EE 240B – Spring 2019

Advanced Analog Integrated Circuits

Lecture 3: Small-Signal Models

Ali Niknejad

Dept. of EECS

MOSFET Models for Design

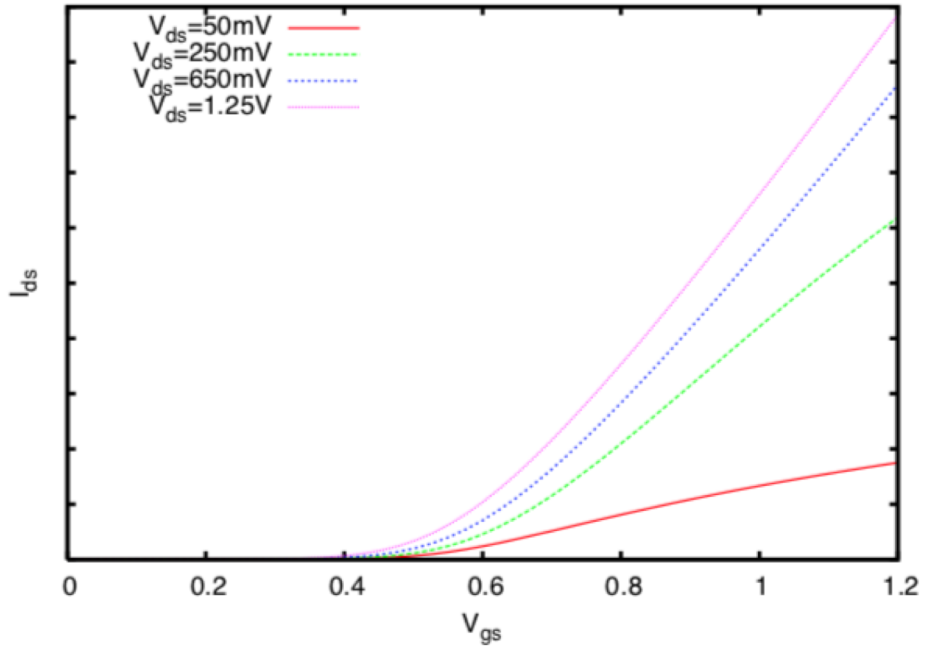
- Hand analysis
  - Square law model for intuition (strong inversion)
  - Small-signal models
- SPICE (BSIM)
  - For verification
  - Device variations
- Challenge
  - Complexity / accuracy tradeoff
  - How can we accurately design when large signal models suitable for hand analysis are off by 50% and more?

Device Variations

- Run-to-run parameter variations:
  - E.g. implant doses, layer thickness
  - Affect V_{TH}, μ, C_{ox}, R_o, …
  - Global variations (wafer to wafer) versus matching (nearby transistors)

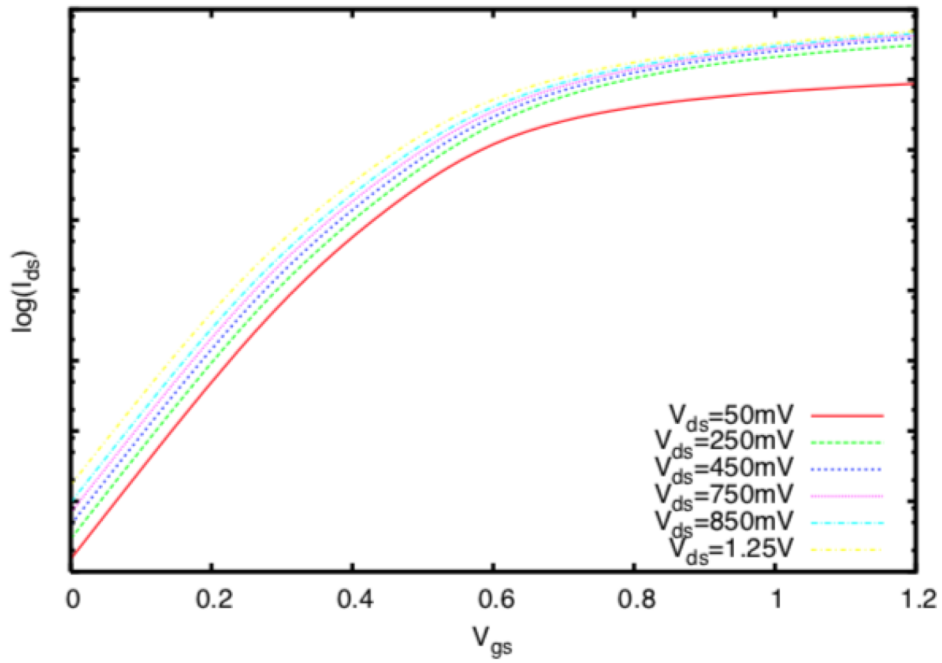
- Nominal / slow / fast parameters
  - E.g. fast: low V_{TH}, high μ, high C_{ox}, low R_o
    - Typical corners TT, SS, FF, SF, FS
  - Combine with supply and temp extremes
  - "PVT" Design: Process/Voltage/Temp
  - Pessimistic but numerically tractable
    → improves chances for working Silicon

I-V Curve



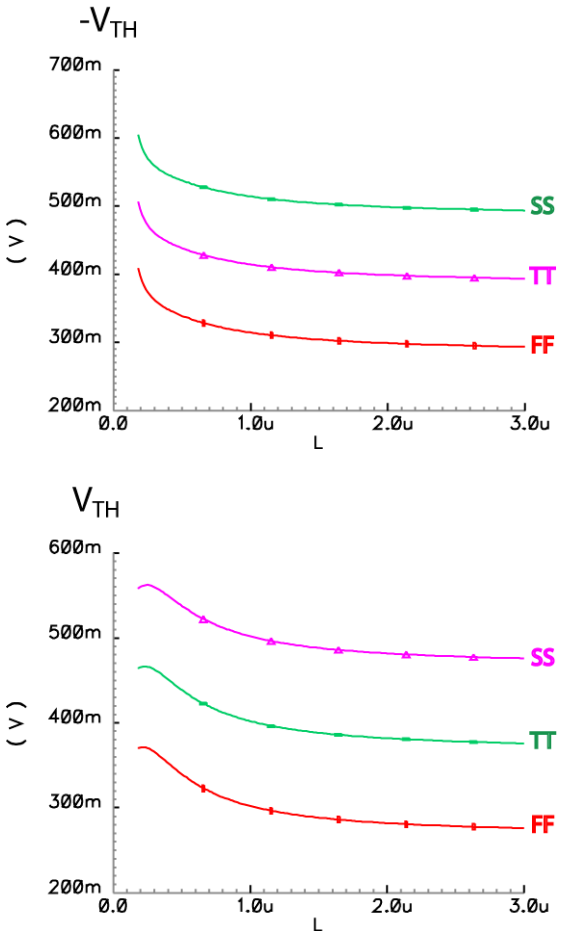
- The I-V curve of a given transistor with fixed dimension (W and L) reveals most of the salient features. For example, a plot of I_{ds} versus V_{gs} for a family of V_{ds} quickly reveals current drive capability. If a device is biased in weak or moderate inversion, then the logarithmic plot is more useful as it expands this regime of operation.

I-V Curve (Log)

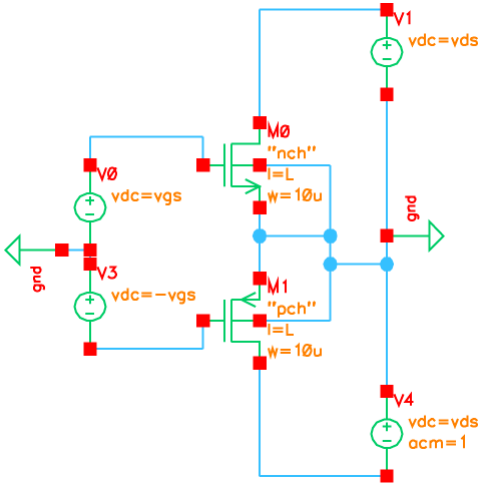


- The leakage currents (sub-threshold slope) is important in analog applications that use the transistor as a switch. If the device cannot be turned off, then it's very difficult to build high precision discrete time circuits.

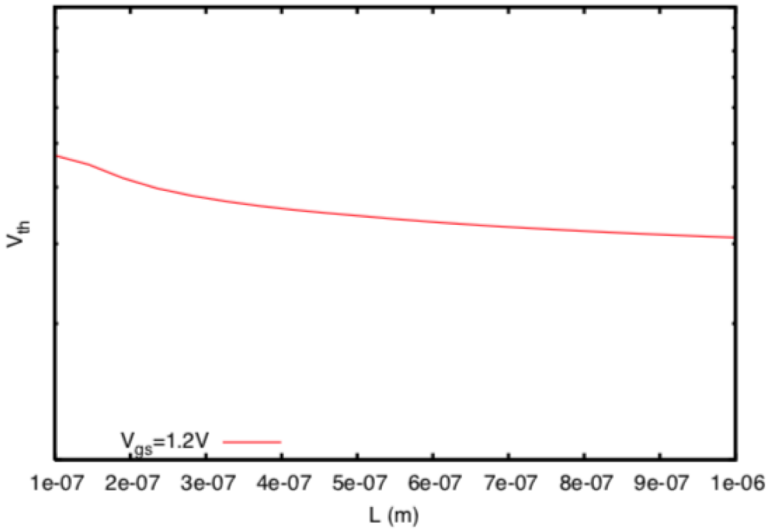
Threshold Voltage V_{TH}



- Strong function of L
- Use long channel for V_{TH} matching
- Process variations
  - Run-to-run
  - How characterize?
  - Slow/nominal/fast
  - Both worst-case & optimistic



Well Proximity Effect (WPE)



- It is well known that CMOS devices exhibit variation in threshold voltage depending on the channel length due to doping variation.
- But modern devices have threshold voltage variation on *W* and even dependence on the distance to the well (WPE) and other structures.

V_{TH} Design Considerations

- Approximate Values (L = 0.5μm)

  V_{THN} = 600mV γ_n=0.5 rt-V

  V_{THP} = -700mV γ_p=0.4 rt-V

- Back-Gate Bias

$$V_T = V_{T0} + \gamma\left(\sqrt{\psi_0 + V_{SB}} - \sqrt{\phi_0}\right) \qquad \phi_0 \approx 2\phi_F$$

  e.g. V_{SB} = 400mV → ΔV_{THN} = 110mV

- Variations: (typical numbers)
  - Run-to-run: +/- 50mV (very process dependent)
  - Device-to-device: σ = 2mV (L > 1μm, common-centroid)
  - Use insensitive designs
    - diff pairs, current mirrors
    - → value of V_{TH} unimportant (if < V_{DD})

Device Parameters for Design

- Region: moderate or strong inversion / saturation
  - Most common region of operation in analog circuits
  - XTR behaves like transconductor: voltage controlled current source
- Key design parameters
  - Large signal
    - Current I_D → power dissipation
    - Minimum V_{DS} → available signal swing
  - Small signal
    - Transconductance g_m → speed / voltage gain
    - Capacitances C_{GS}, C_{GD}, … → speed
    - Output impedance r_o → voltage gain

Low Frequency Model

- A Taylor series expansion of small signal current gives (neglect higher order derivatives)

$$i_{ds} = \frac{\partial I_{ds}}{\partial V_{gs}} v_{gs} + \frac{\partial I_{ds}}{\partial V_{bs}} v_{bs} + \frac{\partial I_{ds}}{\partial V_{ds}} v_{ds}$$

$$i_{ds} = g_m V_{gs} + g_{mb} V_{bs} + g_{ds} V_{ds}$$

- Square law model:

$$g_{m,triode} = \mu C_{ox} \frac{W}{L} \frac{V_{ds}}{\alpha}$$

$$g_m = \mu C_{ox} \frac{W}{L} \frac{(V_{gs} - V_T)}{\alpha}$$

$$g_{ds} = \frac{1}{r_o} = \lambda I_{ds}$$

Transconductance

- Using the square law model we have three equivalent forms for g_m in saturation

$$g_m = \mu C_{ox} \frac{W}{L} \sqrt{\frac{2I_{ds}}{\mu C_{ox} \frac{W}{L}}}$$

$$g_m = \sqrt{2\mu C_{ox} \frac{W}{L} I_{ds}} \propto \sqrt{I_{ds}}$$

$$g_m = \frac{1}{2} \mu C_{ox} \frac{W}{L} (V_{gs} - V_T)^2 \frac{1}{\frac{1}{2}(V_{gs} - V_T)}$$

$$g_m = \frac{2I_{ds}}{V_{gs} - V_T} = \frac{2I_{ds}}{V_{dsat}}$$

Weak Invesion g_m

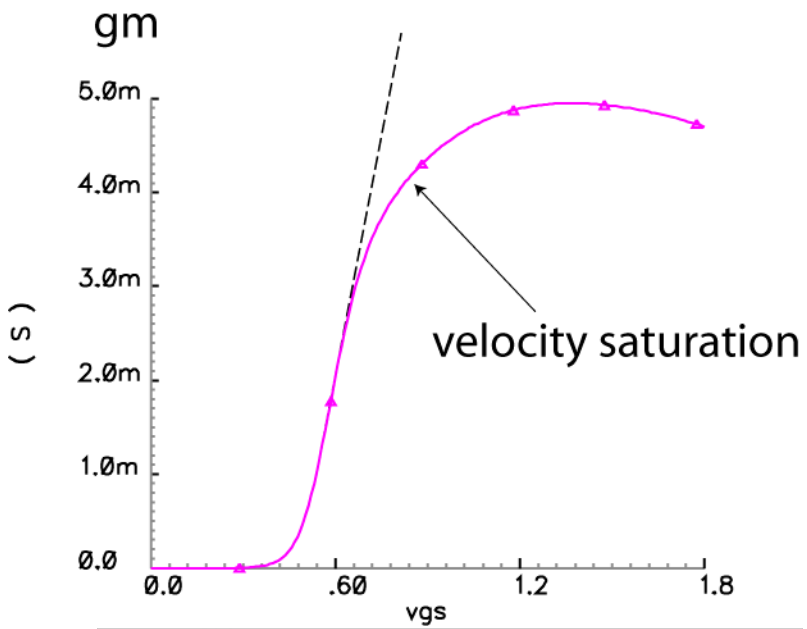
- In weak inversion we have bipolar behavior

