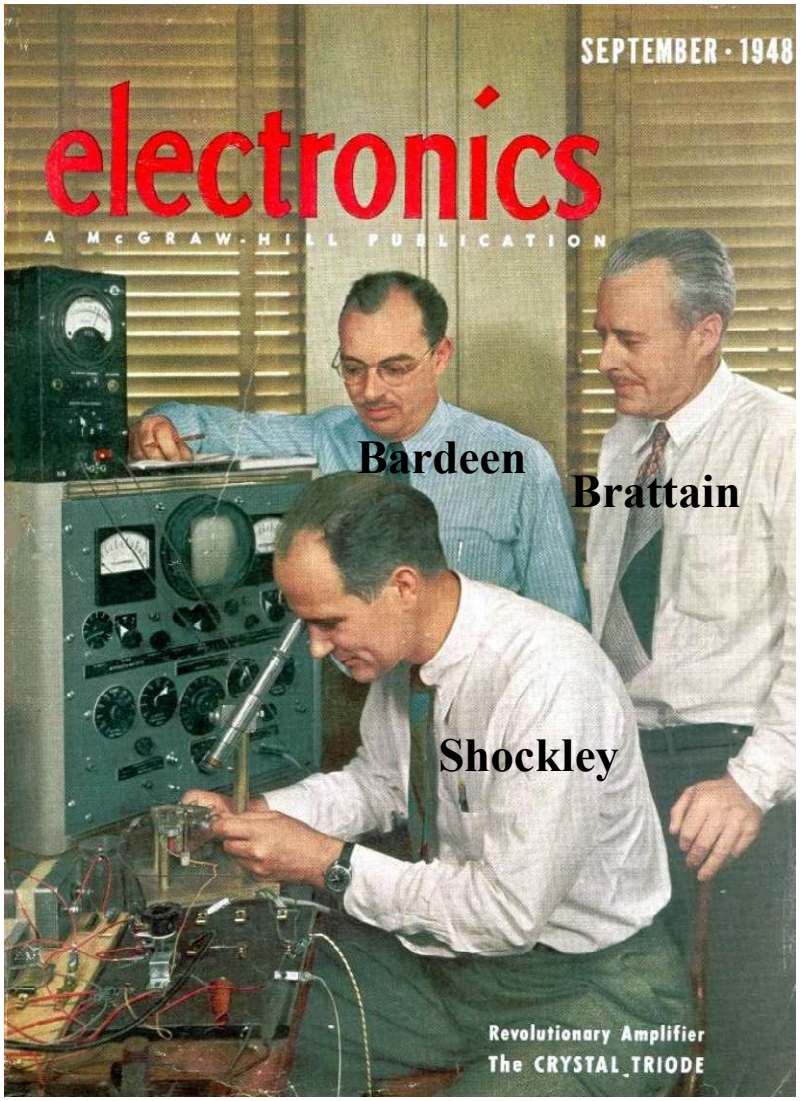
Chapter 6

MOS Transistor

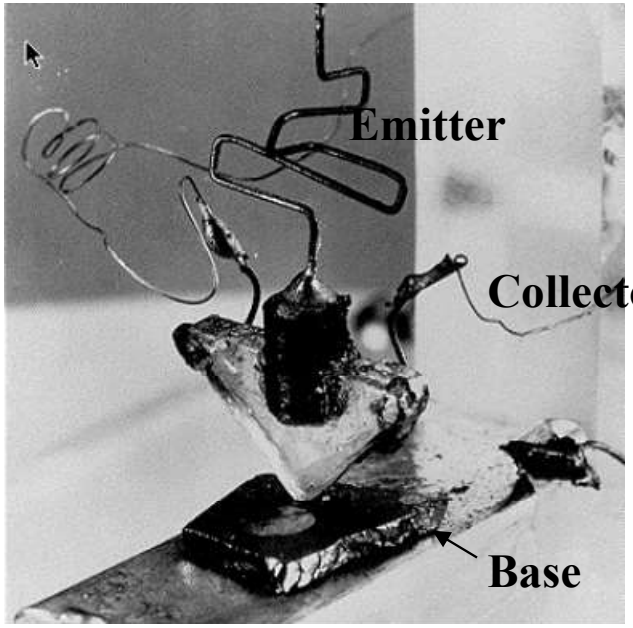
OBJECTIVES

1. Provides a comprehensive introduction to the modern MOSFETS; surface mobility, body effect, a simple I-V theory, and a more complete theory both for long- and short-channel MOSFETs.
2. Introduces the general concept of CMOS circuit speed and power consumption, voltage gain, high-frequency operation, and topics important to analog circuit design (voltage gain, noise).
3. Introduces DRAM, SRAM, and flash nonvolatile memory

Bipolar (Junction) Transistor was invented in 1948 by Bardeen, Brattain and Shockley at the Bell Lab.



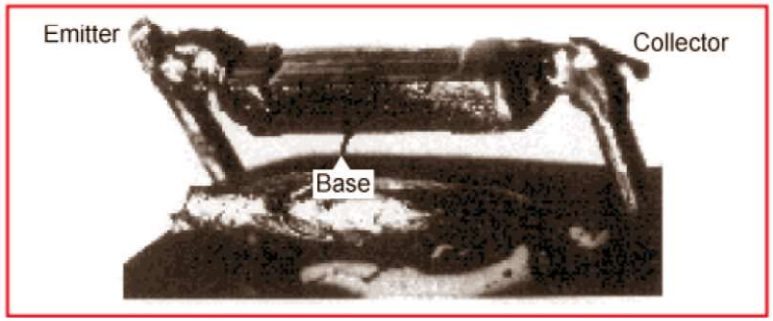
Spring

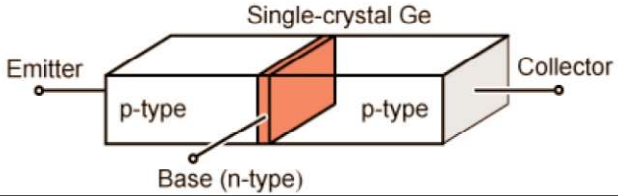


The first point-contact transistor made use of the semiconductor germanium. Paper clips and razor blades were used to make the device.

The First Junction Transistor

First transistor with diffused pn junctions by William Shockley
Bell Laboratories, Murray Hill, New Jersey (1949)

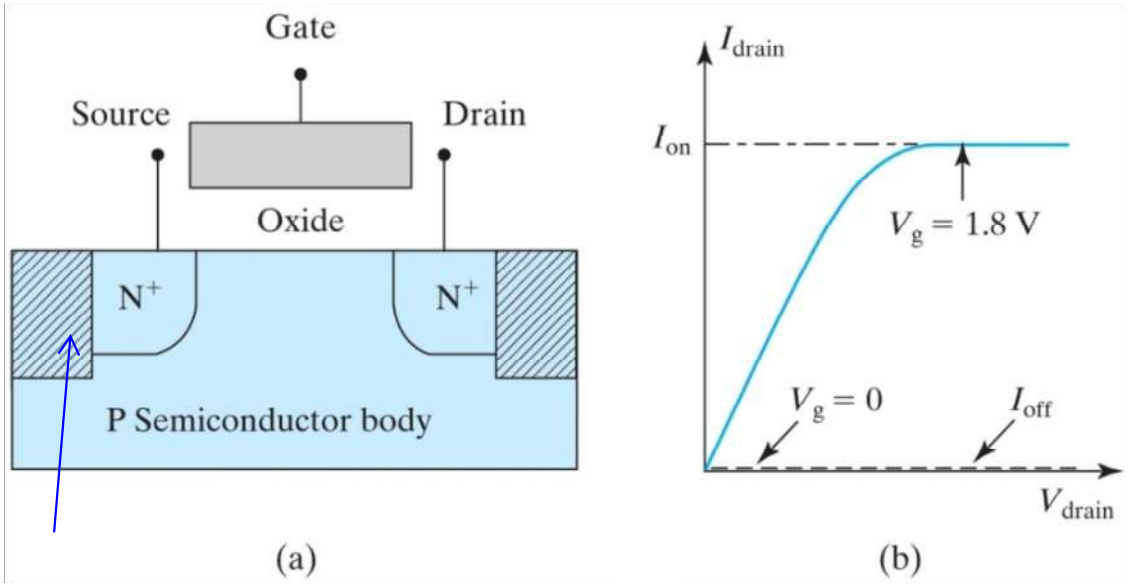




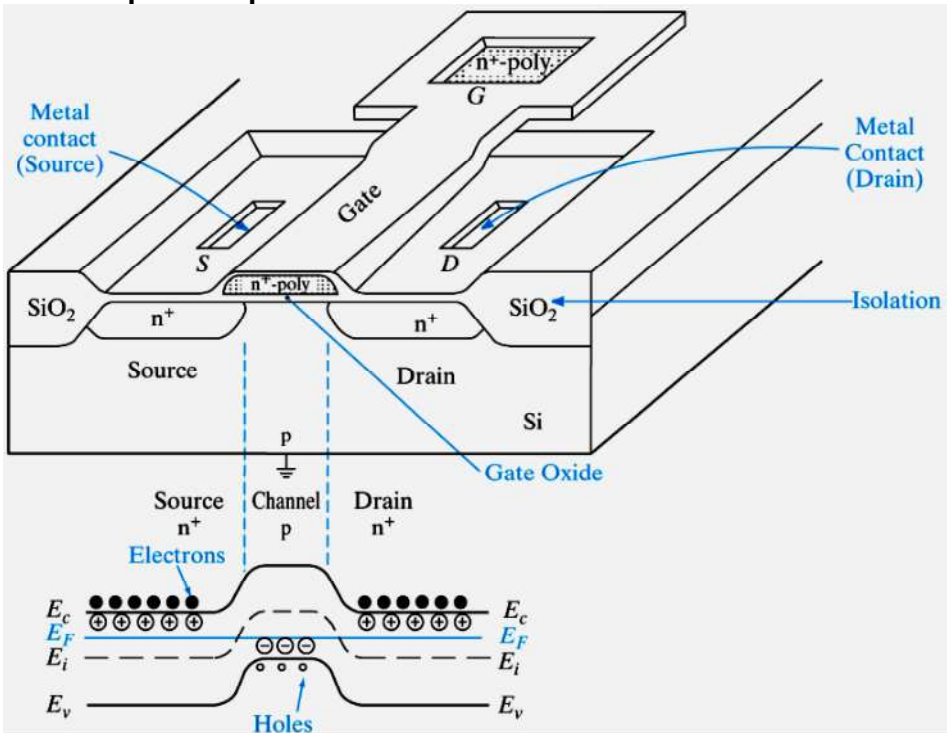
Introduction to the MOSFET

MOSFET: metal-oxide-semiconductor field effect transistor

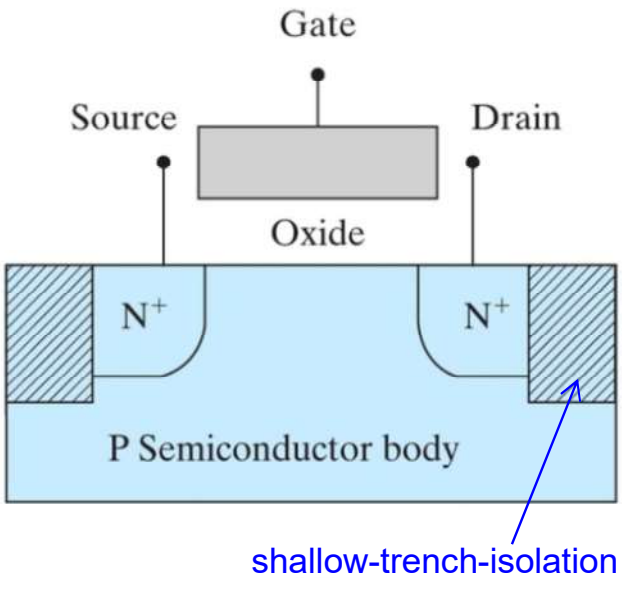
1) the most prevalent semiconductor device in ICs
2) basic building block of digital, analog, and memory circuits
3) small, inexpensive, and dense circuits such as giga-bit (Gb) memory chips
4) low power and high speed chips for GHz computer processor and RF cellular phones

