Unit III - MOS Introduction (Part – II)

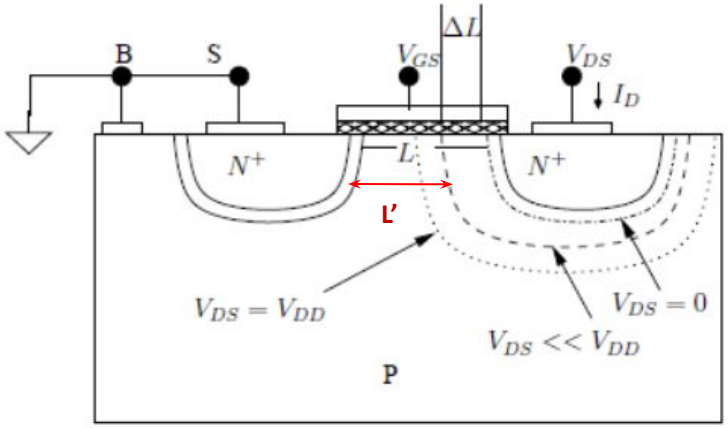
Prepared by

Dr. J.Selvakumar & Dr. A. Maria Jossy

Channel-Length Modulation

- As per equation (9), the transistor in the **saturation mode acts as a perfect current source** or that the current between drain and source terminal is a constant, independent of the applied voltage over the terminals. **This not entirely correct**.
- The effective length of the **conductive channel is actually modulated by the applied V_{DS}**: increasing V_{DS} causes the depletion region at the drain junction to grow, reducing the length of the effective channel.

$$I_{ds} = \beta \left[\frac{V_{ds}^2}{2}\right] \text{ -- (9)}$$



L = L' - ΔL

$$I_D = \frac{k'_n}{2}\frac{W}{L}(V_{GS} - V_T)^2$$

the current increases when the length factor L is decreased

A **more accurate description of the current of the MOS transistor** is therefore given in equation (10)

$$I_D = I_D'(1 + \lambda V_{DS}) \quad \text{---- (10)}$$

with I_D' the current expressions derived earlier, and λ an empirical parameter, called the *channel-length modulation*.

Channel-Length Modulation

- In **shorter transistors**, the drain-junction depletion region presents a larger fraction of the channel, and **the channel-modulation effect is more visible.**

- It is therefore advisable to resort to long-channel transistors if a high-impedance current source is needed.

Velocity Saturation

- The behavior of transistors with very short channel lengths (called *short-channel devices)* deviates considerably from the resistive and saturated models, due to channel length modulation

- **The main culprit for this deficiency is the *velocity saturation effect*.** *Eq. (11)* $\upsilon_n = -\mu_n \xi(x) = \mu_n \frac{dV}{dx}$ states that the velocity of the carriers is proportional to the electrical field, independent of the value of that field.
  - The electric field is given by $dV/dx = V_{ds}/L$, So if the channel length is reduced then dx gets too small blowing up the electric field and hence saturating the velocity.
  - The velocity is saturated beyond a critical field $E_{critical}$. Electrons encounter more collision and hence don't pick up speed.
  - **Maximum velocitv of electrons/holes = 10^5 m/s**.

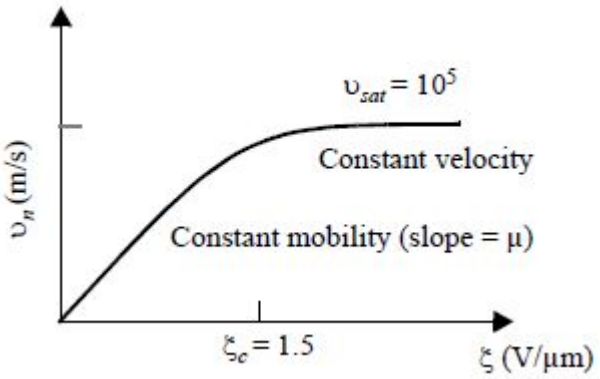


Fig. a : Velocity-saturation effect

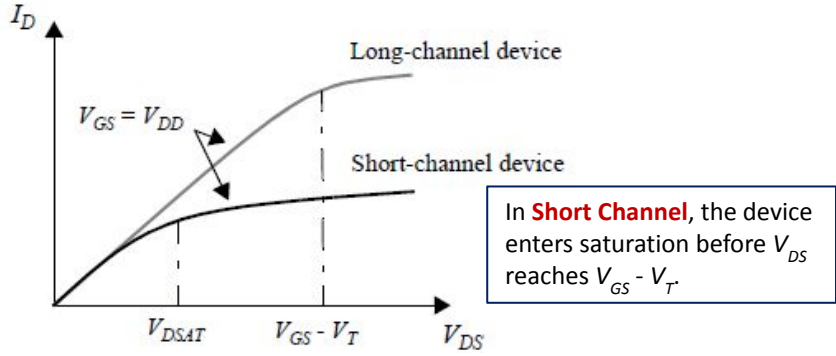


Fig. b: Short-channel devices display an extended saturation region due to velocity-saturation

- In other words, the carrier mobility is a constant. However, **at high field strengths, the carriers fail to follow this linear model.** In fact, **when the electrical field along the channel reaches a critical value** $E_{critical}$, ***the velocity of the carriers tends to saturate*** **due to scattering effects** (collisions suffered by the carriers).

Velocity Saturation

The velocity as a function of the electrical field, plotted in Fig. a can be roughly approximated by the following expression:

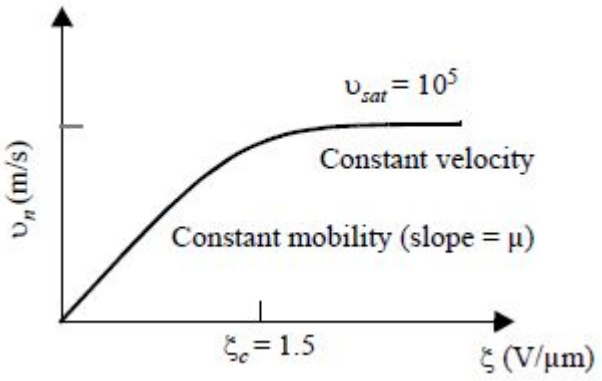


Fig. a : Velocity-saturation effect

$$\upsilon = \frac{\mu_n \xi}{1 + \xi/\xi_c} \quad \text{for} \quad \xi \le \xi_c$$
$$= \upsilon_{sat} \quad \text{for} \quad \xi \ge \xi_c$$

---- (11)

Where, ***Ec*** = Ecritical

Mobility Degradation

There are two reasons for mobility reduction in MOSFET

1. Mobility reduction with the gate voltage due to the vertical electric field
2. Mobility reduction with the drain voltage due to the horizontal electric field

- So far we have only considered the effects of the **tangential field** along the channel due to the V_{DS}, when considering **velocity-saturation effects**.
- However, there **also exists a normal (vertical) field originating from the gate voltage** that further inhibits channel carrier mobility. This effect, which is called ***mobility degradation***, reduces the surface mobility with respect to the bulk mobility. Eq. (12) provides a simple estimation of the mobility reduction.

$$\mu_H = \frac{\mu_0}{1+\frac{V_{ds}}{L_{eff}E_{crit}}} = \frac{\mu_0}{1+\theta_2 V_{ds}} \quad \text{----- (12)}$$

- $1/L_{eff}E_{crtit}$ is referred as drain bias mobility reduction parameter and in some texts denoted as θ_2. E_{crit} is the electric field shown in figure (a)
- For large transistor θ_2 is smaller than 1 thus $\mu_H = \mu_0$.
- **when L_{Eff} decreases, θ_2 increases and $\theta_2 V_{ds}$ becomes important, lowering the mobility below μ_0**

Mobility reduction with the gate voltage due to the vertical electric field

In a MOS transistor, **the current flows very close to the silicon surface**. As a consequence, the **mobility of current carriers is lower than deep inside the substrate** (typically two to three times lower), **due to various scattering mechanisms**

- A **vertical electric field** exists in MOSFET due to the applied gate voltage, which creates the conduction channel.
- When carriers move within the channel under the effect of horizontal electric field, they feel the effect of gate induced vertical electric field, **pushing carriers towards the gate oxide** as shown in figure 2.