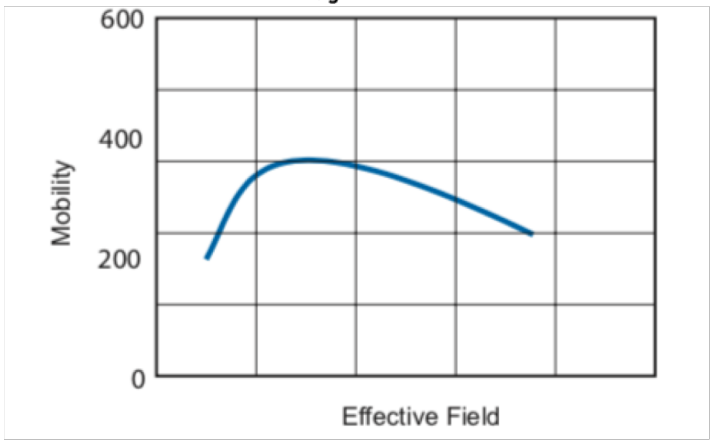
$$I_{ds} \approx \frac{W}{L} I_{ds,0} e^{\frac{q(V_{gs}-V_T)}{nkT}}$$

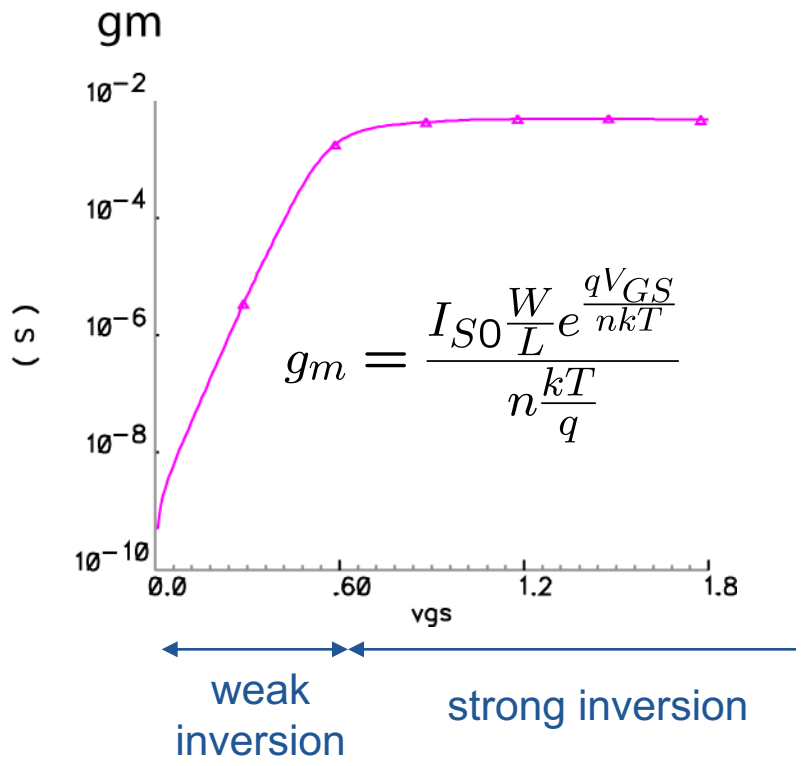
$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}} = \frac{\frac{W}{L} I_{ds,0} e^{\frac{q(V_{gs}-V_T)}{nkT}}}{n\frac{kT}{q}}$$

$$g_m = \frac{I_{DS}}{n\frac{kT}{q}} \propto I_{DS}$$

Transconductance







$$g_{m(sat)} \approx \mu C_{ox}\frac{W}{L}(V_{GS} - V_T)$$

Transconductance (cont)

- The transconductance increases linearly with $V_{gs} - V_T$ but only as the square root of I_{ds}. Compare this to a BJT that has transconductance proportional to current.
- In fact, we have very similar forms for g_m

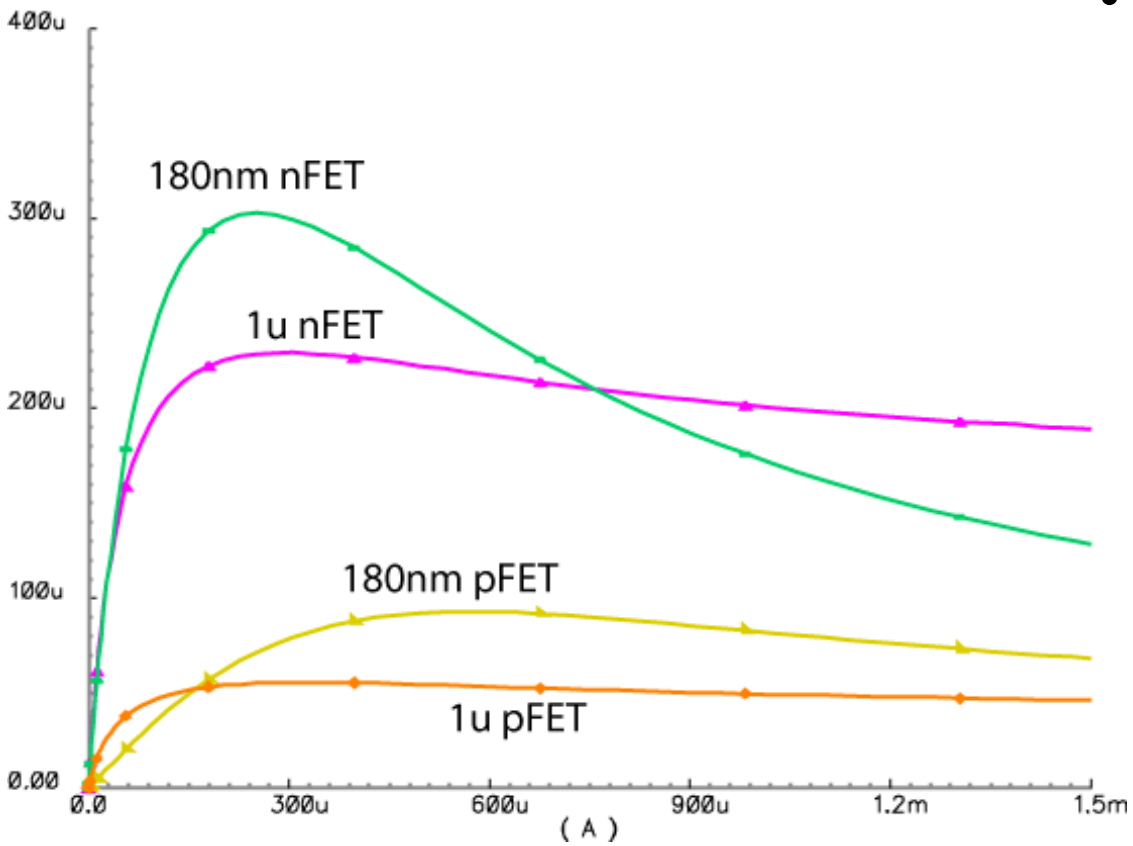
$$g_m^{\text{FET}} = \frac{2I_{ds}}{V_{dsat}} \qquad g_m^{\text{BJT}} = \frac{I_C}{V_t}$$

- Since $V_{dsat} >> V_t$, the BJT has larger transconductance for equal current.
- Why can't we make $V_{dsat} \sim V_t$?

Subthreshold Again…

- In fact, we can make $V_{gs} - V_t$ very small (*or even negative*) and operate in the moderate or sub-threshold region.
- Then the transconductance is the same as a BJT (except the non-ideality *n* factor).
- But as we shall see, the transistor f_T drops dramatically if we operate in this region. Thus we typically prefer moderate or strong inversion for high-speed applications.

μC_{ox}



- Square law:

$$\mu C_{ox} = \frac{g_m^2}{2\frac{W}{L} I_D}$$

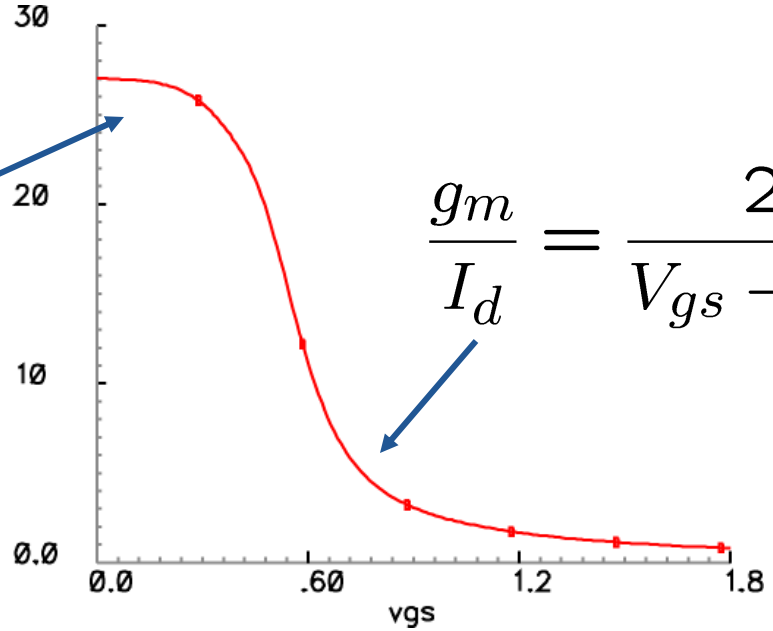
Extracted values strong function of I_D

- Low I_D → weak inversion
- Large I_D → mobility reduction

- Do not use μC_{ox} for design!

Transconductor Efficiency

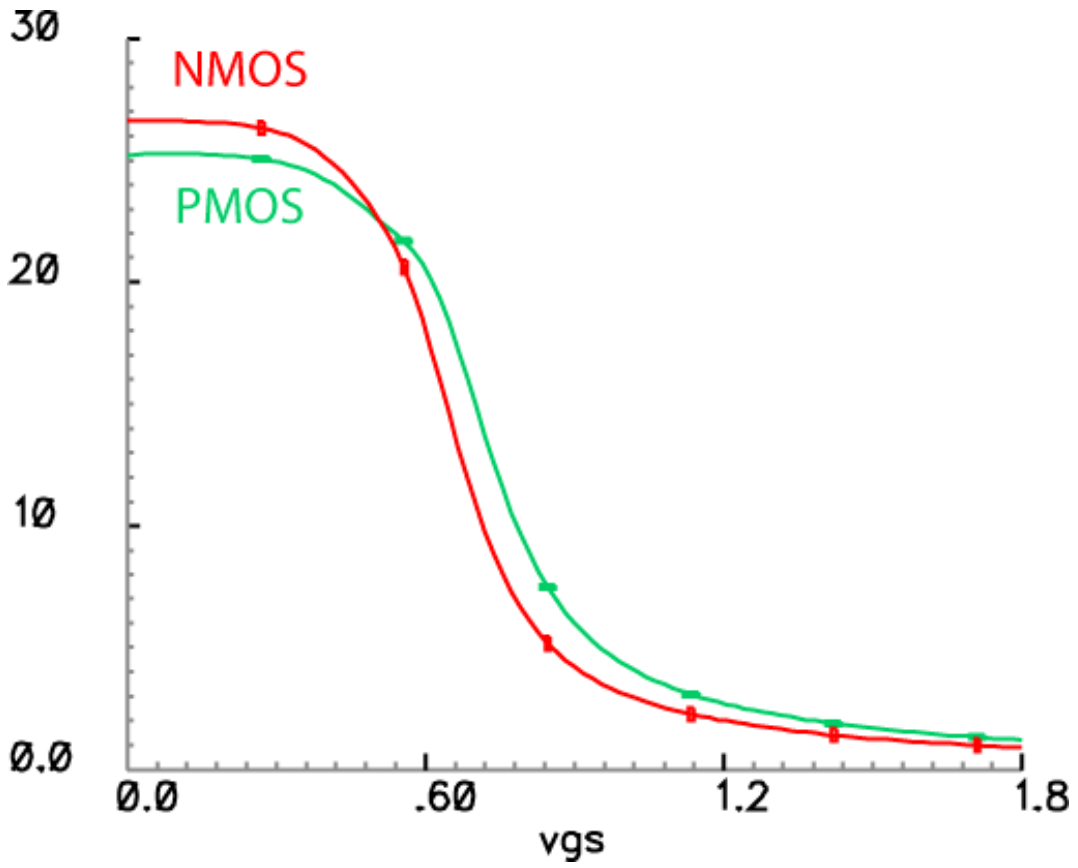
$$\frac{g_m}{I_d} = \frac{1}{n\frac{kT}{q}}$$



$$\frac{g_m}{I_d} = \frac{2}{V_{gs} - V_T} = \frac{2}{V_{dsat}}$$

- A good metric for a transistor is the transconductance normalized to the DC current. Since the power dissipation is determined by and large by the DC current, we'd like to get the most "bang" for the "buck".
- From this perspective, the weak and moderate inversion region is the optimal place to operate.

