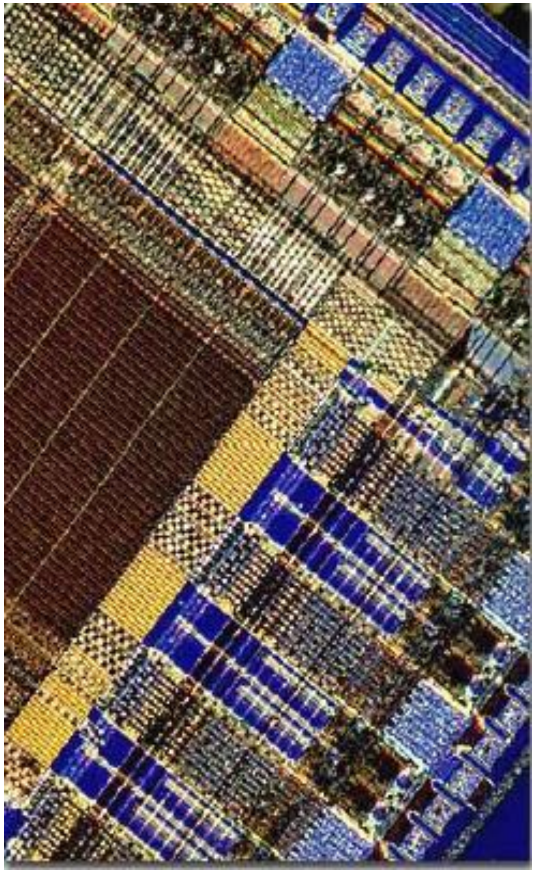
CMOS Digital Integrated Circuits



Lec 4

MOS Transistor II

MOS Transistor II

Goals

- Understand constant field and constant voltage scaling and their effects.
- Understand small geometry effects for MOS transistors and their implications for modeling and scaling
- Understand and model the capacitances of the MOSFET as:
  - » Lumped, voltage-dependent capacitances
  - » Lumped, fixed-value capacitances
- Be able to calculate the above capacitances from basic parameters.

MOS Scaling
What is Scaling?

- Reduction in size of an MOS chip by reducing the dimensions of MOSFETs and interconnects.
- Reduction is symmetric and preserves geometric ratios which are important to the functioning of the chip. Ideally, allows design reuse.
- Assume that ***S*** is the scaling factor. Then a transistor with original dimensions of ***L*** and ***W*** becomes a transistor with dimensions ***L/S*** and ***W/S***.
- Typical values of ***S***: **1.4** to **1.5** per biennium:
- Two major forms of scaling

  Full scaling (constant-field scaling) – All dimensions are scaled by ***S*** and the supply voltage and other voltages are so scaled.

  Constant-voltage scaling – The voltages are not scaled and, in some cases, dimensions associated with voltage are not scaled.

Year	**Channel**
1980	5.00μ
1998	0.25μ
2000	0.18μ
2002	0.13μ
2003	0.09μ

MOS Scaling

Quantity	Sensitivity	Constant Field	Constant Voltage
Scaling Parameters			
Length	L	$1/S$	$1/S$
Width	W	$1/S$	$1/S$
Gate Oxide Thickness	t_{ox}	$1/S$	$1/S$
Supply Voltage	V_{dd}	$1/S$	1
Threshold Voltage	V_{T0}	$1/S$	1
Doping Density	N_A, N_D	S	S^2
Device Characteristics			
Area (A)	WL	$1/S^2$	$1/S^2$
β	W/Lt_{ox}	S	S
D-S Current (I_{DS})	$\beta(V_{dd} - v_T)^2$	$1/S$	S
Gate Capacitance (C_g)	WL/t_{ox}	$1/S$	$1/S$
Transistor On-Resistance (R_{tr})	V_{dd}/I_{DS}	1	S
Intrinsic Gate Delay (τ)	$R_{tr}C_g$	$1/S$	$1/S$
Clock Frequency	f	f	f
Power Dissipation (P)	$I_{DS}V_{dd}$	$1/S^2$	S
Power Dissipation Density (P/A)	P/A	1	S^3

MOS Scaling
How is Doping Density Scaled?

- Parameters affected by substrate doping density:
  - » The depths of the source and drain depletion regions
  - » Possibly the depth of the channel depletion region
  - » possibly $\boldsymbol{V_T}$.
- Channel depletion region and $\boldsymbol{V_T}$ are not good candidates for deriving general relationships since channel implants are used to tune $\boldsymbol{V_T}$.
- Thus, we focus our argument depletion region depth of the source and drain which is given by:

$$x_d = \sqrt{\frac{2\varepsilon_{S_i}}{q}\frac{N_A+N_D}{N_A\bullet N_D}\left|\phi_0 - V\right|}$$

where $\phi_0 = \frac{kT}{q}\ln\left(\frac{N_A+N_D}{n_i^2}\right)$

- Assuming $\boldsymbol{N_A}$ small compared to $\boldsymbol{N_D}$ and $\boldsymbol{\phi_0}$ small compared to $\boldsymbol{V}$ (the reverse bias voltage which ranges from 0 to $-\boldsymbol{V_{DD}}$),

$$X_d \propto \sqrt{\frac{1}{N_A}\left|V\right|}$$

MOS Scaling
How is Doping Density Scaled? (Continued)