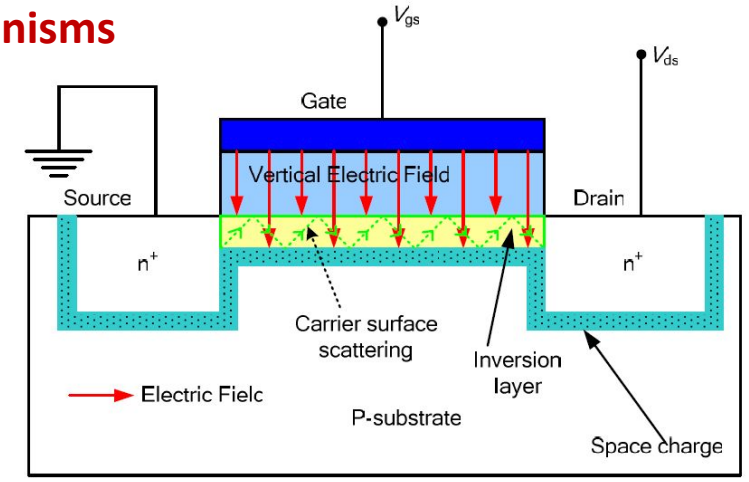


Figure 2: Vertical electric field in a short channel MOSFET and due to that surface scattering

- This provokes carriers to make the collision with the oxide channel interface. The oxide –channel interface is rough and imperfect, thus carriers loses mobility. This **effect is known as surface scattering.**

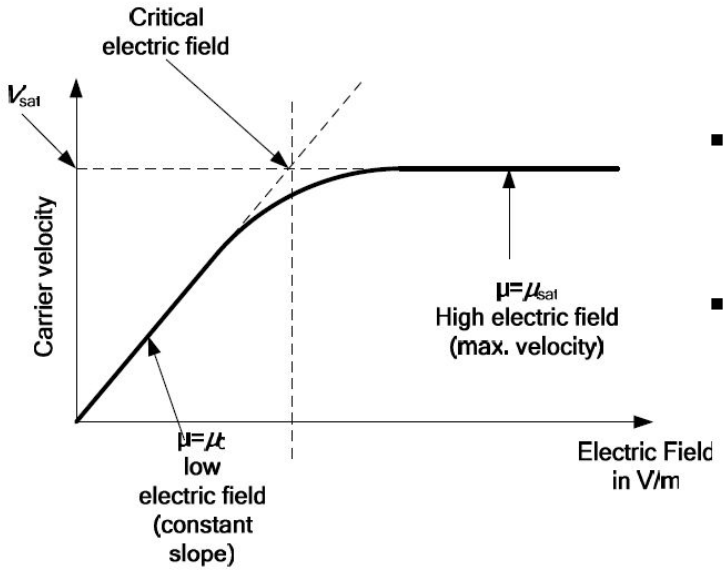
- The **surface mobility depends on how much the electrons interact with the interface**, and therefore, on the vertical electric field which "pushes" the electrons against the interface. We will note **as the surface mobility in absence of such an electric field. The higher the electric field, the lower is the surface mobility.**

Mobility reduction with the gate voltage due to the vertical electric field

The reduction in the surface mobility can be modeled as

$$\mu_H = \frac{\mu_0}{1+\frac{V_{ds}}{L_{eff}E_{crit}}} = \frac{\mu_0}{1+\theta_2 V_{ds}}$$

- $1/L_{eff}E_{crtit}$ is referred as drain bias mobility reduction parameter and in some texts denoted as θ_2. E_{crit} is the electric field shown in figure (a)
- For large transistor θ_2 is smaller than 1 thus $\mu_H = \mu_0$.
- **when L_{Eff} decreases, $\theta_2$increases and $\theta_2 V_{ds}$ becomes important, lowering the mobility below μ_0**



- At **low electric field**, the **electron drift velocity** V_d in the **channel varies linearly** with the electric field intensity.
- However as **the electric field increases above 10^4 V/cm**, the **drift velocity tends to increase more slowly**, and approaches **a saturation value of $V_{d(sat)=}$ 10^7 cm/s** around the electric field =10^5 v/cm at 300k

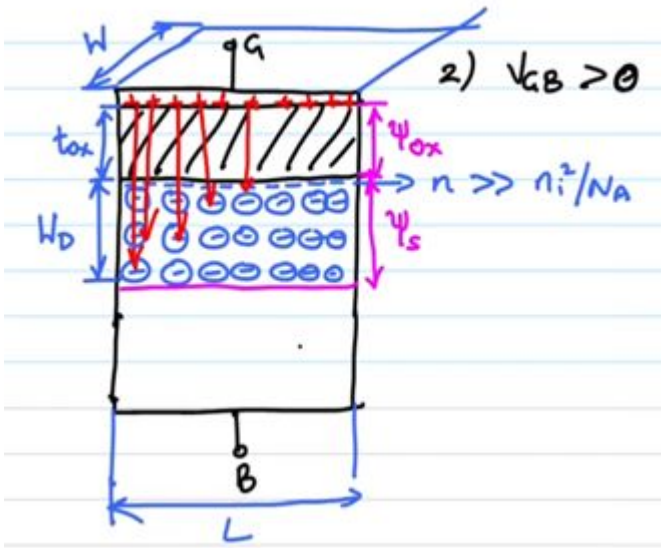
Figure a: Electric field versus Carrier velocity in a solid

Derivation of Threshold Voltage

Gate potential needed to invert the surface is called as the threshold voltage

$$V_{GB} = potential\ across\ the\ oxide\ + potential\ across\ the\ surface$$

$$V_{GB} = \psi ox\ + \psi s \text{ ---------} (1)$$



L – Length of the channel
W – Width of the channel
ψox potential across oxide
ψs potential across surface
t_{ox} thickness of oxide layer
W_D width of the depletion region

- The electric field terminates across the immobile –ive charges and also at the free electrons.
- As the electrons are attracted towards the surface, Beyond a certain point the surface potential is pinned as a small increase in the surface concentration is like 'ln' of that and therefore the ---- stop increasing

$$\ln\left(\frac{n_s}{n_B}\right) = \frac{q\,\psi s}{KT} \text{ ----- (a)}$$

Surface Potential , ***ψs is pinned*** *when the surface concentration is approximately equivalent to the bulk concentration*

$$\text{i.e., } n_s \cong N_A$$

$$\psi s = \frac{KT}{q} \ln\left(\frac{n_s}{n_B}\right) \text{ ------ (2)}$$

Derivation of Threshold Voltage

$As\ per\ law\ of\ mass\ action, np = n_i^2$ ----- (b)

As the surface potential is pinned when ,$n_s \cong N_A$ ----- (c)

$Substituting\ equation\ (c) in\ equation\ (b) N_A\ p = n_i^2$

$$p = \frac{n_i^2}{N_A}$$

$$\psi s = \frac{KT}{q} \ln\left(\frac{N_A}{\frac{n_i^2}{N_A}}\right) = \frac{KT}{q} \ln\left(\frac{N_A}{n_i}\right)^2$$

$$\psi s = 2\frac{KT}{q} \ln\left(\frac{N_A}{n_i}\right) \text{--- (2.a)}$$

Strong inversion **occurs at a voltage equal to twice the** ***Fermi Potential***

$$\psi s = 2\frac{KT}{q} \ln\left(\frac{N_A}{n_i}\right) = 2(\text{Fermi potential}) = 2\emptyset_F$$

