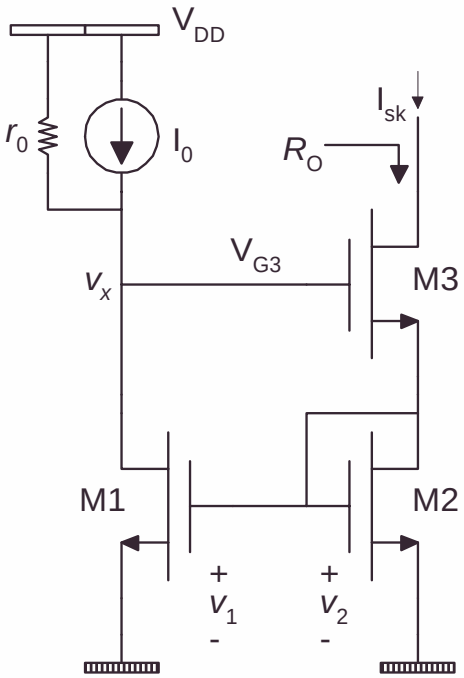
$$g_{m3} = \sqrt{2\beta I_B} = 0.021$$

$$r_{o2} = r_{o3} = \frac{1}{\lambda_N I_B} = 1.82k\Omega$$

$$R_O \approx g_{m3} r_{o3} r_{o2} = 72.8\,k\Omega$$

.. closed to measured slope

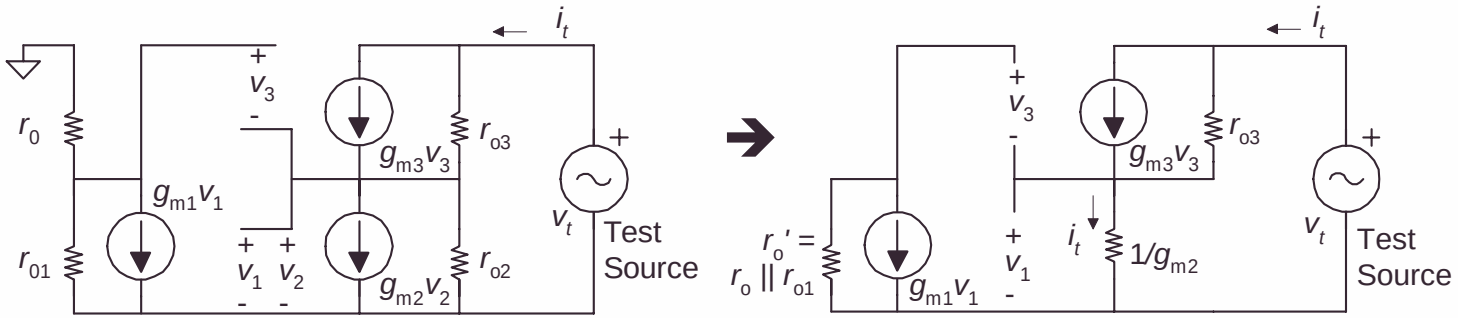
Wilson Current Mirror and Gain Boosting



Wilson current mirror

The Wilson mirror was first intended as a bipolar current mirror, but the same topology can be used with MOS circuits also. This mirror uses feedback to further boost the output impedance.

Any increase in sink current I_{sk} will cause v_2 to increase. This v_2 is then amplified by M1, using I_o as a load, for a large inverting gain between v_2 and v_x. Therefore v_x falls sharply, reducing the drive on M3, and opposing the initial current increase.



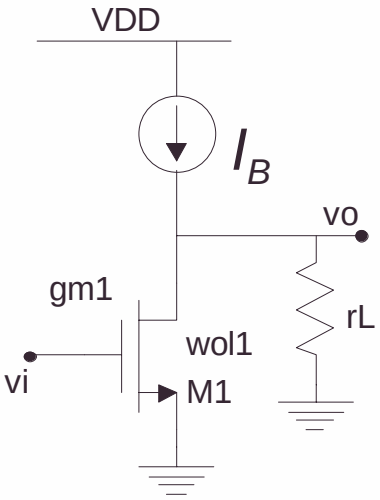
Two Equations: $-g_{m1}(i_t/g_{m2})r'_0=(i_t/g_{m2})+v_3$ and $(i_t-g_{m3}v_3)r_{o3}+(i_t/g_{m2})=v_t$

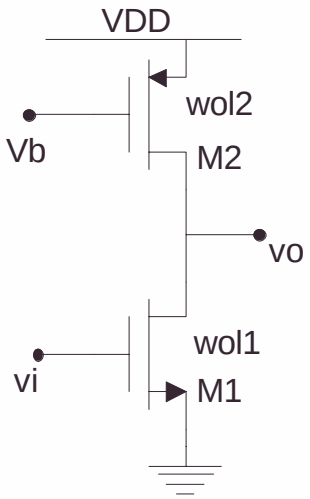
Solution yields: $v_t/i_t=R_o=1/g_{m2}+r_{o3}+r_{o3}g_{m3}r_o'(g_{m1}/g_{m2})+(g_{m3}/g_{m2})r_{o3}$

Ignoring smaller terms: $R_o\approx(g_{m1}r_o')(g_{m3}r_{o3}/g_{m2})$ where $r_o'=r_o\|r_{o1}$

The *normal cascode gain* $(g_{m3}r_{o3}/g_{m2})$ *is boosted by the feedback gain term* $(g_{m1}r_o')$

☐ Common Source Amp : AC Analysis





Concept:

The device M1 uses a current-source load I_B and operates in SAT mode with v_o biased near VDD/2 . r_L is the ac impedance at the output node, such that ac current variations i_D must flow in r_L.

$$i_D = g_{m1} \cdot v_i \qquad v_o = -i_D \cdot r_L \qquad \rightarrow \qquad \frac{v_o}{v_i} = -g_{m1} \cdot r_L$$

Implementation:

In practice, the current source is implemented using a PMOS transistor and the ac load r_L is r_{o1} in parallel with r_{o2}, that is, the combined output Drain impedances of M1 and M2.

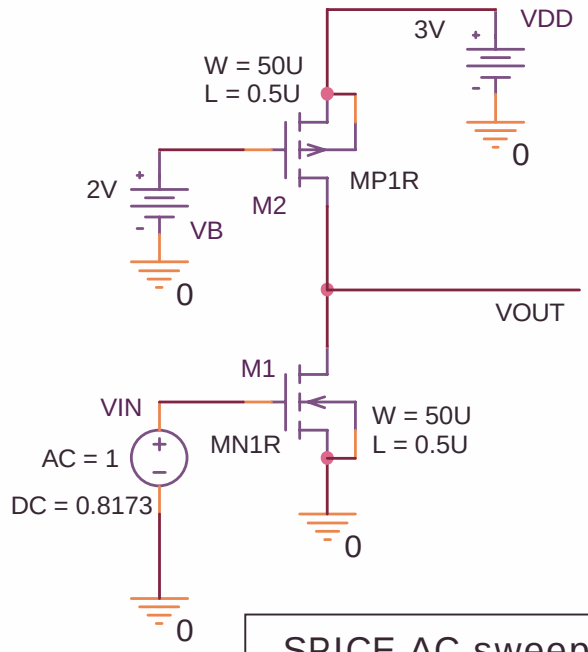
$$\frac{v_o}{v_i} = -g_{m1} \cdot (r_{o1} \,||\, r_{o2}) = \frac{-g_{m1}}{g_{o1} + g_{o2}}$$

with:

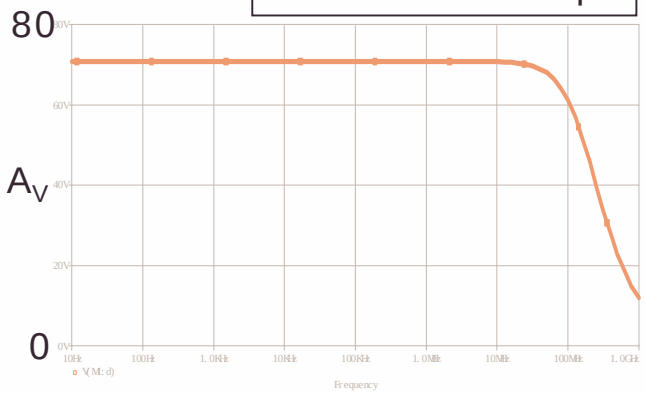
$$g_{m1} = \sqrt{2\beta_1 I_B (1 + \lambda_N V_{DS0})} \qquad g_{o1} = \frac{\lambda_N I_B}{(1 + \lambda V_{DS1})} \qquad g_{o2} = \frac{\lambda_P I_B}{(1 + \lambda V_{DS2})}$$

Now draw a small-signal model and use it as a check on our reasoning in order to verify the gain equation.

☐ Common Source Amp : Revisited



SPICE AC sweep



PMOS Load

For the AC sweep, the DC VIN must first be chosen to set VOUT = VDD/2. The *measured* AC gain was then –70.8.

Small-signal calculations are as follows:

$$I_B = 0.5\beta_p' \frac{W_2}{L_2 - 2LD_P}(V_{SG2} - |V_{TP}|)^2 \cdot (1 + \lambda_P \cdot V_{SD2})$$

$$= 0.5(38.3E-6)\frac{50}{0.5-0.18}(1-0.8)^2 \cdot (1+0.2\cdot 1.5) = 156\,\mu A$$

$$g_{m1} = \sqrt{2\beta_N'(W/L)_1 I_B (1+\lambda_N(1.5))}$$

$$= \sqrt{2(134E-6)(50/0.34)(156E-6)(1+0.1\cdot 1.5)} = 2.66\,mA/V$$

$$g_{o1} = \frac{\lambda_N I_B}{1+\lambda_N \cdot (1.5)} = 13.5\,\mu A/V$$

$$g_{o2} = \frac{\lambda_P I_B}{1+\lambda_P \cdot (1.5)} = 23.9\,\mu A/V$$

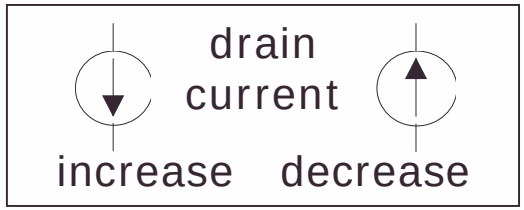
$$A_v = \frac{v_o}{v_i} = \frac{-g_{m1}}{g_{o1}+g_{o2}} = -70.8$$

The calculated ac gain A_V is also the measured value (above), and is quite close to the measured slope of 68 V/V from our earlier large signal analysis ←.

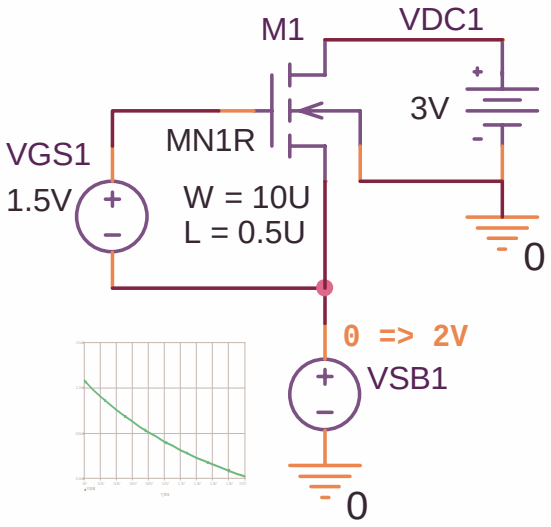
☐ Body Effect Revisited: small-signal models

With body effect, we gave a "back-gate" g_{mb} as well as the "front-gate" g_m.

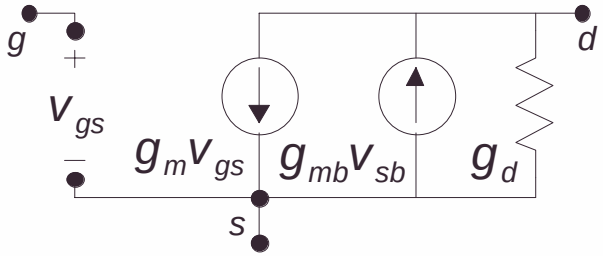
Variations in V_{GS} and in V_{SB} ➔ incremental current changes, which appear in the small-signal models as follows:



We get body-effect when the source *S* moves relative to the bulk *B* (which is ac-grounded)

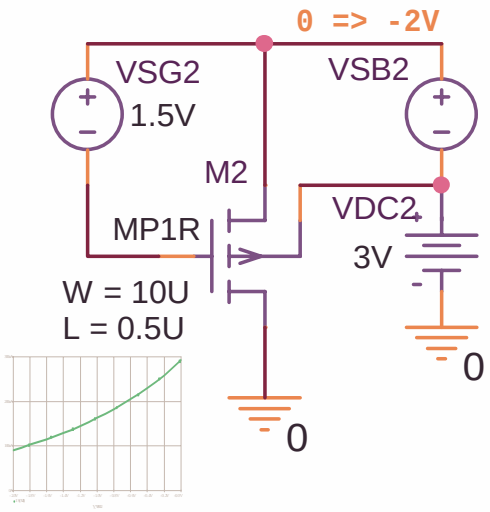


Current falls from 1.26mA (0V) to 0.42mA (2V)

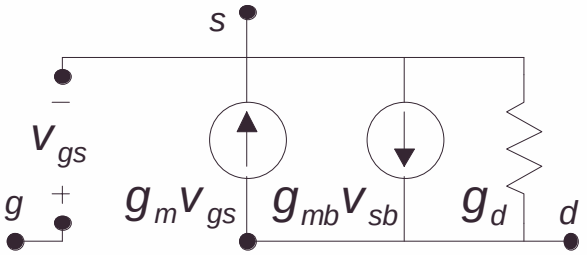


As V_{GS} rises, I_D increases

As V_{SB} rises, I_D decreases



Current falls from 293µA (0V) to 90µA (-2V)

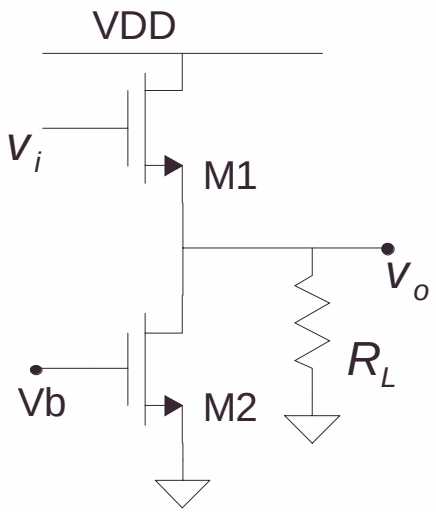


As V_G rises, I_D decreases

As V_{SB} rises, I_D increases

☐ Common Drain Amplifier (Source Follower)

Used as a level shifter (or sometimes as an impedance buffer) in which v_o "follows" v_i (but is V_T+V_e below it). If V_T+V_e were fixed, the ac gain would be 1.0. In practice, both of them vary.



See ac models below. The ac analysis goes:

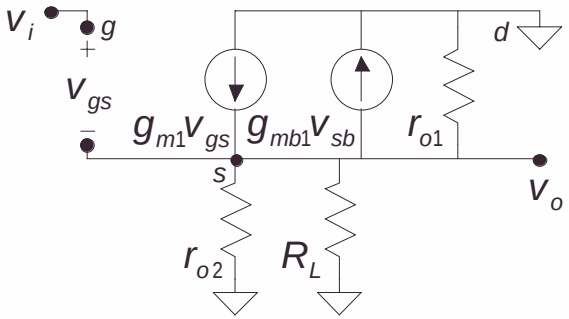
$$i_d = g_{m1}(v_i - v_o) = \frac{v_o}{R_{eq}} \quad \rightarrow \quad \frac{v_o}{v_i} = \frac{g_{m1}}{g_{m1} + \frac{1}{R_{eq}}}$$

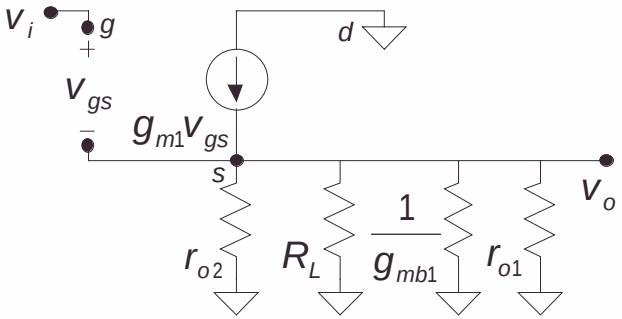
When R_L and r_{o1} and r_{o2} are large: $\frac{v_o}{v_i} = \frac{g_{m1}}{g_{m1}+g_{mb1}} = \frac{1}{1+\delta}$

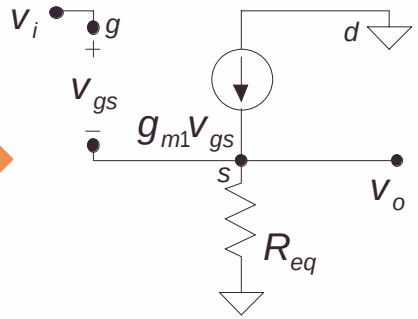
Body effect limits the available gain and reduces linearity.

Use of R_L without the M2 sink makes gain even more non-linear

The V_{GS} of M1 reduces available headroom.



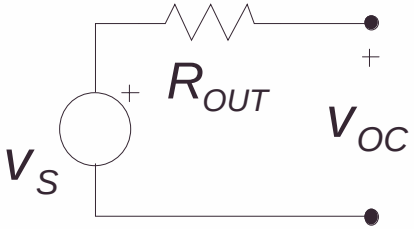




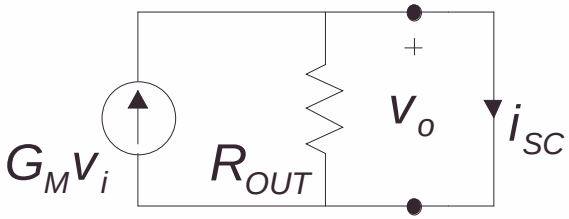
☐ Small Signal Linear Modelling

On an AC equivalent circuit, DC voltage sources become zero and constant current sources go open circuit.

The small-signal models are *linear*. This means that we can view them from their output terminals as Thevenin or as Norton equivalents. For both models, R_{OUT} is the output resistance seen when all independent sources are set to zero.



The Thevenin model : v_{OC} is the open-circuit output voltage



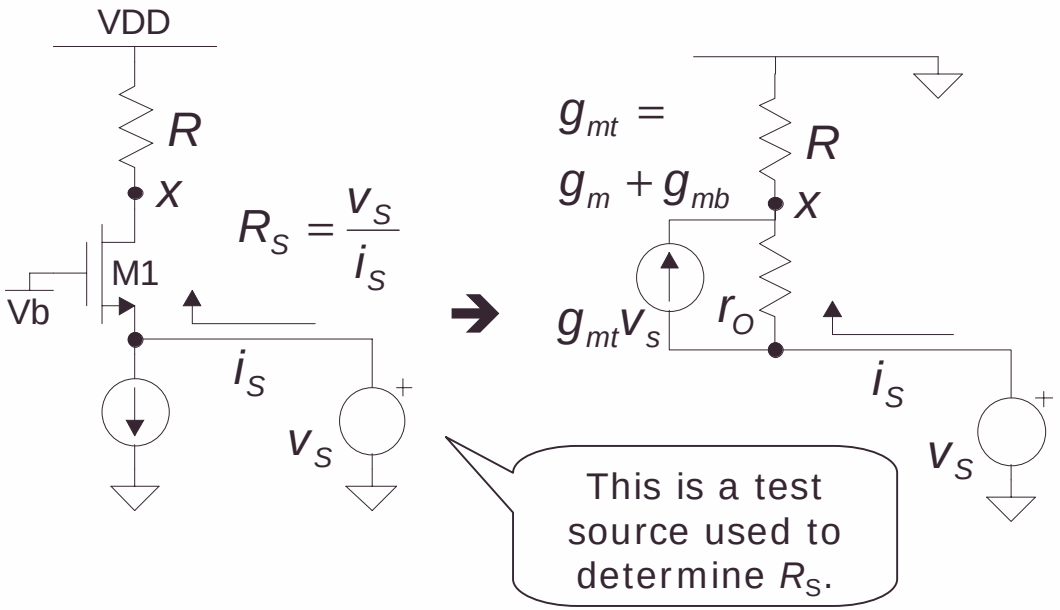
The Norton model : i_{SC} is the short-circuit output current

To use the Norton model, we evaluate R_{OUT} and we short the output terminals to find i_{SC} and hence G_M . Then, for a circuit driven by v_i , we find the gain as $v_O/v_i = G_M R_{OUT}$.

This approach can often reduce the amount of algebra required.

Impedance at a pair of terminals may be calculated by applying a voltage to the terminals in question and measuring the resulting current, with all independent sources set to zero. We will demonstrate this for some common situations ..

☐ Resistor-loaded Source Impedance



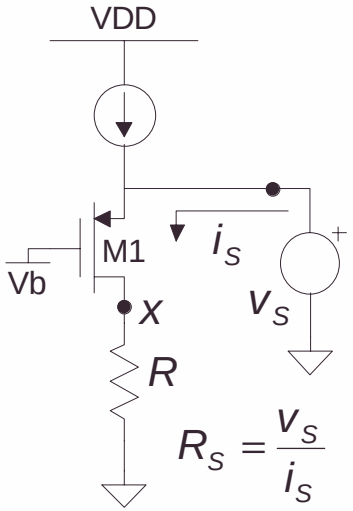
With no AC signal on the Gate, the g_m and g_{mb} will work together as a single $g_{mt} = g_m + g_{mb}$.

On inspection: $v_s = (i_s - g_{mt}v_s)r_O + i_S R$

Re-arranging: $\boxed{R_S \equiv \frac{v_S}{i_S} = \frac{r_O + R}{1 + g_{mt}r_O}}$

Generally: $g_{mt}r_O >> 1$

So, if $r_O >> R$: $R_S \approx \frac{1}{g_{mt}}$

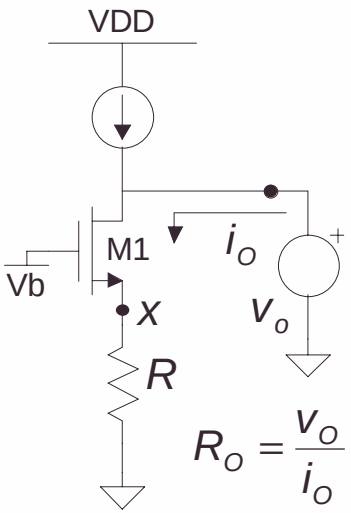


When $\lambda = 0$, r_o = infinity, and $R_S = 1/g_{mt}$, a ***low*** value In most other situations, $r_o >> R$ and R_S remains close to $1/g_{mt}$. But, if the Drain loading R is exceptionally high ($>> r_O$), then the view from the source shows *an attenuated R*, with a value of $R/(g_{mt}r_O)$, higher than the normal $1/g_{mt}$ value.

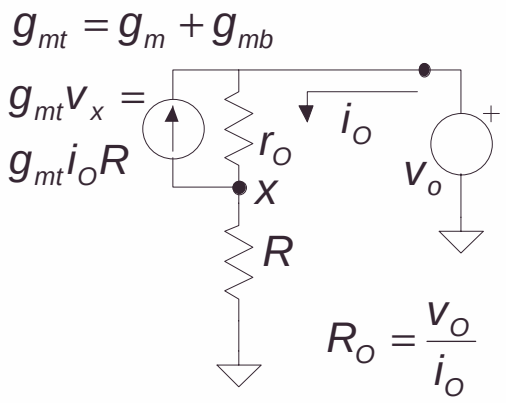
The very same analysis applies to this PMOS circuit ←.

All this remains valid when R is an AC resistance of an active device.

☐ Resistor-loaded Drain Impedance



Here again, the effective g_m is $g_m + g_{mb} = g_{mt}$ (*i.e.* the *total* g_m).

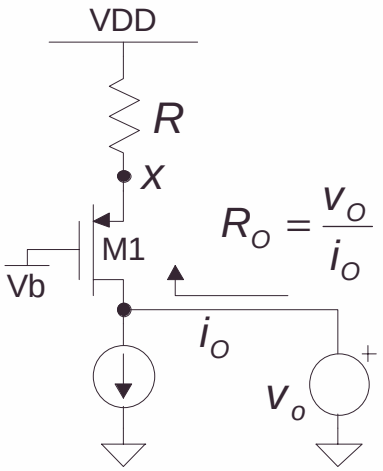


$$v_O = i_O R + r_O\left[i_O + g_{mt} i_O R\right]$$

from which:

$$R_O \equiv \frac{v_O}{i_O} = R + r_O\left[1 + g_{mt} R\right]$$

"the r_O boost factor" $= \left[1 + g_{mt} R\right]$



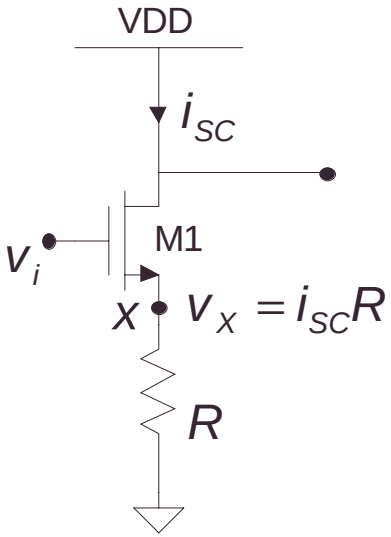
The reaction from R on M1 has the effect of *amplifying* r_o by a factor of $[1 + g_{mt} R]$. Also, on inspection:

$$\frac{v_X}{v_O} = \frac{R}{R + r_O\left[1 + g_{mt} R\right]}$$

Typically, only a small part of v_O appears at node x. This is often helpful as *a de-sensitising effect*.

The same analysis holds for this ← PMOS circuit, and *it still applies when R represents a small-signal resistance* of an active device. Accordingly, these results have wide applicability.

Degenerated CS Amplifier



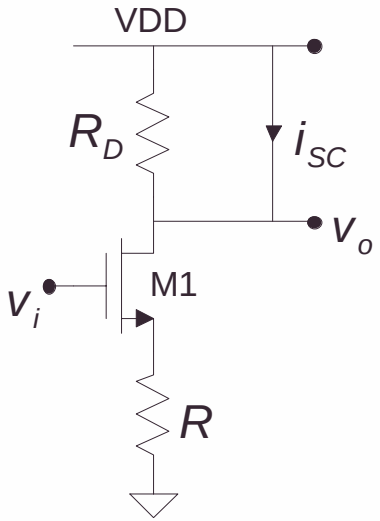
Degeneration (using series R) makes a CS stage more linear.

🡸 On inspection: $i_{SC} = g_m(v_i - v_X) - g_{mb}v_X - v_X / r_O$ with $v_X = i_{SC}R$

Eliminating v_X and re-arranging:

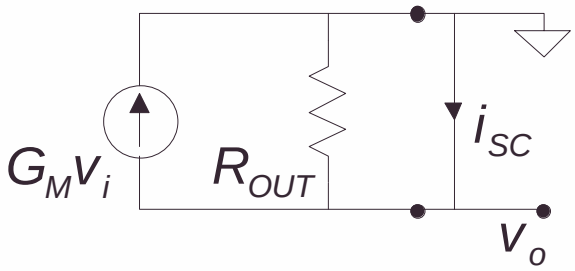
$$\frac{i_{SC}}{v_i} \equiv G_M = \frac{g_m r_O}{R + r_O[1 + g_{mt}R]} \quad \text{with} \quad g_{mt} = g_m + g_{mb}$$

Notice, the denominator is the Resistor-boosted Drain Impedance.
As an "amplifier" this has no output ! , but now consider this ..