Assuming constant voltage scaling with scaling factor ***S***, ***x_d*** scales as follows:

$$\frac{1}{S}x_d \propto \frac{1}{S}\sqrt{\frac{1}{N_A}|V|} = \sqrt{\frac{1}{S^2 N_A}|V|}$$

- Thus, for constant voltage scaling, ***N_A => S^2N_A***, and ***N_D => S^2N_D***
- On the other hand, for full scaling, ***V_A => V/S*** giving:

$$\frac{1}{S}x_d \propto \frac{1}{S}\sqrt{\frac{1}{N_A}|V|} = \sqrt{\frac{1}{S^2 N_A}|V|} = \sqrt{\frac{1}{S\,N_A}\left|\frac{V}{S}\right|}$$

- Thus, for full scaling, ***N_A => SN_A***, and ***N_D => SN_D***

MOS Scaling
Which Type of Scaling Behaves Best?

Constant Voltage Scaling

- Practical, since the power supply and signal voltage are unchanged
- But, I_{DS} => SI_{DS}. W => W/S and x_j => x_j/S for the source and drain (same for metal width and thickness).

  So J_D => S^3J_D, increasing current density by S^3. Causes *metal migration and self-heating in interconnects.*
- Since V_{dd} => V_{dd} and I_{DS} => SI_{DS}, the power P => SP. The area A => A/S^2.

  The power density per unit area increases by factor S^3. Cause *localized heating* and *heat dissipation problems*.
- Electric field increases by factor S. Can cause failures such as *oxide breakdown*, *punch-through*, and *hot electron charging of the oxide*.
- With all of these problems, why not use **full scaling** reducing voltages as well? Done – Over last several years, departure from **5.0 V: 3.3, 2.5, 1.5 V**

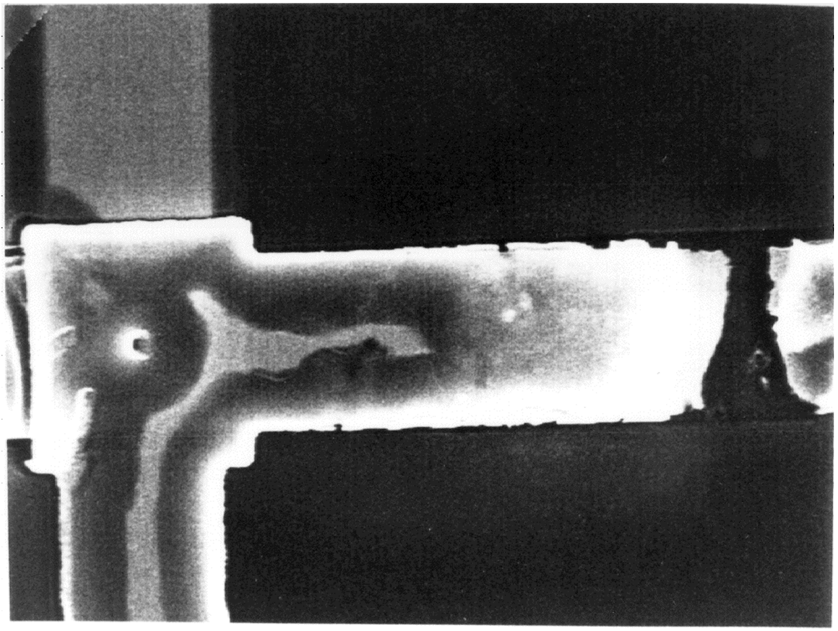
Does Scaling Really Work?

- Not totally as dimension become small, giving us our next topic.

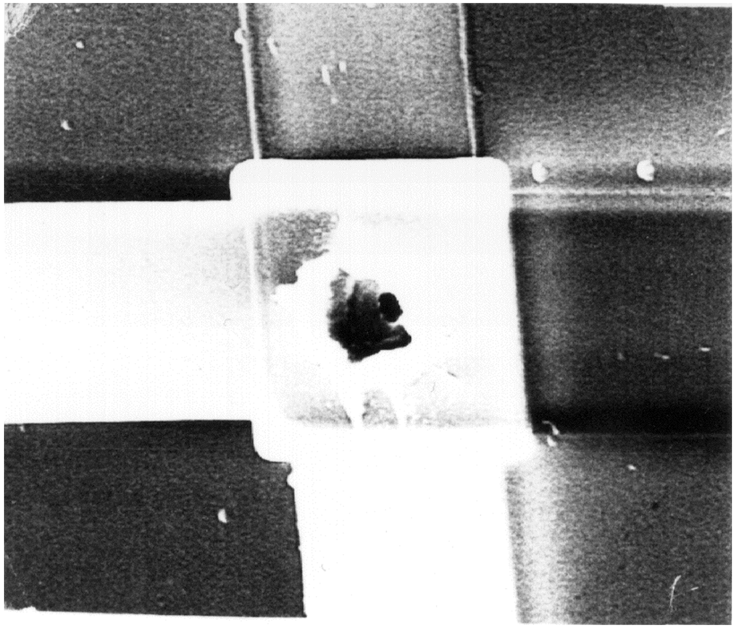
Electromigration Effect

- Wire can tolerate only a certain amount of current density.
- Direct current for a long time causes ion movement breaking the wire over time.
- Contacts are move vulnerable to electromigration as the current tends to run through the perimeter.
- Perimeter of the contact, not the area important.
- Use of copper (heavier ions) have helped in tolerating electromigration.