Derivation of Threshold Voltage

Gate potential needed to invert the surface is called as the threshold voltage

$$V_{GB} = potential\ across\ the\ oxide\ + potential\ across\ the\ surface$$

$$V_{GB} = \psi ox\ + \psi s \text{ ---------} (1)$$

ψ_{ox} - charge by the capacitance ,as the charge at the surface is negative, we can write it as $Q_D^{'}$

Since we have some inversion charge, we can write it as -$(Q_D^{'} + Q_I^{'})$

$$\psi_{ox} = \frac{-(Q_D^{'} + Q_I^{'})}{C_{ox}^{'}} \text{ -----(3)}$$

$Q_D^{'}$ & $Q_I^{'}$ *represent depletion and inversion charges respectively*

$\boldsymbol{Q_D}$ & $\boldsymbol{Q_I}$ *represent depletion and inversion* ***charges per unit area*** *respectively*

In equation (2), $C_{ox}^{'} = \varepsilon_r\ \varepsilon_o\ \frac{A}{d} = \varepsilon_r\ \varepsilon_o\ \frac{WL}{t_{ox}} ---(2.a)$, where $\frac{\varepsilon_r \varepsilon_o}{t_{ox}} = C_{ox}$

Evaluation of $Q_D^{'}$ = *Total depletion charge t*

atom that is ionized is contributing one Q_D ,

the ***total volume of the depletion region***

$$Q_D^{'} = q\ N_A(WL.W_D) ----- (4)$$

Let the expression of W_D can be written as

$$W_D = \sqrt{\frac{2\varepsilon_{si}|\psi_s|}{q.N_A}} \text{ -----} (4.a)$$

Derivation of Threshold Voltage (Cont.)

Using equation (a) in equation (3),

$$Q_D^{'} = q\,N_A\left(WL.\sqrt{\frac{2\varepsilon_{si}|\psi_s|}{q.N_A}}\right) \text{--- (5}$$

$$Q_D^{'} = \left(\sqrt{2\varepsilon_{si}|\psi_s|q\,N_A}\right) WL \text{-----(6)}$$

Q_D = Charge per unit area = $\left(\sqrt{2\varepsilon_{si}|\psi_s|q\,N_A}\right)$ -- -(7)

From equation (1); $V_{GB} = \psi ox + \psi s$

In equation (1), *Substituting* ψs *and* ψ_{ox}, *where* $\psi_{ox} = \frac{-(Q_D^{'}+Q_I^{'})}{C_{ox}^{'}}$ -----(3)

$$V_{GS} = \left(\psi s - \frac{Q_D^{'}}{C_{ox}}\right) - \frac{Q_I^{'}}{C_{ox}}$$

Gate potential needed to invert the surface is given as

$$\frac{Q_I'}{C_{ox}} = \psi s - \frac{\left(\sqrt{2\varepsilon_{si}|\psi_s|q\,N_A}\right)}{C_{ox}} = Vth \text{ ---(8)}$$

Derivation of Threshold Voltage (Cont.)

$$\frac{Q'_I}{C_{ox}} = \psi s - \frac{\left(\sqrt{2\varepsilon_{si}|\psi_s| q\, N_A}\right)}{C_{ox}} = V_{th} \quad ----(8)$$

$where, \psi s = \; 2\frac{KT}{q} \ln \left(\frac{N_A}{n_i}\right) = \; 2(\text{Fermi potential}) = \; 2\,\emptyset_F$, ***Strong inversion*** **occurs at a voltage equal to twice the *Fermi Potential***

In general the value of V_{GS} where strong inversion occurs is called the *threshold voltage* V_T

V_T is a function of several components, most of which are material constants such as

- the difference in work-function between gate and substrate material
- the oxide thickness
- the Fermi voltage
- the charge of impurities trapped at the surface between channel and gate oxide
- the dosage of ions implanted for threshold adjustment

in case a substrate bias voltage V_{SB} is applied, then $\psi s = \; |-2\emptyset_F + V_{SB}|$

V_{SB} is normally positive for *n*-channel devices

$$In\ general\ , V_{th} = V_{To} + \gamma\left(\sqrt{|-2\emptyset_F + V_{SB}|} - \sqrt{|-2\emptyset_F|}\right) \quad ---(9)$$

empirical parameter V_{T0}, which is the threshold voltage for V_{SB} = 0, the parameter γ (gamma) is called the *body-effect coefficient*, and expresses the impact of changes in V_{SB}.
Observe that the threshold voltage has a **positive** value for a typical **NMOS** device, while it is **negative** for a normal **PMOS** transistor.

Effects of Threshold Voltage

$$In\ general, V_{th} = V_{To} + \gamma\left(\sqrt{|-2\emptyset_F + V_{SB}|} - \sqrt{|-2\emptyset_F|}\right) \quad ---(9)$$

- In equation(9) states that the threshold voltage is only a function of the manufacturing technology and the applied body bias V_{SB}. The threshold can therefore be considered as a constant over all NMOS (PMOS) transistors in a design.
- **As the device dimensions are reduced, this model becomes inaccurate, and the threshold potential becomes a function of *L*, *W*, and V_{DS}.**

- In a **long channel device**, the channel formation is controlled by the gate and the substrate. **The gate voltage** will control essentially **all the space charge induced in the channel region**.
- As the channel length decreases, the charge control of the channel is shared by the four terminals (gate, substrate, source and drain), called charge sharing.
- The total charge below the gate controlled by the gate voltage in a short-channel device is correspondingly less than that controlled by the gate in a long-channel device.
- Consequently, **a lower gate voltage is required to attain threshold in a short-channel device**. Now as the drain voltage increases the reversed biased space charge region at the drain extends further into channel area and the gate will control even less bulk charge.
- i.e., in a short channel device the n+ type source and the drain induce a significant amount of the depletion charge that cannot be neglected. The depletion regions of the source and the drain are very close to one another.

Effects of Threshold Voltage

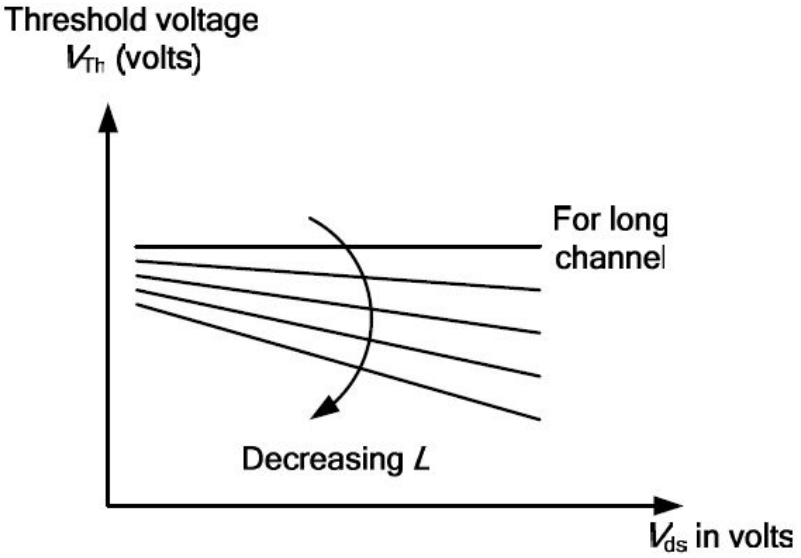


Figure 1: Effective threshold voltage as a function of *V*ds taking *L* as parameter for an n-channel MOSFET

- The deeper depletion region is accompanied by larger surface potential, which makes the channel more attractive for electrons. Thus, the device can conduct more current.
- This effect can be considered as the reduction of V_{Th} as drain current is the function of (V_{gs}-V_{Th}).
- Increase in V_{ds} and reduction of channel length will decrease the effective threshold voltage as shown in figure 1.
- Curve representing the reduction of V_{Th} with decreasing effective channel length is known as V_{Th} roll off. This adverse roll-off effect is perhaps the most daunting roadblock in future MOSFET design . The minimum acceptable channel length is primarily determined by this roll-off.