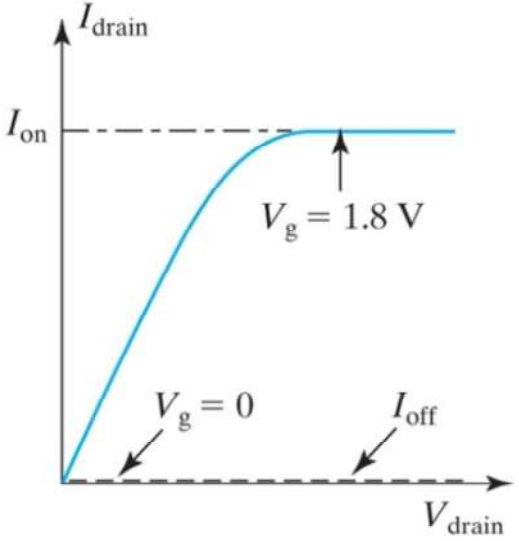


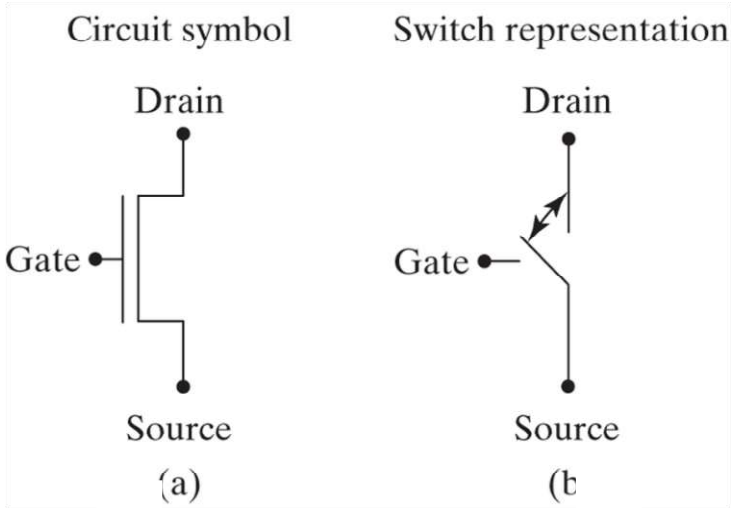
shallow-trench-isolation oxide region



The MOSFET is a device that presents a high input resistance to the signal source, drawing little input power, and a low resistance to the output circuit.







The source and the drain supplies the electrons or holed to the transistor and drains them away, respectively.

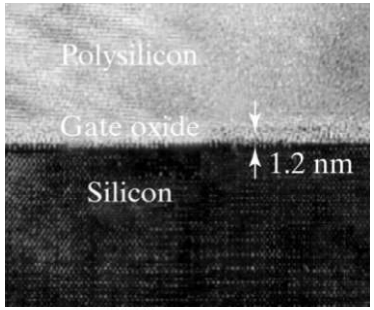
The gate turns the transistor (inversion layer) on and off with an electric field through the oxide. ➡ Field Effect Transistor or FET

Depending on the gate voltage, the MOSFET can be off (conducting only a very small off-state leakage current, I_{off}) or on (conducting a large on-state current, I_{on})

At the most basic level, a MOSFET may be thought of as an on-off switch.

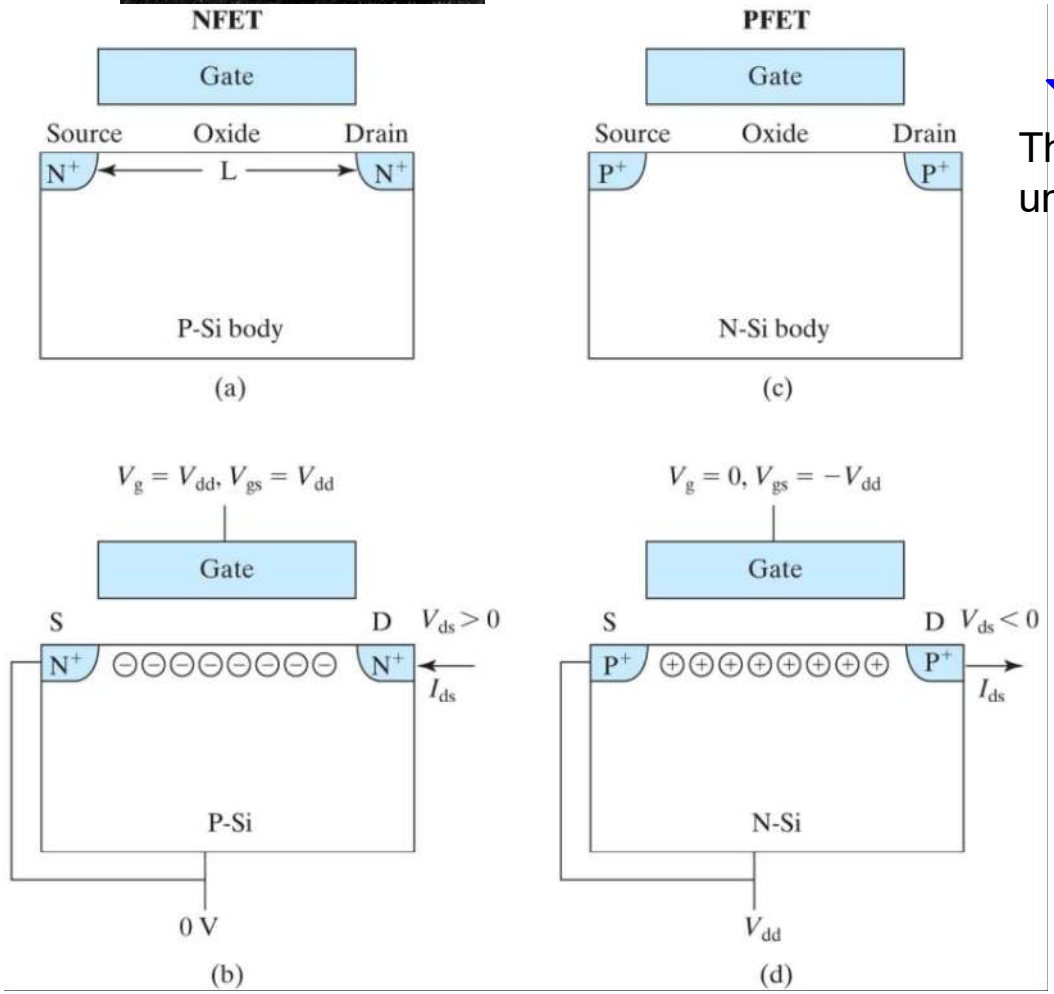
➡ The first MOSFET was demonstrated in 1960 by Kahng and Atalla due to technology gap.

Complementary MOS (CMOS) Technology



Gate oxides as thin as 1.2 nm can be manufactured reproducibly. Individual Si atoms are visible in the substrate and in the polycrystalline gate. (From [3]. © 1999 IEEE.)

In both cases, V_g and V_d swing between 0 V and V_{dd}, the power supply voltage.



The PN junctions are always reverse-biased or unbiased and do not conduct forward diode current.

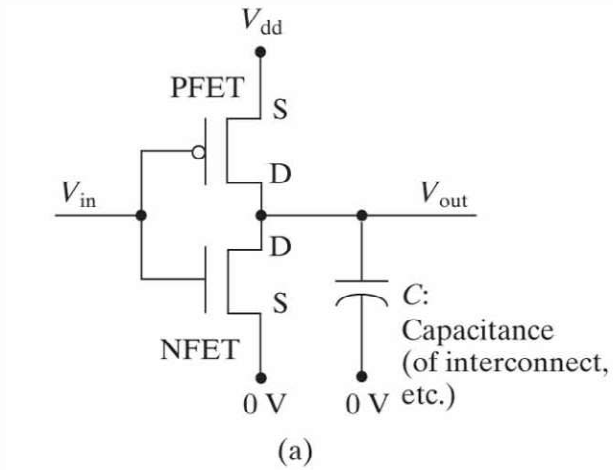
When $V_g = V_{dd}$, the N-MOSFET is on and P-MOSFET is off.

When $V_g = 0$, the P-MOSFET is on and the N-MOSFET is off.

The complementary nature of N-MOSFETs and P-MOSFETs makes it possible to design low-power circuits called **CMOS** or **complementary MOS** circuits

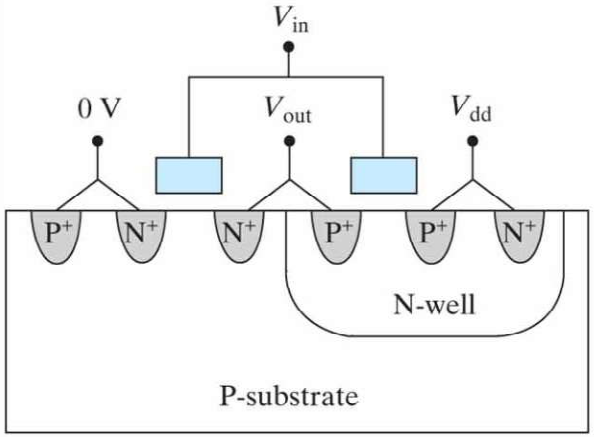
Schematic drawing of an N-channel MOSFET in the off state (a) and the on state (b). (c) and (d) show a P-channel MOSFET in the off and the on states.

CMOS Inverter



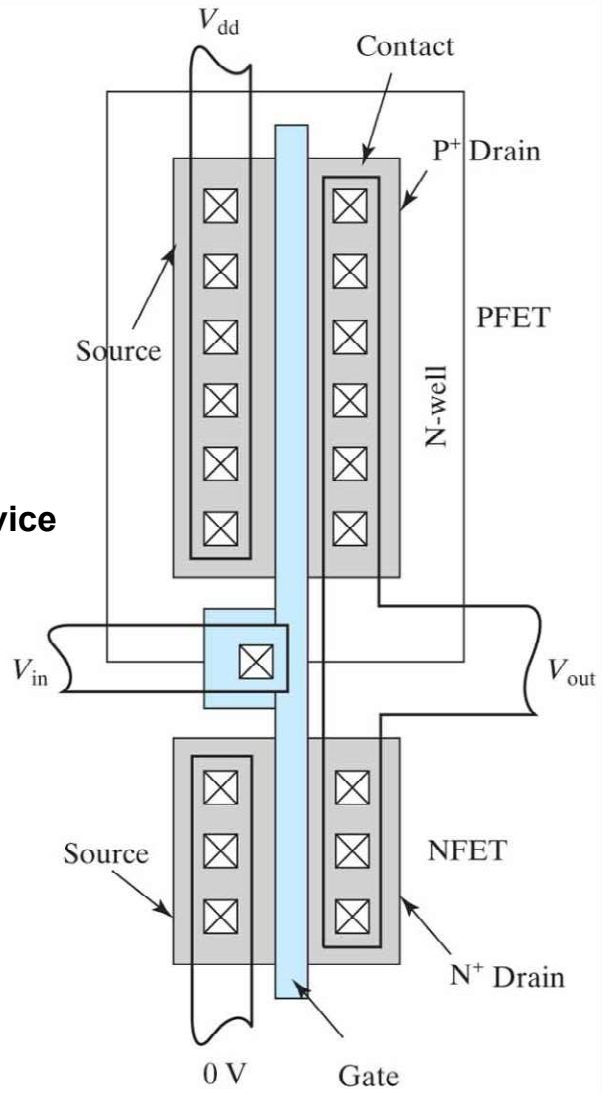
(a)

(a) A CMOS inverter consists of a PFET pull-up device and an NFET pull-down device.



(b)

(b) Integration of NFET and PFET on the same chip. For simplicity, trench isolation, which fills all the surface area except for the diffusion regions and the channel regions, is not shown.