🡸 This circuit has the same i_{SC} , but, when the short-circuit (S/C) is removed, it becomes a degenerated CS amp. It then has a Norton model 🡺 for which:

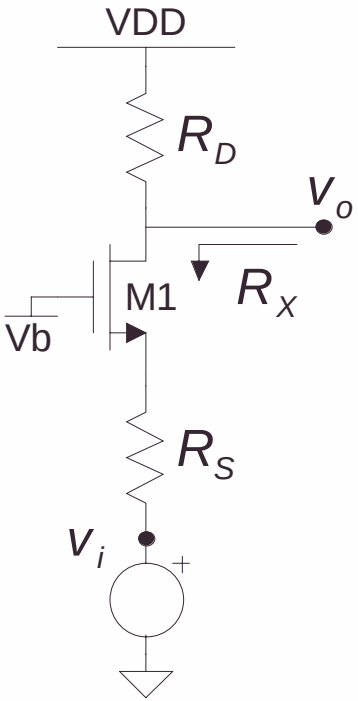
$$R_{OUT} = R_D \,||\, \{R + r_O[1 + g_{mt}R]\}$$



The Norton model with S/C removed shows that:

$$\frac{v_O}{v_i} = -G_M R_{OUT} = \frac{-g_m r_O}{R + r_O[1 + g_{mt}R]} \cdot [R_D \,||\, \{R + r_O[1 + g_{mt}R]\}]$$

☐ Common Gate Amplifier

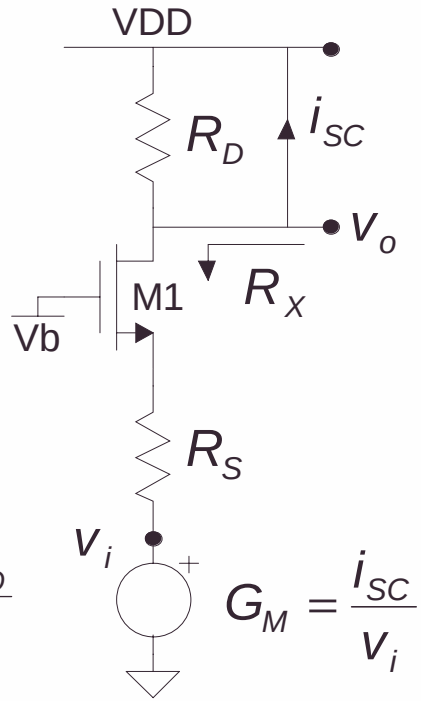


The CG amp ← features a positive gain > 1.0 with low input impedance (which is often undesirable but it ***is*** suitable for wideband operation into a 50Ω load). We'll use a Norton source model → to simplify the analysis. We can find the gain as:

$$\frac{v_O}{v_i} = G_M R_{OUT} = G_M (R_D \,||\, R_X)$$

where: $R_X = R_S + r_O[1 + g_{mt} R_S]$ and ..

$$G_M = \frac{i_{SC}}{v_i} = \frac{1}{R_S + \dfrac{r_O}{1 + g_{mt} r_O}} = \frac{1 + g_{mt} r_O}{R_S(1 + g_{mt} r_O) + r_O} = \frac{1 + g_{mt} r_O}{R_X}$$

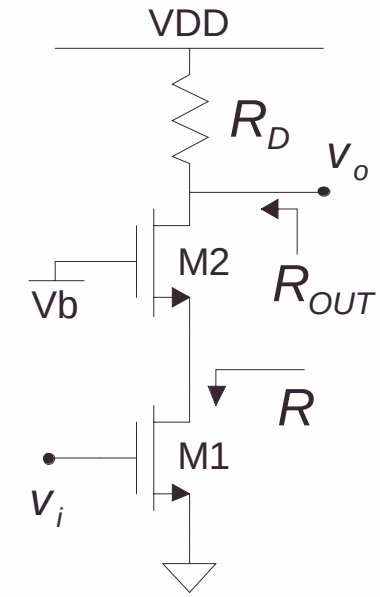


Thus: $$\frac{v_O}{v_i} = \frac{1 + g_{mt} r_O}{R_X}(R_D \,||\, R_X) = (1 + g_{mt} r_O)\frac{R_D}{R_D + R_X} = \frac{1 + g_{mt} r_O}{1 + \dfrac{R_X}{R_D}} = \frac{1 + g_{mt} r_O}{1 + \dfrac{R_S}{R_D} + \dfrac{r_O}{R_D}[1 + g_{mt} R_S]}$$

Quite often: $\boxed{\dfrac{v_O}{v_i} \approx g_{mt} R_D}$ subject to $R_S = 0, \quad r_O >> R_D, \quad g_{mt} r_O >> 1$

The CG amp is also part of the widely-used cascode amplifier stage, which comes next ..

☐ Cascode Amplifier



M1 is a CS amp
M2 is a CG amp

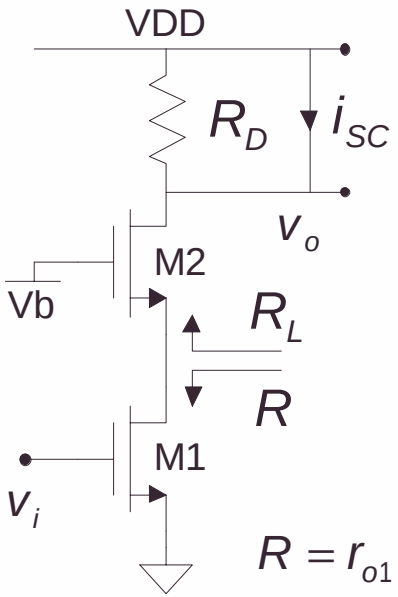
The Cascode amp ← features a higher output impedance and better separation between v_i and v_o due to the de-sensitizing effect of M2. Using a Norton source model → , we can find the gain as:

$$\frac{v_O}{v_i} = -G_M R_{OUT} = -G_M (R_D \,||\, R_X)$$

where: $R_X = R + r_{o2}[1 + g_{mt2} R]$ and ..

$$G_M = \frac{i_{SC}}{v_i} = g_{m1} \frac{r_{o1}}{r_{o1} + R_L}$$

with R and R_L as shown →



$R = r_{o1}$

$$R_L = \frac{r_{o2}}{1 + g_{mt2} r_{o2}}$$

Substituting for R and R_L :

$$\frac{v_o}{v_i} = \frac{-g_{m1} r_{o1}}{r_{o1} + \dfrac{r_{o2}}{1 + g_{mt2} r_{o2}}} \cdot R_D \,||\, \{r_{o1} + r_{o2}[1 + g_{mt2} r_{o1}]\} = \frac{-g_{m1} r_{o1}(1 + g_{mt2} r_{o2})}{r_{o1}(1 + g_{mt2} r_{o2}) + r_{o2}} \cdot R_D \,||\, \{r_{o1} + r_{o2}[1 + g_{mt2} r_{o1}]\}$$

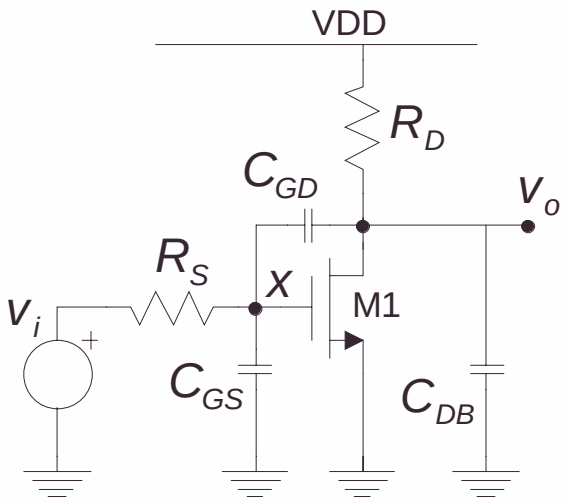
Typically: $g_{mt} r_O >> 1$ → $G_M \approx g_{m1}$ →

$$\boxed{\frac{v_o}{v_i} = -g_{m1} \cdot R_{OUT}}$$

Cascoding gives high R_{OUT} and *high gain*, provided R_D also is made high by cascoding !

MOS AMPLIFIER FREQUENCY RESPONSE

CS Amplifier Frequency Response



There are two defining equations (summing currents at circuit nodes ➔) :

$$\frac{v_X - v_i}{R_S} + V_X C_{GS} s + (v_x - v_O) C_{GD} s = 0 \qquad \text{at node } x$$

$$(v_O - v_X) C_{GD} s + g_m v_X + v_O \left(\frac{1}{R_D} + C_{DB} s \right) = 0 \qquad \text{at outout node}$$

Eliminating v_X and solving we obtain:

$$\frac{v_O}{v_i} = \frac{(C_{GD} s - g_m) R_D}{1 + Bs + As^2}$$

where:

$$B = R_S [1 + g_m R_D] C_{GD} + R_S C_{GS} + R_D (C_{GD} + C_{DB})$$

and:

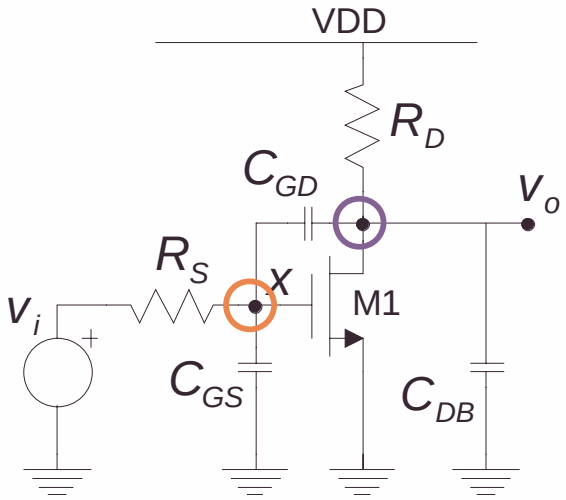
$$A = R_S R_D (C_{GS} C_{GD} + C_{GS} C_{DB} + C_{GD} C_{DB})$$

This is all rather complicated, but a computer can quickly find the roots of the vector [1 *B A*] thus informing us of the pole locations for this amplifier ➔

$$\frac{v_O}{v_i} = \frac{(C_{GD} s - g_m) R_D}{\left(1 + \frac{s}{\omega_{p1}}\right)\left(1 + \frac{s}{\omega_{p2}}\right)}$$

Notice: we also have a rhp zero at: $\omega_z = g_m / C_{GD}$

CS Amplifier Miller Approximation



If we set $s = 0$ we find the DC gain as:

$$A_V = \frac{v_O}{v_i} = \frac{(C_{GD}s - g_m)R_D}{1 + Bs + As^2} \rightarrow A_V = -g_m R_D$$

(A_V is negative and $|A_V| >> 1.0$)

At low to medium frequencies: $v_O \approx A_V v_X$

The voltage across C_{GD} becomes: $v_X - v_O = v_X(1 - A_V)$

An equivalent C from v_X to ground would be: $C_{GD}(1 - A_V) \approx |A_V| C_{GD}$

An equivalent C from v_O to ground would be: $C_{GD}\left(1 - A_V^{-1}\right) \approx C_{GD}$

This gives rise to the following pole approximations: $\omega_{p1} = \frac{1}{R_s C_{IN}}$ $\omega_{p2} = \frac{1}{R_D C_{OUT}}$

with: $C_{IN} = C_{GS} + C_{GD}[1 - A_V] = C_{GS} + C_{GD}[1 + g_m R_D]$ C_{IN} is the "Miller Capacitance"

and: $C_{OUT} = C_{DB} + C_{GD}$ ← this treats node X as if it were a "virtual zero"

These give useful insights, but are often inaccurate.

☐ CS Amplifier Split-pole Approximation

"Pole-splitting" is usually the result of a deliberate action designed to push the poles apart, in order to make: $\omega_{p2} >> \omega_{p1}$ (often by 3 or 4 decades)

Then: $$\frac{v_O}{v_i} = \frac{(C_{GD}s - g_m)R_D}{\left(1+\frac{s}{\omega_{p1}}\right)\left(1+\frac{s}{\omega_{p2}}\right)} = \frac{(C_{GD}s - g_m)R_D}{\left(1+\frac{s}{\omega_{p1}}+\frac{s}{\omega_{p2}}+\frac{s^2}{\omega_{p1}\omega_{p2}}\right)} \approx \frac{(C_{GD}s - g_m)R_D}{\left(1+\frac{s}{\omega_{p1}}+\frac{s^2}{\omega_{p1}\omega_{p2}}\right)} = \frac{(C_{GD}s - g_m)R_D}{1+Bs+As^2}$$

and: $$\omega_{p1} \approx \frac{1}{B} \approx \frac{1}{R_S[1+g_mR_D]C_{GD} + R_SC_{GS} + R_D(C_{GD}+C_{DB})}$$ the input pole

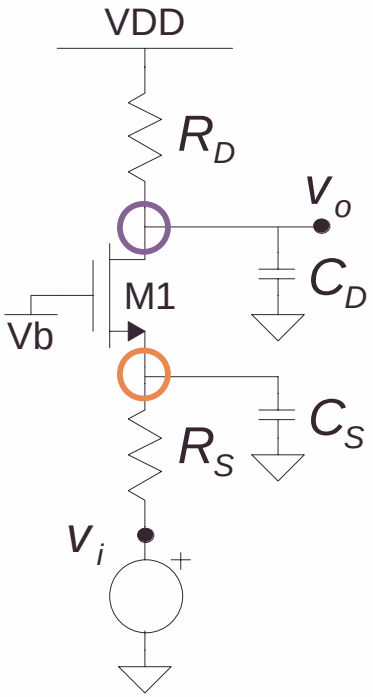
$$\omega_{p2} \approx \frac{B}{A} \approx \frac{R_S[1+g_mR_D]C_{GD} + R_SC_{GS} + R_D(C_{GD}+C_{DB})}{R_SR_D(C_{GS}C_{GD} + C_{GS}C_{DB} + C_{GD}C_{DB})}$$ the output pole

For the Miller approximations to be valid, the boxed parameters should be small enough to neglect. Only then could we say:

$$\omega_{p1} \approx \frac{1}{R_sC_{IN}} \approx \frac{1}{R_s(C_{GS} + C_{GD}[1+g_mR_D])} \qquad \omega_{p2} \approx \frac{1}{R_DC_{OUT}} = \frac{1}{R_D(C_{DB}+C_{GD})}$$

The ω_{p2} approximation can be used only if C_{GS} is large.

CG Amplifier Frequency Response



We saw that: $$\frac{v_O}{v_i} = \frac{1+g_{mt}r_O}{1+\frac{R_S}{R_D}+\frac{r_O}{R_D}\left[1+g_{mt}R_S\right]}$$

If r_O is high (meaning negligible λ effect), the low-frequency gain is approximately: $$\frac{v_O}{v_i} \approx \frac{g_{mt}R_D}{1+g_{mt}R_S}$$

The frequency response can be approximated as:

$$\frac{v_O}{v_i} = \frac{\frac{g_{mt}R_D}{1+g_{mt}R_S}}{\left(1+\frac{s}{\omega_{p1}}\right)\left(1+\frac{s}{\omega_{p2}}\right)}$$

with:

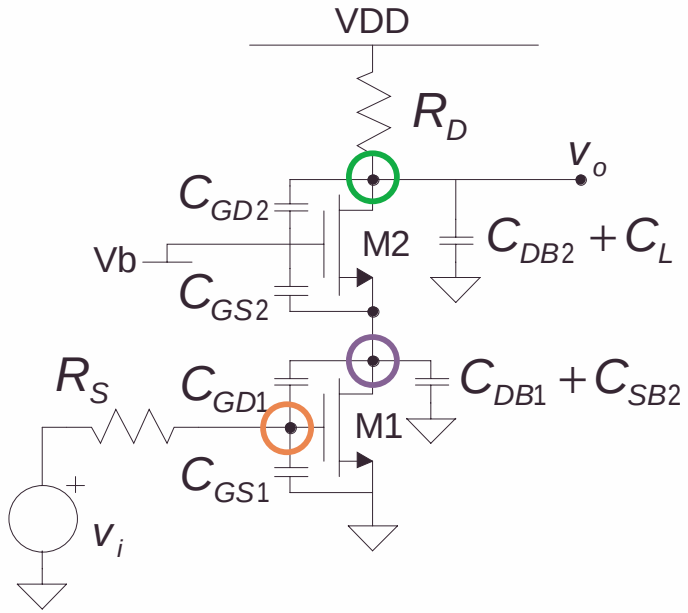
$$C_S = C_{GS} + C_{SB}$$

$$C_D = C_{GD} + C_{DB}$$

$$\omega_{p1} \approx \frac{g_{mt}+1/R_S}{C_S} \qquad \omega_{p2} \approx \frac{1}{R_D C_D}$$

This promises a wideband response (no Miller effect), but at the expense of a low input impedance. The cascode amp, which is next, avoids the low impedance penalty ..

Cascode Amplifier Frequency Response



We saw that: $\dfrac{v_o}{v_i} \approx -g_{m1} \cdot R_{OUT}$ .. at low frequency

When poles are included ..

$$\frac{v_O}{v_i} \approx \frac{-g_{m1} \cdot R_{OUT}}{\left(1+\dfrac{s}{\omega_{p1}}\right)\left(1+\dfrac{s}{\omega_{p2}}\right)\left(1+\dfrac{s}{\omega_{p3}}\right)}$$

On inspection:

○ $C_1 = C_{GS1} + C_{GD1}(1-A_V)$

○ $C_2 = C_{DB1} + C_{SB2} + C_{GS2} + C_{GD1}\left(1-A_V^{-1}\right)$

○ $C_3 = C_{GD2} + C_{DB2} + C_L$

with: $A_V = -\dfrac{g_{m1}}{g_{mt2}}$ (often close to –1. 0)

$$\omega_{p1} \approx \frac{g_{m1}+1/R_s}{C_1}$$

$$\omega_{p2} \approx \frac{g_{mt2}}{C_2}$$

$$\omega_{p3} \approx \frac{1}{R_D C_3}$$

.. where all λ effects have been neglected

CMOS BAND-GAP REFERENCES

☐ Bandgap Voltage References

A voltage reference must be *stable over temperature*, and must be *immune to supply* changes and noise sources.

The term "bandgap" refers to the gap between levels in a Silicon Energy-level diagram. The gap value is VG0 = 1.206 volts, and the voltage references to be described are of a closely similar value, with the *bandgap voltage* as the key component thereof.

Bandgap voltage references are the most widely used kind, for bipolar technologies, and for CMOS as well.

Bandgap voltage references depend on the temperature behaviour of the V_{BE} of a bipolar transistor. In CMOS there are no purpose-built bipolars, but an n-well can be made to operate as the base of a grounded pnp, as we shall see.

We will first set out the main ideas and some circuit implementations, with very little maths. Afterwards, we will show the maths in more detail.

☐ Bandgap Reference Basics

Bipolar transistors follow very well-defined rules from semiconductor physics. Some key features are:

The V_{BE} of a bipolar at constant IC has a temp coefficient of about -2.0 mV/degC.

The difference between two such V_{BE} values is directly proportional to absolute temperature. In fact, it follows a simple expression ➔

$$\Delta V_{BE} = V_{BE2} - V_{BE1} = \frac{kT}{q} Ln\left(\frac{J_2}{J_1}\right)$$

J_1 and J_2 are the current densities in the transistors, Q_1 and Q_2. If the transistors are identical, we can use the currents I_1 and I_2 in place of J_1 and J_2.

Suppose I_2 = 10 I_1. Then: $$\Delta V_{BE} = \frac{kT}{q} Ln\left(\frac{10}{1}\right) = \frac{1.38E-23(300)}{1.6E-19}(2.30) = 59.5mV$$

This is at T = 300 deg K, which is 27 degC. At 37 degC, the answer is 61.5 mV. That's a 2 mV increase over 10 deg C, so the temp coefficient is 0.2 mV per deg C.

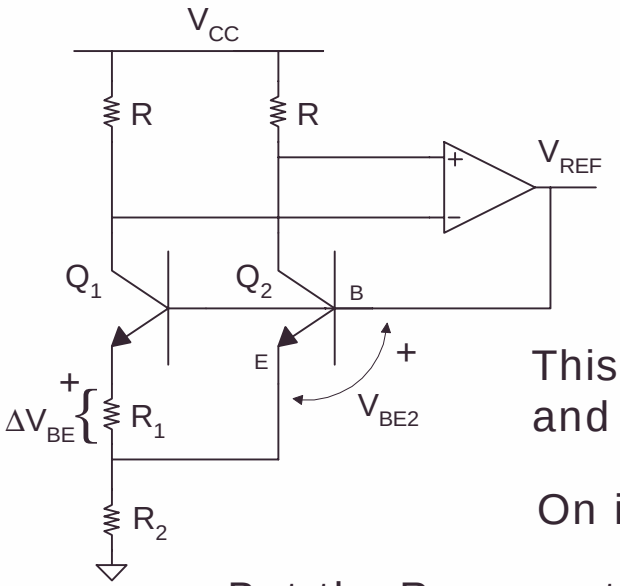
Compare this with the -2.0 mV/degC of a V_{BE} at constant current. To get a constant voltage we need to amplify a ΔV_{BE} by about 10 and add it to a V_{BE} value. That is:

$$_{REF} = V_{BE} + K \cdot \Delta V_{BE}$$

This sums up the main idea. For the current ratio that we suggested, the value of *K* is about 10. We will now look at some circuits to implement the idea.

A Bipolar Bandgap Reference

This bipolar bandgap reference is due to Paul Brokaw of Analog Devices (1974).



The op-amp's virtual zero, and the two identical resistors R, mean that currents in Q_1 and Q_2 are identical.

To make the current densities different, we will assume differing transistor emitter areas by specifying that:

$A_{e1} >> A_{e2}$ The actual area ratio may be about 10.

This causes V_{BE1} to be far less than V_{BE2}. The difference is ΔV_{BE} and it appears across R_1 , as shown on the diagram ←

On inspection: $V_{REF} = V_{BE2} + \Delta V_{R2}$

But the R_2 current is twice the R_1 current, so the R_2 drop is twice the R_1 drop, scaled by (R2/R1), ➔

$$V_{REF} = V_{BE2} + \frac{2R_2}{R1} \cdot \Delta V_{BE}$$

Based on our numeric example, if the emitter area ratio is close to 10, the value of *K* will be close to 10 as well. With a ΔV_{BE} of about 60 mV, this will result in:

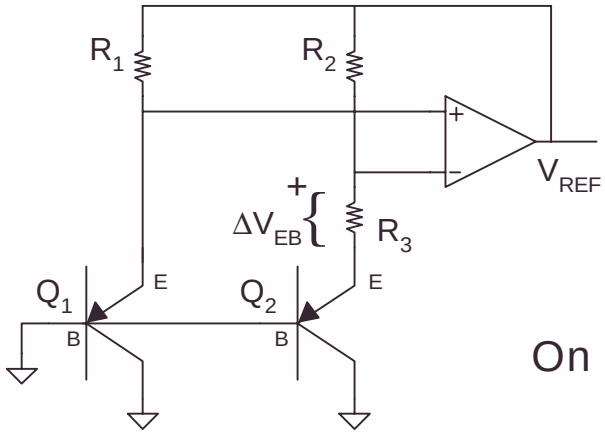
$$V_{REF} \approx V_{BE2} + 10 \cdot \Delta V_{BE} \approx 640\,mV + 10 \cdot 60\,mV = 1.24V$$

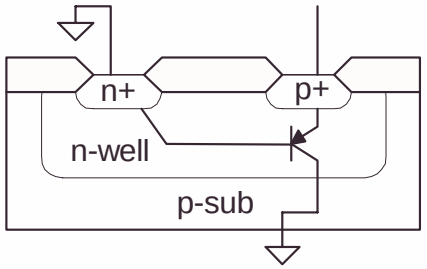
Fine-tune by laser-trimming R_1, R_2.

This is slightly higher than the bandgap voltage (V_{G0} = 1.206 V).

☐ A CMOS Bandgap Reference

In an n-well CMOS implementation, we have two pnp's with grounded base and collector, because these terminals are n-well and p-bulk regions respectively.





Structure of Q_1, Q_2 above

This pnp has the essential temperature dependence that we require.

The op-amp virtual zero forces equal voltages across R_1 and R_2 , that is: $\Delta V_{R1} = \Delta V_{R2}$

For this circuit, we'll assume that both transistors have identical areas. We'll make $R_2 >> R_1$ so as to make $I_1 >> I_2$. Typically: $I_1 / I_2 \approx 10$

On inspection: $V_{REF} = V_{EB1} + \Delta V_{R1} = V_{EB1} + \Delta V_{R2}$

Also on inspection: $\Delta V_{R2} = \Delta V_{R3} \cdot \frac{R_2}{R_3}$ because $I_{R2} = I_{R3} = I_{Q1}$.

Therefore: $V_{REF} = V_{EB1} + \Delta V_{R3} \cdot \frac{R_2}{R_3} = V_{EB1} + \Delta V_{EB} \cdot \frac{R_2}{R_3}$

This has the form that we require, with a gain value of $K = R_2/R_3$.

Here again, V_{REF} will be slightly higher than the bandgap voltage (V_{G0} = 1.206 V).

Fine-tune K by laser-trim for exact temp comp

☐ Mathematics of the bi-polar V_{BE}

The behaviour of a junction is strongly governed by semi-conductor physics, it is well defined and repeatable. For a detailed mathematical development, refer to Grey & Meyer (1993). The temperature dependence of the junction is summed up by Johns & Martin (1997) in the form:

$$_{BE} = V_{G0}\left(1-\frac{T}{T_0}\right)+V_{BE0}\frac{T}{T_0}+\frac{mkT}{q}\cdot Ln\left(\frac{T_0}{T}\right)+\frac{kT}{q}Ln\left(\frac{J_C}{J_{C0}}\right)$$

with V_{G0} = 1.206, the band-gap voltage. T_0 and J_{C0} refer to a "reference temperature", typically 300 deg K. T is the prevailing temperature, and J the prevailing current density. The factor m is a constant, with m = 2.3 approx.

This equation is all that we need to proceed. For example, it immediately accounts for the equation in which the difference between two V_{BE} values was given as:

$$\Delta V_{BE} = V_{BE2} - V_{BE1} = \frac{kT}{q}Ln\left(\frac{J_2}{J_1}\right)$$

The V_{BE} Eqn above allows us to see V_{BE} and it negative TC of about -2mV/deg C.

The ΔV_{BE} Eqn allows us to view ΔV_{BE} and its typical positive TC of about +0.2 mV/degC

☐ Mathematics of the bi-polar V_{BE} (Cont'd)

We defined the bi-polar VREF as: $_{REF} = V_{BE} + K \cdot \Delta V_{BE}$ with $K = \frac{2R_2}{R1}$

We can now substitute for V_{BE} and ΔV_{BE} to find V_{REF}. We can then set: $\frac{\delta V_{REF}}{\delta T} = 0$

.. and solve for the value of *K* which makes this possible. We can then backsubstitute the *K* value to get our final V_{REF} expression, one which has a TC of zero at the nominated temperature.

All this leads to the same circuits and procedures that we described. But it also predicts the value of *K* that is "best" for a given design.

CMOS LOW VOLTAGE TOPICS

☐ Low Voltage CMOS

Smaller device sizes allow denser circuitry (more functionality per mm^2), and this favours lower voltages as well.

Lower voltages reduce power on a square-law basis. There is strong motivation for this also due to increasing use of battery-powered equipment, such as mobile phones and portable computers.

Lower voltages make analogue design more difficult.

Lower currents increase the noise as well.

We will look at some design issues that are affected by low-voltage considerations.

Under low voltage operation, the values of V_{GS} and V_{DS} are critical and must be kept as low as possible, in order to conserve the available *headroom*.

The V_{GS} has two components:

$V_{GS} = V_T + V_e$ = Threshold + Effective value (also called “overdrive voltage”)

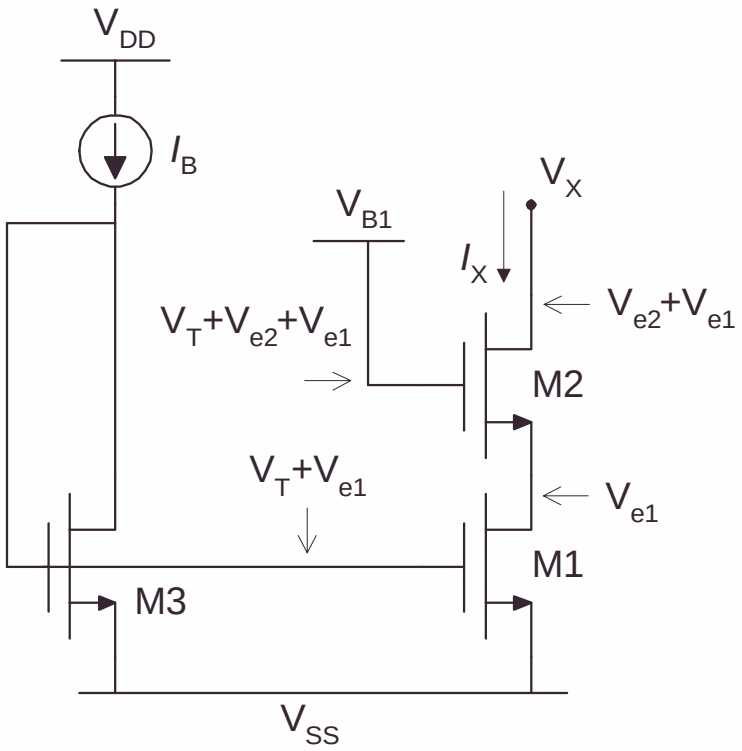
V_T is fixed (with perhaps 0.1V tolerance). Several factors make lower V_T impractical.

Typical values are $V_T = 0.7$, $V_e = 0.3$. (We try to keep V_e as low as possible).

☐ Low Voltage Cascode Current Mirror

The normal cascode current mirror uses up ($V_T+V_e+V_e$) from the available headroom.

We now show a circuit which reduces this total to (V_e+V_e), or about 0.6 volts



For a given MOSFET, Drain voltage V_D can fall below gate voltage V_G by as much as V_T before the MOSFET goes out of saturation (into the NON-SAT or the *triode* region).