Down Scaling of CMOS

CMOS have been scaled to achieve

- **Higher density**
- **Higher performance**
- **Lower power consumption**

Disadvantage :

- Electric field within the gate oxide grow lager
- Short channel effcets

Moore's Law and MOSFET Scaling

Moore's Law

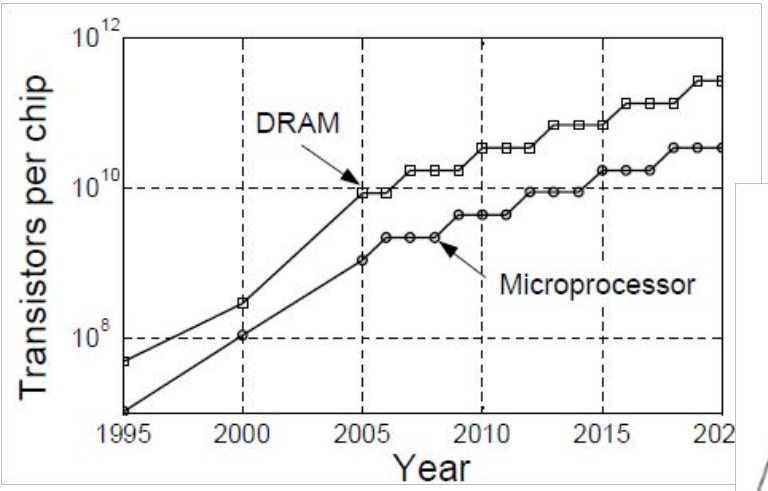


Figure 1 Evolution of the number of transistors per chip (Moore's law) predicted by the ITRS for DRAMs and high-performance microprocessors

Dennard's Scaling Rule

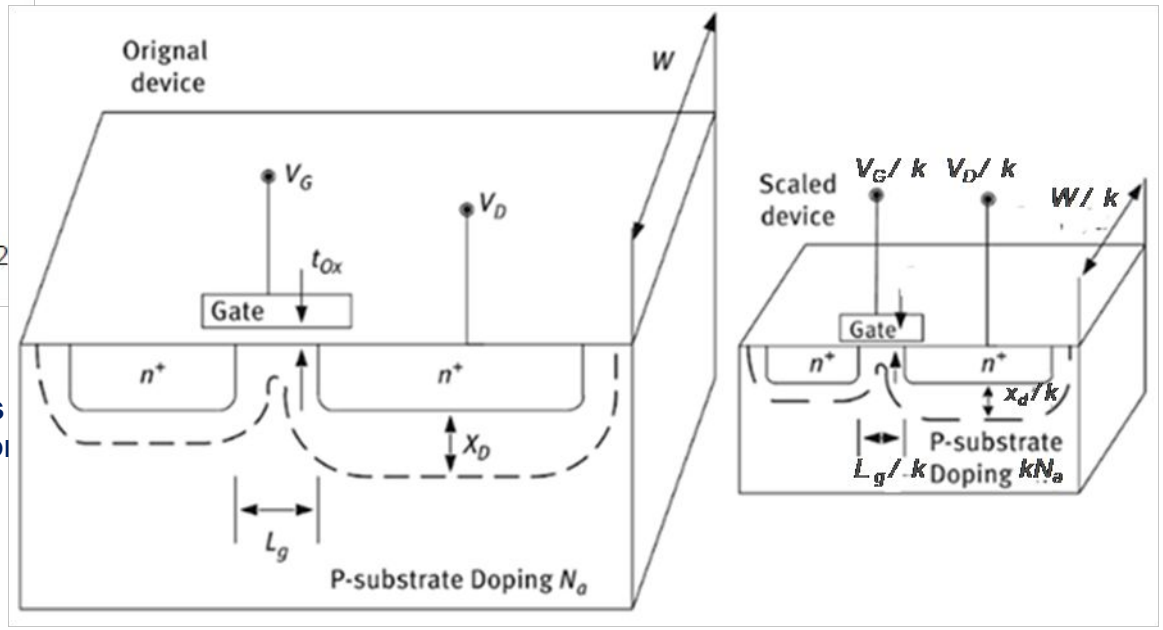


Figure 2 MOSFET scaling rule [2]

Reducing the critical dimensions while keeping the electrical field constant yields

- **higher speed**
- **reduced power** consumption of a digital MOS circuit

Types of Scaling

1. constant field scaling
2. constant voltage scaling

1. Constant field scaling

- **Dennard et al**. presented their pioneering research work on the scaling of MOSFET devices at the International Electron Device Meeting (IEDM) 1972 and published a comprehensive paper on the scaling of MOS transistors in 1974 , from which the **"constant field scaling" theory has emerged**.
- The basic principle which they employ is that **in order to increase the performance of a MOSFET** we must
  - **reduce linearly the size of the transistor**, together with the supply voltage
  - and increase the doping concentration in a way which keeps the electric field in the device constant - hence the name "constant field scaling" .

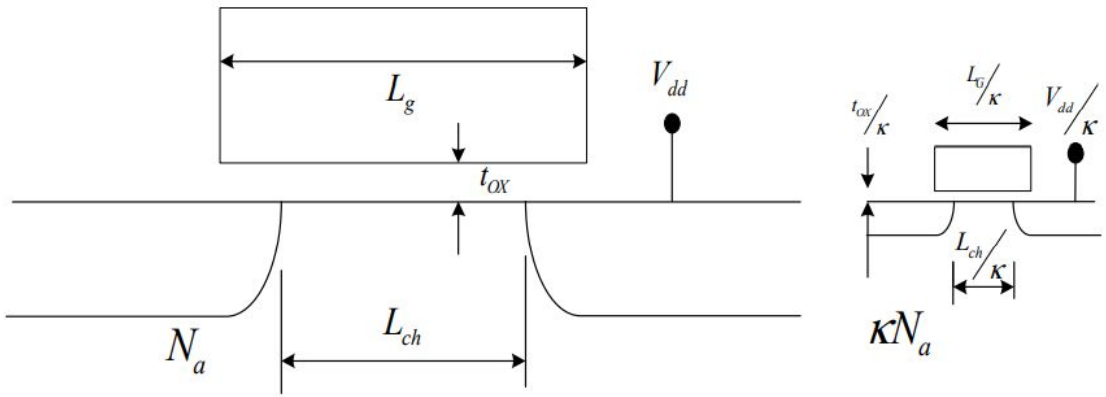


Figure : Illustration of MOSFET miniaturisation. The sketch on the right hand is the scaled device according to the constant field rule.

Constant field scaling yields the largest reduction in the power-delay product of a single transistor. However, it requires a reduction in the power supply voltage as one decreases the minimum feature size.

1. Constant field scaling

- Constant field scaling yields the **largest reduction in the power-delay product of a single transistor.**
- However, it requires **a reduction in the power supply voltage** as one decreases the minimum feature size.

2. Constant voltage scaling

Constant voltage scaling does not have this problem and is therefore the **preferred scaling method since it provides voltage compatibility with older circuit technologies**.

- The **disadvantage of constant voltage scaling** is that the **electric field increases as the minimum feature length is reduced.**
- This leads to **velocity saturation, mobility degradation, increased leakage currents and lower breakdown voltages**.

