Chapter 7

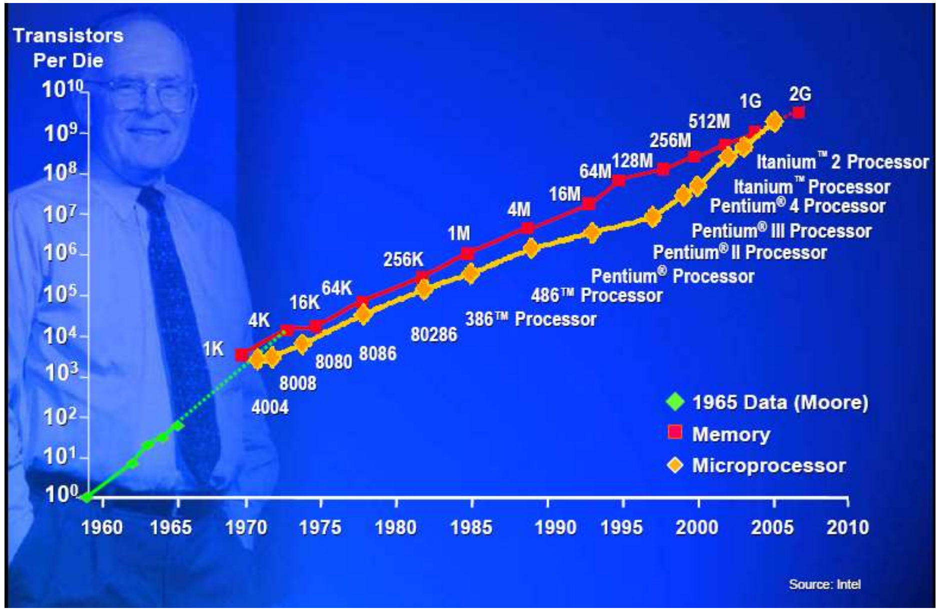
MOSFETs in ICs-Scaling, Leakage, and Other Topics

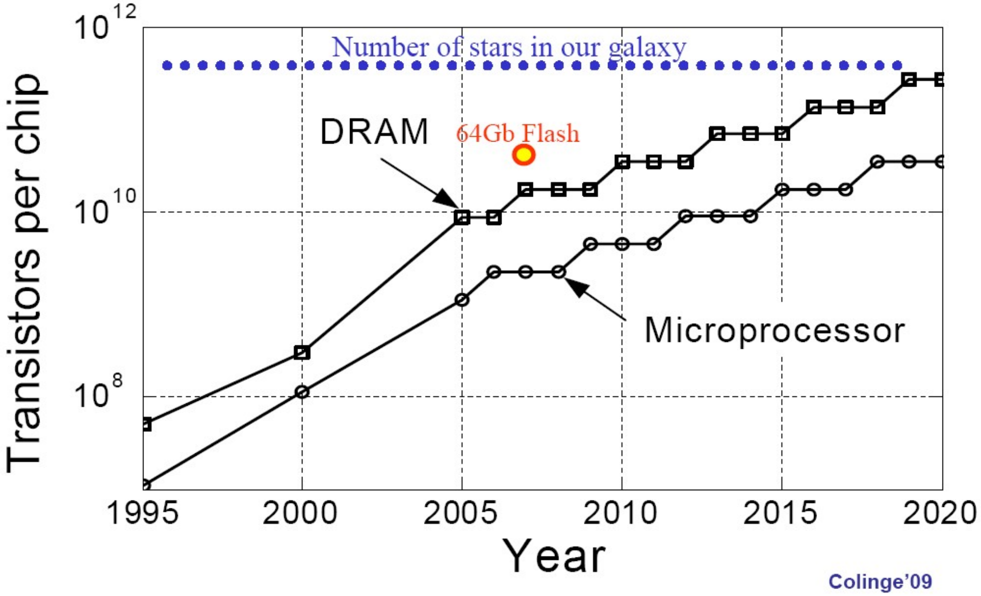
OBJECTIVES

1. Understanding the off-state current or the leakage current of the MOSFETs:
   subthreshold leakage and its impact on device size reduction,
   trade-off between I_{on} and I_{off}, and effects on the circuit design.
2. Understanding of the opportunities for future MOSFET scaling including mobility enhancement, high-*k* dielectric and metal gate, SOI, multigate MOSFET, metal source/drain, etc.
3. Introducing device simulation and MOSFET compact model for circuit simulation

Technology Scaling: - for Cost, Speed, and Power Consumption

Moore's law: The number of devices on a chip doubles every 18 to 24 months or so.





Colinge'09

Technology node or **Technology generation** in every two or three years

minimum metal line width: 0.18 μm, 0.13 μm, 90 nm, 64 nm, 45 nm…
(Poly-Si gate lengths may be even smaller)

Scaling At each new node, all features in the circuit layout are reduced in size to 70 % of the previous node. This practice of periodic size reduction is called **scaling**.

70 % of previous line width means ~ 50 % reduction in area, i.e.,0.7 × 0.7 = 0.49

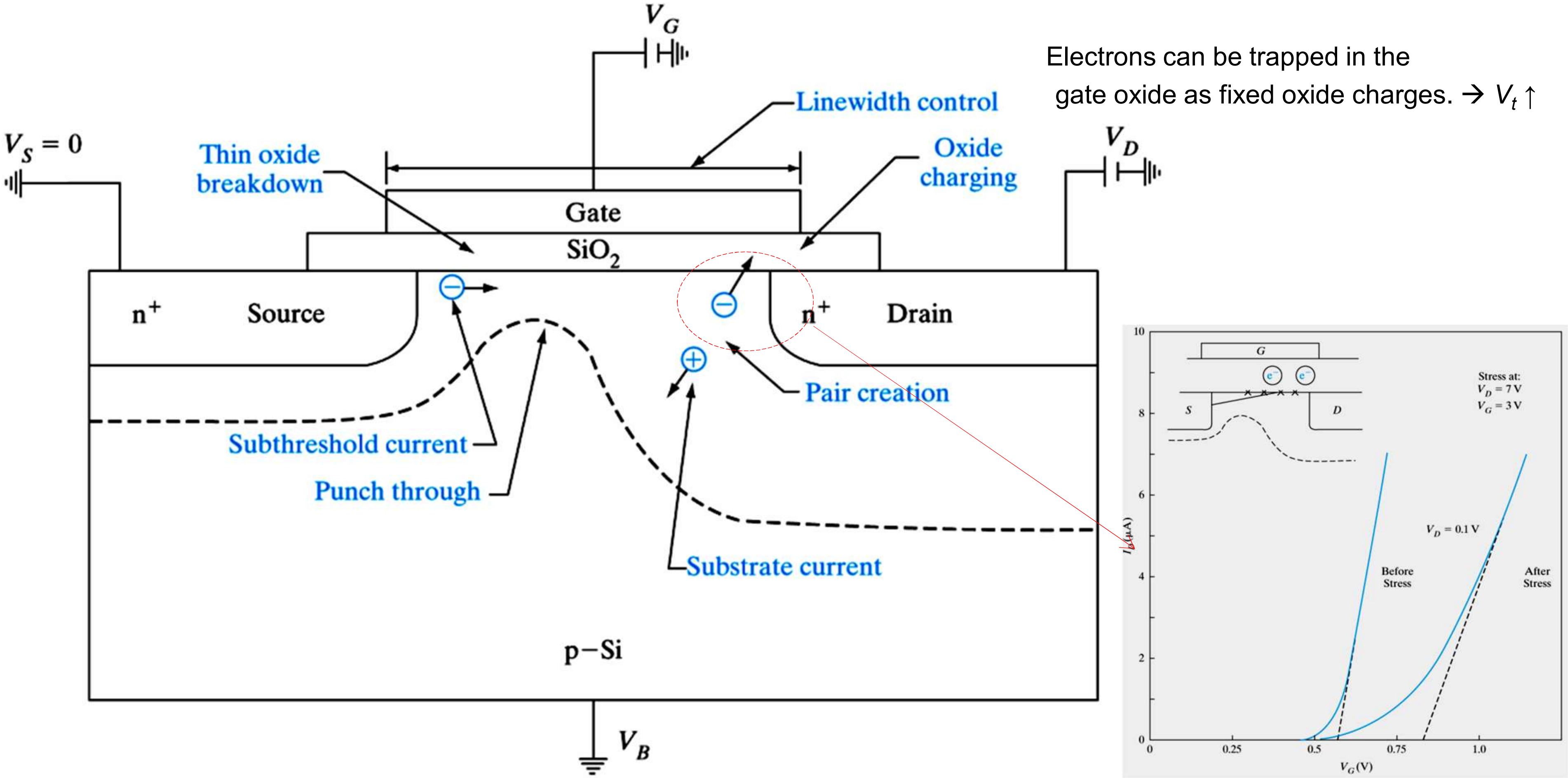
↳ drives down the cost of ICs. (1/100 million since 1965) & Power & speed

MOSFET Scaling and Hot Electron Effects

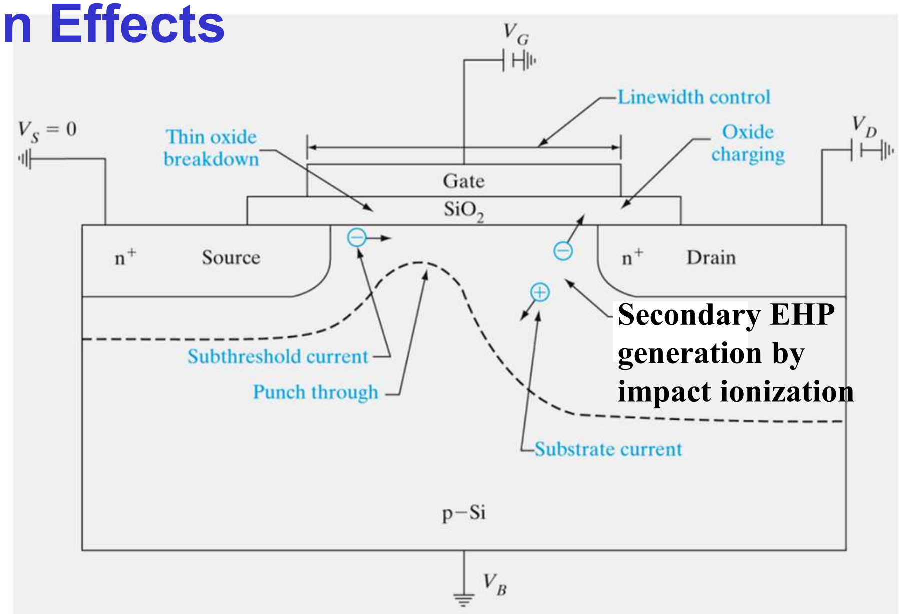
When an electron travels along the channel, it gains kinetic energy at the expense of electrostatic potential energy in the pinch-off region, and becomes a **hot electron.**

Some hot electrons can go through the gate oxide and be collected as gate current, reducing the input impedance.

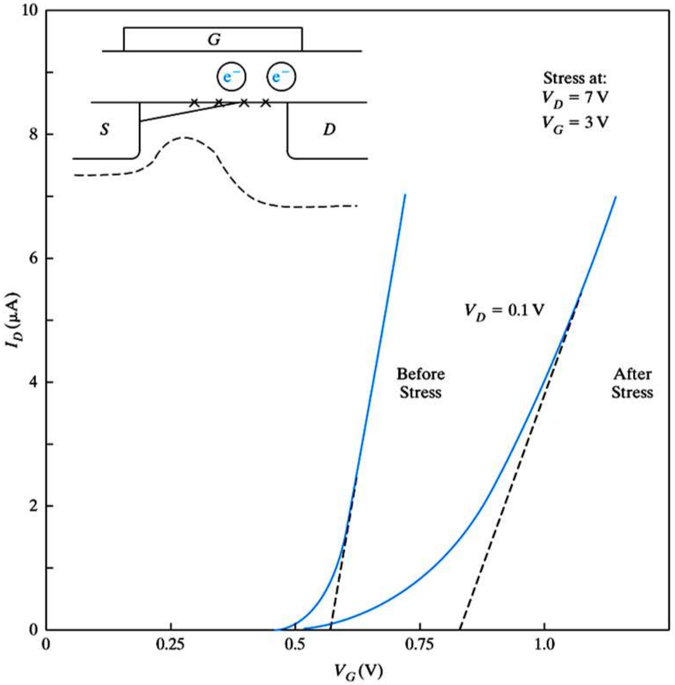
Electrons can be trapped in the gate oxide as fixed oxide charges. → V_t ↑

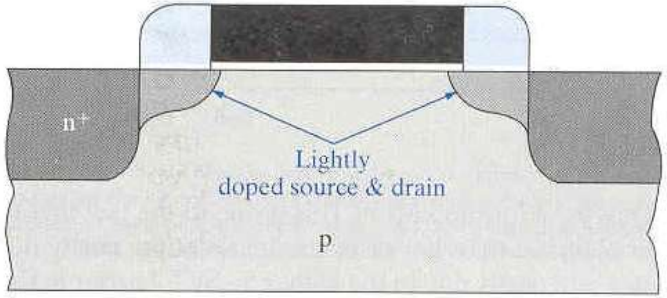


• MOSFET Scaling and Hot Electron Effects



- ✓ The energetic hot carriers can rupture Si-H bonds that exists at the Si-SiO_2 interface, creating interface states that degrade MOSFET parameters with stress (device aging).
- ✓ A method to reduce hot carrier generation
  - → LDD (Lightly Doped Drain) in n-channel MOSFET
  - → Peak field reduction





Scaling rules:

1) The horizontal (*L* , *W*) and vertical dimensions such as the gate oxide thickness are scaled by the same **scaling factor, *K*.**

2) The power supply voltage, V_{dd}, is also scaled to keep the internal electric fields more or less constant (The reductions are chosen such that the transistor current density (I_{on}/W) increases with each new node).

⬇

lead to smaller capacitance and hence cause the circuit delays to drop.

$$\tau_d \approx \frac{CV_{dd}}{4}\left(\frac{1}{I_{onN}}+\frac{1}{I_{onP}}\right) \Rightarrow \downarrow \Leftarrow \begin{cases} C=\dfrac{\varepsilon A}{T_{ox}} \Rightarrow \dfrac{1/K^2}{1/K} \Rightarrow \dfrac{1}{K} \\ V_{dd} \Rightarrow \downarrow \end{cases}$$

Surface Dimensions (*L*,*W*)	*1/K*
Vertical Dimensions (T_{ox}, x_j)	*1/K*
Impurity Concentrations	*K*
Current, Voltages	*1/K*
Current Density	*K*
Capacitance (C_{ox}, per unit area ~ *K*)	*1/K*
Transconductance	*1*
Circuit Delay Time	*1/K*
Power Consumption	$1/K^2$
Power Density	*1*
Power-Delay Product	$1/K^3$

IC speed has increased roughly 30 % at each new technology node.

Scaling is also very effective in reducing power consumption due to reduction in *C* and V_{dd}.

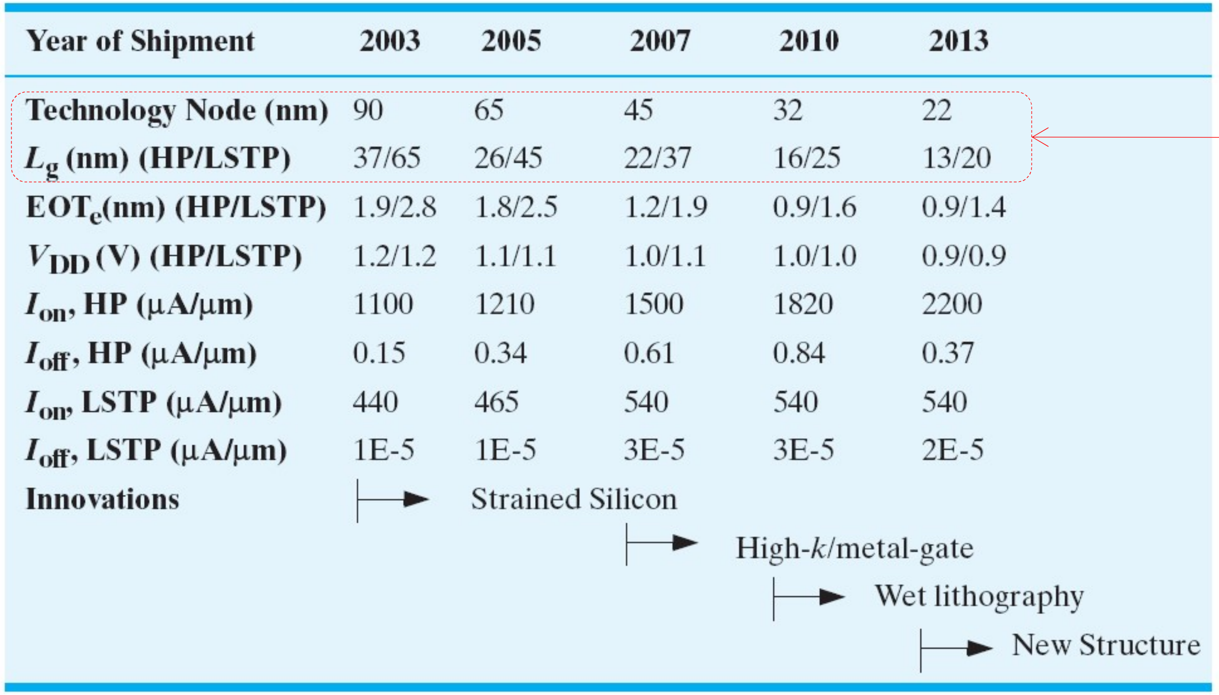
$$P_{dynamic} = V_{dd} \times (average\,current) = kCV_{dd}^2 f \Rightarrow \downarrow \Leftarrow \begin{cases} C \Rightarrow \dfrac{1}{K} \\ V_{dd} \Rightarrow \downarrow \\ f \Rightarrow K \end{cases}$$

In summary, scaling improves **cost**, **speed**, and **power consumption** per function with every new technology node.