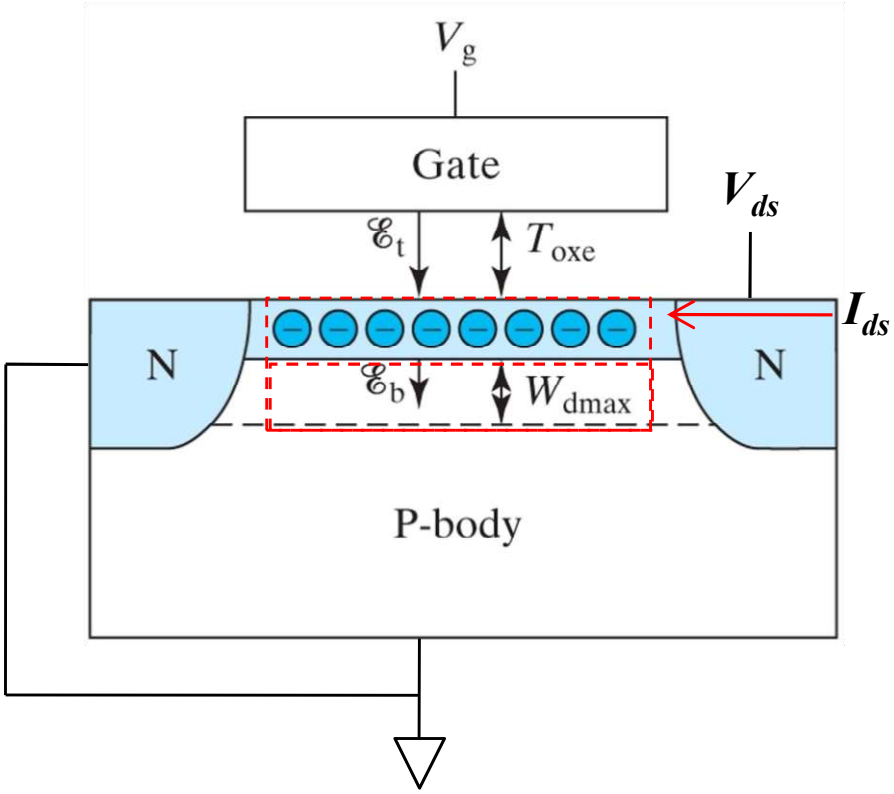
(c)

(c) Layout of a CMOS inverter.

Surface Mobilities and High-Mobility FETS

The electron and hole mobility in the surface inversion layer:
important factor that determines the MOSFET current and circuit speed.

Surface Mobilities



When a small V_{ds} is applied, the drain to source current, I_{ds},

$$I_{ds} = qnv = W \cdot Q_{inv} \cdot v = WQ_{inv}\mu_{ns}\mathcal{E}$$
$$= WQ_{inv}V_{ds}\mu_{ns} / L = WC_{oxe}(V_{gs} - V_t)\mu_{ns}V_{ds} / L$$

where, qn = electron charge density in inversion channel (C/cm^3)

Q_{inv} = inversion charge density (C/cm^2)

W = channel width

$\mathcal{E}$ = channel electric field

L = channel length

μ_{ns} = electron *surface mobility* or *effective mobility*

All quantities besides μ_{ns} are known or can be measured, and ,therefore μ_{ns} can be determined.

Applying Gauss's law to a box encloses only the depletion layer, $E_b = -Q_{dep} / \varepsilon_s$

and to a box encloses the depletion layer and the inversion layer, $E_t = -(Q_{dep} + Q_{inv}) / \varepsilon_s$

From MOS capacitor, $V_t = V_{fb} + \phi_{st} - Q_{dep} / C_{oxe}$

$$\therefore E_b = \frac{C_{oxe}}{\varepsilon_s}(V_t - V_{fb} - \phi_{st}) \text{ and}$$

$$E_t = E_b - Q_{inv} / \varepsilon_s = E_b + \frac{C_{oxe}}{\varepsilon_s}(V_{gs} - V_t)$$

$$= \frac{C_{oxe}}{\varepsilon_s}(V_{gs} - V_{fb} - \phi_{st})$$

$$\frac{1}{2}(E_b + E_t) = \frac{C_{oxe}}{2\varepsilon_s}(V_{gs} + V_t - 2V_{fb} - 2\phi_{st})$$

$$= \frac{C_{oxe}}{2\varepsilon_s}(V_{gs} + V_t + 0.2V)$$

$$= \frac{\varepsilon_{ox}}{2\varepsilon_s T_{oxe}}(V_{gs} + V_t + 0.2V)$$

$$= \frac{(V_{gs} + V_t + 0.2V)}{6T_{oxe}}$$ for N+ poly-gate

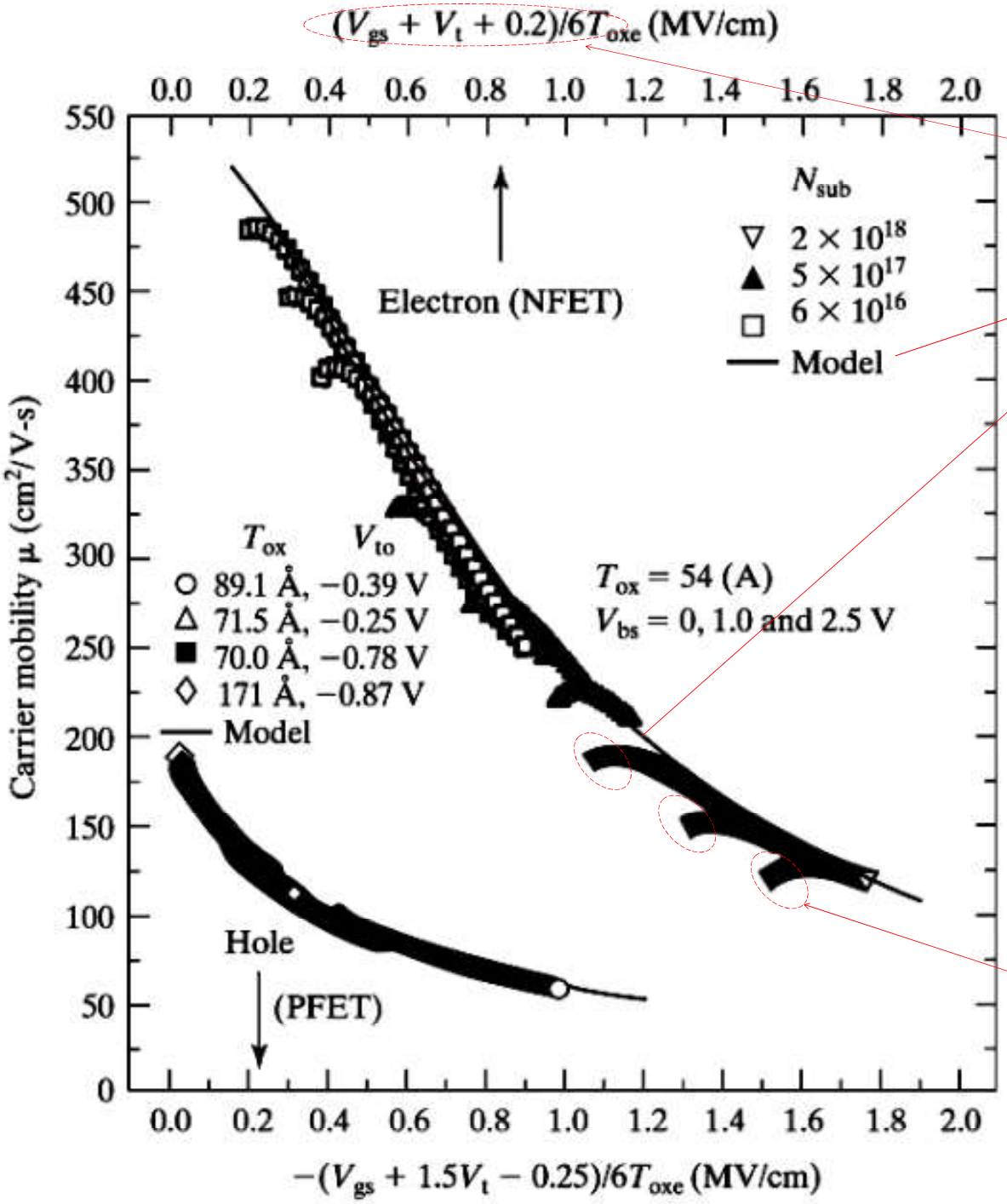
Empirically,

$$\mu_{ns} = \text{function of } \frac{1}{2}(E_b + E_t) \quad \text{or} \quad \text{function of } Q_{dep} + \frac{Q_{inv}}{2}$$

$$\mu_{ps} = \text{function of } \frac{1}{2}(1.5E_b + E_t)$$

$$\mu_{ns} = \frac{540\,cm^2/Vs}{1 + \left(\frac{V_{gs} + V_t + 0.2V}{5.4T_{oxe}}\right)^{1.85}}$$

$$\mu_{ps} = \frac{185\,cm^2/Vs}{1 - \left(\frac{V_{gs} + 1.5V_t - 0.25\,V}{3.38T_{oxe}}\right)}$$



When V_{gs}, V_t, and T_{oxe} are properly considered, all silicon MOSFETs exhibits essentially the same surface mobility, called "**universal effective mobility.**"

Surface mobility is lower than the bulk mobility because of **surface roughness scattering.** $(\mu_{ns} < \mu_n \quad and \quad \mu_{ps} < \mu_p)$

⬇

Mobilities decrease as the field in the inversion layer (E_b, E_t) becomes stronger and the charge carriers are confined closer to the Si-SiO_2 interface.

$$\mu_{ns} \; and \; \mu_{ps} \propto T^{-3/2}$$

$$\Rightarrow characteristic \; of \; phonon \; scattering$$

Surface mobility around $V_g \sim V_t$, especially in the heavily doped semiconductor, is lower than the universal mobility, due to dopant ion scattering.

⬇

At higher V_g, dopant ion scattering effect is screened out by the inversion layer carriers.

• Mobility Models

✓ The mobility of carriers in the channel of a MOSFET is lower than in bulk semiconductors because there are additional scattering mechanism.

✓ Additional scattering factors

; surface roughness at semiconductor-oxide interface

; coulombic interaction with fixed charges in the gate oxide

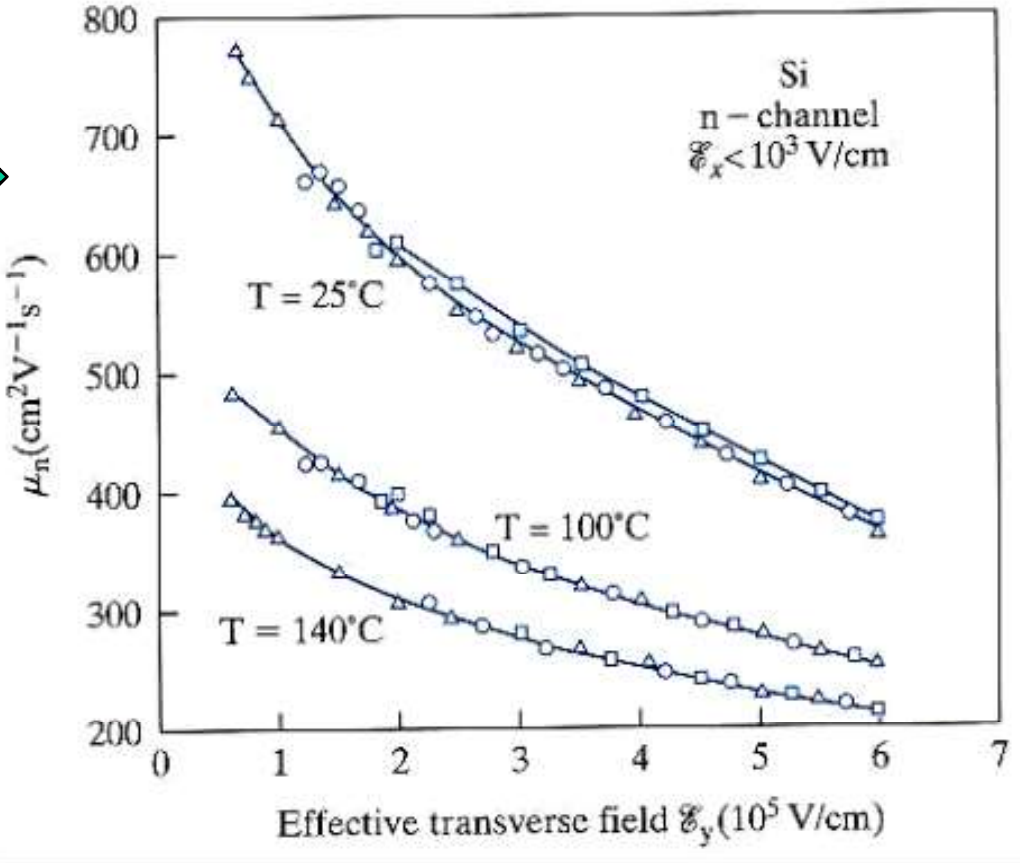
✓ Gate bias voltage (or transverse field) increases

; more carriers are drawn to the oxide-silicon interface

; effective carrier mobility in channel decreases

$$\mu_{ns} \text{ and } \mu_{ps} \propto T^{-3/2}$$

$$\Rightarrow \textit{characteristic of phonon scattering}$$



Inversion layer electron mobility versus effective transverse field, at various temperatures. The triangles, circles, and squares refer to different gate oxide thickness and channel dopings.