Types of Scaling

After scaling, the different MOSFET parameters will be converted as given by table below:
Before Scaling After Constant Field Scaling After Constant Voltage Scaling

Before Scaling	*After Constant Field Scaling*	*After Constant Voltage Scaling*
L	$L' = L/s$	$L' = L/s$
W	$W' = W/s$	$W' = W/s$
t	$t'_{ox} = t_{ox}/s$	$t'_{ox} = t_{ox}/s$
xi	$x'_i = x_i/s$	$x'_i = x_i/s$
VDD	$V'_{DD} = V_{DD}/s$	$V'_{DD} = V_{DD}$
VTh	$V'_{Th} = V_{Th}/s$	$V'_{Th} = V_{Th}$
Na or *Nd*	$N'_a = N_a * s$ or $N'_d = N_d * s$	$N'_a = N_a * s^2$ or $N'_d = N_d * s^2$
Cox	$C'_{ox} = C_{ox} * s$	$C'_{ox} = C_{ox} * s$
IDS	$I'_{DS} = I_{DS}/s$	$I'_{DS} = I_{DS} * s$
PD	$P'_D = P_D/s^2$	$P'_D = P_D * s$

Where **s** = scaling parameter of MOS

Components of Leakage Power due to scaling of MOSFET

$$P_{dynamic} = C_L \cdot V_{DD}^2 \cdot f \qquad P_{static} = I_{leak} \cdot V_{DD}$$

- Supply voltage has been scaled down - to keep power consumption under control
- The threshold voltage scaling results in substantial increase of the leakage current

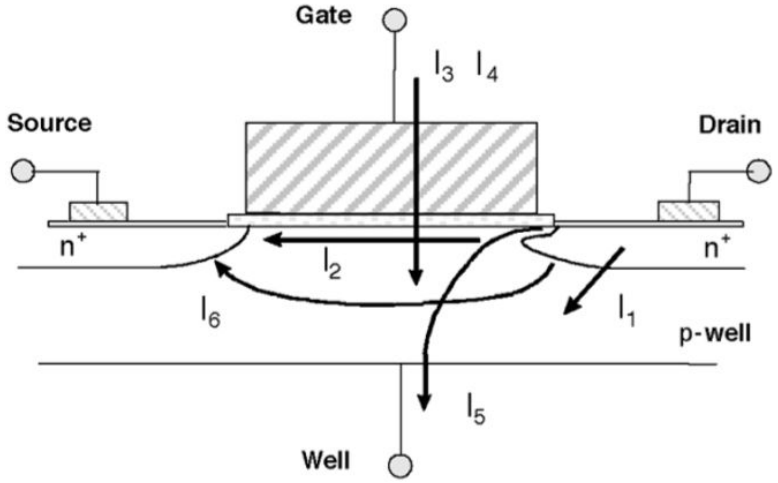


Figure b Components of leakage power due to short channel effects

- pn Junction Reverse-Bias Current (I_1)
- Subthreshold Leakage (I_2)
- Tunneling into and Through Gate Oxide(I_3)
- Injection of Hot Carriers from Substrate to Gate Oxide (I_4)
- Gate-Induced Drain Leakage (I_5)
- Punchthrough(I_6)

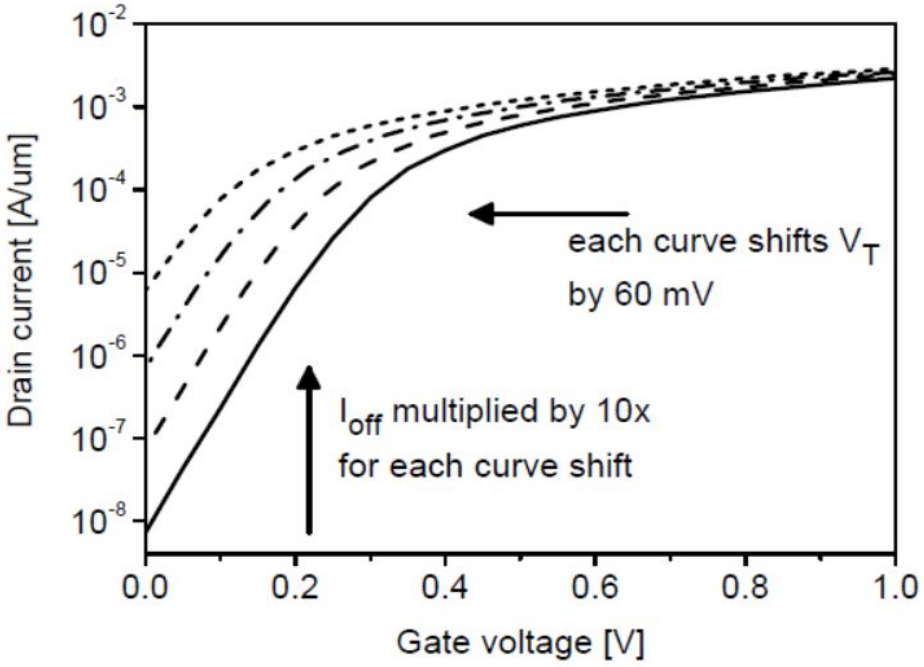


Figure c Exponential increase in the off-current of MOSFET due to the downscaling of threshold voltage reduction

Short Channel effects

The sources of leakage in short channel devices are

- ✔Reverse-biased diode leakage
- ✔Gate-oxide tunneling
- ✔Gate induced drain leakage (GIDL)
- ✔subthreshold leakage
- ✔Drain-induced barrier lowering(DIBL)

Short Channel effects

The sources of leakage in short channel devices are

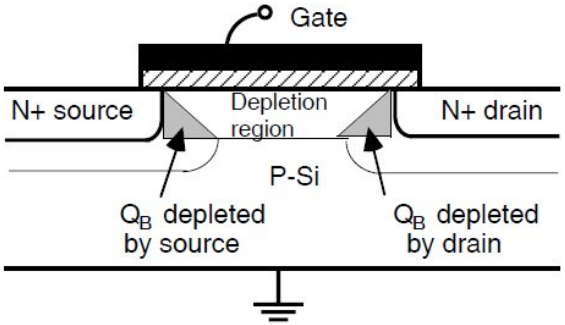
- ✔Reverse-biased diode leakage
- ✔Gate-oxide tunneling
- ✔Gate induced drain leakage (GIDL)
- ✔subthreshold leakage
- ✔Drain-induced barrier lowering(DIBL)

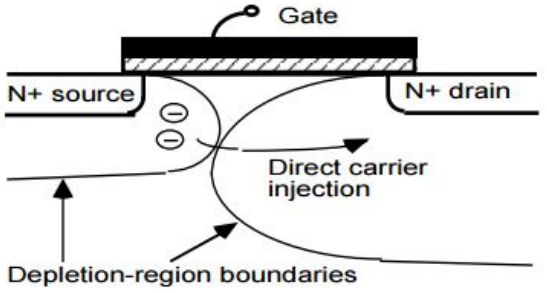
DIBL and Punch Through

One of the major challenges in transistor scaling are the ***"Short Channel Effects"*** with channel length (L_G) < 100nm.

DIBL occurs when **high drain voltage** is applied to short-channel devices, the depletion region of the drain interact with the source depletion region as the **source potential barrier height is lowered resulting in injection of the carrier into to the channel from the source** due to the reduced threshold voltage as it is not dependent of gate voltage.

- **Punch Through** – For short Channel, the depletion region from the drain can reach the source side and reduces the barrier for electron injection
- **Drain Induced Barrier Lowering(DIBL)** - the barrier for electron injection from source to drain decreases. This is known as drain induced barrier lowering (DIBL)





Reverse-biased diode leakage

- The current due to the reverse-biased drain-substrate and source-substrate junction is relatively small, but there is an exponential increase in current when large forward biased is applied to the substrate region.
- To limit the effects due to scaling, the p-region and n-region are heavily doped and results in BTBT leakage current and dominates the reverse biased diode leakage.
- It is evident from Figure 1.4, that the tunneling of electrons from the p-region valence band to the n-region conduction band causes a flow of current through the junction as a result of the high electric field across the junction.