Efficiency g_m/I_D



- High efficiency is good for low power
- Higher g_m/I_D at low V_{GS}
- Approaches BJT for $V_{GS} < V_{TH}$

  $g_m/I_C = 1/V_t \sim 40\ V^{-1}$
- NMOS / PMOS about same

Efficiency g_m/I_D

- **Let's define**

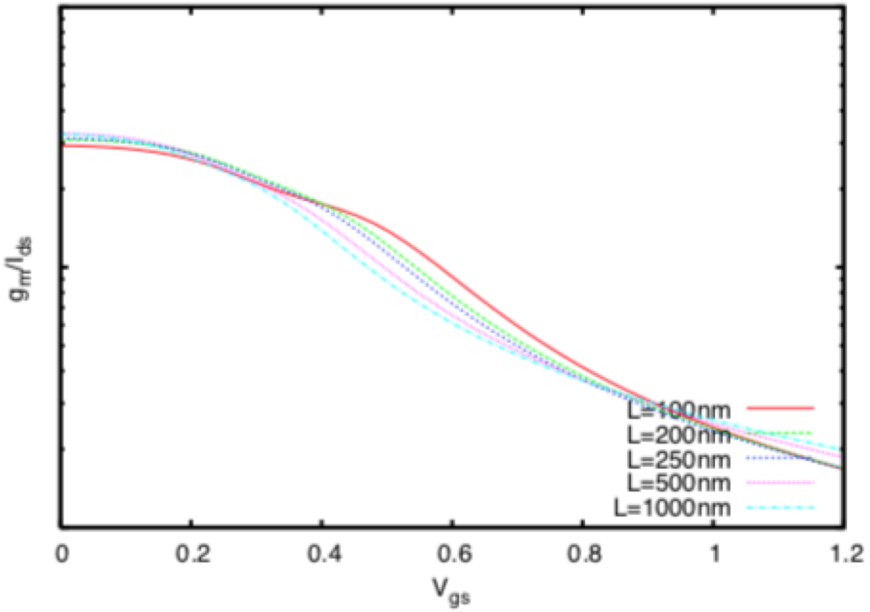
$$V^* = \frac{2I_D}{g_m} \quad \Leftrightarrow \quad \frac{g_m}{I_D} = \frac{2}{V^*}$$

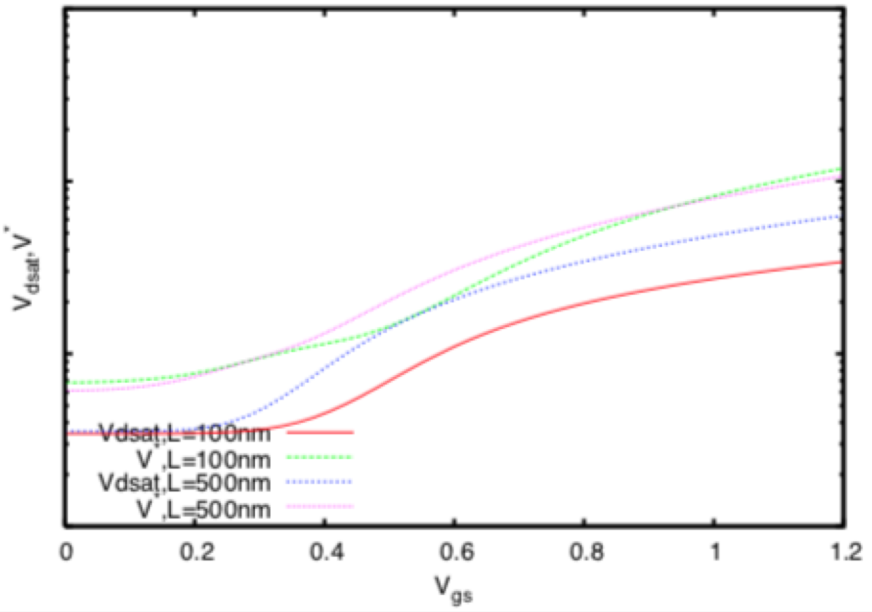
e.g. V* = 200mV → g_m/I_D = 10 V^{-1}

- **Square-law devices: $V^* = V_{GS}-V_{TH} = V_{dsat}$**

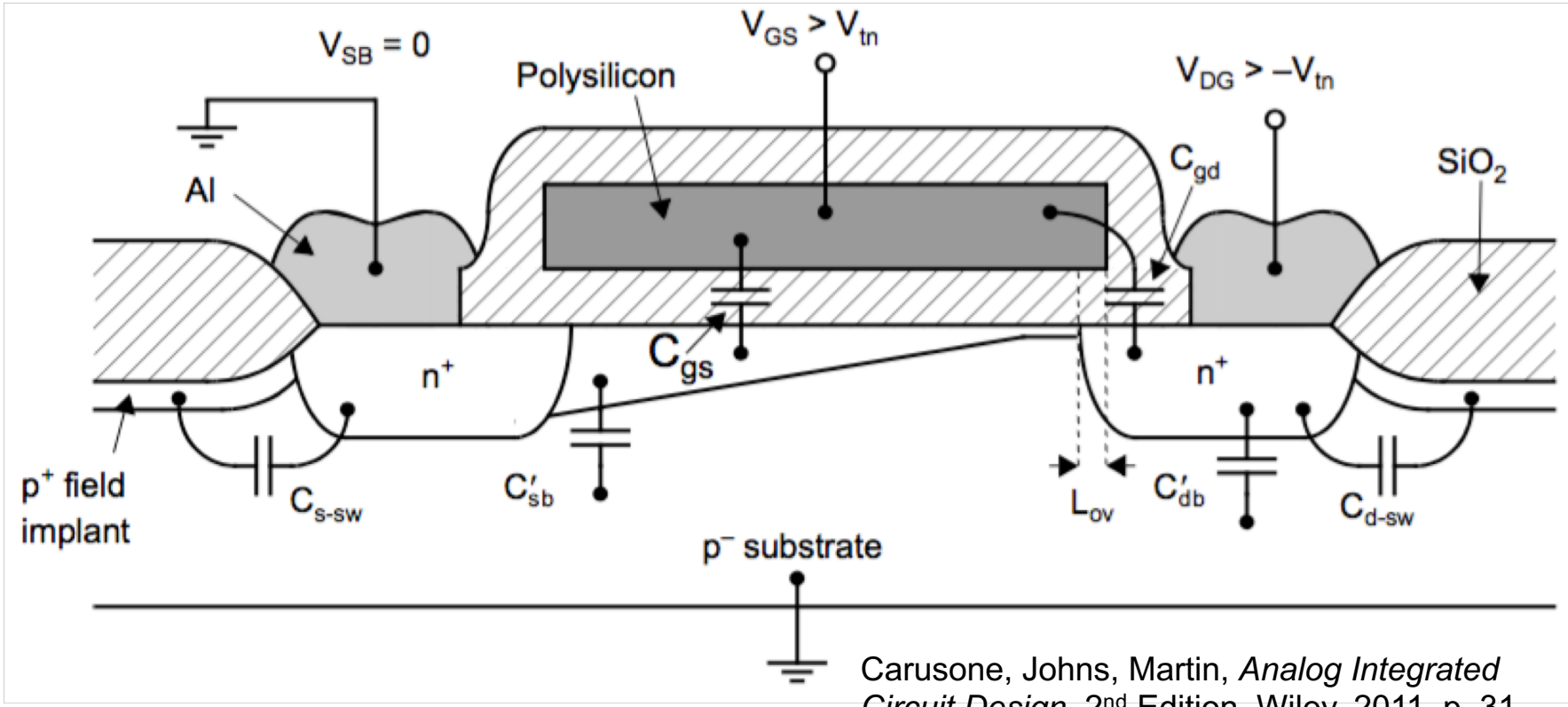
$$\text{Square law :} \qquad g_m = \frac{2I_D}{V_{GS}-V_{TH}} = \frac{2I_D}{V^*}$$

Effective Overdrive V^*





Device Capacitance



Carusone, Johns, Martin, *Analog Integrated Circuit Design,* 2nd Edition, Wiley, 2011, p. 31.

SPICE Charge Model

- **Charge conservation**

- **MOSFET:**
  - **4 terminals: S, G, D, B → 16 derivatives**
  - **4 charges: $Q_S + Q_G + Q_D + Q_B = 0$ (3 free variables)**
  - **3 independent voltages: V_{GS}, V_{DS}, V_{SB}**
  - **9 derivatives: $C_{ij} = dQ_i / dV_j$, e.g. $C_{G,GS} \sim C_{GS}$**
  - **C_{ij} != C_{ji}**

Ref: HSPICE manual, "Introduction to Transcapacitance", pp. 15:42, Metasoft, 1996.

Why C_{jk} is negative ?

- **Increasing the body voltage increases the threshold voltage of an NMOS, so the gate current should be negative (*positive charges leaving the gate*)**
  - $i_G = C_{gg}\frac{dV_{GS}}{dt} + C_{gd}\frac{dV_{DS}}{dt} + C_{gb}\frac{dV_{BS}}{dt}$
- **If this bothers you (it does bother a lot of people), define things so that it becomes positive**
  - $i_G = C_{gg}\frac{dV_{GS}}{dt} - C_{gd}\frac{dV_{DS}}{dt} - C_{gb}\frac{dV_{BS}}{dt}$

Why C_{ij} != C_{ji} (Non Reciprocity)

- **Suppose you vary the gate voltage and hold the drain voltage constant and observe charge flowing into gate and drain (saturation)**
  - **Charge flows into gate since inversion layer is modified. Drain is isolated and no charge flows into the drain…**
- **Now hold gate constant and vary the drain**
  - **The gate charge does change due to DIBL, or drain-induced barrier lowering.**

Non-Reciprocity

Charge Partition

- **50/50 Fake but convenient**
- **0/100 Also fake but convenient**
- **40/60 Reality but not fixed**
- **Partition depends on region of operation and changes smoothly from 50/50 (triode) to 40/60 (saturation)**

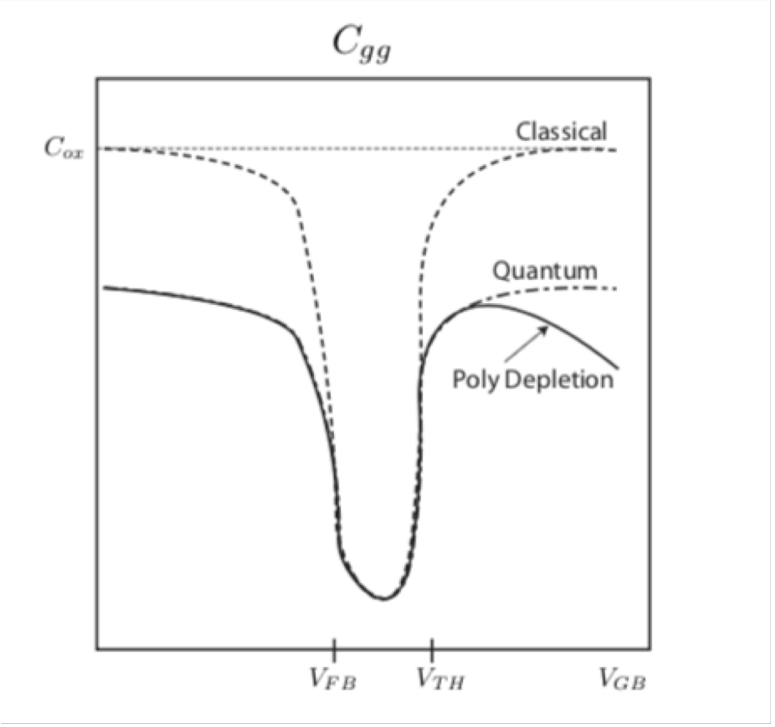
Small Signal Capacitances

	Subthreshold	Triode	Saturation
C_{GS}	C_{ol}	$C_{GC}/2 + C_{ol}$	$2/3\ C_{GC} + C_{ol}$
C_{GD}	C_{ol}	$C_{GC}/2 + C_{ol}$	C_{ol}
C_{GB}	$C_{GC} \parallel C_{CB}$	0	0
C_{SB}	C_{jSB}	$C_{jsB} + C_{CB}/2$	$C_{jsB} + 2/3\ C_{CB}$
C_{DB}	C_{jDB}	$C_{jDB} + C_{CB}/2$	C_{jDB}

$$C_{GC} = C_{ox}WL$$

$$C_{CB} = \frac{\varepsilon_{Si}}{x_d}WL$$

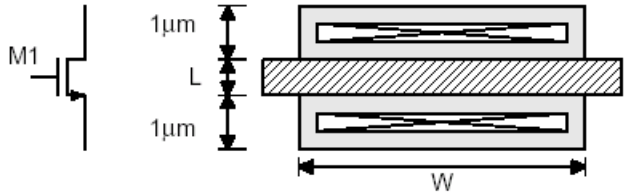
Non-Linear Capacitance



- The MOS capacitor of modern devices does not follow classical equations due to polysilicon depletion and quantum effects ...