Observed Electromigration Failure



A wire broken off due to electromigration



A contact (via) broken up due to electromigration

These figures are derived from *Digital integrated circuit – a design perspective*, J. Rabaey Prentice Hall

Self Heating

- Electromigration depends on the directionality of the current.
- Self-heating is just proportional to the amount of current the wire carries.
- Current flow causes the wire to get heated up and can result in providing enough energy to carriers to make them hot carriers.
- Self-heating effect can be reduced by sizing the wire (same as electromigration).

Small Geometry Effects
Short-Channel Effects

- **General – Due to small dimensions**
  - Effects always present, but masked for larger channel lengths
  - Effects absent until a channel dimension becomes small
  - Many, but not all of these effects represented in SPICE, so a number of the derivations or results influence SPICE device models.
- **Short – Channel Effects**
  - What is a short-channel device? The effective channel length (L_{eff}) is the same order of magnitude as the source and drain junction depth (x_j).

Velocity Saturation and Surface Mobility Degradation

- Drift velocity v_d for channel electrons is proportional to electric field along channel for electric fields along the channel of 10^5 V/cm

Small Geometry Effects
Short-Channel Effects (Cont.)

(as occur as L becomes small with V_{DD} fixed), v_d saturates and becomes a constant $v_{d(SAT)}$ = 10^7 cm/s. This reduces $I_{D(SAT)}$ which no longer depends quadratically on V_{GS}.

$$v_{d(SAT)} = \mu_n E_{SAT} = \mu_n V_{DSAT}/L \Rightarrow V_{DSAT} = L\, v_{d(SAT)}/\mu_n$$

- Hence,

$$\begin{aligned} I_{D(SAT)} &= I_D(V_{DS}=V_{DSAT}) \\ &= \mu_n C_{ox}(W/L)[(V_{GS}-V_T)\, V_{DSAT}-V_{DSAT}^2/2] \\ &= v_{d(SAT)}\, C_{ox} \mathrm{W}(V_{GS}-V_T-V_{DSAT}/2) \end{aligned}$$

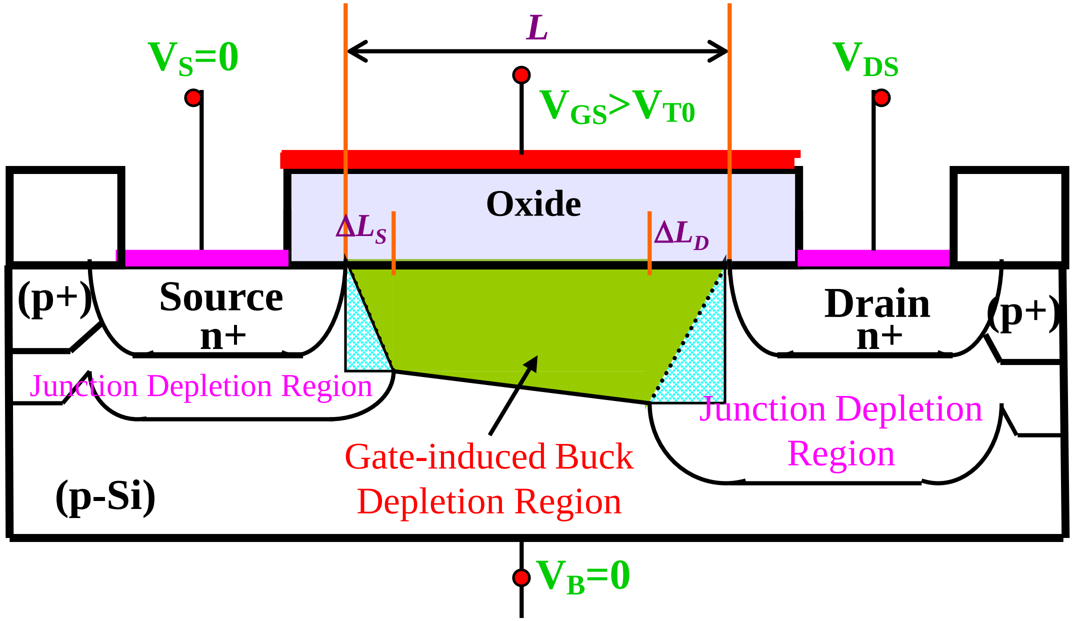
Therefore, the drain current is *linearly dependent on* V_{GS} when fully velocity saturated.

- The vertical field (E_x) effects cause μ_n to decline represented by effective surface mobility $\mu_{n(eff)}$.
- Empirical formulas for $\mu_{n(eff)}$ on p.120 of Kang and Leblebici use parameters Θ and η.

Small Geometry Effects
Short-Channel Effects (Cont.)