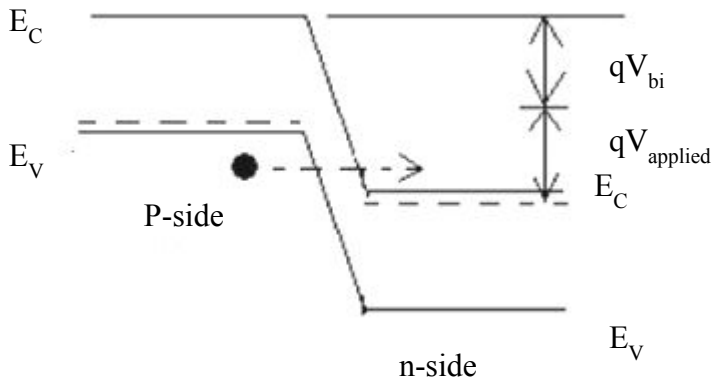


Figure 1.4: BTBT due to high electric field across reverse-biased junction [5]

It occurs when the total voltage drop across the junction, applied reverse bias ($V_{applied}$) and the built-in potential (V_{bi}) is larger than the band-gap

Gate-Oxide Tunneling

- With scaling of transistor to nanometer dimensions the thickness of the oxide layer is scaled down and it leads to increase in electric field across the oxide layer
- Resulting in a **leakage current from gate to substrate regio**n through the thin oxide layer due to the negative gate bias and also leakage current from substrate region to gate due to positive gate bias.

Gate Induced Drain Leakage

A high field effect in MOS transistor drain junction results in GIDL as the field is crowded near the surface due to the narrowing of the depletion layer at the silicon surface when large negative-biased is applied to the gate .

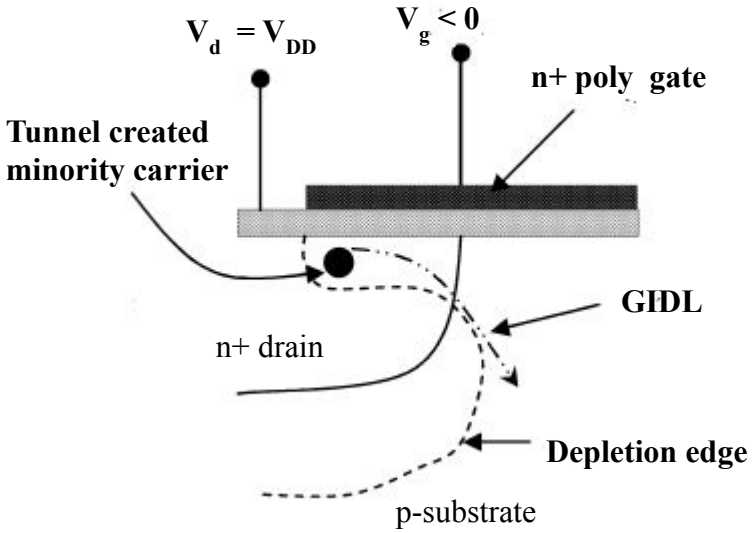


Figure 1.7: n+ region inverted depleted due to the high negative gate bias

As shown in Figure 1.7 the n+ drain region under the gate can be depleted and even inverted. GIDL can be minimized by increasing the drain doping concentration and also by abrupt doping as it gives lower series resistance required for high transistor drive currents[6].

Subthreshold Leakage

- As shown in Figure 1.8, when $V_{GS} < V_{TH}$, weak inversion current flows from the source region to drain is defined as **subthreshold current and it is can be estimated using subthreshold swing (*SS*).**
- A derivative of the log of drain current versus gate voltage is defined as subthreshold slope and its inverse is defined the subthreshold swing(*SS*) and expressed as Equation (1.5),

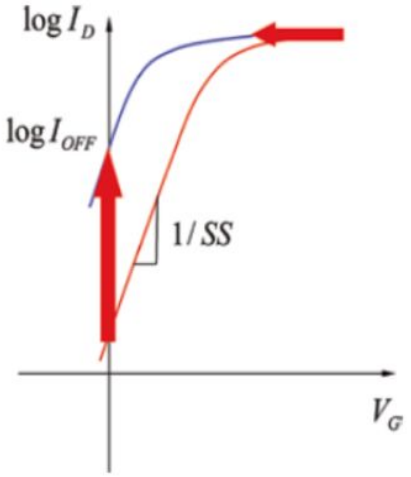


Figure 1.8: Exponential increase in the OFF current (I_{OFF}) of MOSFET due to downscaling of Threshold voltage reduction

$$SS = \frac{dV_G}{d(\log I_D)} = \frac{K_B T}{q \log_e}\left(1 + \frac{C_{ch}}{C_{ox}}\right) \qquad (1.5)$$

- At room temperature SS value equals 60 mV/decade. Where K_B is the Boltzmann constant, T is the room temperature; C_{ch} and C_{ox} are the channel and oxide capacitance per unit area respectively.

- The presence of the subthreshold slope implicates **that for every 60mV voltage reduction, weak inversion current decreases with a factor of 10.**
- So if the current has to be 5 orders of magnitude below the on-state currents, the threshold voltage has to be at least 5 x 60mV = 0.3V.
- The limit in subthreshold slope of MOSFET due to the increase in OFF-state current provides the quest for energy efficiency in computing

CMOS Energy Efficiency

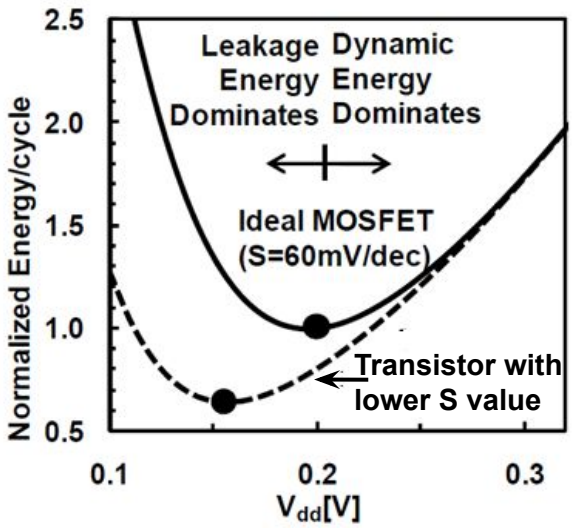


Figure 7 Dependence of MOSFET energy consumption on the power supply voltage for a given switching speed [6]

- Dynamic energy reduces quadratically as the supply voltage is scaled down. To maintain a certain switching speed, the threshold voltage of the MOSFET must be scaled down as well, which increases the leakage energy.
- Therefore there exists an optimal V_{dd} *that minimizes the energy dissipation.*
- Transistor designs with lower *S value reduce the leakage energy; they therefore* improve the energy efficiency.

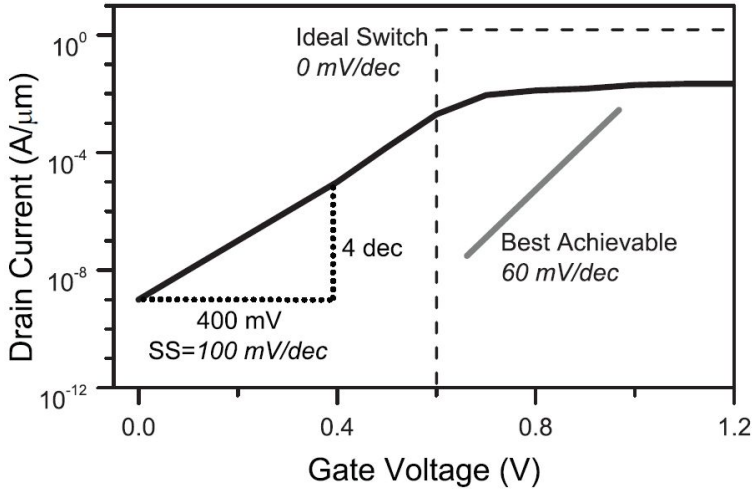


Figure 8 Transfer characteristics of a typical NMOS and comparison with an ideal switch (shown in dashed line) [7]

To overcome the CMOS energy efficiency limit, alternative transistor designs which can achieve a steeper sub-threshold swing (*i.e. more abrupt* transition between on- and off-states) have been proposed.