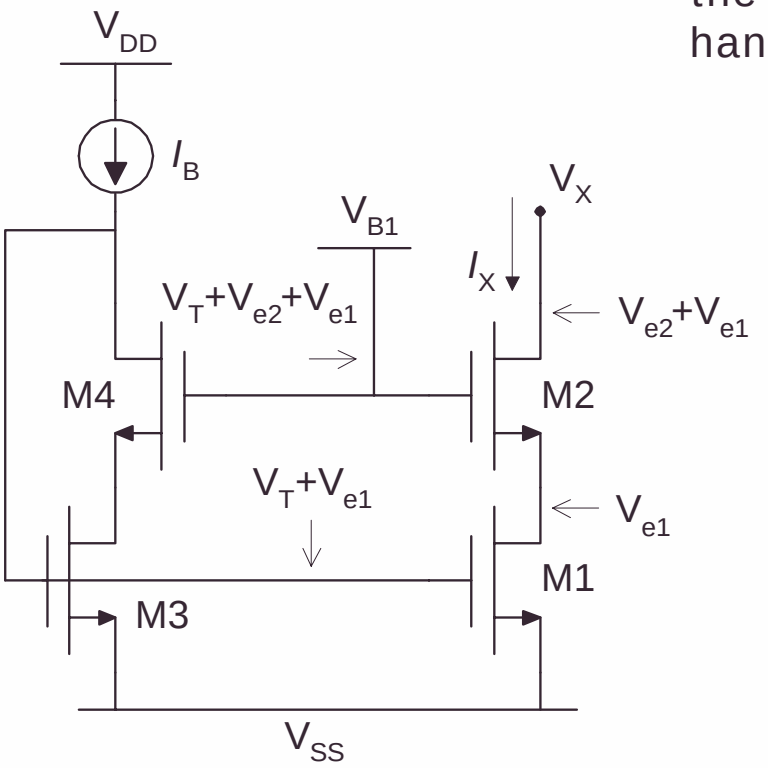
The voltages shown are the minimum values that place M1 and M2 at the edge of saturation. They are based on the relationship:

$$_{GS} = V_T + V_e = V_T + \sqrt{\frac{2I}{\beta'(W/L)}}$$

If M1 or M2 enters the triode region, the output impedance falls off rapidly. To avoid this, in practice, we must allow a small margin for safety.

The conclusion is that the current sink terminal voltage V_X can go as low as $V_{e1}+V_{e2}$ before its impedance begins to collapse.

Modified Low Voltage Current Mirror

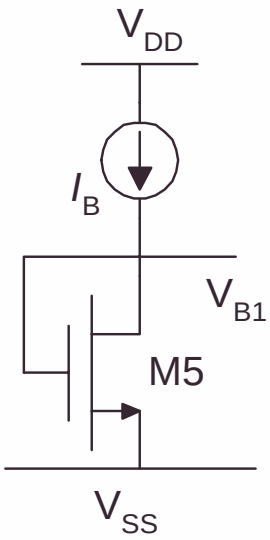


In practice, the two currents will differ slightly because of the differing drain voltages of M1 and M3. There's a handy remedy for this, which we show now.

We'll assume, for simplicity, that all transistors are identical, making V_e values identical too.

Here we've added another identical transistor, M4. It does'nt change the action of the diode-connected M3, but it does ensure that the drains of M1 and M3 are at the same potential. This, in turn, ensures that $I_X = I_B$.

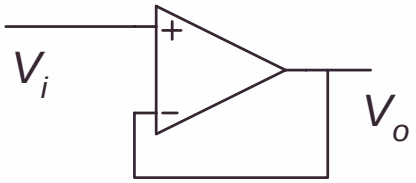
The bias voltage V_{B1} must be provided too. One option is to use another diode-connected transistor M5 with a current source of the same value I_B that we used previously ➔



We now require: $V_{B1} = V_T + 2V_e$

For the same bias current, and if M1 to M4 are identical, this calls for a $(W/L)_5$ that is four times smaller than that of the other transistors.

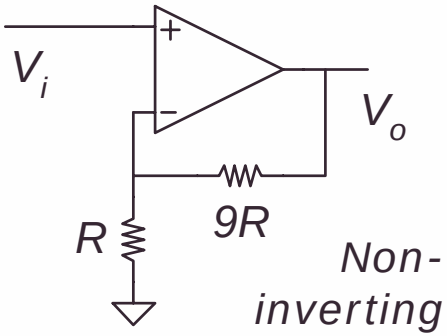
☐ Rail To Rail Operation



Unity-gain

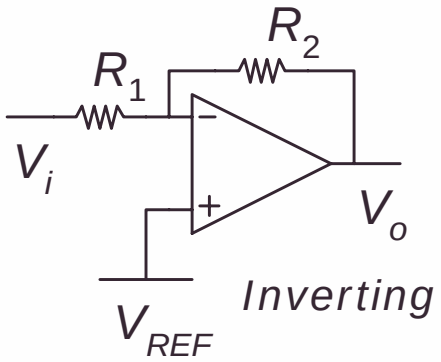
Many circuits require "rail-to-rail" operation, or something close to that.

This unity-gain buffer ← is a good example. The output must track the input over as wide a range as possible. That means designing op-amp circuits with input stages that go rail-to-rail, and output stages that go rail-to-rail too.



Non-inverting

We do'nt always need rail-to-rail working at the input. For example, this amplifier ← has a non-inverting gain of 10, so the input signal range is only one-tenth of the output signal range. That simplifies the input stage design considerably.

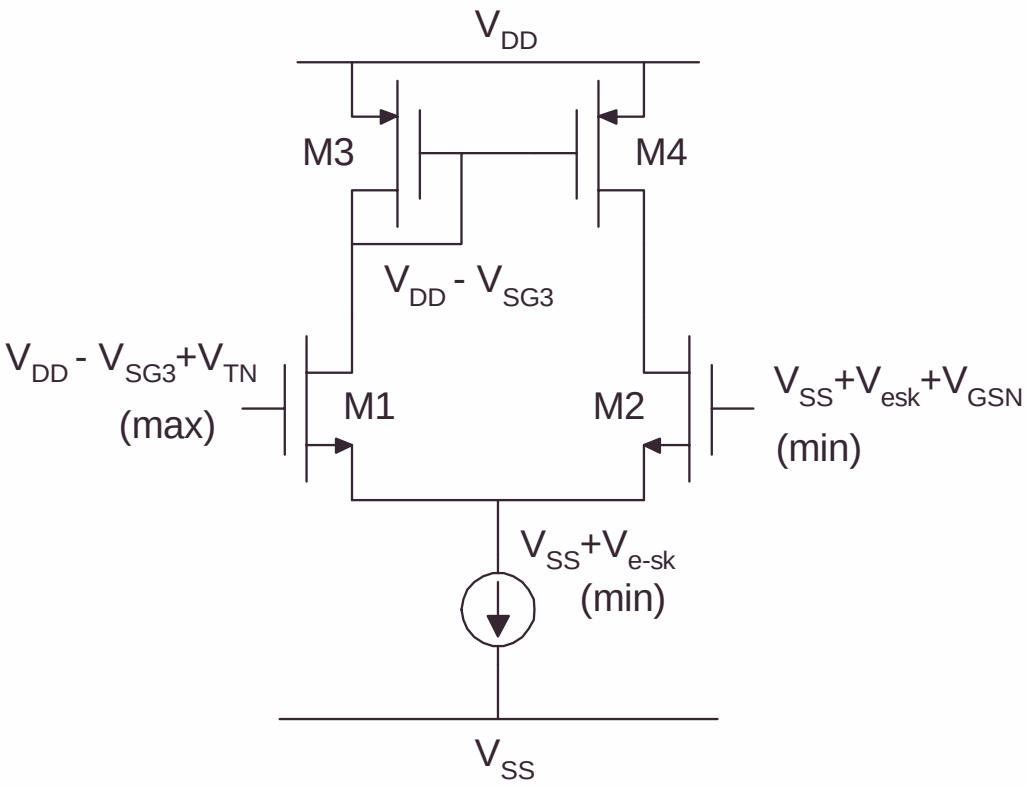


Inverting

The inverting amplifier is even better. Here ← the input signal stays close to V_{REF} at all times, irrespective if the closed-loop gain value. The typical V_{REF} is mid-way between the power supply values.

We will look at the implications of "rail-to-rail" operation for input stages, and also for output stages.

Differential Input CM Range



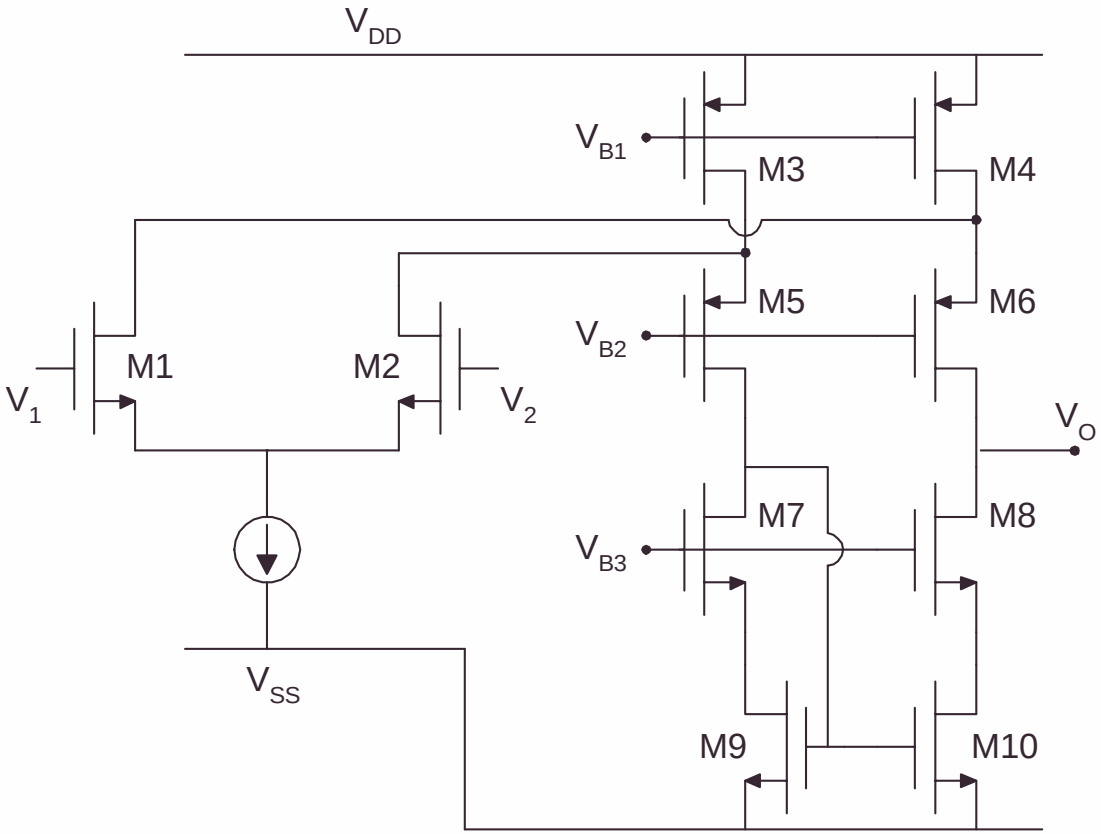
This circuit gives some insights into input common-mode range limitations.

The upper limit (max) is set by the current mirror (M3,M4) to about one V_e value below V_{DD}.

The lower limit (min) is set by the current sink and by (M1,M2) to be about $V_{TN}+V_e+V_{e\text{-}sk}$ above V_{SS}.

All this falls far short of rail-to-rail operation ..

Differential Folded Cascode

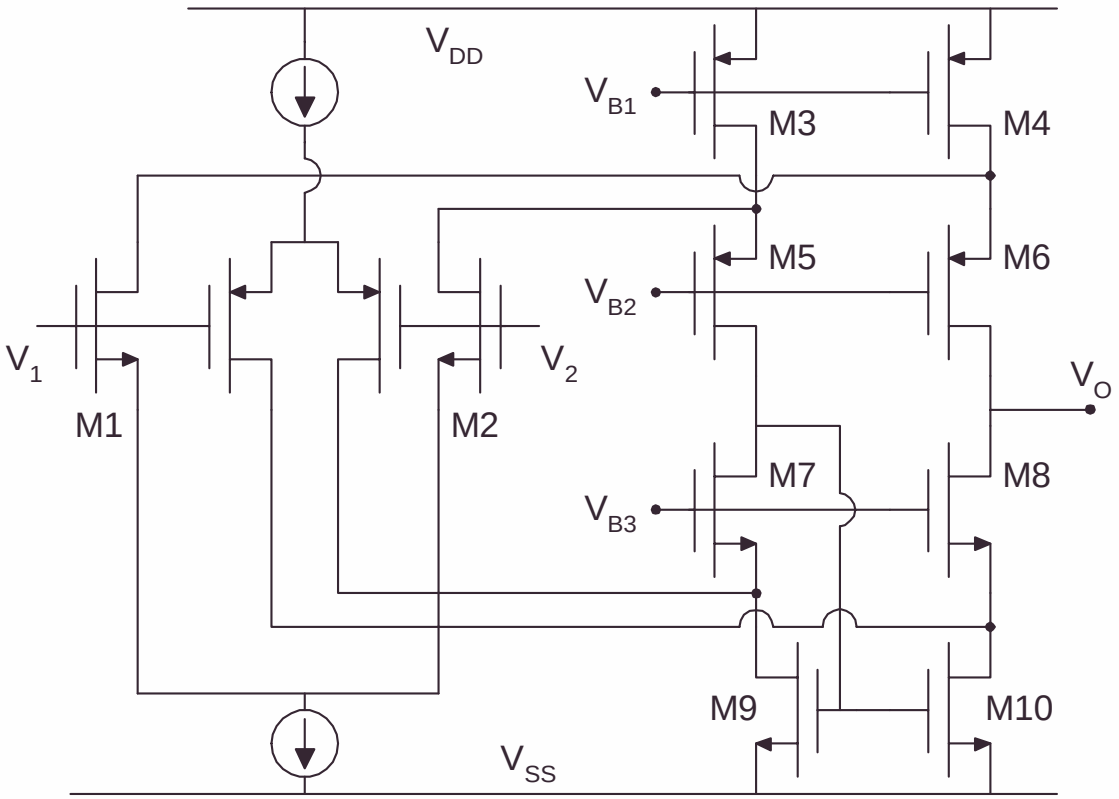


In this folded-cascode circuit, the drains of M3 and M4 can go to within V_e of the V_{DD} rail. This allows the inputs V_1 and V_2 to go all the way up to V_{DD}, a marked improvement.

The differential currents from (M1, M2) add to the fixed currents from (M3, M4) to form differential currents in (M5, M6). These are mirrored by the cascode mirror comprising (M7,M8,M9,M10).

The input terminals can operate up to the V_{DD} rail, but not down to the V_{SS} rail. Only a PMOS differential pair could meet this latter requirement.

☐ Complementary Differential Inputs



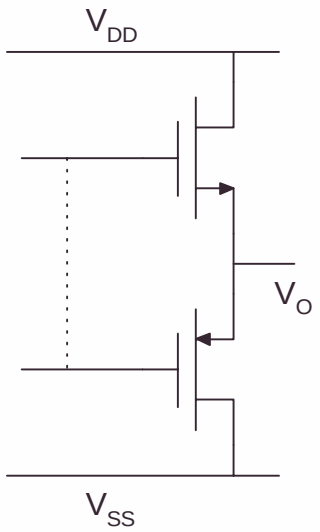
In order to have true rail-to-rail operation at the input, it is necessary to have both an NMOS and a PMOS differential pair working in parallel, as this diagram illustrates.

When the common-mode input is close to V_{DD}, the NMOS differential pair still works correctly, but the PMOS differential pair is cut off.

When the common-mode input is close to V_{SS}, the PMOS differential pair still works correctly, but the NMOS differential pair is cut off.

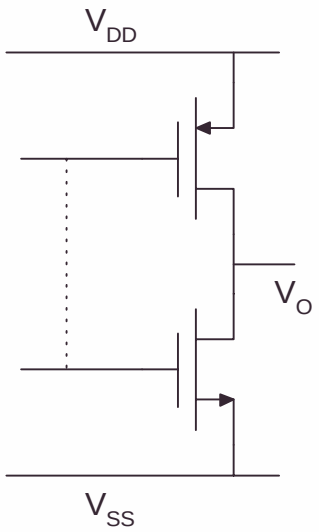
At intermediate values of common-mode input voltage, both differential pairs work together. One drawback from this is that the effective transconductance changes a lot over the range, and is a maximum when both pairs work together. There are ways in which to counter this problem, and to arrive at a more uniform transconductance.

☐ Rail To Rail Output Stages



This is the normal kind of output stage, when rail-to-rail working is not a requirement. It gives low output impedance, as we would prefer. But it is limited in its output voltage swing, in that the distance from V_O to either rail is at least a full V_{GS} value, that is, the sum of a V_T and a V_e.

This would be called a "common-drain" stage.

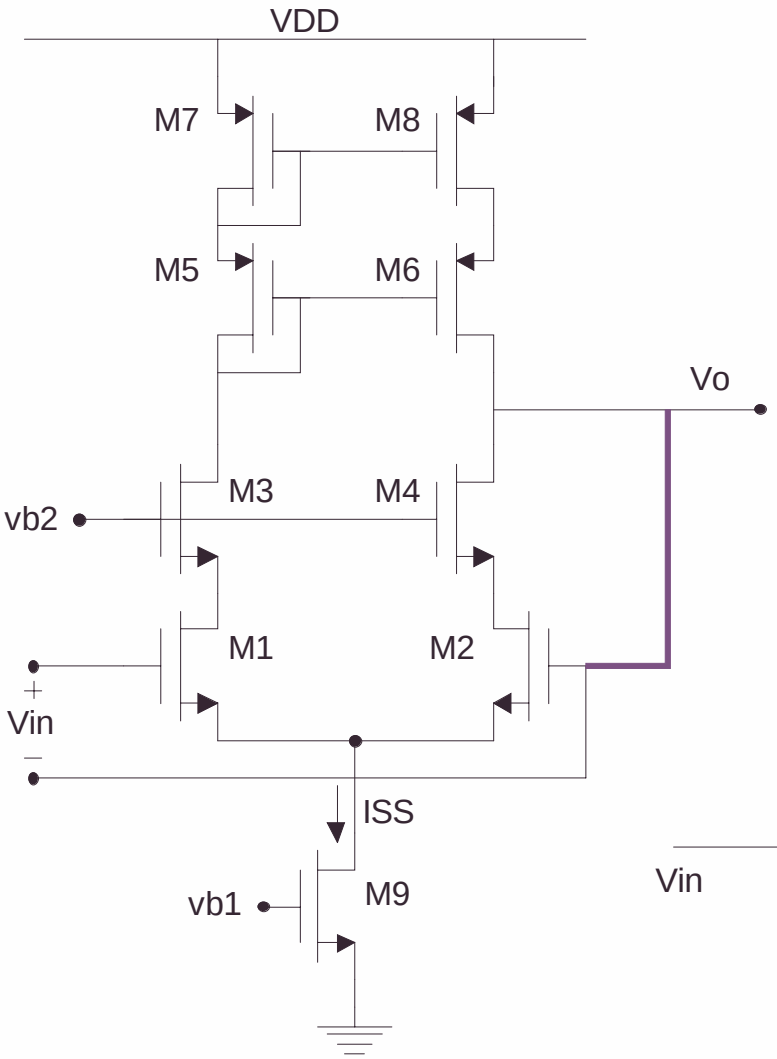


This output stage is basically a current-output stage, so its impedance characteristics are very different. But the output V_O can now go to within a V_e value of either rail. This is only a few tenths of a volt, or less if suitably large transistors are in use.

This is the popular output stage for low-voltage operation. It would be called a "common-source" stage.

CMOS OPERATIONAL AMPLIFIERS

☐ Telescopic Op-Amp : single-ended output

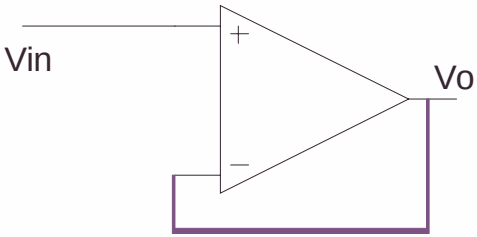


Bias voltages and device sizing must place devices in SAT region, but close to the NON-SAT boundary.

Output swing is limited to (VDD – five Ve values – one V_T value), where Ve = V_{GS} – V_T , but Ve values will not all be the same.

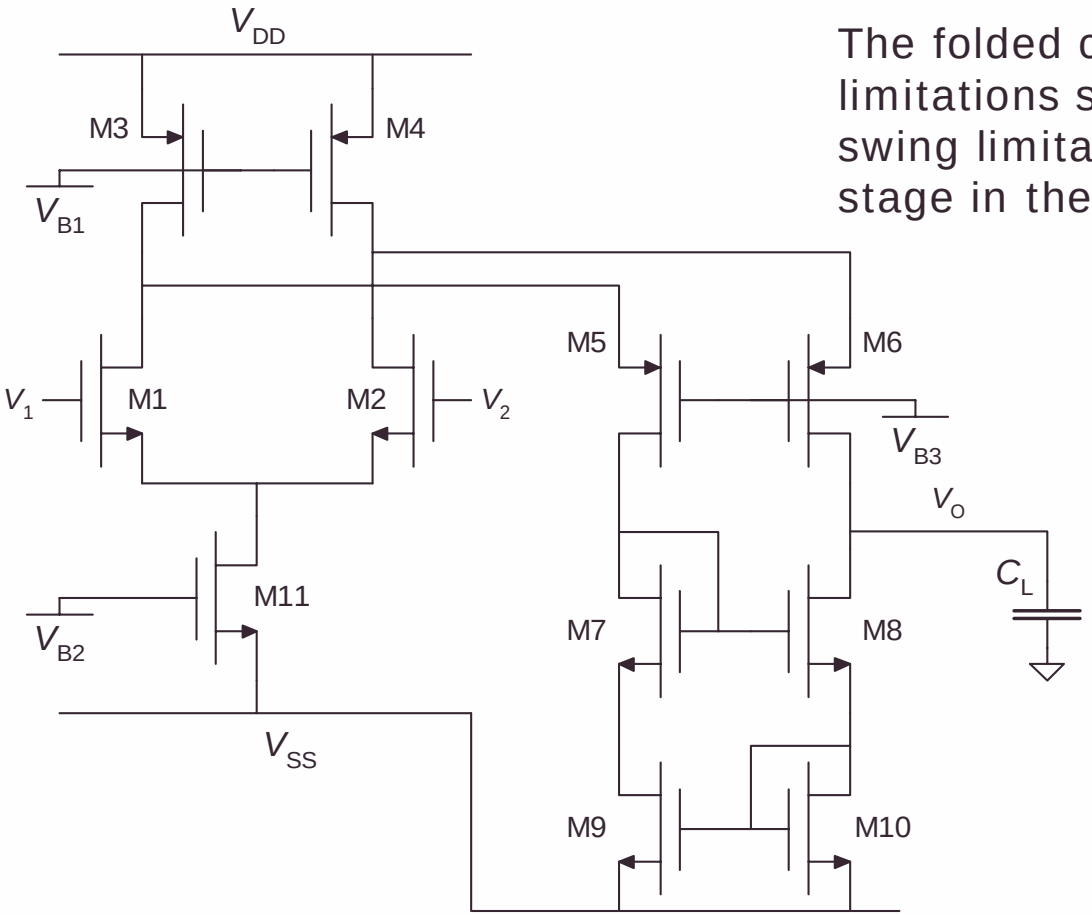
The open-loop gain at low frequency will be approximately:

$$A_{v0} = g_{mN} R_O \approx g_{mN} \left[g_{mN} r_{oN}^{\,2} \,||\, g_{mP} r_{oP}^{\,2} \right]$$



In unity-gain mode, Vin = Vo, but the CM range is severely limited. See example ↓.

Folded Cascode Op-Amp

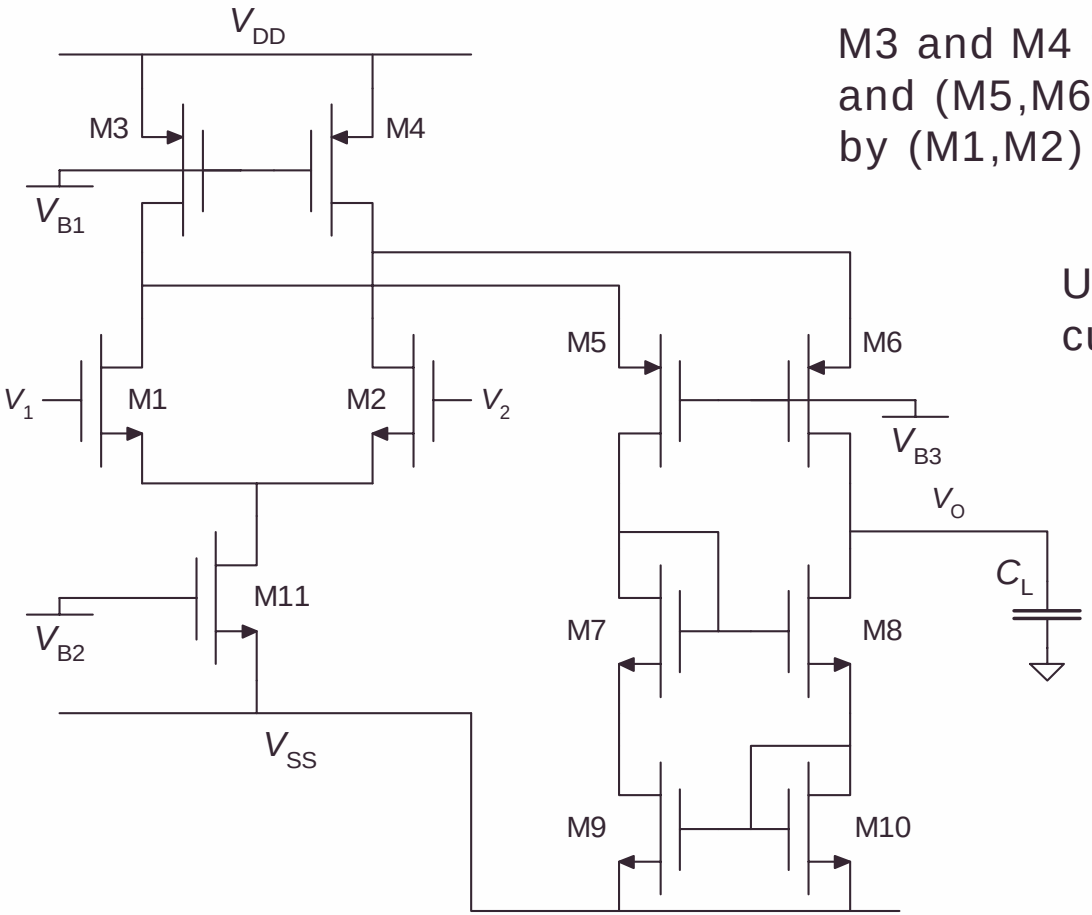


The folded cascode alleviates input CM swing limitations somewhat. It also avoids the CM swing limitation of the single-ended telescopic stage in the unity-gain feedback mode.

In common with other single-stage op-amps, the folded cascode has a single output node Vo at high impedance. Looking downward from Vo, this is provided by M8 and M10, with feedback via M7 and M9. These four transistors constitute a Wilson current mirror.

Looking upward from Vo, the cascode devices are M6 and M4. The net effect is a very high impedance at the Vo node, as is essential for high voltage gain.

Folded Cascode Op-Amp (cont'd)

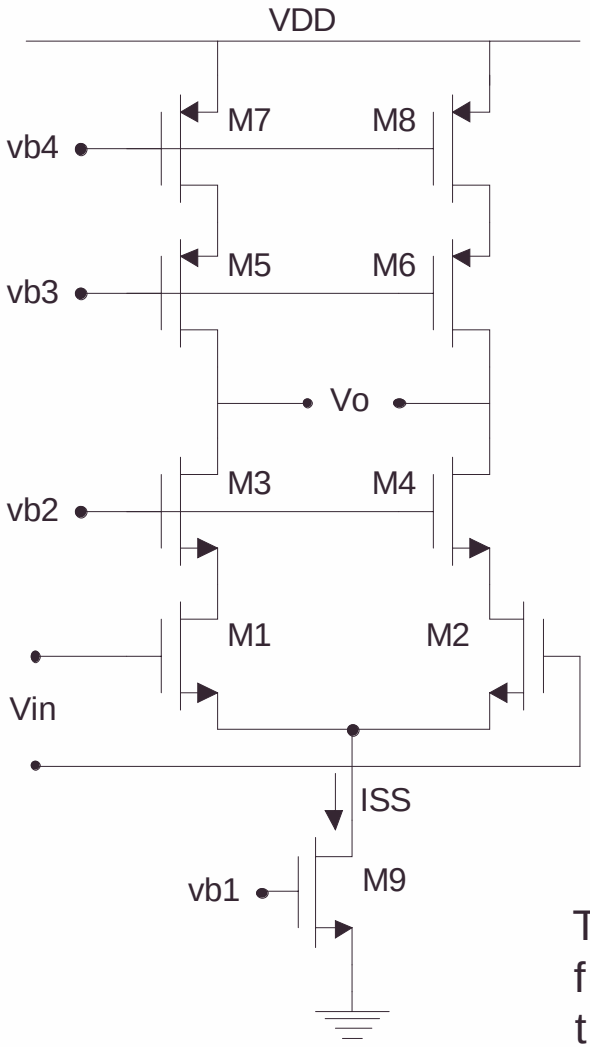


M3 and M4 provide bias current for both (M1,M2) and (M5,M6). V_{B2} decides the bias current taken by (M1,M2) and the remainder goes to (M5,M6).