Impact of Layout

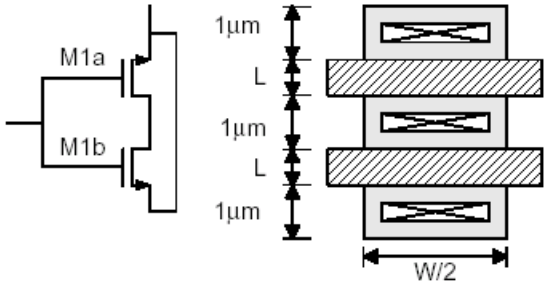
Individual devices:



AS = AD =1µm * W
PS = PD = 2µm + W
e.g. NMOS, W=20µm, V_{sb}=0V
$C_{sb} = C_{db}$ = 28fF

HSPICE geo = 0 (default)

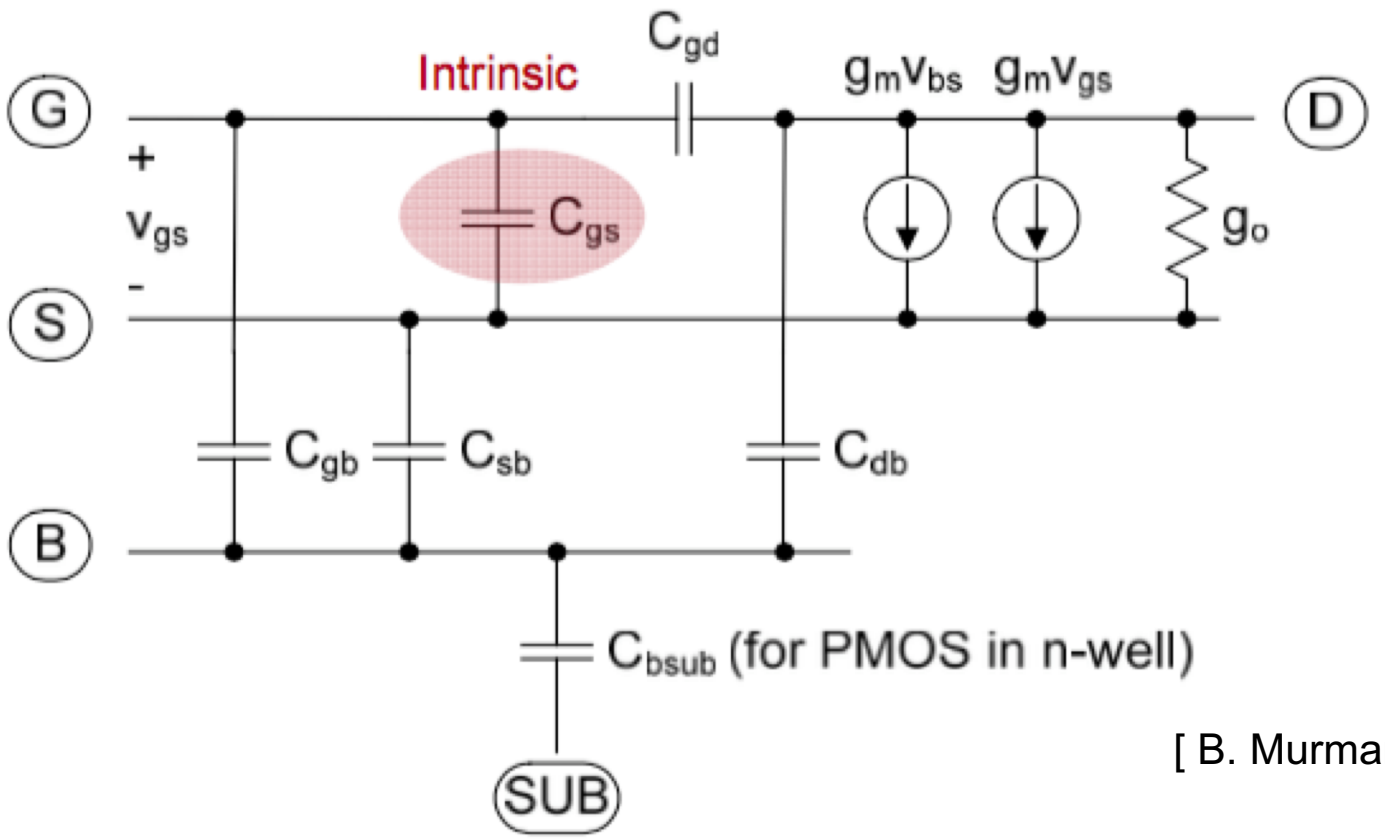
Wide devices consisting of multiple individual ones wired in parallel:



AS = 1µm * W
PS = 4µm + W
AD = 1µm * W/2
PD = 2µm
e.g. NMOS, W=20µm, V_{sb}=0V
C_{sb} = 29fF
C_{db} = 10fF

HSPICE geo = 3

“Complete” Small Signal Model



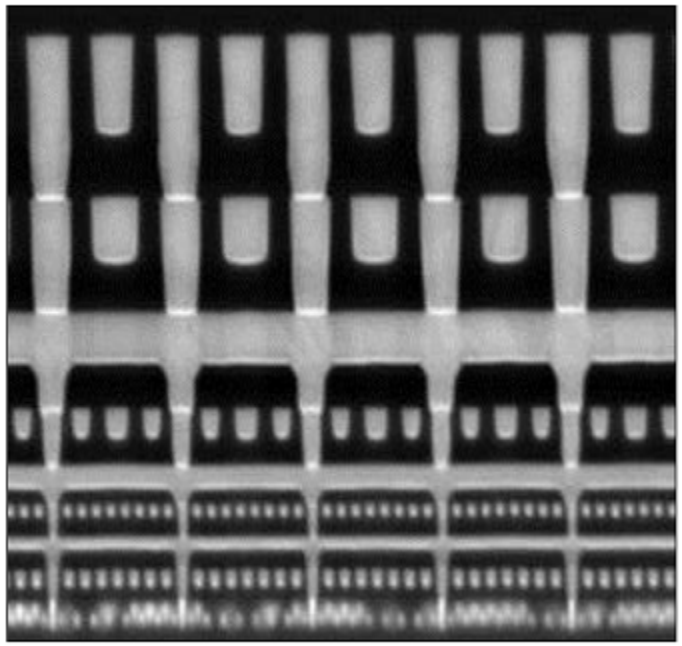
[B. Murmann]

$$C_{gg} = C_{gs} + C_{gb} + C_{gd} \qquad C_{dd} = C_{db} + C_{gd}$$

To Make Matters Worse…

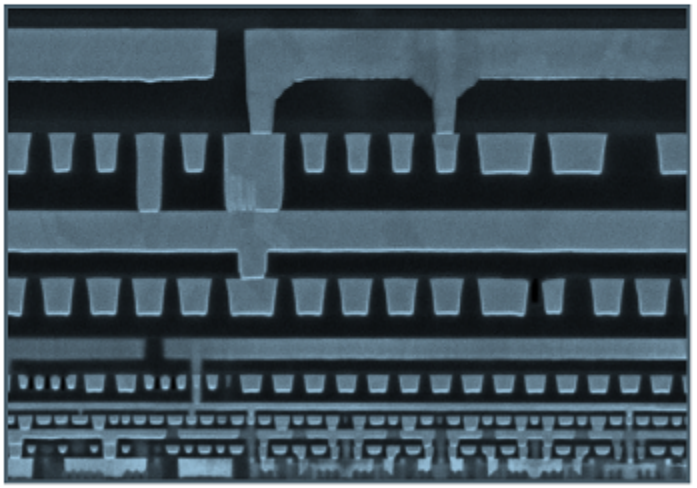
Interconnects

22 nm Process



80 nm minimum pitch

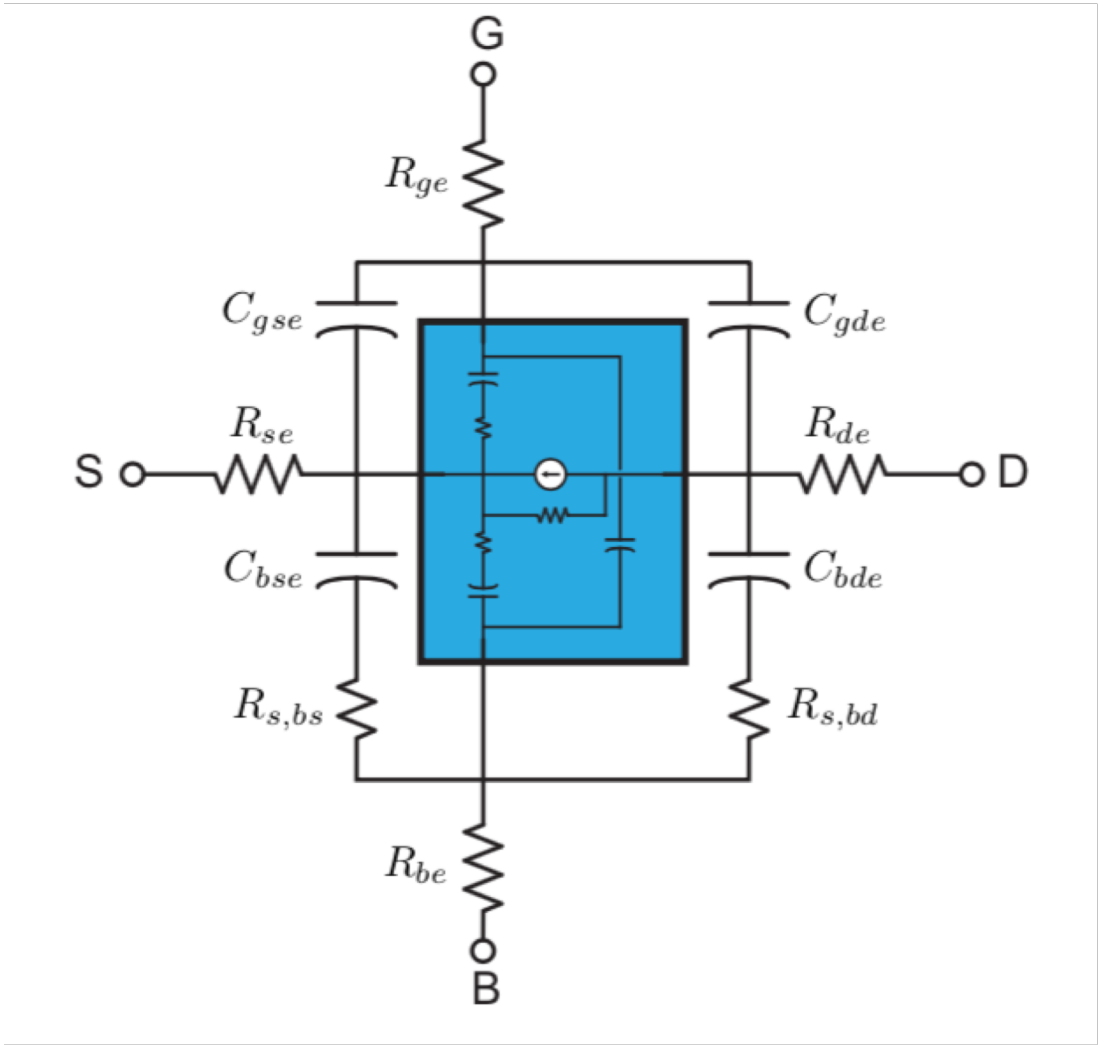
14 nm Process



52 nm (0.65x) minimum pitch

Image from Intel/www.legitreviews.com

“More Complete” SS Model



High Frequency Figures of Merit

- Unity current-gain bandwidth

$$\omega_T = \frac{g_m}{C_{gs} + C_{gd}}$$

$$\omega_T = \frac{3}{2}\frac{\mu V_{dsat}}{L^2} = \frac{3}{2}\omega_0$$ (Long channel model)