Channel Depletion Region Charge Reduction

- Often viewed as the short channel effect
- At the source and drain ends of the channel, channel depletion region charge is actually depletion charge for the source and drain
- For ***L*** large, attributing this charge to the channel results in small errors
- But for short-channel devices, the proportion of the depletion charge tied to the source and drain becomes large



Small Geometry Effects
Channel Depletion Region Charge Reduction (Cont.)

- The reduction in charge is represented by the change of the channel depletion region cross-section from a rectangle of length L and depth x_{dm} to a trapezoid with lengths L and $L-\Delta L_S-\Delta L_D$ and depth x_{dm}. This trapezoid is equivalent to a rectangle with length:

$$L\left(1-\frac{\Delta L_S+\Delta L_D}{2L}\right)$$

- Thus, the channel charge per unit area is reduced by the factor:

$$1-\frac{\Delta L_S+\Delta L_D}{2L}$$

- Next, need ΔL_S and ΔL_D in terms of the source and drain junction depths and depletion region junction depth using more geometric arguments. Once this is done, the resulting reduction in threshold voltage V_T due to the short channel effect can be written as:

Small Geometry Effects
Channel Depletion Region Charge Reduction (Cont.)

$$(x_j+x_{dD})^2=x_{dm}^2+(x_j+\Delta L_D)^2$$

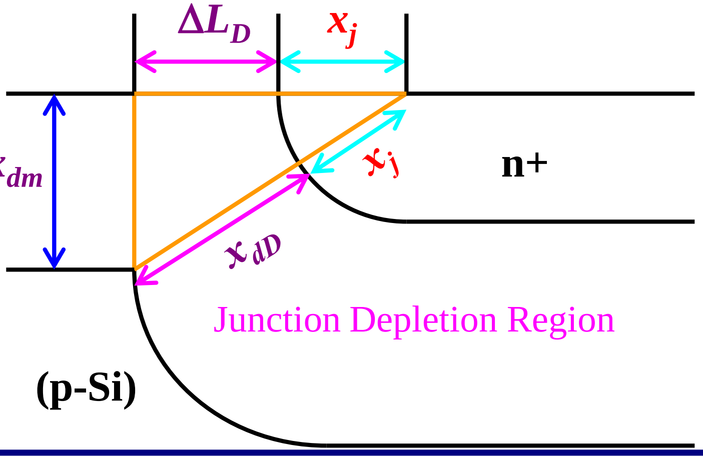
$$\Rightarrow \Delta L_D=x_j+\sqrt{x_j^2-(x_{dm}^2-x_{dD}^2)+2x_jx_{dD}} \approx x_j\left(\sqrt{1+2x_{dD}/x_j}-1\right)$$

- Similarly,

$$\Delta L_S=x_j+\sqrt{x_j^2-(x_{dm}^2-x_{dS}^2)+2x_jx_{dS}} \approx x_j\left(\sqrt{1+2x_{dS}/x_j}-1\right)$$

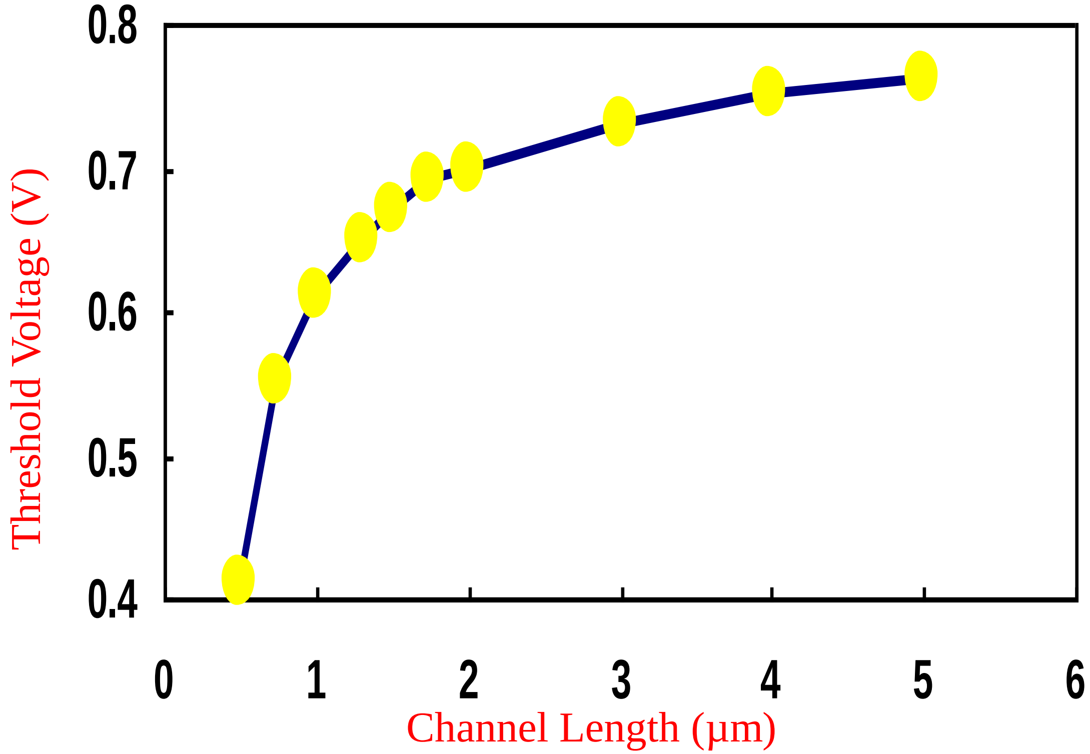
- Therefore,

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



Small Geometry Effects
Channel Depletion Region Charge Reduction (Cont.)