Under balance conditions, the net current to C_L is zero.

Suppose we apply a small positive input (v_1 - v_2). This generates a $+\Delta I$ in M1 and a corresponding $-\Delta I$ in M2. Because I_{D3} and I_{D4} are fixed, this adds $+\Delta I$ to M6, and a corresponding $-\Delta I$ to M5. The $+\Delta I$ of M6 goes directly to C_L, and the $-\Delta I$ of M5 is mirrored by (M7,M8,M9,M10). The net result is to supply $+2\Delta I$ to C_L.

Telescopic Op-Amp : differential output



Bias voltages and device sizing must place devices in SAT region, but close to the NON-SAT boundary.

The differential output *doubles* the available output swing.

Output swing is limited to 2(VDD – five Ve values), where Ve = $V_{GS} - V_T$, but Ve values will not all be the same.

The open-loop gain at low frequency will be approximately:

$$A_{v0} = g_{mN} R_O \approx g_{mN} \left[g_{mN} r_{oN}^{\,2} \,||\, g_{mP} r_{oP}^{\,2} \right]$$

This op-amp can achieve higher gain and bandwidth than the folded cascode, but it requires a CM feedback circuit to set the output CM level appropriately.

☐ Single-stage Op-Amp frequency response

Because the differential input voltage is divided between M1 and M2, the overall stage transconductance is simply: $G_m = g_{m1} = g_{m2}$

Single-stage op-amps generally feature a single very high impedance node, and one dominant pole, with other poles removed to high frequencies (well above crossover). The result is a single-pole roll-off dominated by the load C_L and by the output node impedance R_O. The dynamics are simply stated as:

$$A_{v0} = G_m R_O \qquad \text{and} \qquad \frac{V_O}{V_{IN(diff)}} = \frac{A_{vo}}{1 - \frac{s}{s_{p1}}} = \frac{A_{vo}}{1 + sR_O C_L}$$

with $s_{p1} = \frac{-1}{R_O C_L}$ and $f_{p1} = -s_{p1}/2\pi$

Because s_{p1} is a low-frequency pole, the gain over most frequencies is, approximately: $\frac{V_O}{V_{IN(diff)}} \approx \frac{A_{vo}}{sR_O C_L} = \frac{G_m}{sC_L}$

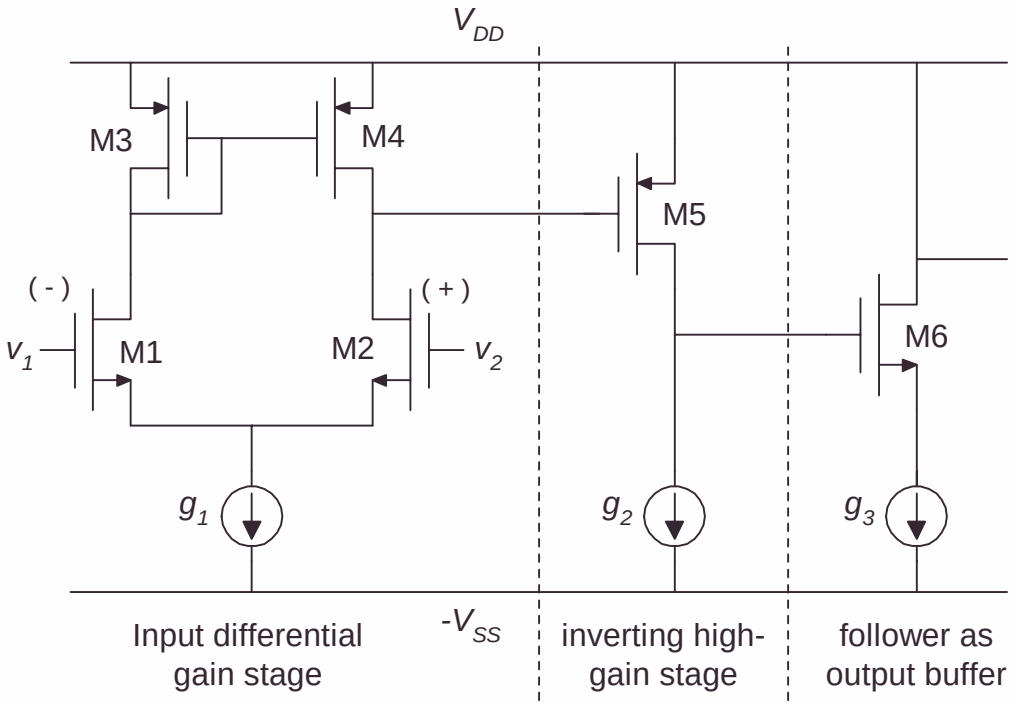
For unity gain, $|A_V(s)| = 1$, occurring at $s = j\,\omega_u$, we obtain: $\omega_u = G_m / C_L$

$G_m = \sqrt{2\beta'(W/L)I_D}$ Because G_m decides the bandwidth, input devices must have a (W/L) value and a bias current sufficient to meet the target ω_u value.

A sufficiently large C_L is the only requirement for closed-loop stability.

☐ Two-stage Op-Amp Overview

Two-stage op-amps try to boost gain by having a second gain stage. This may be instead of cascoding (where supply voltage will not allow it), or in addition to cascoding (to give further gain enhancement). The main drawback is the feedback stability concerns which arise in a 2-pole system.

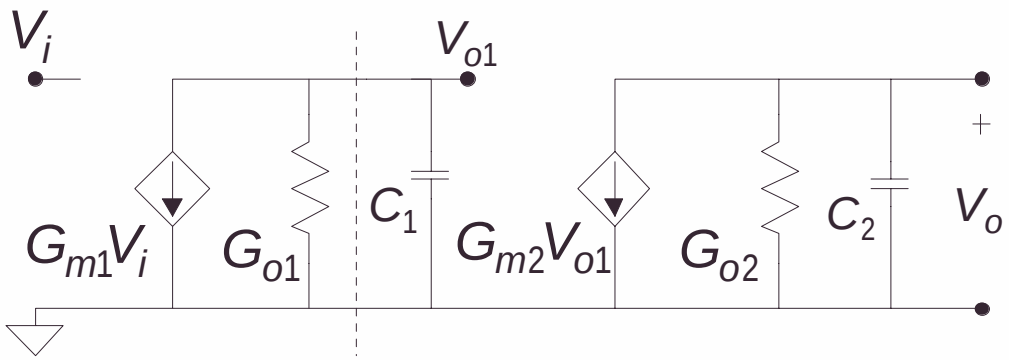


This circuit is symbolic of two gain stages, with an unity-gain buffer at the output. The buffer reduces output impedance without much effect on the frequency response.

If the op-amp loading is low, the buffer can be omitted. In such cases, the output parameter is a *current*, and low output impedance is achieved by feedback alone. An op-amp without an output buffer might be called an Operational Transconductance Amplifier, or OTA for short.

Two-stage small-signal model

A two-stage op-amp has a small-signal model as shown.



V_i represents the differential input, and higher-order poles are ignored. The two poles included are the output poles from stage 1 and stage 2. The pole frequencies are at $f_{p1} = -s_{p1}/2\pi$ and at $f_{p2} = -s_{p2}/2\pi$.

For stage 1: $a_{o1} = G_{m1}/G_{o1}$ $s_{p1} = -\dfrac{G_{o1}}{C_1}$

For stage 2: $a_{o2} = G_{m2}/G_{o2}$ $s_{p2} = -\dfrac{G_{o2}}{C_2}$

Then :

$$\frac{v_{o1}}{v_i} = \frac{-a_{o1}}{\left(1-\dfrac{s}{s_{p1}}\right)} = \frac{-(G_{m1}/G_{o1})}{\left(1+\dfrac{sC_1}{G_{o1}}\right)}$$

$$\frac{v_{o2}}{v_{o1}} = \frac{-a_{o2}}{\left(1-\dfrac{s}{s_{p2}}\right)} = \frac{-(G_{m2}/G_{o2})}{\left(1+\dfrac{sC_2}{G_{o2}}\right)}$$

For both stages in cascade:

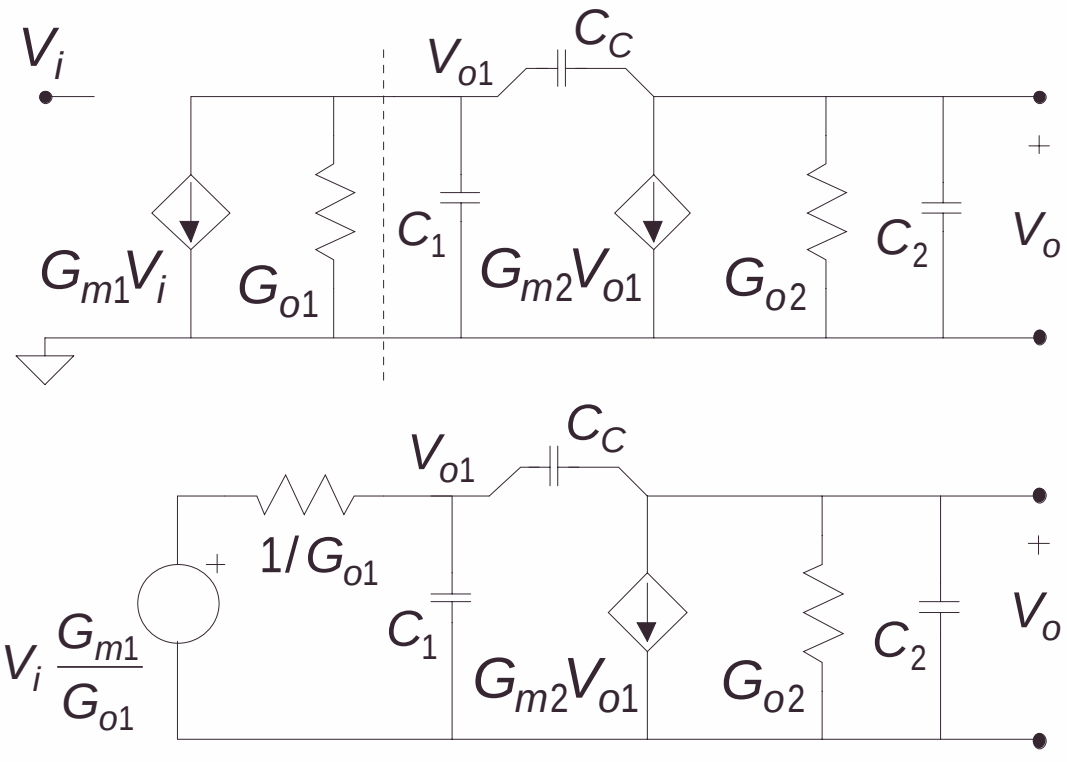
$$\frac{v_o}{v_i} = \frac{A_{vo}}{\left(1-\dfrac{s}{s_{p1}}\right)\left(1-\dfrac{s}{s_{p2}}\right)}$$

where: $A_{vo} = a_{01}a_{02} = \dfrac{G_{m1}}{G_{o1}} \cdot \dfrac{G_{m2}}{G_{o2}}$

(the DC open-loop gain)

Two-stage Pole-Splitting

In general, the op-amp poles are not widely separated. This is problematic because, under feedback, the loop gain phase-angle approaches -360 deg (-180 + two amplifier lags) while the gain is still high, and this leads to unstable behaviour.

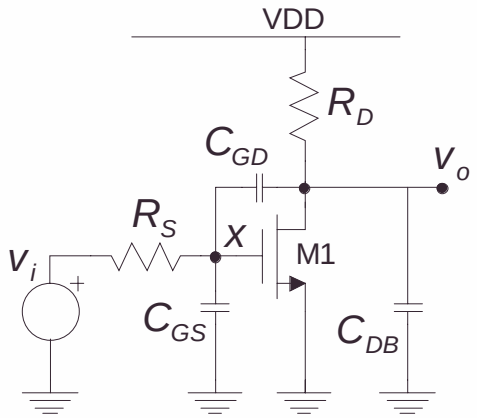


The common solution is to introduce a compensating capacitor C_C, as shown ←. We will show that this behaves as *a pole-splitting capacitor*. It causes the first pole to move down to a far lower frequency, while the second pole goes up to a far higher frequency. We will then have *a dominant pole* condition, and this is desirable for purposes of stability.

C_C complicates the analysis, but here ← we show a Thevenin equivalent for the left side of the dotted line boundary.

Its still complicated, but now it has the same model structure as the CS apmlifier that we analyzed before. That can save us a lot of effort ...

CS Amplifier Analysis Re-visited



For this circuit ←, we've already shown that →

$$\frac{v_O}{v_i} = \frac{(C_{GD}s - g_m)R_D}{1 + Bs + As^2}$$

where:

$$B = R_S[1 + g_m R_D]C_{GD} + R_S C_{GS} + R_D(C_{GD} + C_{DB})$$

and:

$$A = R_S R_D(C_{GS}C_{GD} + C_{GS}C_{DB} + C_{GD}C_{DB})$$

Our simplified two-stage op-amp is topologically the same ↙. We can re-use the results making the substitutions shown ←, to obtain:

$$\frac{v_O}{v_i} = \frac{G_{m1}}{G_{o1}} \cdot \frac{(C_C s - G_{m2})/G_{o2}}{1 + Bs + As^2}$$

$$_i \rightarrow V_i \cdot \frac{g_{m1}}{g_{o1}}$$

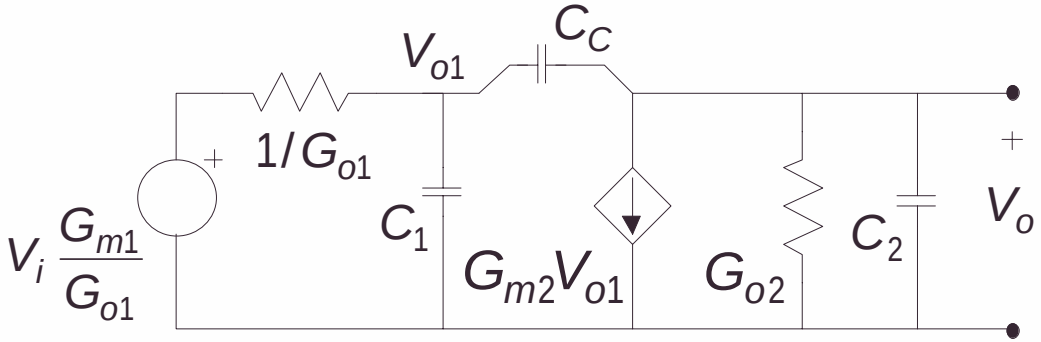
$g_m \rightarrow g_{m2}$ $\quad C_{GS} \rightarrow C_1$

$R_s \rightarrow 1/g_{o1}$ $\quad C_{GD} \rightarrow C_C$

$R_D \rightarrow 1/g_{o2}$ $\quad C_{DB} \rightarrow C_2$

where: $B = [1 + G_{m2}/G_{o2}]C_C/G_{o1} + C_1/G_{o1} + (C_C + C_2)/G_{o2}$

and: $A = (C_1C_C + C_1C_2 + C_CC_2)/G_{o1}G_{o2}$



In this model, C_1 is the inter-stage node capacitance. C_2 is generally much larger as it includes the external loading C_L. The G_m and G_o values are seen from inspection of the circuit.

Pole-Splitting Model Analysis

After pole-splitting, the poles will be far apart, such that: $s_{p2} >> s_{p1}$

By assuming that this is the case, we can simplify our analysis, because then ...

$$\left(1-\frac{s}{s_{p1}}\right)\left(1-\frac{s}{s_{p2}}\right)=\left(1-\frac{s}{s_{p1}}-\frac{s}{s_{p2}}+\frac{s^2}{s_{p1}s_{p2}}\right)\approx\left(1-\frac{s}{s_{p1}}+\frac{s^2}{s_{p1}s_{p2}}\right)=1+Bs+As^2$$

We see that:

$$s_{p1}\approx -1/B=\frac{-G_{o1}G_{o2}}{C_C(G_{m2}+G_{o2})+C_1G_{o2}+(C_2+C_C)G_{o1}} \qquad s_{p1}\approx\frac{-G_{o1}G_{o2}}{C_CG_{m2}}=\frac{-G_{o1}}{a_{02}C_C}$$

$$s_{p2}\approx\frac{1}{As_{p1}}\approx-\frac{C_C(G_{m2}+G_{o2})+C_1G_{o2}+(C_2+C_C)G_{o1}}{(C_1C_C+C_1C_2+C_CC_2)} \qquad s_{p2}\approx-\frac{C_CG_{m2}}{(C_1C_C+C_1C_2+C_CC_2)}$$

The approximations on the right often suffice because, normally: $G_m >> G_o$

If we can also assume that C_CC_2 >> either C_CC_1 or C_1C_2, then: $s_{p2}\approx-\frac{G_{m2}}{C_2}$

Before C_C was introduced, we had: $s_{p1}=-\frac{G_{o1}}{C_1}$ $s_{p2}=-\frac{G_{o2}}{C_2}$

The comparison is quite revealing ..

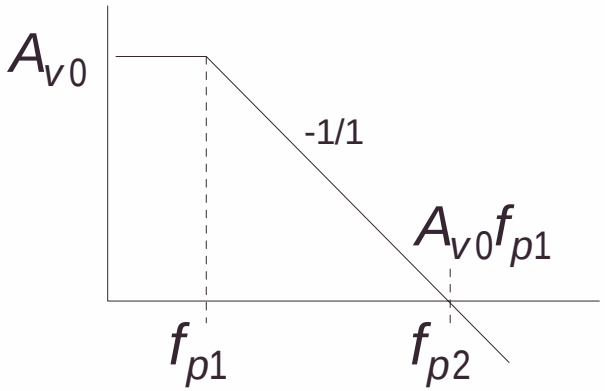
The Consequences of Pole-Splitting

Before C_C was introduced	After C_C was introduced	
$s_{p1} = -\frac{G_{o1}}{C_1}$	$s_{p1} \approx \frac{-G_{o1}}{a_{02}C_C}$	We will generally find that : $C_2 >> C_C >> C_1$
		a_{02} is the stage-2 gain, generally large, and $C_C >> C_1$ usually. The pole s_{p1} moves down – perhaps by two decades or so.
$s_{p2} = -\frac{G_{o2}}{C_2}$	$s_{p2} \approx -\frac{G_{m2}}{C_2}$	Because $G_{m2} >> G_{o2}$, the pole s_{p2} moves up – it may move as much as two decades.

The downward movement of pole-1 is the result of a large deliberate Miller effect.
The upward movement of pole-2 is the result of capacitive feedback around stage-2.

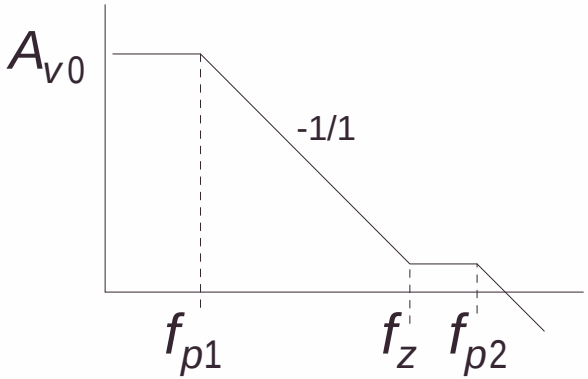
The poles are now far apart, making s_{p1} *a dominant pole*.

The idea is that the second pole f_{p2} should occur at or beyond crossover (i.e. the unity-gain frequency). If f_{p2} occurs at crossover, the total phase lag becomes –180-45 = -225 deg, wheich leaves a phase margin of 45 degrees ➔



The Influence of the Zero

We generally need a phase margin of 60 deg or more, and the *zero* of the transfer function is an added complication ➔. It is a *right-half-plane* zero, it *halts* the falling gain slope while sending the phase even *more negative*. Both these factors conspire to reduce the phase margin. This makes it important that f_z be at least a decade above the *unity gain frequency or the crossover frequency* (often called f_u).



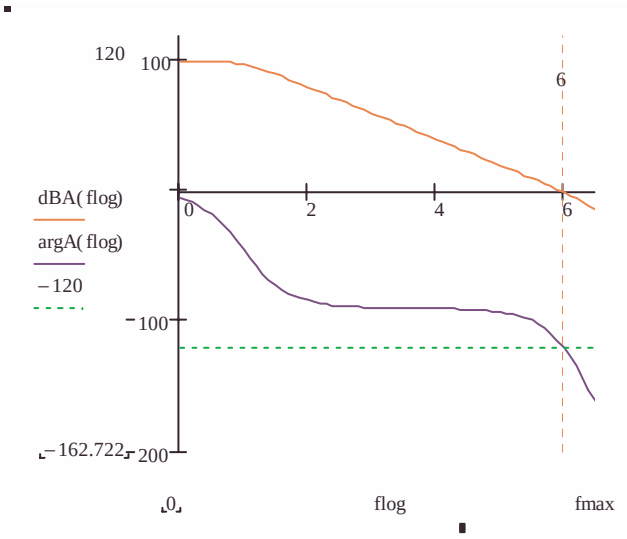
Allen & Holberg (Ref: AH, Ch 6) point out that if $f_z = 10\ f_u$, where f_u is the unity-gain (crossover) frequency, then we must have $f_{p2} > 2.2\ f_u$ to get a 60 deg phase margin.

$A_{vo} = 10^5$. $f_u = 10^6$. $f_{p2} = 2.2\ f_u$. $f_z = 10 f_u$.

This is easily checked using the response function shown here ➔

$$A_v(f) = A_{v0} \frac{\left(1 - j\frac{f}{f_z}\right)}{\left(1 + j\frac{f}{f_u / A_{vo}}\right)\left(1 + j\frac{f}{f_{p2}}\right)}$$

For the same pole locations, the simulation also shows how phase margin changes if f_z moves up or down :

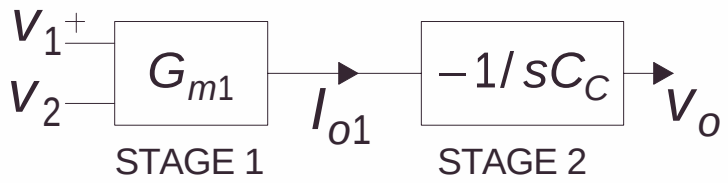


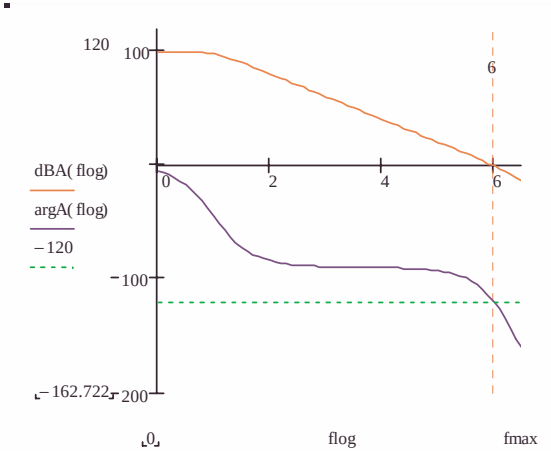
$f_z = 100 f_u$ ➔ 65 deg margin (higher f_z has little effect)
$f_z = 1 f_u$ ➔ 40 deg margin

☐ Simplified 2-stage Design Method

With a dominant pole, we can approximate the response as being simply a single-pole roll-off toward f_u , one in which stage-2 acts like a integrator with integrating capacitor C_C , while current injection from stage-1 sets the gain level. We model it like this …

$$I_{o1} = G_{m1}(v_1 - v_2) \qquad v_o = \frac{-I_{o1}}{sC_c} \quad \rightarrow \quad \frac{v_o}{v_1 - v_2} = \frac{-G_{m1}}{sC_c}$$





To find f_u, we set $s = j\omega_u = j2\pi f_u$, then set |gain| to unity, and solve for f_u to find : $\omega_u = 2\pi f_u = \frac{G_{m1}}{C_c}$

We take this approach because f_u is generally specified as a design target. For a given f_u and DC gain A_{v0} , the first pole will occur at f_u / A_{v0} .

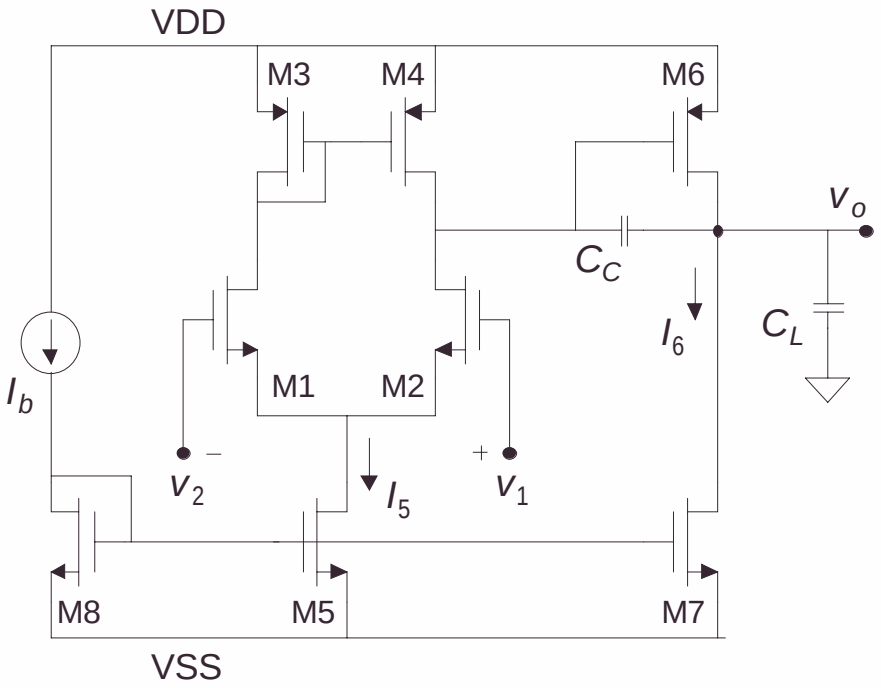
For a 60 deg phase margin, we must then make sure that : $f_{p2} \geq 2.2\, f_u$ and $f_z \geq 10\, f_u$

We also know that: $\omega_z = \frac{G_{m2}}{C_c}$ and $\omega_{p2} \approx \frac{G_{m2}}{C_2}$

By meeting these conditions, we get an op-amp that is stable even for unity feedback.

Two-stage Op-Amp and Specification

The typical op-amp performance specification might include the following :



Load Capacitance C_L.

Crossover f_u (= Gain-Bandwidth GB)

Slew Rate *SR*.

DC Gain A_{v0}.

Power dissipation *Pd*.

Input CM range

Output voltage max and min

For the simple topology shown, the slew rate is : $SR = \frac{I_5}{C_C}$

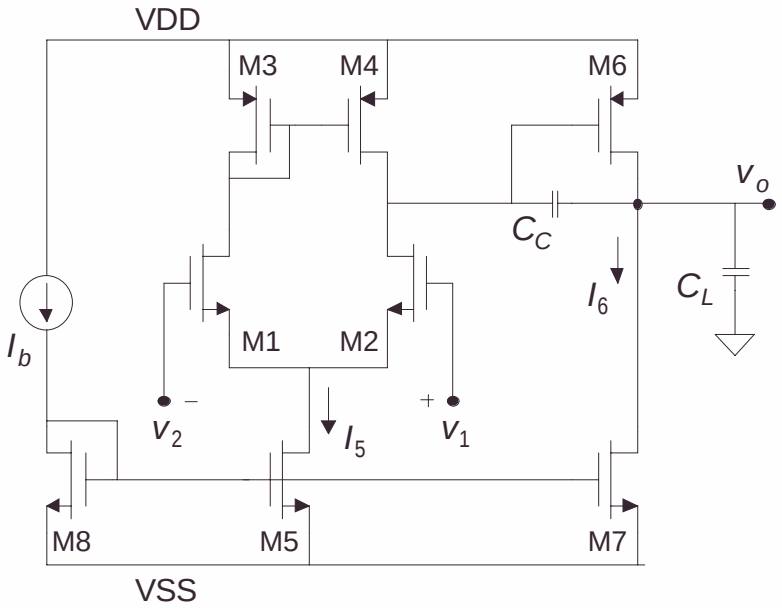
The SR spec (if given) will determine I_5.

Also on inspection :

$$G_{m1} = g_{m1} = g_{m2} \qquad G_{m2} = g_{m6}$$

$$G_{o1} = g_{o2} + g_{o4} \qquad G_{o2} = g_{o6} + g_{o7}$$

Two-stage Op-Amp Design Procedure



DESIGN EQUATIONS

$\omega_{p2} \geq 2.2\ \omega_u$ [1] $\omega_z \geq 10\ \omega_u$ [2]

$\omega_u = \frac{G_{m1}}{C_c}$ [3] $\omega_z = \frac{G_{m2}}{C_c}$ [4]

$\omega_{p2} \approx \frac{G_{m2}}{C_L}$ [5] $SR = \frac{I_5}{C_C}$ [6]

Choose $C_C > C_L$ (2.2/10) [1], [2], [4], [5]

Choose $I_5 = (SR)C_C$. [6]

Choose $G_{m1} = 2\pi f_u C_C$. [3]

Choose $G_{m2} = 2.2 G_{m1} C_L / C_C$. [1], [3], [5]

We can use I_5 and G_{m1} to find $(W/L)_{1,2}$.

