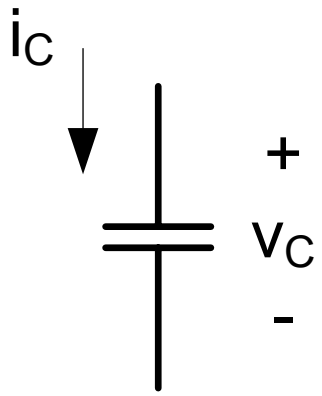
Intrinsic Capacitance BW – Supply Current Tradeoff

Claudio Talarico
Gonzaga University

The Culprit

Source: B. Murmann

- In practical circuits, the presence of capacitance prevents us from building circuits that can run "infinitely" fast
  - Sometimes inductors can be used improve the situation



$$i_C = C\frac{dv_C}{dt}$$

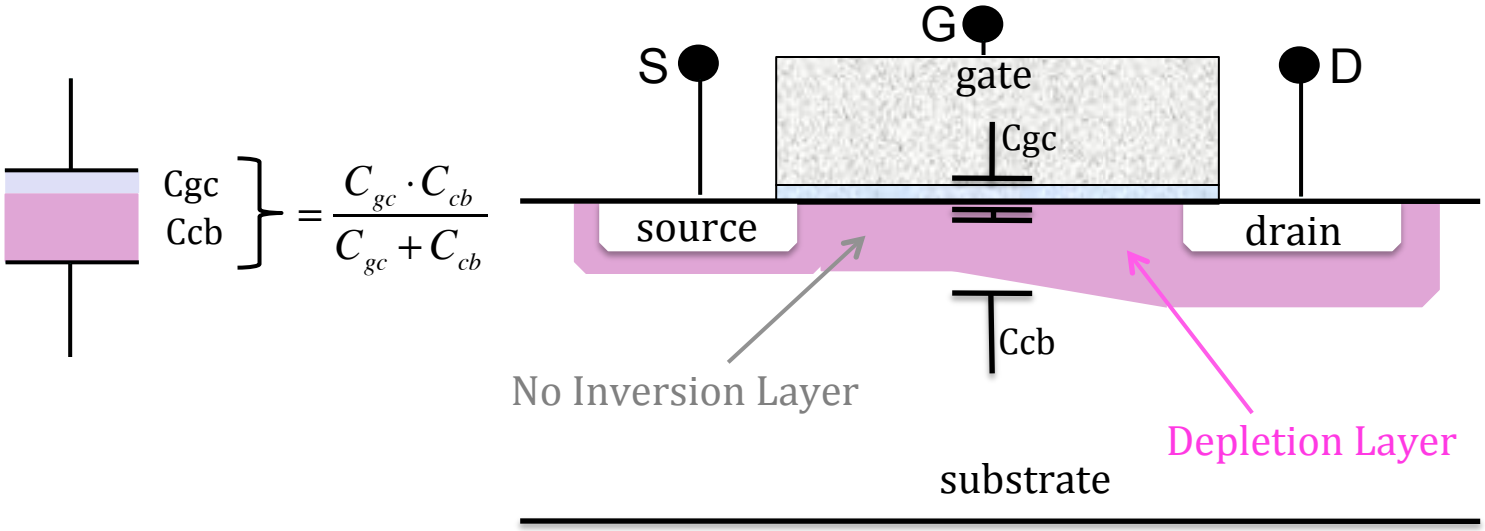
- Intuition
  - High frequency results in large dv_C/dt and large i_C
  - Capacitor becomes a "short" for high frequencies

Modeling MOSFET Capacitance (1)

Source: B. Murmann

- To predict the frequency response of our CS amplifier, we need to understand the capacitances associated with the MOS device
- We will first look at "intrinsic gate capacitance"
  - Intrinsic means that this capacitance is unavoidable and required for the operation of the device
- There are plenty of "extrinsic" capacitances as well
  - We'll neglect these for the time being and discuss them in detail next lecture
- To model gate capacitance, we must distinguish the MOSFET's operating regions
  - Transistor on
    - Triode and active regions
  - Transistor "off"
    - Sub-threshold operation