- This is related to the channel transit time: $\tau_0 = 1/\omega_0$
- For degenerate short channel device

$$\omega_T = \frac{3}{2}\frac{\nu_{sat}}{L} = \frac{3}{2}\frac{1}{\tau_{sat}}$$

Why f_T

- Since $i_d = g_m v_{gs}$, and $v_{gs} j\omega(C_{gs} + C_{gd}) = i_s$, taking the ratio we have

$$A_i = \frac{i_d}{i_s} = \frac{g_m}{j\omega(C_{gs} + C_{gd})} \tag{4}$$

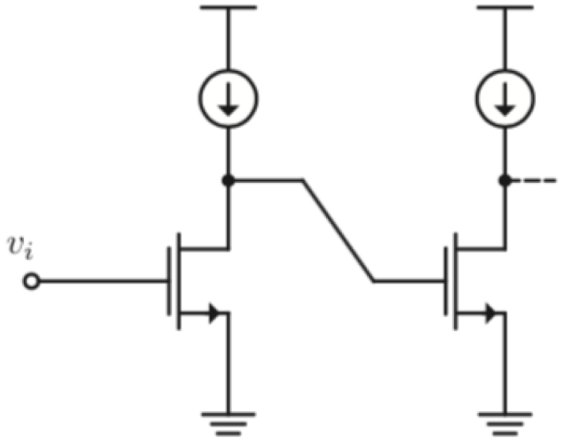
  Solving for $|A_i = 1|$, we arrive at the unity gain frequency

$$\omega_T = 2\pi f_T = \frac{g_m}{C_{gs} + C_{gd}} \tag{5}$$

- It is not at first obvious why f_T plays such an important role in analog (and digital) circuits. To see this, consider the simple cascade amplifier, where a single stage amplifier drives an identical copy. The gain of this first stage is given by the intrinsic gain of the amplifier, and the 3-dB bandwidth is limited by the pole at the high impedance node

$$\omega_0 = \frac{1}{r_{o,1}((1 + |A_2|)C_{gs,2} + C_{d1,tot})} \tag{6}$$

Cascade Bandwidth



- Here $C_{d1,tot}$ is the total drain capacitance ($C_{gd} + C_{ds} + C_{wire} + \cdots$) and the effect of Miller multiplication is captured by boosting the second stage input capacitance by the gain A_2. If the second stage is a cascode with a small voltage gain between the gate/drain, then A_2 can be made smaller than unity to minimize its impact. So if we neglect the Miller effect, we have

$$\omega_0 = \frac{1}{r_o(C_{gs} + C_{d,tot})} \tag{7}$$

Gain/Bandwidth Product

- In most applications, we intend to embed this amplifier in a feedback loop, so the gain-bandwidth product is of interest.
- For a feedback system, we know that we can approximately trade gain for bandwidth. So if we place the amplifiers in a feedback loop, we have

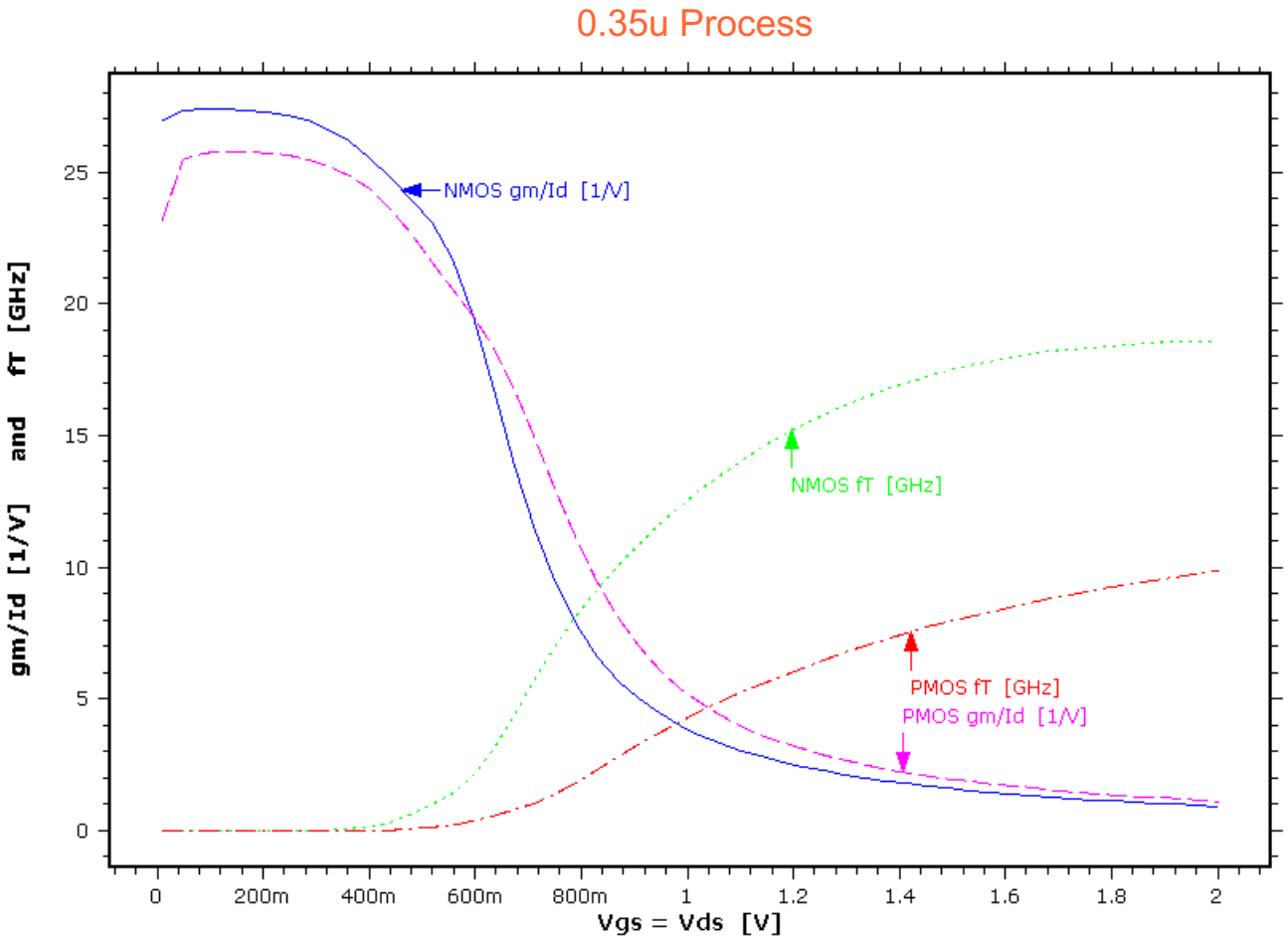
$$Gain = A_0 \tag{8}$$

$$BW = \omega_0 \tag{9}$$

$$Gain \times BW = A_0\omega_0 < g_m r_o \frac{1}{r_o(C_{gs} + C_{d,tot})} \tag{10}$$

$$= \frac{g_m}{C_{gs} + C_{d,tot}} \approx \omega_T \tag{11}$$

Efficiency g_m/I_D versus f_T



Speed-Efficiency Tradeoff

NMOS faster than PMOS

Weak Inversion Freq Response

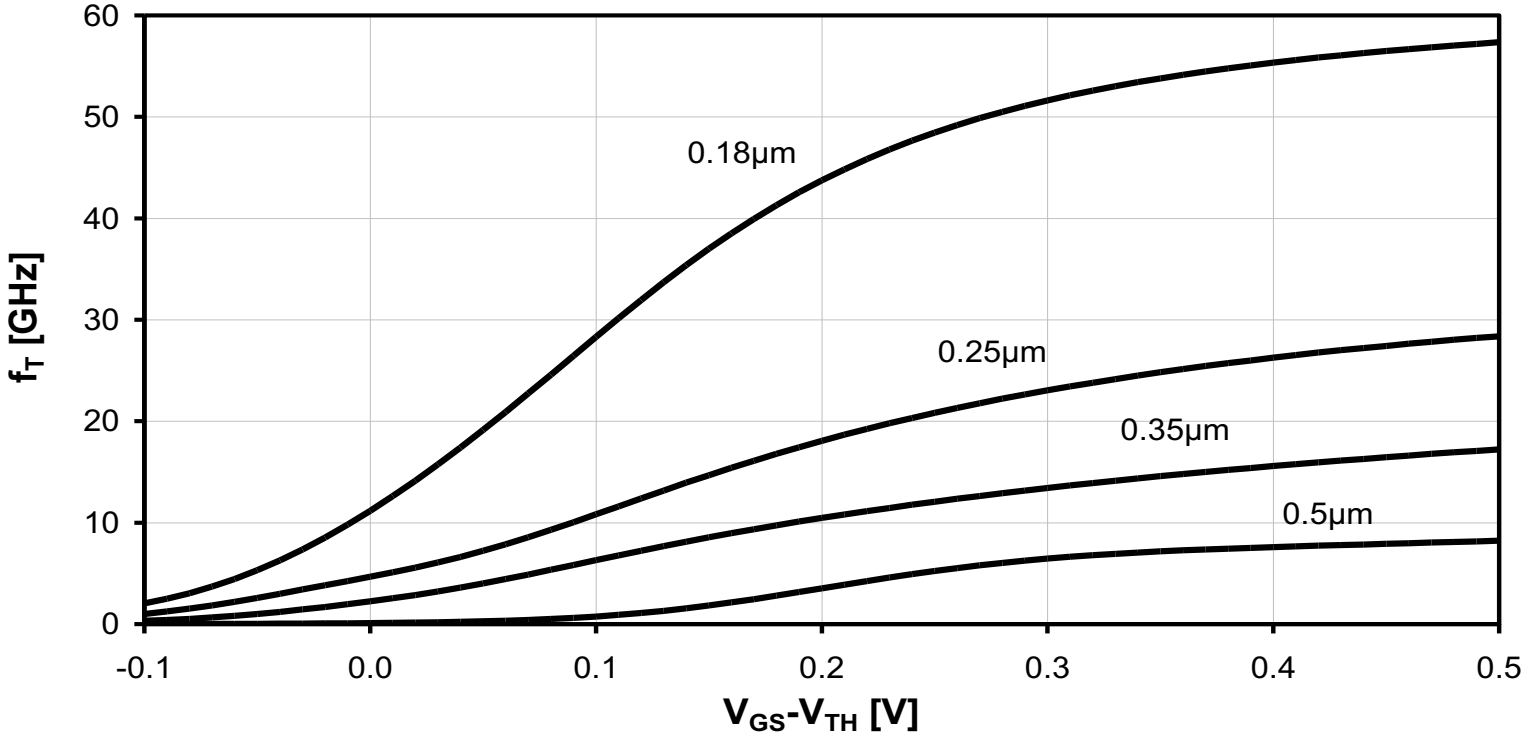
- The gate capacitance in weak inversion is given by

$$C_{gb} = C_{ox}\frac{\gamma}{2\sqrt{\gamma^2/4 + V_{GB} - V_{FB}}}$$

$$\omega_T = \frac{\mu\frac{kT}{q}}{L^2}\left(\frac{I_{DS}}{I_M}\right)$$

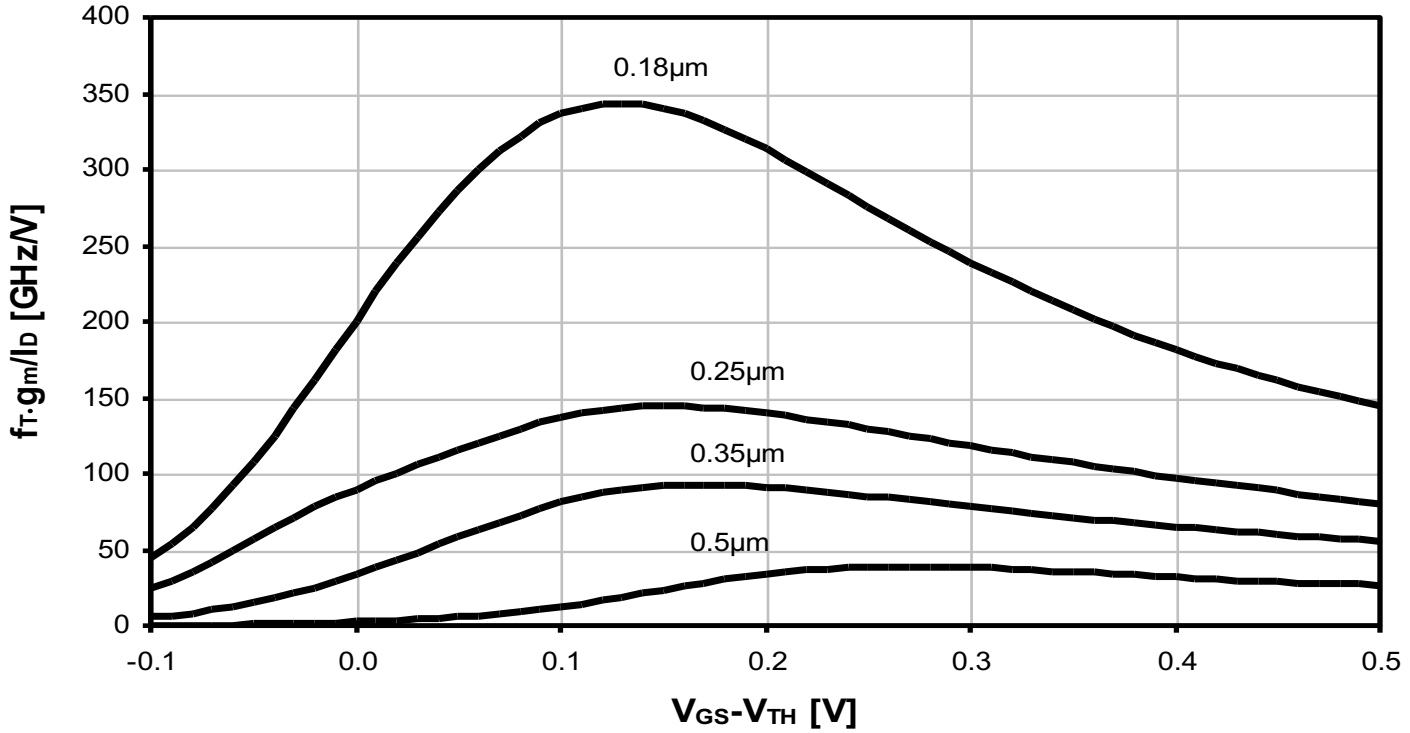
- I_M is the maximum achievable current in weak inversion so the factor ( ) < 1

