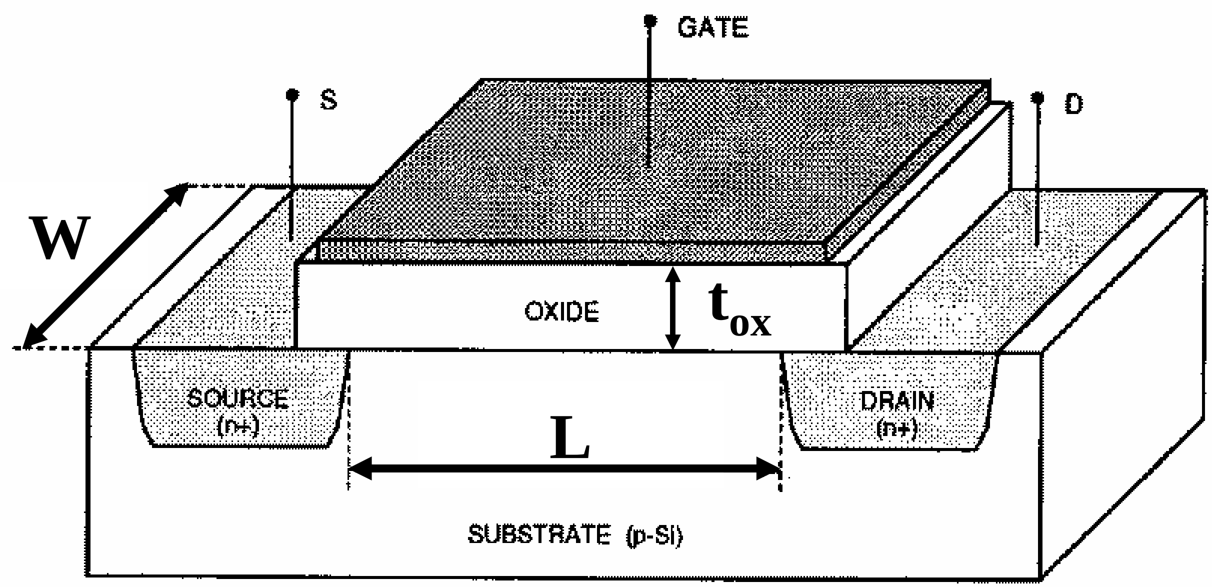
ECE 225
High-Speed Digital IC Design
Lecture 8

CMOS Devices: Basic Operation, Sub-micron Effects, Parasitic Effects, Emerging Structures

Prof. Kaustav Banerjee
Electrical and Computer Engineering
E-mail: kaustav@ece.ucsb.edu

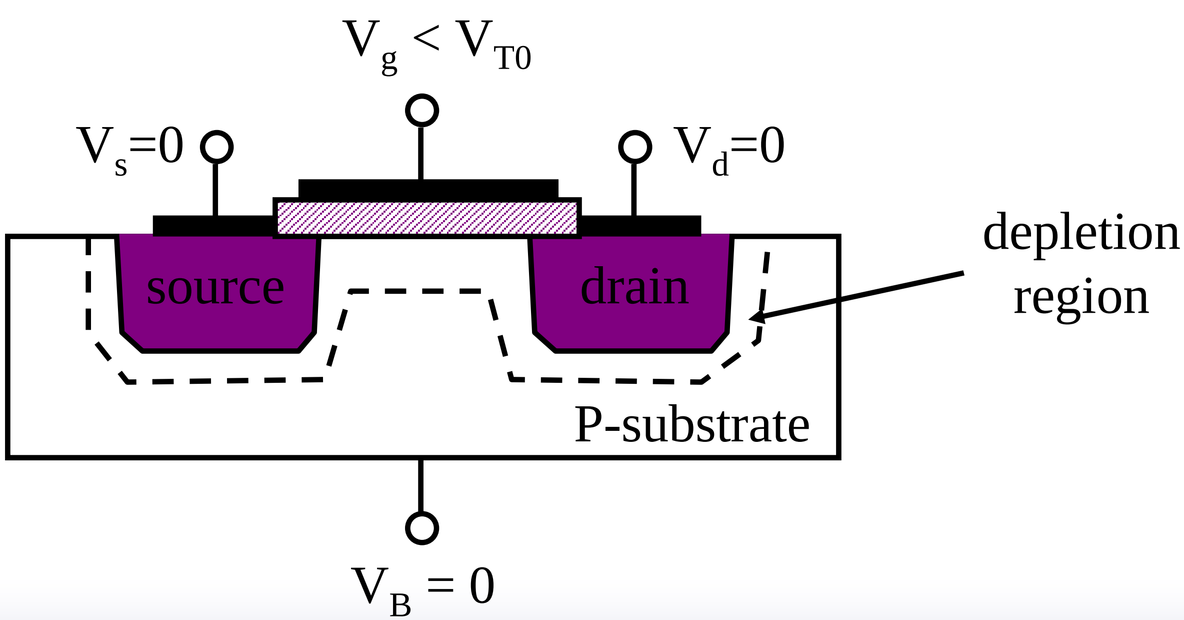
MOS Transistor

- Important transistor physical characteristics
  - Channel length L
  - Channel width W



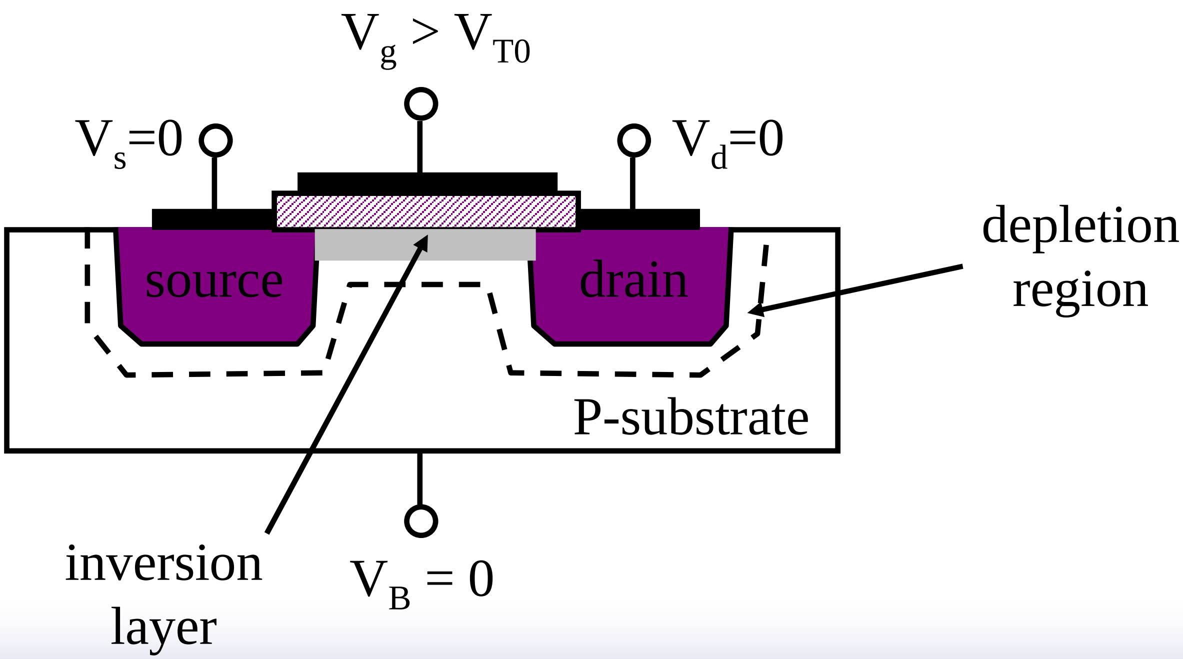
MOS Transistor Operation

- Simple case: $V_D = V_S = V_B = 0$
  - Operates as MOS capacitor
- When $V_{GS}<V_{T0}$ (but positive) , depletion region forms
  - No carriers in channel to connect S and D
- V_{T0} is known as the *threshold voltage*



MOS Transistor Operation

- When $V_{GS} > V_{T0}$, inversion layer forms
- Source and drain connected by conducting n-type layer (for NMOS)



Physical Parameters that Affect V_{T0}

- Threshold voltage (V_{T0}): voltage between gate and source required for inversion
  - NMOS Transistor is "off" when $V_{GS} < V_{T0}$
- Components:
  - Work function difference between gate and channel (Flat-band voltage)
  - Gate voltage to change surface potential
  - Gate voltage to offset depletion region charge
  - Gate voltage to offset fixed charges in gate oxide and in silicon-oxide interface

Threshold voltage, summary

- If $V_{SB} = 0$ (no substrate bias):

$$V_{T0} = \Phi_{GC} - 2\phi_F - \frac{Q_{B0}}{C_{ox}} - \frac{Q_{ox}}{C_{ox}}$$

- If $V_{SB} \neq 0$ (non-zero substrate bias)

$$V_T = V_{T0} + \gamma\left(\sqrt{\left|-2\phi_F + V_{SB}\right|} - \sqrt{\left|2\phi_F\right|}\right)$$

- Body effect (substrate-bias) coefficient:

$$\gamma = \frac{\sqrt{2qN_A\varepsilon_{Si}}}{C_{ox}}$$

+ for NMOS
- for PMOS

- Threshold voltage increases as V_{SB} increases!

Threshold Voltage (NMOS vs. PMOS)

	NMOS	PMOS
Substrate Fermi potential	$\phi_F < 0$	$\phi_F > 0$
Depletion charge density	$Q_B < 0$	$Q_B > 0$
Substrate bias coefficient	$\gamma > 0$	$\gamma < 0$
Substrate bias voltage	$V_{SB} > 0$	$V_{SB} < 0$
Threshold voltage (enhancement devices)	$V_{T0} > 0$	$V_{T0} < 0$

Threshold voltage adjustment

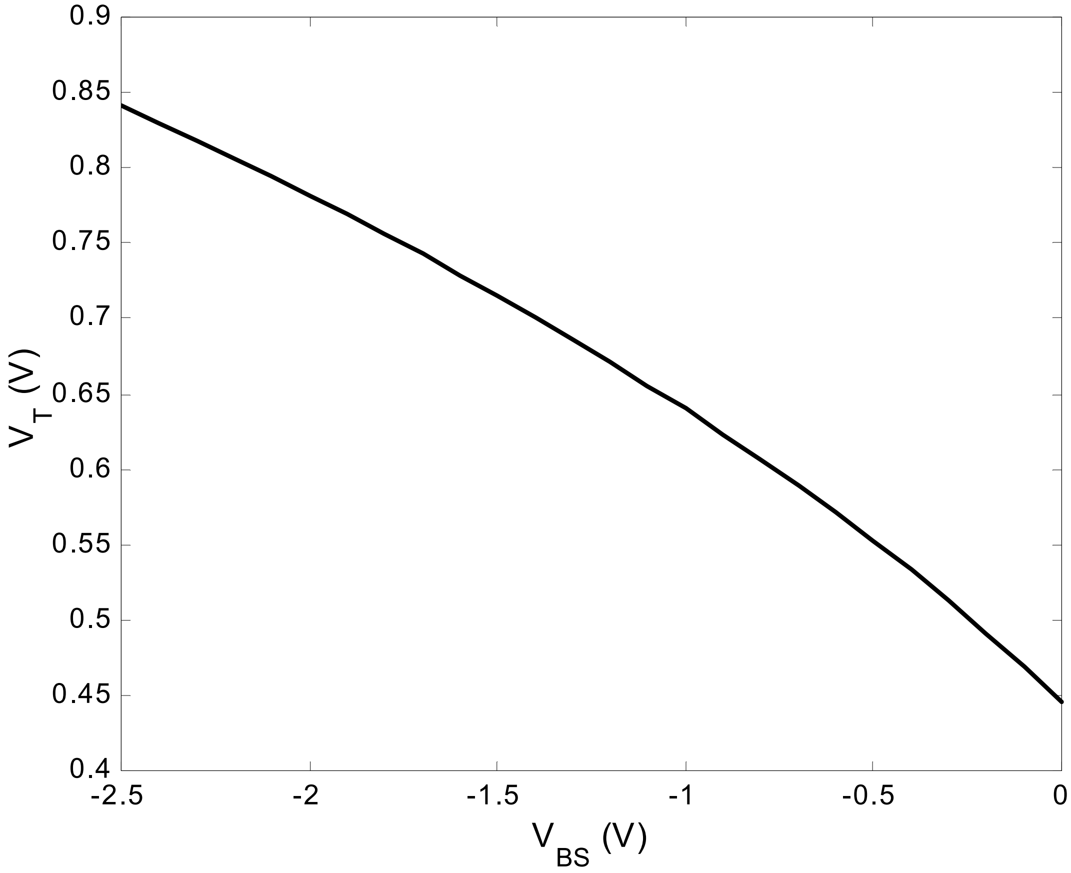
- Threshold voltage can be changed by doping the channel region with donor or acceptor ions
- For NMOS:
  - V_T increased by adding acceptor ions (p-type)
  - V_T decreased by adding donor ions (n-type)
  - Opposite for PMOS
- Approximate change in V_{T0}:
  - Density of implanted ions = N_I [cm^{-2}]
  - Assume all implanted impurities are ionized

$$\Delta V_{T0} = \frac{qN_I}{C_{ox}}$$

Example: V_{T0} Adjustment

- Consider an NMOS device:
  - P-type substrate: $N_A = 2 \times 10^{16}$ cm^{-3}
  - Polysilicon gate: $\Phi_{GC} = -0.92V$
  - $t_{ox} = 600$ Å (1Å = 1×10^{-8} cm)
  - $N_{ox} = 2 \times 10^{10}$ cm-2
  - $\varepsilon_{Si} = 11.7\ \varepsilon_{0,}\ \varepsilon_{ox} = 3.97\ \varepsilon_0$
- (a) Find V_{T0}
- (b) Find amount and type of channel implant to get $V_{T0} = 0.4$ V

The Body Effect



Kaustav Banerjee

Transistor Currents (NMOS)

Cutoff Region: $I_{ds} = 0,\ V_{gs} < V_t$

$I_{ds} = Q_{channel}$ /carrier velocity(v)