- For 5 $\boldsymbol{\mu}$, effect is negligible. But at 0.5 $\boldsymbol{\mu}$, $\boldsymbol{V_{T0}}$ reduced to 0.43 from 0.76 volts ($\Delta \boldsymbol{V_{T0}}$=0.33V)

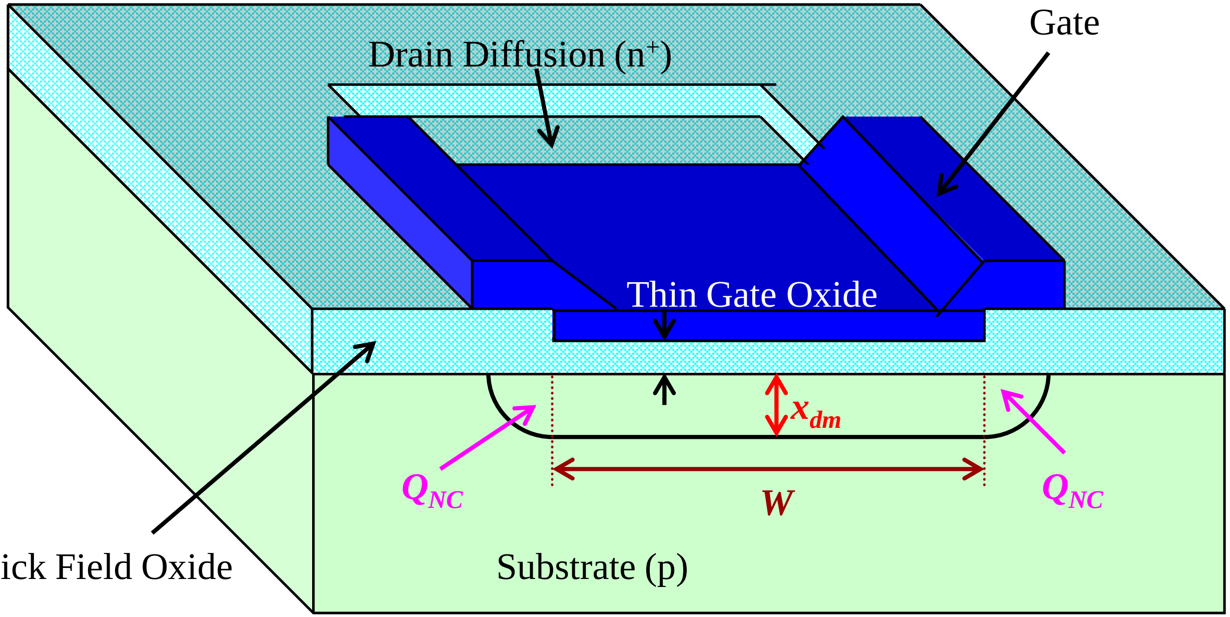


Small Geometry Effects
Narrow-Channel Effect

- ***W*** is on the same order of the maximum depletion region thickness x_{dm}.
- The channel depletion region spreads out under the polysilicon at its rises over the thick oxide. Thus, there is extra charge in the depletion region.
- The increase in V_{T0} due to this extra charge is

$$\Delta V_{T0} = \frac{1}{C_{ox}} \sqrt{2q\varepsilon_{Si} N_A |2\phi_F|} \frac{\kappa x_{dm}}{W}$$

- κ is an empirical parameter dependent upon the assumed added charge cross-section. This increase of V_{T0} may offset much of the short channel effect which is subtracted from V_{T0}.



Small Geometry Effects
Subthreshold Condition

- The potential barrier that prevents channel formation is actually controlled by both the gate voltage V_{GS} and the drain voltage V_{DS}
- V_{DS} lowers this potential, an effect known as ***DIBL (Drain-Induced Barrier Lowering).***
- If the barrier is lowered sufficiently by V_{GS} and V_{DS}, then there is channel formation for $V_{GS} < V_{T0}$.
- Subthreshold current is the result.
- Upward curvature of the I_D versus V_{GS} curve for $V_{GS} < V_T$ with $V_{DS} \neq 0$.

$$I_D(subthreshold) \cong \frac{qD_n W x_c n_0}{L_B} \cdot e^{\frac{q\phi_r}{kT}} \cdot e^{\frac{q}{kT}(A \cdot V_{GS} + B \cdot V_{DS})}$$

Small Geometry Effects
Other Effects

Punch-Through

- Merging of depletion regions of the source and drain.
- Carriers injected by the source into the depletion region are swept by the strong field to the drain.
- With the deep depletion, a large current under limited control of V_{GS} and V_{SB} results.
- *Thus, normal operation of devices in punch-through not feasible.*
- Might cause permanent damage to transistors by localized melting of material.