$(W/L)_{3,4}$ is chosen to satisfy CM requirements. It also affects the op-amp offset voltage.

$(W/L)_5$ is chosen to satisfy CM requirements

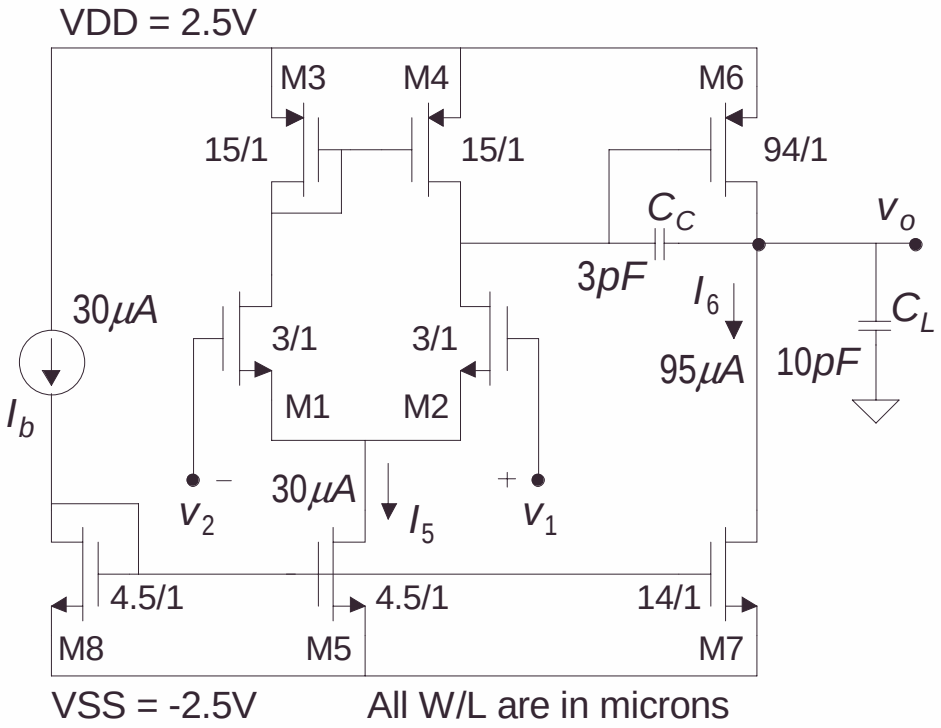
$(W/L)_6$ and I_6 are chosen so that $g_{m6} = G_{m2}$ and so that V_{GS6} is acceptable (close to V_{GS3}).

$(W/L)_7$ is chosen as a scaled M6 designed to carry I_6 rather than I_5.

$(W/L)_8$ sets bias in conjunction with I_b.

$I_b + I_5 + I_6$ sets power dissipation Pd.

Two-stage Op-Amp Numeric Example



This ← is an op-amp design from Allen & Holberg (Ref: AH, Ch 6). To give a general impression of performance, we now quote several numeric results from that design, as follows:

$SR \geq 10V/\mu\text{sec}$ [spec]

$f_u = 5MHz$ [spec]

$A_{v0} = 7696$

$_{o(\min)} = 0.35$

$Pd = 0.62mW$

$g_{m1} = 94.2\mu S$

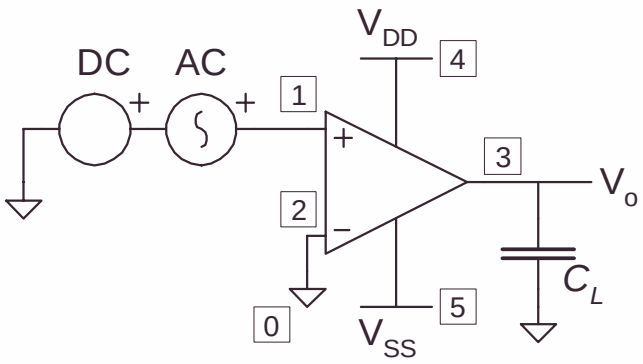
$g_{m6} = 942\mu S$

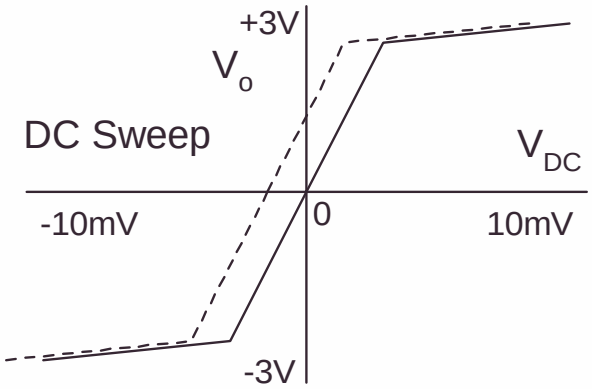
Note that open-loop gain A_{v0} can be improved by using longer devices to reduce the λ values, at the cost of greater die area.

The detailed calculations can be found in Example opa_2s_1. It also shows that use of a 2-micron channel length would increase open-loop gain to around 165,000.

A similar design using the Razavi 0.5um LEVEL 1 model can be found in opa_2s_2.

☐ Open-Loop Simulation Tests





This ← is the set-up for open-loop tests, in which the op-amp is implemented as a sub-circuit.

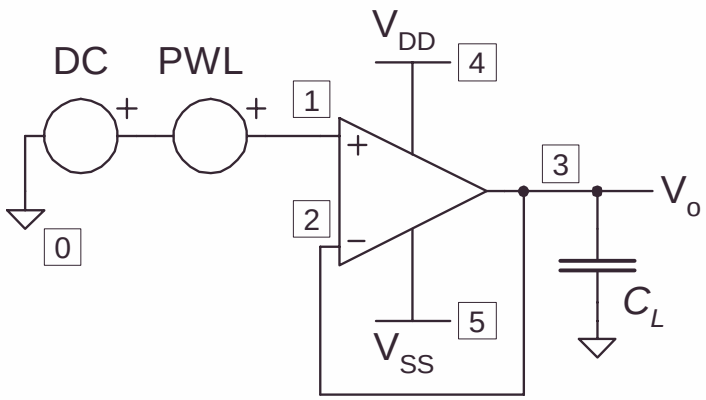
The DC source is set to 0V, but it is also used for a DC sweep, using a range of about (-10mV to +10mV).

The DC sweep plot looks something like this ↙. The initial sweep result will be offset from zero to left or right (the dotted curve). Size adjustments of (M3,M4) will change this offset value.

The slope through the origin gives the op-amp's DC gain.

At this point, we can run the AC sweep which gives us a Bode plot of Gain magnitude and phase over frequency. It confirms the DC gain value, and it also gives the crossover frequency f_u and the phase margin (*PM*).

☐ Closed-Loop Simulation Tests



This ← is the set-up for closed-loop tests, in which the op-amp is implemented as a sub-circuit.

The DC test shows the range of compliance for the unity-gain connection.

The transient test uses a PWL square wave. The initial PWL amplitude is large (e.g. ±2V) and it gives the available slew rate. It should be quite close to the target *SR* value.

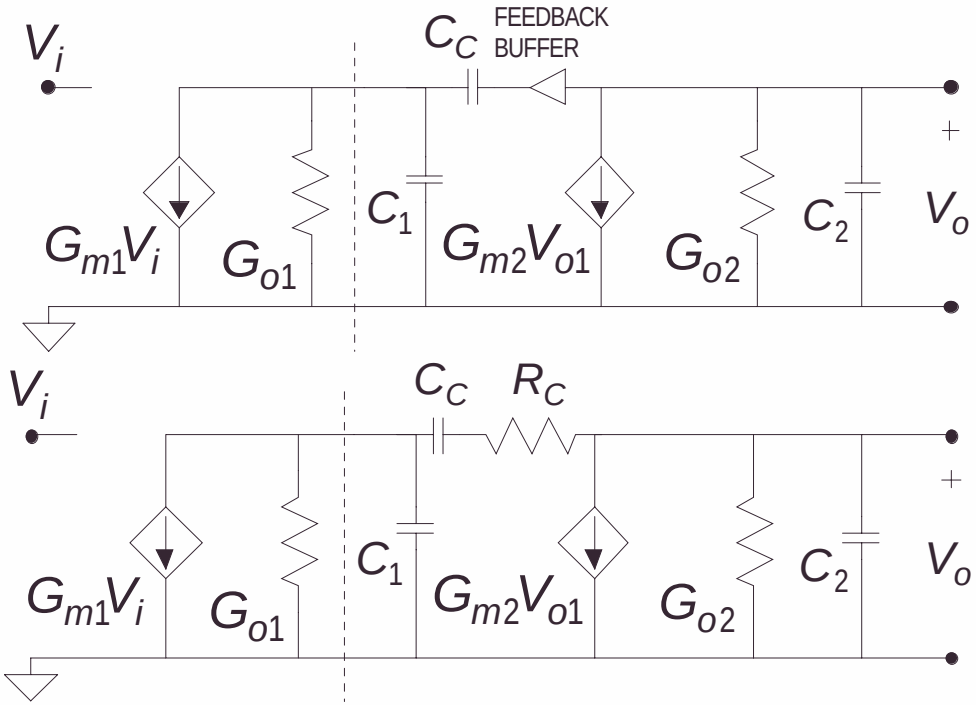
The later PWL amplitude is small (e.g. ±0.1V) and this is so as to observe the linear behaviour, and to look at the *overshoot* and the *settling time*.

The *overshoot* and the *settling time* should correspond with the observed phase margin.

If the phase margin (or the gain-bandwidth) is not sufficient, other measures can be used to improve the performance ...

☐ Removing the Effects of the Zero

To achieve greater bandwidth the Zero can be eliminated, or it can be moved so that it cancels the 2nd pole.



METHOD A. ← We place a buffer in the feedback path so that C_C cannot feed forward. This eliminates the zero.

METHOD B. ↙ We place a resistor R_C in the feedback path. The action is more subtle, but also very effective. It can be used to send the zero to infinity, or to cause it to cancel the second pole …

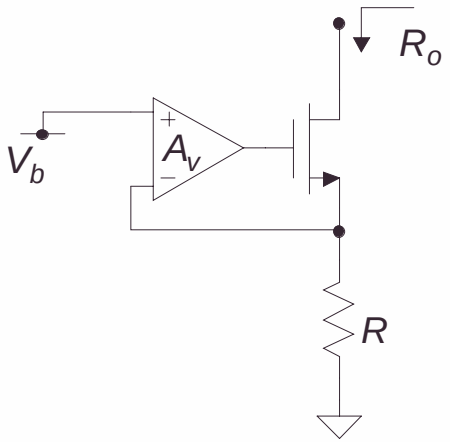
Method B alters the zero location and introduces a third pole such that …

$$s_z = \frac{-1}{(R_c - 1/G_{m2})C_c} \qquad s_{p3} = \frac{-1}{R_c}\left(\frac{1}{C_1}+\frac{1}{C_2}+\frac{1}{C_c}\right) \approx \frac{-1}{R_c C_1}$$

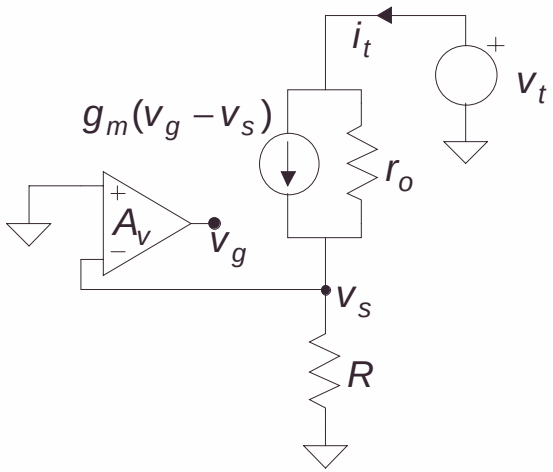
Setting $R_C = 1/G_{m2}$ will remove the zero to infinity. R_C is implemented by using a non-saturated MOSFET.

See Ref GT Ch 4 for greater detail.

☐ The Gain-Boosting Approach



The pursuit of high gain ***and*** high speed presents some difficulties. A single-stage approach is more conducive to high speed, but further cascoding for even higher gain would reduce the headroom excessively. However, it *is* possible to increase gain in another way. Cascoding, in itself, is a form of feedback and, in the *gain boosting* approach shown here ➔, an auxiliary feedback loop is used to further increase the gain. We've seen this idea already in the form of the Wilson Current mirror. For the AC model shown …

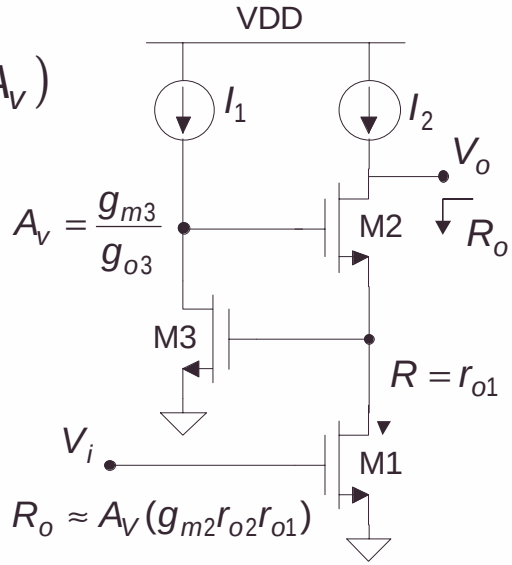


$$v_t = i_t R + r_o\left[i_t - g_m\left(v_g - v_s\right)\right] \qquad v_g = -A_v v_s \qquad v_s = i_t R$$

leading to .. $$\frac{v_t}{i_t} = R_o = r_o + R + g_m r_o R\left(1 + A_v\right)$$

$$R_o \approx A_v\left(g_m r_o R\right)$$

The cascode gain is further boosted by A_v.

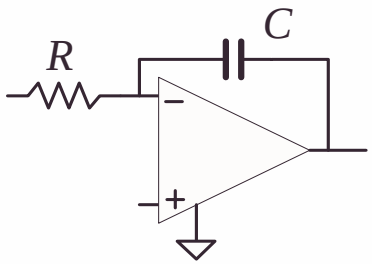


This amplifier ➔ has a gain-boosted R_o, but current-source I_2 needs a similar implementation to achieve the desired high gain. Op-amps would use differential versions of the same idea.

SWITCHED CAPACITOR METHODS

Active Filter Components

Board-level active filters use R's and C's and op-amps.

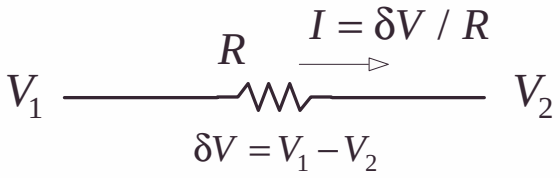


The RC integrator.
A key building block

Integration of RC components presents some difficulties:

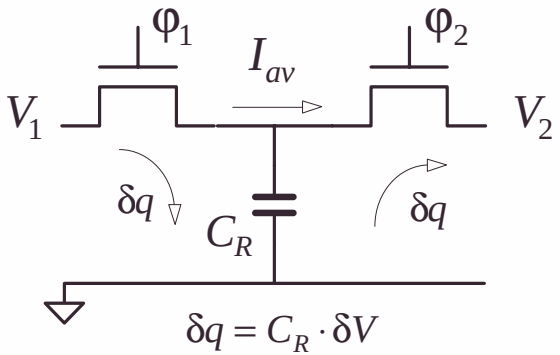
- Low-frequency filters require large RC products
- Large R's and large C's are space-consuming
- MOSFET R's are non-linear
- R's and C's have large absolute-value tolerances
- R's and C's don't track over temperature or voltage

The Switched Capacitor



An actual resistor, R

A resistor carries current in a *continuous* flow



The switched alternative, C_R.

A switched capacitor carries current in *discrete bucket-fulls*

$$I_{av} = \delta q / T = (C_R \cdot \delta V) / T$$

Equate the two currents:

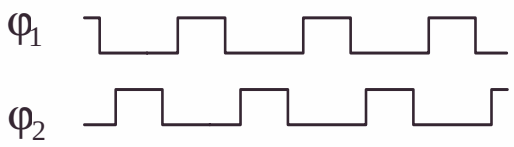
$$I_{av} = I$$

$$I_{av} = (C_R \cdot \delta V) / T = I = \delta V / R$$

yielding:

$$C_R = T / R$$

Non-overlapping Clocks

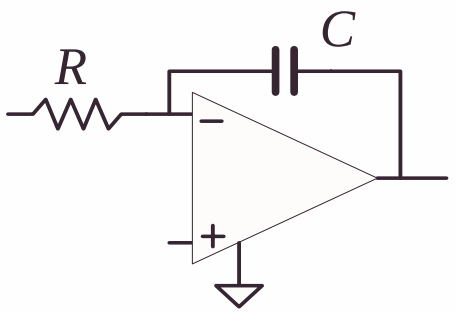


SC Time Constants

Because: $C_R = T / R$

the time-constant becomes:

$$\tau = RC = T \cdot \left(\frac{C}{C_R} \right) = \left(\frac{C}{C_R} \right) / \mathbf{f}_c$$

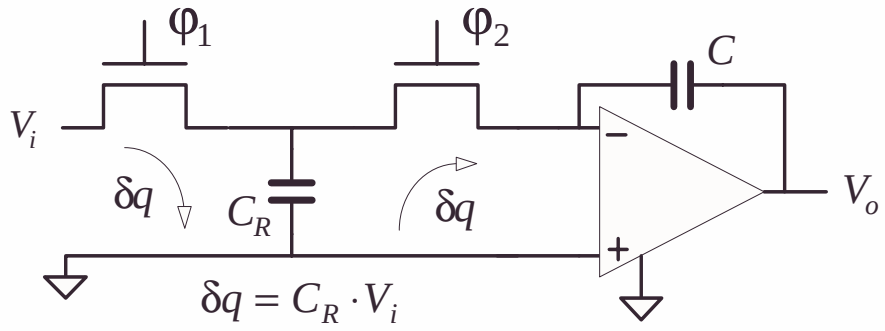


The RC integrator.
A key building block

This has big advantages:

- Accurate Clock frequencies (from crystal oscillators)
- Capacitor Ratios -- small caps, low tolerances, good tracking
- Clock-adjustable break points

A Switched Integrator



$$\delta q = V_i[n-1]\cdot C_R$$

$$\delta V_o[n] = V_o[n] - V_o[n-1] = -\delta q / C$$

Substitute: $$_o[n] - V_o[n-1] = -V_i[n-1]\cdot C_R / C$$

(the difference equation)

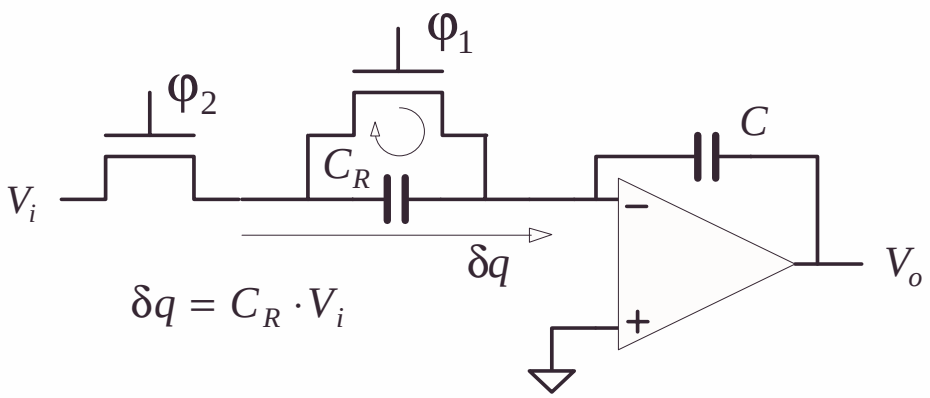
Translate into the z-domain: $$_o(z)\cdot\left(1-z^{-1}\right) = -z^{-1}\cdot V_i(z)\cdot\left(C_R / C\right)$$

The Transfer Function: $$H(z) = \frac{V_o(z)}{V_i(z)} = -\left(C_R / C\right)\cdot\frac{z^{-1}}{\left(1-z^{-1}\right)}$$

This is a *Forward Difference* Integrator

This Integrator samples the input once per cycle. This makes it part of a *sampled-data* system. It is therefore subject to *aliasing* considerations.

Another Switched Integrator



$$\delta q = V_i[n] \cdot C_R$$

$$\delta V_o[n] = V_o[n] - V_o[n-1] = -\delta q / C$$

Substitute: $$_o[n] - V_o[n-1] = -V_i[n] \cdot C_R / C$$

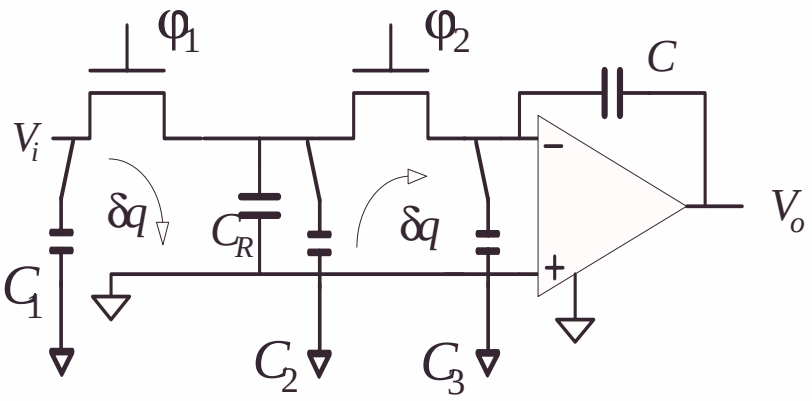
(the difference equation)

Translate into the z-domain: $$_o(z) \cdot \left(1 - z^{-1}\right) = -V_i(z) \cdot \left(C_R / C\right)$$

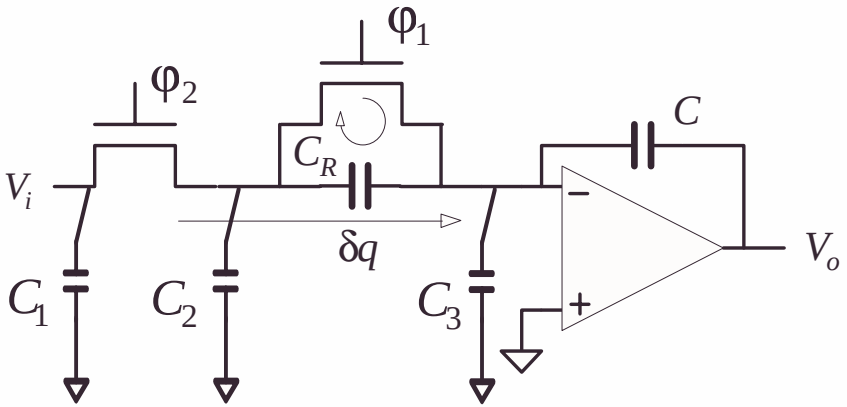
The Transfer Function: $$H(z) = \frac{V_o(z)}{V_i(z)} = -\left(C_R / C\right) \cdot \frac{1}{\left(1 - z^{-1}\right)}$$

This is a *Backward Difference* Integrator

Parasitic Capacitances



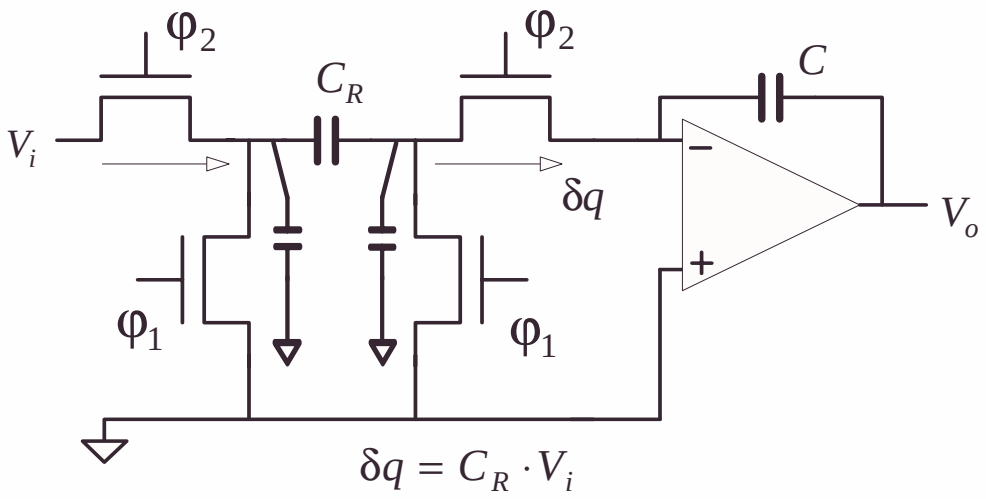
Some of these parasitic capacitances will cause errors in the data transfer.



Can you identify them, and say how they cause problems ?

If parasitic capacitance is a problem, we must employ capacitors that are *much larger* than the parasitics, with serious *area implications* resulting.

Parasitic-Insensitive Switch



Which of the parasitics is a potential hazard ?

These parasitics are grounded on every cycle -- which renders then harmless

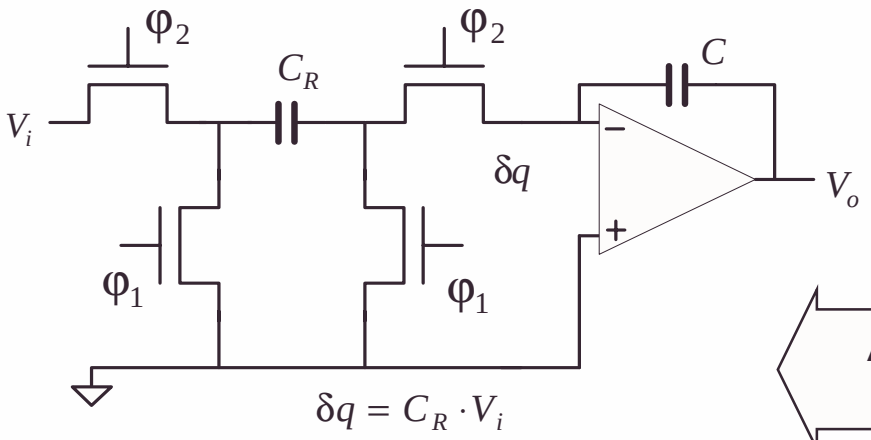
The difference equation: $_o[n] - V_o[n-1] = -V_i[n] \cdot C_R / C$

Translate into the z-domain: $_o(z) \cdot \left(1 - z^{-1}\right) = -V_i(z) \cdot \left(C_R / C\right)$

The Transfer Function: $$H(z) = \frac{V_o(z)}{V_i(z)} = -\left(C_R / C\right) \cdot \frac{1}{\left(1 - z^{-1}\right)}$$

This *Backward Difference* Integrator is stray-insensitive. We can use smaller capacitors.