$v = \mu E$

Linear Region: $V_{gs} > V_t,\ V_{ds} < V_{gs} - V_t$

$E = V_{ds}/L$

$I_{ds} = \mu\, C_{ox}\, W/L\ (V_{gs} - V_t - V_{ds}/2) V_{ds}$

$\beta = \mu\, C_{ox}\, W/L$

Saturation Region: $V_{gs} > V_t,\ V_{ds} > V_{gs} - V_t$

$I_{ds} = \beta/2\, (V_{gs} - V_t)^2$

Note: for PMOS $V_{tp} \neq V_{tn}$

$\mu_p < \mu_n$, hence $(W/L)_{PMOS} \sim 2\, (W/L)_{NMOS}$

NMOS Characteristics

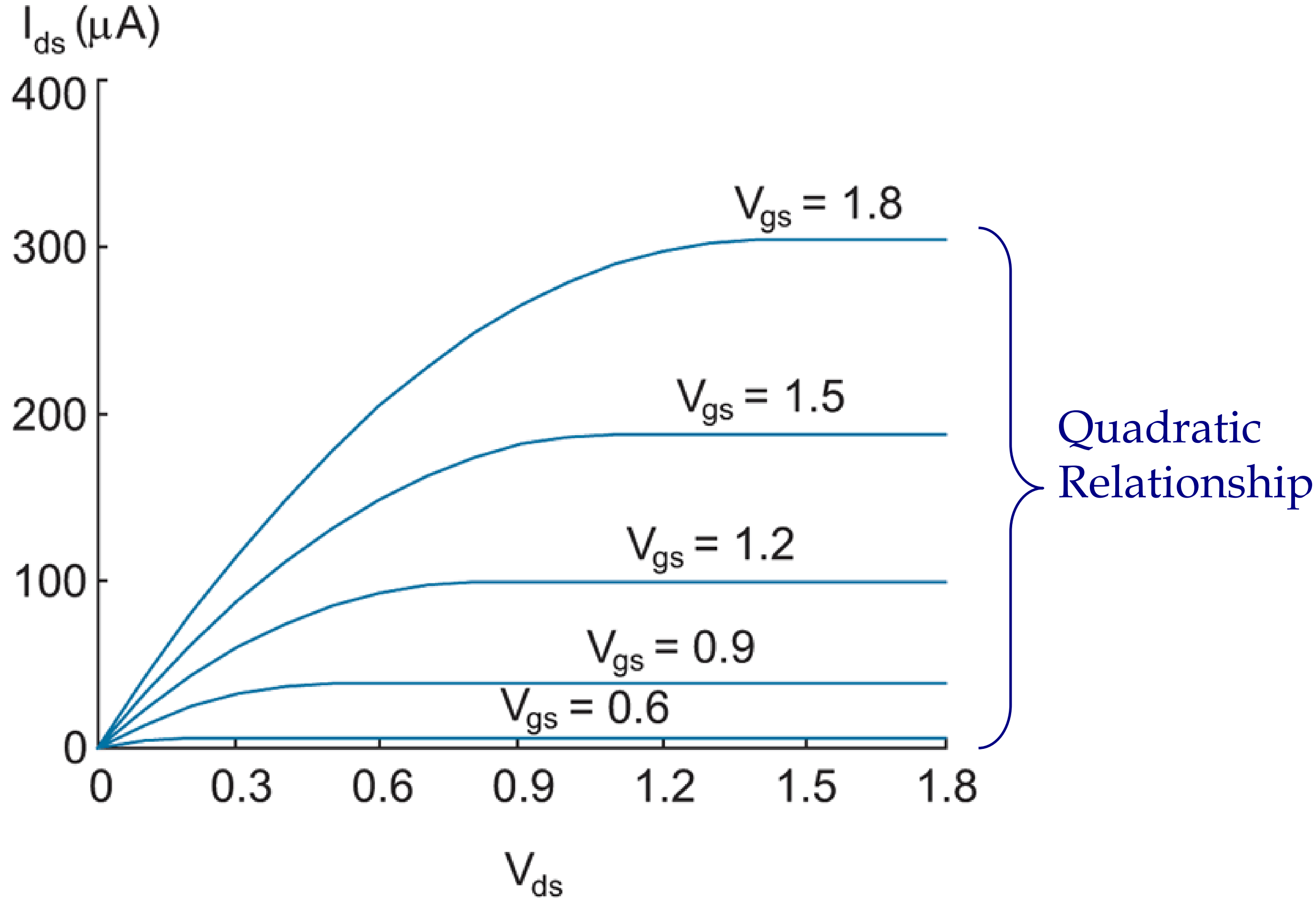


FIG 2.7 I-V characteristics of ideal nMOS transistor

PMOS Characteristics

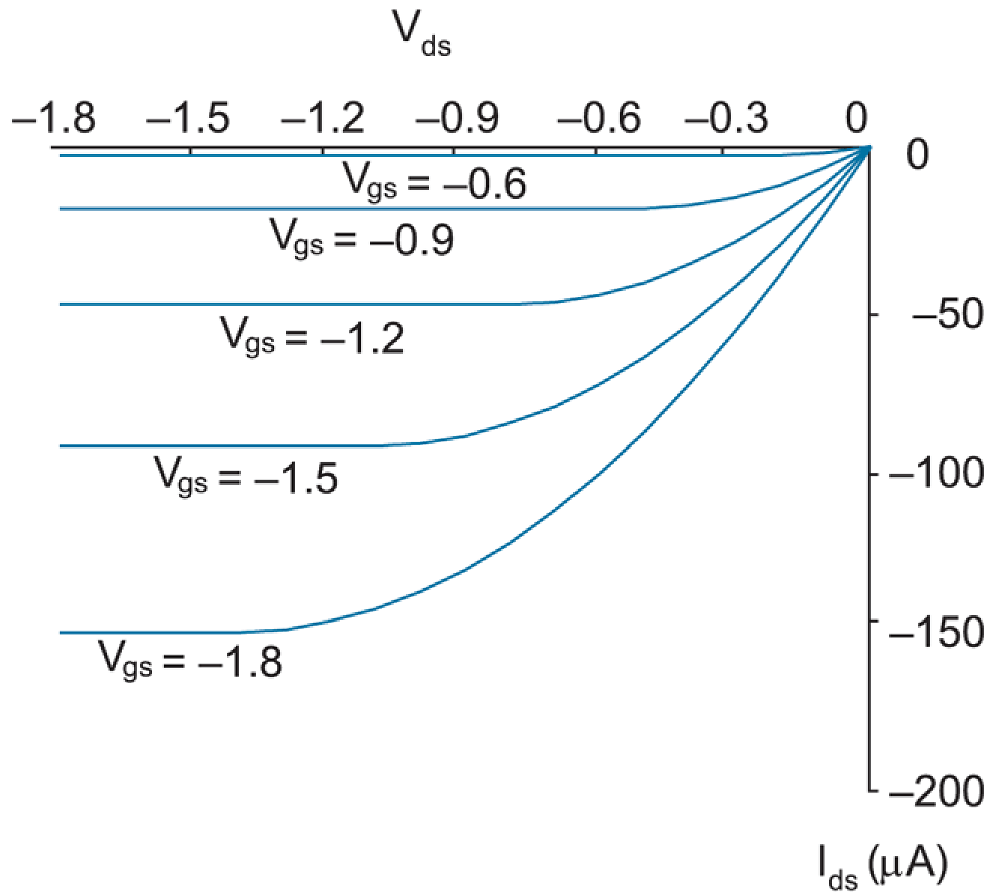


FIG 2.8 I-V characteristics of ideal pMOS transistor

Channel Length Modulation

- In saturation, pinch-off point moves
  - As V_{DS} is increased, pinch-off point moves closer to source
  - Effective channel length becomes shorter
  - Current increases due to shorter channel

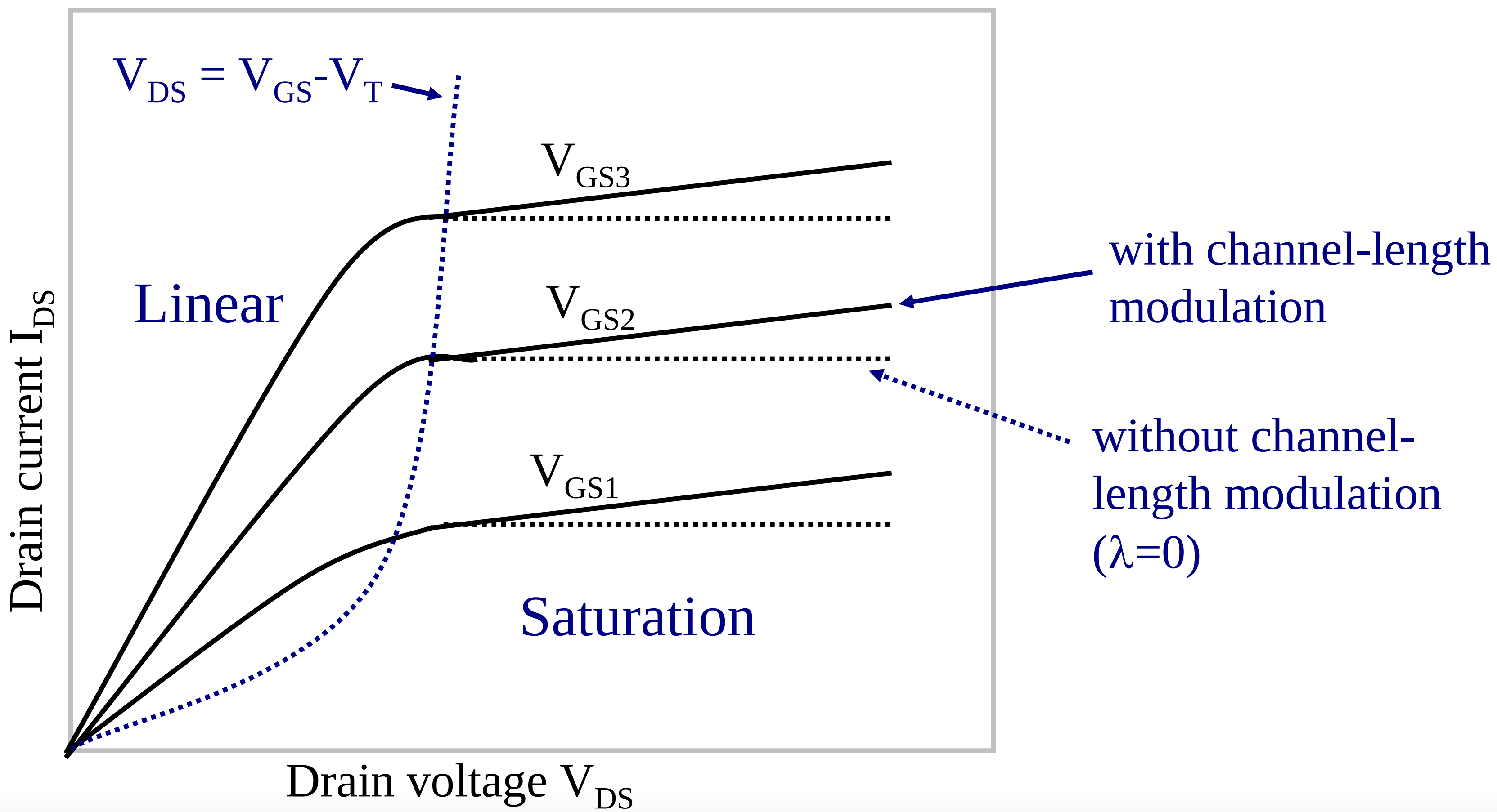
$$L' = L - \Delta L$$

$$I_D = \tfrac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TN})^2(1 + \lambda V_{DS})$$

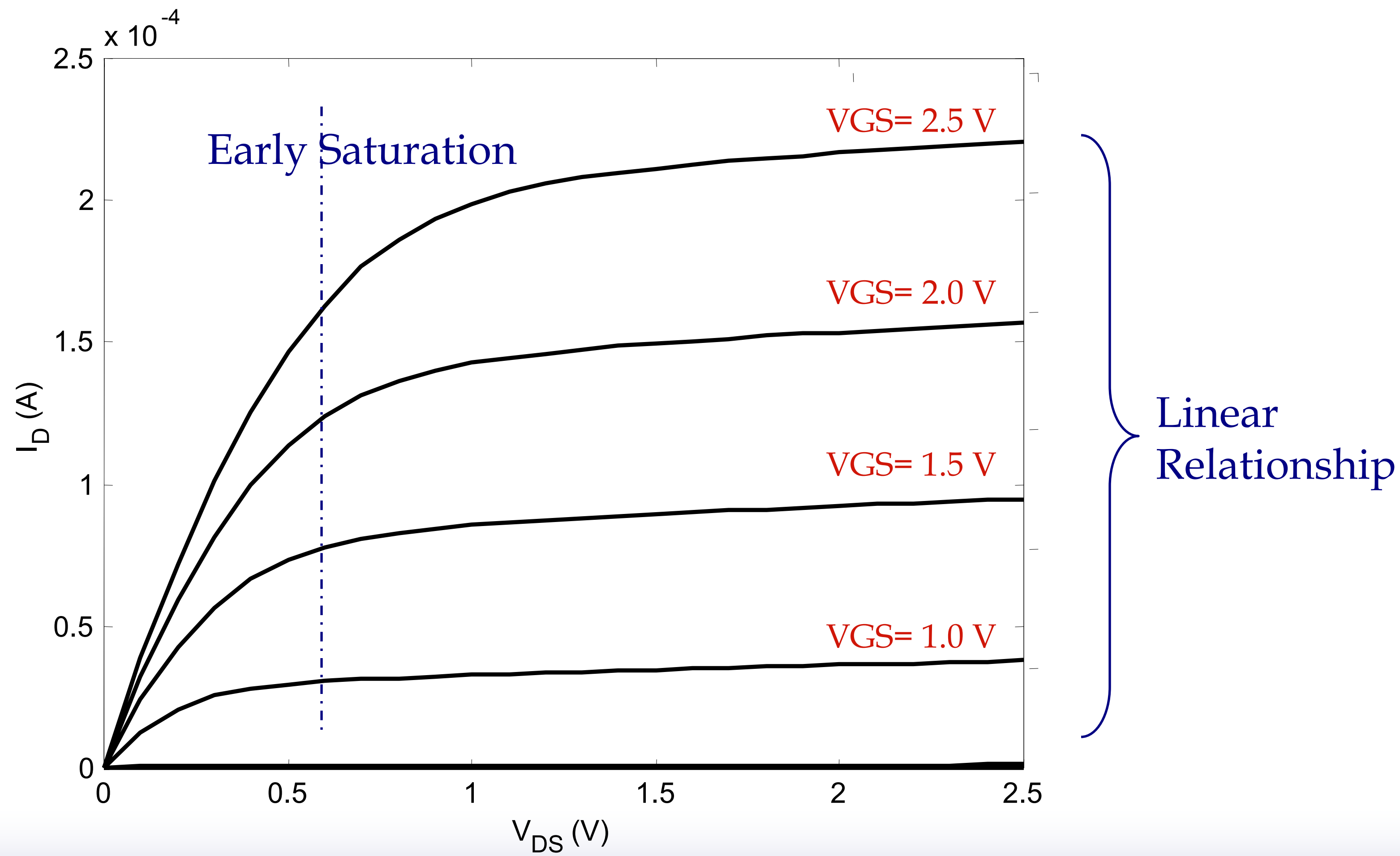
λ = channel length modulation coefficient