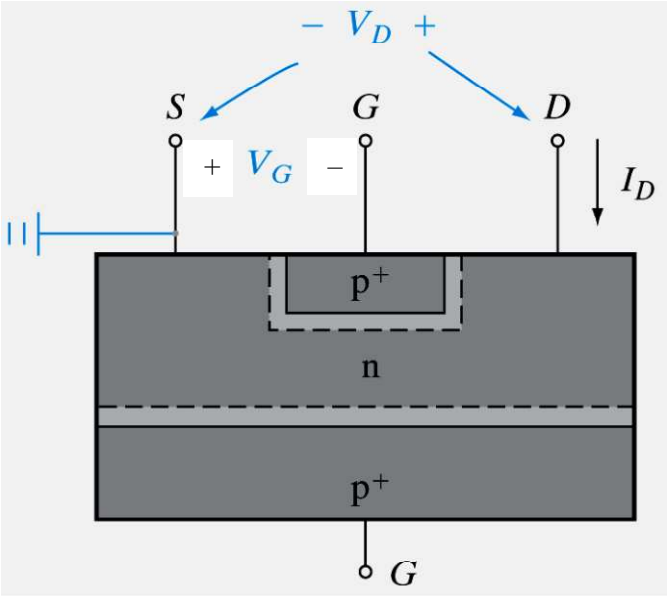
JFET(junction field-effect transistor)

Before the advent of MOSFET, ICs were built with bipolar transistors, which have forward diode at the input and draw significant input current. The high input current and capacitances were quite undesirable for some circuits.

JFET provided a low input current and capacitance device because its input is a reverse-biased diode. JFET can be fabricated with bipolar transistors and coexist in the same IC chip.

A reverse bias (gate voltage) between the p^+ regions and the channel causes the depletion region to intrude into the n channel, and therefore the effective width of the channel can be restricted changing the channel resistance.

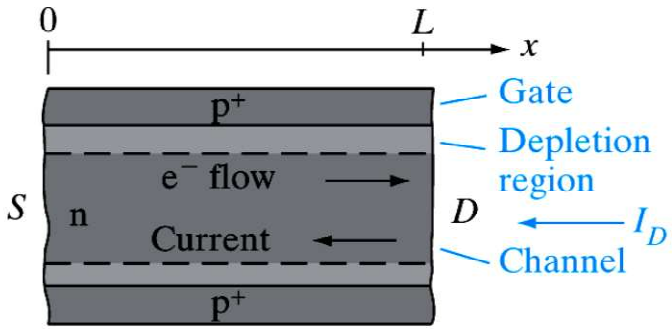
→ The JFET current can be controlled with the gate voltage.

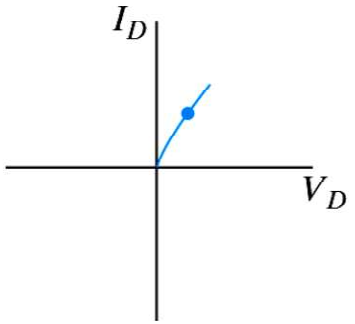


• Pinch-off and Saturation

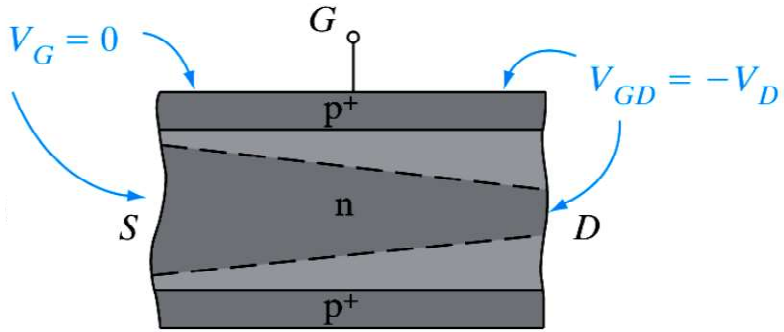
$V_G = 0\,V \rightarrow$ G short circuited to S

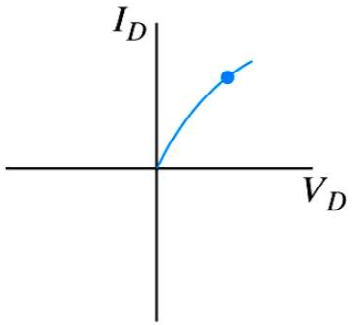
linear range



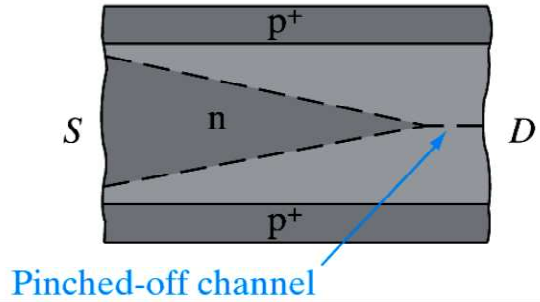


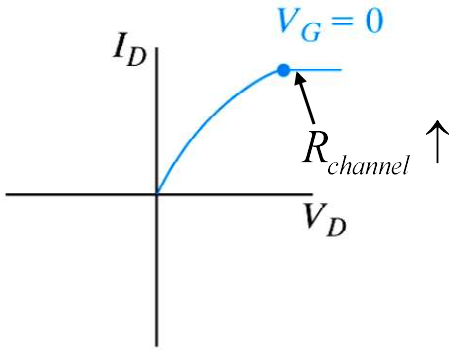
near pinch-off





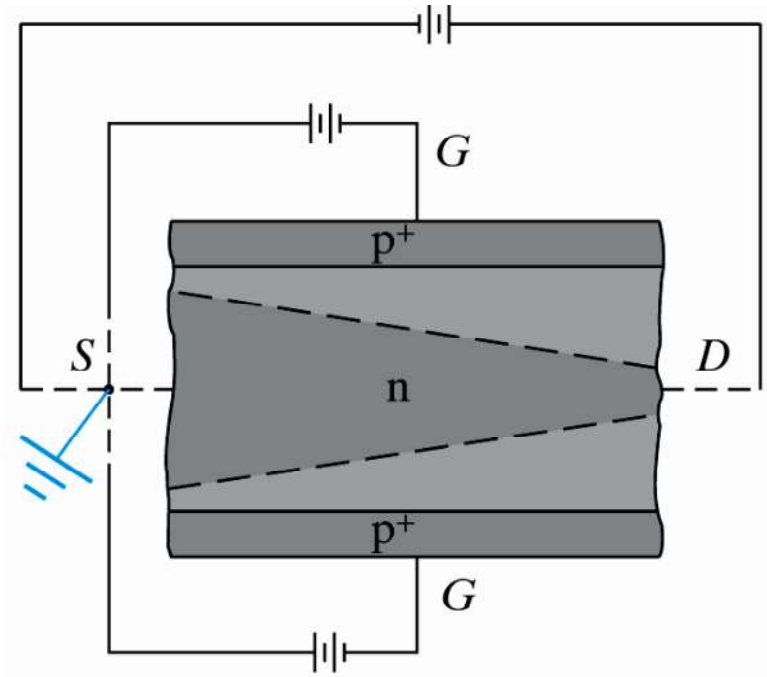
beyond pinch-off

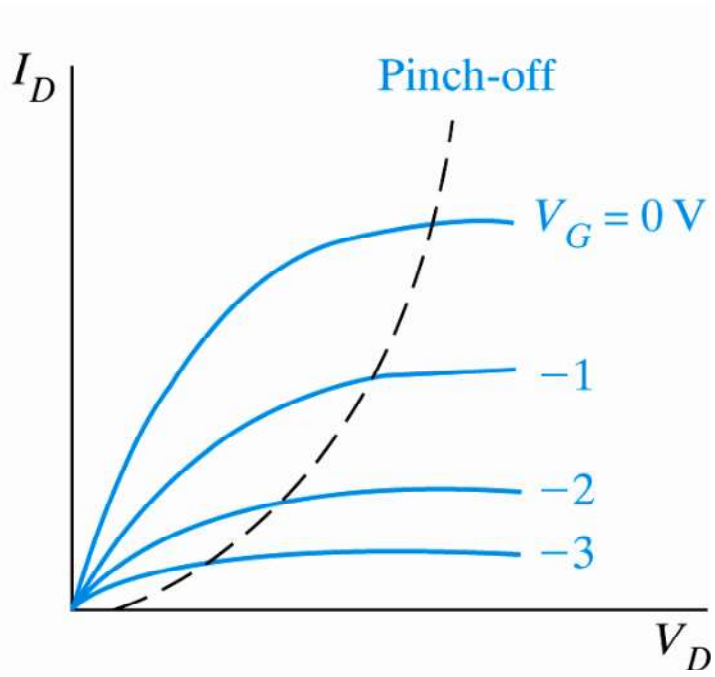




Depletion regions in the channel of a JFET with zero gate bias for several values of V_D

- **Gate Control**
  - ✓ The effect of a negative gate bias $-V_G$ is to increase the resistance of the channel and induce pinch-off at a lower value of a current.
  - ✓ As V_G is varied, a family of curves is obtained for the *I-V* characteristic of the channel.
  - ✓ Beyond the pinch-off voltage, I_D is controlled by V_G → amplification
  - ✓ Since the input control voltage V_G appears across the reverse-biased gate junctions, the input impedance of the device is high.





GaAs MESFET (metal-semiconductor field-effect transistor)

Higher carrier mobility allows the carriers to travel faster and the transistors to operate at higher speeds.