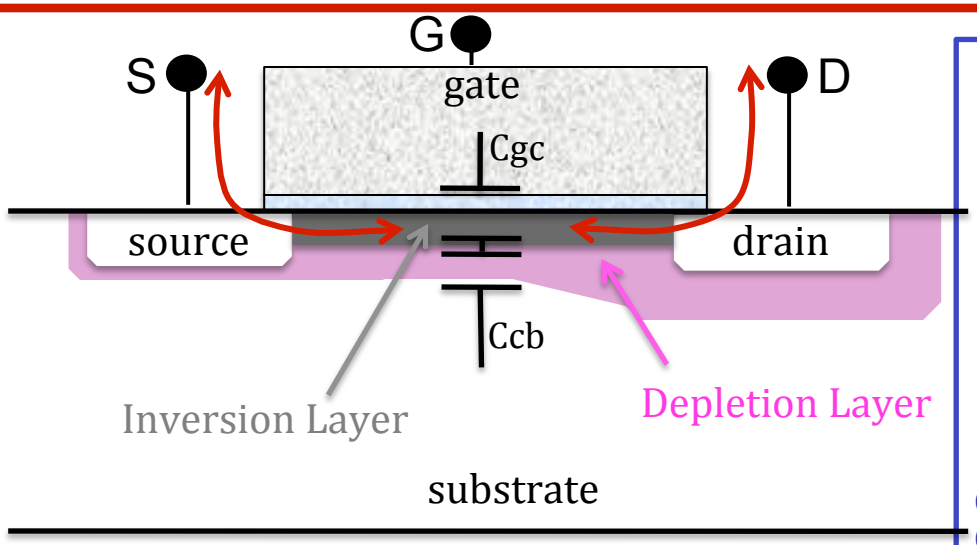
Transistor Off

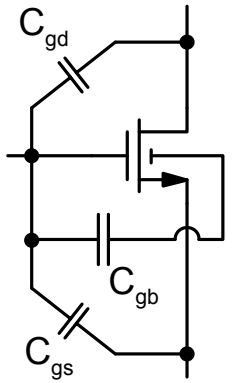


- There is no conductive channel
  - Gate sees a capacitor to substrate, equivalent to the series combination of the gate oxide capacitor and the depletion capacitance
- If the gate voltage is taken negative, the depletion region shrinks, and the gate-substrate capacitance grows
  - With large negative bias, the capacitance approaches WLC_{OX}

Transistor in Triode Region

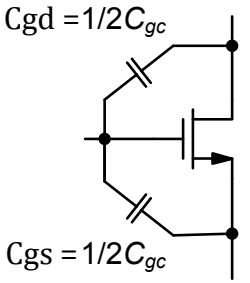
Electrons respond to gate via the S and D; this effectively "short out" the effects of C_{CB} (Bulk charge is not changing with VGB: any increment in gate charge is mirrored in the inversion layer)





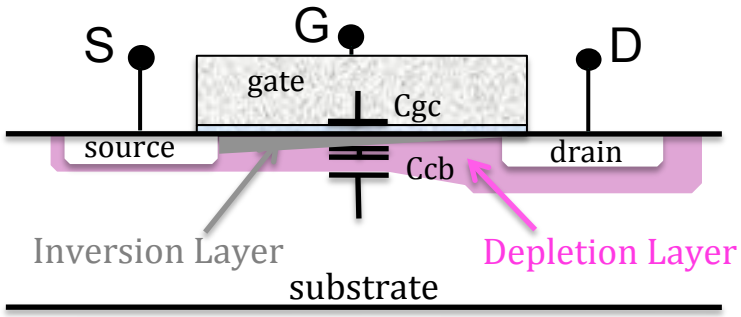
Cgs and Cgd model Cgc
Cgb models Ccb

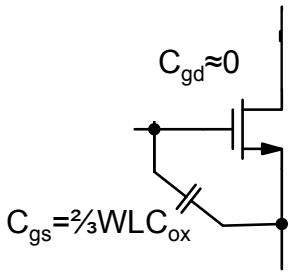
- Gate terminal and conductive channel form a parallel plate capacitor across gate oxide $C_{GC} = WL\varepsilon_{ox}/t_{ox} = WLC_{ox}$
  - We can approximately model this using lumped capacitors of size ½ C_{GC} each from gate-source and gate-drain
    - Changing either voltage will change the channel charge
- The depletion capacitance C_{CB} adds extra capacitance from drain and source to substrate
  - Usually negligible (see above comment)