Innovations Enables Scaling

International Technology Roadmap for Semiconductors (ITRS)

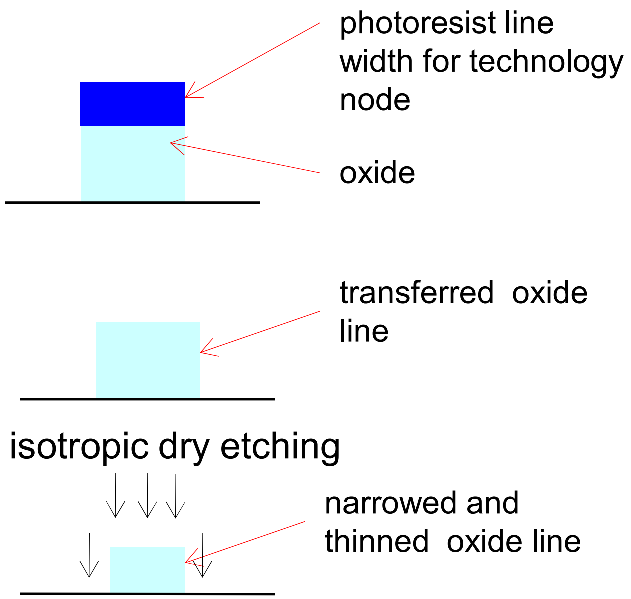
TABLE 7–1 • Scaling from 90 nm to 22 nm and innovations that enable the scaling.



Year of Shipment	2003	2005	2007	2010	2013
Technology Node (nm)	90	65	45	32	22
L_g (nm) (HP/LSTP)	37/65	26/45	22/37	16/25	13/20
EOT_e(nm) (HP/LSTP)	1.9/2.8	1.8/2.5	1.2/1.9	0.9/1.6	0.9/1.4
V_{DD} (V) (HP/LSTP)	1.2/1.2	1.1/1.1	1.0/1.1	1.0/1.0	0.9/0.9
I_{on}, HP (μA/μm)	1100	1210	1500	1820	2200
I_{off}, HP (μA/μm)	0.15	0.34	0.61	0.84	0.37
I_{on}, LSTP (μA/μm)	440	465	540	540	540
I_{off}, LSTP (μA/μm)	1E-5	1E-5	3E-5	3E-5	2E-5
Innovations	→ Strained Silicon		→ High-*k*/metal-gate	→ Wet lithography	→ New Structure

HP: High-Performance technology. LSTP: Low Standby Power technology for portable applications. EOT_e: Equivalent electrical Oxide Thickness, i.e., equivalent T_{oxe}. I_{on}: NFET I_{on}.

The physical gate length, L_g, is actually smaller than the technology node.



photoresist line width for technology node

oxide

transferred oxide line

isotropic dry etching

narrowed and thinned oxide line

Using the narrowed oxide lines as the new etch mask, they produce the gate patterns by etching.

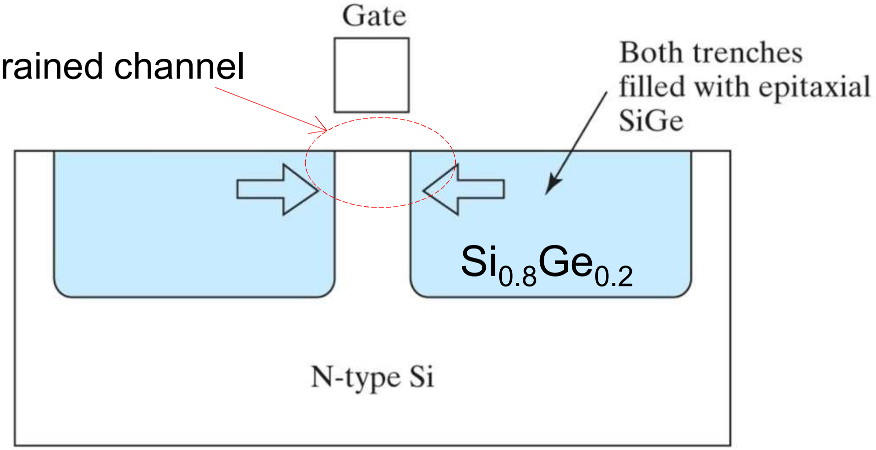
Strained Silicon and Other Innovations

Strained silicon technology (90 nm node): increases I_{on}.
High-k/metal gate (45 nm node): reduces EOTe (electrical equivalent oxide thickness)
Wet lithography (32 nm node): improves fine pattern
New structures (22nm node): reverse the trend of increasing I_{off}.

For example,

The strain changes the lattice constant of the silicon crystal and therefore the *E-k* relationship, which in turn determines the effective mass and the mobility.

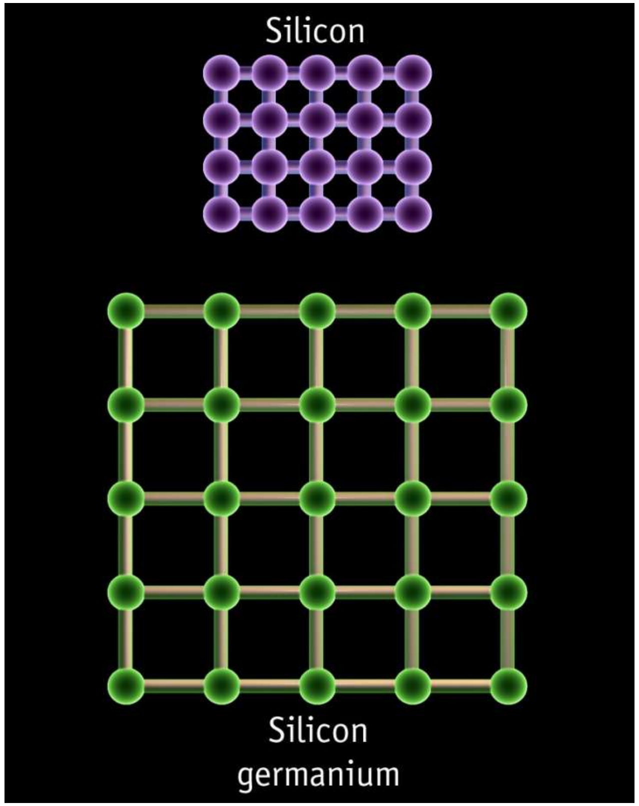
$$\frac{1}{m^*} = \frac{d^2E / dk^2}{\hbar^2}, \quad \mu = \frac{q\tau_{mn}}{m^*}$$

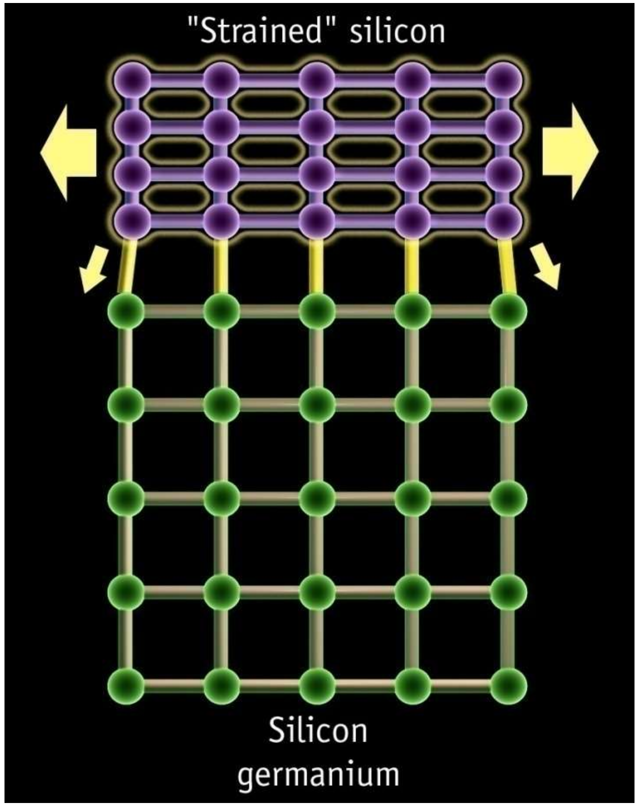

compressively strained channel
Gate
Both trenches filled with epitaxial SiGe
$Si_{0.8}Ge_{0.2}$
N-type Si

Hole mobility can be raised with a compressive mechanical strain illustrated with the arrows pushing on the channel region.

It is also attractive to incorporate a thin film of Ge material in the channel itself because Ge has higher carrier mobility.

Promise a new Dimension for Strained Silicon





Subthreshold Current-“ Off” Is Not Totally “Off”

Circuit speed improves with increasing I_{on}; therefore, it would be desirable to use a small V_t. Can we set V_t at arbitrarily small value, say 10 mV? The answer is no.

At $V_{gs} < V_t$, an N-channel MOSFET is in the off state.

$$I_{dsat} = \frac{W}{2mL} C_{oxe} \mu_{ns} (V_{gs} - V_t)^2 \to 0$$

However, a leakage current can still flow between the drain and the source.

The MOSFET current observed at $V_{gs} < V_t$ is called the **subthreshold current**.

$$I_{ds} \propto \frac{W}{L}\left(1 - e^{-qV_{ds}/kT}\right)\left(e^{q\left(V_{gs}-V_t\right)/\eta kT}\right) \propto \exp(V_{gs}), where\, \eta = \left[1 + \frac{C_{dep}}{C_{oxe}}\right]$$

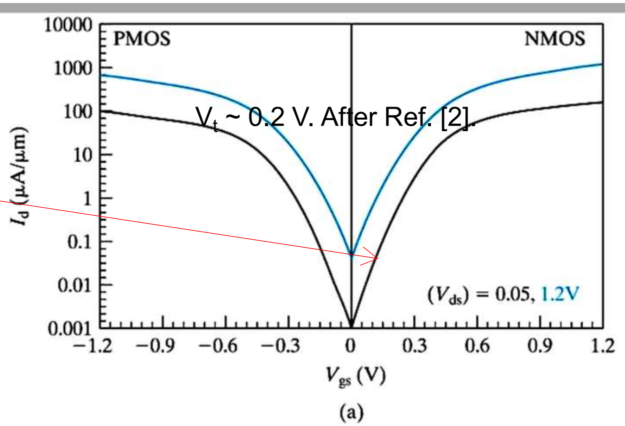
V_{ds} has little influence once V_{ds} exceeds a few *kT/q*.

This is the main contributor to the MOSFET **off-state current**, ***I_{off}***, which is the I_{ds} measured at V_{gs} = 0 and V_{ds} = V_{dd}.

It is important to keep I_{off} very small in order to minimize the static power that a circuit consumes when it is in the standby mode.

$$I_{ds} \propto \exp(V_{gs}) \Rightarrow \log I_{ds} \propto V_{gs}$$

straight line in *semi-log* I_{ds} vs. V_{gs}.



V_t ~ 0.2 V. After Ref. [2].

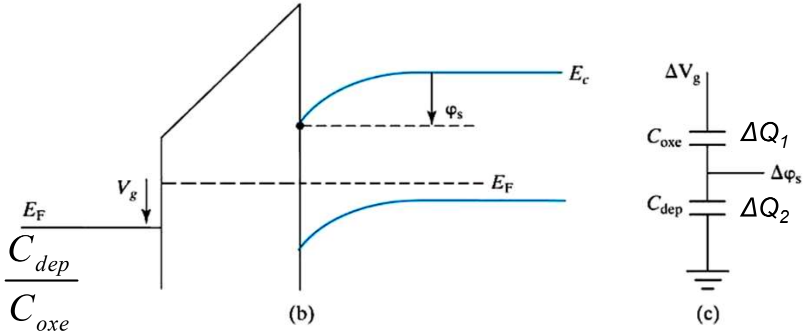
When V_{gs} is increased, E_c at the surface is pulled closer to E_F, causing n_s and I_{ds} to rise. From the equivalent circuit,

$$\Delta Q_1 = C_{oxe}(\Delta V_{gs} - \Delta\varphi_s),\ \ \Delta Q_2 = C_{dep}(\Delta\varphi_s)$$

$$\Delta Q_1 - \Delta Q_2 = 0 \Rightarrow C_{oxe}(\Delta V_{gs} - \Delta\varphi_s) - C_{dep}(\Delta\varphi_s) = 0$$

$$\Rightarrow C_{oxe}\left(1 - \frac{\Delta\varphi_s}{\Delta V_{gs}}\right) - C_{dep}\left(\frac{\Delta\varphi_s}{\Delta V_{gs}}\right) = 0$$

$$\Rightarrow \frac{\Delta\varphi_s}{\Delta V_{gs}} = \frac{C_{oxe}}{C_{oxe} + C_{dep}} \Rightarrow \frac{d\varphi_s}{dV_{gs}} = \frac{C_{oxe}}{C_{oxe} + C_{dep}} \equiv \frac{1}{\eta},\quad where\ \eta = 1 + \frac{C_{dep}}{C_{oxe}}$$



equivalent capacitance network.

$$\therefore \varphi_s = constant + \frac{V_{gs}}{\eta}$$

I_{ds} proportional to n_s, therefore

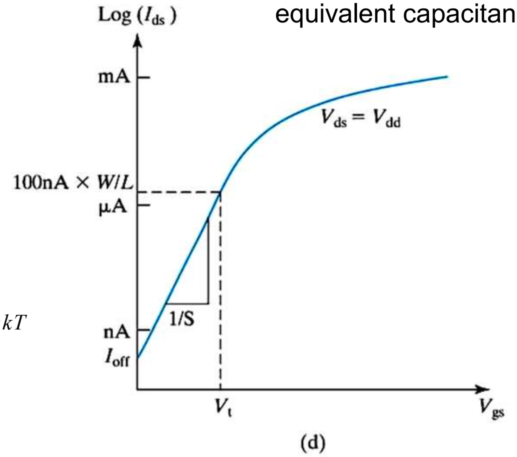
$$I_{ds} \propto n_s \propto e^{q\varphi_s/kT} \propto e^{q(constant + V_{gs}/\eta)/kT} = constant \cdot e^{qV_{gs}/\eta kT}$$

A practical and common definition of V_t is the V_{gs} at which I_{ds} = 100 nA × *W/L*.

$$I_{ds}(nA) = 100 \cdot \frac{W}{L}(nA) = constant \cdot e^{q\,V_t/\eta kT}$$

$$\therefore constant = 100 \cdot \frac{W}{L}(nA) \cdot e^{-q\,V_t/\eta kT}$$

$$\boxed{I_{ds}(nA) = 100 \cdot \frac{W}{L} \cdot e^{q(V_{gs} - V_t)/\eta kT}}$$



Subthreshold I-V with V_t and I_{off}. Swing, S, is the inverse of the slope in the subthreshold region.

Subthreshold swing, S

$$\boxed{S(mV/decade) = \frac{dV_{gs}}{d(\log I_{ds})} = \ln 10 \frac{dV_{gs}}{d(\ln I_{ds})} = 2.3\frac{kT}{q}\cdot\eta = \eta\cdot 60\,mV\cdot\frac{T}{300\,k}}$$

At room temperature, exp(qV_{gs}/kT) changes by 10 for every η × 60 mV change in V_{gs}.

$$I_{ds}(nA) = 100\cdot\frac{W}{L}\cdot e^{q(V_{gs}-V_t)/\eta kT} = 100\cdot\frac{W}{L}\cdot e^{2.3(V_{gs}-V_t)/S} = 100\cdot\frac{W}{L}\cdot 10^{(V_{gs}-V_t)/S}$$

$$I_{off}(nA) = 100\cdot\frac{W}{L}\cdot e^{q(-V_t)/\eta kT} = 100\cdot\frac{W}{L}\cdot 10^{-V_t/S}$$

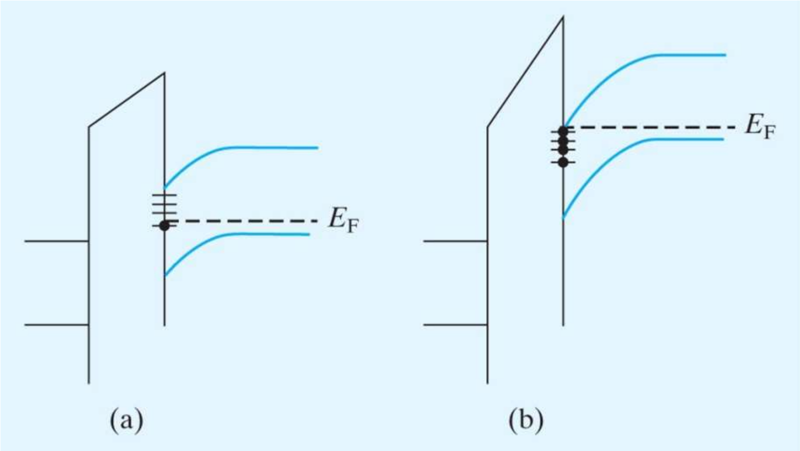
$$\eta = 1 + \frac{C_{dep}}{C_{oxe}}$$

How to minimize I_{off} for given W and L?

1) Choose a large V_t. This is not desirable because a large V_t reduces I_{on} and therefore degrades the circuit speed.
2) Reduce the subthreshold swing, *S*, which can be reduced by reducing η. That can be done by increasing C_{oxe}, i.e., using a thinner T_{ox}, and by decreasing C_{dep}, i.e., increasing W_{dep}.
3) Operate the transistor at significantly lower than the room temperature. This is rarely used because cooling add a considerable cost.

The effect of interface states on the subthreshold swing

The subthreshold swing is degraded when interface states are present.

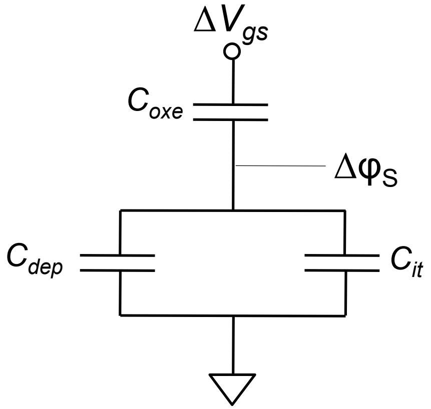


(a) If φ_S is small, most of the interface states are empty because they are above E_F.

(b) If φ_S is large at another V_{gs}, most of the interface states are filled with electrons.