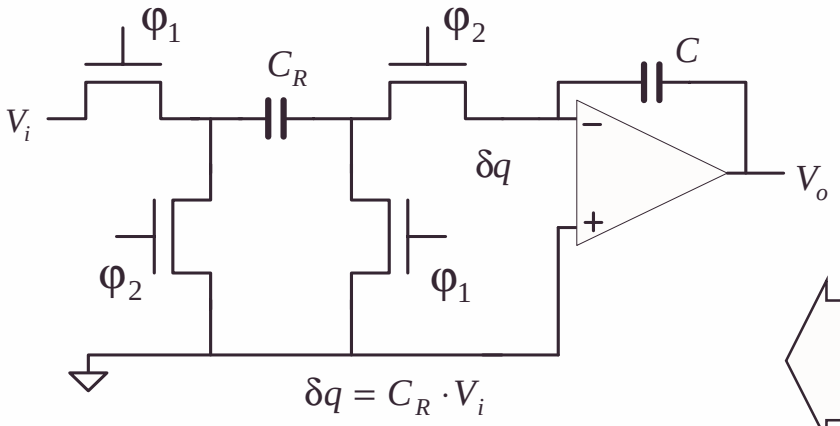
A Useful Pair of Integrators



$$v_o[n] - V_o[n-1] = -V_i[n] \cdot C_R / C$$

$$H(z) = \frac{V_o(z)}{V_i(z)} = -(C_R / C) \cdot \frac{1}{(1 - z^{-1})}$$

An inverting integrator

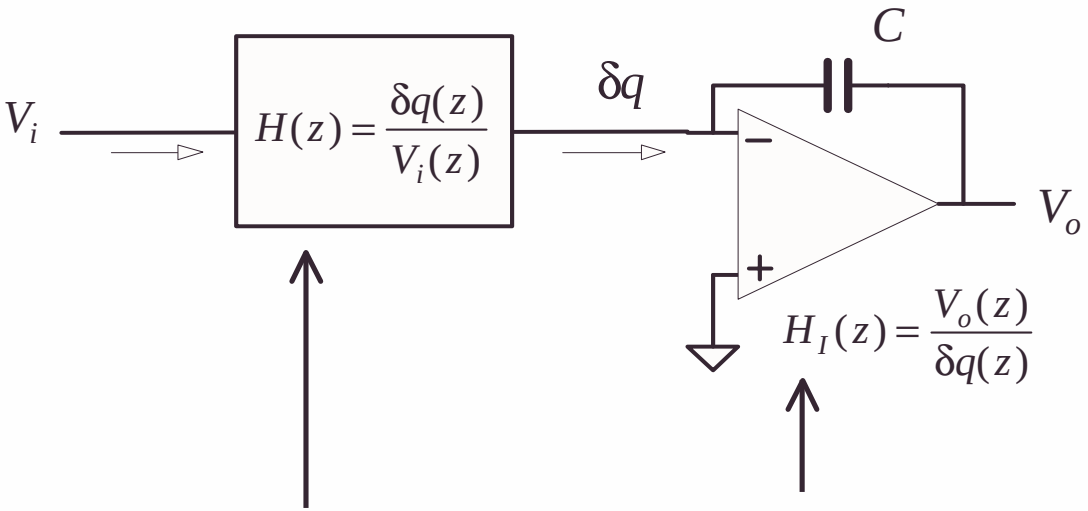


$$v_o[n] - V_o[n-1] = +V_i[n-1] \cdot C_R / C$$

$$H(z) = \frac{V_o(z)}{V_i(z)} = +(C_R / C) \cdot \frac{z^{-1}}{(1 - z^{-1})}$$

A non-inverting integrator

SC Building Blocks – the Integrator Block



Input Block, H

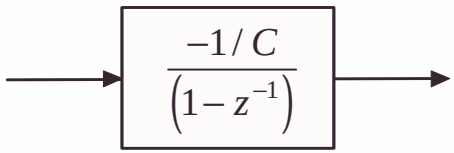
Integrator Block, H_I

We will use the same integrator with a number of different input blocks.

$$_o[n] - V_o[n-1] = -\delta q[n]/C$$

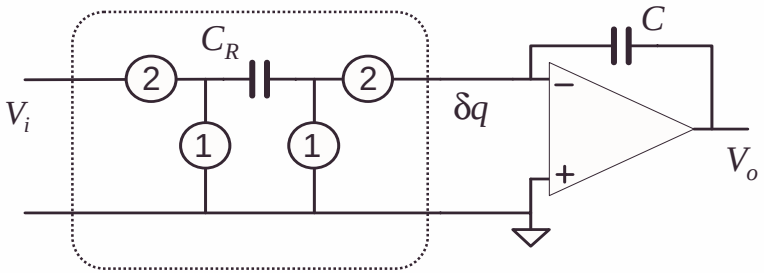
$$_o(z) \cdot \left(1 - z^{-1}\right) = -\delta q(z)/C$$

$$H_I(z) = \frac{V_o(z)}{\delta q(z)} = \frac{-1/C}{\left(1 - z^{-1}\right)}$$



The Integrator Block

SC Building Blocks – three Input Blocks

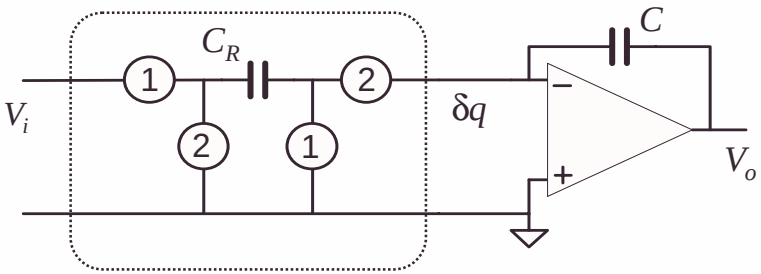


$$\delta q[n] = C_R \cdot V_i[n]$$

$$\delta q(z) = C_R \cdot V_i(z)$$

$$H(z) = \frac{\delta q(z)}{V_i(z)} = C_R$$

Non-inv Input Block

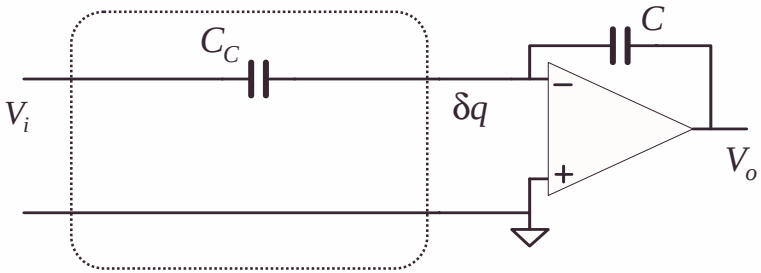


$$\delta q[n] = -C_R \cdot V_i[n-1]$$

$$\delta q(z) = -C_R \cdot V_i(z) \cdot z^{-1}$$

$$H(z) = \frac{\delta q(z)}{V_i(z)} = -C_R \cdot z^{-1}$$

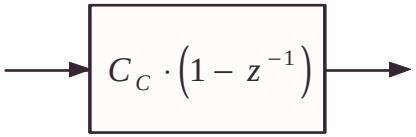
Inverting Input Block



$$\delta q[n] = C_C \cdot \left(V_i[n] - V_i[n-1]\right)$$

$$\delta q(z) = C_C \cdot V_i(z) \cdot \left(1 - z^{-1}\right)$$

$$H(z) = \frac{\delta q(z)}{V_i(z)} = C_C \cdot \left(1 - z^{-1}\right)$$



Capacitor Input Block

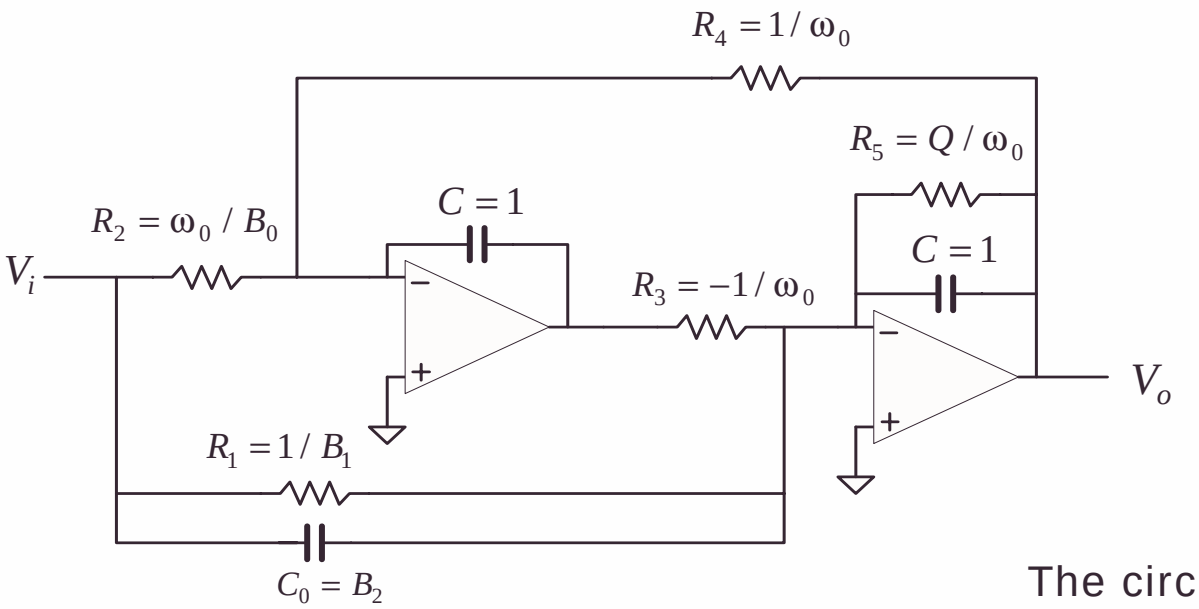
An Active Bi-quad Topology

$$H(s) = -\frac{B_2 \cdot s^2 + B_1 \cdot s + B_0}{s^2 + (\omega_0 / Q) \cdot s + \omega_0^{\,2}}$$

A general-purpose active-filter building block.

Passband gain = 1

Type	B_0	B_1	B_2
LP	$\omega_0^{\,2}$	0	0
HP	0	0	1
BP	0	$\frac{\omega_0}{Q}$	0
BS	$\omega_0^{\,2}$	0	1



The integrators convert current to voltage with a gain of (-1/s).

The circuit *H*(*s*) can be checked by writing its system equation ..

SC Bi-quad Realization

$$H(s) = -\frac{B_2 \cdot s^2 + B_1 \cdot s + B_0}{s^2 + (\omega_0 / Q) \cdot s + \omega_0^{\,2}}$$

Capacitor values are relative to C = 1

$$C_0 = B_2 = 1$$

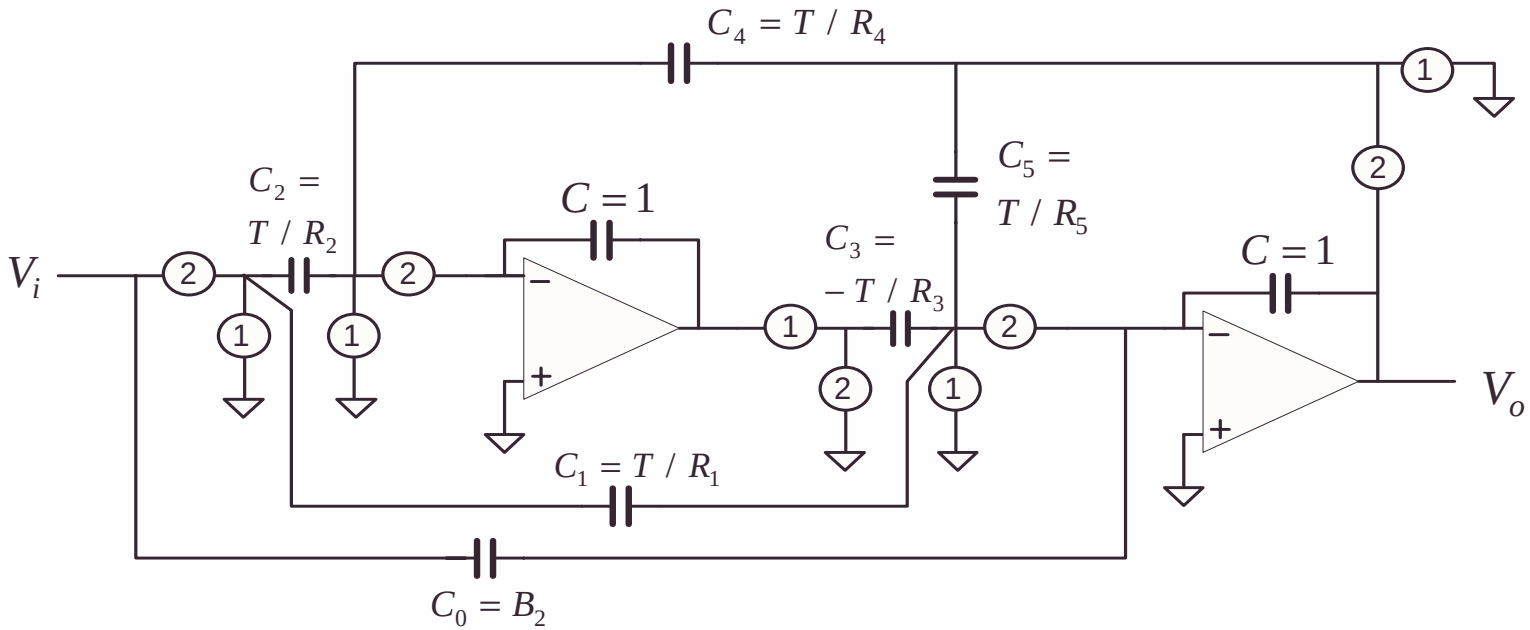
$$C_1 = TB_1 = \omega_0 T / Q$$

$$C_2 = TB_0 / \omega_0 = \omega_0 T$$

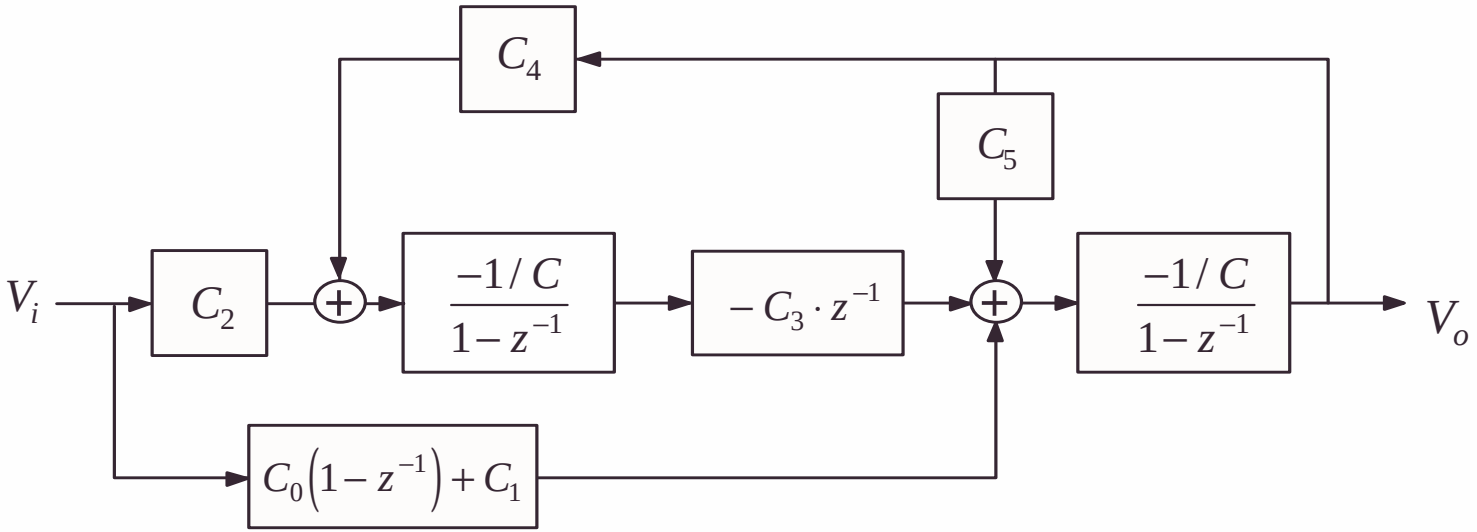
$$C_3 = \omega_0 T$$

$$C_4 = \omega_0 T$$

$$C_5 = \omega_0 T / Q$$



SC Bi-quad Data Flow



Here we do a *precise* analysis of the SC bi-quad

As with a digital filter, the transfer function is *periodic*.

The defining equation:

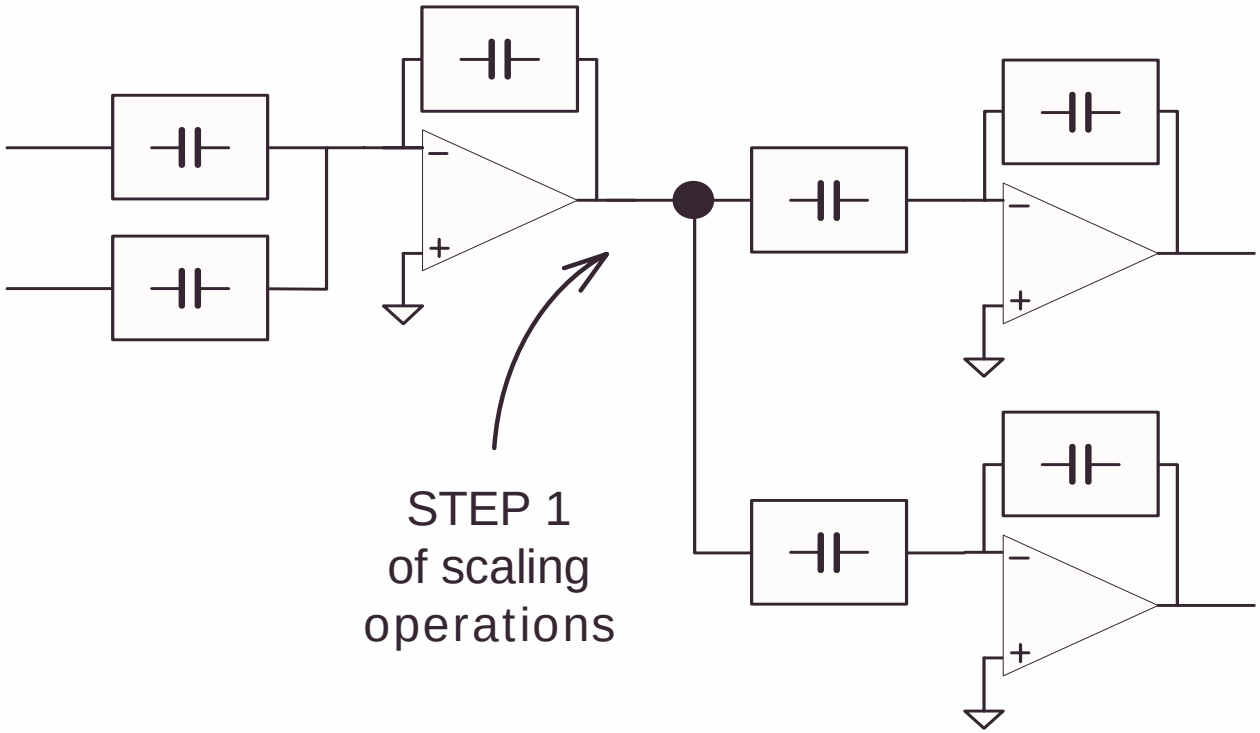
$$_o = \frac{-1/C}{\left(1-z^{-1}\right)}\left[C_5V_o + C_0\left(1-z^{-1}\right)V_i + C_1V_i + \frac{z^{-1}C_3/C}{\left(1-z^{-1}\right)}\cdot\left(C_2V_i + C_4V_o\right)\right]$$

On setting C = 1, and after re- arrangement:

$$H(z) = \frac{V_o(z)}{V_i(z)} = -\frac{\left(C_0+C_1\right)+\left(C_2C_3-C_1-2C_0\right)\cdot z^{-1}+C_0\cdot z^{-2}}{\left(1+C_5\right)+\left(C_3C_4-C_5-2\right)\cdot z^{-1}+z^{-2}}$$

Range Scaling of SC Circuits

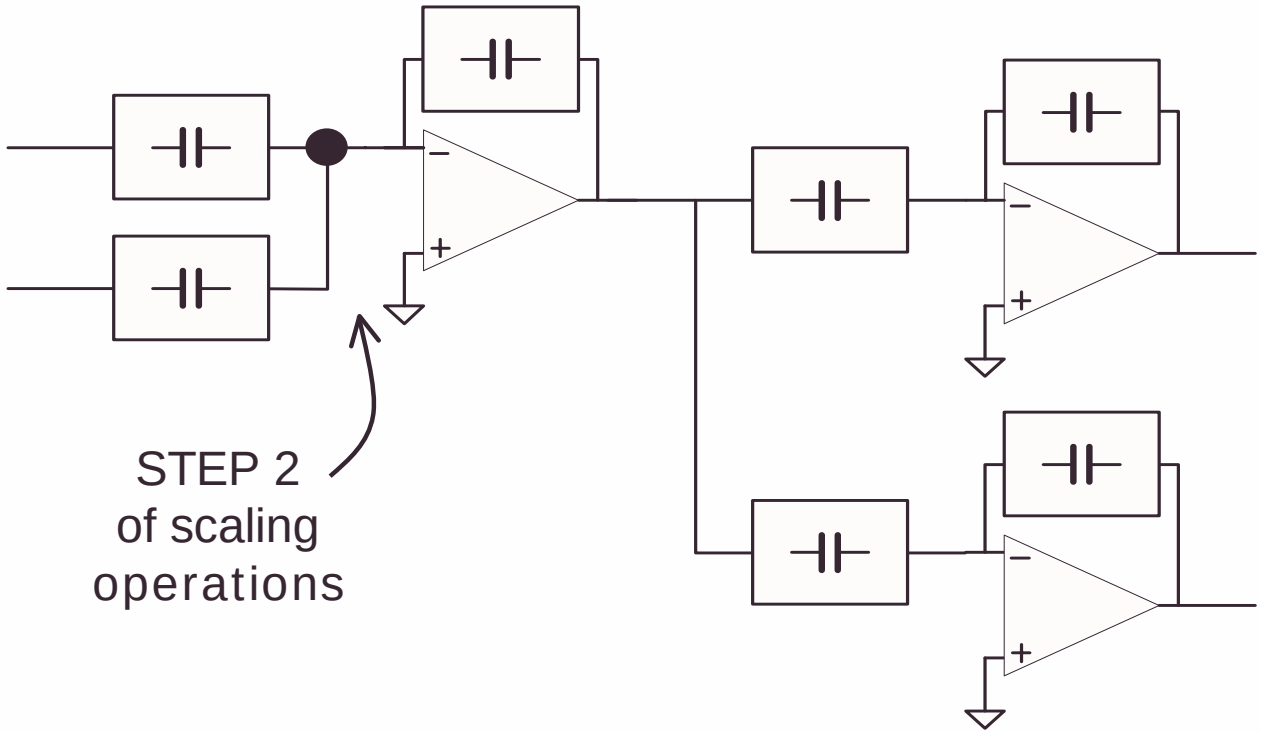
When the design is completed, we can do additional scaling in two steps, as follows. This first step helps us make better use of the op-amp dynamic range and thus improves the SNR.



Scale all caps connected to an op-amp OUTPUT node. This adjusts the op-amp output swings, for optimum dynamic range.

Capacitance Scaling of SC Circuits

If the caps connected to an input node are all above the minimum size, we can scale them all down, thus reducing the space requirements.



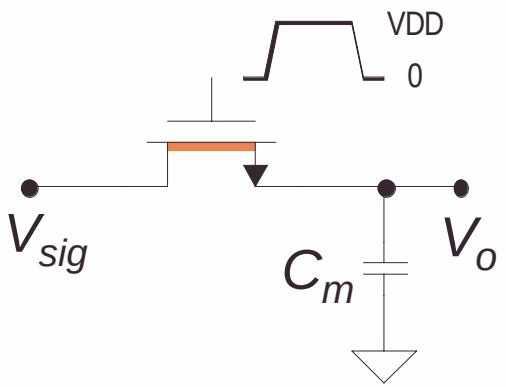
Scale all caps connected to an op-amp INPUT node. This changes only the charge packets, but allows minimisation of capacitor sizes.

SWITCHED CAPACITOR CIRCUITS

☐ Preamble

Until now, we've taken little account of the circuit aspects that limit SC performance. This chapter will attempt to redress that imbalance. It will also point to some other applications of SC technology, notably SC amplifiers and comparators, which can offer improved performance over the more traditional methods.

☐ Channel Charge Injection Errors



When the gate-clock is high, V_o tracks the signal V_{sig}.

When the gate-clock goes low, V_o becomes "frozen" in time.

The "memory" capacitor C_m "remembers" the signal value.

This is an effective sample/hold action, except that ..

The "ON" switch requires a channel charge of : $\Delta q = WLC_{ox}(VDD - V_{sig} - V_T)$

When the switch goes ON, this charge must be supplied. It gets *absorbed* through S and D terminals from adjacent circuitry.

When the switch goes OFF, this charge (electrons for an NFET) must vanish. It gets *injected* through S and D terminals to the adjacent circuitry.

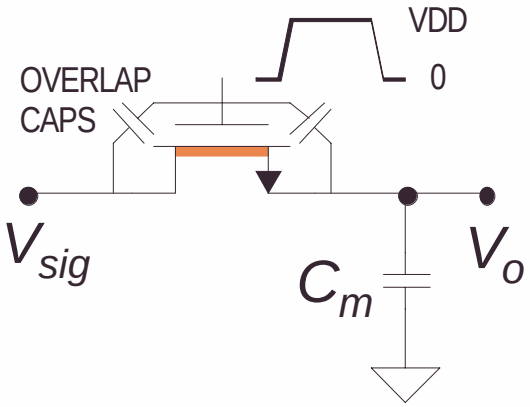
Charge injection induces an error ➔ Δv in the remembered value on C_m.

$$\Delta v = -\frac{\Delta q}{C_m} = -\frac{WLC_{ox}(VDD - V_{sig} - V_T)}{C_m}$$

(negative for an NFET)

Some of this error is just a fixed *offset*. Body-effect on V_T gives some *non-linearity*. The V_{sig} part is a *signal-dependent error*. Offsets can often be eliminated. Some non-linearity may be unavoidable. The signal-dependent error is the most troublesome.

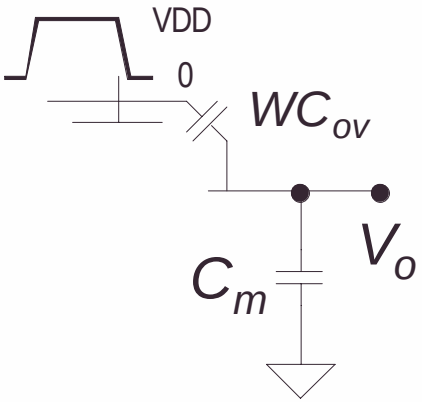
Clock Feed-through Errors



Consider the events during SWITCH-OFF ...

Charge-injection from the channel is not the only feature.

There is also a charge transfer through the *overlap capacitance*, particularly after the channel has disappeared. The effect on the output is:



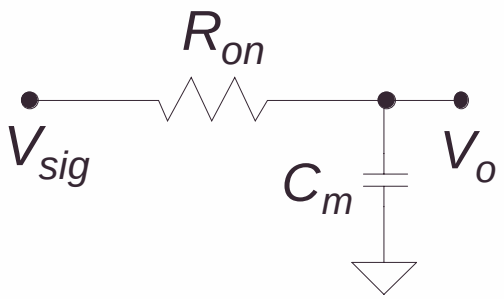
W = channel width
C_{ov} = over-lap capacitance per unit width

$$\Delta V_o = \Delta V_G \frac{WC_{ov}}{WC_{ov} + C_m}$$

The *effective* gate transition ΔV_G is generally less than VDD. Even if we use a pessimistic ΔV_G = VDD, this feed-through will probably be much less than channel charge injection because $WC_{ov} << WLC_{ox}$.

The clock-feedthrough error is not signal-dependent. It contributes a constant offset, and can be lumped with the constant part of channel charge injection.

☐ Noise (kT/C) Errors



The error in Vo will include the noise generated in the switch resistance R_{on}. Resistor noise is "white", i.e. a flat spectrum with a mean-power spectral density of $|V_R|^2 = 4kTR$ V²/Hz.

The noise voltage V_R is in series with R and is scaled by $Z_C/(R+Z_C)$ before reaching V_o. Thus ➔

$$\frac{V_o}{V_R} = \frac{1/sC}{R + 1/sC} = \frac{1}{1+sRC}$$

Because only mean power quantities are quantifiable in a noise context :

$$\left|\frac{V_o}{V_R}\right|^2 = \left|\frac{1}{1+sRC}\right|^2 = \frac{1}{1+\omega^2R^2C^2} = \frac{1}{1+4\pi^2f^2R^2C^2}$$

Thus : $$|V_o(f)|^2 = \frac{4kTR}{1+4\pi^2f^2R^2C^2}$$ and $$P_o = \int_0^\infty |V_o(f)|^2 .df = \frac{kT}{C}$$ .. the total mean noise power at the output

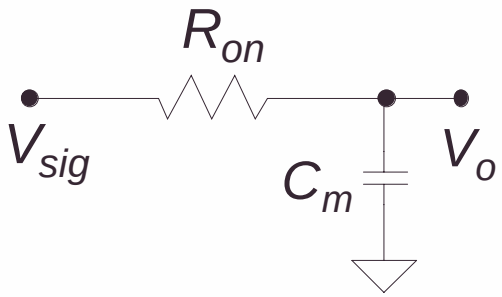
Surprising ? The mean output noise power is independent of R, even though the generated power is proportional to R . This is because any increase in R also reduces the bandwidth, causing less of the generated power to reach the output.

EXAMPLE: RMS noise voltage on C=1pF at room temp:

$$_n = \sqrt{kT/C} = \sqrt{1.38E-23(300)/1E-12} = 64\mu V \text{ rms}$$

A sobering thought : the only way to reduce kT/C noise is to increase C !

The Speed / Precision Trade-off



Switch resistance R_{on} must be low enough to allow V_o to settle before being sampled. Thus, the *clock speed* is limited by the time constant $R_{on}C_m$.