Summary: MOS I/V

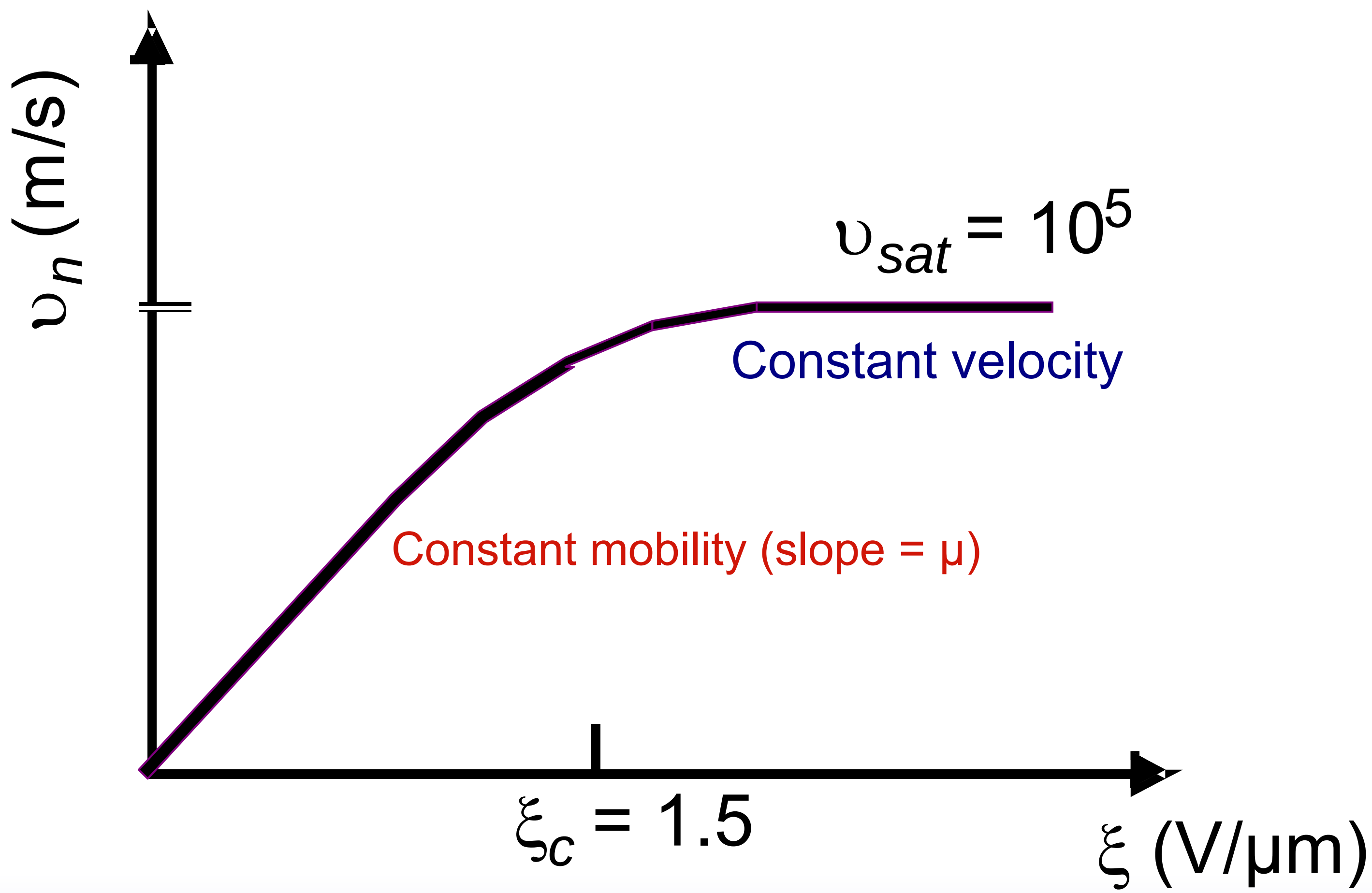
I/V curve for NMOS device:



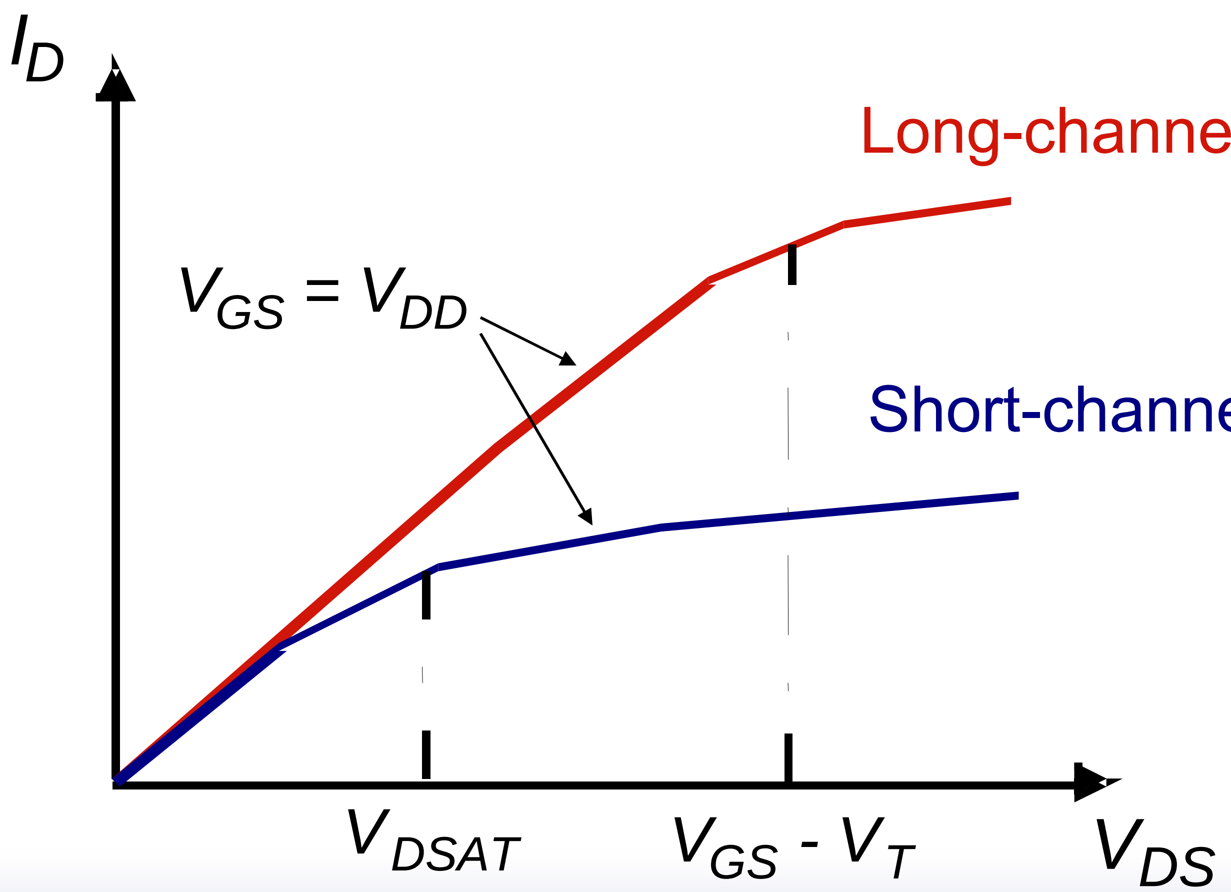
Current-Voltage Relations Short-Channel Transistors