The interface traps change from being empty to being occupied by electrons. This change of charge in response to change of voltage (φ_S) has the effect of a capacitor which is parallel to C_{dep}.

$$\eta = \left[1 + \frac{C_{dep} + C_{it}}{C_{oxe}}\right] \qquad S(mV/decade) = \eta \cdot 60mV \cdot \frac{T}{300k}$$

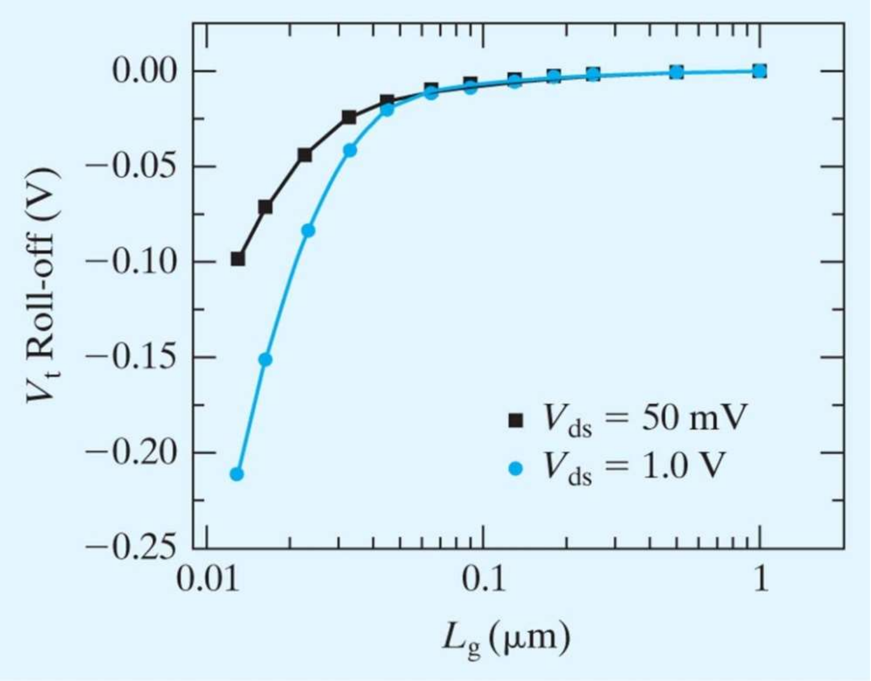


If the interface trap density, D_{it}, is high, η ↑ and S ↑

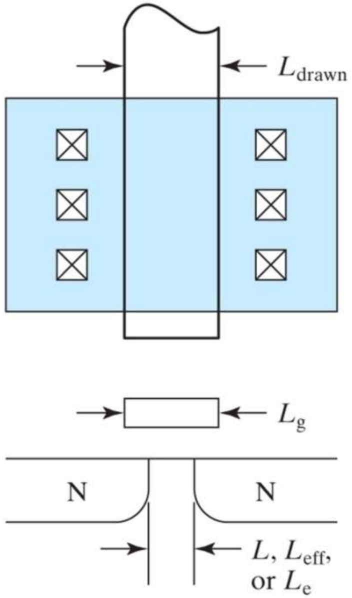
The subthreshold swing is often degraded after a MOSFET is electrically stressed and new interface states are generated.

V_t Roll-Off — Short-Channel MOSFETs Leak More

$|V_t|$ decreases at very small L_g, which is called ***V_t* roll-off**.
; ***V_t*** becomes the function of V_{DS} at very small L_g (**$|V_{DS}|$** ↑ → **$|V_t|$** ↓ → **$|I_{off}|$** ↑)



Gate Length (L_g) vs. Electrical Channel length (L)



L_{drawn} (drawn gate length in the circuit layout)

≈ ***L_g*** (patterned physical gate length)

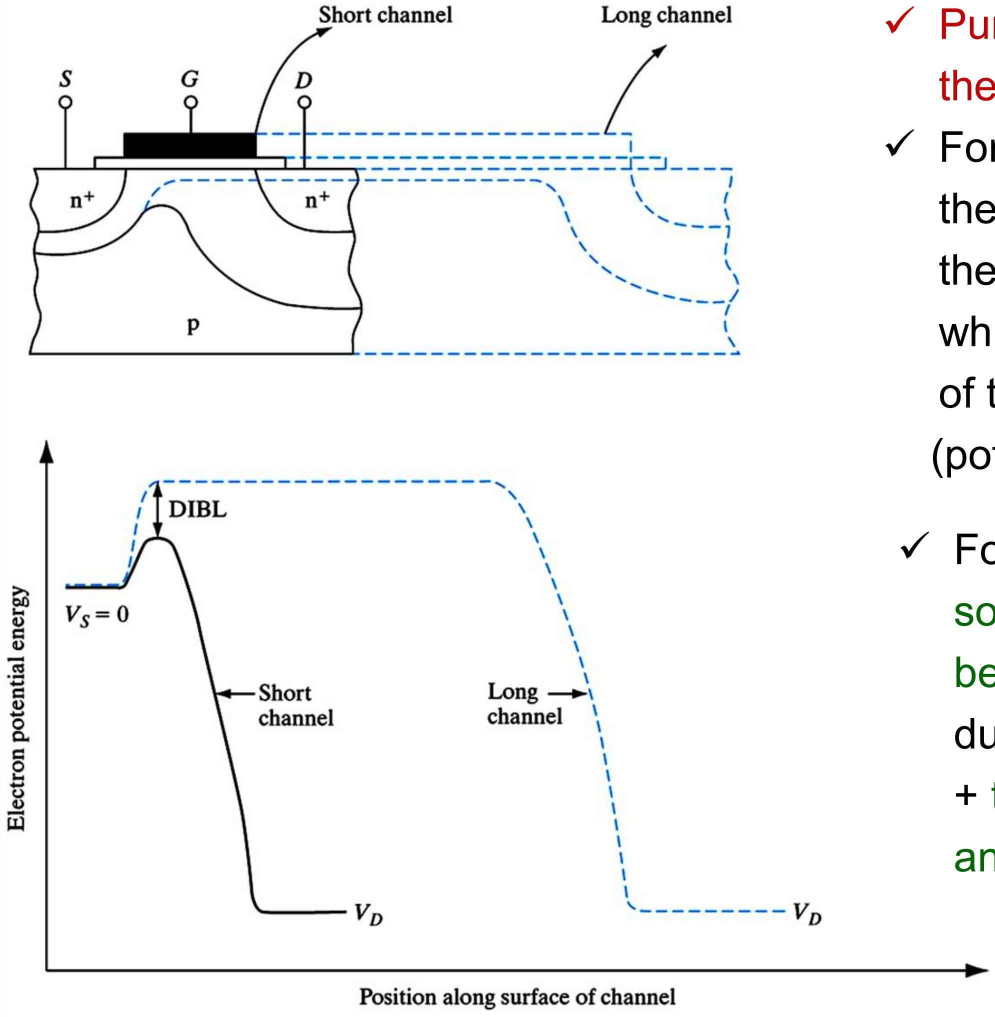
L (channel length),
L_{eff} (effective channel length)
L_e (electrical channel length)
≠ ***L_{drawn}***

When V_t drops too much, I_{off} becomes too large and that channel length is not acceptable.

Device development engineer must design the device such that the V_t roll-off does not prevent the use of the targeted minimum L_g.

Why does V_t decrease with decreasing L?

If short channel length MOSFETs are not scaled properly, there can be unintended electrostatic interactions between the source and the drain. The drain can lower the source-channel barrier and reduce V_t, which is called **drain induced barrier lowering** or **DIBL**.

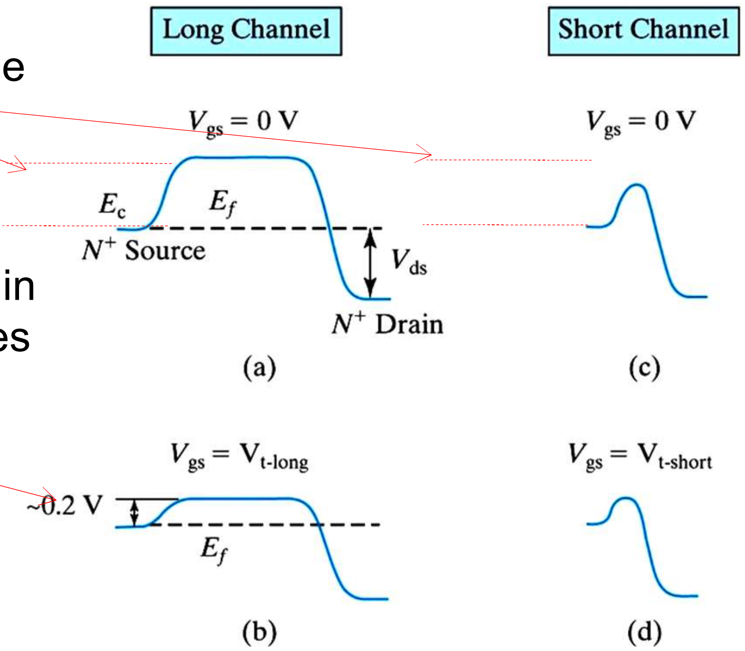


- ✓ Punch-through leakage, breakdown between the source and the drain, loss of gate control
- ✓ For a long channel MOSFET,
  the drain bias does not affect
  the source-to channel potential barrier,
  which corresponds to the built-in potential
  of the source-channel p-n junction
  (potential barrier is controlled by gate bias)
- ✓ For a short channel MOSFET,
  source-junction potential barrier is lowered
  below the built-in potential
  due to the drain voltage
  + the drain depletion region can expand
  and merge with the source depletion region

1) First Approach

If V_{ds} is not zero, E_c in the short channel is pulled lower than that in the long channel and therefore is closer to the E_c in the source.

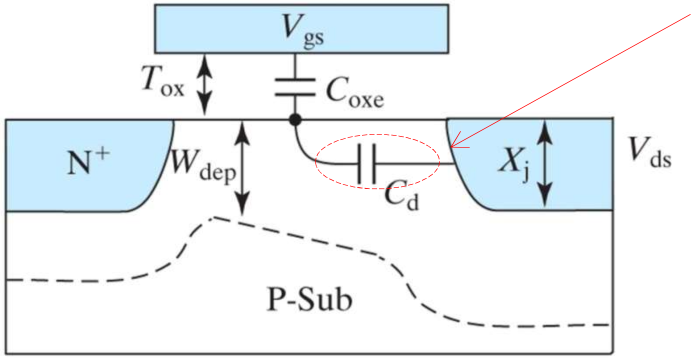
When the channel E_c is only ~ 0.2 eV higher than the E_c in the source (which is also ~ E_{Fn}), n_s in the channel reaches ~ 10^{17} cm^{-3} and inversion threshold condition is reached.



As a result, a smaller V_{gs} is needed in the short channel device than in the long channel device to pull the barrier down to 0.2 eV.

In other word, V_t is lower in the short channel device than the long channel device. This explains the V_t roll-off.

2) Second Approach



Schematic two-capacitor network in MOSFET.

C_d models the electrostatic coupling between the channel and the drain.

(capacitive coupling between the drain and the channel barrier point)

As the channel length is reduced,
drain to “channel” distance is reduced; therefore, C_d increases.

For the long channel device, C_d = 0.

From this two-capacitor equivalent circuit, it is evident that the drain voltage has a similar effect on the channel potential as the gate voltage.

When V_{ds} is present, less V_{gs} is needed to pull the barrier down to 0.2 eV; therefore, V_t is lower by definition..

$$V_t = V_{t-long} - V_{ds} \cdot \frac{C_d}{C_{oxe}}$$

More accurately, V_{ds} should be supplemented with a constant that represents the combined effects of the 0.2 eV built-in potential between the N$^-$ inversion layer and both the N$^+$ drain and source at the threshold condition.

$$V_t = V_{t-long} - (V_{ds} + 0.4V) \cdot \frac{C_d}{C_{oxe}}$$

Solution of the Poisson's equation indicated that C_d is an exponential function of L in the two-dimensional structure.

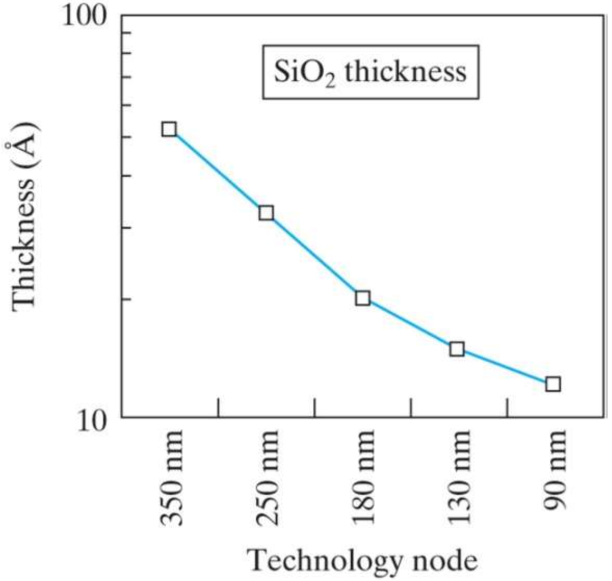
$$V_t = V_{t-long} - (V_{ds} + 0.4V) \cdot e^{-L/l_d}$$

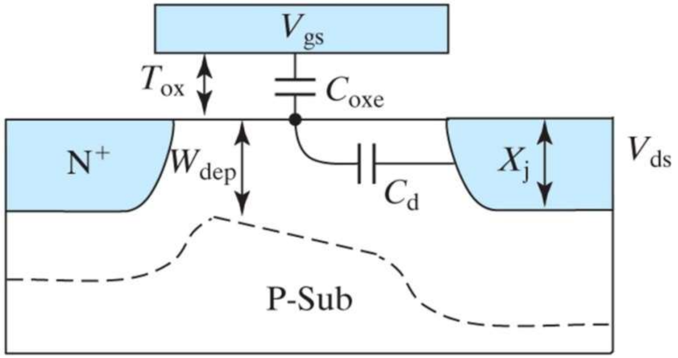
At a very large L, V_t is equal to $V_{t\text{-}long}$ as expected.
The roll-off is an exponential function of L and also is larger at larger V_{ds}.
The acceptable minimum L is several times of l_d.

$$where \quad l_d \propto \sqrt[3]{T_{oxe} W_{dep} X_j}\ , \ called\ the\ DIBL\ characteristic\ length$$

The vertical dimensions in a MOSFET (T_{ox}, W_{dep}, X_j) must be reduced in order to support the reduction of the gate length L.

Reducing T_{ox} increases the gate control or C_{oxe}.
Reducing X_j decreases C_d by reducing the size of drain electrode.
Reducing W_{dep} also reduces C_d by introducing a ground plane(the neutral region of the substrate or the bottom of the depletion region) that tends to electrostatically shield the channel from the drain.

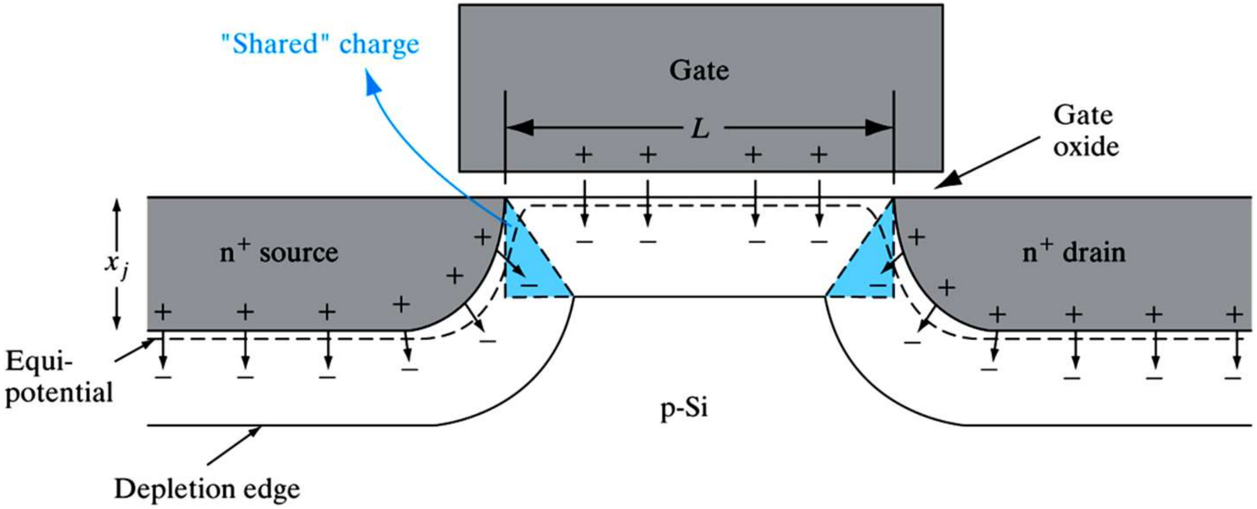




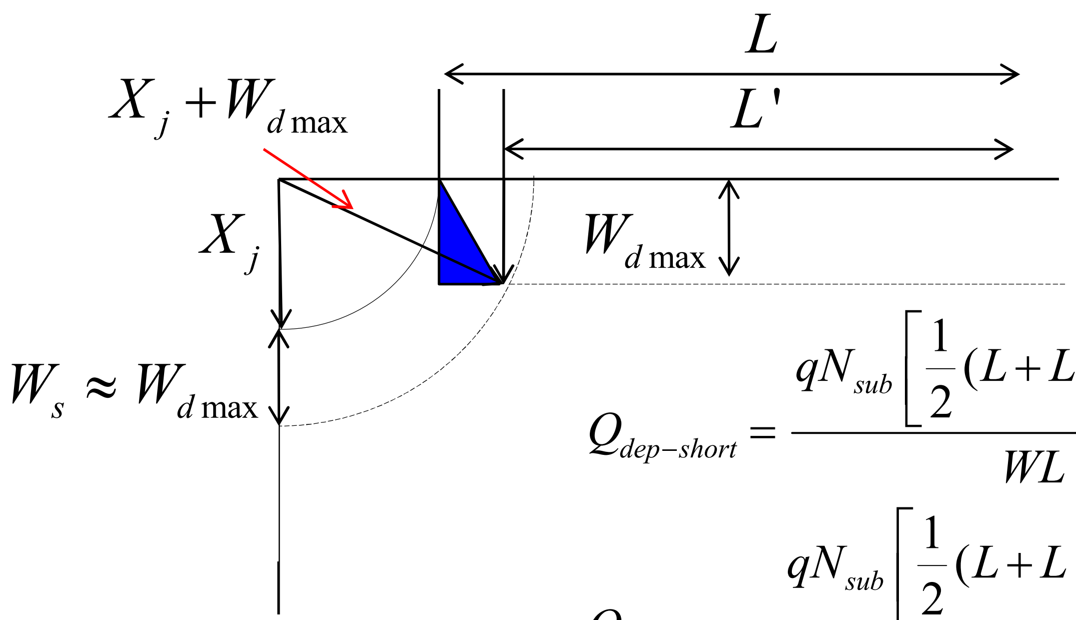
In the past,
the gate oxide thickness has been scaled roughly in proportion to the line width.