Precision is limited mainly by channel charge injection according to:

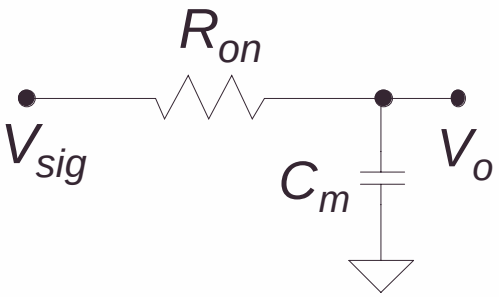
$$\Delta v = -\frac{\Delta q}{C_m} = -\frac{WLC_{ox}(VDD - V_{sig} - V_T)}{C_m}$$

The (speed x precision) product is a constant, as follows (noting that β' = C_{ox} μ) :

$$(\text{speed x precision}) = \frac{1}{R_{on}C_m} \cdot \frac{1}{\Delta v} = \frac{\beta'(W/L)(VDD - V_{sig} - V_T)}{C_m} \cdot \frac{C_m}{WLC_{ox}(VDD - V_{sig} - V_T)} = \frac{\mu}{L^2}$$

For *high speed*, we use *wider switches* to reduce R_{on}, and *smaller capacitors* C_m . The larger switches then have larger WLC_{ox} and larger WC_{ov}, for a worsening capacitor ratio, and a corresponding loss of precision. Without special measures, only advances in technology (smaller L) can beat this very basic limitation. An example will illustrate ..

☐ Speed / Precision Example



The clock speed is itself dictated by precision needs. Suppose we allow m time-constants ($m\tau$ sec) for settling in a 3V system. The sampling error becomes 3 $e^{-m\tau}$ volts. Examples:

$$m = 4 \rightarrow 55mV, \qquad m = 5 \rightarrow 20mV$$

If we choose m = 5, then each half-clock-period is of 5τ sec duration, and a clock period becomes 10τ sec. Thus:

$$f_{ck} = \frac{1}{10 R_{on} C_m}$$

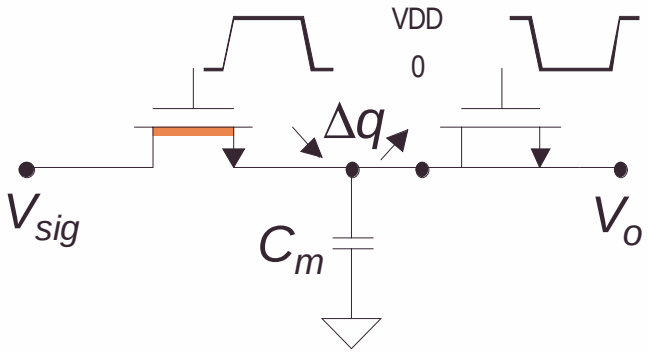
It follows that : $$(\text{speed x precision}) = 10 \cdot f_{ck} \cdot \frac{1}{\Delta v} = \frac{\overline{\mu}}{L^2} \quad \rightarrow\rightarrow \quad f_{ck} = \frac{\overline{\mu}}{10 L^2} \cdot \Delta V$$

The NMOS μ is typically 350 cm^2/volt-sec, or 0.035 in SI units. We can now tabulate clock speed versus channel length for some allowable ΔV, remembering that these ΔV errors may accumulate …

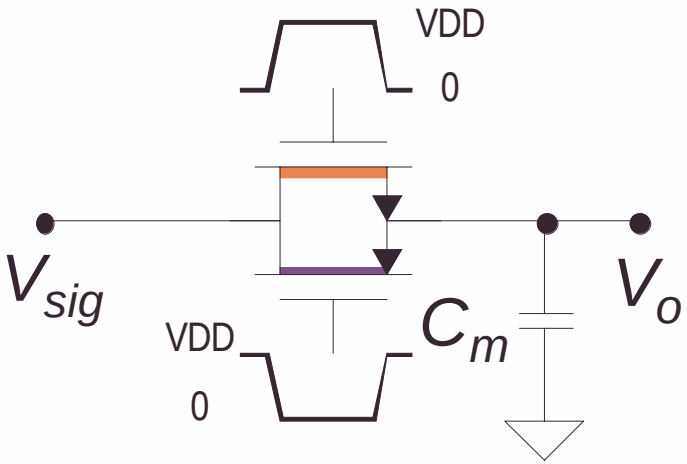
CLOCK LIMITS	L = 0.5 um	L = 0.1 um
ΔV = 1 mV	f_{ck} = 14 MHz	f_{ck} = 350 MHz
ΔV = 0.2 mV	f_{ck} = 2.8 MHz	f_{ck} = 70 MHz

Performance is set to improve considerably as channel lengths grow shorter.

☐ Error Reduction Strategies



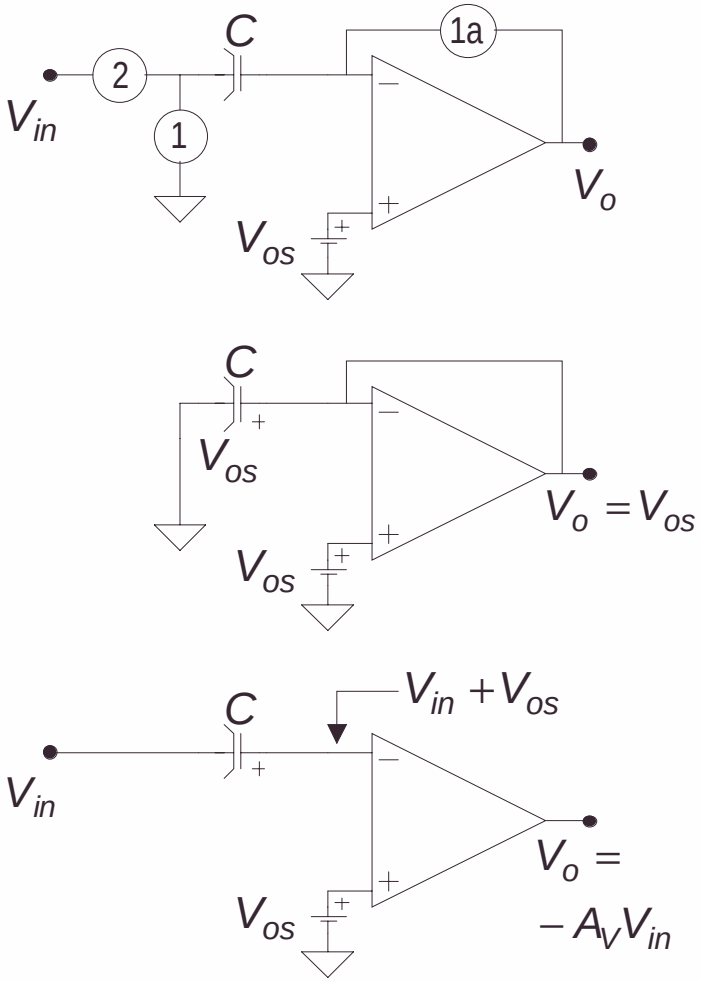
🡸 Addition of a dummy S-D-shorted MOSFET, with gate driven in anti-phase, seeks to eliminate Δq by having the dummy device absorb it. Main snag: hard to say how much of channel charge is injected toward C_m.



🡸 Addition of a PMOS device in parallel with gate driven in anti-phase. This is a standard approach to reducing channel resistance to allow wider signal swing. But it also means that both positive and negative charge injection occurs during switch-off. These charges can be made to cancel, but cancellation is exact only for one value of V_{sig}.

In addition to these and other similar measures, *differential operation* is widely used to eliminate the *offsets* caused by charge injection. It can also reduce the non-linear effects.

An SC Comparator



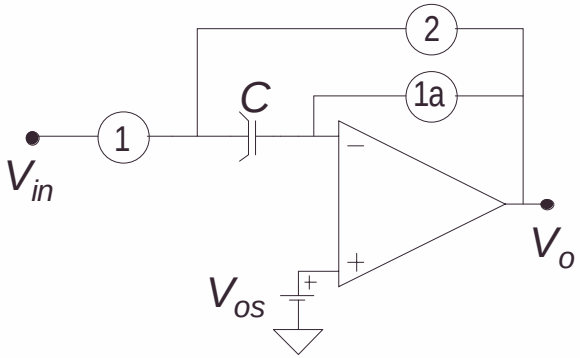
This comparator compensates for the offset of the op-amp. It uses two-phase non-overlapping clocks.

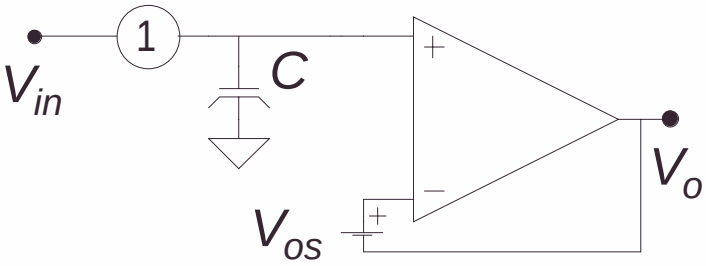
Phase 1a means phase 1 slightly advanced in time. (It turns off just before phase 1 does). The reason for this will be clarified very shortly.

RESET PHASE : Capacitor C is pre-charged to V_{os}.

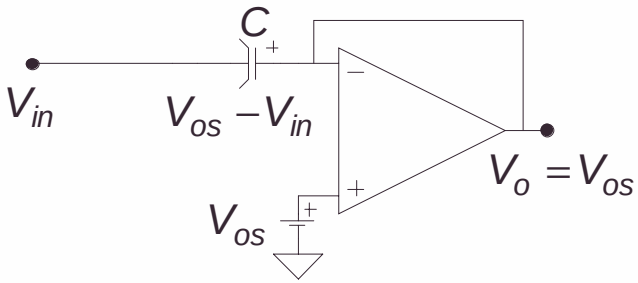
COMPARISON PHASE : Right side of capacitor C is raised to $V_{in}+V_{os}$. The output switches when $V_{in} = 0$. The output level saturates close to the supply levels.

☐ Offset-free SC Buffer

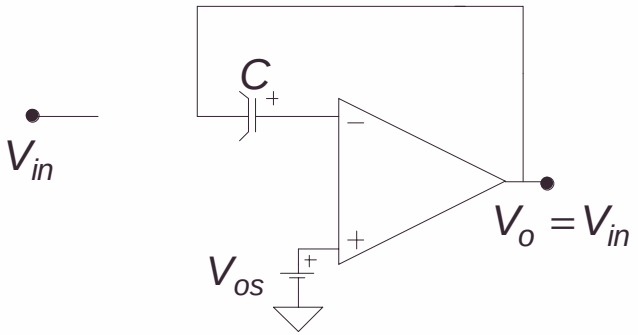




This circuit ➔ is a simple sampler-buffer that does not eliminate V_{os} and it also suffers from signal-dependent charge injection by the ϕ_1 switch.



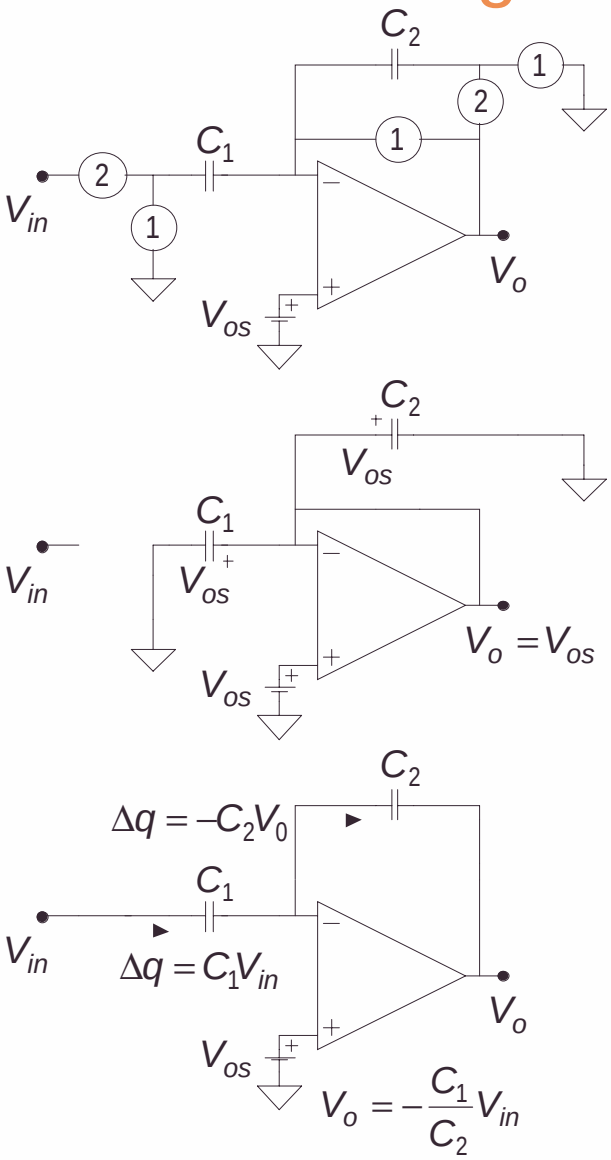
Here, ➔, both problems are resolved. The small timing advance of phase 1a is essential to eliminating the signal-dependent charge injection which phase 1 at the input V_{in} could otherwise generate. More on this point later.



We still have charge injection as phase 1a switches off but, as its channel is at a virtual zero, there is no signal dependence and no non-linearity from body effect. It contributes only an offset. The offset can be eliminated by differential working.

Notice the *bottom-plate* of *C* is NOT on the summing node where its capacitance to substrate (up to 20%) would degrade the op-amp speed and also its precision.

☐ Inverting SC Op_Amp



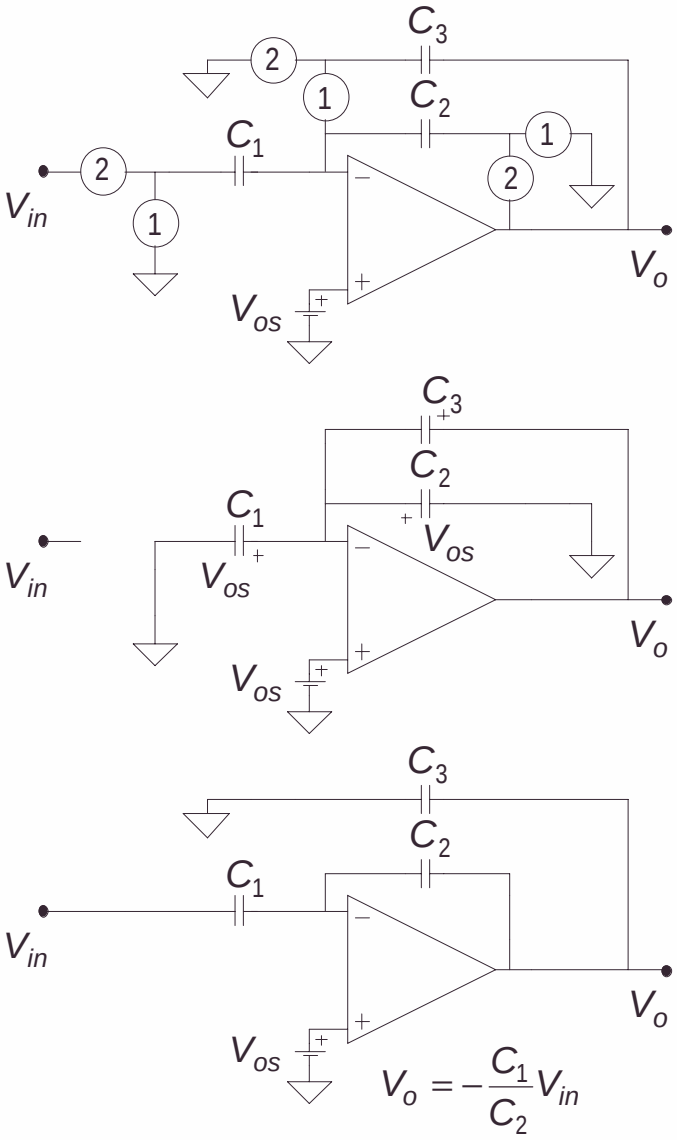
Here, 🡐, we use two non-overlapping clocks phases, ϕ_1 and ϕ_2. The RESET phase ϕ_1 samples the offset and is followed by an EVALUATE phase ϕ_2 that amplifies V_{in} with inversion, and with a gain of $-C_1/C_2$.

The RESET phase 🡐 pre-charges C_1 and C_2 to V_{os} for offset cancellation.

The EVALUATE phase 🡽 transfers charge Δq through the system and V_o steps to its new value. After settling, the output can be read.

The switching of V_o between the levels of V_{os} and an amplified V_{in} requires a rapid slewing behaviour, and this is often difficult to achieve.

☐ Inverting Non-slewing SC Op_Amp



This circuit ← uses an additional capacitor C_3 to maintain the output close to the valid V_o at all times, and in this way *avoids the need for rapid slewing*.

The RESET phase ↙ pre-charges C_1 and C_2 to V_{os} for offset cancellation. Meanwhile, V_o stays at its earlier valid output value stored on C_3. Feedback (which provides a virtual zero) is maintained by C_3 and not by C_2.

The EVALUATE phase ↙ transfers charge Δq through the system via C_2, and V_o takes a small step to its new value, while updating the charge on C_3 . After settling, the output can be read.

An additional small capacitor from the left side of C1 to the output Vo maintains feedback while all switches are open and has a useful de-glitching effect (see Ref).

☐ Minimising Charge Injection Effects

We recall that ..

Charge injection induces an error ➔ Δv in the remembered value on C_m.

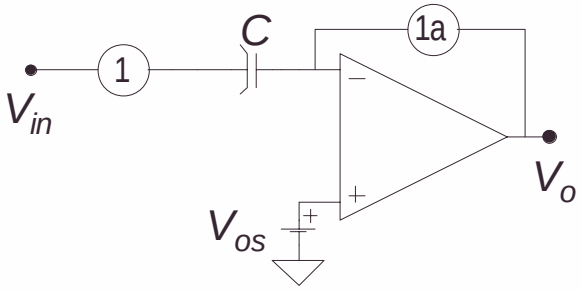
$$\Delta v = -\frac{\Delta q}{C_m} = -\frac{WLC_{ox}(VDD - V_{sig} - V_T)}{C_m}$$

(negative for an NFET)

It can yield an *offset* error, a *signal-dependent* error, and a *linearity* error from V_T.

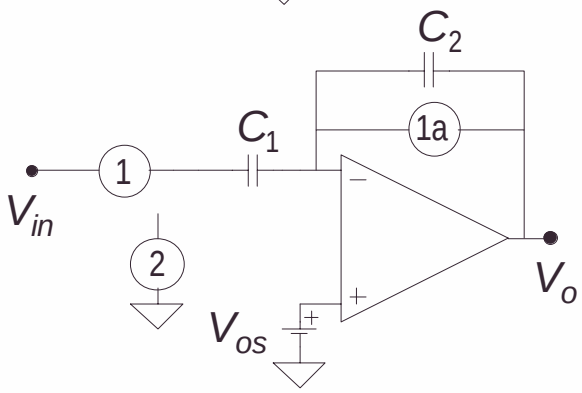
But if the channel voltage is fixed (e.g. at ground or at virtual zero), only the offset error remains, and this is generally eliminated by differential working.

Switches at fixed potential are opened first. This can render other switch openings harmless, as follows ..

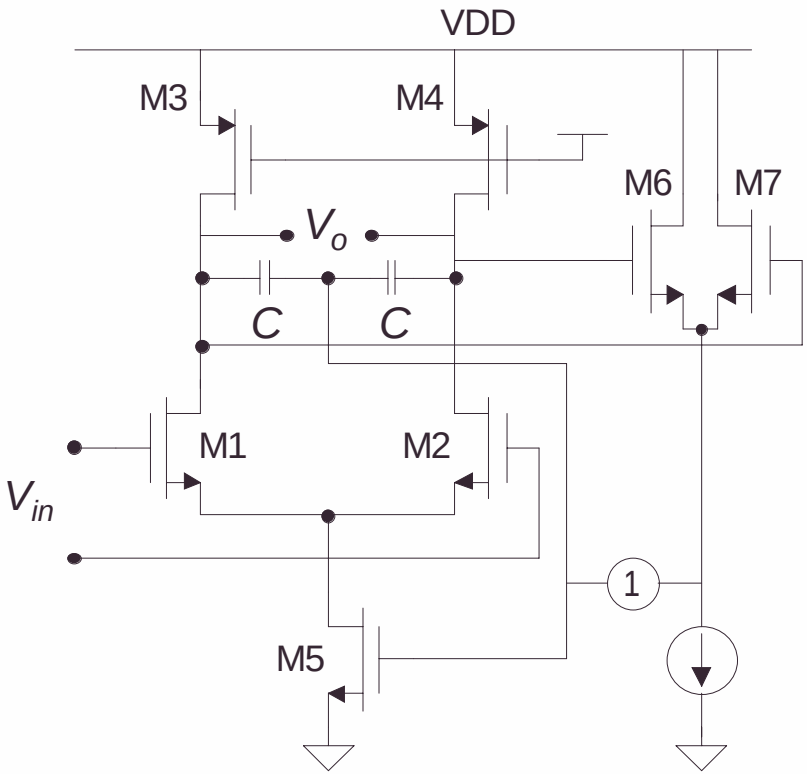


In this circuit ➔, if (1a) opens before (1), then (1) cannot inject into C because the right side of C is open-circuited. Only an offset due to (1a) need be considered.

In this circuit ➔, if (1a) opens before (1), then the charge on the (C_1 ,C_2) node cannot change because there is no resistive exit path. Only an offset due to (1a) need be considered.



SC CM Feedback Amp



SC methods are well-suited to the provision of common-mode feedback.

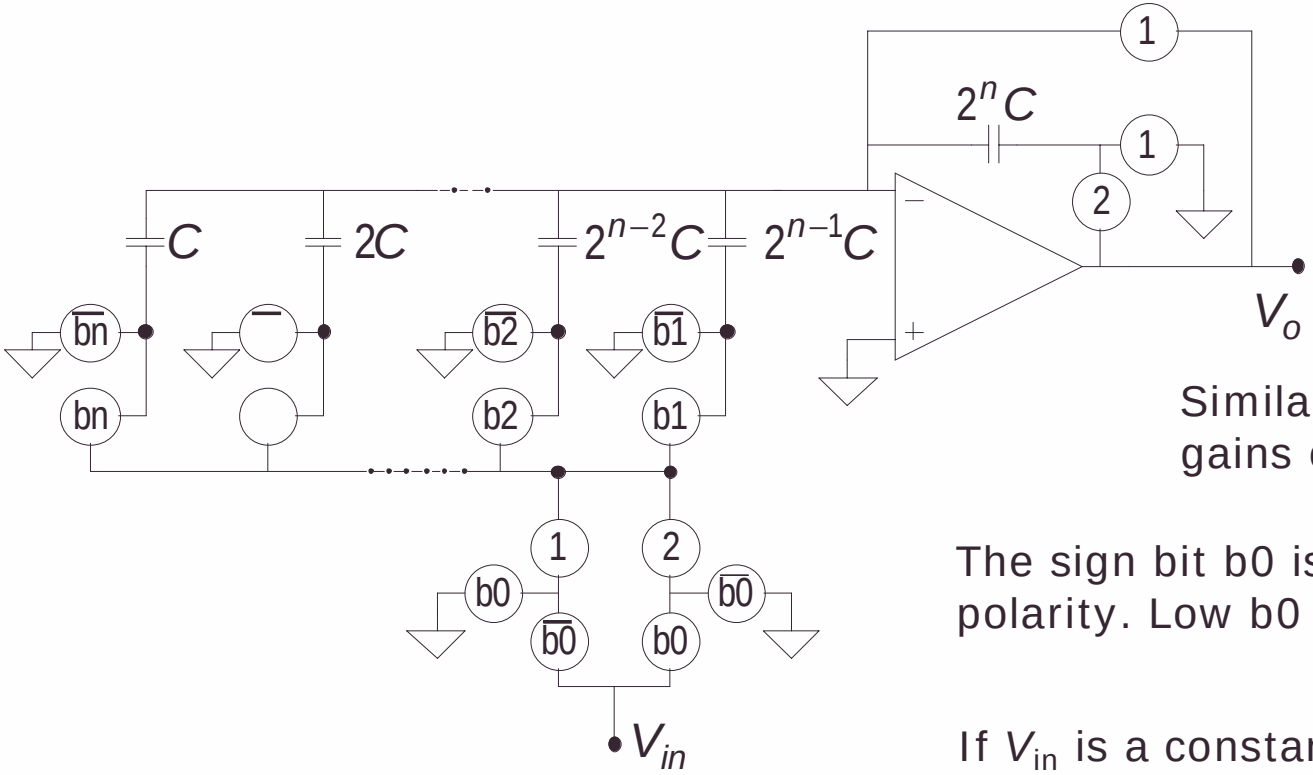
This amplifier ← has a differential output V_o and, because M3 and M4 are in SAT mode (for higher gain), a CM stabilizing loop is required.

To set up suitable conditions across capacitors C, switch (1) is turned on while the differential V_{in} is zero. Then the common-mode outputs (the capacitor end-points) drive (M6,M7) to adjust V_{GS5} until the CM output level stabilizes at $V_{GS5}+V_{GS6,7}$, with the capacitor mid-point at V_{GS5}.

In the *normal feedback mode*, switch (1) is open. Then, common-mode variations are sensed at the mid-capacitor node and fed back to the gate of M5, creating a negative feedback loop that provides CM correction.

A periodic refresh maintains correct CM level.

SC Multiplying DAC



The gain of bit b1 is :

$$|g_1| = \frac{2^{n-1}C}{2^nC} = \frac{1}{2}$$

Similarly for other bits, yielding gains of ¼ for b2, 1/8 for b3, etc

The sign bit b0 is used for switching gain polarity. Low b0 yields a non-inverting gain.

If V_{in} is a constant = V_{REF}, we have a DAC.

If V_{in} is variable, we have a *multiplying* DAC.

$$_o = \pm(\text{digital word value}) \cdot V_{in}$$

If the word length n is large, the capacitor area becomes excessive. This can be resolved by using two separate “banks” with different weightings.

DELTA SIGMA METHODS

Suggested Reading

Delta-Sigma data Converters : Theory, Design and Simulation
Ed: Norsworthy , Schreier, Temes. IEEE Press 1997

☐ Preamble

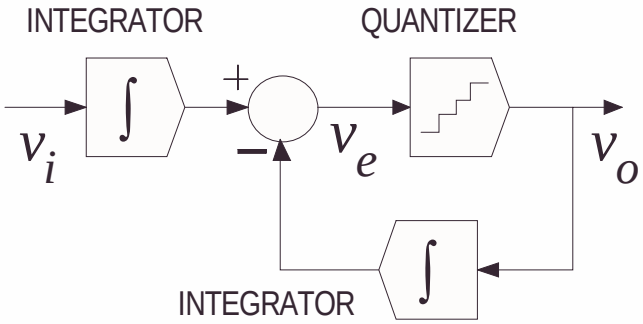
A slow-changing but high-precision signal can have an equivalent description in a form that changes very rapidly but requires far less precision.

This latter form is compatible with fast-but-simple digital circuitry, and Δ-Σ methods are the key to translating a signal to this new form.

Conventional A/D and D/A converters rely on precision analogue parts but, nowadays, Δ-Σ methods have led to converters that are mostly digital, yet they still achieve high precision.

To do this, their digital circuits must run very much faster than the analogue signals that they work with. Consequently, converters that use Δ-Σ methods must operate at low to medium sampling rates.

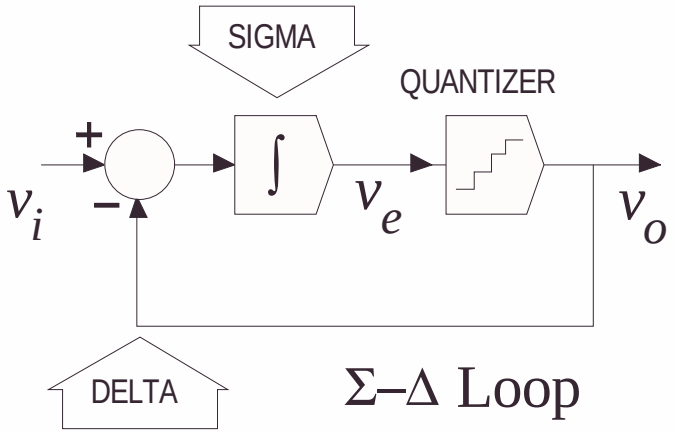
A Δ-Σ Feedback Loop



Σ–Δ Loop Concept

This is a feedback system ← that can convert a slow-varying high-precision voltage-signal v_i into a fast-changing but low-precision alternative called v_o. The output signal v_o includes a high level of quantization noise, but we will see that most of that noise can later be removed by filtering.

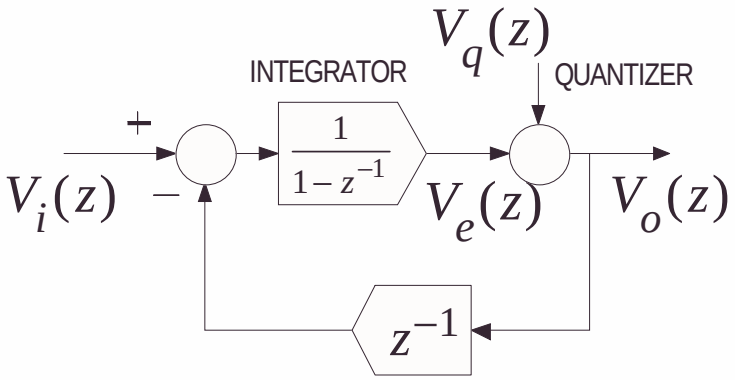
The difference between the integrator outputs becomes the error signal v_e in a negative feedback loop, and the loop acts to minimize that error.