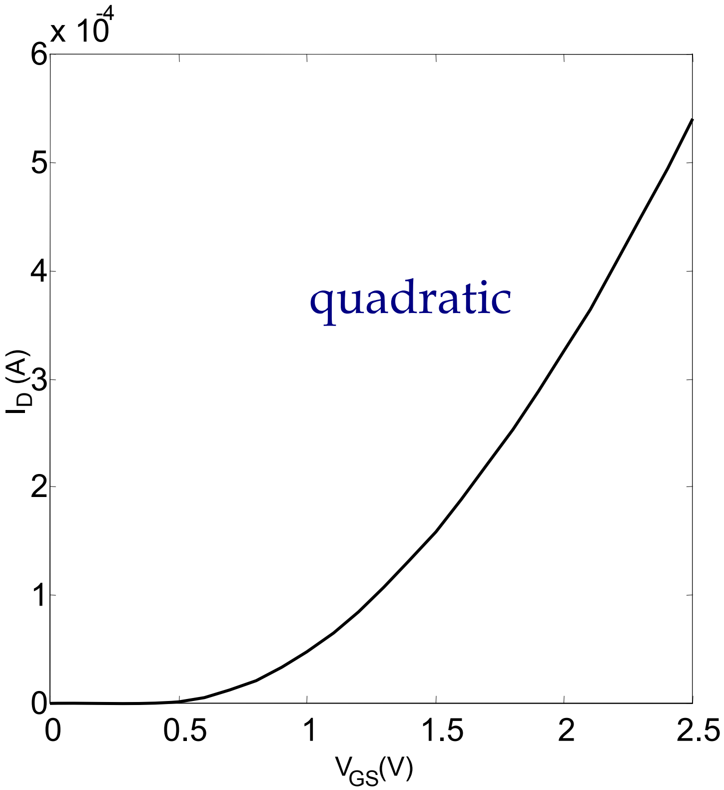
Velocity Saturation



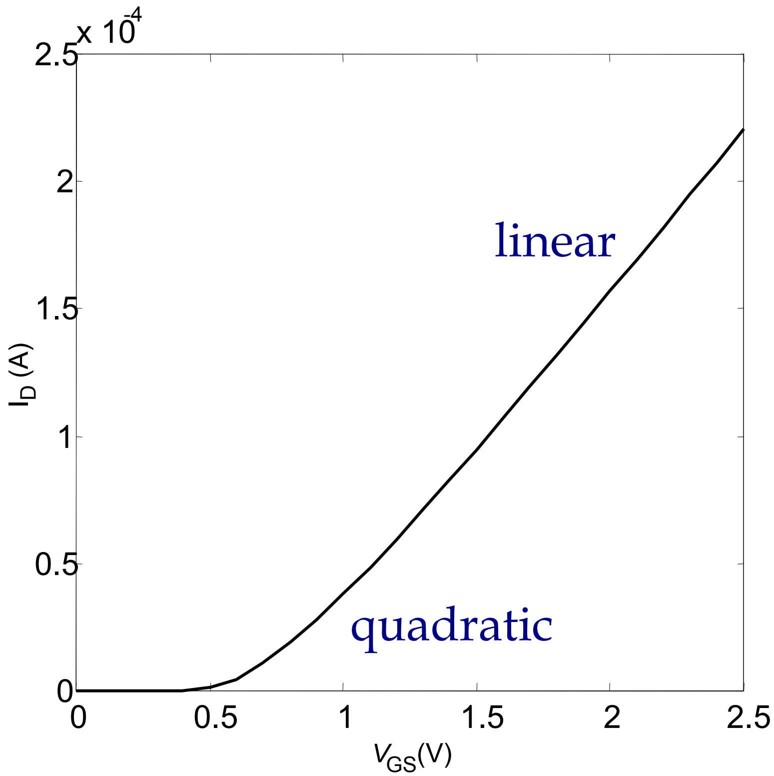
Perspective



I_D versus V_{GS}

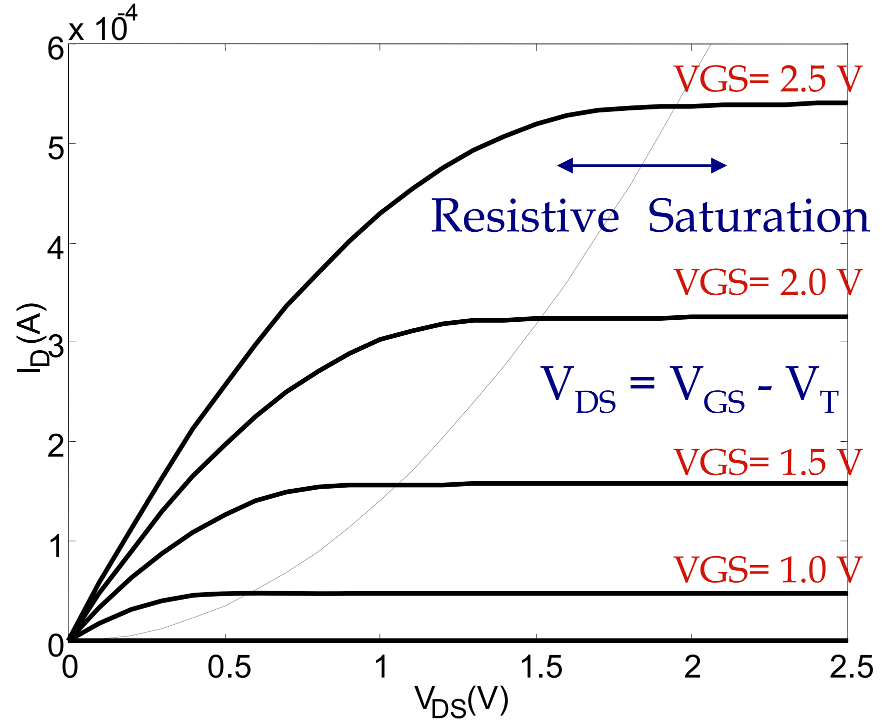


Long Channel

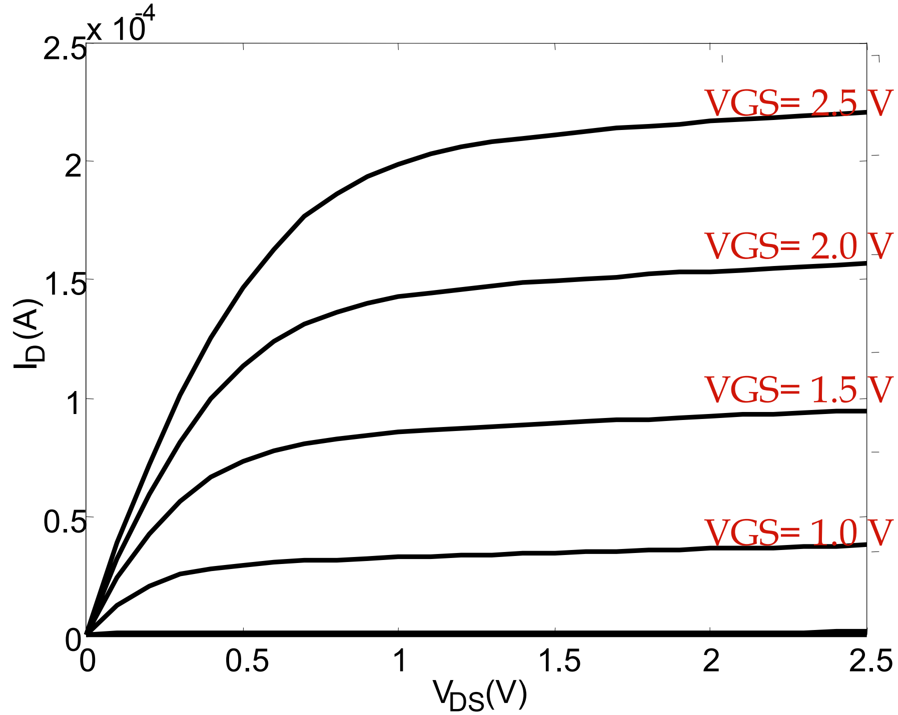


Short Channel

I_D versus V_{DS}

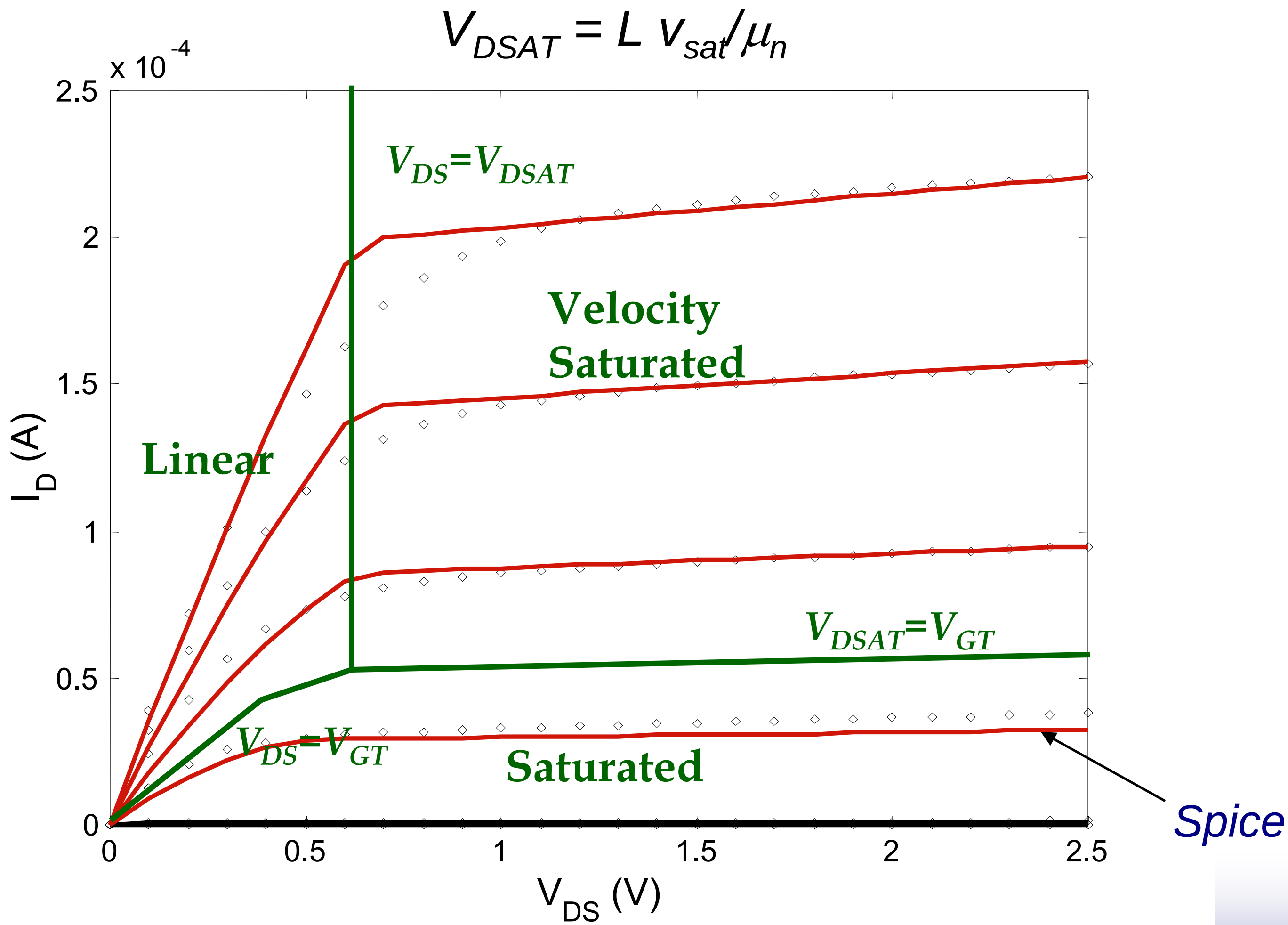


Long Channel

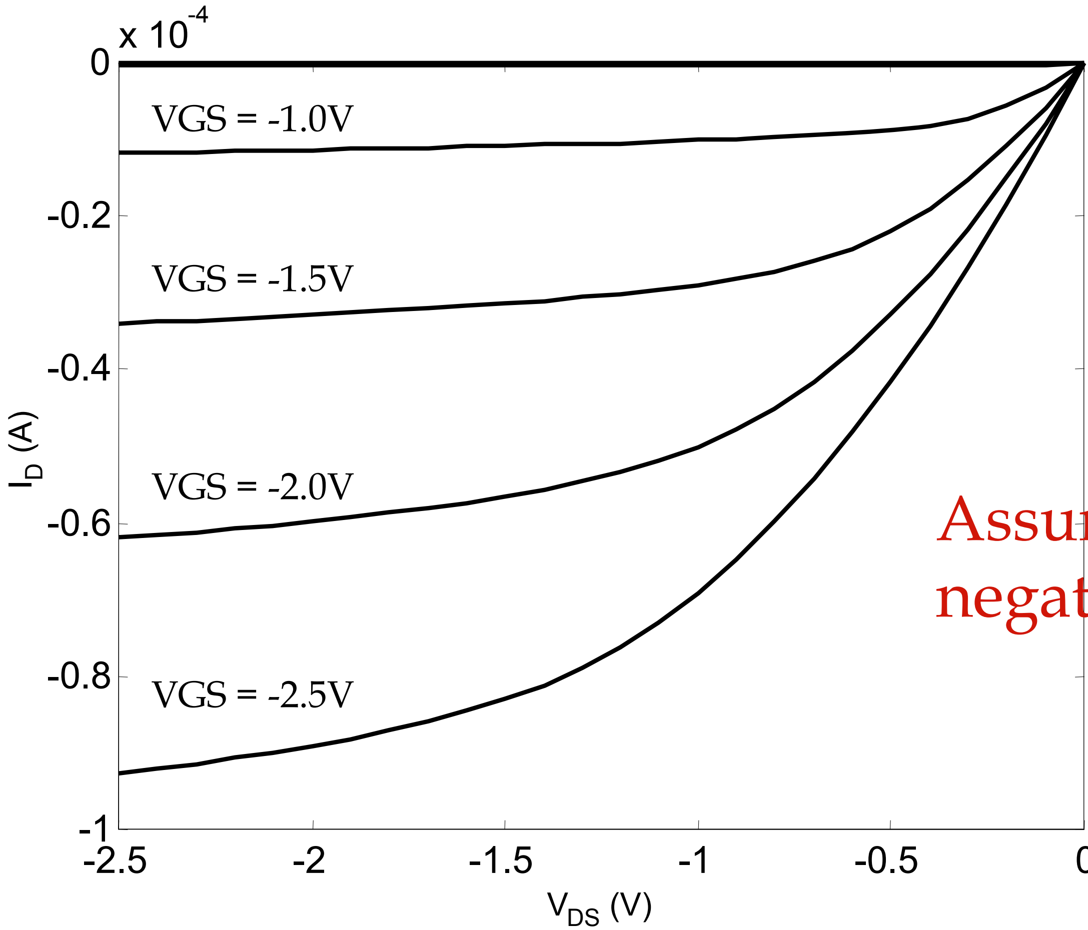


Short Channel

Simple Model versus SPICE



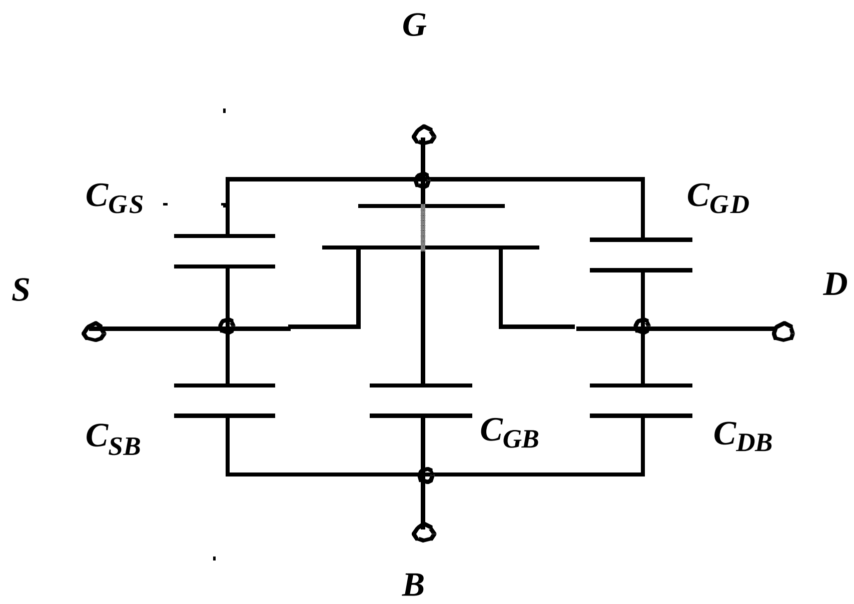
A PMOS Transistor (short-channel)



Assume all variables negative!

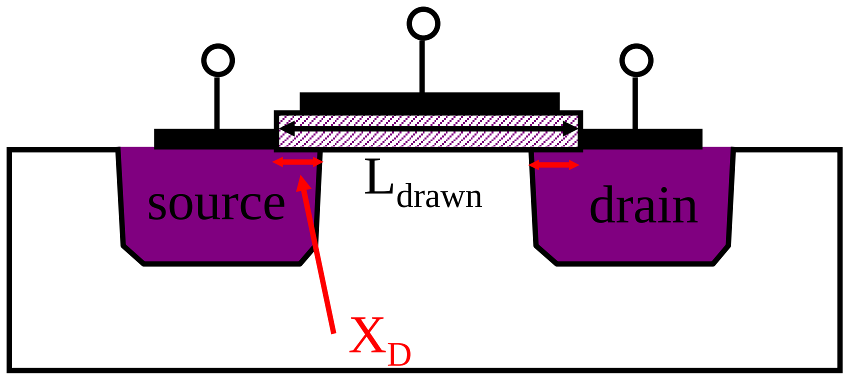
Dynamic Behavior of MOS Transistor

- Oxide Capacitance
  - Gate to Source overlap
  - Gate to Drain overlap
  - Gate to Channel/Bulk
- Junction Capacitance
  - Source to Bulk junction
  - Drain to Bulk junction



Oxide capacitances

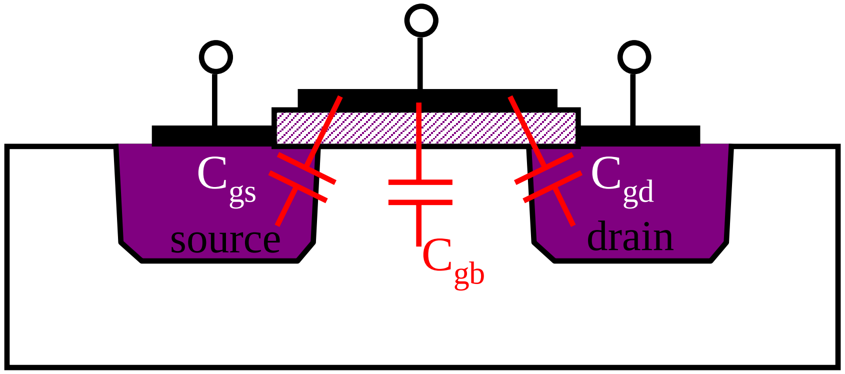
Overlap



- Overlap capacitances
  - gate electrode overlaps source and drain regions
  - X_D is overlap length on each side of channel
  - $L_{eff} = L_d - 2X_D$
  - Total overlap capacitance:

$$C_{overlap} = C_{GSO} + C_{GDO} = 2C_{ox}WX_D$$

Oxide capacitances
Channel



- Channel capacitances
  - Gate-to-source: C_{gs}
  - Gate-to-drain: C_{gd}
  - Gate-to-bulk: C_{gb}

- Cutoff:
  - No channel connecting source and drain
  - $C_{gs} = C_{gd} = 0$
  - $C_{gb} = C_{ox}WL_{eff}$
  - Total channel capacitance $C_{GC} = C_{ox}WL_{eff}$

Oxide capacitances

Channel

❑ Linear mode

- Channel spans from source to drain
- Capacitance split equally between S and D

$$C_{GS} = \frac{1}{2} C_{ox} W L_{eff} \qquad C_{GD} = \frac{1}{2} C_{ox} W L_{eff} \qquad C_{GB} = 0$$

– Total channel capacitance $C_{GC} = C_{ox} W L_{eff}$

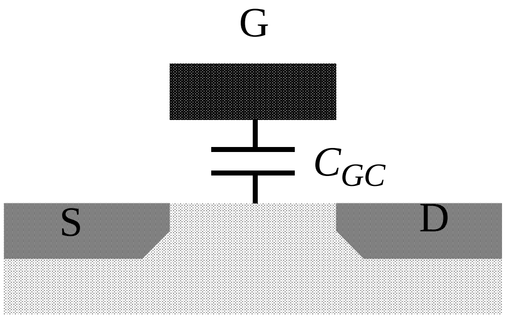
❑ Saturation mode

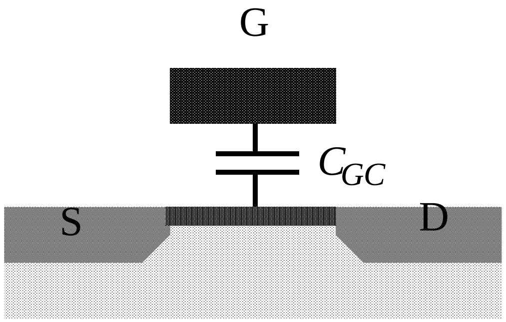
– Channel is pinched off:

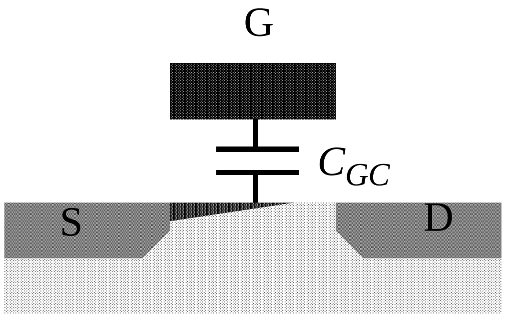
$$C_{GD} = 0 \qquad C_{GS} = \frac{2}{3} C_{ox} W L_{eff} \qquad C_{GB} = 0$$

– Total channel capacitance $C_{GC} = 2/3\ C_{ox} W L_{eff}$

Gate-to-Channel Capacitance (summary)



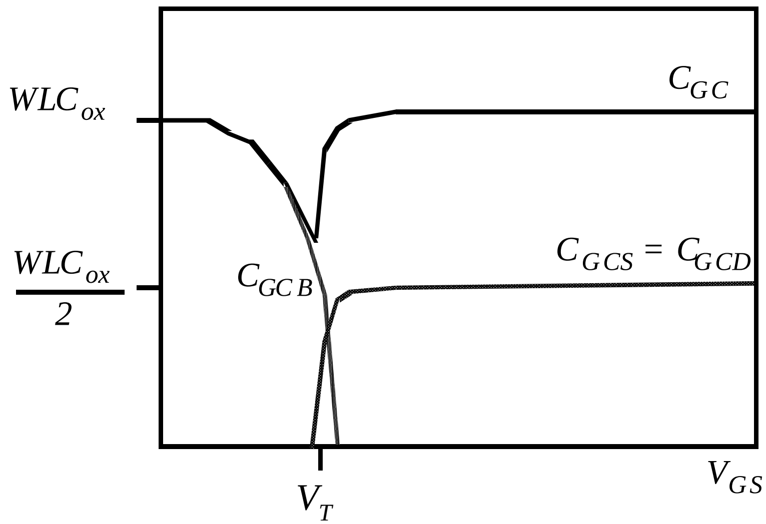


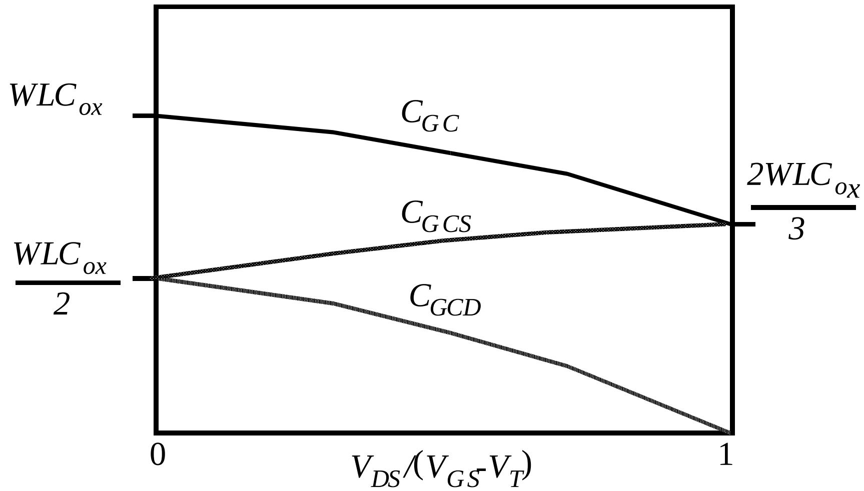


$$C_{GC} = C_{gb} + C_{gs} + C_{gd}$$

Operation Region	C_{gb}	C_{gs}	C_{gd}
Cutoff	$C_{ox}WL_{eff}$	0	0
Resistive	0	$C_{ox}WL_{eff}/2$	$C_{ox}WL_{eff}/2$
Saturation	0	$(2/3)C_{ox}WL_{eff}$	0