Thinning of t_{ox}

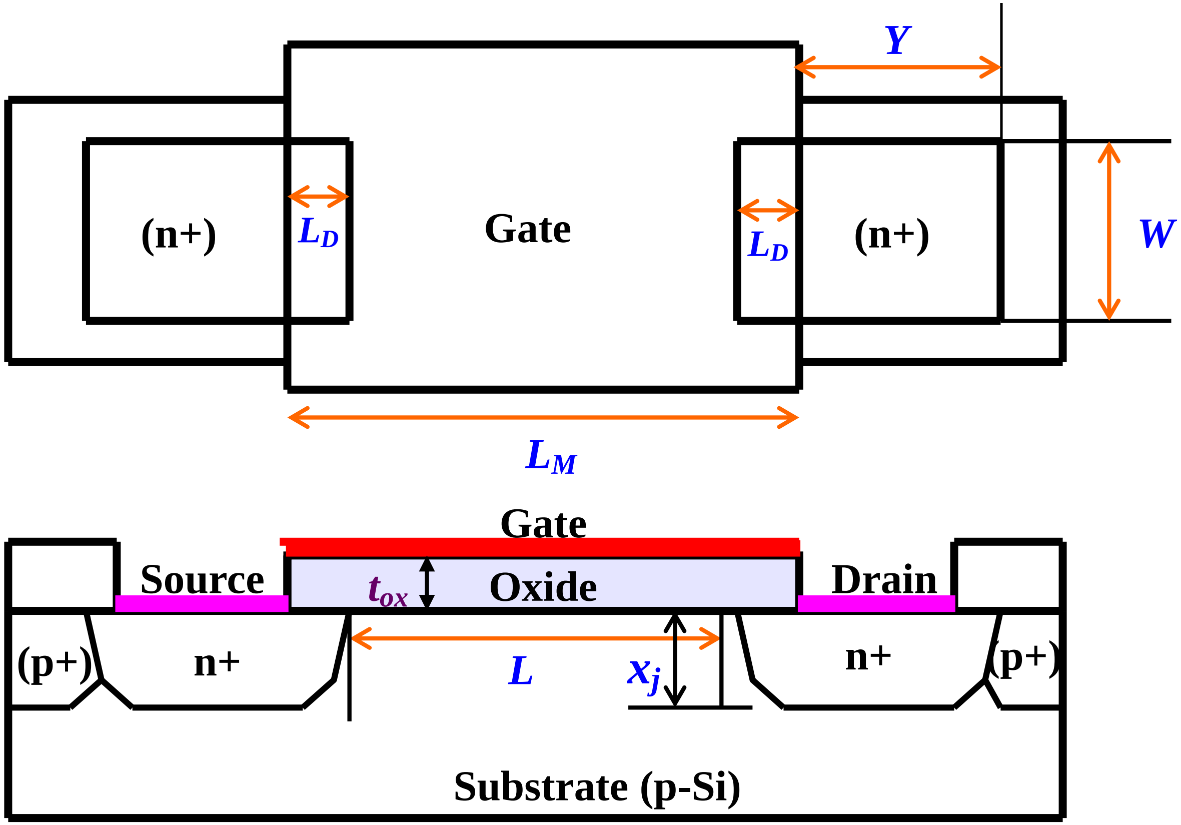
- As oxide becomes thin, localized sites of nonuniform oxide growth (***pinholes***) can occur.
- Can cause electrical shorts between the gate and substrate.
- Also, dielectric strength of the thin oxide may permit **oxide breakdown** due to application of an electric field in excess of **breakdown field**.
- May cause permanent damage due to current flow through the oxide.

Small Geometry Effects
Other Effects

Hot Electron Effects

- High electric fields in both the channel and pinch-off region for short channel lengths occur for small L.
- Particularly apparent in the pinch-off region where voltage $V_{DS} - V_{D(SAT)}$ large with $L - L_{eff}$ small causes very high fields.
- High electric fields accelerate electrons which have sufficient energy with the accompanying vertical field to be injected into the oxide and are trapped in defect sites or contribute to interface states.
- These are called **hot electrons**. See Kang and Leblebici – Fig.3.27.
- Resulting trapped charge increases V_T and otherwise affects transconductance, reducing the drain current. Since these effects are concentrated at the drain end of the channel, the effects produce asymmetry in the I-V characteristics seen in Kang and Leblebici – Fig.3.28.
- Effect further aggravated by **impact ionization**.