Transistor in Saturation Region

- Assuming a long channel model, if we change the source voltage
  - The voltage difference between the gate and channel at the drain end remains at V_t, but the voltage at the source end changes
  - This means that the "bottom plate" of the capacitor does not change uniformly
- Detailed analysis shows that in this case $C_{gs}=^2/_3WLC_{ox}$
  - For your amusement you can derive this; write Q(V(y)) and integrate from 0 to L
- In the long channel model for active operation, the drain voltage does not affect the channel charge
  - This means $C_{gd}=0$ in the saturation region!
    - Neglecting second order effects and extrinsic caps, of course

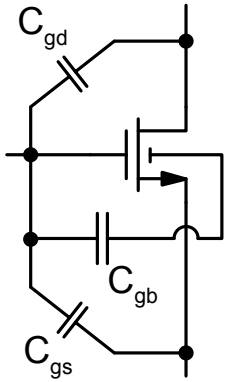




Source:
B. Murmann

Intrinsic MOS Capacitor Summary

source: B. Murmann



	Subthreshold	Triode	Saturation
C_{gs}	0	$\frac{1}{2} WLC_{ox}$	$\frac{2}{3} WLC_{ox}$
C_{gd}	0	$\frac{1}{2} WLC_{ox}$	0
C_{gb}	$\left(\frac{1}{C_{CB}} + \frac{1}{WLC_{ox}}\right)^{-1}$	0	0

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

x_d is the width of the depletion region at the silicon surface

- For $V_{bi} \le V_{GB} \le V_t$

$$x_d(V_{GB}) = \sqrt{\frac{2\varepsilon_{Si}(V_{bi} - V_{GB})}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)} = x_{do}\sqrt{1 - \frac{V_{GB}}{V_{bi}}}$$

$$x_{do} \triangleq \sqrt{\frac{2\varepsilon_{Si} V_{bi}}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)}$$

$$V_{bi} = \frac{KT}{q} \ln \frac{N_a N_d}{n_i^2}$$

Performance Considerations (1)

Source: B. Murmann

- Suppose we want to maximize the bandwidth of our circuit
  - And keep the current consumption (I_D) as low as possible
- Further assume that R_i, R and A_{v0} are fixed
  - Fixed by the particular application in which the circuit is used
- For the bandwidth, we can write

$$\omega_{3dB} = \frac{1}{R_i C_{gs}} = \frac{1}{R_i \cdot \frac{2}{3} WLC_{ox}}; \qquad g_m = \mu C_{ox} \frac{W}{L} \cdot V_{OV}; \qquad |a_{v0}| = g_m R$$

$$\Rightarrow \omega_{3dB} = \frac{3}{2} \frac{\mu}{L^2} \cdot \frac{R}{|a_{v0}| \cdot R_i} \cdot V_{OV}$$

Technology Specs Design "knob"

Performance Considerations (2)

Source: B. Murmann

- For the current we can write

$$g_m = \frac{2I_D}{V_{OV}}; \qquad |a_{v0}| = g_m R \qquad \Rightarrow I_D = \frac{1}{2}\frac{|a_{v0}|}{R}\cdot V_{OV}$$

Specs — Design "knob"

- Observations
  - Larger V_{OV} means larger bandwidth
  - Unfortunately larger V_{OV} also results in larger I_D
- Part of your job as an analog designer is to choose V_{OV} such that you get sufficient bandwidth while using as little current as possible

Performance Considerations (3)

Source: B. Murmann

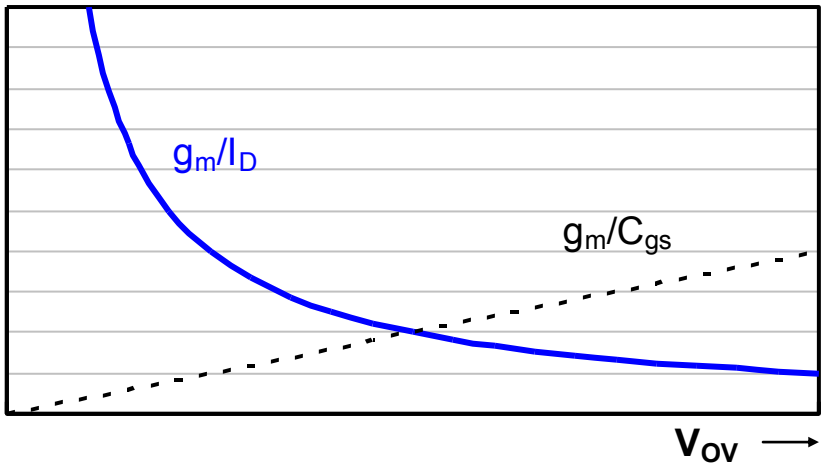
- Even though we've come to this observation using a very simple example, this tradeoff tends to hold in general
  - Of course, additional considerations and second order dependencies will factor in as you learn more about analog circuit design…
- The reasons behind this tradeoff lie in the fundamental properties of the transistor itself
- To see this, think about what we really want from the MOSFET
  - Large g_m without investing much current
    - I.e. large g_m/I_D
  - Large g_m without having large C_{gs}
    - I.e. large g_m/C_{gs}