Gate-to-Channel Capacitance



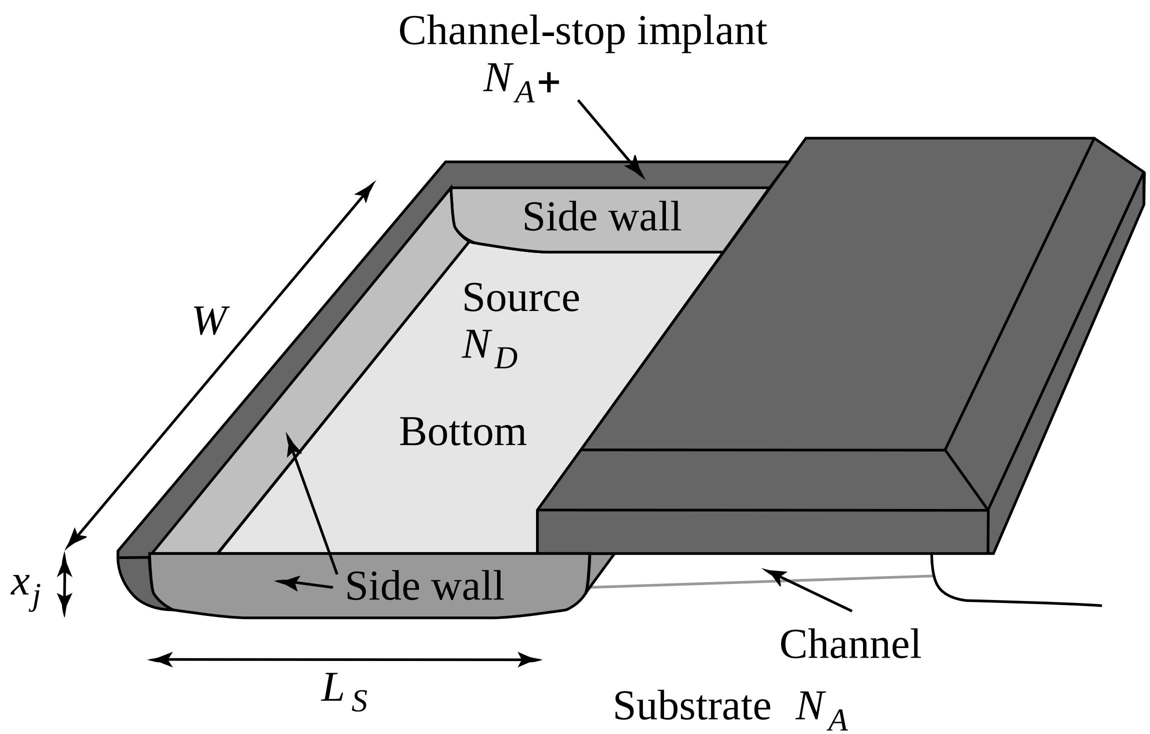


Capacitance as a function of VGS (with VDS = 0)

Capacitance as a function of the degree of saturation

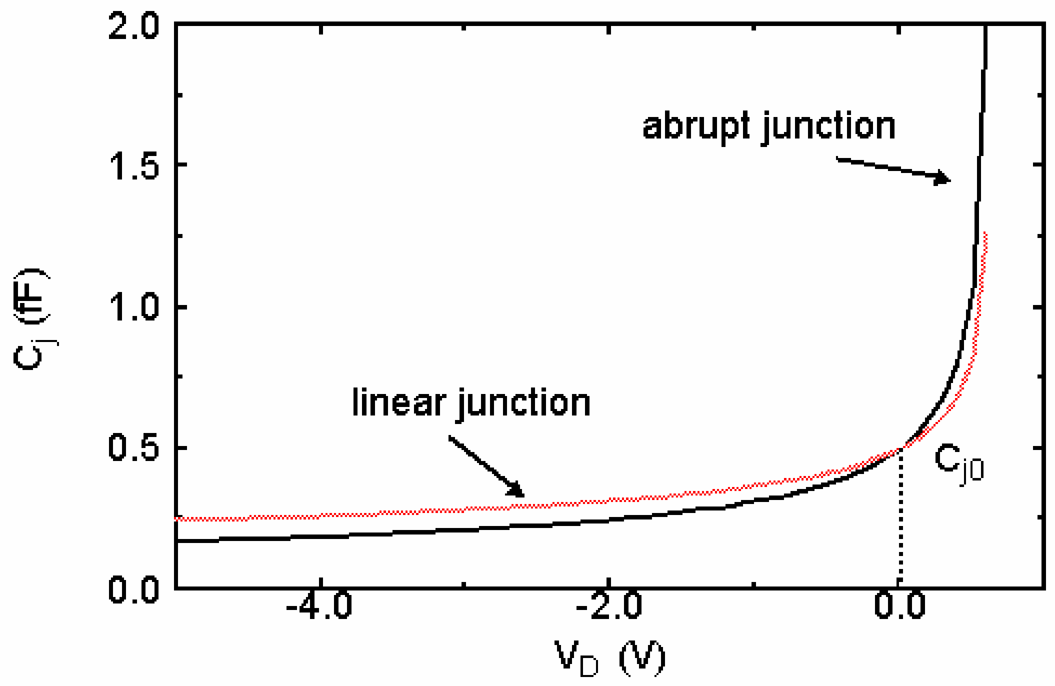
Bottom Line: Cap. components are non-linear

Diffusion Capacitance



$$C_{diff} = C_{bottom} + C_{sw} = C_j \times AREA + C_{jsw} \times PERIMETER$$
$$= C_j L_S W + C_{jsw}(2L_S + W)$$

Junction Capacitance



$$C_j = \frac{C_{j0}}{(1 - V_D / \phi_0)^m}$$

m = 0.5: abrupt junction
m = 0.33: linear junction

Linearizing the Junction Capacitance

Replace non-linear capacitance by
large-signal equivalent linear capacitance
which displaces equal charge
over voltage swing of interest

$$C_{eq} = \frac{\Delta Q_j}{\Delta V_D} = \frac{Q_j(V_{high}) - Q_j(V_{low})}{V_{high} - V_{low}} = K_{eq} C_{j0}$$

$$K_{eq} = \frac{-\phi_0^m}{(V_{high} - V_{low})(1-m)}[(\phi_0 - V_{high})^{1-m} - (\phi_0 - V_{low})^{1-m}]$$

Capacitances in 0.25 μm CMOS process

	C_{ox} (fF/μm^2)	C_O (fF/μm)	C_j (fF/μm^2)	m_j	ϕ_b (V)	C_{jsw} (fF/μm)	m_{jsw}	ϕ_{bsw} (V)
NMOS	6	0.31	2	0.5	0.9	0.28	0.44	0.9
PMOS	6	0.27	1.9	0.48	0.9	0.22	0.32	0.9