Device Scaling



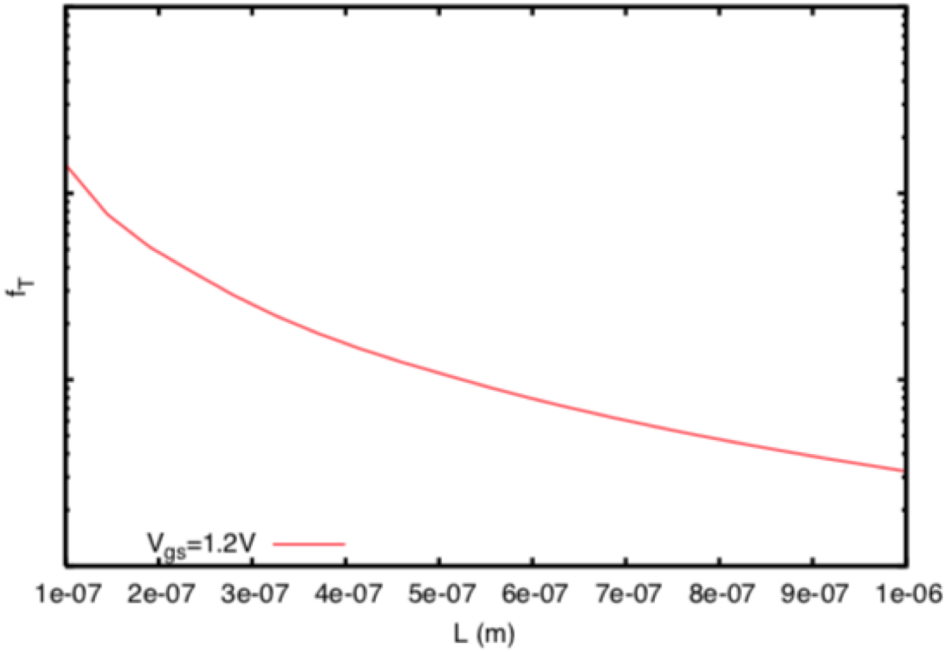
Short channel devices are significantly faster!

Device Figure-of-Merit



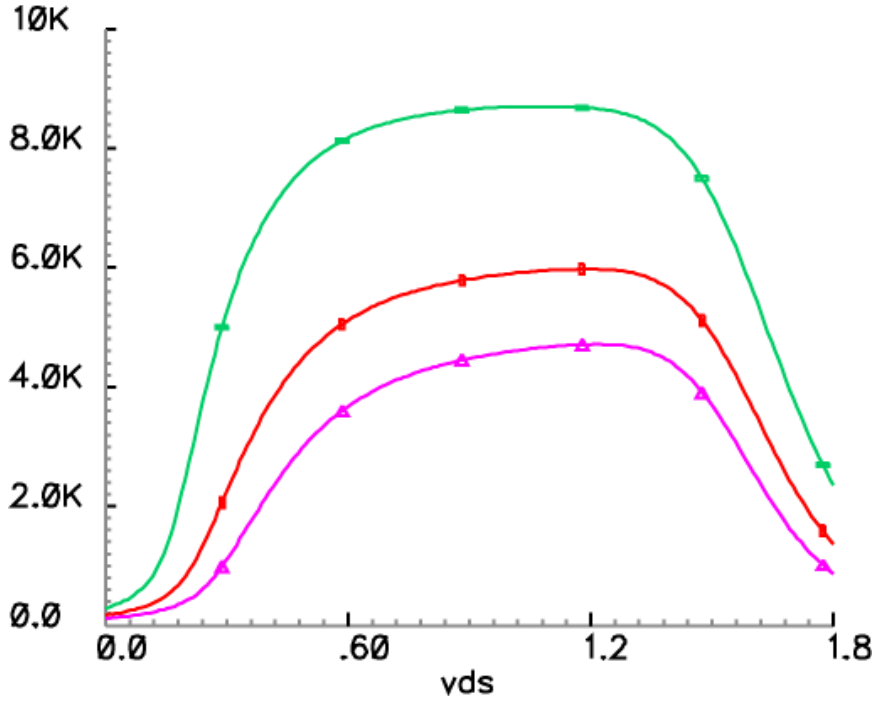
Peak performance for low V_{GS}-V_{TH} (V*)

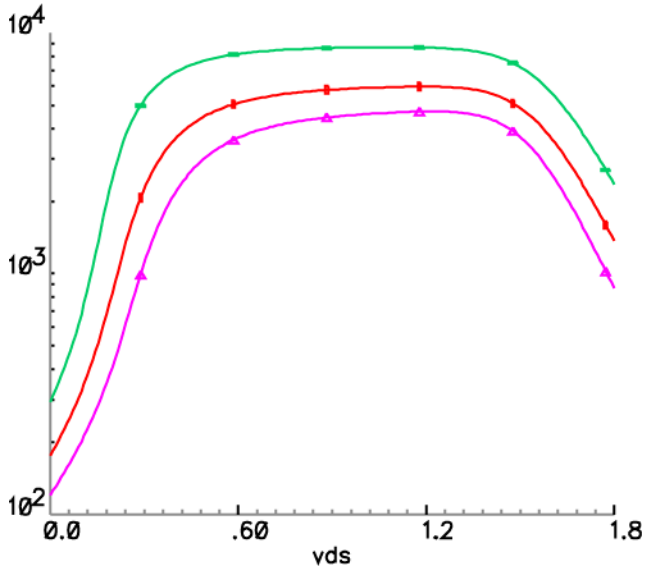
f_T vs Dimension in Same Node



- **Going to 2X or 3X channel length has a lot of benefits to analog specifications but we take a 10X hit in f_T**

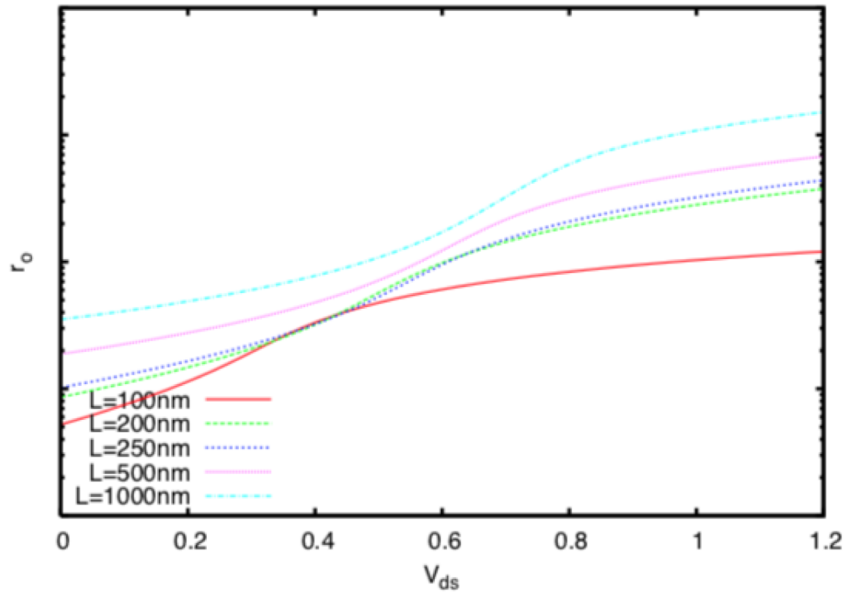
Output Resistance r_o





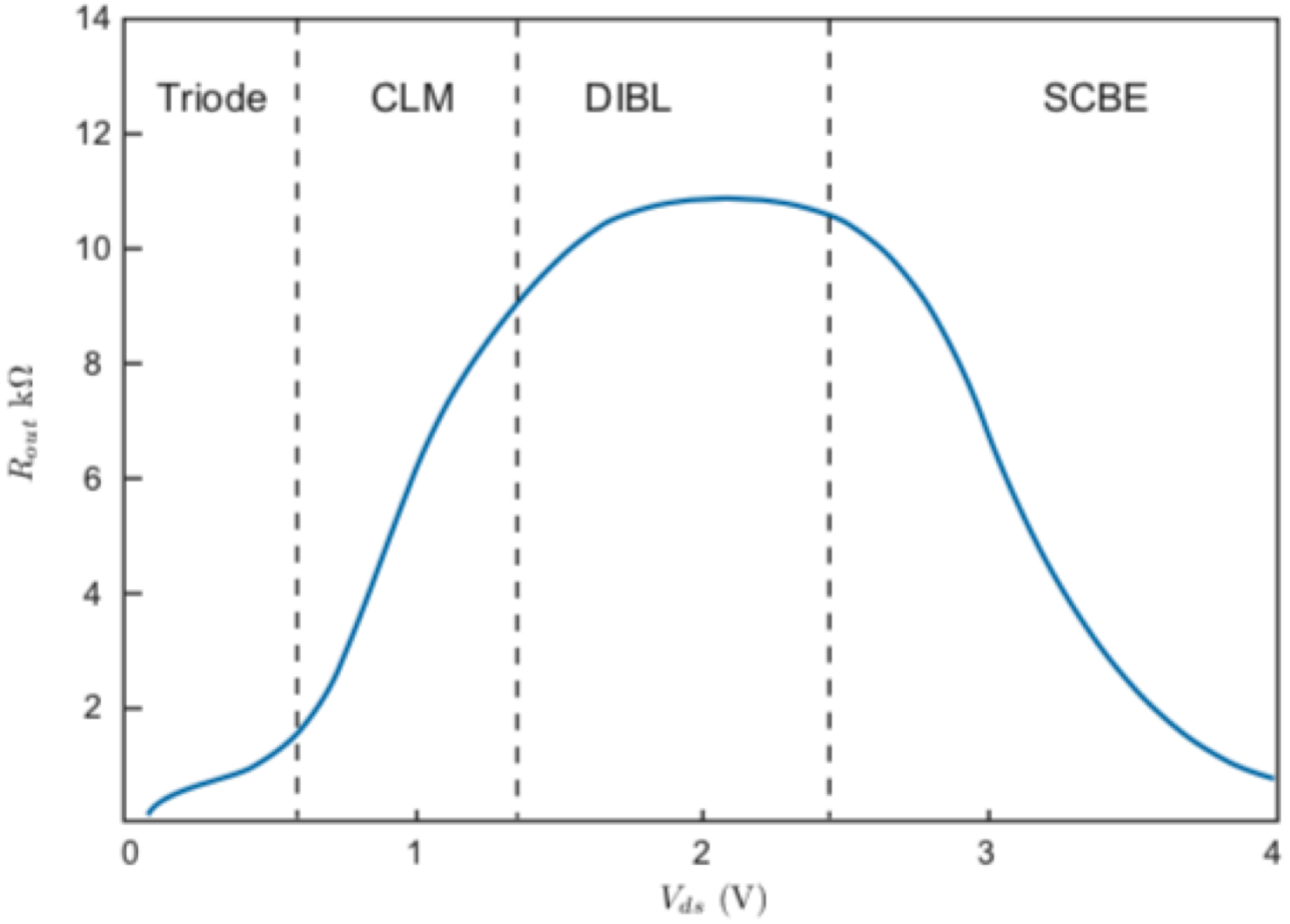
Hopeless to model this with a simple equation
(e.g. $g_{ds} = \lambda I_D$)

Output Resistance vs L

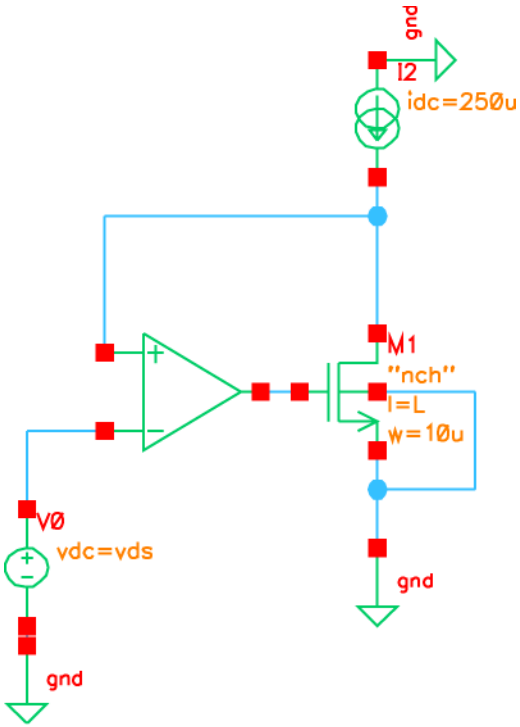


- The drain still modulates the depletion region width, which indirectly impacts the device through channel length modulation (CLM).
- In short channel devices, both because of the increase in the relative magnitude of the change in channel length δL relative to L, but also due to Drain Induced Barrier Lowering (DIBL).

Why is Output Resistance Complicated?