3) Third Approach: Charge Sharing Model

- ✓ The mechanism of SCE is due to charge sharing between the S/D and the gate.
- ✓ There are **shared charges** by both G and S/D.
- ✓ This shared region should not be counted in the V_T expression.
  → Replace the original Q_d in the rectangular region underneath the gate by a lower Q_d in the trapezoidal region.
- ✓ For a long channel device, the triangular depletion charge regions near the S and D are a very small fraction of the total depletion charge underneath the gate.
- ✓ However, as the channel length is reduced, the shared charge becomes a larger fraction of the total.



- **V_T roll-off** as a function of L



$$(X_j + W_{d\max})^2 = (X_j + \frac{L-L'}{2})^2 + W_{d\max}^2$$

$$\therefore L' = L - 2X_j\left[\sqrt{1+\frac{2Wd\max}{X_j}} - 1\right]$$

$$Q_{dep-short} = \frac{qN_{sub}\left[\frac{1}{2}(L+L')WW_{d\max}\right]}{WL} = qN_{sub}W_{d\max}\frac{L+L'}{2L}$$

$$Q_{dep-long} = \frac{qN_{sub}\left[\frac{1}{2}(L+L')WW_{d\max}\right]}{WL} \approx \frac{qN_{sub}(WLW_{d\max})}{WL} = qN_{sub}W_{d\max}$$

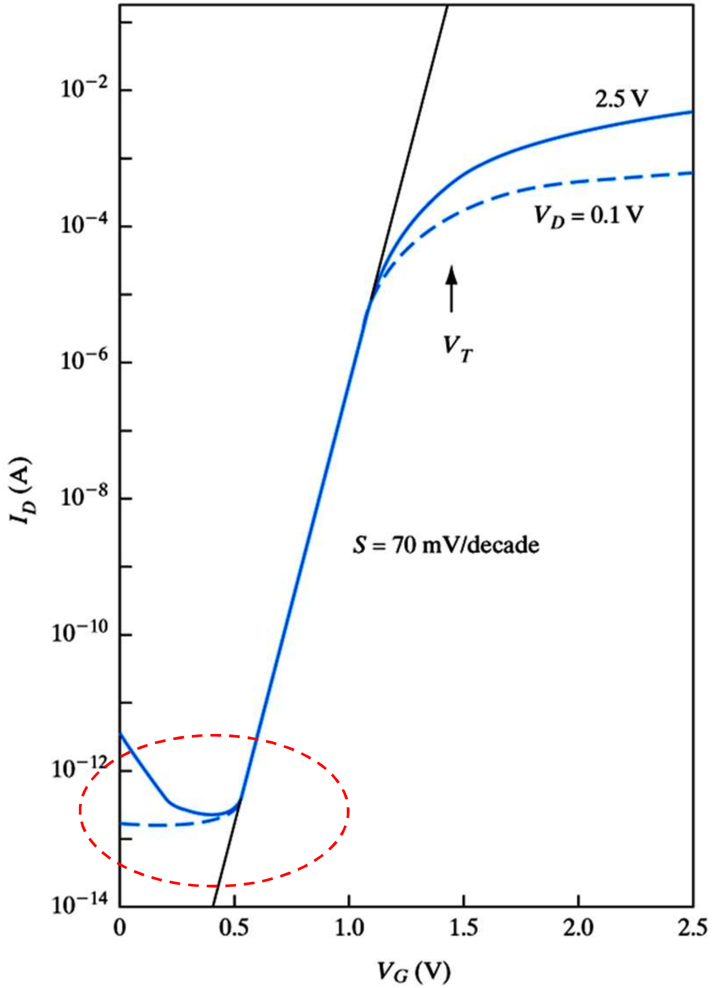
$$\therefore \Delta V_t = V_{t-long} - V_{t-short} = \frac{Q_{dep-long}}{C_{ox}} - \frac{Q_{dep-short}}{C_{ox}} = \frac{qN_{sub}W_{d\max}}{C_{ox}}\left(1 - \frac{L+L'}{2L}\right)$$

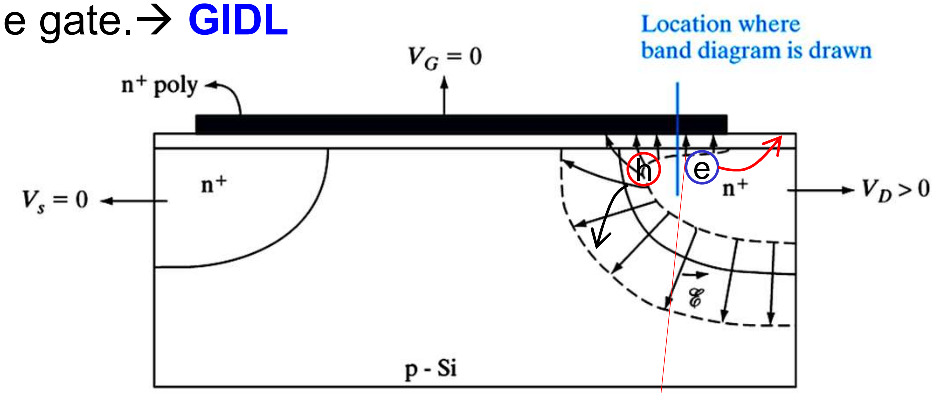
$$= \frac{qN_{sub}W_{d\max}}{C_{ox}}\frac{X_j}{L}\left[\sqrt{1+\frac{2W_{d\max}}{X_j}} - 1\right]$$

Short channel effect can be minimized by **reducing T_{ox}** (increasing C_{ox}), **reducing X_j**, and **reducing W_{dmax}**.

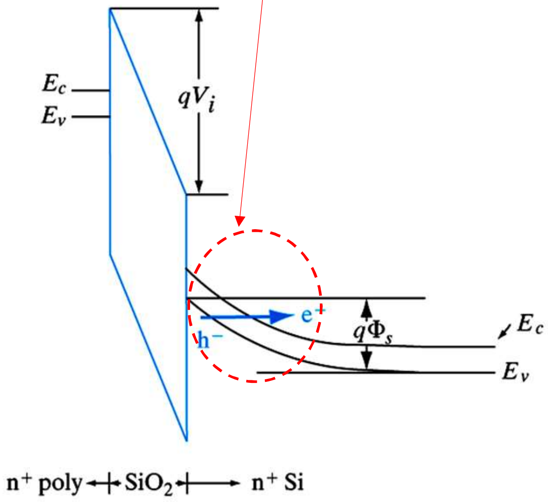
Gate-Induced Drain Leakage (GIDL)

As the gate voltage is reduced below V_t, the sub-threshold current drops and then bottoms out at a level determined by the S/D diode leakage. However, for even more negative gate biases, the off-state leakage current actually goes up as we try to turn off the MOSFET more for high V_{ds}, due to the direct tunneling of electrons from the valence band to the conduction band in the drain region under the gate.→ **GIDL**



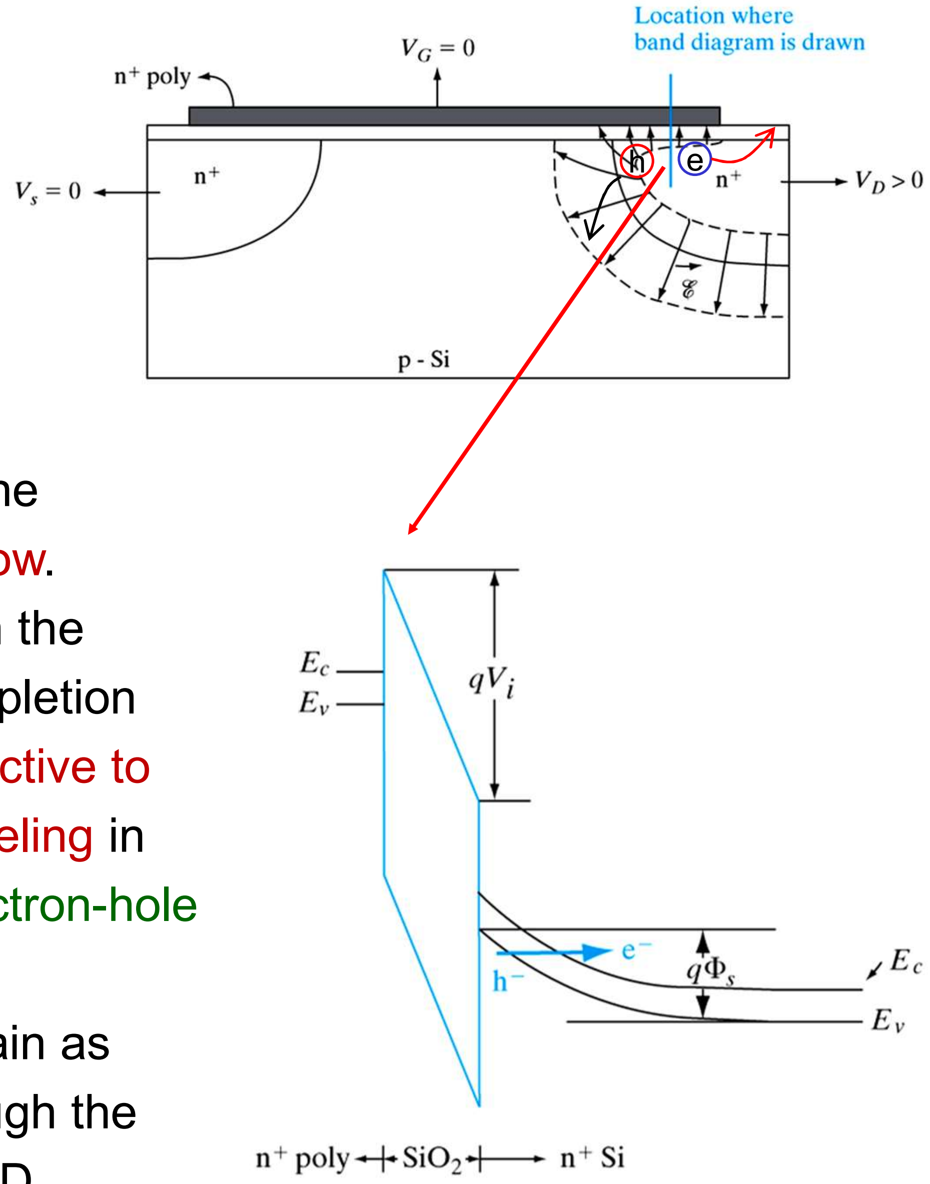


alignment overlapping between the gate electrode & D contacts.→ **GIDL**



• Gate-Induced Drain Leakage (GIDL)

- ✓ As the gate is made more negative (or alternatively, for a fixed gate bias, the drain is made more positive), a depletion region forms in the n-type drain.
- ✓ Since the drain doping is high, the depletion widths tend to be narrow.
- ✓ If the band bending is more than the bandgap Eg across a narrow depletion region, the conditions are conductive to band-to-band field-induced tunneling in this region, thereby creating electron-hole pairs.
- ✓ The electrons then go to the drain as GIDL. This tunneling is not through the gate oxide, but entirely in the Si D.



Reducing Gate-Insulator Electrical Thickness and Tunneling Leakage

300 nm for 10 μm tech.

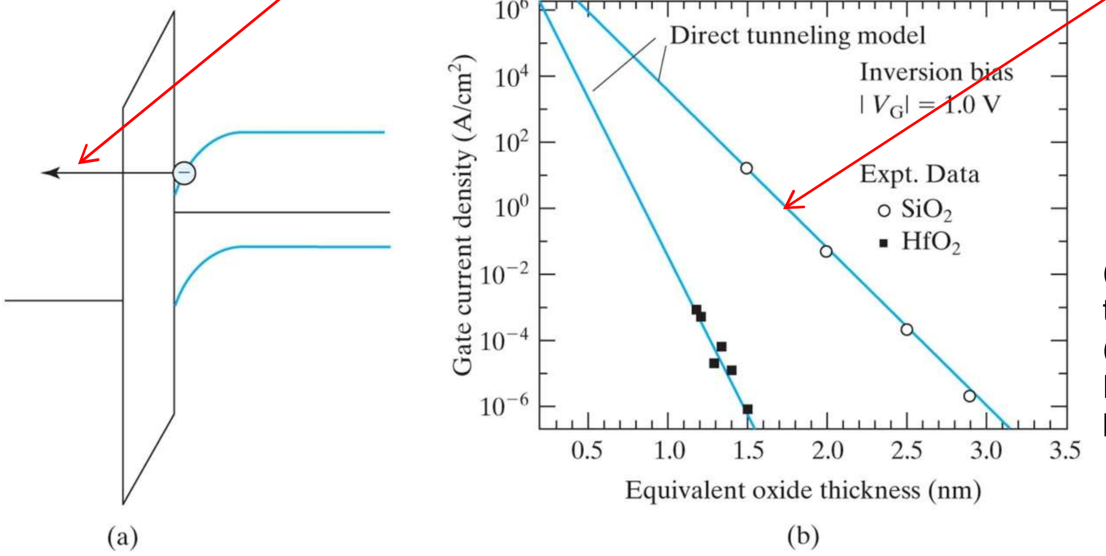
1.2 nm for 65 nm tech.

Two reasons for the relentless drive to reduce the oxide thickness;

1) A thinner oxide, i.e., a larger C_{ox} raises I_{on} and a large I_{on} raises the circuit speed.
2) A thinner oxide reduces V_t roll-off (and therefore the subthreshold leakage) in the presence of a shrinking *L*.

Limiting factors for using a thinner oxide;

1) Oxide breakdown (electric field in the thin oxide can be so high as to cause destructive breakdown).
2) Long term reliability (long term operation at high field breaks the weaker chemical bonds at the Si-SiO$_2$ interface thus creating oxide charge and V_t shift).
3) Tunneling leakage current (most serious limiting factor for SiO$_2$ films thinner than 1.5 nm).



Exponential rise of the SiO$_2$ leakage current with decreasing thickness. (The leakage current can be reduced by about 10 × with the addition of nitrogen into SiO$_2$.

(a) Energy band diagram in inversion showing electron tunneling path through the gate oxide;
(b) 1.2 nm SiO$_2$ conducts 10^3 A/cm^2 of leakage current. High-k dielectric such as HfO$_2$ allows several orders lower leakage current to pass. (After [6]. © 2003 IEEE.)

$$V_t = V_{t-long} - (V_{ds} + 0.4V)\cdot e^{-L/l_d},\ where\ l_d \propto \sqrt[3]{T_{oxe} W_{dep} X_j},\ T_{oxe} = T_{ox} + T_{poly-dep} + T_{inv}$$

High-*k* dielectric technology to replace SiO_2: HfO_2, ZrO_2, Al_2O_3……….

HfO_2: *k* ~ 24 (six times larger than that of SiO_2)

Equivalent oxide thickness (or **EOT**) of 6 nm-thick HfO_2 is 1 nm in the sense of producing same C_{ox} in SiO_2.
However, the HfO_2 film presents a much thicker tunneling barrier to the electrons and holes and allows the leakage current with several orders of magnitude smaller than that through SiO_2.

The difficulties of adopting high-*k* dielectrics:

1) chemical reactions between them and the silicon substrate,
2) lower surface mobility than the Si-SiO_2 system, and
3) more oxide charges

} Inserting a thin SiO_2 interfacial layer

Metal gate technology: The poly-Si gate depletion layer thickness also needs to be minimized. Metal is a much better gate material in this respect.

NFET and PFET gates may require two different metals (with metal work functions close to those of N^+ and P^+ poly-Si) in order to achieve the optimal V_ts).

Reduction of T_{inv}: The material parameters that determine T_{inv} is the effective mass. A larger effective mass (or a lower mobility) leads to a thinner T_{inv}.

The effective mass in the direction normal to the oxide interface determines T_{inv}, while the effective mass in the direction of the current flow determines the surface mobility.

It may be possible to build a transistor with a wafer orientation that offers larger m_n and m_p normal to the oxide interface but smaller m_n and m_p in the direction of the current flow.

How To Reduce W_{dep}

$$W_{dep} = \sqrt{\frac{2\varepsilon_s \varphi_{st}}{qN_{sub}}}$$

W_{dep} can be reduced by increasing the substrate doping concentration, N_{sub}.

$$V_t = V_{fb} + \phi_{st} - \frac{Q_{dep}}{C_{ox}} = V_{fb} + \phi_{st} + \frac{\sqrt{qN_{sub} 2\varepsilon_s \phi_{st}}}{C_{ox}}$$

If V_t is not to increase, N_{sub} must not be increased unless C_{ox} is increased, i.e., T_{ox} is reduced.

Eliminating N_{sub},

$$V_t = V_{fb} + \phi_{st}\left(1 + \frac{2\varepsilon_s T_{ox}}{\varepsilon_{ox} W_{dep}}\right)$$

W_{dep} can only be reduced in proportional to T_{ox}.