Data Dependency

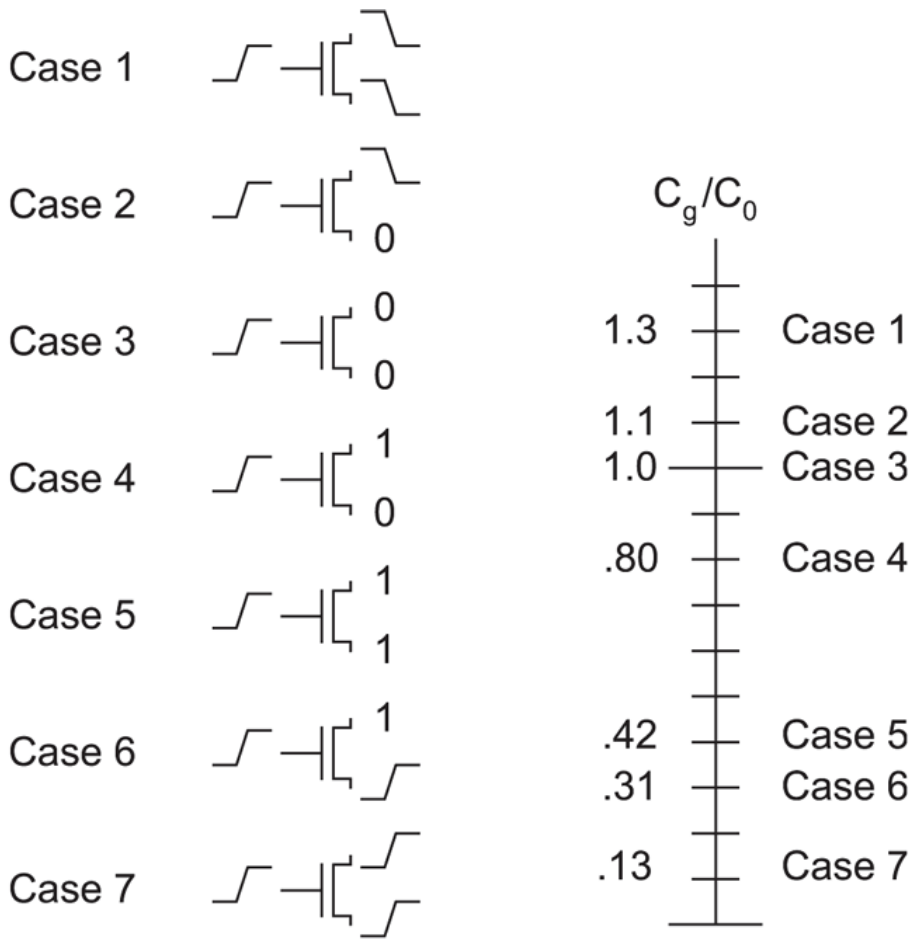


FIG 2.12 Data-dependent gate capacitance

MOS Cap. Summary

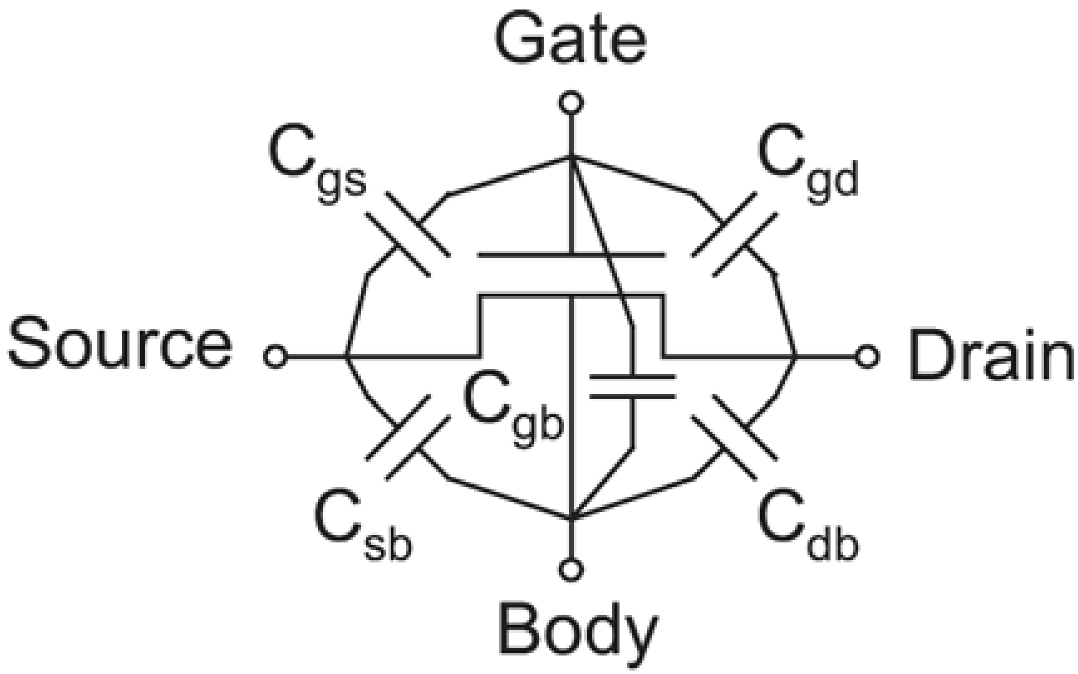


FIG 2.14 Capacitances of an MOS transistor

Alpha-Power MOSFET Model

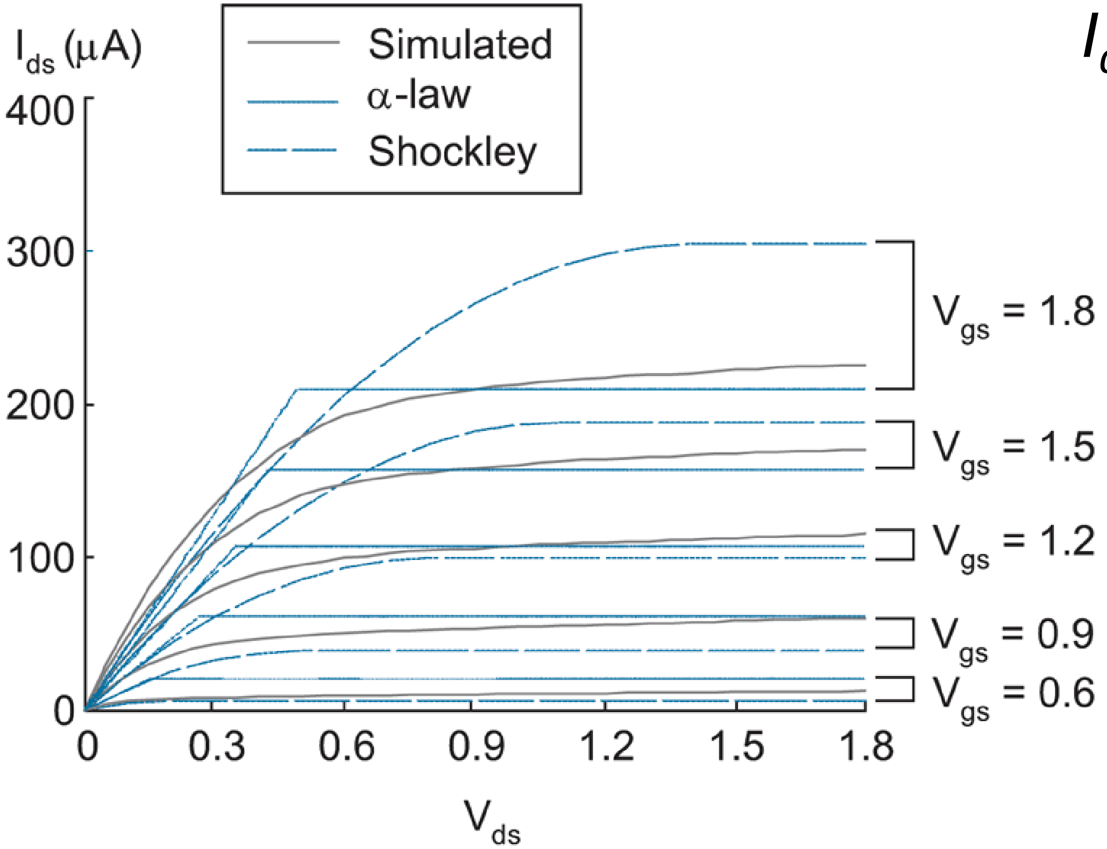


FIG 2.17 I-V characteristics for nMOS transistor with velocity saturation

$$I_{ds} \propto (V_{gs} - V_t)^{\alpha}$$

$$1 < \alpha < 2$$

Subthreshold Leakage

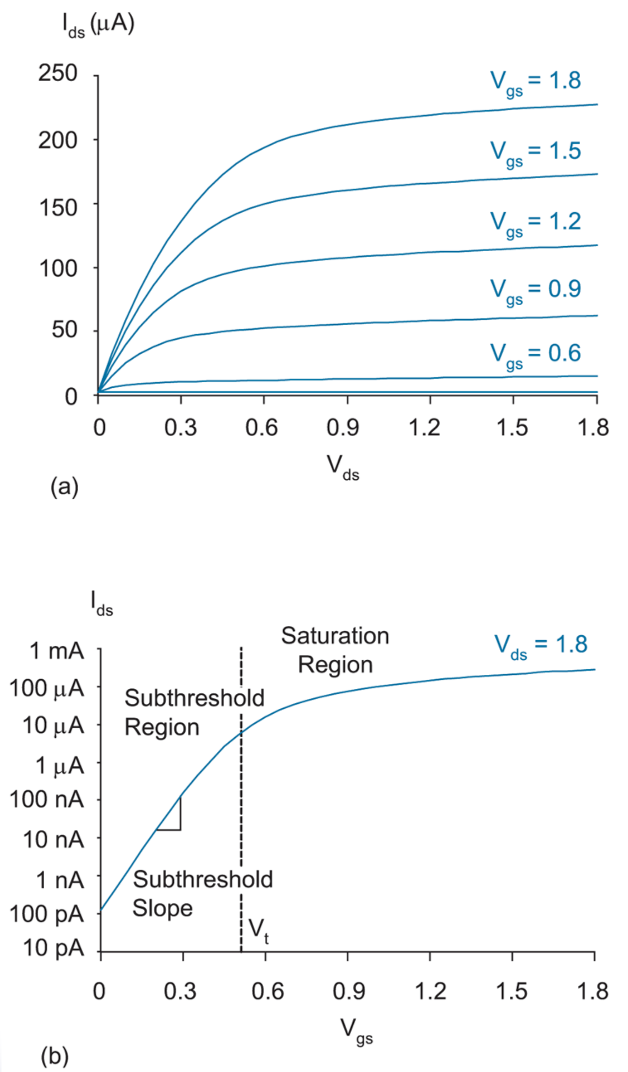


FIG 2.15 Simulated I-V characteristics

- *Dominant leakage mechanism*
- *Increases exponentially with temperature and Vt*

Gate Leakage

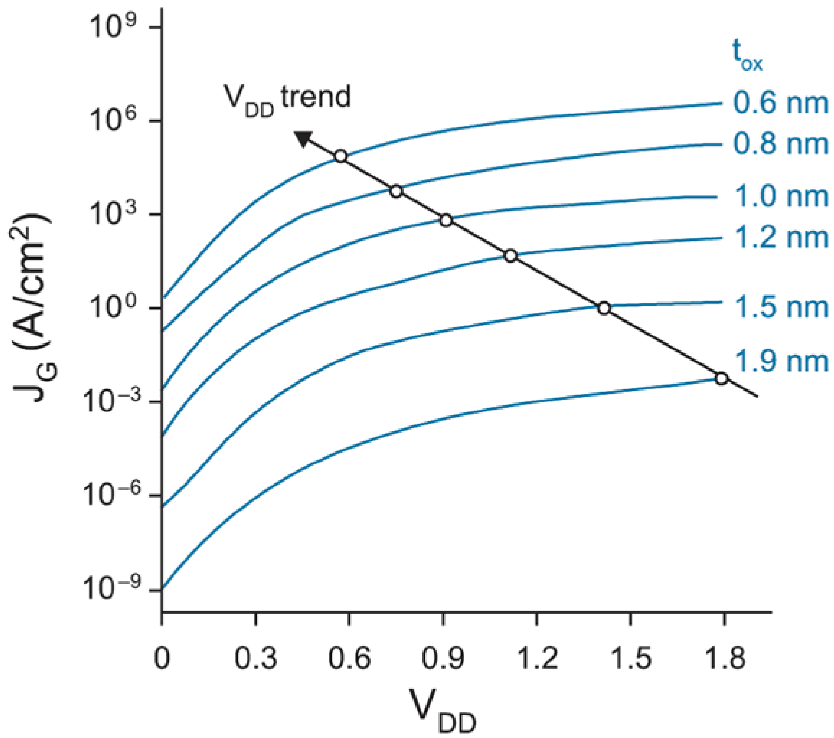


FIG 2.20 Gate leakage current from [Song01]

- *Increases with gate oxide (SiO2) scaling*
- *High-k gate oxides can be used to lower gate leakage*
- *Independent of temperature*

Junction Leakage

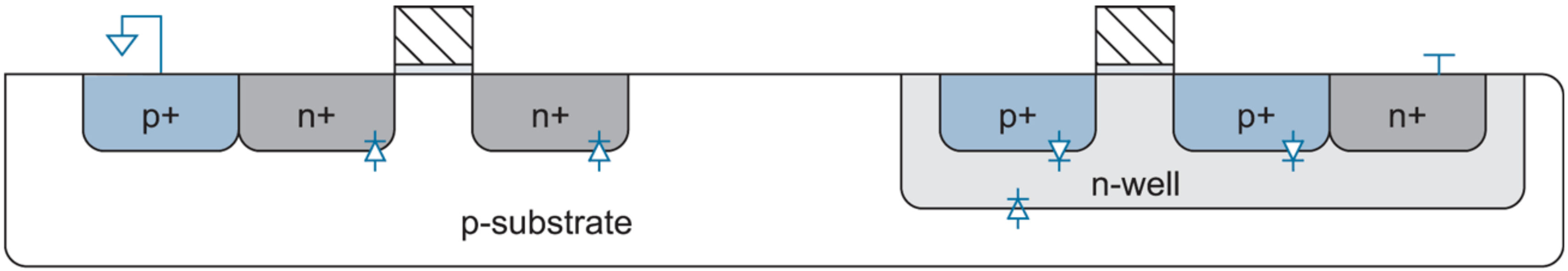


FIG 2.19 Reverse-biased diodes in CMOS circuits

- *Less significant than gate and subthreshold leakage*
- *Increases with temperature*

Temperature Effects

- *Mobility decreases with increase in T*
- *Vt decreases linearly with T*

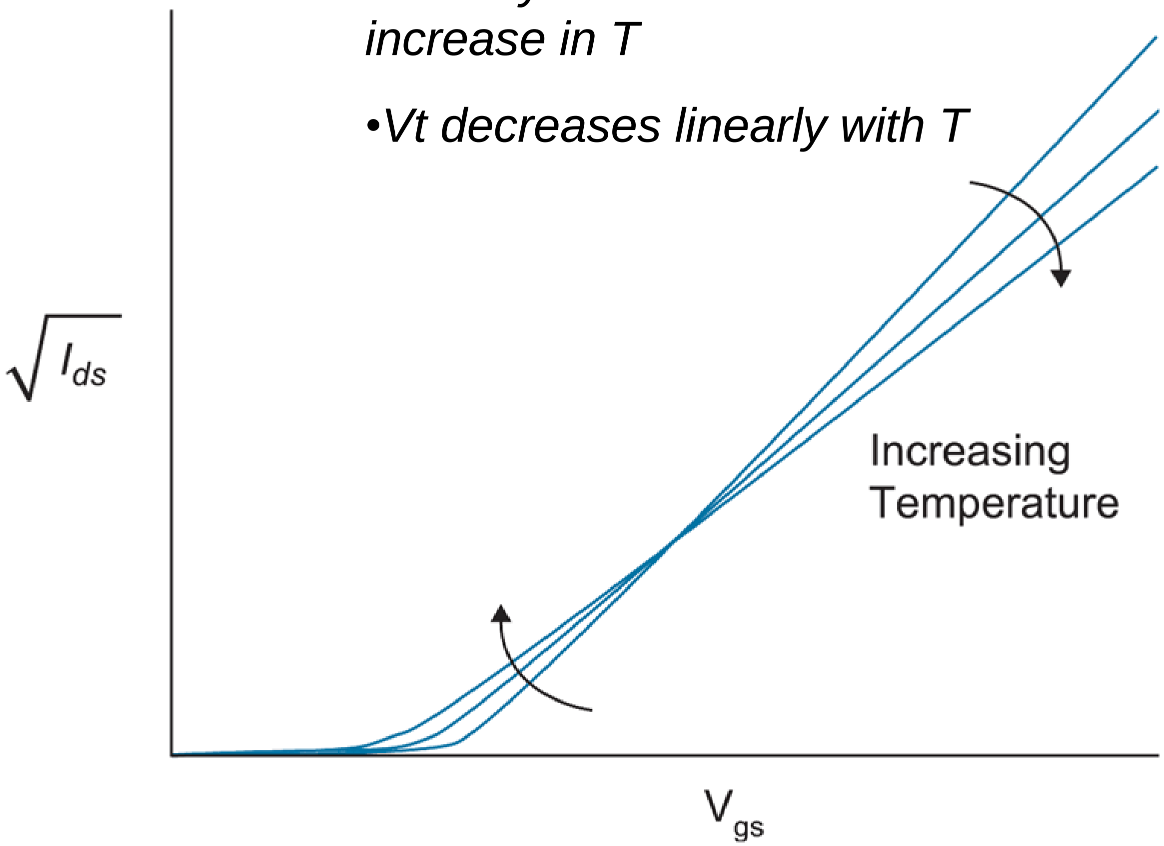


FIG 2.21 I–V characteristics of nMOS transistor in saturation at various temperatures

Temperature Effects

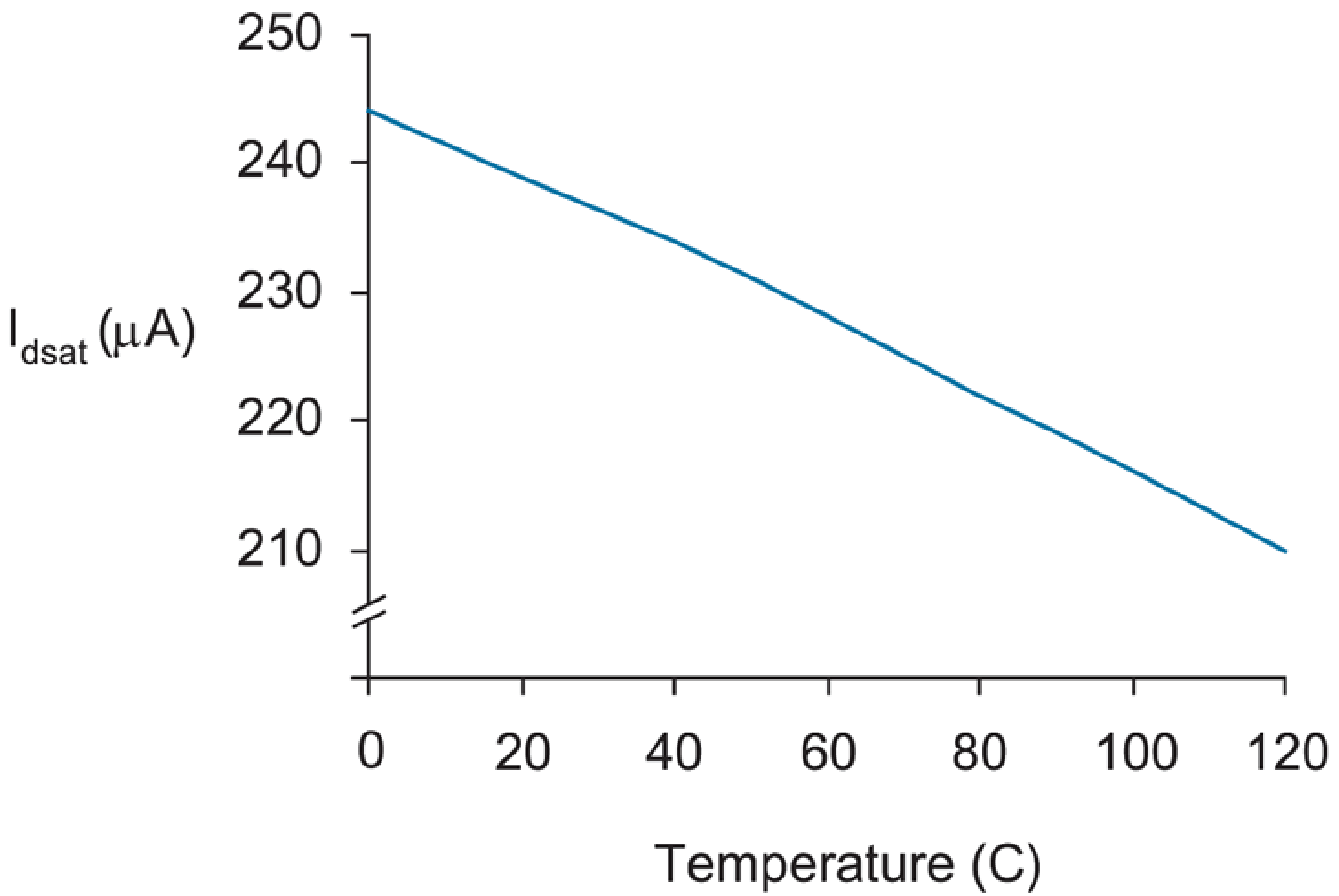


FIG 2.22 I_{dsat} vs. temperature

Temperature Effects

Chip Cooling can:

1. *Improve Circuit performance*
    - *speed up transistors*
    - *decrease the delay of interconnects since metal resistance decreases with temperature*
    - *Lowers junction capacitance (increases depletion width)*
2. *Decrease leakage (mainly subthreshold)*
3. *Improve reliability of the chip*

The Sub-Micron MOS Transistor

- **Threshold Variations**
- **Subthreshold Conduction/Leakage**
- **Parasitic Resistances**