Unified Current Model

- Useful to combine all effects into one equation
- Voltage values determine the governing equations

$$I_{DS} = \begin{cases} 0 & |V_{GS}| < |V_{TH}| \\ k' \frac{W}{L} \left((V_{GS} - V_{TH}) V_{min} - \frac{V_{min}^2}{2} \right) (1 + \lambda V_{DS}) & |V_{GS}| > |V_{TH}| \end{cases}$$

Where

$$V_{min} = min(V_{DS}, (V_{GS} - V_{TH}), V_{DS-SAT}) \ldots NMOS$$
$$V_{min} = max(V_{DS}, (V_{GS} - V_{TH}), V_{DS-SAT}) \ldots PMOS$$
$$V_{TH} = V_{TH0} + \gamma(\sqrt{|V_{SB} + \Psi_S|} - \sqrt{\Psi_S})$$

$(V_{TH0}, k', V_{DSAT}, \gamma, \lambda)$ - Technology parameters

Dynamic behavior of a MOSFET transistor

- The dynamic response of a MOSFET transistor is a **sole function of the time it takes to (dis)charge the parasitic capacitances** that are intrinsic to the device, and the extra capacitance introduced by the interconnecting lines.
- A profound understanding of the nature and the behavior of these intrinsic capacitances is essential for the designer of high-quality digital integrated circuits.
- **They originate from three sources:**
  - the basic MOS structure
  - the channel charge
  - and the depletion regions of the reverse-biased *pn*-junctions of drain and source.
- Aside from the MOS structure capacitances, all capacitors are nonlinear and vary with the applied voltage, which makes their analysis hard.

Gate Capacitance

- The gate of the MOS transistor is isolated from the conducting channel by the gate oxide that has a capacitance per unit area equal to $C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$.
- We learned earlier that from a *IV* perspective it is useful to have C_{ox} as large as possible, or to keep the oxide thickness (t_{ox}) very thin.
- The total value of this capacitance is called the *gate capacitance* C_g

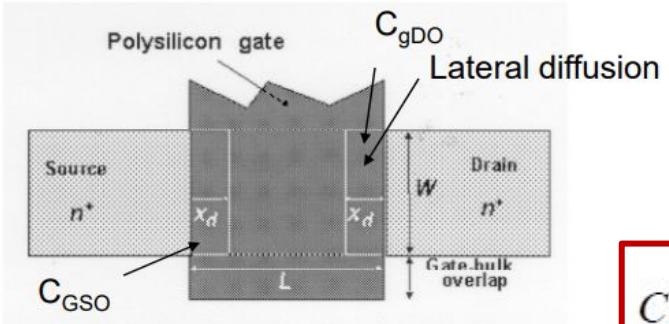
***Gate capacitance C_g is* decomposed into two elements**, each with a different behavior.

- one part of C_g contributes to the channel charge
- another part is C_g solely due to the topological structure of the transistor.

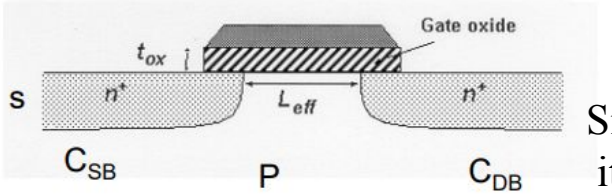
Gate Capacitance (Cont.)

- Ideally, the source and drain diffusion should end right at the edge of the gate oxide. In reality, **both source and drain tend to extend somewhat below the oxide by an amount x_d**, called the ***lateral diffusion***.
- Hence, the effective channel of the transistor L becomes shorter than the drawn length L_d (or the length the transistor was originally designed for) by a factor of $\Delta L = 2\,x_d$.
- It also gives rise to a parasitic capacitance between gate and source (drain) that is called the *overlap capacitance.* This capacitance is strictly linear and has a fixed value

(a) Top view



(b) Cross-section



$$C_{GSO} = C_{GDO} = C_{ox} x_d W = C_o W$$

Since x_d is a technology-determined parameter, it is customary to combine it with the oxide capacitance to yield the overlap capacitance per unit transistor width C_o (more specifically, C_{gso} and C_{gdo}).

$$C_{gate} = \frac{\varepsilon_{ox}}{t_{ox}} WL$$

Can be decomposed into a number of elements each with a different behavior

Channel Capacitance

- Perhaps the most significant MOS parasitic circuit element, the **gate-to-channel capacitance** C_{GC} **varies** in both magnitude and in its division into three components depending upon the operation region and terminal voltages.
  - C_{GCS} (gate-to-source capacitance)
  - C_{GCD} (gate-to-drain capacitance)
  - C_{GCB} (gate-to-body capacitances)

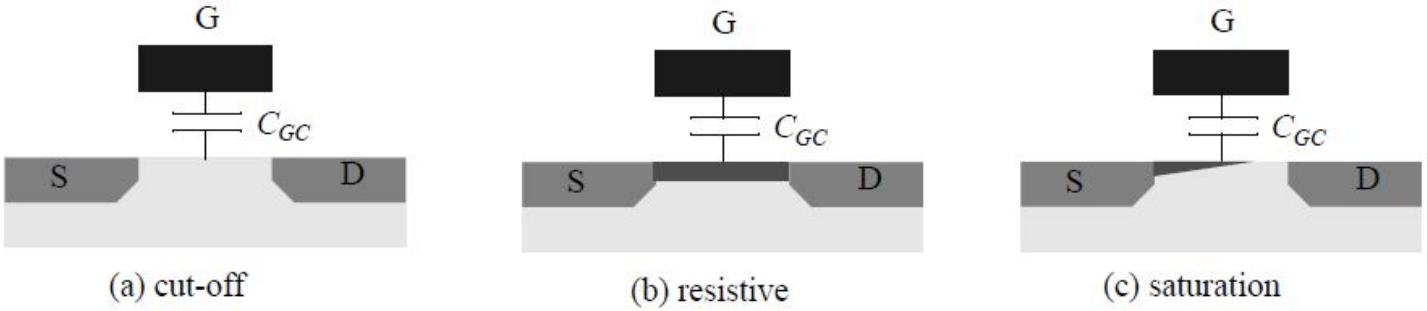


Figure 1: The gate-to-channel capacitance and how the operation region influences is distribution over the three other device terminals

Channel Capacitance: C_{gs}, C_{gd}, and C_{gb}

Cut-Off: no channel, total capacitance = $C_{ox}WL_{eff}$
appears between gate and bulk

Triode Region: Inversion layer - acts as conductor $\therefore C_{gb} = 0$

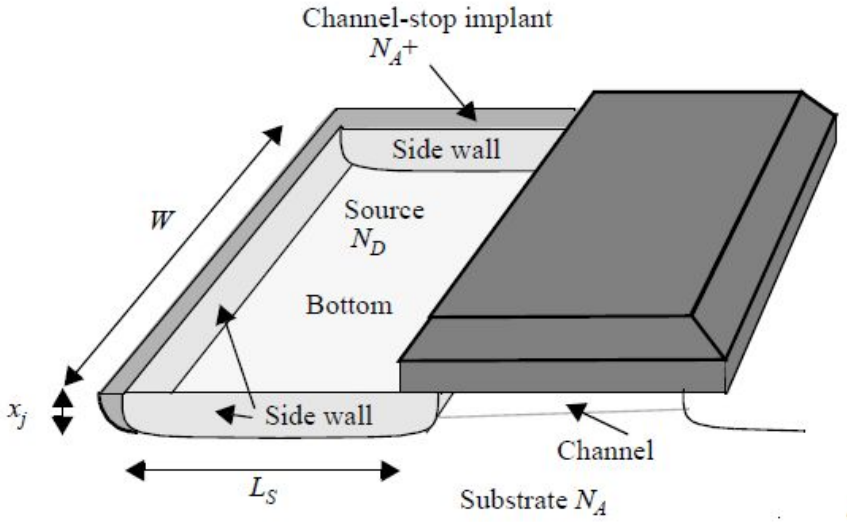
Symmetry dictates $C_{gs} \approx C_{gd} \approx \frac{C_{ox}WL_{eff}}{2}$

Saturation: Pinch off, $\therefore C_{gd} \approx 0, C_{gb} = 0$

C_{gs} averages $(2/3)C_{ox}WL_{eff}$

Junction Capacitance

- A final capacitive component is **contributed by the reverse-biased source-body and drain body *pn*-junctions.**
- The **depletion-region capacitance is nonlinear** and decreases when the reverse bias is raised as discussed earlier.
- To understand the components of the junction capacitance (often called the *diffusion capacitance*), we must look at the source (drain) region and its surroundings.
- The detailed picture, shown in Figure 2, shows that the junction consists of two components: *bottom-plate* junction and *side-wall* junction



The *bottom-plate* junction, which is formed by the source region (with doping N_D) and the substrate with doping N_A.