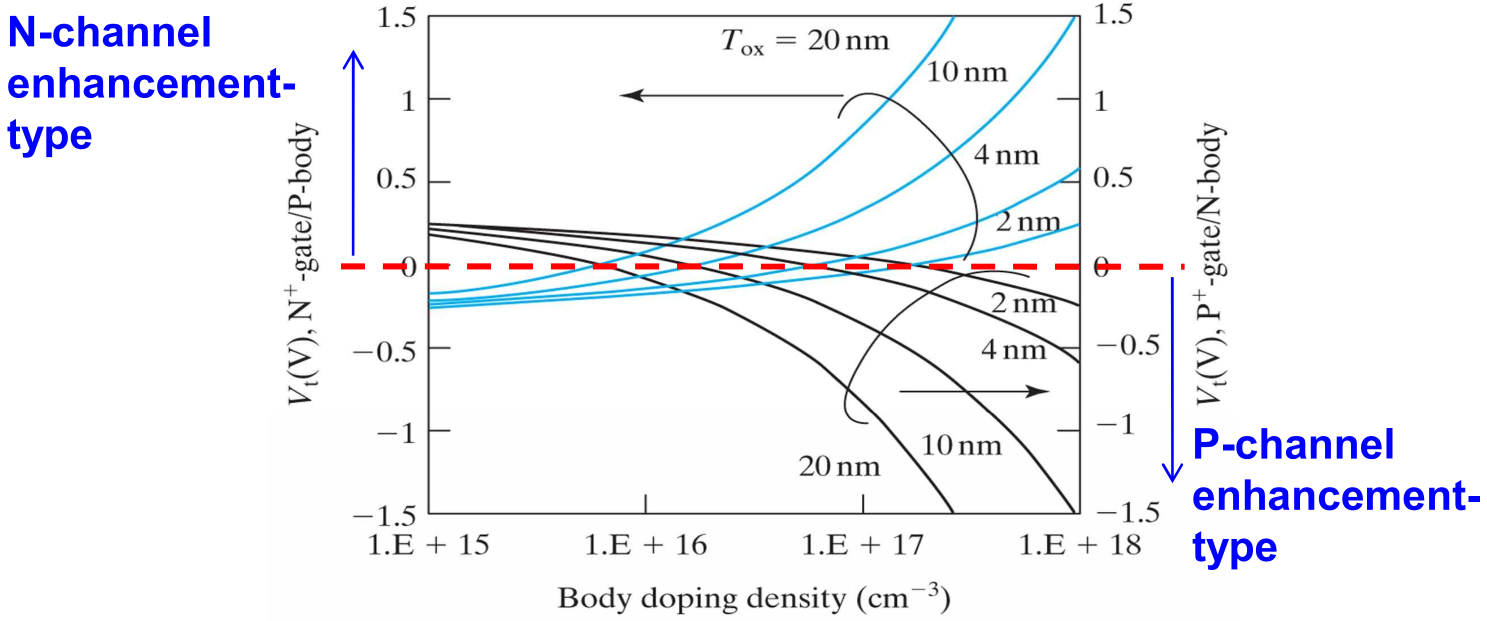
This fact establishes T_{ox} as the main enabler of L reduction.

In Chapter 5

Choice of V_t and Gate Doping Type

To make circuit design easier, it is routine to set V_t at a small positive value, e.g., 0.4 V, so that, at $V_g = 0$, the transistor does not have an inversion layer and current does not flow between the two N^+ regions.

Enhancement-Type Device



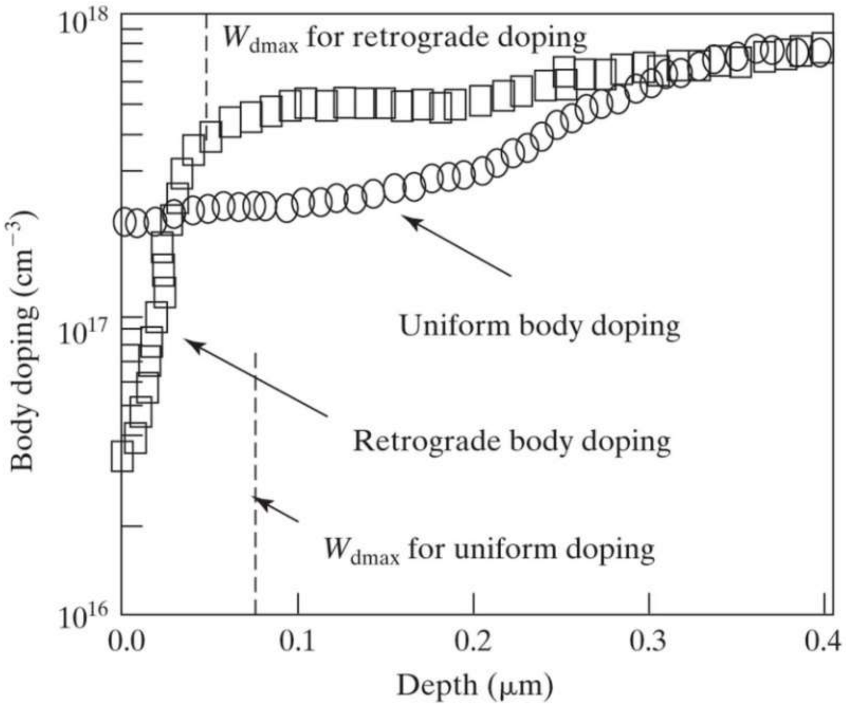
N-channel enhancement-type

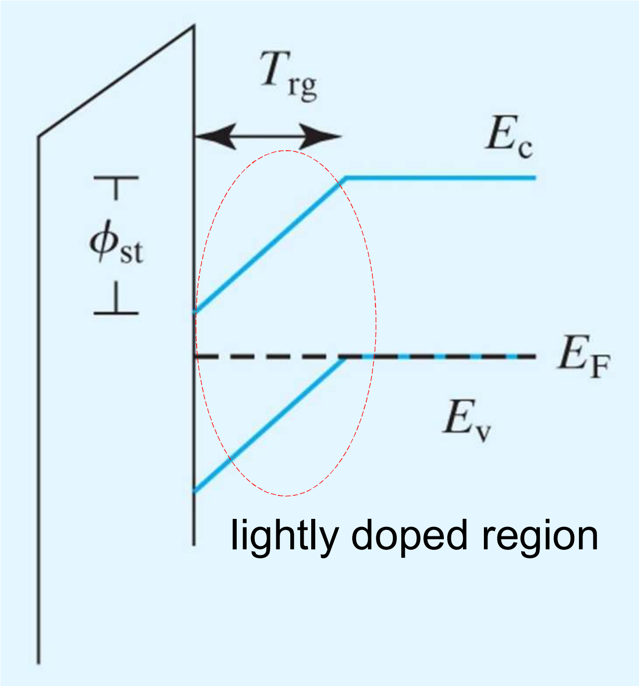
P-channel enhancement-type

P-type body is almost always paired with N^+ gate to achieve a small positive threshold voltage, and N-type body is normally paired with P^+ gate to achieve a small negative threshold voltage.

For the case of P-type body paired with P^+ gate, V_t would be too large (over 1 V) and necessitate a larger power supply voltage. This would lead to larger power consumption and heat generation.

Steep retrograde doping: another way of reducing $W_{dep.}$



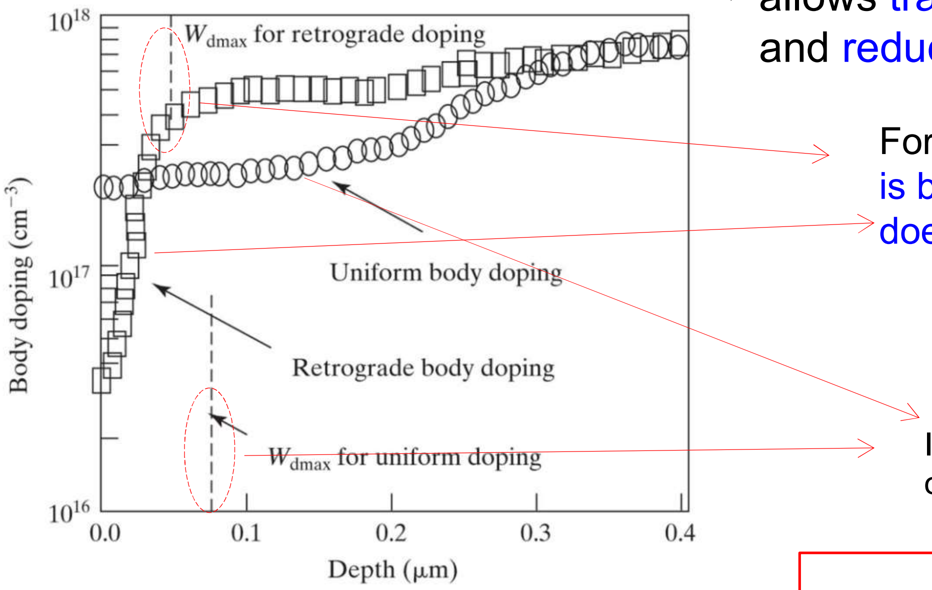


lightly doped region

Steep Retrograde Doping

; light doping in a thin surface layer and very heavy doping underneath

⟹ allows transistor shrinking to smaller size for cost reduction and reduces impurity scattering.



For steep retrograde doping, the depletion-layer thickness is basically the thickness of the lightly doped region and does not significantly change as V_{sb} increases.

In earlier generation of MOSFETs, the body doping density is more or less uniform and W_{dmax} varies with V_{sb}.

For steep retrograde doping,

$$\boxed{V_t(V_{sb}) = V_{t0} + \frac{C_{dep}}{C_{oxe}} V_{sb} = V_{t0} + \alpha V_{sb}}$$

$$\boxed{C_{dep} \;\; and\; \alpha \approx constants \Rightarrow W_{dmax} \;\; and\; C_{dep}/C_{oxe} \;\; ratio: independent\; of\; the\; body\; bias}$$

$$V_t(V_{sb}) = V_{t0} + \frac{C_{dep}}{C_{oxe}} V_{sb} = V_{t0} + \alpha V_{sb}: linear\;\; relationship\;\; between\;\; V_t \;\; and\;\; V_{sb}$$

$$V_{t0} = V_{fb} + 2\phi_B + \frac{\sqrt{qN_a 2\varepsilon_s 2\phi_B}}{C_{ox}}$$

α : **body-effect coefficient**

In Chapter 6

Steep retrograde doping: another way of reducing $W_{dep.}$

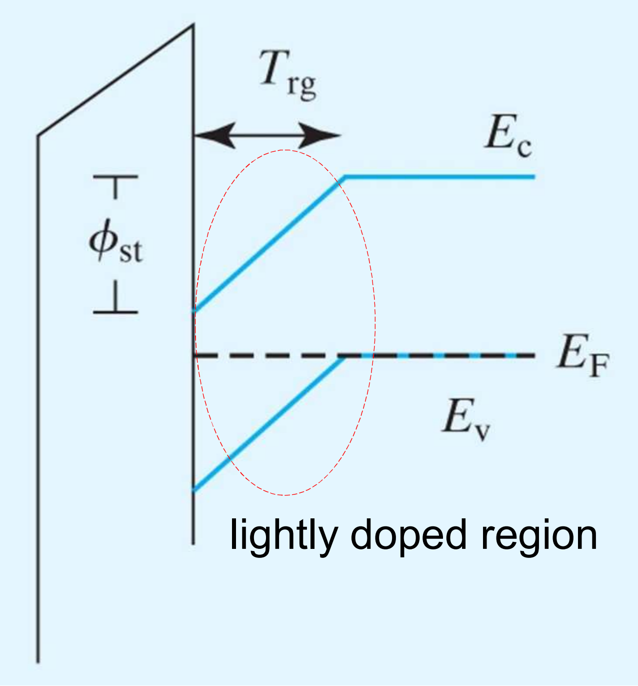
The band bending, ϕ_{st} , is dropped uniformly over T_{rg}, the thickness of the lightly doped depletion layer, creating an electric field, $\mathscr{E}_s = \phi_{st} / T_{rg}$.

$$V_{ox} = T_{ox}\mathscr{E}_{ox} = T_{ox}\mathscr{E}_s \cdot \frac{\varepsilon_s}{\varepsilon_{ox}} = \phi_{st}\frac{\varepsilon_s T_{ox}}{\varepsilon_{ox} T_{rg}}$$

$$\Leftarrow V_g = V_{fb} + \phi_s + V_{ox}$$

$$\therefore V_t = V_{fb} + \phi_{st} + V_{ox} = V_{fb} + \phi_{st}\left(1 + \frac{\varepsilon_s T_{ox}}{\varepsilon_{ox} T_{rg}}\right)$$

Again, T_{rg}, can only be scaled in proportion to T_{ox}.



Energy diagram of a steep-retrograde doped MOSFET at the threshold condition.

Uniform doping case : $V_t = V_{fb} + \phi_{st}\left(1 + \frac{2\varepsilon_s T_{ox}}{\varepsilon_{ox} W_{dep}}\right)$

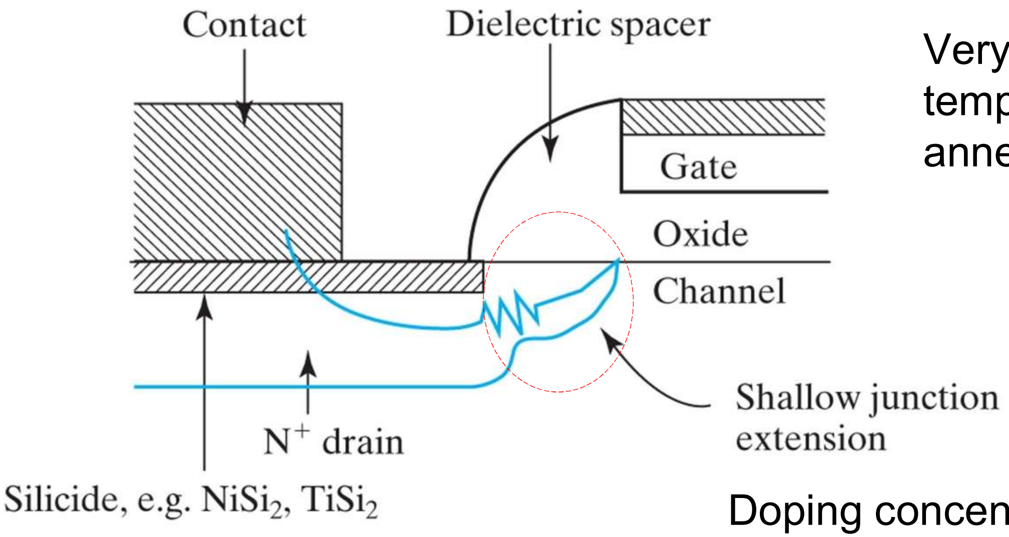
Advantages;

1) T_{rg}, the W_{dep} of an ideal retrograde device, can be about half the W_{dep} of a uniformly doped device, which yield the same V_t.

2) Ionized impurity scattering in the inversion layer is reduced and the surface mobility can be higher.

Shallow Junction and Metal Source/Drain MOSFET

Shallow junction is needed because the drain junction depth must be kept small.



Very short annealing at the lowest necessary temperature is used to activate the dopant and anneal out the implantation damage.

Doping concentration in the shallow junction extension is kept much lower than the N^+ doping density.

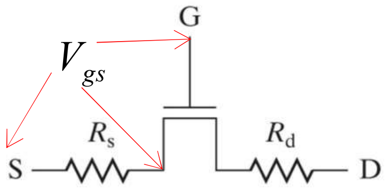
The **shallow junction extension** next to the channel helps to **suppress the V_t roll-off**.

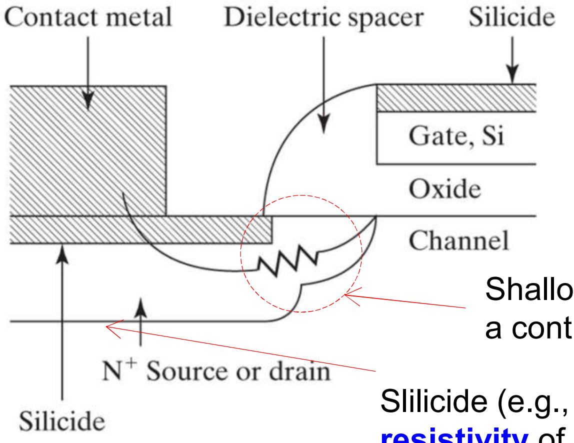
However, shallow junction and light doping combine to produce an undesirable parasitic resistance that reduces the precious I_{on}.

A price to pay for suppressing V_t roll-off and the subthreshold leakage current.

Parasitic Source-Drain Resistance

Shallow junction is needed to prevent excessive off-state leakage I_{ds} in short channel transistor.





Shallow diffusion under the dielectric spacer is a contributor to the parasitic resistance.

Slilicide (e.g., $TiSi_2$ or $NiSi_2$) reduces the **sheet resistivity** of N^+ (or P^+) source-drain regions by a factor of ten. It also reduces **contact resistance**.