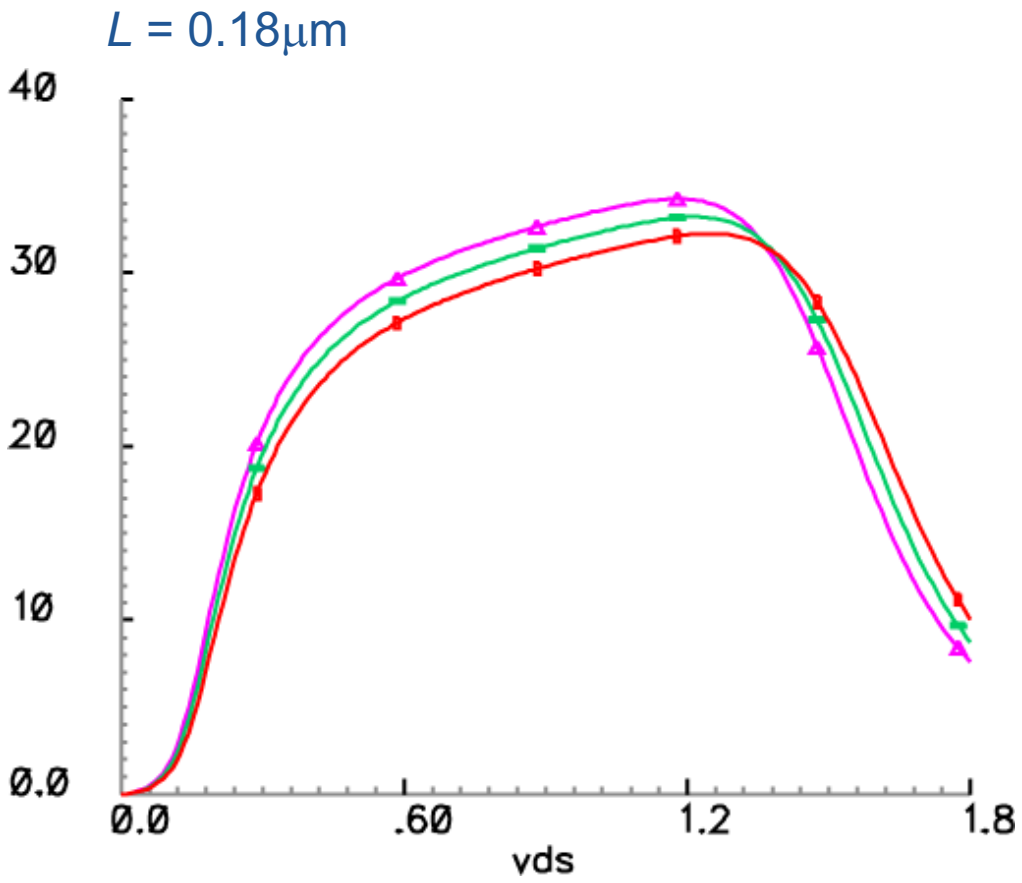
Open-loop Gain a_{v0}



- More useful than r_o
- Represents maximum attainable gain from a transistor

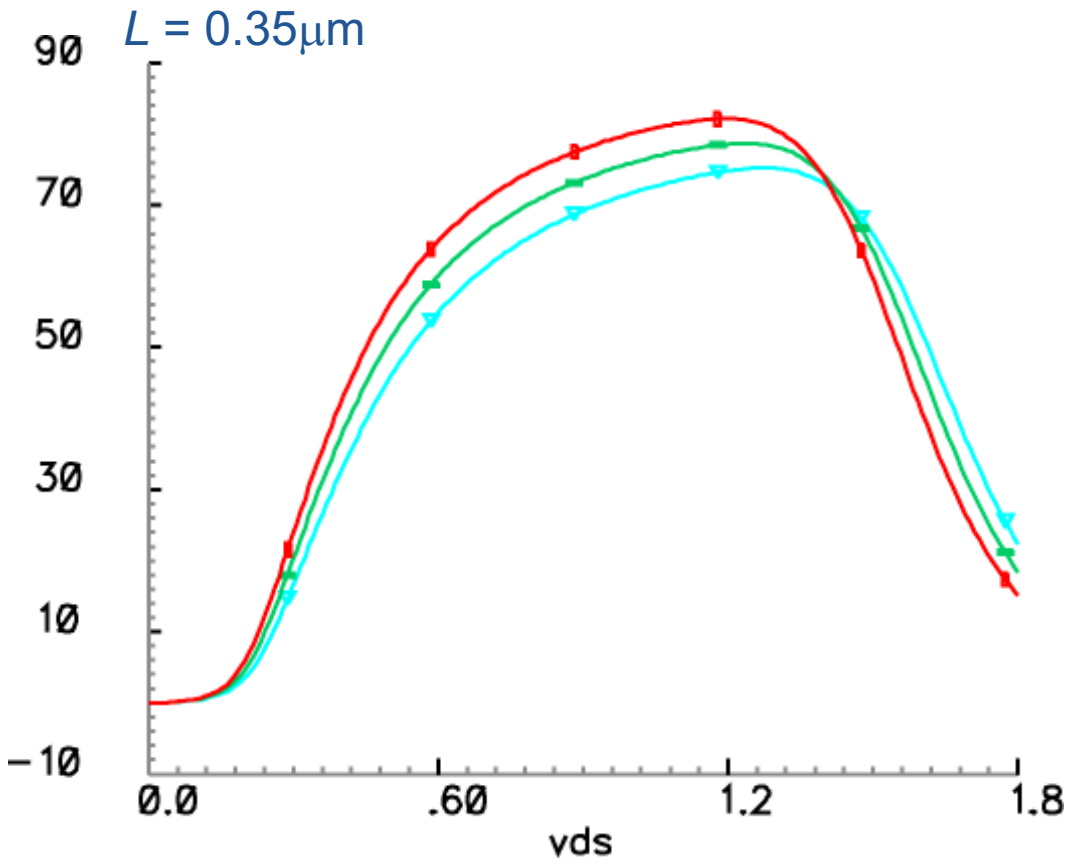
- Simulation Notes:
- Use feedback to bias $V_{ds} = V_{gs}$
- Use relatively small gain (100) for fast DC conversgence

Gain, av0 = gm ro



- Strong tradeoff: a_{v0} versus V_{DS} range
- Create such plots for several device length' for design reference

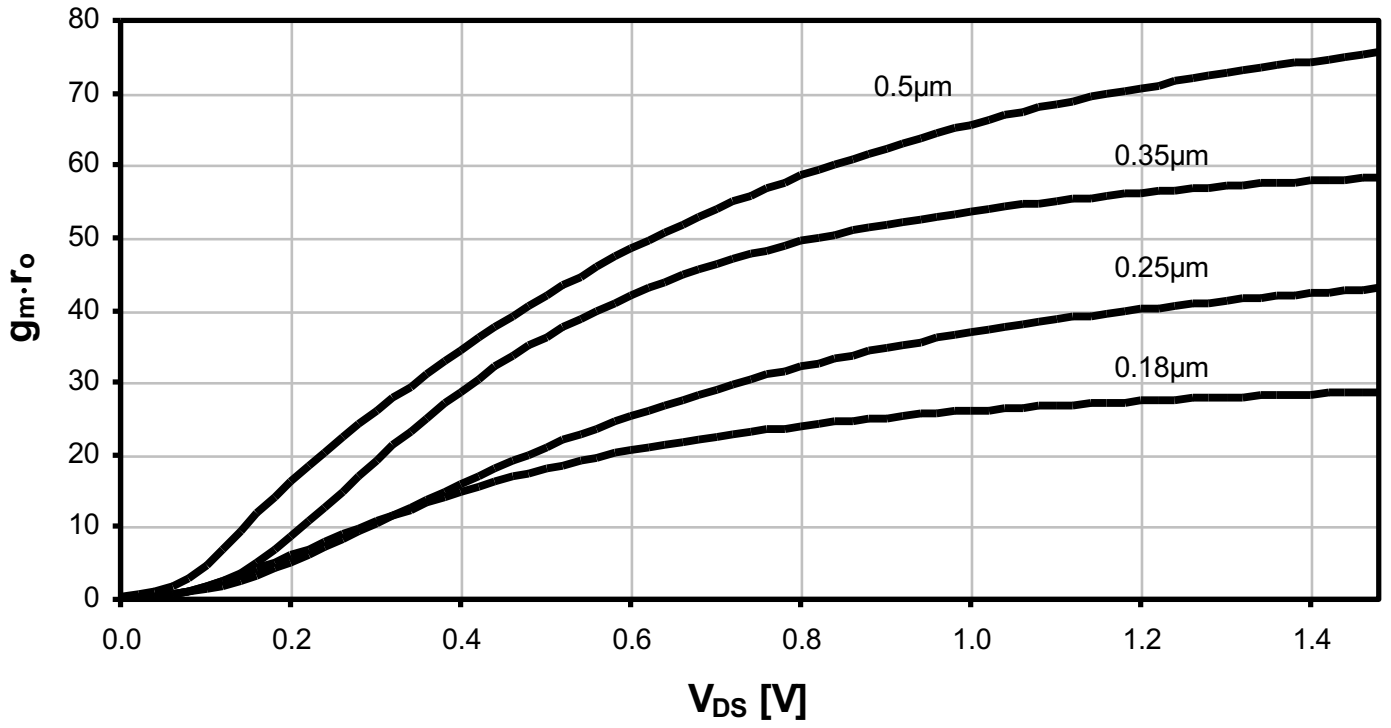
Long Channel Gain



L ↑ → a_{v0} ↑

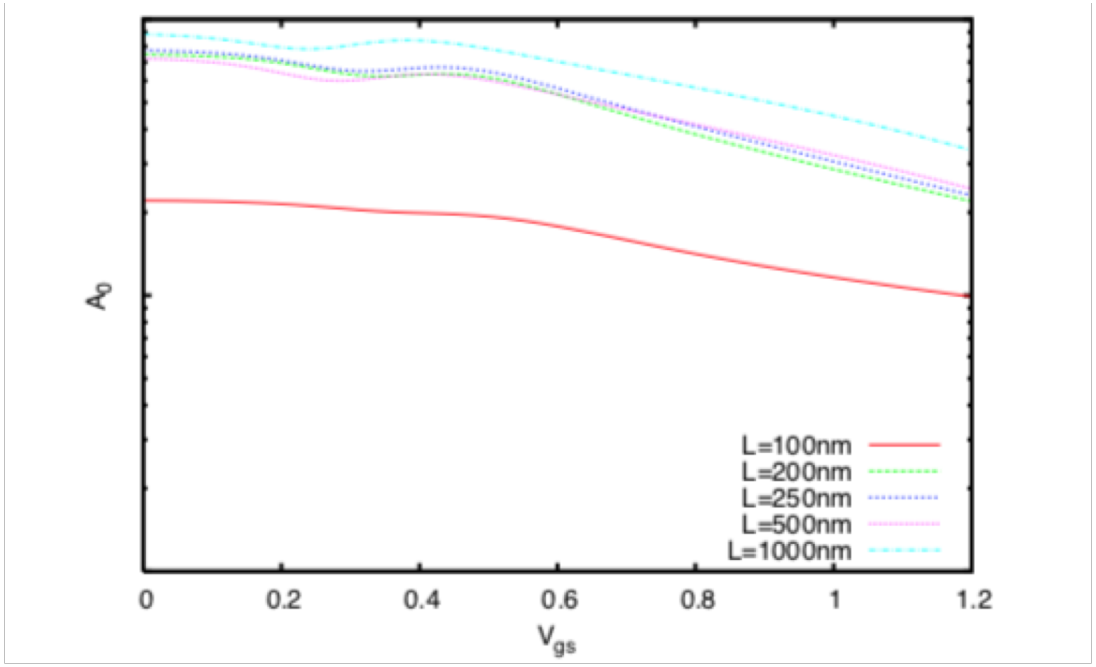
like long channel device

Technology Trend



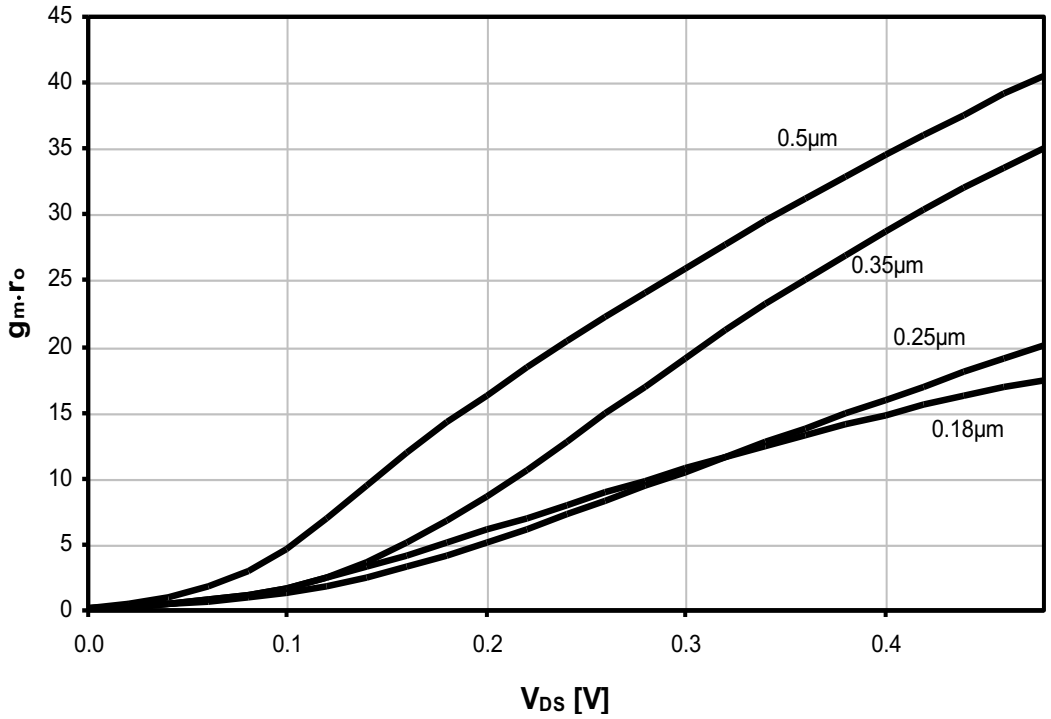
Short channel devices suffer from reduced per transistor gain

Gain vs L (Same Node)



- **Analog circuits often use non-minimum L (100nm in this example) to achieve much better performance**

Transistor Gain Detail



For practical V_{DS} the effect the "short-channel" gain penalty is less severe

(remember: worst case V_{DS} is what matters!)