Figure 2: Diffusion (Junction Capacitance)

Junction Capacitance

- **Bottom plate**

$$C_{bottom} = C_j W L_s,$$

- **Side-wall junctions** - formed by source (N_D) and P^+ channel stop (N_A^+)

- graded junction (m=1/3)

$$C_{sw} = C'_{jsw} x_j (w + 2L_s)$$

$$= C_{jsw}(W + 2L_s)$$

$$C_{jsw} = C'_{jsw} x_j \quad , \qquad x_j = \text{junction depth}$$

- $C_{diff} = C_{bottom} + C_{sw}$

$$= C_j * \text{Area} + C_{jsw} \times \text{Perimeter}$$

$$= C_j L_s W + C_{isw} (2L_s + W)$$

Capacitive Device Model

All the above contributions can be combined in a single capacitive model for the MOS transistor, which is shown Figure 3. Its components are readily identified on the basis of the preceding discussions.

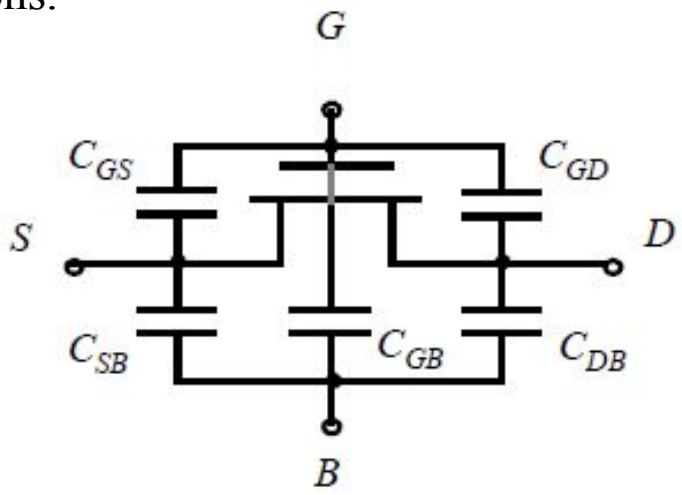


Figure 3: MOSFET capacitance model.

$$C_{GS} = C_{GCS} + C_{GSO};\ C_{GD} = C_{GCD} + C_{GDO};\ C_{GB} = C_{GCB}$$

$$C_{SB} = C_{Sdiff};\ C_{DB} = C_{Ddiff}$$

It is essential for the designers of high-performance and low-energy circuits to be very familiar with this model as well as to have an intuitive feeling of the relative values of its components.

MOS – Parasitic and Contact Resistance

Source-Drain Resistance

The performance of a CMOS circuit may further be affected by another set of parasitic elements, being the resistances in series with the drain and source regions, as shown in Figure 4a.

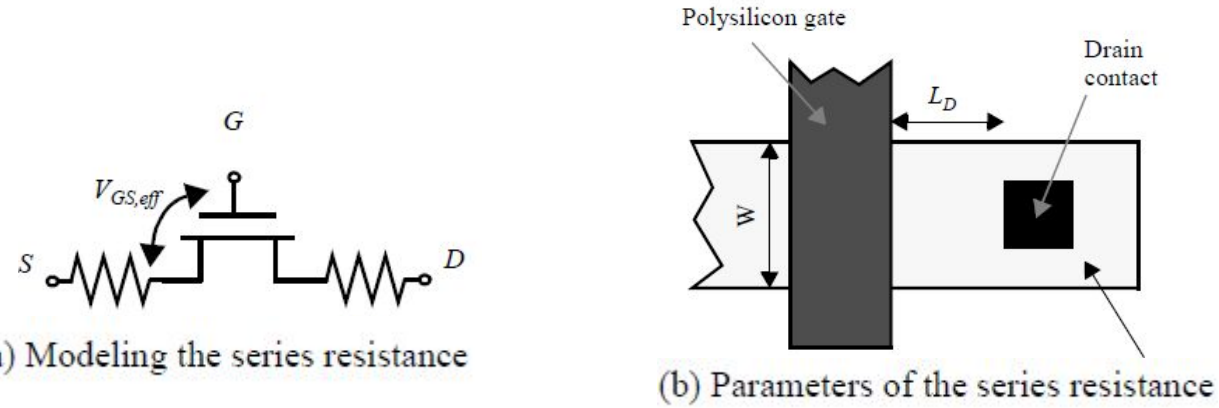
This effect become more pronounced when transistors are scaled down, as this leads to shallower junctions and smaller contact openings become smaller.

The resistance of the drain (source) region can be expressed as

$$R_{S,D} = \frac{L_{S,D}}{W} R_{\square} + R_C$$

with R_C the contact resistance, W the width of the transistor, and $L_{S,D}$ the length of the source or drain region (Figure 4b). $R_{\square}$ is the *sheet resistance* per square of the drain source diffusion, and ranges from 20 to 100 Ω/❑.

Observe that the resistance of a square of material is constant, independent of its size



(a) Modeling the series resistance

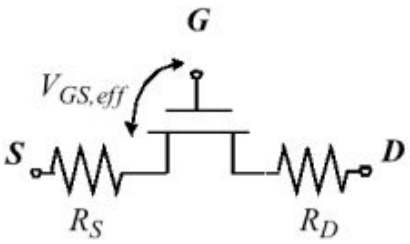
(b) Parameters of the series resistance

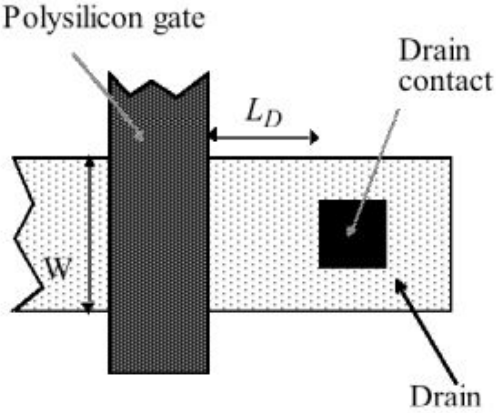
Figure 4 Series drain and source resistance.

Source-Drain Resistance

- The series resistance causes a deterioration in the device performance, as it reduces the drain current for a given control voltage.
- Keeping **its value as small as possible is thus an important design goal for both the device and the circuit engineer.**
- One option, popular in most contemporary processes, is to **cover the drain and source regions with a low-resistivity material such as titanium or tungsten. This process is called *silicidation*** and effectively reduces the sheet resistance to values in the range from 1 to 4 $\Omega/\square$.

Parasitic Resistances





$$R_S = (L_S/W)R_{\square} + R_C$$

$$R_D = (L_D/W)R_{\square} + R_C$$

R_C: contact resistance

$R_{\square}$: sheet resistance per square of drain-source diffusion