Σ–Δ Loop

Rather than integrating v_i and v_o separately, and then subtracting them, we can get the same result by subtracting them first and then integrating the difference. That is the approach taken here ←, in what is called *a Delta-Sigma loop*.

A Digital Δ-Σ System



Σ–Δ System Diagram

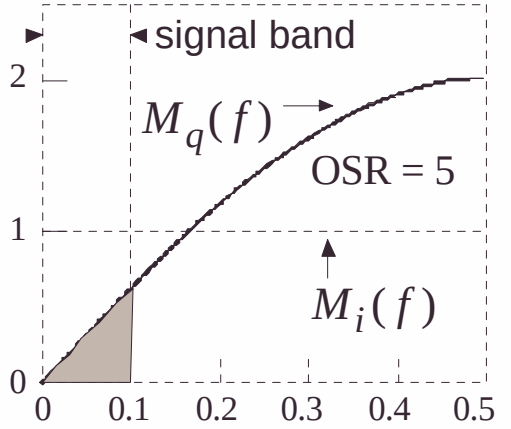
$$v_e[n] = v_e[n-1] + v_i[n] - Q(v_e[n-1])$$

$$v_o[n] = Q(v_e[n])$$

We could use these equations to build a z-domain description, or we can do so more directly by inspection:

$$_o(z) = V_q(z) + \frac{1}{1-z^{-1}}\left[V_i(z) - z^{-1}\cdot V_o(z)\right]$$

$$_o(z) = V_i(z) + (1-z^{-1})\cdot V_q(z)$$



$\Delta-\Sigma$ Noise Filter (1^{st} Order)

We identify: $H_s(z) = 1, \quad H_q(z) = (1-z^{-1})$ signal gain H_s noise gain H_q

Examining the noise gain:

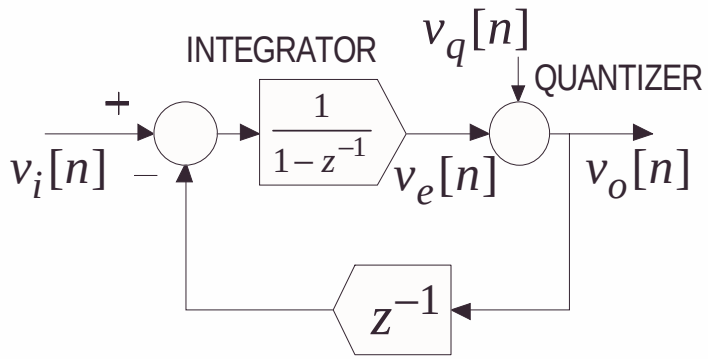
$$M_q(f) = \left|H_q\left(e^{j2\pi f}\right)\right| = \left|1-e^{-j2\pi f}\right|$$

$$= \left|e^{-j\pi f}\left(e^{j\pi f} - e^{-j\pi f}\right)\right| = 2\cdot\sin(\pi f)$$

$$A_1 = \int_{-0.5/\mathrm{OSR}}^{+0.5/\mathrm{OSR}} M_q(f)^2.df \qquad A_1 \approx \frac{\pi^2}{3\cdot \mathrm{OSR}^3}$$

OSR	$1/A_1$
5	39
16	1247
64	79692
256	5099700

☐ Δ-Σ System in Time Domain



$\Delta - \Sigma$ System Diagram

We can implement an iterative procedure as follows:

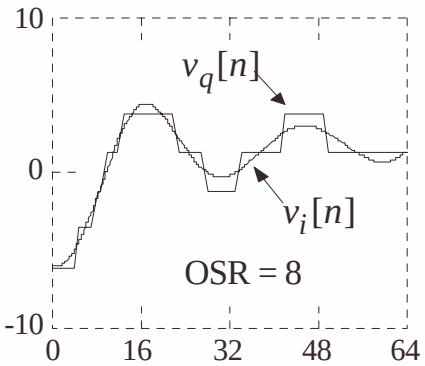
For $n = 0..N-1$

$$v_e[n] = v_e[n-1] + v_i[n] - v_o[n-1]$$
$$v_o[n] = Q(v_e[n])$$

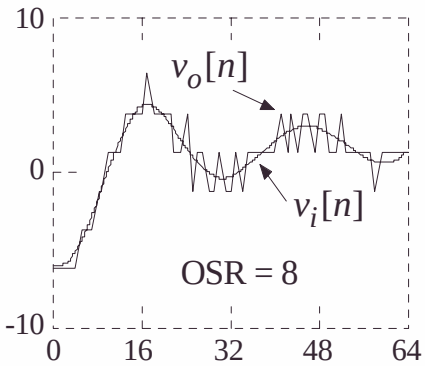
From this, we can show that:

$$v_e[n] = v_i[n] - v_q[n-1]$$
$$v_o[n] = v_i[n] + (v_q[n] - v_q[n-1])$$

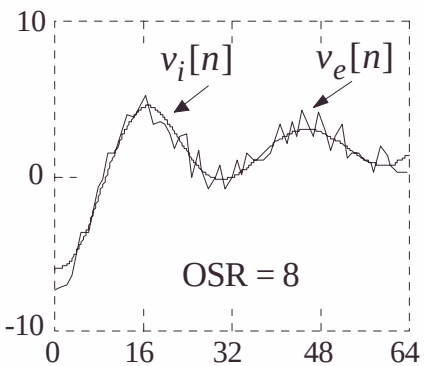
To avoid "overload", the range of $v_i[n]$ must be smaller by q than the quantizer range.



quantization yields $v_q[n]$



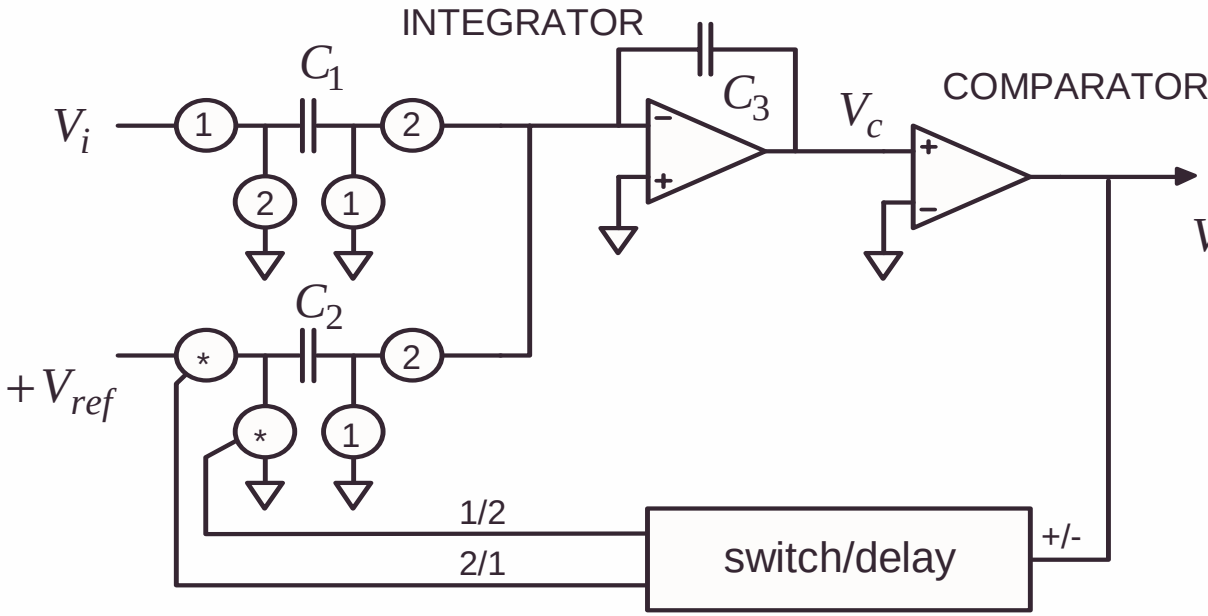
noise-shaping yields $v_o[n]$



integral error signal $v_e[n]$

sample plots from the iteration

1st Order Loop using an SC Modulator



The input SC unit behaves like a high-precision sampler. The feedback controls the polarity of the applied Vref.

$$_c[k] = V_c[k-1] + \left(\frac{C_1}{C_3}\right)\cdot V_i[k-1] - \left(\frac{C_2}{C_3}\right)\cdot V_{ref}\cdot Sgn(V_c[k-1])$$

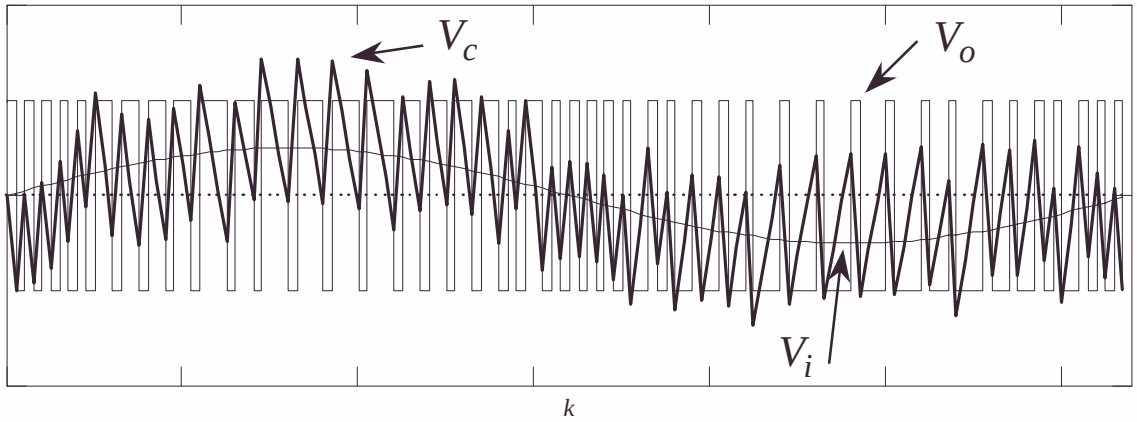
V_i is the analogue input signal
V_{ref} is a DC reference level
V_c is the integrator output. V_o is the digital output

We run this difference equation to see the action of the filter.

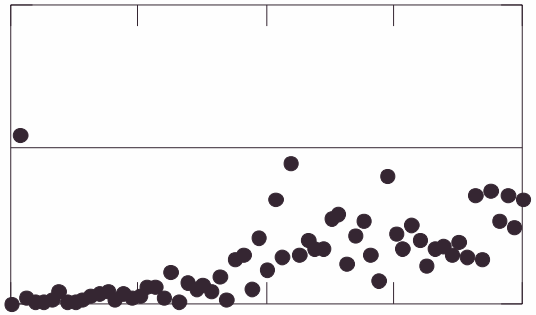
☐ 1st Order Loop Experimental Data

These results are from the difference equation, with all capacitors equal. They show how even a 1-bit quantizer can form *a high-speed binary output* which, after further digital processing in *a decimation filter*, can deliver *high-precision digital low-rate data values*.

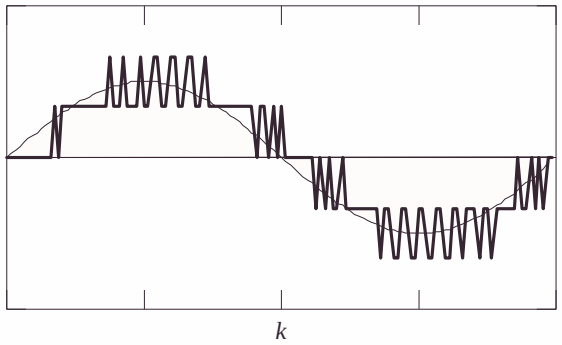
REAL-TIME OPERATION WITH SINE-WAVE INPUT



128-PT FFT OUTPUT SPECTRUM



OUTPUT AFTER 6-POINT AVERAGING



The SC modulator forms the front end of a Δ-Σ ADC.

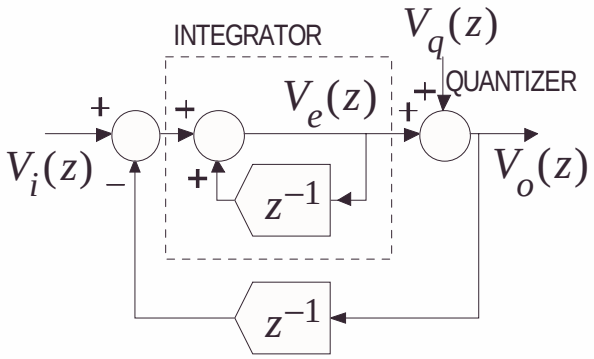
The modulator output is a binary bit-stream. This removes any *linearity* considerations.

One-bit systems are widely used but require high oversampling ratios.

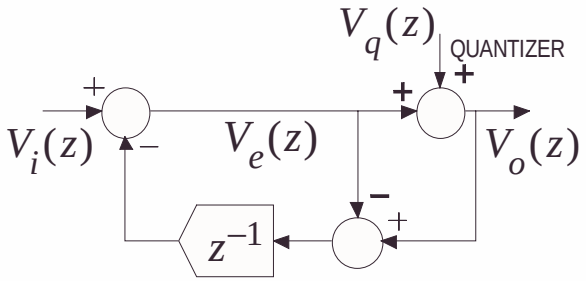
This FFT ← shows evidence of noise-shaping

Unlike simple averaging ←, a decimation filter removes nearly all of the noise.

An Alternative Loop Description



$\Delta-\Sigma$ with Integrator Detail



$\Delta-\Sigma$ Equivalent System

This is the loop as we already described it ←, except that the details of the integrator have been included in the diagram, and this highlights the presence of *two* feedback paths, and of two z^{-1} delay blocks.

We could use just one of these z^{-1} blocks to serve both V_e and V_o in the manner shown ↙

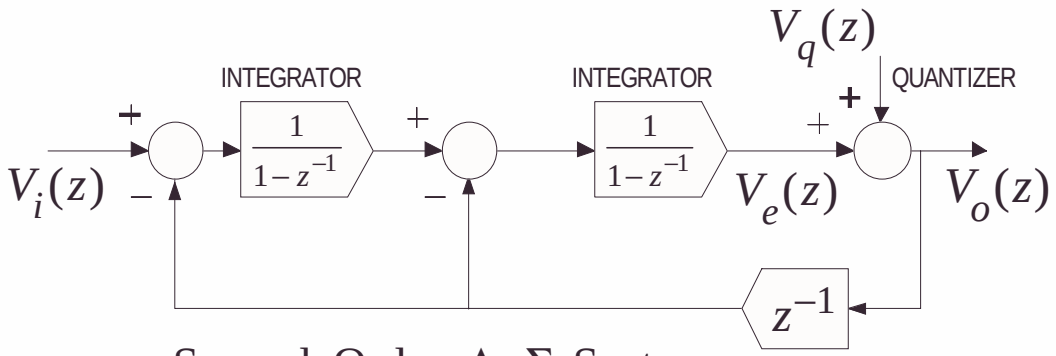
The quantizer injects the quantization error V_q, but the feedback loop removes it again, one step later. Because quantization is the *stripping of bits* from the end of a word, what this diagram describes is *a re-circulation* of those bits so as to include them once more, after a 1-sample delay. The precision is not discarded, rather, the lower bits are held back for re-distribution over a longer time span.

$$v_e[n] = v_i[n] + \{v_e[n-1] - Q(v_e[n-1])\} = v_i[n] + \text{re-circulated bits}$$

It also means that: $\Delta v_e[n] = v_e[n] - v_e[n-1] = v_i[n] - v_o[n-1]$

The error v_e is updated on every cycle to include the latest difference between v_i and v_o. Thus, all deviations of v_o from v_i are accounted for over the long term.

A 2nd Order Δ-Σ System



Second Order Δ-Σ System

In this 2nd Order system, the noise is differentiated twice:

$$_o(z) = V_i(z) + V_q(z)\cdot\left(1-z^{-1}\right)^2$$

This enhances the noise shaping effect. To prove it, we write:

(This is visible on inspection ➔)

$$_o(z) = V_q(z) + \frac{1}{1-z^{-1}}\left\{\frac{1}{1-z^{-1}}\left(V_i(z) - z^{-1}\cdot V_o(z)\right) - z^{-1}\cdot V_o(z)\right\}$$

Noting also that:

$$_o(z) = V_e(z) + V_q(z)$$

the iteration loop can be found to be:

For $n = 0\,..\,N-1$

$$\begin{cases} v_e[n] = v_i[n] + 2v_e[n-1] - v_e[n-2] - 2v_o[n-1] + v_o[n-2] \\ v_o[n] = Q\left(v_e[n]\right) \end{cases}$$

and we can also see that:

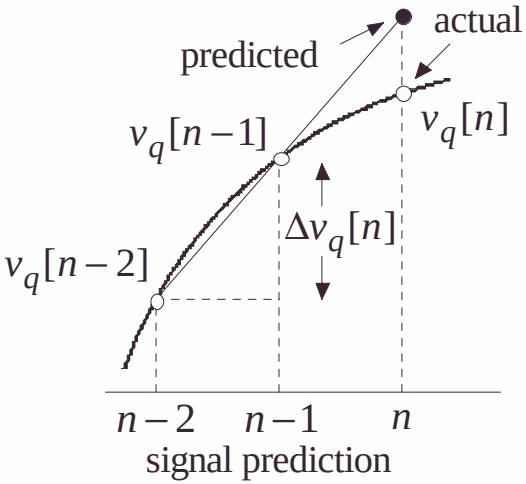
$$v_o[n] = v_i[n] + v_q[n] - 2v_q[n-1] + v_q[n-2]$$

Here, the noise term is a *second difference* of $v_q[n]$:

2nd Order Noise Shaping

Using $y_q[n]$ as the noise term: $v_o[n] = v_i[n] + \{v_q[n] - 2v_q[n-1] + v_q[n-2]\} = v_i[n] + y_q[n]$

Thus: $y_q[n] = v_q[n] - 2v_q[n-1] + v_q[n-2]$ equivalent to ➔ $H_q(z) = \left(1 - z^{-1}\right)^2 = 1 - 2z^{-1} + z^{-2}$



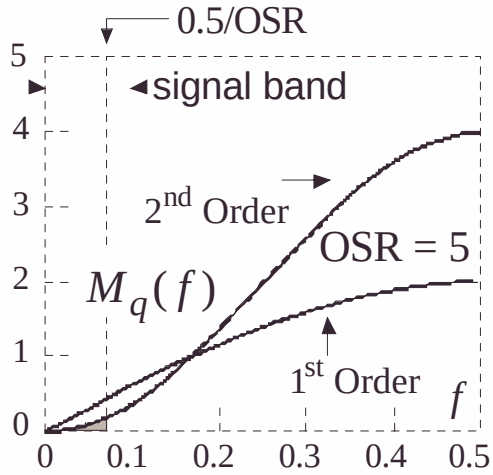
We can write: $y_q[n] = v_q[n] - p[n]$

where: $p[n] = v_q[n-1] + (v_q[n-1] - v_q[n-2])$

This $p[n]$ is a *prediction*. It is just a *linear extrapolation* from the previous two samples ➔.

For a 2nd Order predictor: $M_{q2}(f) = (2 \cdot \sin(\pi f))^2$

and: $A_2 = \int_{-0.5/\mathrm{OSR}}^{+0.5/\mathrm{OSR}} M_{q2}(f)^2 . df$

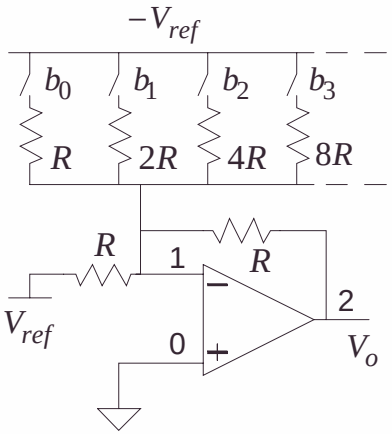


$\Delta - \Sigma$ Noise Filters

1st Order: $A_1 \approx \dfrac{\pi^2}{3 \cdot \mathrm{OSR}^3}$ ➔ 9 dB per OSR doubling

2nd Order: $A_2 \approx \dfrac{\pi^4}{5 \cdot \mathrm{OSR}^5}$ ➔ 15 dB per OSR doubling

A 3-bit Δ-Σ DAC

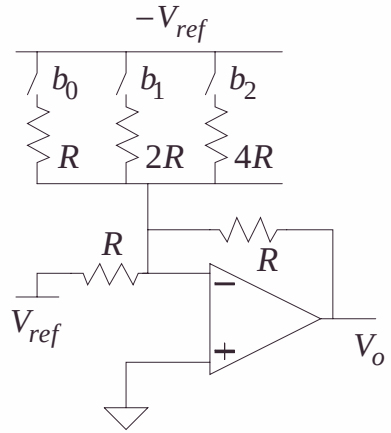


multi-bit bi-polar DAC

The normal method ← becomes problematic as the word length is increased.

The Delta-Sigma approach uses very few bits, resulting in large quantization errors, and depends on suitable shaping of the error noise to make it easy to remove later by filtering.

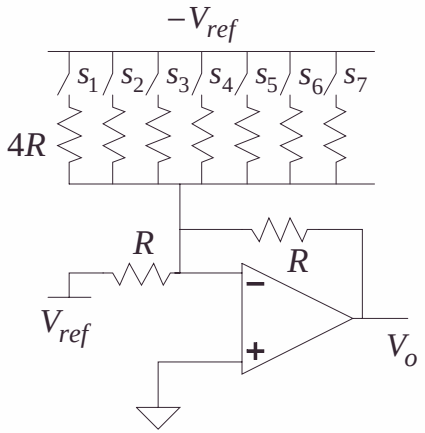
For noise-shaping to be possible, we must first interpolate the data to a far higher rate (an all-digital operation).



a 3-bit bi-polar DAC

Then we run the data through a digital "bit-stripping" Δ–Σ loop, reducing the precision from perhaps 16 bits down to 3-bits. The 3-bit output has large quantization noise but it is suitably shaped. We can apply it to a fast 3-bit DAC ← and the DAC output V_o is then passed through a fairly simple analogue filter to remove the noise.

A 3-bit Thermometer Δ-Σ DAC



a 3-bit thermometer DAC

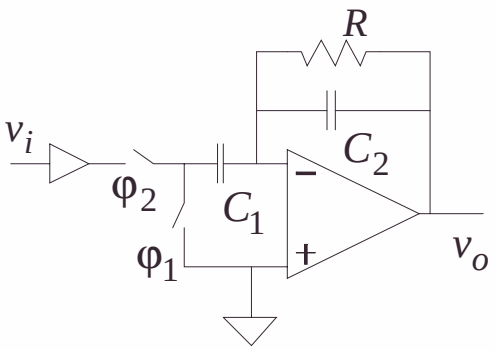
With so few bits, we can consider using a *thermometer* DAC instead. An *N*-bit thermometer DAC uses 2^N–1 identical resistors, instead of *N* binary-weighted resistors. For a 16-bit DAC, we would need 65535 resistors, and that is quite impractical, but it is much easier in the 3-bit case where only 7 resistors are required ← .

We now have 7 switches, and the 3-bit binary word must be digitally converted to a 7-bit word in the manner that this table indicates ↙ .

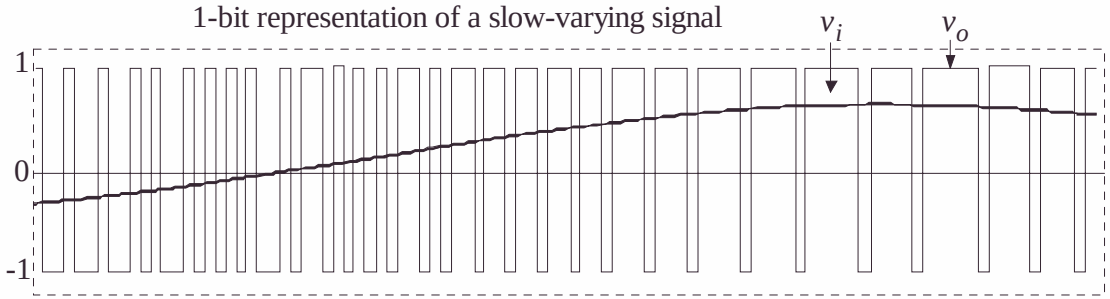
b_0	b_1	b_2	s_1	s_2	s_3	s_4	s_5	s_6	s_7
0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0
0	1	0	1	1	0	0	0	0	0
0	1	1	1	1	1	0	0	0	0
1	0	0	1	1	1	1	0	0	0
1	0	1	1	1	1	1	1	0	0
1	1	0	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1	1

High precision is easier to achieve when all resistors are identical. We can also cause the resistors to change places continuously, in a randomised manner. The effects of small differences in resistor values cancel out over time, assisted by the high rate of oversampling. The short word length, the use of identical resistors, and the randomization of resistor roles, all conspire to yield a high-performance DAC with less emphasis on the analogue design aspects. The resistors can be replaced by semiconductor current sinks, but the advantages of the Δ-Σ method remain unchanged.

☐ A 1-bit Δ-Σ DAC



one-bit DAC concept



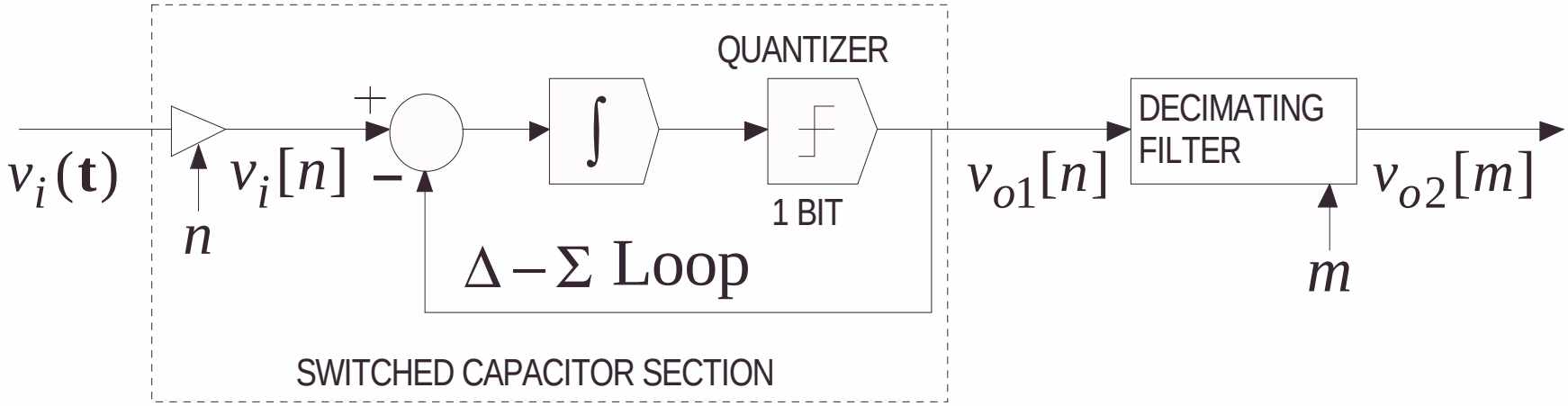
A 1-bit DAC uses a lossy integrator as shown ←

For each new bit-value $v_i[n]=\pm 1$ at the input, capacitor C_1 is first grounded through the switch labelled φ_1, then, after φ_1 has re-opened, φ_2 closes and connects C_1 to $v_i[n]$. This applies a charge of $\pm q=\pm|v_i[n]/C_1|$ to C_2 during each new bit-period. Each pulse of charge causes $v_o[n]$ to take a very small step, either up or down. The output is a stair-case waveform, but the resistor R "knocks off the edges", it filters and smoothes the output waveform. This alone is insufficient, it still requires an analogue filter on the output.

This could be a 1-bit DAC for audio CD use. The input comprises 16-bit music samples from the CD at the 44.1 kHz rate. The data is first interpolated, perhaps by $U = 256$ which yields a rate of 11.2896 MHz. This may be followed by 1-bit quantization in a 2^{nd}-order noise shaper. The 1-bit data enters a 1-bit DAC and the DAC output is filtered by an analogue LP filter, possibly a 3^{rd}-order, or similar.

☐ A 1-bit Δ-Σ ADC

Δ–Σ ADCs combines a small and relatively simple SC analogue section with an all-digital decimating filter that does most of the work.



The SC section converts a slow high-precision analogue input signal $v_i(\mathbf{t})$ into a high-speed binary output sequence $v_{o1}[n]$, with appropriate noise shaping in a Δ–Σ loop.

The digital decimating filter removes the noise and lowers the sample rate to a new rate (index m) consistent with signal bandwidth requirements. This could mean down-sampling by, say, D=256, and may be accomplished in two stages, a decimation by D_2=64 using a sinc^2 filter, and a subsequent decimation by D_1=4 using a conventional LP filter. As the filtering and decimation proceeds, the word length grows longer, reaching a final length of 16 bits or more, at a final rate of > (2 x 20kHz), e.g. 44.1kHz for CD, 48kHz for DAT.