g_m/I_D and g_m/C_{gs}

Source: B. Murmann

$$\frac{g_m}{I_D} = \frac{2}{V_{OV}}$$

$$\frac{g_m}{C_{gs}} = \frac{3}{2}\frac{\mu V_{OV}}{L^2}$$



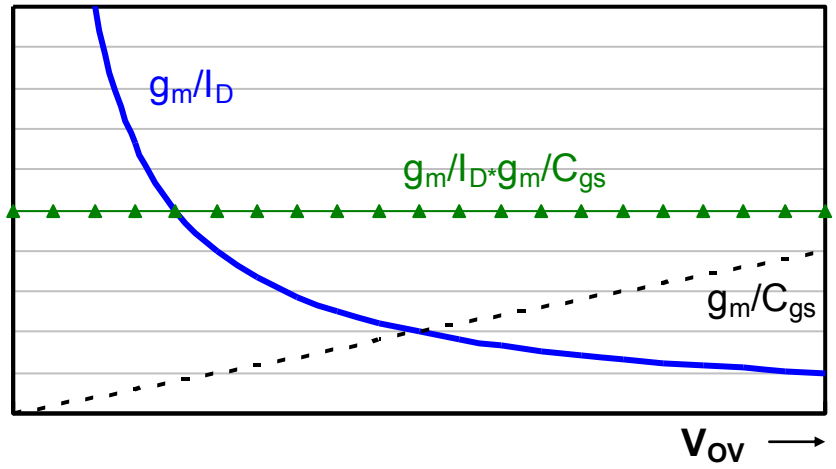
	V_{OV} small	V_{OV} large
g_m/I_D	Large	Small
g_m/C_{gs}	Small	Large

Product

Source: B. Murmann

- In cases where we want to get the "best of both worlds", it is interesting to look at the product of our two figures of merit

$$\frac{g_m}{I_D} \cdot \frac{g_m}{C_{gs}} = \frac{3\mu}{L^2}$$



- While this result looks boring, it shows that using smaller channel lengths improves circuit performance
  - Either or both speed and current efficiency