If $R_s = 0$,

$$I_{dsat0} \approx WC_{oxe}v_{sat}(V_{gs} - V_t - m\mathcal{E}_{sat}L)$$

If $R_s \neq 0$,

$$I_{dsat} \approx WC_{oxe}v_{sat}(V_{gs} - I_{dsat}R_s - V_t - m\mathcal{E}_{sat}L) = I_{dsat0} - WC_{oxe}v_{sat}I_{dsat}R_s \Rightarrow I_{dsat}(1 + WC_{oxe}v_{sat}R_s) = I_{dsat0}$$

$$\Rightarrow \boxed{I_{dsat} = \frac{I_{dsat0}}{1 + WC_{oxe}v_{sat}R_s} = \frac{I_{dsat0}}{1 + I_{dsat0}R_s/(V_{gs} - V_t - m\mathcal{E}_{sat}L)}}$$

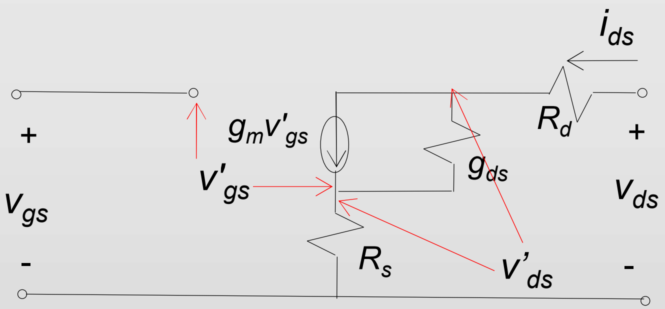
Parasitic resistance significantly reduces I_{dsat} and increases V_{ds}.

$$\boxed{V_{dsat} = V_{dsat0} + I_{dsat}(R_s + R_d)}$$

In Chapter 6

Effect of R_s and R_d

With low frequency equivalent circuit,



$$v_{ds} = v'_{ds} + (R_s + R_d)i_{ds}$$

$$v_{gs} = v'_{gs} + R_s i_{ds}$$

$$i_{ds} = g_{ds} v'_{ds} + g_m v'_{gs}$$

$$i_{ds} = \left[\frac{g_m}{1 + R_s g_m + (R_s + R_d)g_{ds}}\right] v_{gs} + \left[\frac{g_{ds}}{1 + R_s g_m + (R_s + R_d)g_{ds}}\right] v_{ds}$$

$$g'_m = g_{meff} = \left[\frac{g_m}{1 + R_s g_m + (R_s + R_d)g_{ds}}\right]$$

$$g'_{ds} = g_{dseff} = \left[\frac{g_{ds}}{1 + R_s g_m + (R_s + R_d)g_{ds}}\right]$$

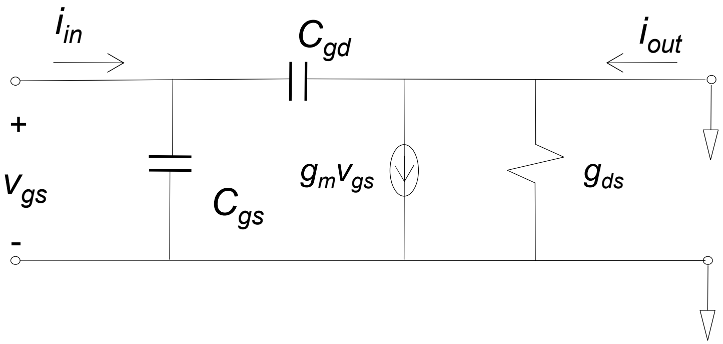
R_s and R_d *affect* g'_m and g'_{ds}.

Cutoff Frequency, f_T (unity current gain frequency)

: the frequency where the MOSFET is no longer amplifying the input signal

$$\left|\frac{i_{out}}{i_{in}}\right| = 1, \ \ with\ ouput\ short\ circuted$$

With high frequency equivalent circuit,



$$i_{in} = j\omega(C_{gs} + C_{gd})WL_g v_{gs} \approx j(2\pi f)C_{ox}WL_g v_{gs}$$

$$i_{out} \approx g_m v_{gs}$$

$$\left|\frac{i_{out}}{i_{in}}\right| = \frac{g_m v_{gs}}{2\pi f C_{ox} W L_g v_{gs}}\bigg|_{f=f_T} = 1$$

$$\boxed{f_T = \frac{g_m}{2\pi C_{ox} W L_g}}$$

In Chapter 6

MOSFETs with Metal Source/Drain

A **metal source/drain MOSFET** or **Schottky source/drain MOSFET** can have very shallow junctions (good for the short-channel effect) and low series-resistance because the silicide is ten times more conductive than N^+ or P^+ Si.

↓

The ultimate way to reduce the increasingly important parasitic resistance.

The energy band diagram in the off state (at $V_g = 0$) is similar to that of a conventional MOSFET.

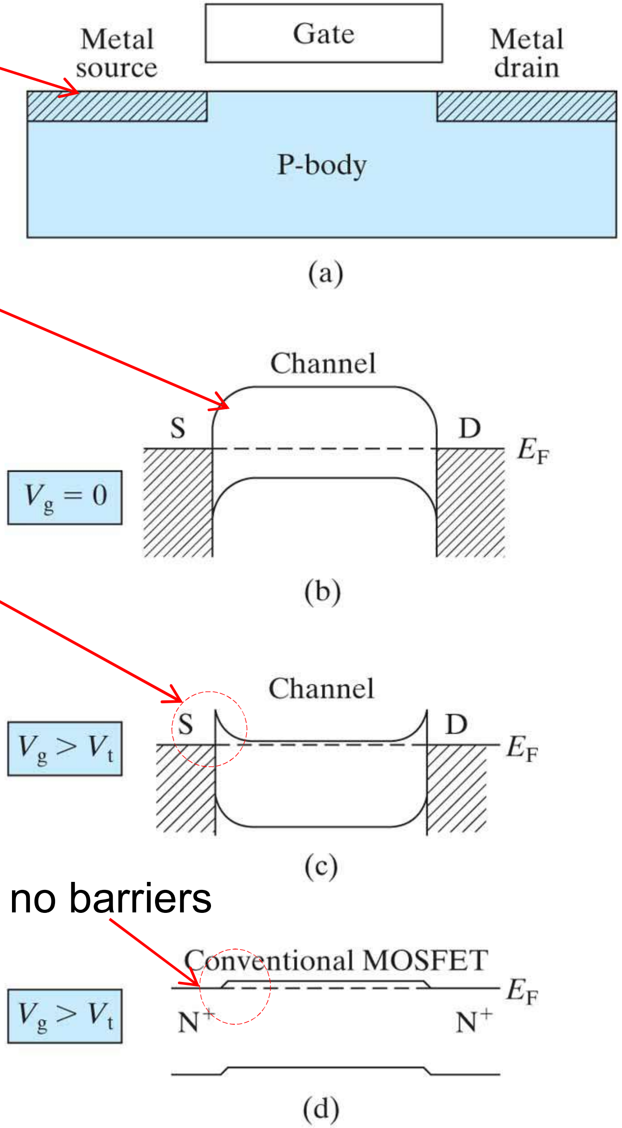
In the on state, there may be energy barriers ϕ_B impeding current flow and must be minimized.

The only problem is that the Schottky-S/D MOSFET would have a lower I_d than the regular MOSFET if ϕ_B is too large to allow easy flow of carriers from the source into the channel.

To unleash the full potential of Schottky S/D MOSFT;

A very low-ϕ_B Schottky junction technology should be used.
A thin N^+ region can be added between the metal and the channel.

Attention must be paid to reduce the large reverse leakage current of a low-ϕ_{Bn} Schottky drain to body junction.



no barriers

Trade-Off between I_{on} and I_{off} and Design for Manufacturing

Using a higher V_t can decrease subthreshold I_{off} .

That is not acceptable because a high V_t would reduce I_{on} and therefore reduce circuit speed.

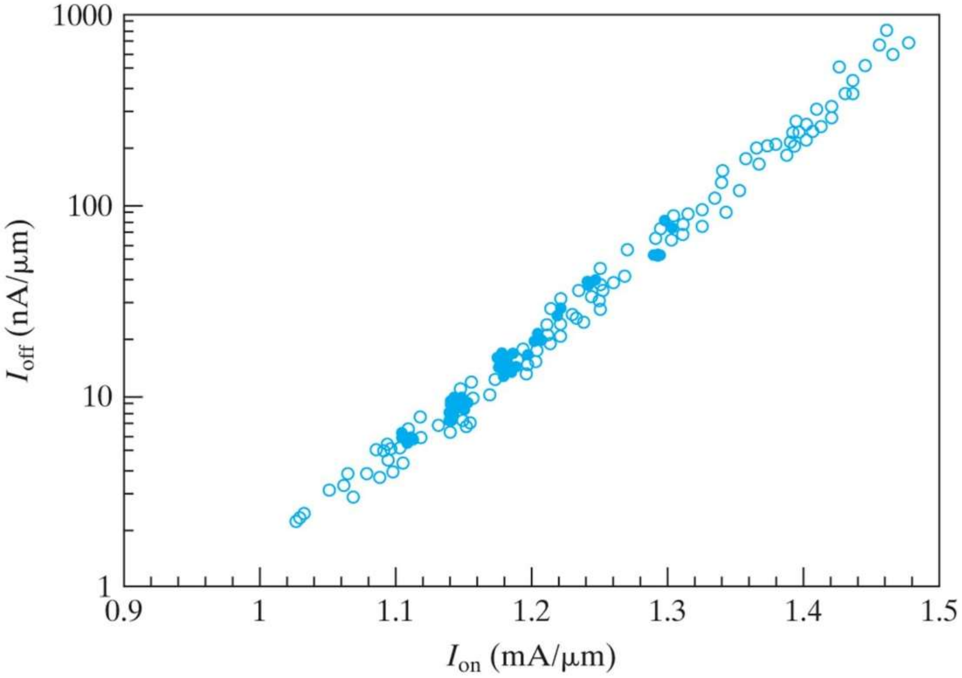
Using a larger V_{dd} can raise I_{on}.

That is not acceptable either because it would raise the power consumption.

Decreasing L can raise I_{on}.

That is not acceptable because it would also reduce V_t and raise I_{off}.

Which, if any, of the following changes lead to both subthreshold leakage reduction and I_{on} enhancement? A larger V_t. A larger L. A smaller V_{dd}.



Trade-off between I_{on} and I_{off}, *i.e.*, **between speed and standby power consumption**.

⬇

Higher I_{on} goes hand-in-hand with larger I_{off}:

Log I_{off} vs. linear I_{on}. The spread in I_{on} (and I_{off}) is due to the presence of several slightly different drawn L_{gs} and unintentional manufacturing variations in L_g and V_t. (After [2]. © 2003 IEEE.)

Techniques to address the trade-off between I_{on} and I_{off}

Multiple (two, three, or even more) V_ts.

A large circuit may be designed with only the high-V_t devices first. Circuit timing simulations are performed to identify those signal paths and circuits where speed must be tuned up.

Intermediate-V_t devices are substituted into them.

Finally, low-V_t devices are substituted into those few circuits that need speed.

Multiple V_{dd}.

A higher V_{dd} is provided to a small number of circuits that need speed while a lower V_{dd} is used in the other circuits.

The larger V_{dd} provides higher speed and/or allows a larger V_t to be used (to suppress leakage).

The dynamic power consumption can be kept low because most of the circuits operate at the lower V_{dd}.

Well bias technique.

In a large circuit such as a microprocessor, only some circuit blocks need to operate at high speed at a given time and other circuit blocks operate at lower speed or are idle.

V_t can be set relatively low to produce large I_{on} so that circuits that need to operate at high speed can do so.

A well bias voltage, V_{sb}, is applied to the other circuit blocks to raise the V_t and suppress the subthreshold leakage. This technique requires intelligent control circuits to apply V_{sb} where and when needed.

A well bias technique also provides a way to compensate for the chip-to-chip and block-to-block variations in V_t that results from nonuniformity among devices due to inevitable variations in manufacturing equipment and process.

Design for manufacturing or **DFM**

Many techniques at the border between manufacturing and circuit design can help to ease the problem of manufacturing variations, collectively known as **design for manufacturing** or **DFM**.

↳ mainly due to the imperfect control of L_g in the lithography process

Systematic variation (more or less predictable)

Distortion in lithography due to the interference of neighboring patterns of light and darkness

Elaborate mathematical optical proximity correction or OPC reshapes each pattern in the photomask to compensate for the neighboring patterns.

Variation of the carrier mobility and the current due to the mechanical stress effect (created by nearby structure, e.g., shallow trench isolation or other MOSFETs.)

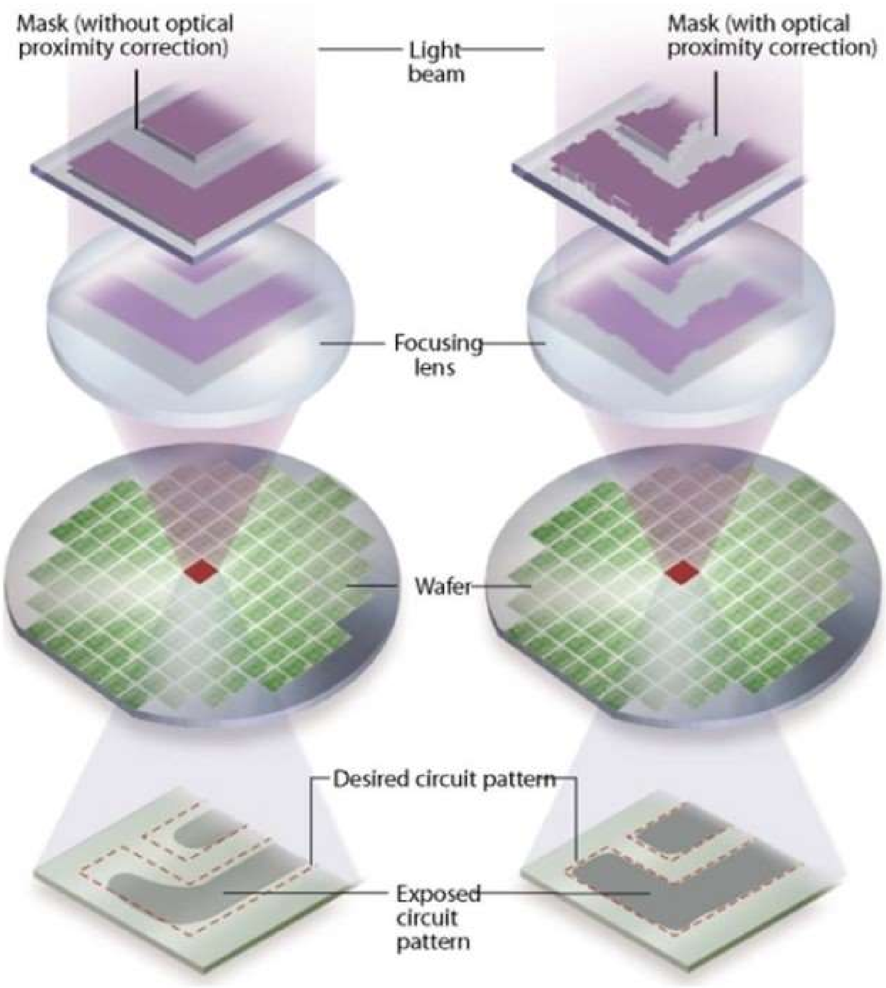
Sophisticated simulation tools can analyze the mechanical strain and predict the I_{on} based on the neighboring structure and feed the I_{on} information to circuit simulators to obtain more accurate simulation results.

Random variation (unpredictable)

Gate edge roughness or waviness (caused by the graininess of the photoresist and the poly-crystalline Si).

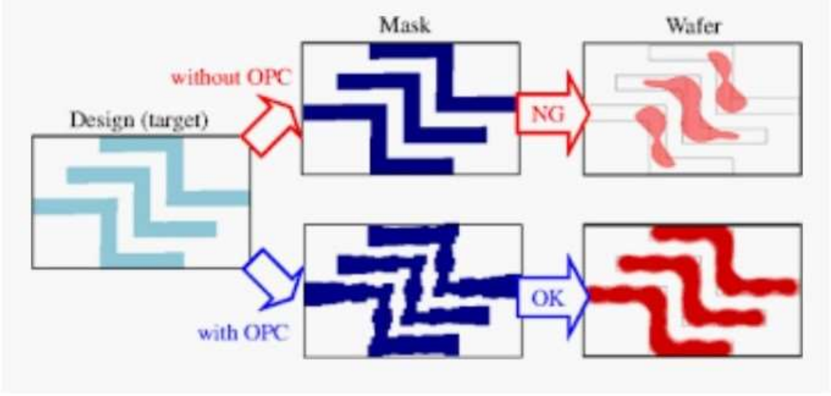
Random dopant fluctuation phenomenon (The statistical variation of the number of dopant atoms and their location in small size MOSFET creates significant variations in the threshold voltage).

OPC (optical proximity correction)

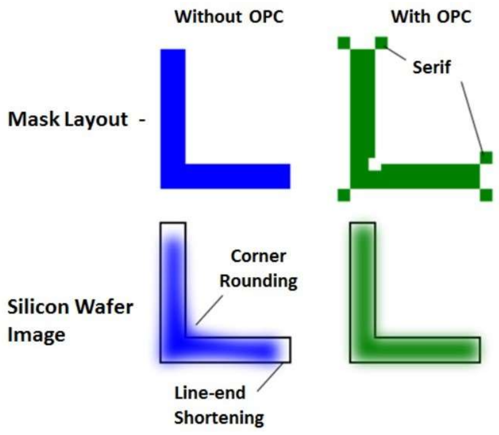

Mask (without optical proximity correction)
Light beam
Mask (with optical proximity correction)
Focusing lens
Wafer
Desired circuit pattern
Exposed circuit pattern

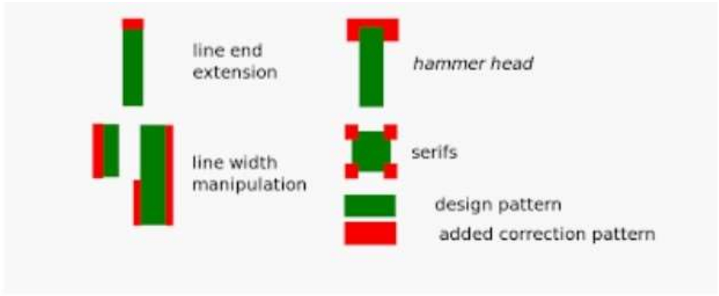
OPC for Semiconductor Manufacturing

(Image source: IEEE Spectrum, 2003)



Optical proximity correction with hierarchical Bayes model
spiedigitallibrary.org





File:Optical proximity correction structures.svg - Wikimedia Commons
commons.wikimedia.org

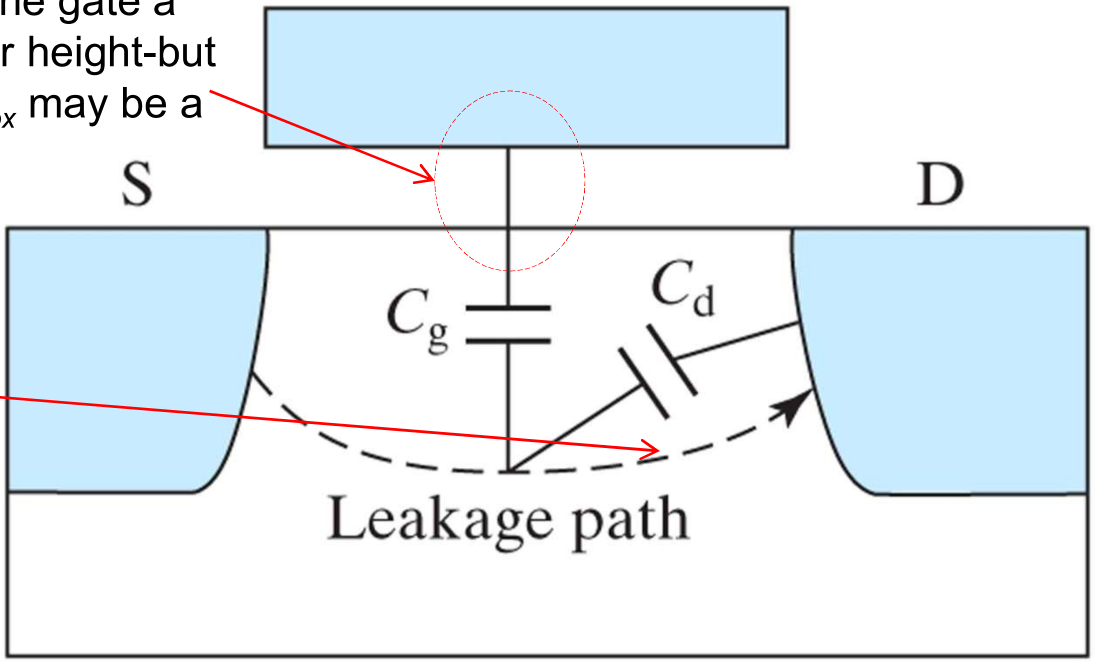
Ultra-Thin-Body SOI and Multigate MOSFETs

To suppress V_t roll-off, it is required to maximize the gate-to-channel capacitance and minimize the drain-to-channel capacitance.