Threshold voltage variation

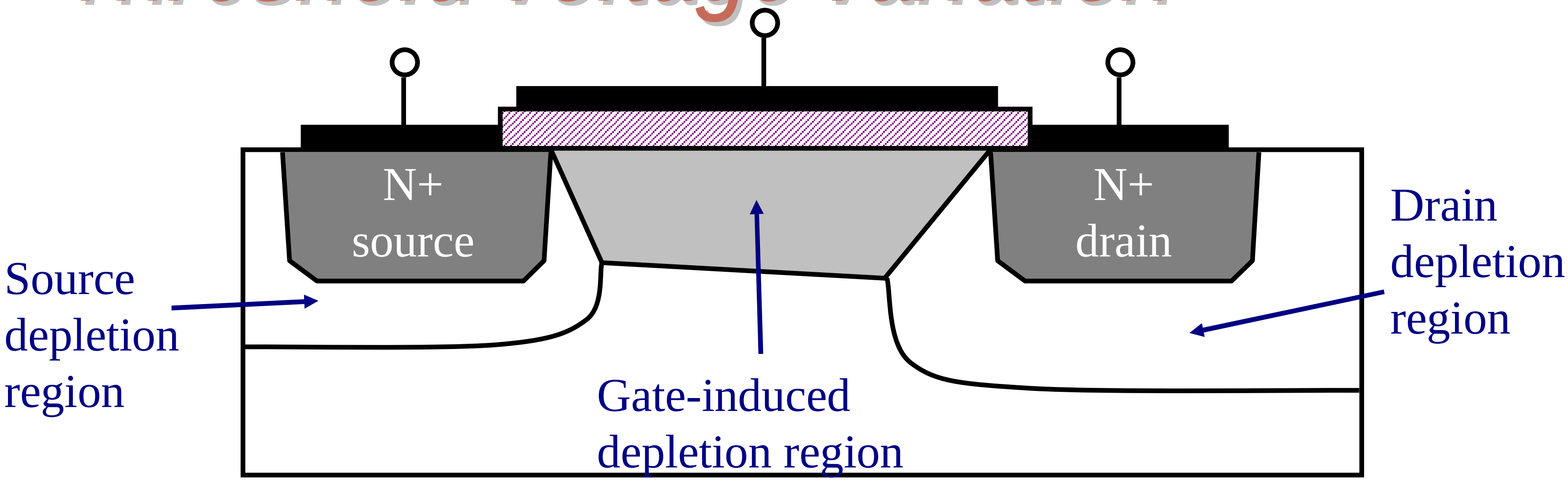
- Until now, threshold voltage assumed constant
  - V_T changed only by substrate bias V_{SB}
- In threshold voltage equations, channel depletion region assumed to be created by gate voltage only
- Depletion regions around source and drain neglected: valid if channel length is much larger than depletion region depths
- In short-channel devices, depletion regions from drain and source extend into channel

Threshold voltage variation

Short-channel effects cause threshold voltage variation:

- V_T rolloff
  - As channel length L decreases, threshold voltage decreases
- Drain-induced barrier lowering
  - As drain voltage V_{DS} increases, threshold voltage decreases
- Hot-carrier effect
  - Threshold voltages drift over time
- Negative-Bias Temperature Instability (NBTI)
  - Issue in PMOS transistors
  - Vt drifts over time
  - Typical stress temperature 100-150 C
  - Typical oxide electric fields of 5-6 MV/cm

Threshold voltage variation



- Even with V_{GS}=0, part of channel is already depleted
- Bulk depletion charge is smaller in short-channel device → V_T is smaller

Threshold voltage variation

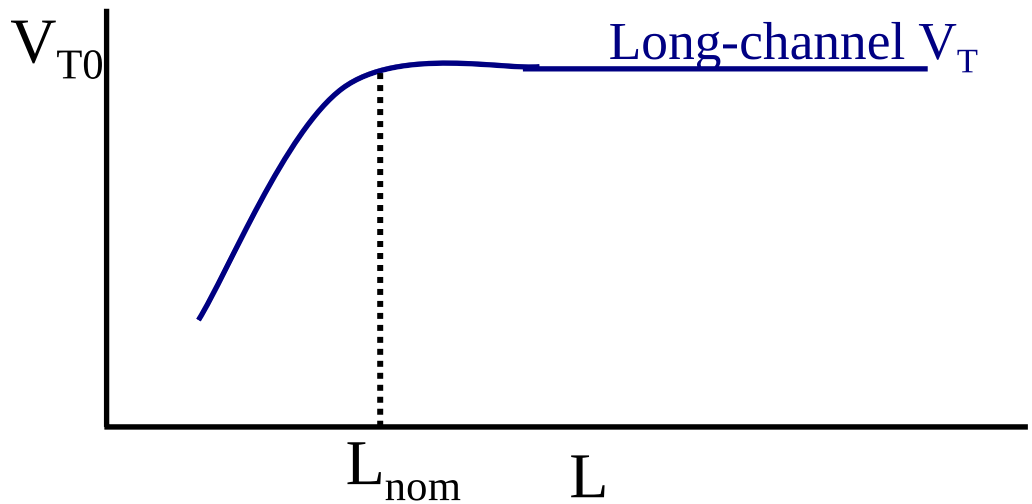
- Change in V_{T0}:
  - x_{dS}, x_{dD}: depth of depletion regions at S, D
  - x_j: junction depth

$$\Delta V_{T0} = \frac{1}{C_{ox}}\sqrt{2q\varepsilon_{Si}N_A|2\phi_F|} \bullet \frac{x_j}{2L}\left[\left(\sqrt{1+\frac{2x_{dS}}{x_j}}-1\right)+\left(\sqrt{1+\frac{2x_{dD}}{x_j}}-1\right)\right]$$

- ΔV_{T0} is proportional to (x_j/L)
  - For short channel lengths, ΔV_{T0} is large
  - For large channel lengths, term approaches 0

Threshold voltage variations

Graphically: V_{T0} versus channel length L



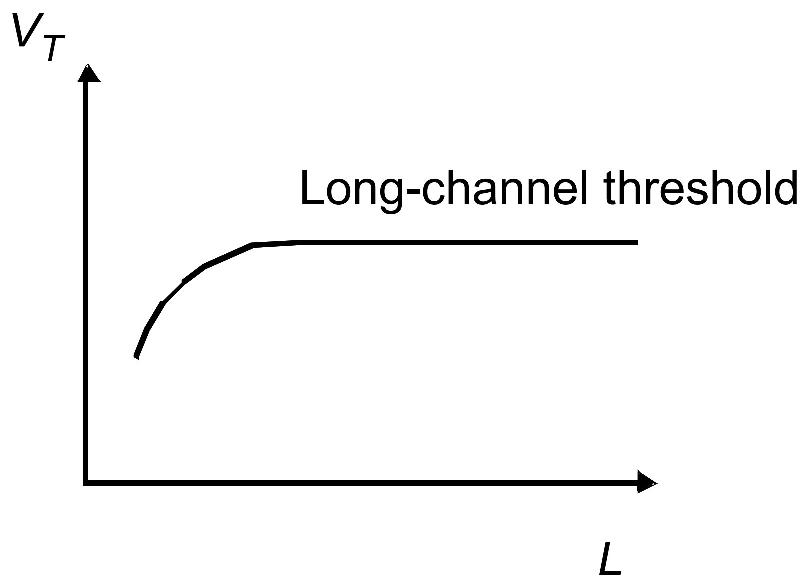
V_T Roll-off:

V_T decreases rapidly with channel length

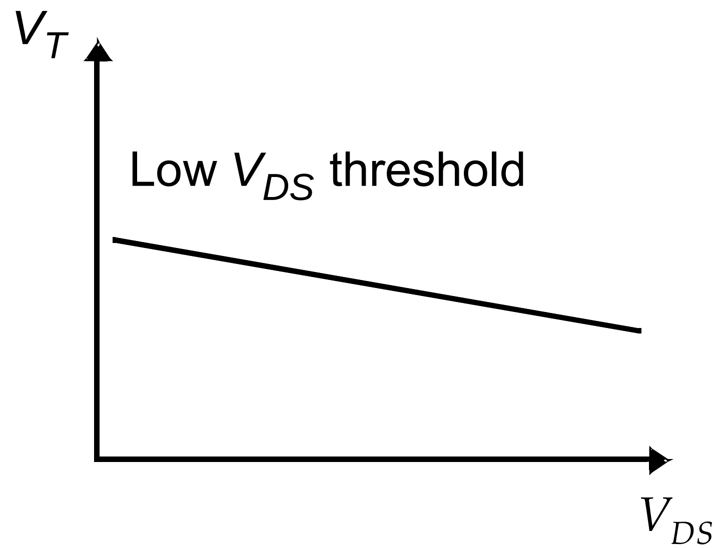
DIBL

- Drain-induced barrier lowering (DIBL)
  - Drain voltage V_{DS} causes change in threshold voltage
  - As V_{DS} is increased, threshold voltage decreases
- Cause: depletion region around drain
  - Depletion region depth around drain depends on drain voltage
  - As V_{DS} is increased, drain depletion region gets deeper and extends further into channel
  - For very large V_{DS}, source and drain depletion regions can meet → *punch-through*!
- Issue: results in uncertainty in circuit design

Threshold Variations



Threshold as a function of the length (for low V_{DS})



Drain-induced barrier lowering (DIBL) (for low L)

Threshold voltage variation

- Hot-carrier effect
  - increased electric fields causes increased electron velocity
  - high-energy electrons can tunnel into gate oxide
  - This changes the threshold voltage (increases V_T for NMOS)
  - Can lead to long-term reliability problems

Threshold voltage variation

- Hot electrons
  - High-velocity electrons can also impact the drain, dislodging holes
  - Holes are swept towards negatively-charged substrate → cause substrate current
  - Called *impact ionization*
  - This is another factor which limits the process scaling → voltage must scale down as length scales

Threshold voltage variations

- Summary of threshold variations in short-channel devices
  - V_T rolloff: threshold voltage reduces as channel length L reduces
  - DIBL: threshold voltage reduces as V_{DS} increases
  - Hot-carrier effect: threshold voltage drifts over time as electrons tunnel into oxide
  - NBTI—causes Vt increase in PMOS transistors, strong dependence on Temperature.

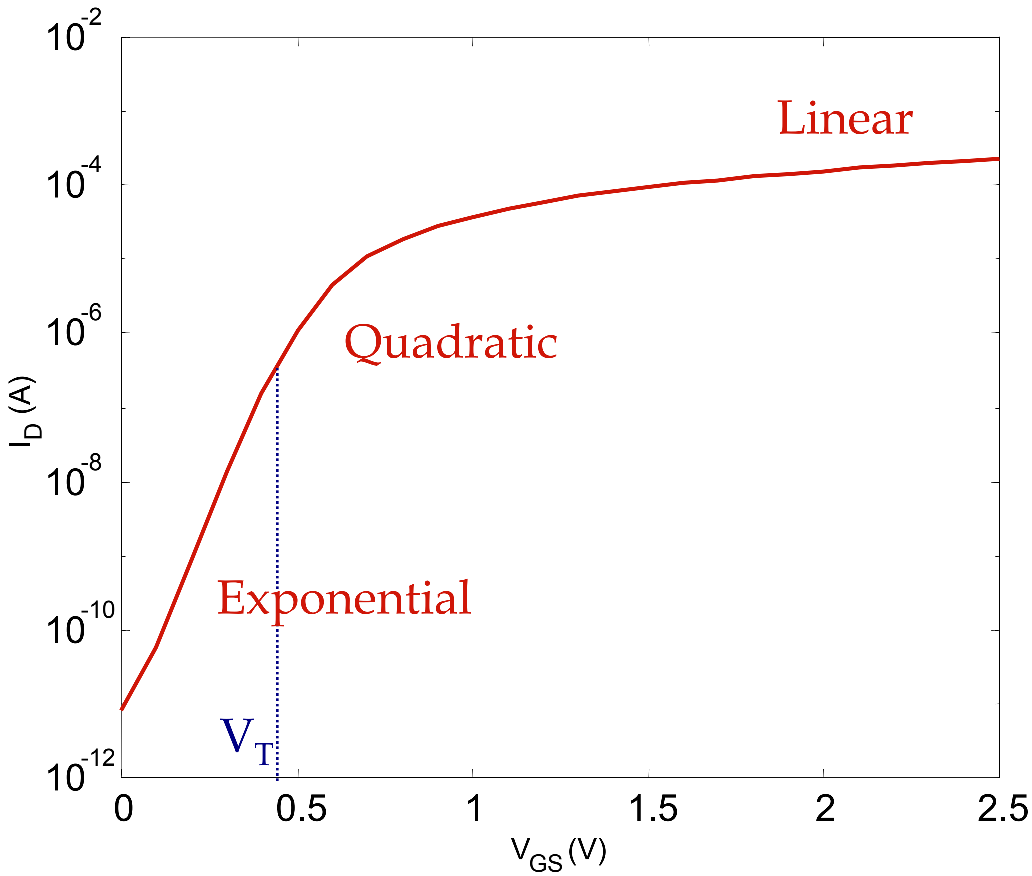
Sub-threshold conduction (1)

- When $V_{GS} < V_T$, transistor is "off"
  - However, small drain current I_D still flows
  - Called *subthreshold leakage* current
- Model for subthreshold current:

$$I_D(subthreshold) = I_S W e^{\frac{q}{kT}(AV_{GS}+BV_{DS})}$$

  - Increases as V_{GS} increases (potential barrier lowered)
  - Increases as V_{DS} increases (DIBL)

Sub-Threshold Conduction (2)



The Slope Factor

$$I_D \sim I_0 e^{\frac{qV_{GS}}{nkT}}, \quad n = 1 + \frac{C_D}{C_{ox}}$$

S is ΔV_{GS} for $I_{D2}/I_{D1} = 10$

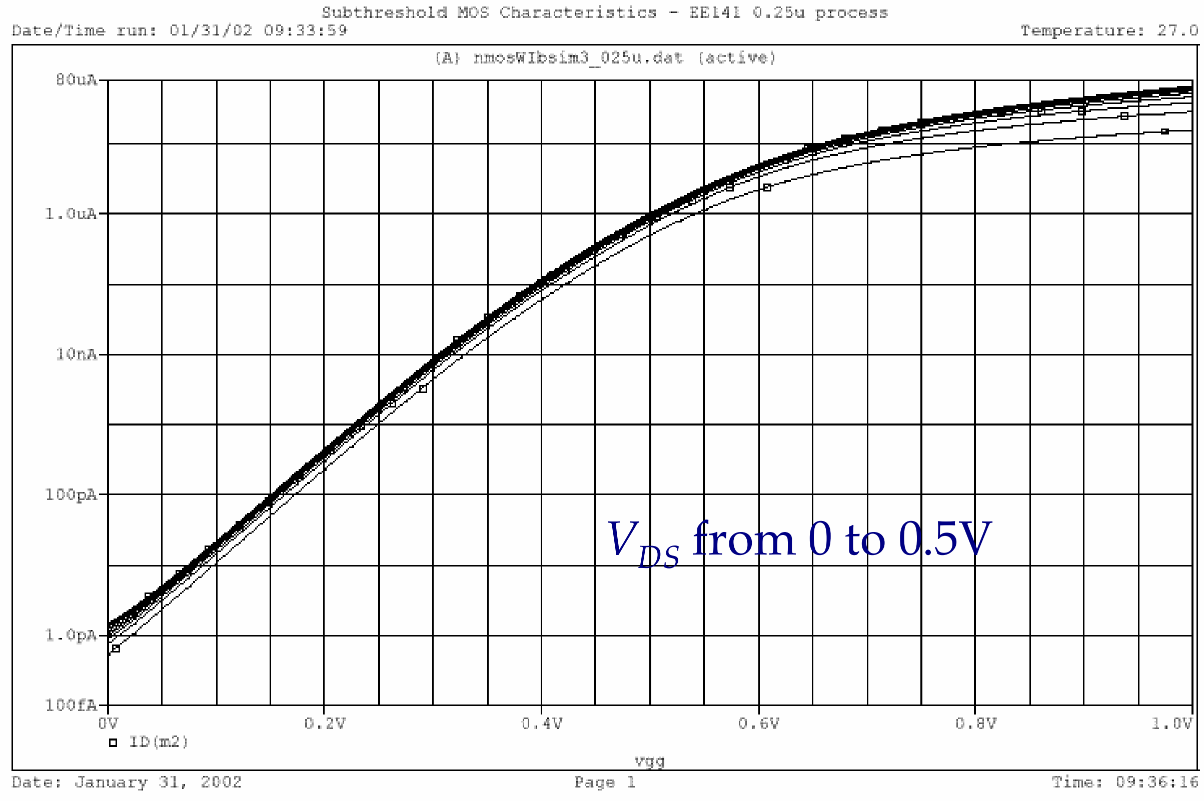
$$S = n\left(\frac{kT}{q}\right)\ln(10)$$