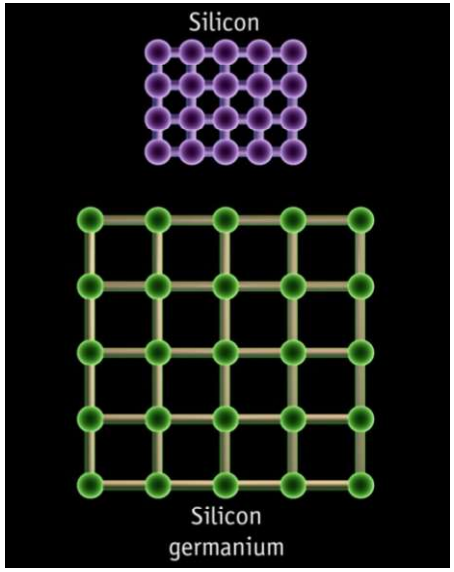
⬇

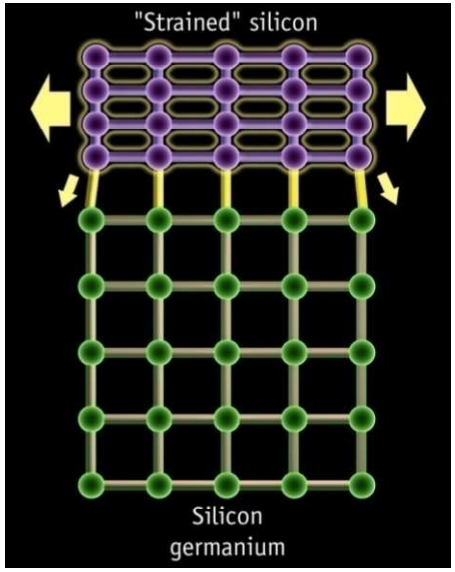
High speed devices not only improve the throughput of electronic equipment but also open up new applications such as inexpensive microwave communication.

The most obvious way to improve speed is to use a semiconductor having higher mobility than Si such as **Ge**, **SiGe alloy** (grown epitaxially over Si substrates), or **strained Si**.

⬇

The extension of Si technology is a promising way to improve the device speed.





GaAs: much higher electron mobility than Si.

too many charge traps at the semiconductor/dielectric interface for MOSFET application.

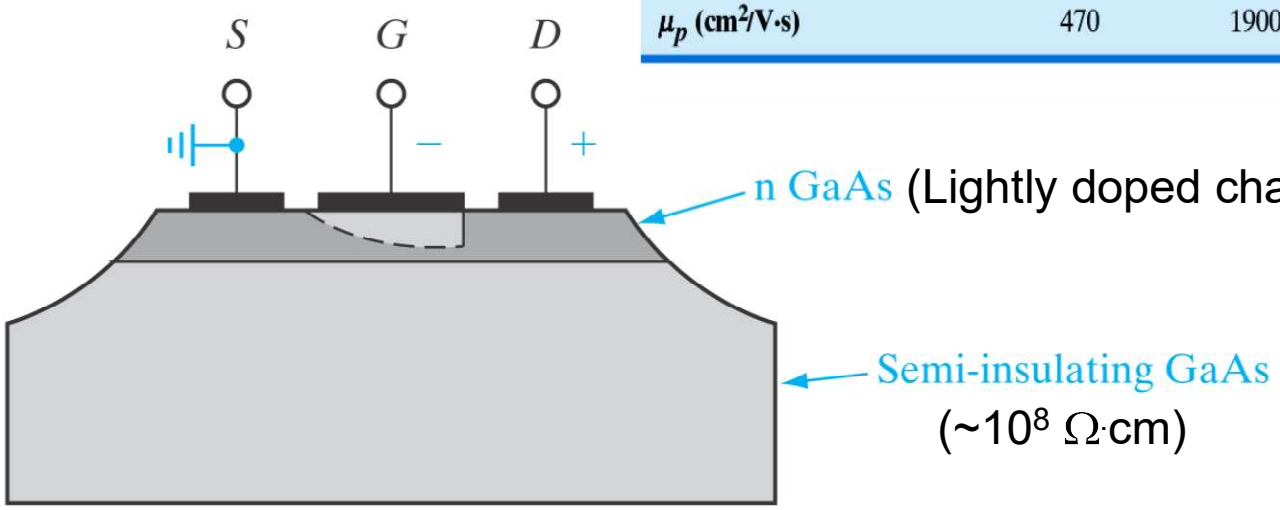
↳ Fortunately, a Schottky junction can serve as the control gate of a GaAs FET(mostly, N-channel) in place of MOS gate.

• The GaAs MESFET

- ✓ Ohmic contacts(Au-Ge) for S/D and Schottky barrier (Al or Au) gate (reverse bias)
  - → By reverse biasing the Schottky gate, the channel can be depleted
  - → similar I-V characteristics to JFET device
- ✓ By using GaAs instead of Si, a higher electron mobility is available, and GaAs can be operated at higher temperatures (therefore higher power levels).
- ✓ The source and drain contacts may be improved by further n^+ implantation

TABLE 2–1 • Electron and hole mobilities at room temperature of selected lightly doped semiconductors.

	Si	Ge	GaAs	InAs
μ_n ($cm^2/V \cdot s$)	1400	3900	8500	30, 000
μ_p ($cm^2/V \cdot s$)	470	1900	400	500

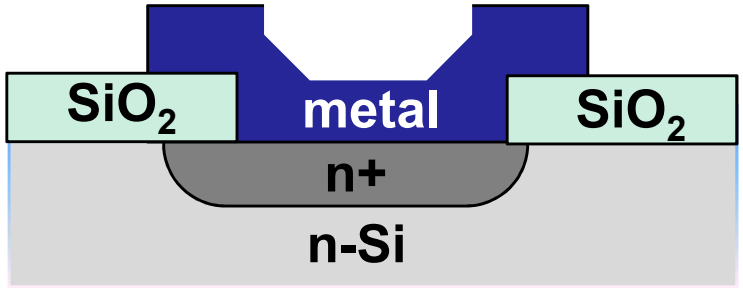


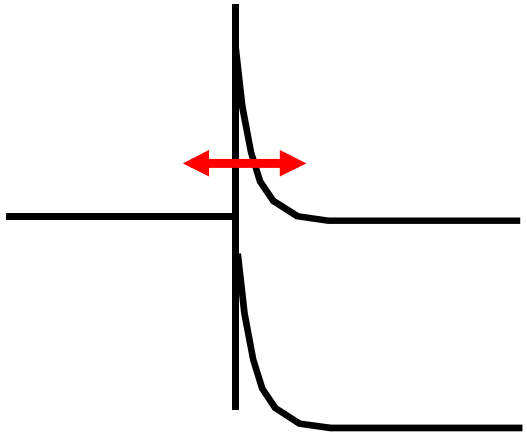
GaAs MESFET

❑ Metal-Semiconductor Junctions

❑ Ohmic contacts : practical method

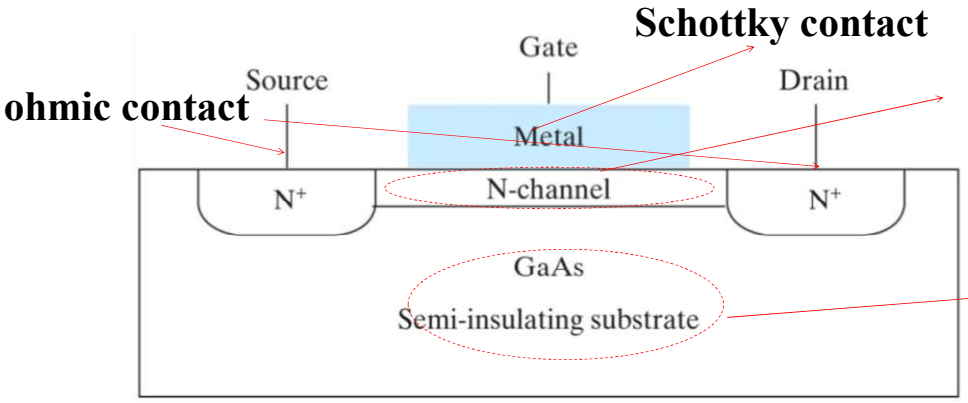
- ✓ A practical method for forming Ohmic contacts regardless of the biasing polarity and relative work functions between S & M is by doping the semiconductor heavily ($10^{17}/cm^3 \sim 10^{19}/cm^3$) in the contact (surface) region.
- ✓ Barrier height is not affected by an increase in the semiconductor doping. Thus if a barrier exists at the interface, the depletion width is small enough to allow carriers to tunnel through the barrier.





Electrical nature of Ideal MS contacts

	n-type Semiconductor	p-type Semiconductor
$\Phi_M > \Phi_S$	Rectifying	Ohmic
$\Phi_M < \Phi_S$	Ohmic	Rectifying



A large Schottky barrier height (for example, Au) is desirable for minimizing the input gate current.

GaAs has a large E_g and small n_i.

⬇

Undoped GaAs has a very high resistivity and can be considered an insulator.

When a reverse-bias voltage or a small forward voltage is applied to the gate, the depletion region under the gate expands or contracts.

⬇

This modulate the thickness of the conductive channel and, in turn, modulates the channel current I_{ds}.

I_{ds} does not flow in a surface inversion layer. ➡ *The electron mobility is not degraded by surface scattering. This further enhances GaAs MESFET's speed advantage*

Depletion-mode: N-channel thickness > depletion-layer width at $V_g = 0$.
requires a reverse gate voltage to turn it off.
easy to make.

Enhancement-mode: N-channel thickness < depletion-layer width at $V_g = 0$ and
requires a forward gate voltage to turn it on.
difficult to make, but much easier to make circuit design.

(Modern Si MOSFET are all enhancement-mode.)

• The High Electron Mobility Transistor (HEMT)

✓ Bandgap engineering is available with hetero-junctions.

✓ To obtain high transconductance, the conductivity of the channel can be increased by

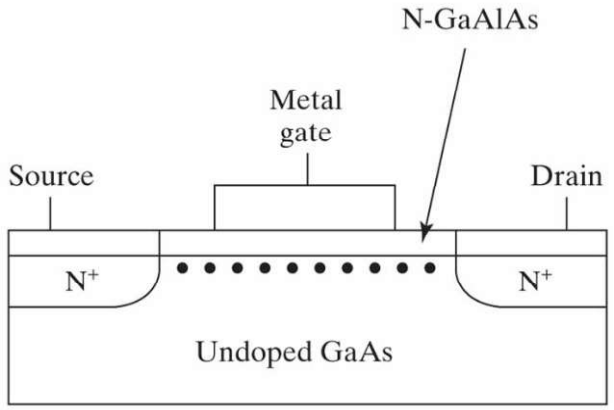
(1) high doping in the channel, *but*

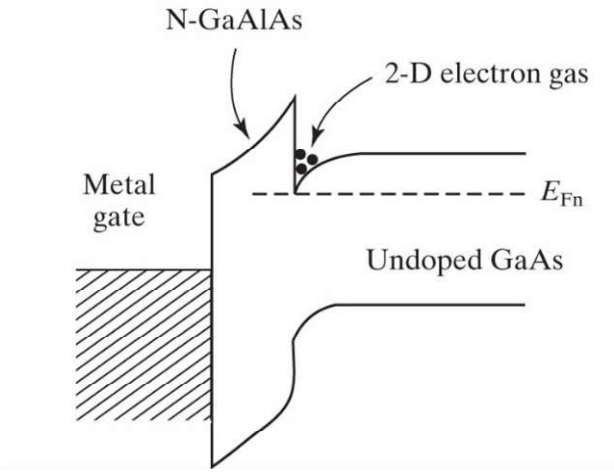
→ scattering by the ionized impurities → low mobility

(2) growing of a thin undoped well (e.g., GaAs) bounded by wider bandgap, doped barriers (e.g., AlGaAs)

→ ***modulation doping*** : conductive GaAs when electrons from the doped AlGaAs barriers fall into the well and become trapped there