To do former, we reduce T_{ox} as much as possible. To accomplish the latter, we reduce W_{dep} and X_j as much as possible. It is increasingly difficult to make these dimensions smaller.

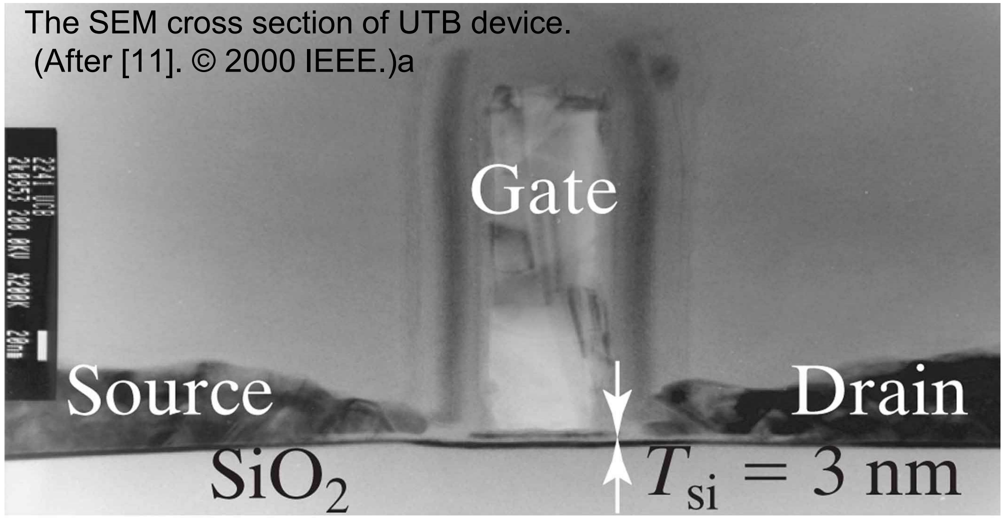
An infinitesimally small T_{ox} would give the gate a perfect control over the potential barrier height-but only right at the Si surface, because T_{ox} may be a small part T_{oxe} and the T_{inv} is large.

The drain could still have more control than the gate along another leakage current path that is some distance below the Si surface.

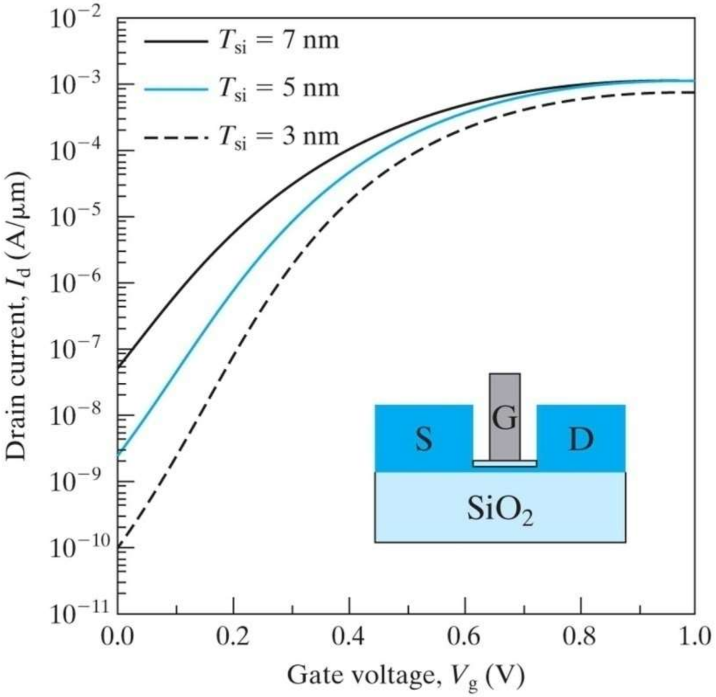


There are two transistor structures that can eliminates the leakage paths that are far away from the gate. One is called the **ultra-thin-body MOSFET** or **UTB MOSFET**. The other is **multigate MOSFET**.

Ultra-Thin-Body MOSFET and SOI


The SEM cross section of UTB device.
(After [11]. © 2000 IEEE.)a

Since the film of the UTB structure is very thin, no leakage path is very far from the gate.



The subthreshold leakage is reduced as the Si film (transistor body) is made thinner. L_g = 15 nm. (After [11]. © 2000 IEEE.)

T_{si} should take the places of W_{dep} and X_j such that L_g can be scaled roughly in proportion to T_{Si}.

$$V_t = V_{t-long} - (V_{ds} + 0.4V)\cdot e^{-L/l_d}, \ where\ l_d \propto \sqrt[3]{T_{oxe} W_{dep} X_j}$$

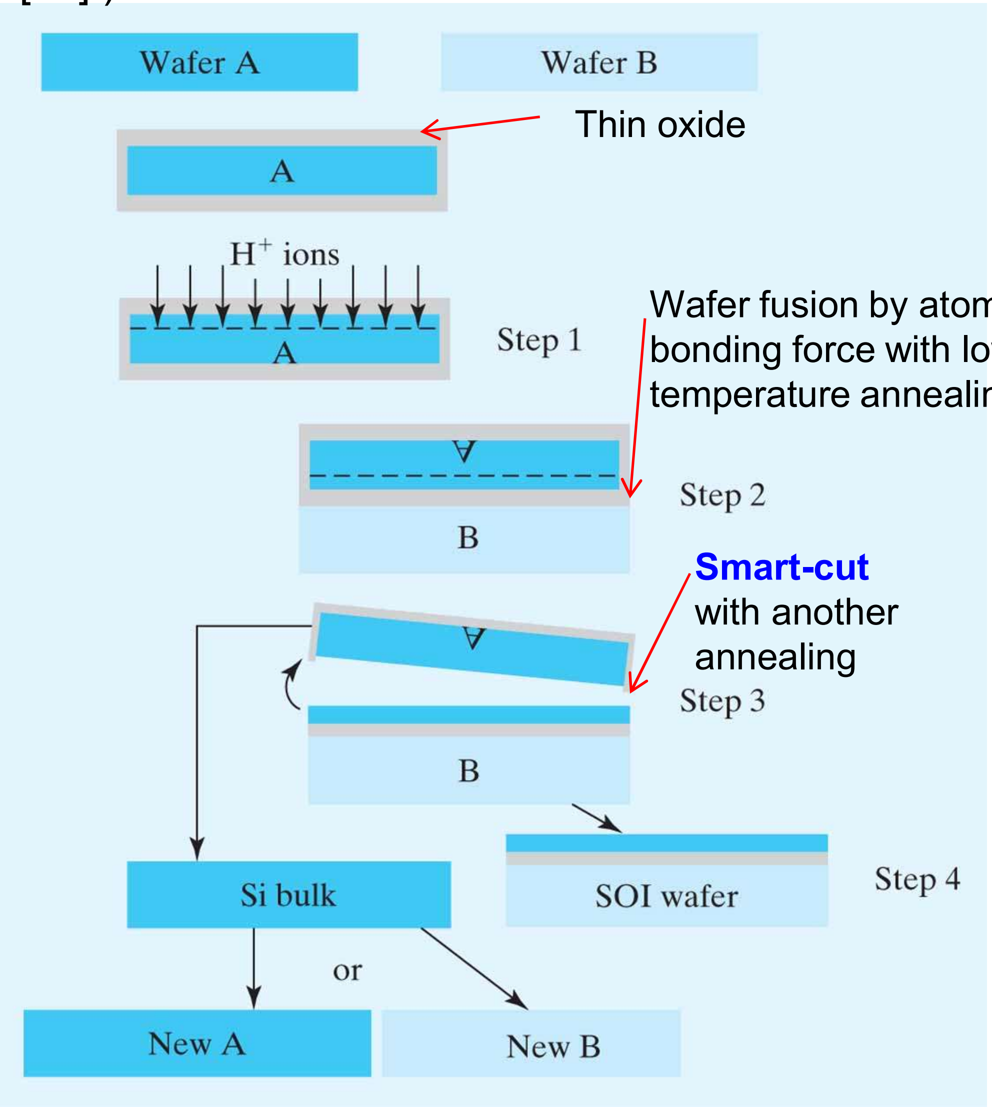
Additional device benefits of the UTB MOSFETs;

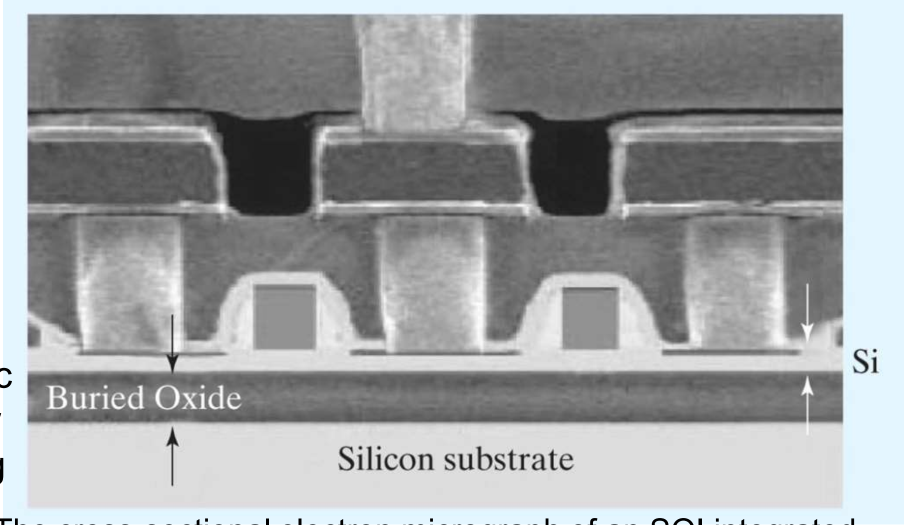
1) Carrier mobility is improved, because small I_d can be obtained without heavy doping.

2) Body effect that is detrimental to circuit speed is eliminated because the body is **fully depleted** and floating and has no fixed voltage.

One challenge posed by UTB MOSFETs is the large source/drain resistance due to their thinness. The solution is to use the thicker **raised source and drain** with epitaxial deposition.

Steps of making an SOI wafer. (After [12].)


Wafer A
Wafer B
Thin oxide
A
H+ ions
A
Step 1
Wafer fusion by atomic bonding force with low temperature annealing
A
B
Step 2
Smart-cut with another annealing
A
Step 3
B
Si bulk
SOI wafer
Step 4
or
New A
New B

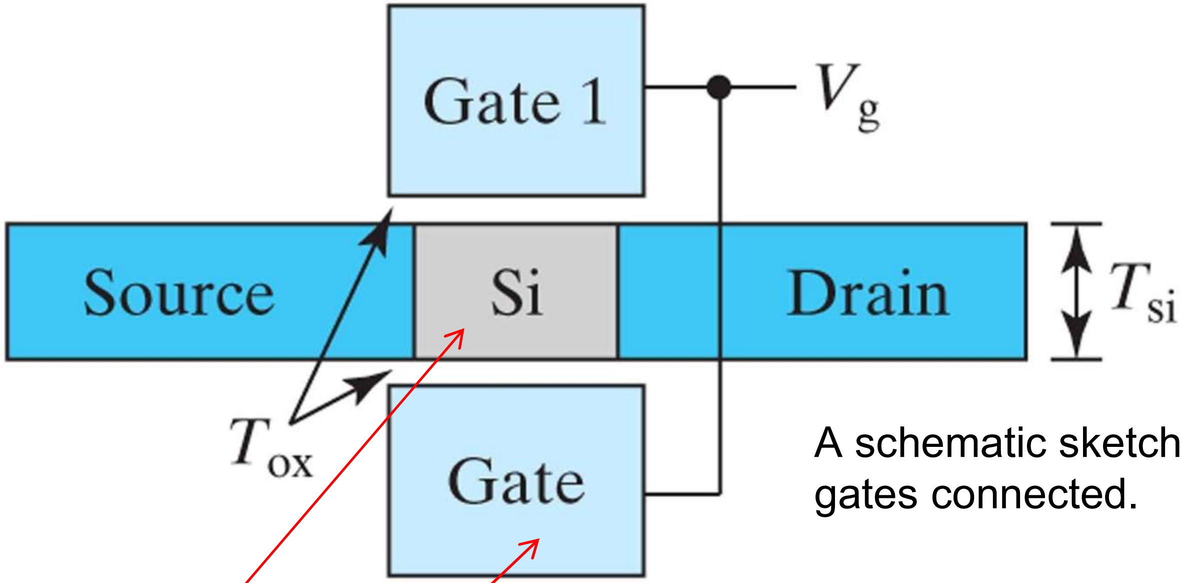


The cross-sectional electron micrograph of an SOI integrated circuit. The lower level structures are transistors and contacts. The upper two levels are the vias and the interconnects, which employ multiple layers of materials to achieve better reliability and etch stops.

SOI provides a speed advantage because the source/drain to body junction capacitance is practically eliminated as the source and drain diffusion regions extends vertically to the buried oxide.

The cost of an SOI wafer is higher than an ordinary Si wafer and increases the cost of IC chips.

FinFET– Multigate MOSFET



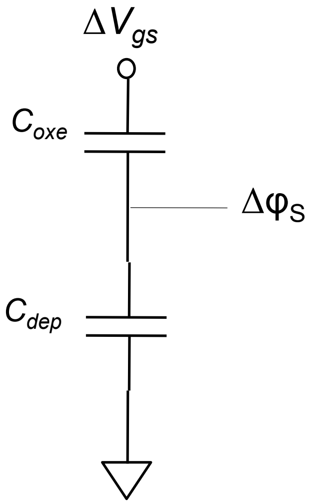
A schematic sketch of a **double-gate MOSFET** with gates connected.

The Si film is very thin so that no leakage path is far from one of the gates. (the worst-case path is along the center of the Si film.)

The gate (s) can suppress leakage current more effectively than the conventional MOSFET. Because there are more than one gate, the structure may be called **multigate MOSFET**.

Shrinking T_{si} automatically reduces W_{dep} and X_j and V_t roll-off can be suppressed to allow L_g to shrink to as small as a few nm.

Because the top and bottom gates are at the same voltage and the Si film is fully depleted, the Si surface potential moves up and down with V_g mV for mV in the subthreshold region. The voltage divider effect does not exist and η is desired unity and I_{off} is very low.



$$\frac{\Delta\varphi_s}{\Delta V_{gs}} = \frac{C_{oxe}}{C_{oxe}+C_{dep}} \Rightarrow \frac{d\varphi_s}{dV_{gs}} = \frac{C_{oxe}}{C_{oxe}+C_{dep}} \equiv \frac{1}{\eta}, \quad \therefore \varphi_s = constant + \frac{V_{gs}}{\eta}$$

$$where \;\; \eta = 1 + \frac{C_{dep}}{C_{oxe}} \approx 1$$

$$I_{ds} \propto n_s \propto e^{q\varphi_s/kT} \propto e^{q(constant+V_{gs}/\eta)/kT} = constant \cdot e^{qV_{gs}/\eta kT}$$

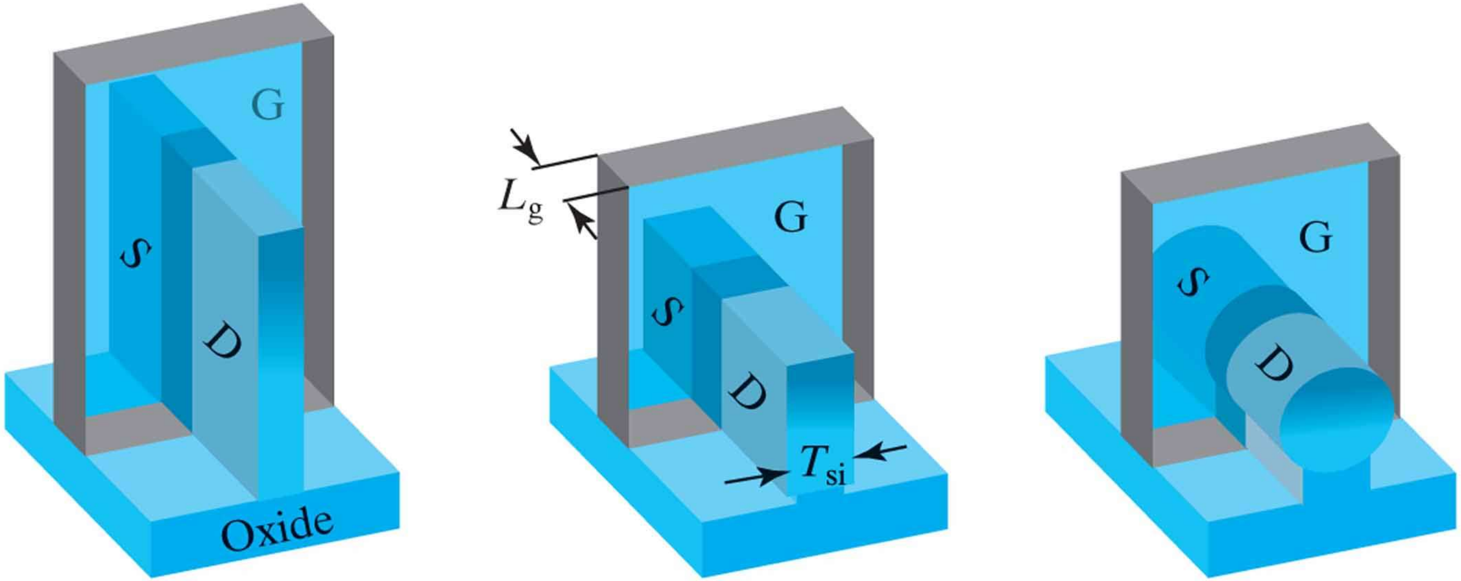
$$I_{off}(nA) = 100 \cdot \frac{W}{L} \cdot e^{q(-V_t)/\eta kT} = 100 \cdot \frac{W}{L} \cdot 10^{-V_t/S}$$

There is no need for heavy doping in the channel to reduce W_{dep}. This leads to low vertical field and less impurity scattering; as a result the mobility is higher.