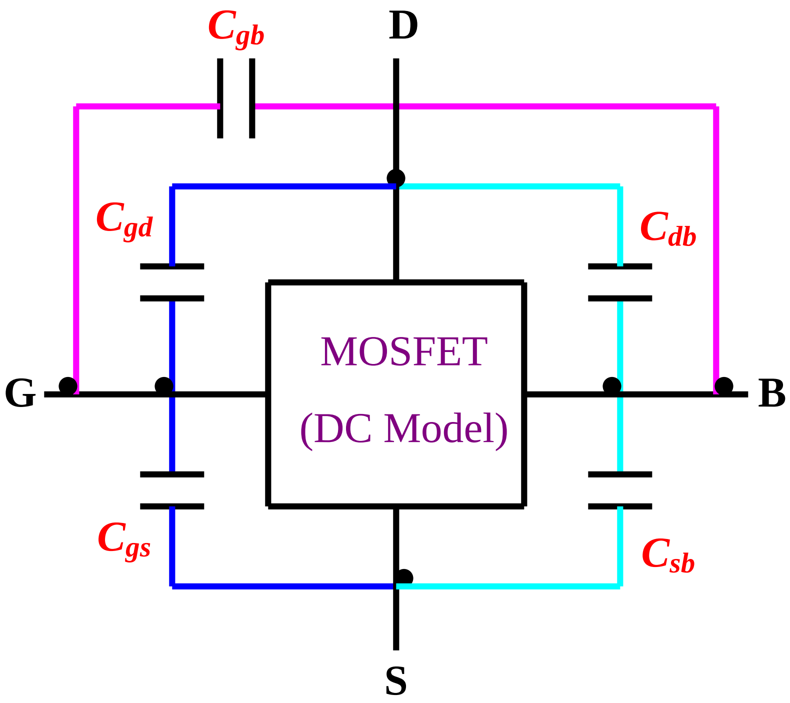
MOSFET Capacitances
Transistor Dimensions



- L_M: mask length of the gate
- L: actual channel length
- L_D: gate-drain overlap
- Y: typical diffusion length
- W: length of the source and drain diffusion region

MOSFET Capacitances
Oxide Capacitances

- Parameters studied so far apply to steady-state (DC) behavior. We need add parameters modeling transient behavior.
- MOSFET capacitances are *distributed and complex*. But, for tractable modeling, we use *lumped approximations*.
- Two categories of capacitances: 1) oxide-related and 2) junction. Inter-terminal capacitances result as follows:



MOSFET Capacitances
Overlap Capacitances

- Capacitances C_{gb}, C_{gs}, and C_{gd}
- Have the thin oxide as their dielectric

Overlap Capacitances

- Two special components of C_{gs} and C_{gd} caused by the lateral diffusion under the gate and thin oxide

$$C_{GS(overlap)} = C_{ox}WL_D$$

$$C_{GD(overlap)} = C_{ox}WL_D$$

L_D: lateral diffusion length

W : the width of channel

$C_{ox} = \varepsilon_{ox}/t_{ox}$: capacitance per unit area

- Theses overlap capacitances are *bias independent* and are added components of C_{gs} and C_{gd}.

MOSFET Capacitances
Gate-to-Channel Charge Capacitances

- Remaining oxide capacitances not fixed, but are dependent in the mode of operation of the transistor; referred to as being *bias-dependent*.
- Capacitances between the gate and source, and the gate and drain are really distributed capacitances between the gate and the channel apportioned to the source and drain.

Cutoff

- No channel formation => $C_{gs} = C_{gd} = 0$. The gate capacitance to the substrate

$$C_{gb} = C_{ox}\, W\, L$$

Linear

- The channel has formed and the capacitance is from the gate to the source and drain, **not** to the substrate. Thus $C_{gb}=0$ and

$$C_{gs} \approx C_{gd} \approx (C_{ox}\, W\, L)/2$$

MOSFET Capacitances
Gate-to-Channel Charge Capacitances (Cont.)

Saturation

- In saturation, the channel does not extend to the drain. Thus, $C_{gd}=0$ and

$$C_{gs} \approx (C_{ox}\,W\,L)*2/3$$

These capacitances as a function of V_{GS} (and V_{DS}) can be plotted as in Kang and Leblebici – Fig.3.32. Note that the capacitance seen looking into the gate is C_g:

$$C_{ox}W(\,2L/3+2L_D) \le C_g = C_{gb} + C_{gs} + C_{gd} \le C_{ox}W(L+2L_D)$$

- For manual calculations, we approximate C_g as its maximum value.

Capacitance	Cut-off	Linear	Saturation
C_{gb}(total)	$C_{ox}WL$	0	0
C_{gd}(total)	$C_{ox}WL_D$	$C_{ox}WL/2+C_{ox}WL_D$	$C_{ox}WL_D$
C_{gs}(total)	$C_{ox}WL_D$	$C_{ox}WL/2+ C_{ox}WL_D$	$2C_{ox}WL/3+ C_{ox}WL_D$

- This component of input capacitance is directly proportional to L and W and inversely proportional to t_{ox}.

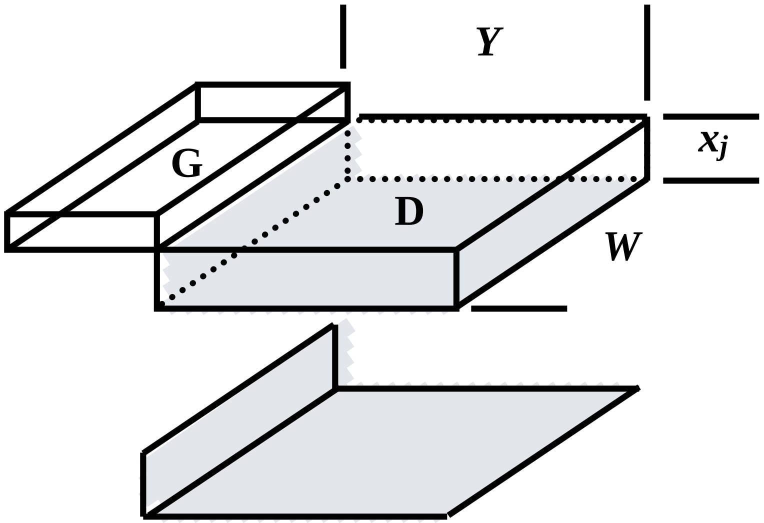
MOSFET Capacitances
Gate-to-Channel Charge Capacitances (Cont.)