Typical values for S:
60 .. 100 mV/decade

Kaustav Banerjee

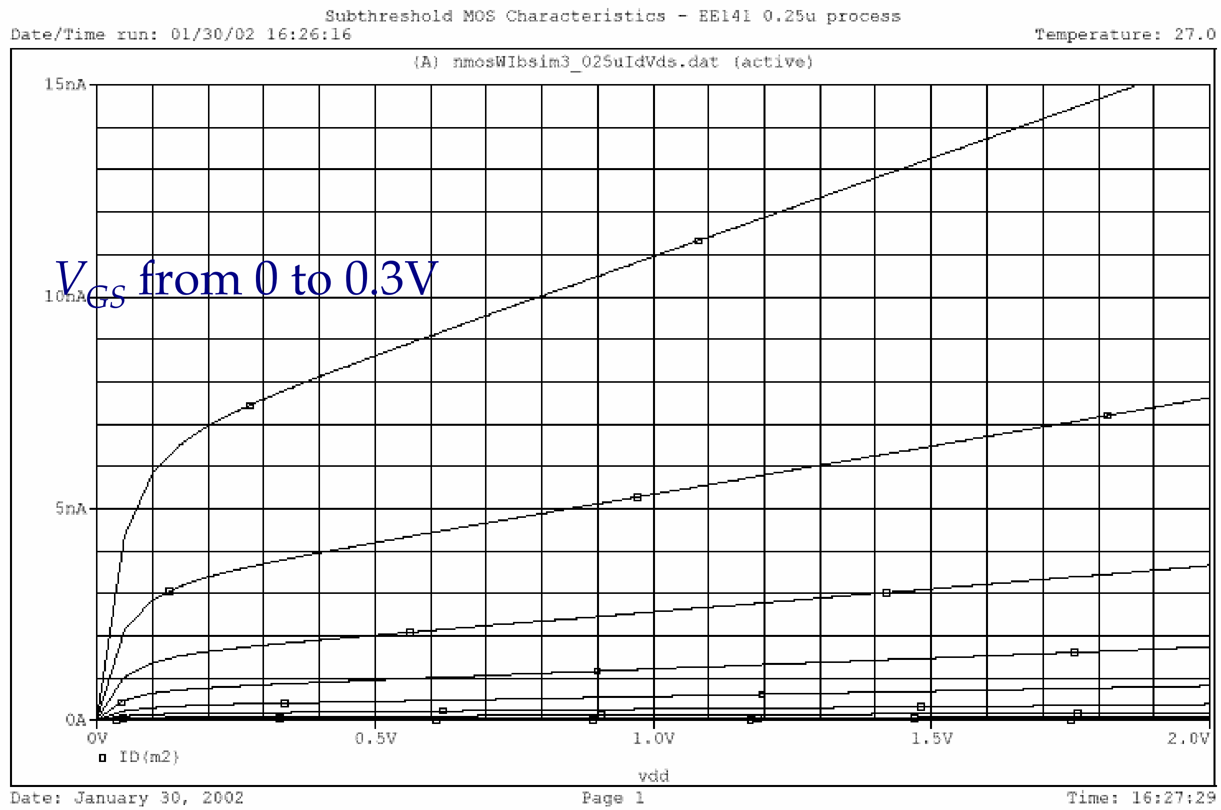
Sub-Threshold I_D vs V_{GS}

$$I_D = I_0 e^{\frac{qV_{GS}}{nkT}} \left(1 - e^{-\frac{qV_{DS}}{kT}}\right)$$



Sub-Threshold I_D vs V_{DS}

$$I_D = I_0 e^{\frac{qV_{GS}}{nkT}} \left(1 - e^{-\frac{qV_{DS}}{kT}}\right)\left(1 + \lambda \cdot V_{DS}\right)$$



Leakage

- Effect of leakage current
  - "Wasted" power: power consumed even when circuit is inactive
  - Leakage power raises temperature of chip
  - Can cause functionality problem in some circuits: memory, dynamic logic, etc.
- Reducing transistor leakage
  - Long-channel devices
  - Small drain voltage
  - Large threshold voltage V_T

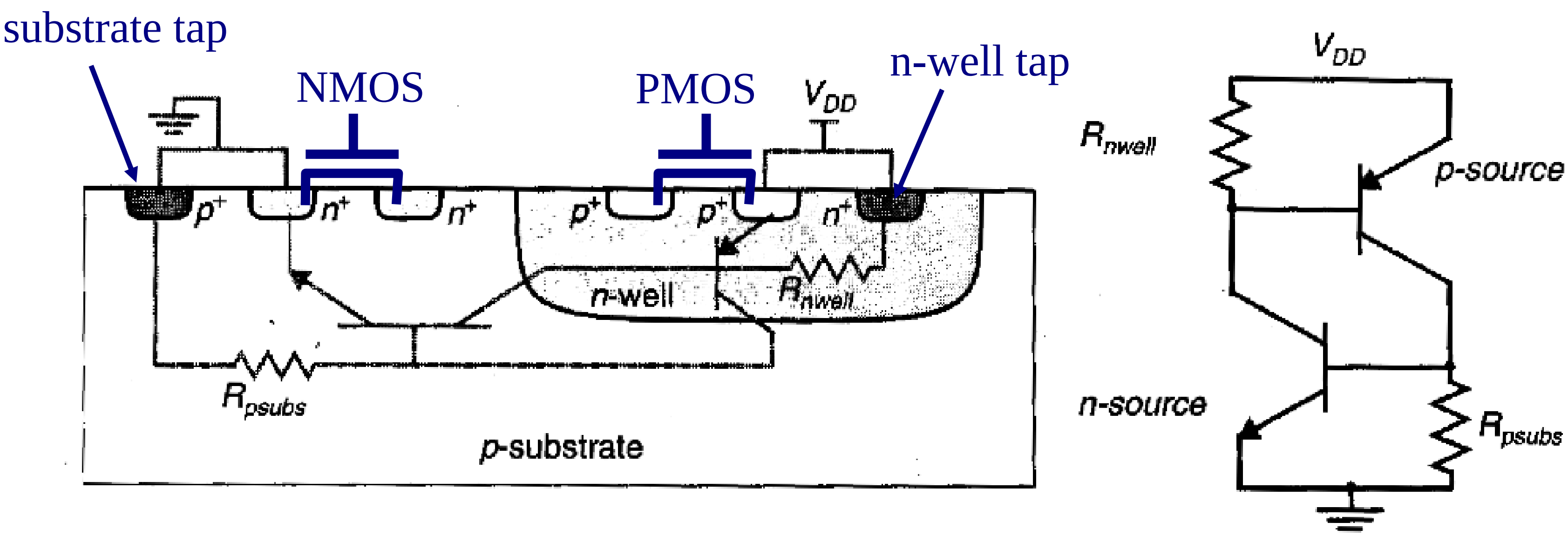
Leakage

- Leakage vs. performance tradeoff:
  - For high-speed, need small V_T and L
  - For low leakage, need high V_T and large L
- Process scaling
  - V_T reduces with each new process (historically)
  - Leakage increases ~10X!
- One solution: dual-V_T process
  - Low-V_T transistors: use in critical paths for high speed
  - High-V_T transistors: use to reduce power

Latchup

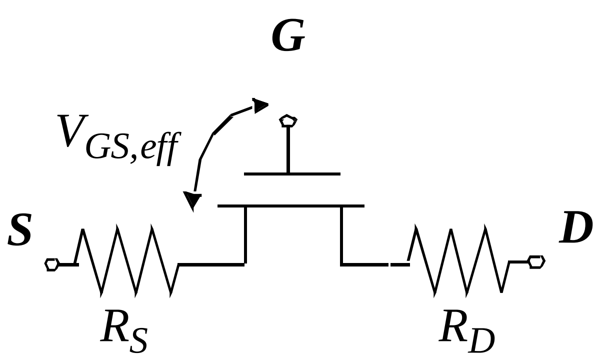
- CMOS process contains parasitic bipolar transistors
- Under certain conditions, these parasitic transistors can turn on, shorting power and ground rails and usually destroying the chip → *latchup*
- Avoiding latchup requires certain layout design rules, and careful control of process
- Latchup was a major problem in early CMOS processes
- Now, latchup is mainly issue for I/O circuits, with high current demands and possibly noisy voltages

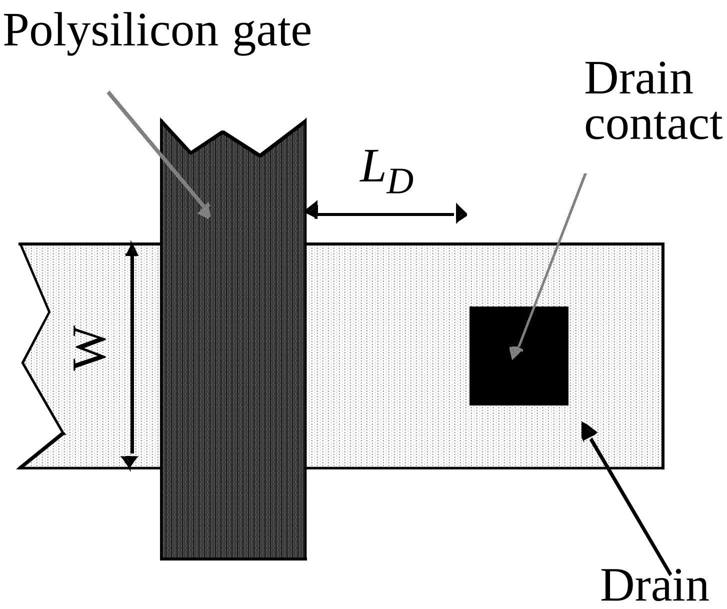
Latchup



- Current flowing in well or substrate can forward-bias bipolar transistor
- Positive feedback between transistors: when one turns on, V_{dd} and Gnd are connected
- Solution: reduce R_{nwell} and R_{psubs}: use many substrate taps in layout
- High-current circuits use *guard rings*

Parasitic Resistances

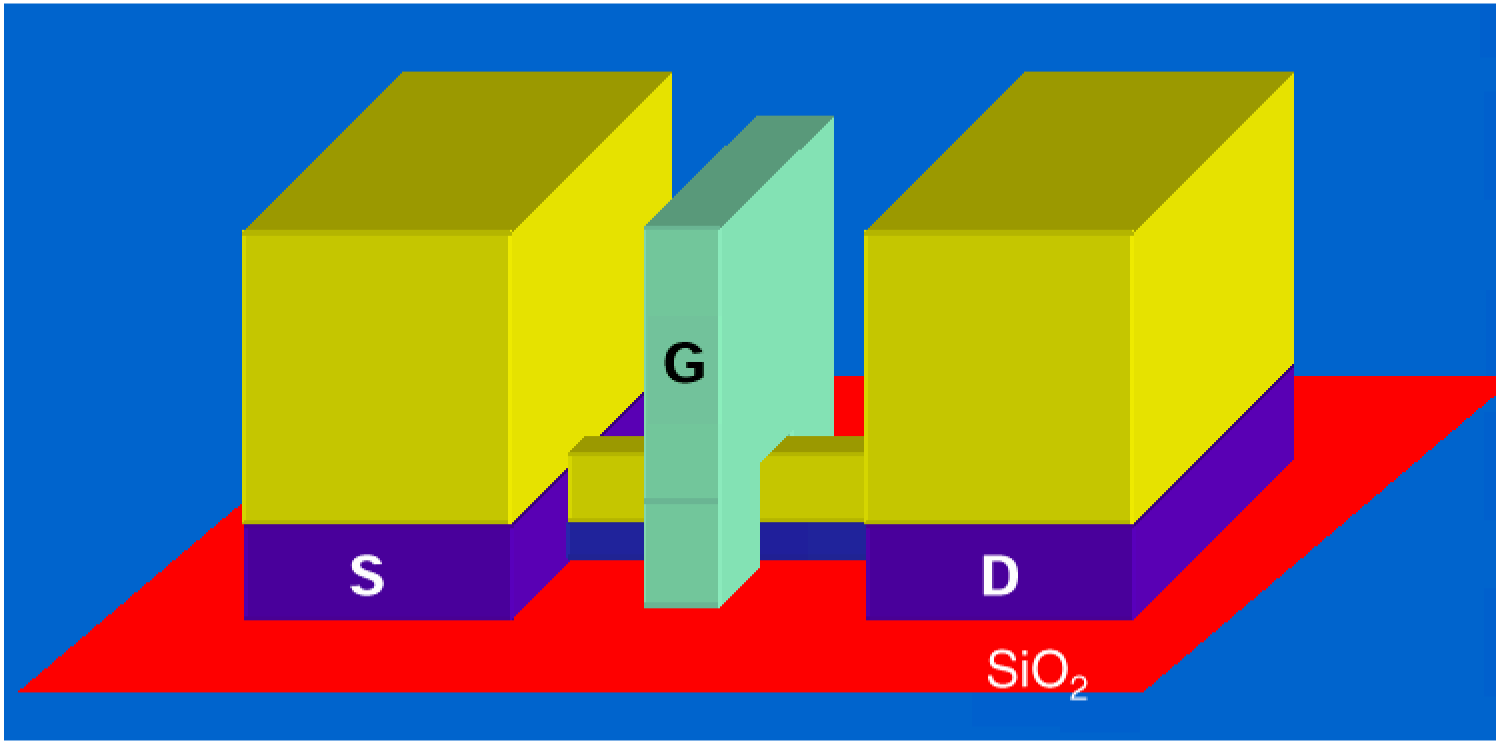




Problem can be alleviated by silicided source/drain contacts

Silicide thickness is an important factor….

Future Perspectives……



25 nm FINFET MOS transistor (Berkeley)