→ reduced scattering → high μ

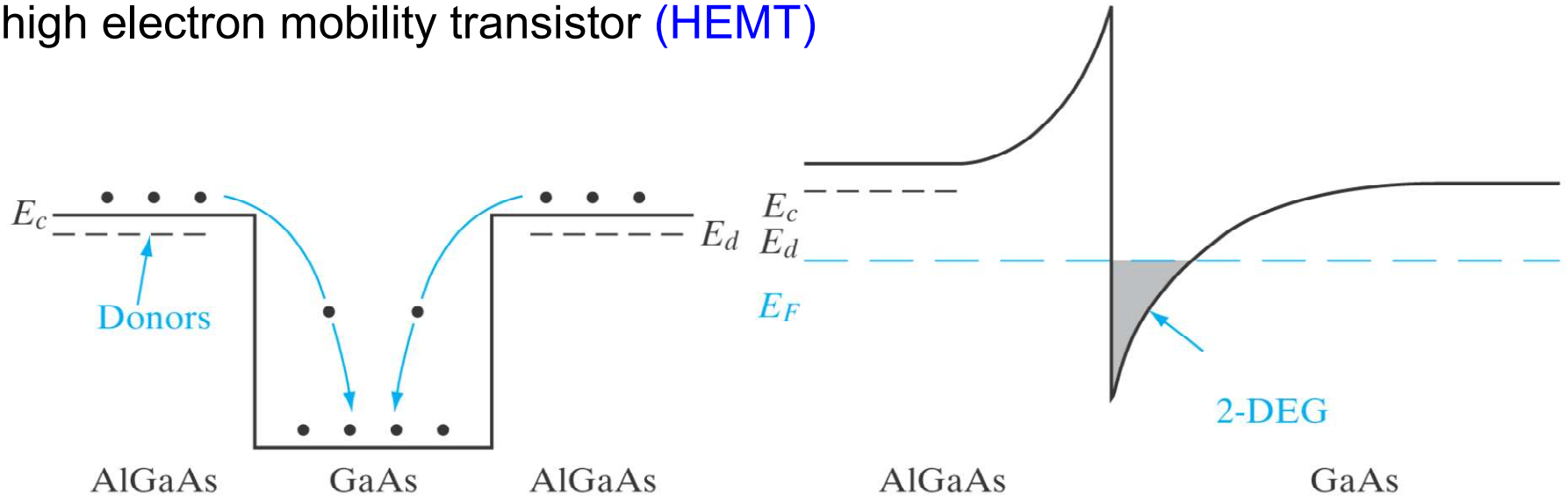




The thin sheet of free electrons at the interface due to modulation doping forms a two-dimensional electron gas (2-DEG),which do not suffer from ionized impurity scattering and hence the channel mobility can be much higher.

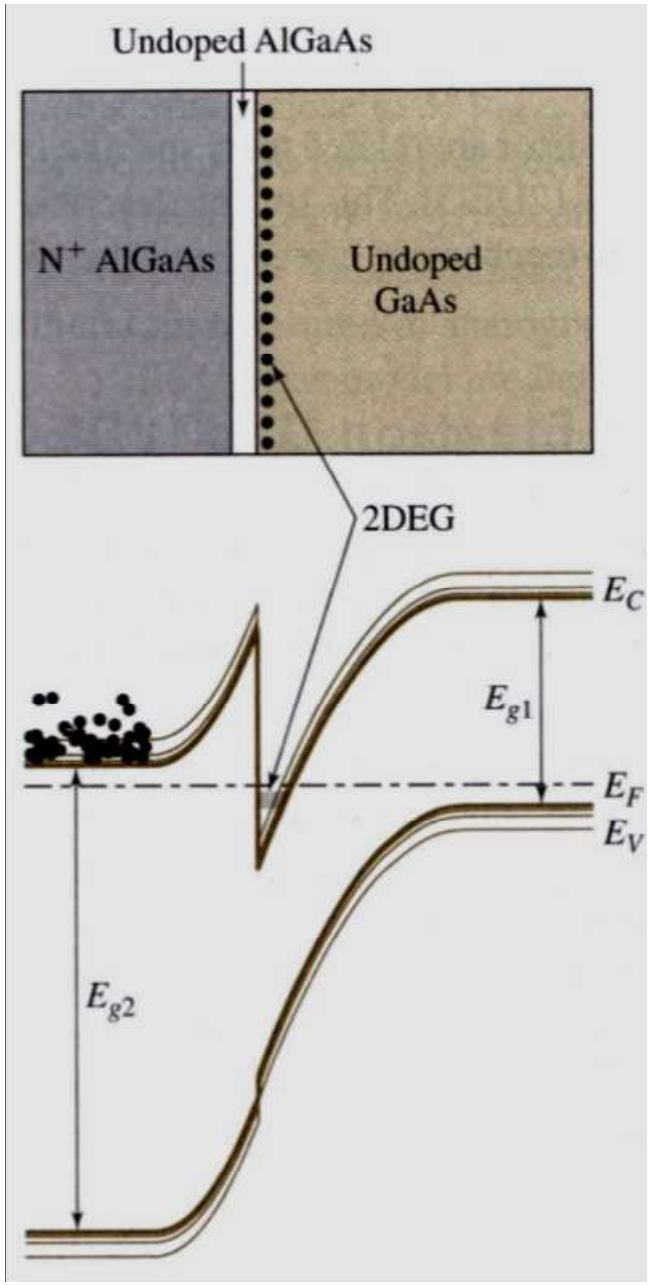
⬇

can be exploited in modulation doped field effect transistor (MODFET) or high electron mobility transistor (HEMT)

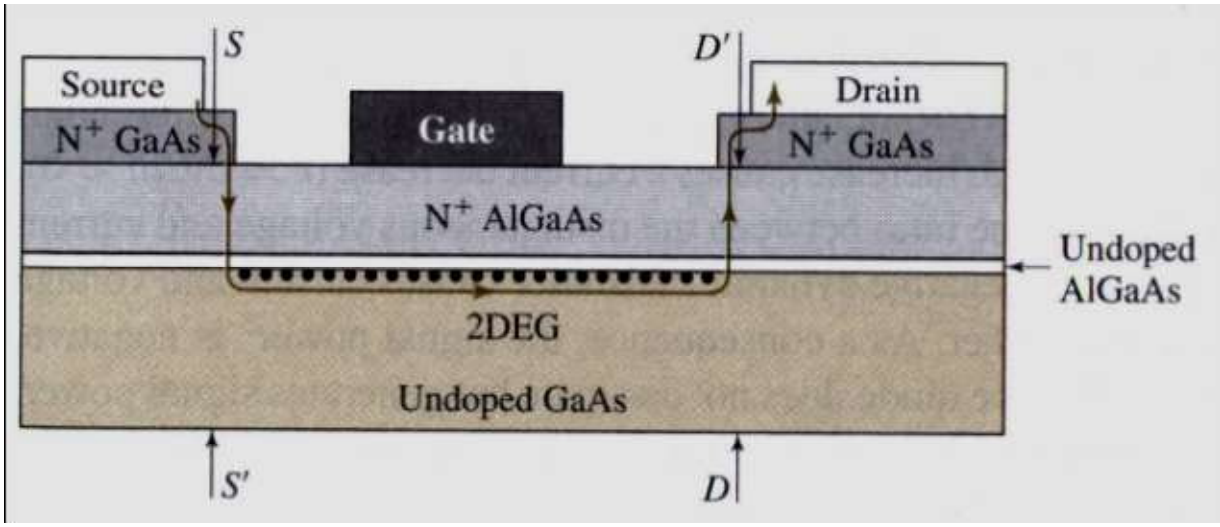


SQW with double heterostructure TQW with single heterostructure

(a) Simplified view of modulation doping, showing only the conduction band. Electrons in the donor-doped AlGaAs fall into the GaAs potential well and become trapped. As a result, the undoped GaAs becomes n-type, without the scattering by ionized donors which is typical of bulk n-type material. (b) Use of a single AlGaAs/GaAs hetero-junction to trap electrons in the undoped GaAs. The thin sheet of charge due to free electrons at the interface forms a two-dimensional electron gas (2-DEG),which can be exploited in HEMT device.



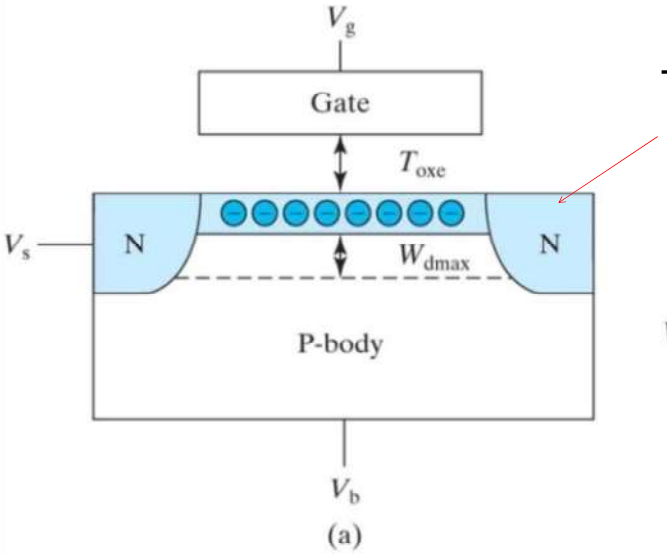
The basic HEMT structure. The large band gap AlGaAs functions like the SiO_2 in a MOSFET. The conduction channel is in the undoped GaAs.



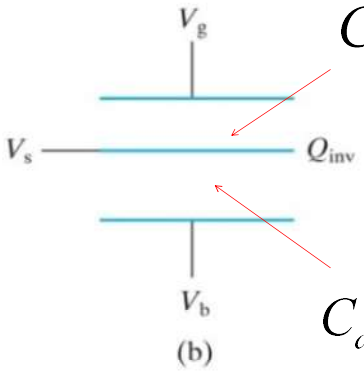
- ✓ The electrons are accumulated at the corners of the well due to band-bending at the hetero-junction.
- ✓ The donors in the AlGaAs layer are purposely separated from the interface by ~100 Å. The channel region is spatially separated from the ionized impurities.
- ✓ The advantage of a HEMT are its ability to locate a large electron density (~10^{12} cm^{-2}) in a very thin layer (< 10 nm thick) very close to the gate while simultaneously eliminating ionized impurity scattering.
- ✓ The advantages of the HEMT over the Si MOSFET are the *higher mobility* and *maximum electron velocity* in GaAs compared with Si, and the smoother interfaces possible with an AlGaAs/GaAs hetrojunction compared with the Si/SiO_2 interface
  → extremely high cut-off frequency (10GHz operation) & high gain & ultra low noise < 1dB
- ✓ *Pseudomorphic*: growth of a thin layer on a crystal with different lattice constant. The thin layer is under strain, but keeps crystalline.

  - ✓ Two dimensional electron gas FET(2-DEG FET, TEGFET) ; conduction along the channel occurs in a thin sheet of charge.
  - ✓ Separately doped FET (SEDFET) ; doping occurs in a separate region from the channel.

MOSFET V_t, Body Effect, and Steep Retrograde Doping



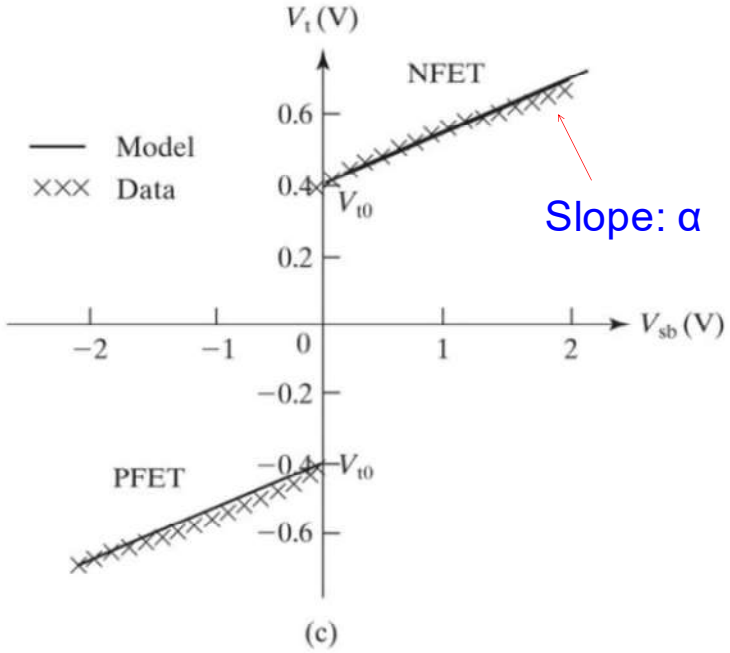
The drain is open-circuited



$$C_{oxe} = \frac{\varepsilon_{ox}}{T_{oxe}}$$

$$C_{dep} = \frac{\varepsilon_s}{W_{dmax}}$$

The inversion layer can be viewed as a conductive film that is coupled to V_g through the oxide capacitance and coupled to V_b through the depletion-layer capacitance.