Finally, there are two channels (top and bottom) to conduct the transistor current. For these reasons, a multigate MOSFET can have shorter L_g, lower I_{off}, and larger I_{on} than a single-gate MOSFET.

But there is one problem–how to fabricate the multigate MOSFET structure.

Variations of FinFET



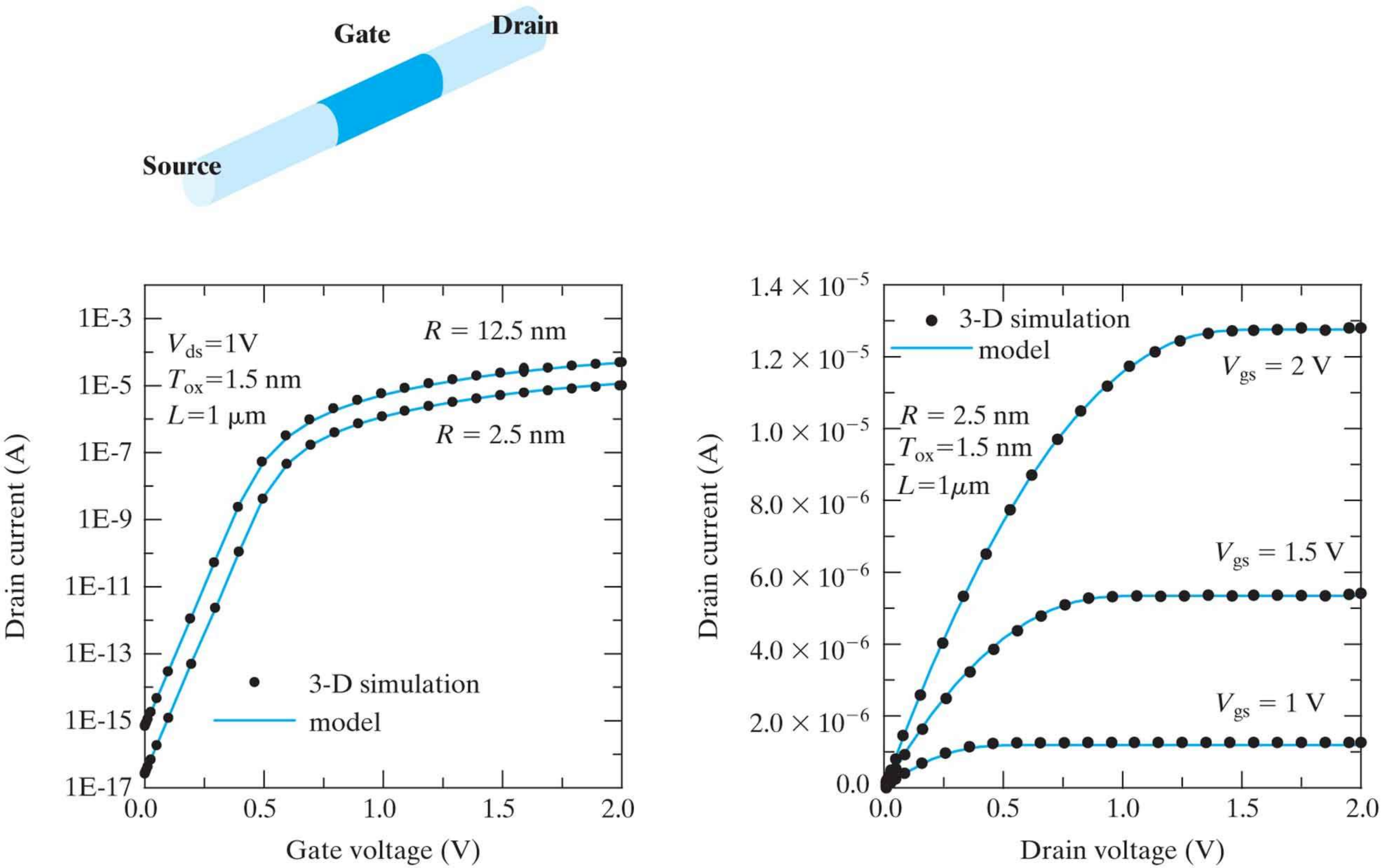
Tall FinFET　Short FinFET　Nanowire FET

Tall FinFET has the advantage of providing a large W and therefore large I_{on} while occupying a small footprint.

Short FinFET has the advantage of less challenging lithography and etching. The top surface of the fin contributes significantly to the suppression of V_t roll-off and the leakage control. This structure is also known as a **triple-gate MOSFET**.

Nanowire FET gives the gate even more control over the transistor body by surrounding it.

FinFETs can also be fabricated on bulk Si substrates.



Simulated I–V curves of a nanowire MOSFET. R is the nanowire radius. (After [16].)

Output Conductance

Output conductance limits the transistor voltage gain. However, its cause and theory are intimately related to those of V_t roll-off.

$$Maximum\ Voltage\ Gain = \frac{g_{msat}}{g_{ds}} \Leftarrow \quad v_{out} = \frac{-g_{msat}}{g_{ds} + 1/R} \times v_{\text{in}}$$

Output conductance,

$$g_{ds} \equiv \frac{dI_{dsat}}{dV_{ds}} = \frac{dI_{dsat}}{dV_t} \cdot \frac{dV_t}{dV_{ds}}$$

$$= g_{msat} \times e^{-L/l_d}$$

Since I_{ds} is a function of V_{gs} - V_t , it is obvious that

$$\frac{dI_{dsat}}{dV_t} = \frac{-dI_{dsat}}{dV_{gs}} = -g_{msat}$$

From $V_t = V_{t-long} - (V_{ds} + 0.4V) \cdot e^{-L/l_d}$, $where\ l_d \propto \sqrt[3]{T_{oxe} W_{dep} X_j}$

$$\frac{dV_t}{dV_{ds}} = -e^{-L/l_d}$$

$$\boxed{Intrinsic\ voltage\ gain = \frac{g_{msat}}{g_{ds}} = e^{+L/l_d}}$$

Increasing V_{ds} would reduce V_t . That is why I_{ds} continues to increase without saturation.

The output conductance is caused by the drain/channel capacitive coupling, the same mechanism that is responsible for V_t roll-off.That is why I_{ds} continues to increase without saturation.

This is why g_{ds} is larger in a MOSFET with shorter L. To reduce g_{ds} or to increase the intrinsic voltage gain, we can use a large L and/or reduced l_d.

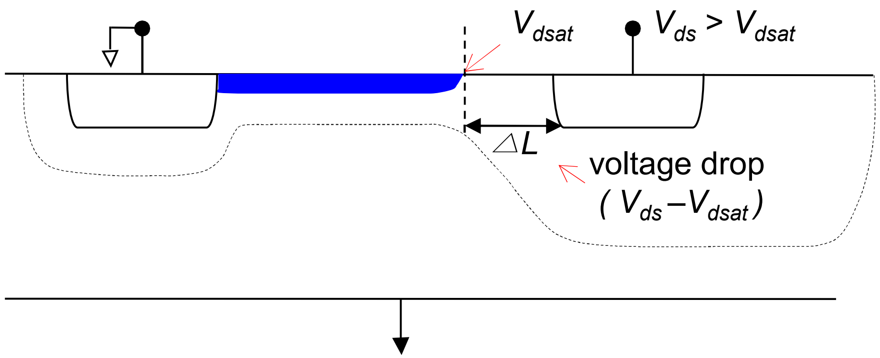
Every design change that improves the suppression of the V_t roll-off also suppresses g_{ds} and improves the voltage gain.

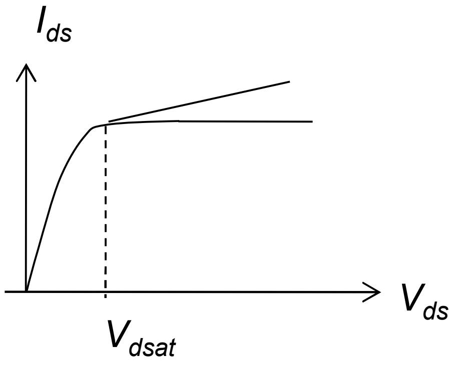
Channel length modulation

V_t dependence on V_{ds} is the main cause of output conductance in very short MOSFETs. For larger L and V_{ds} close to V_{dsat}, another mechanism may be the dominant contributor to g_{ds} – **channel length modulation.**

Strong saturation, $V_{gs} > V_t$

and $V_{ds} > (V_{gs}-V_t)/m$





The effective channel length decreases with increasing V_{ds}. I_{ds}, which is inversely proportional to L, thus increases without true saturation.

$$I_{ds} \propto \frac{1}{L-\Delta L} = L^{-1}(1-\frac{\Delta L}{L})^{-1} \square \frac{1}{L}\left(1+\frac{\Delta L}{L}\right), \quad for \;\; large \;\; L$$

$$I_{dsat} \approx \frac{W}{2mL}\mu_n C_{ox}(V_{gs}-V_t)^2(1+\frac{\Delta L}{L})$$

$$g_{ds} = \frac{\partial I_{dsat}}{\partial V_{ds}} = I_{dsat} \cdot \frac{\partial \Delta L}{\partial V_{ds}} \approx \frac{l_d \cdot I_{dsat}}{L(V_{ds}-V_{dsat})}$$

This component of g_{ds} can also be suppressed with **larger *L* and smaller *T_{ox}*, *X_j*, *and* *W_{dep}*.**

Device and Process Simulation

Device simulation is an important tool that provides the engineers with quick feedback about device behaviors. This narrows down the number of variables that need to be checked with expensive and time-consuming experiments.

Most of the equations are solved simultaneously, e.g.,
Fermi-Dirac probability
incomplete ionization of dopants
drift and diffusion currents
current continuity equation
Poisson equation……..

Related to device simulation is process simulation.

The input that a user provides to process simulation program are
lithography mask pattern
implantation dose and energy
temperatures and times for oxide growth
annealing steps………..

The process simulator generates a two- or three-dimensional structure with all the deposited or grown and etched thin film and doped regions. This output may be fed into a device simulator together with the applied voltages and the operating temperature as the input to the device simulator.

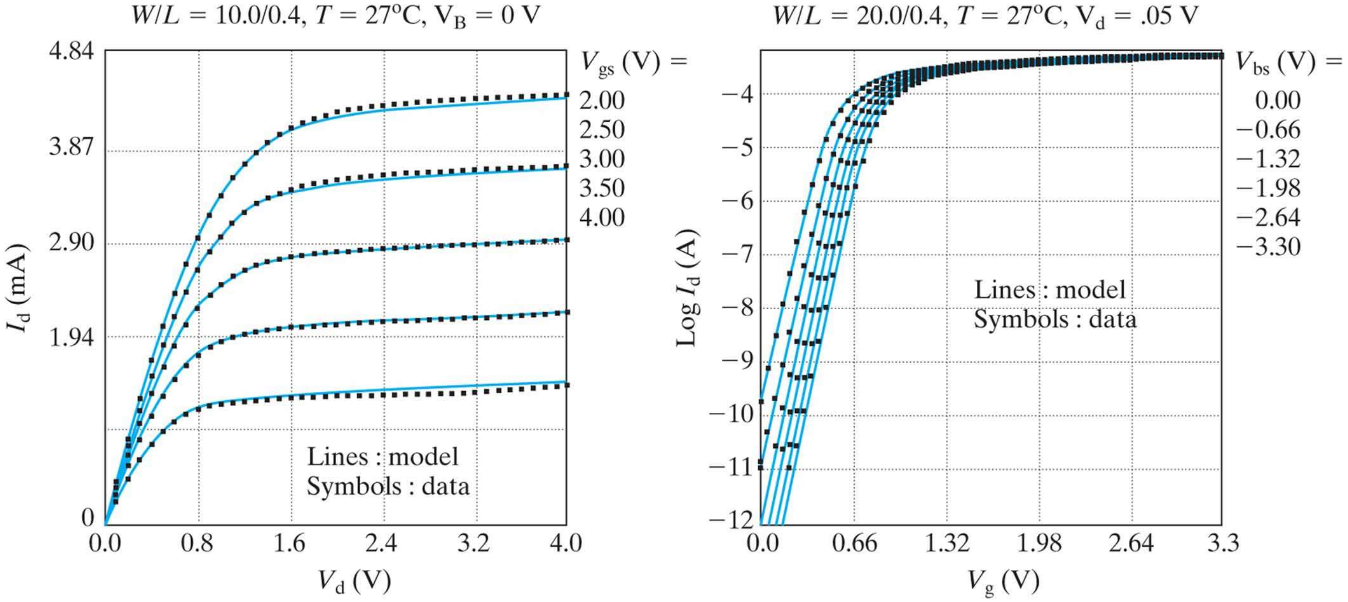
MOSFET Compact Model for Circuit Simulation

In circuit simulations, MOSFETs are modeled with analytical equations much like the ones introduced in this and the previous chapters. More details are introduced in the model equations than this textbook can introduce. These models are called **compact models** to highlight their computational efficiency in contrast with the device simulators described in Section 7.10.

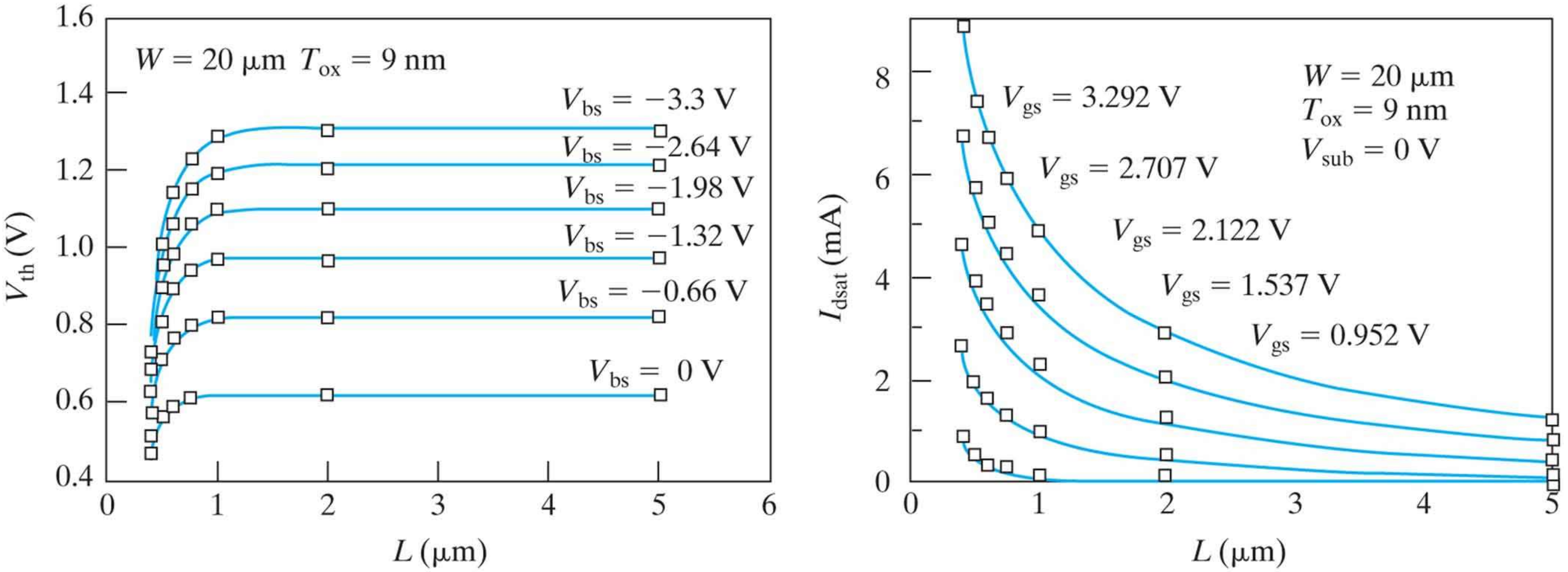
A compact model must capture all the subtle behaviors of the MOSFET over wide ranges of voltage, *L, W*, and temperature and present them to the circuit designers in the form of equations.

Some circuit-design methodologies, such as anlog circuit design, use circuit simulation directly. Other design methodologies use **cell libraries**. A cell library is a collection of hundreds of small building blocks of circuits that have been carefully designed and characterized beforehand using circuit simulation.

In 1977, an industry standard setting group selected **BSIM** as the first industry standard model.



Selected comparisons of BSIM and measured device data to illustrate the accuracy of a compact model. (After [18].)



A compact model needs to accurately model the transistor behaviors for any L and W that circuit designers may specify. (After [19]. © 1997 IEEE.)