Saturation Voltage vs V*

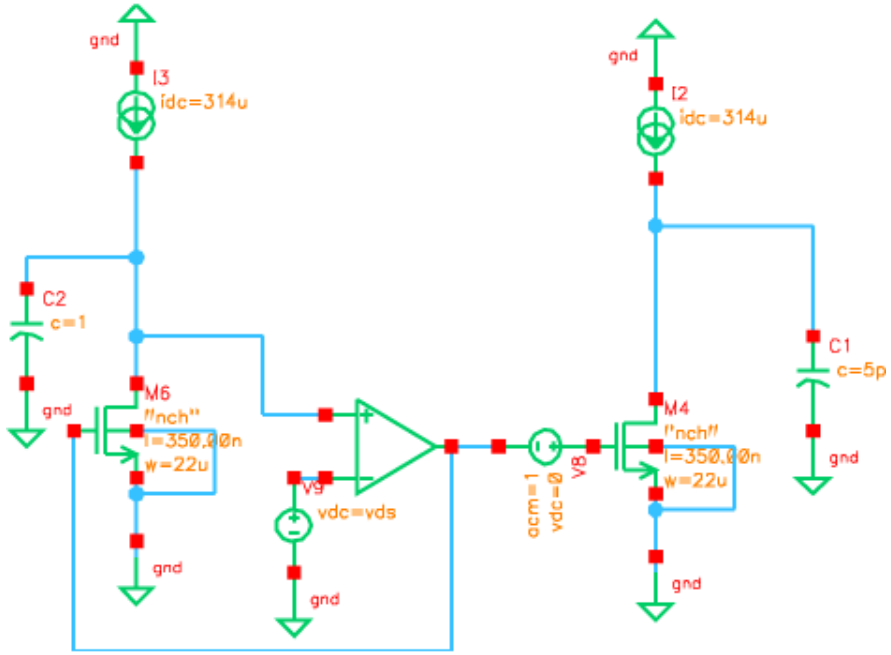
- Saturation voltage
  - Minimum VDS for “high” output resistance
  - Poorly defined: transition is smooth in practical devices

- “Long channel” (square law) devices:
  - VGS – VTH = Vdsat = Vov = V*
  - Significance:
    - Channel pinch-off
    - ID ~ V*2
    - Boundary between triode and saturation
    - ro “large” for VDS > V*
    - CGS, CGD change
    - V* = 2 ID / gm

- “Short channel” devices:
  - All interpretations of V* are approximations
  - Except V* = 2 ID / gm (but V* $\neq$ Vdsat)

Design Example

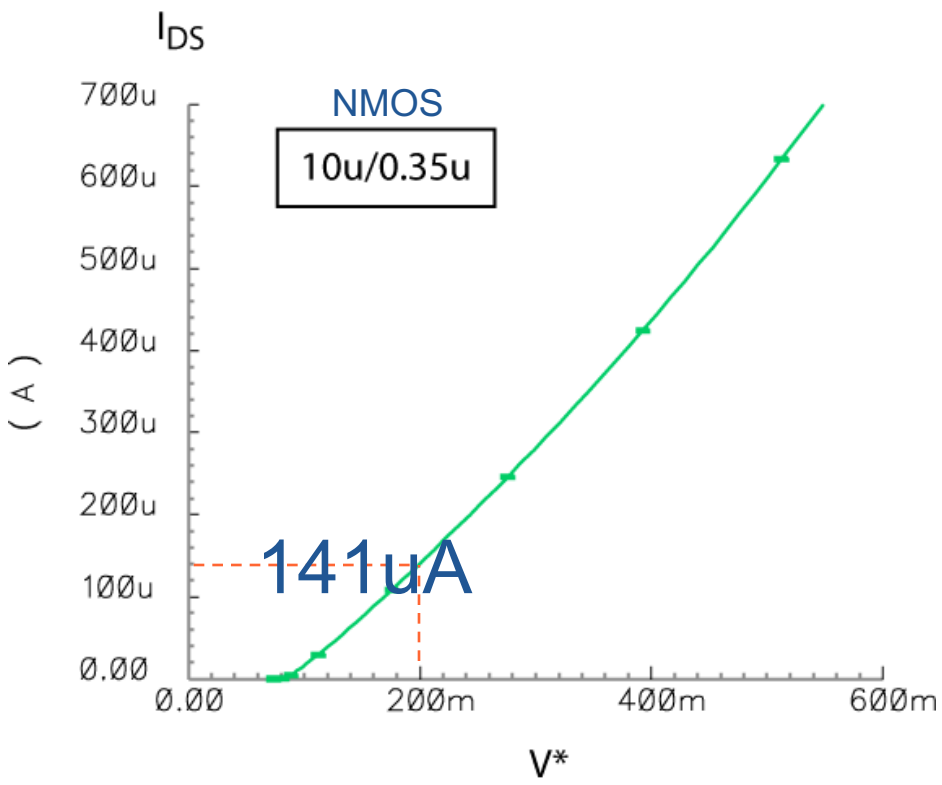
Example: Common-source amp
a_{v0} > 70, f_u = 100MHz for C_L = 5pF



$$g_m \approx 2\pi f_u C_L = 3.14\text{mS}$$

$$I_D = \frac{g_m V^*}{2} = 314\mu\text{A}$$

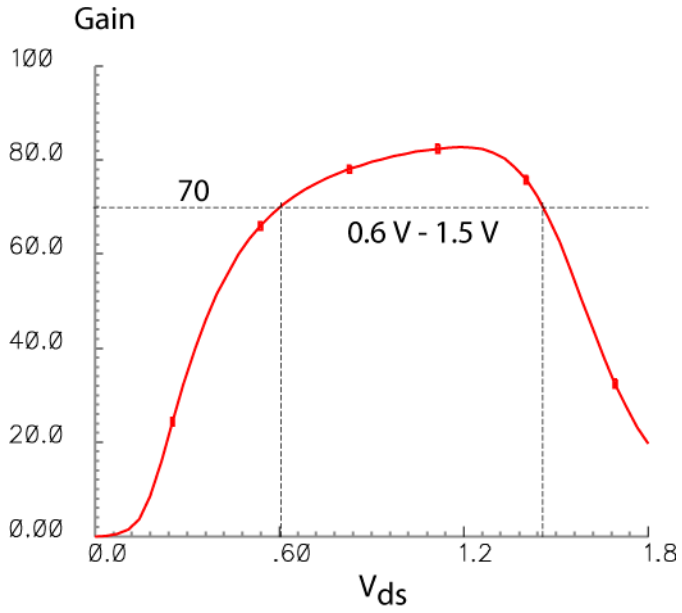
Device Sizing

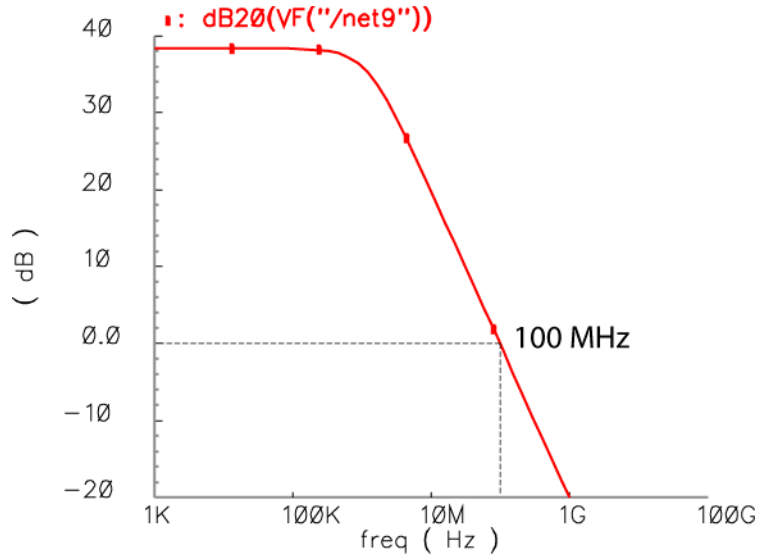


- Pick L 0.35um
- Pick V* 200mV
- Determine gm 3.14mS
- ID = 0.5 gm V* 314uA
- W from graph or BAG (generate with SPICE)

  → W = 10um (314uA /141uA)
  = 22um
- Create such graphs for several device lengths for design reference

Common Source Sims



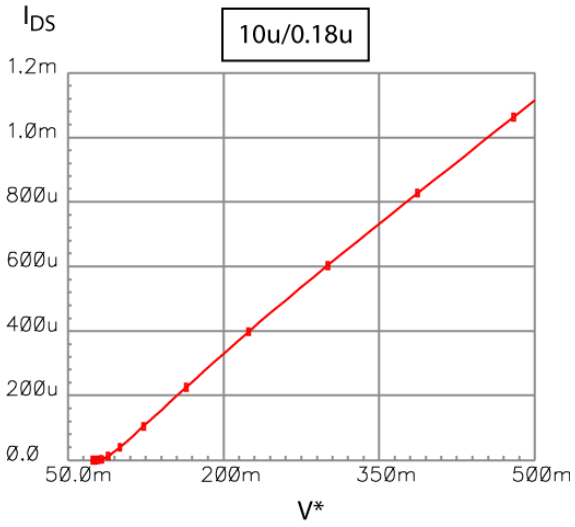


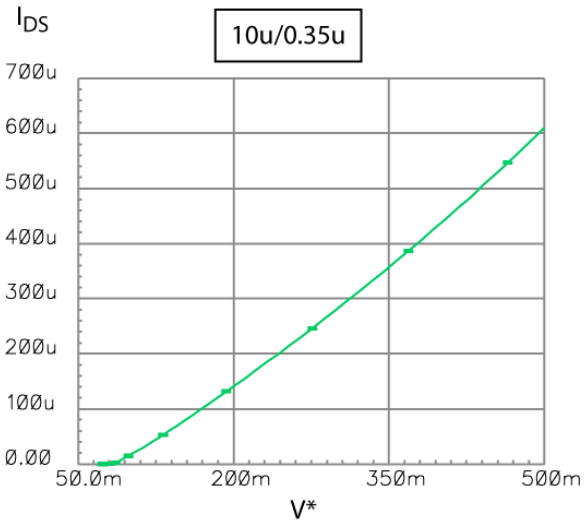
- Amplifier gain > 70
- Amplifier unity gain frequency is “dead on”
- Output range limited to 0.6 V – 1.5 V to maintain gain (about 0.45V swing)

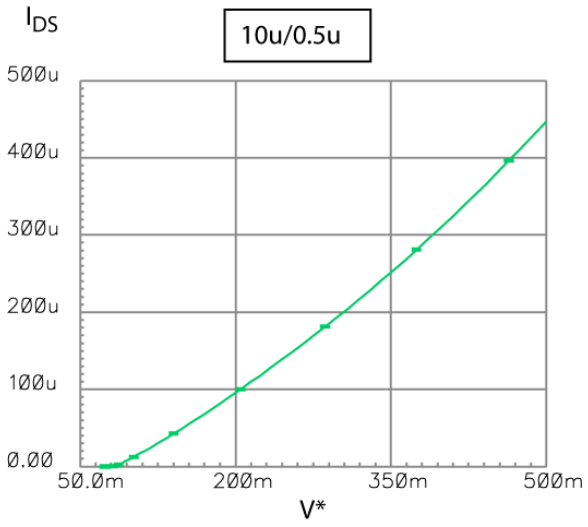
Small Signal Design Summary

- Determine gm (from design objectives)
- Pick L
  - Short channel → high fT
  - Long channel → high ro, av0
- Pick V* = 2ID/gm
  - Since V* is approximately the saturation voltage
  - Small V* → large signal swing
  - High V* → high fT
  - Also affects noise (see later)
- Determine ID (from gm and V*)
- Determine W (SPICE / plot)
- Accurate for short channel devices → key for design

Device Sizing Chart







Generate these curves for a variety of L's and device flavors (NMOS, PMOS, thin oxide, thick oxide, different VT)

Device Parameter Summary

Device Parameter	Circuit Implications
V*	• Current efficiency, g_m/I_D • Power dissipation (I_D) • Speed (g_m) • Cutoff frequency, f_T → phase margin, noise • Headroom, $V_{DS,min}$
L	• Cutoff frequency, f_T → phase margin, noise • Intrinsic transistor gain (a_{v0})
W	• Obtain from L, I_D • Self loading (C_{GS}, C_{DB}, …)

Another Approach

- **Using frequency dependent Y-Parameters, you can extract equivalent "hybrid-pi" models for transistor in region of interest**
  - **242A/142 students will be familiar with this approach**
- **BAG has some tools for doing this in an automated way. This way is good for optimization and less useful for design insights.**