Slope: α

V_t is an approximately linear function of the body to source bias voltage. The polarity of the body bias is normally that which would reverse bias the body-source junction. V_t becomes more positive and negative for NFET and PFET with V_{sb}, respectively.

MOSFET V_t

From Chapter 5, with $V_b = V_s$, $Q_{inv} = -C_{oxe}(V_{gs} - V_t)$

Since the body and the channel are coupled by C_{dep}, V_{sb} induces a charge in the inversion layer, $C_{dep}V_{sb}$.

$$Q_{inv} = -C_{oxe}(V_{gs} - V_t) + C_{dep}V_{sb}$$

$$= -C_{oxe}\left(V_{gs} - \left(V_t + \frac{C_{dep}}{C_{oxe}}V_{sb}\right)\right) = -C_{oxe}(V_{gs} - (V_t(V_{sb}))$$

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

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

$$\alpha = \frac{C_{dep}}{C_{oxe}} = \frac{\varepsilon_s / W_{dmax}}{\varepsilon_{ox} / T_{oxe}} = \frac{3T_{oxe}}{W_{dmax}}$$ called **body-effect coefficient**

Body Effect

The fact that V_t is a function of the body bias is called the **body effect**.

When multiple NFETs (or PFETs) are connected in series in a circuit, they share a common body (the silicon substrate) but their sources do not have the same voltage. Clearly some transistors' source-body junctions are reverse biased. This raises their V_t and reduces I_{ds} and the circuit speed.

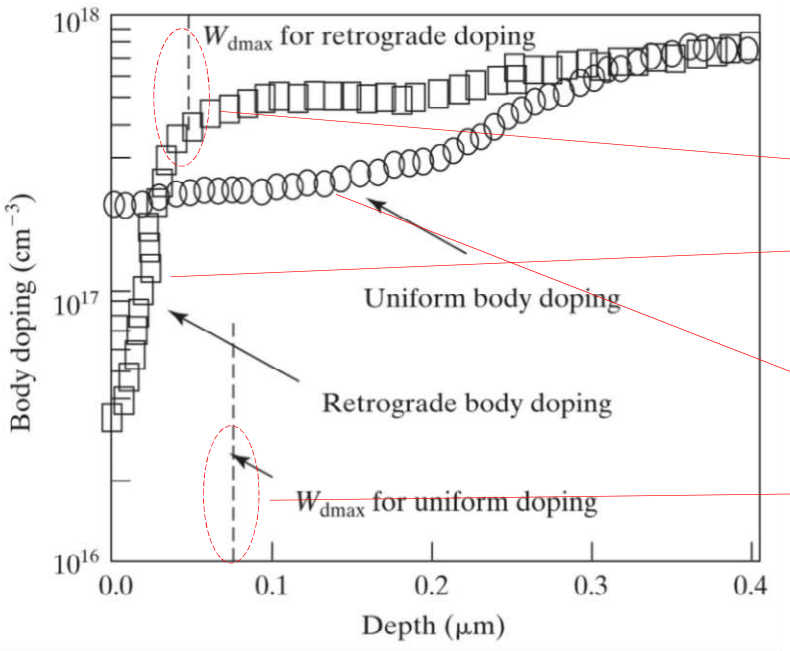
↳ V_t should be insensitive to V_{sb} as possible, i.e., the body effect should be minimized.

This can be accomplished by minimizing the α or T_{oxe}/W_{dmax} ratio.

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,

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

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

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

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

Uniform Doping

• Substrate Bias Effects (Body effect)

✓ With a reverse bias between the substrate and the source (V_B negative for an n-channel device), the depletion region is widened and the threshold gate voltage required to achieve inversion must be increased to accommodate the larger Q_d.

$$V_{t0} = V_{fb} + 2\phi_B - \frac{-\sqrt{qN_a 2\varepsilon_s 2\phi_B}}{C_{ox}} \Rightarrow Q_d = -qN_a W_{d\max} = -(qN_a 2\varepsilon_s 2\phi_B)^{1/2}$$

$$Q'_d = -qN_a W'_{d\max} = -(qN_a 2\varepsilon_s (2\phi_B + V_{SB}))^{1/2}$$

For uniform doping, V_t can be obtained by replacing the $2\phi_B$ term in V_t equation with $2\phi_B + V_{sb}$.

$$V_t = V_{t0} + \frac{\sqrt{qN_a 2\varepsilon_s}}{C_{oxe}}(\sqrt{2\phi_B + V_{sb}} - \sqrt{2\phi_B}) \equiv V_{t0} + \gamma(\sqrt{2\phi_B + V_{sb}} - \sqrt{2\phi_B}),$$

$$\gamma \text{ is called the body-effect parameter.}$$

V_t is a sublinear function of V_{sb}.

This equation can be linearlized by Taylor expansion so that V_t is expressed as a linear function of V_{sb}.

✓ If the V_{SB} (typically, ~0.6 V) is much larger than $2\phi_B$,

$$\Delta V_T \; ; \; \frac{\sqrt{2\varepsilon_s qN_a}}{C_{oxe}}(V_{SB})^{1/2} \text{ (n channel)}$$

$$\Delta V_T \; ; \; -\frac{\sqrt{2\varepsilon_s qN_d}}{C_{oxe}}(-V_{SB})^{1/2}$$

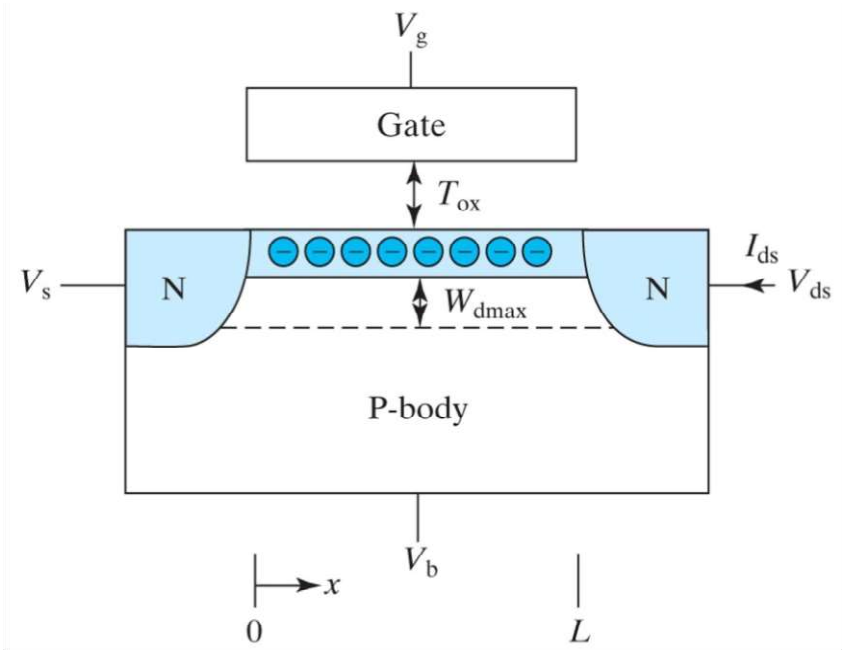
(p - channel)

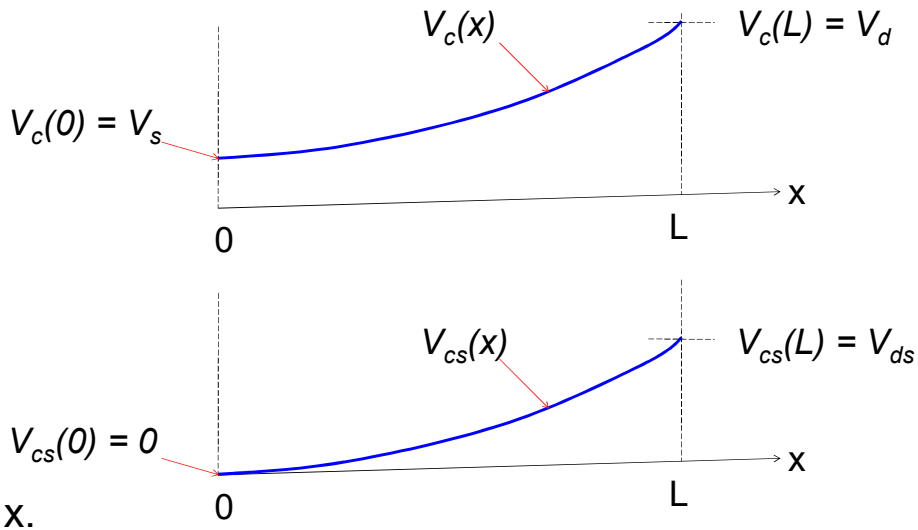
- $V_B \uparrow \rightarrow \Delta V_T$ becomes more positive
- The effect of this bias becomes more dramatic as the substrate doping is increased.

$$\Delta V_T \propto \sqrt{N_a}$$

✓ The substrate bias effect increases V_T for either type of device.

Q_{INV} in MOSFET





When $V_{ds} \neq 0$, the channel voltage V_c is a function of x.

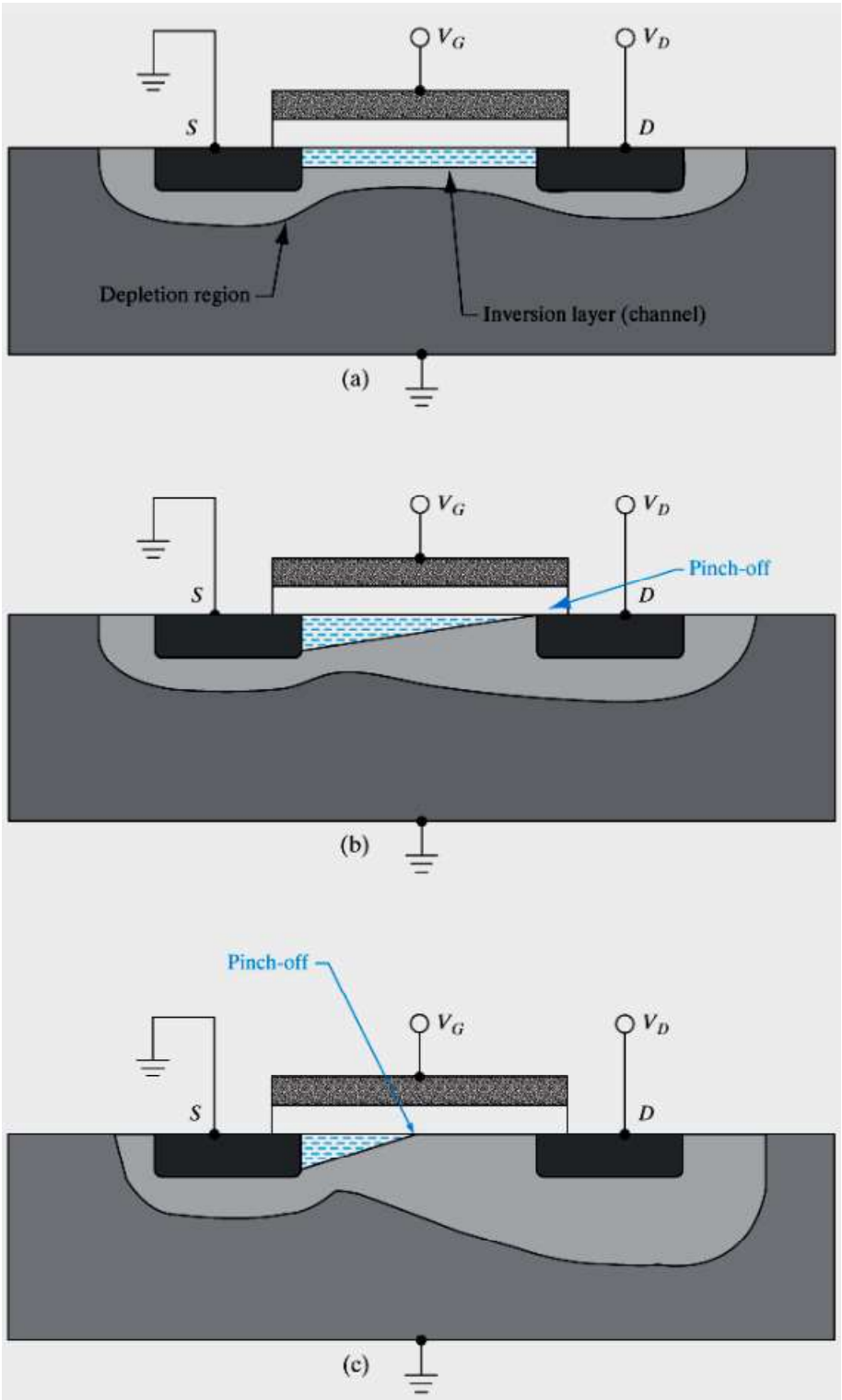
Because the voltage in the middle of the channel is higher at V_c (x), there is less voltage across the oxide capacitor (and across the depletion layer capacitor). Therefore, there will be fewer electrons on the capacitor electrode (the inversion layer)