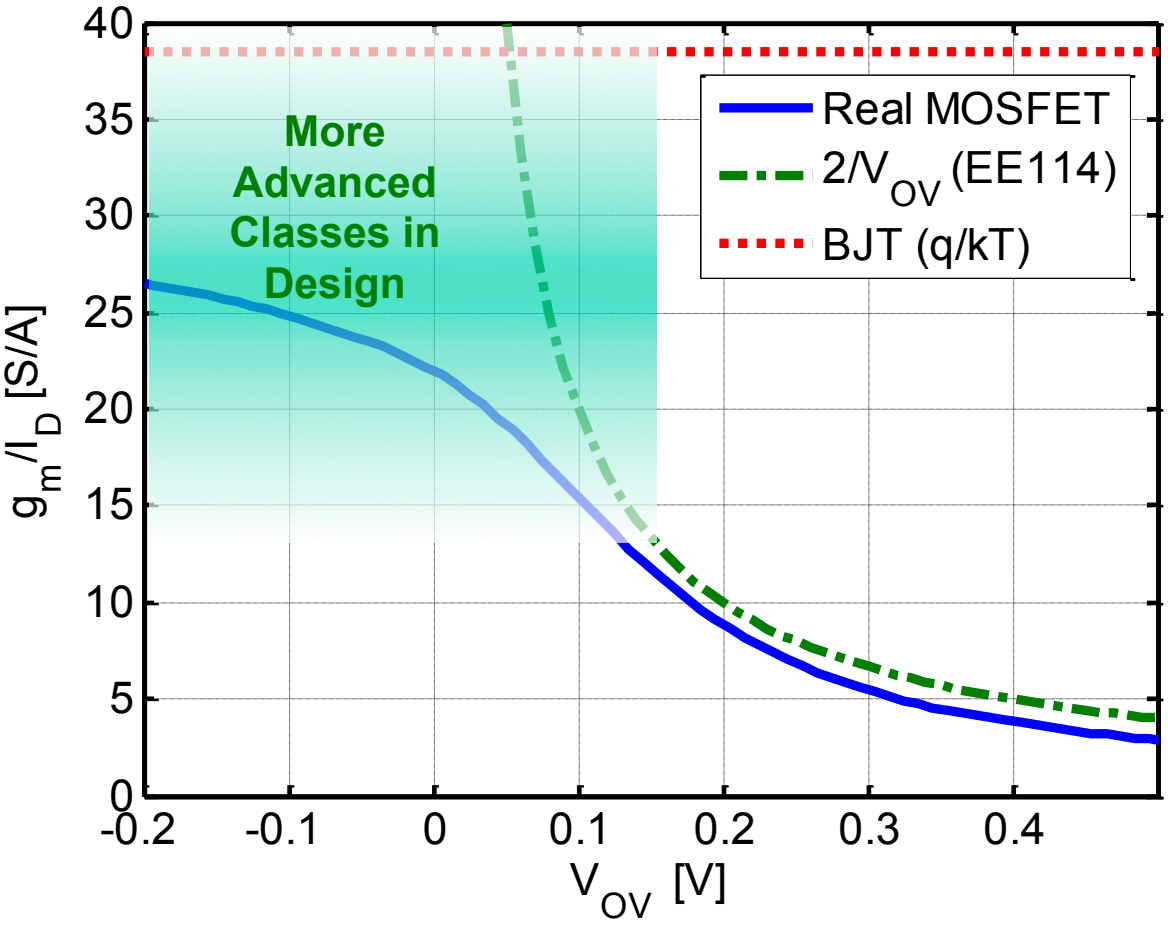
Scaling Impact

Source: B. Murmann

- Thanks to "Moore's Law" feature sizes and thus the available minimum channel length has been shrinking continuously
  - L_{min} has decreased roughly 2x every 5 years
  - L_{min}=10mm in 1970, L_{min}=45nm in 2007
- From the above discussion, it is clear that we can exploit technology scaling in different ways
  - Build faster circuits (higher g_m/C_{gs}), while keeping power efficiency constant (g_m/I_D)
    - E.g. A/D converter for a disk drive - want to maximize bandwidth/ throughput
  - Build more efficient circuits (higher g_m/I_D), while keeping the bandwidth constant (g_m/C_{gs})
    - E.g. A/D converter for video signals - bandwidth fixed by a certain standard

Aside: g_m/I_D of a "Real Transistor"

Source: B. Murmann

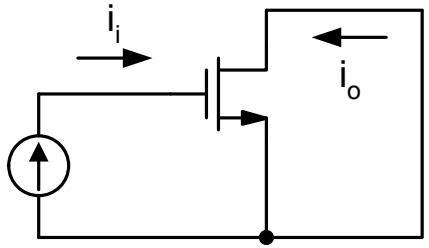


- Long channel predication ($2/V_{OV}$) is fairly close to "real device" for V_{OV} > 150mV
- Unfortunately g_m/I_D does not approach infinity for $V_{OV} \rightarrow 0$
- In this course limit V_{OV} >= 150mV in to avoid non-physical design outcomes

Aside: Transit Frequency (ω_T)

Source: B. Murmann

- The transit frequency of a transistor has "historically" been defined as the frequency where the magnitude of the common source current gain ($|i_o/i_i|$) falls to unity



(Biasing not shown)

- Ignoring extrinsic capacitance, it follows that

$$\omega_T = \frac{g_m}{C_{gs}} = \frac{3}{2}\frac{\mu V_{OV}}{L^2}$$

- Incidentally, this metric is identical to the figure of merit we considered on the previous slides

Aside: Transit Frequency Interpretation

Source: B. Murmann

- The transit frequency is only useful as a figure of merit in the sense that it quantifies g_m/C_{gs}
- It does not accurately predict up to which frequency you can use the device
  - At high frequencies, many assumptions in our "lumped" transistor model become invalid
  - Rule of thumb: lumped model is good up to about $\omega_T/5$
- At higher frequencies, device modeling becomes more challenging and many effects depend on how exactly you layout and connect the device