Junction Capacitances

- Capacitances associated with the source and the drain
- Capacitances of the reversed biased substrate-to-source and substrate-to-drain p-n junctions.
- Lumped, but if the diffusion used as a conductor of any length, both its capacitance and resistance need to be modeled in a way that tends more toward a distributed model which is used for resistive interconnect.

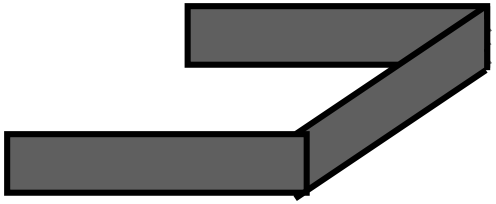
MOSFET Capacitances
Junction Capacitance Geometry

The Geometry



Junction between p substrate and n+ drain (**Bottom**)

Area: $W(Y+x_j)$ = AD



Junction between p+ channel stop and n+ drain (**Sidewalls**)

Area: $x_j(W+2Y) = x_j$ PD

MOSFET Capacitances
Junction Capacitance Geometry (Continued)

- Since the diffusion also enters into contacts at a minimum here, actual geometries will be more complex, but the fundamental principles remain.
- Why separate bottom and sides? The carrier concentration in the channel stop area is an order of magnitude higher **(~$10N_A$)** than in the substrate (N_A). This results in a higher capacitance for the sidewalls.
- The bottom and channel edge can be treated together via **AD** in the SPICE model but often channel edge either ignored or included in **PD**.
- All other areas are treated together via the length of the perimeter **PD** in the SPICE model. The capacitance in this case is per meter since dimension x_j is incorporated in the capacitance value.
- Same approach for source.

MOSFET Capacitances
Junction Capacitance/Unit Area

- Two junction capacitances per unit area for each distinct diffusion region, the bottom capacitance and the sidewall. Equations are the same, but values different.
- Thus, we use a single value ***C_j*** which is the capacitance of a p-n junction diode.
- Recall that most of the depletion region in a diode lies in the region with the lower impunity concentration, in this case, the p-type substrate.
- Finding the depletion region thickness in term of basic physical parameters and ***V*** the applied voltage (note that ***V*** is negative since the junction is reversed biased).

$$x_d = \sqrt{\frac{2\varepsilon_{Si}}{q}\frac{N_A+N_D}{N_A N_D}\left(\phi_0 - V\right)}$$

- The junction potential in this equation is

MOSFET Capacitances
Junction Capacitance/Unit Area (Continued)

- The junction potential in this equation is

$$\phi_0 = \frac{kT}{q} \ln\left(\frac{N_A N_D}{n_i^2} \right)$$

The total depletion region charge can be calculated by using ***x_d***:

$$Q_j = Aq \frac{N_A N_D}{N_A + N_D} x_d = A \sqrt{2 \varepsilon_{Si} q \frac{N_A N_D}{N_A + N_D} \left(\phi_0 - V \right)}$$

The capacitance found by differentiating ***Q_j*** with respect to ***V*** to give:

$$C_j(V) = \left| \frac{dQ_j}{dV} \right| = \frac{A C_{j0}}{\left(1 - V / \phi_0 \right)^{1/2}}$$

with

$$C_{j0} = \sqrt{ \frac{\varepsilon_{Si} q}{2} \left(\frac{N_A N_D}{N_A + N_D} \right) \frac{1}{\phi_0} }$$

MOSFET Capacitances
Junction Capacitance - Approximations

Approximation for Manual Calculations

- The voltage dependence of $C_j(V)$ makes manual calculations difficult. An *equivalent large-signal capacitance* for a voltage change from V_1 to V_2 can be defined as

$$C_{eq} = \Delta Q/\Delta V = (Q_j(V_2)-Q_j(V_1))/(V_2-V_1)$$

- The formula of this *equivalent large-signal capacitance* is derived in the book with the final version:

$$C_{eq} = AC_{j0}K_{eq}$$

where K_{eq} $(0<K_{eq}<1)$ is the *voltage equivalence factor*,

$$K_{eq} = \frac{2\sqrt{\phi_0}}{V_1 - V_2}\left(\sqrt{\phi_0 - V_2} - \sqrt{\phi_0 - V_1}\right)$$

Summary

- Full scaling (constant field scaling) better than constant voltage scaling if the power supply value can be changed.
- Scaling is subject to small geometry effects that create new limitations and requires new modeling approaches.
- The short-channel effect, narrow-channel effect, mobility degradation, and subthreshold conduction all bring new complications to the modeling of the MOSFET.
- Geometric and capacitance relationships developed permit us to calculate:

  the two overlap capacitances due to lateral diffusion,

  the three transistor-mode dependent oxide capacitances

  the voltage-dependent bottom and sidewall junction capacitances for the sources and drain, and

  fixed capacitance source and drain capacitances values for a voltage transition in manual calculations.