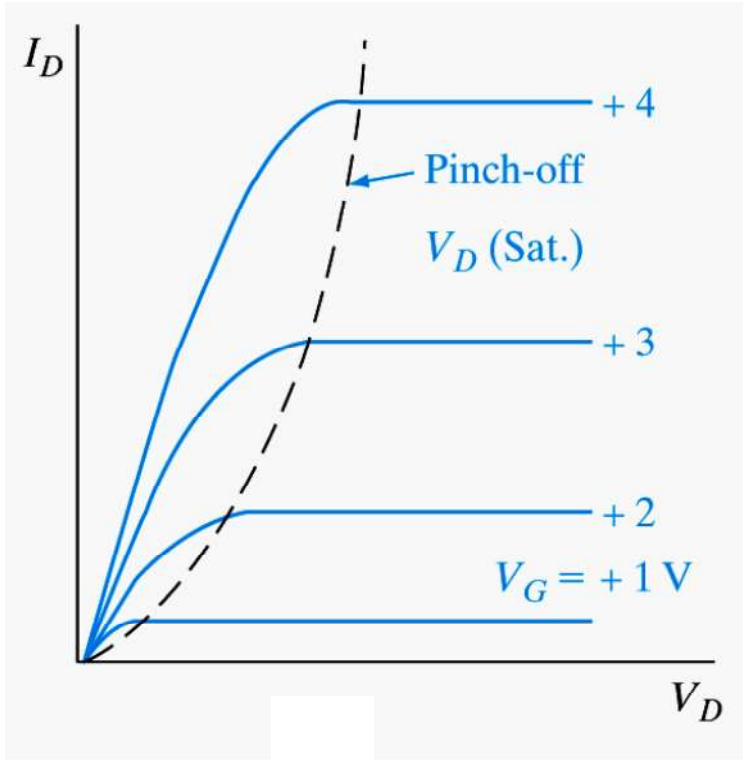
When the drain is open-circuited, $Q_{inv} = -C_{oxe}(V_{gs} - (V_t(V_{sb})), \quad V_t(V_{sb}) = V_{t0} + \alpha V_{sb}$

When the drain bias, V_d, is applied, V_{gs} term should be replaced by $V_{gc}(x)$ or $V_{gs} - V_{cs}(x)$ and V_{sb} by $V_{cb}(x)$ or $V_{sb} + V_{cs}(x)$.

$$Q_{inv}(x) = -C_{oxe}\{V_{gc}(x) - V_t(V_{cb}(x))\} < Q_{inv}(0)$$

$$\Rightarrow Q_{inv}(x) = f(x)$$


V_G
V_D
S
D
Depletion region
Inversion layer (channel)
(a)
V_G
V_D
S
D
Pinch-off
(b)
Pinch-off
V_G
V_D
S
D
(c)


I_D
+ 4
Pinch-off
V_D (Sat.)
+ 3
+ 2
V_G = + 1 V
V_D

$$Q_{inv}(x) = -C_{oxe}\{V_{gc}(x) - V_t(V_{cb}(x))\}$$
$$= -C_{oxe}(V_{gs} - V_{cs}(x) - V_{t0} - \alpha(V_{sb} + (V_{cs}(x))$$
$$= -C_{oxe}(V_{gs} - V_{cs}(x) - (V_{t0} + \alpha V_{sb}) - \alpha V_{cs}(x))$$
$$\therefore \quad \boxed{Q_{inv}(x) = -C_{oxe}(V_{gs} - mV_{cs}(x) - V_t)}$$

$$\boxed{m \equiv 1 + \alpha = 1 + \frac{C_{dep}}{C_{oxe}} = 1 + \frac{3T_{oxe}}{W_{dmax}},}$$ **m** is the **bulk-charge factor**, typically ~ 1.2.

The body is sometimes called the **back gate** since it clearly has a similar though weaker effect on the channel charge.

The back-gate effect on Q_{inv} is often called the **bulk-charge effect**, closely linked to the body-effect.

Basic MOSFET I-V Model

$$I_{ds} = -W \cdot Q_{inv}(x) \cdot v = -W \cdot Q_{inv}(x)\mu_{ns}E = WC_{oxe}(V_{gs} - mV_{cs} - V_t)\mu_{ns}\frac{dV_{cs}}{dx}$$

$$\Rightarrow \int_0^L I_{ds}dx = WC_{oxe}\mu_{ns}\int_0^{V_{ds}}(V_{gs} - mV_{cs} - V_t)dV_{cs}$$

$$I_{ds}L = WC_{oxe}\mu_{ns}(V_{gs} - V_t - \frac{m}{2}V_{ds})V_{ds}$$

$$I_{ds} = \frac{W}{L}C_{oxe}\mu_{ns}(V_{gs} - V_t - \frac{m}{2}V_{ds})V_{ds}$$

$I_{ds} \propto W(channel\ \ width),$

$\mu_{ns},$

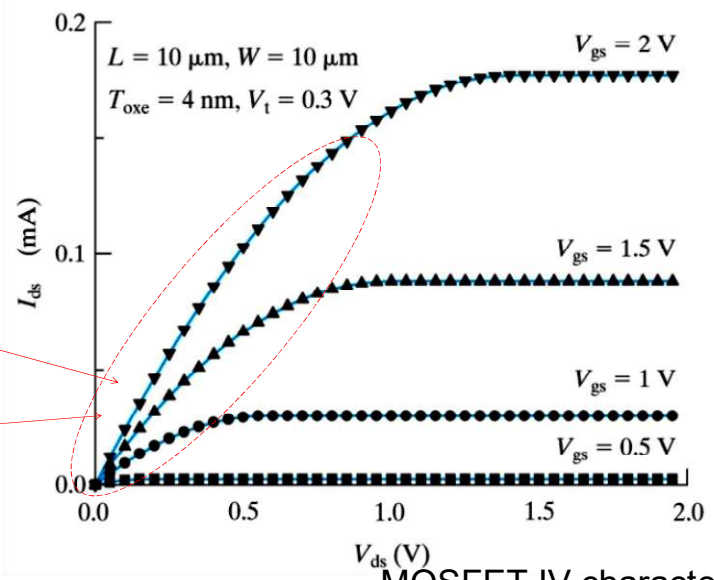
$V_{ds} / L(average\ field\ in\ the\ channel),$

$C_{oxe}(V_{gs} - V_t - \frac{m}{2}V_{ds})(average\ Q_{inv}\ in\ the\ channel).$

When V_{ds} is very small, the $mV_{ds}/2$ term is negligible, then

$$I_{ds} \approx \frac{W}{L}C_{oxe}\mu_{ns}(V_{gs} - V_t)V_{ds} \propto V_{ds}$$

(The transistor behaves as a resistor, called linear region operation.)



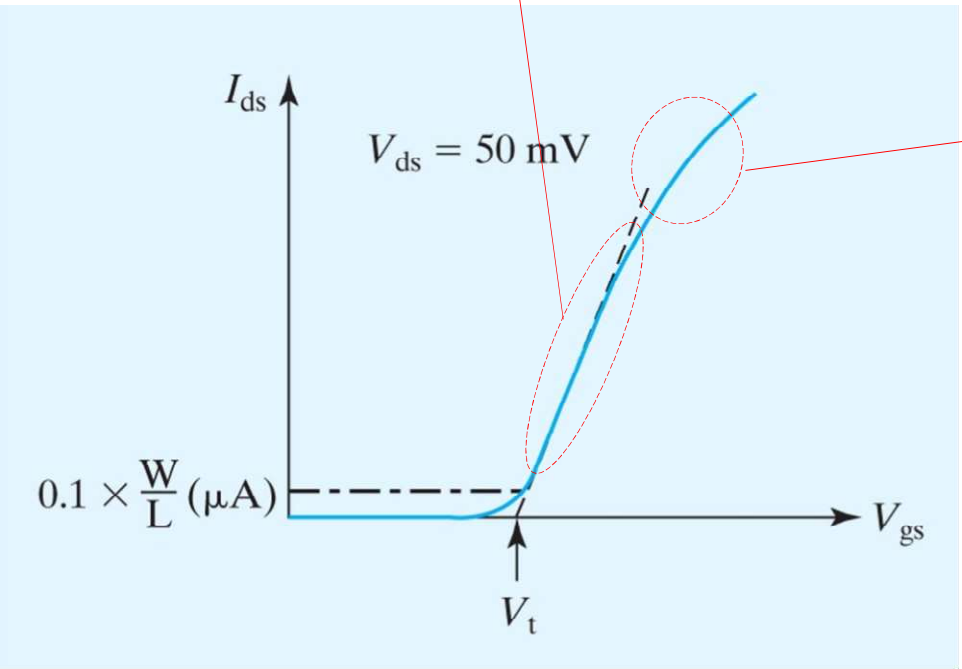
MOSFET IV characteristics.

How to measure the V_t of a MOSFET?

When V_{ds} is very small (for example ~ 50 mV),

$$I_{ds} \approx \frac{W}{L} C_{oxe} \mu_{ns} (V_{gs} - V_t) V_{ds}$$

At $V_{gs} > V_t$, I_{ds} increases linearly with (V_{gs} - V_t), if μ_{ns} is constant.



The curve becomes sublinear, because μ_{ns} decreases with increasing V_{gs}.

V_t can be measured by extrapolating the I_{ds} vs. V_{gs} curve to I_{ds} = 0.
Alternatively, it can be defined as the V_{gs}, at which I_{ds} is a small fixed amount such as

$$I_{ds} = 0.1 \mu A \times \frac{W}{L}$$

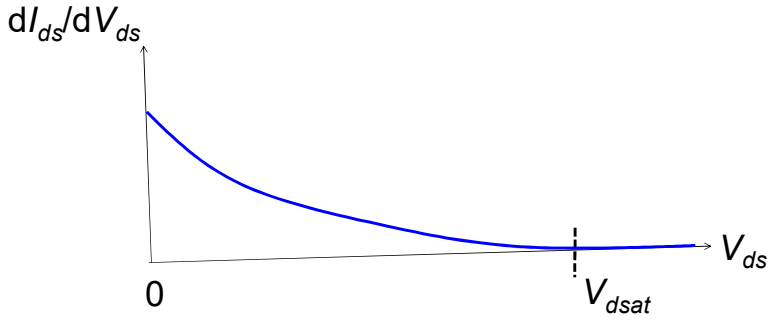
Saturation

$$I_{ds} = \frac{W}{L} C_{oxe} \mu_{ns} \left(V_{gs} - V_t - \frac{m}{2} V_{ds}\right) V_{ds}$$

As V_{ds} increases, the average Q_{inv} decreases and dI_{ds}/dV_{ds} decreases.

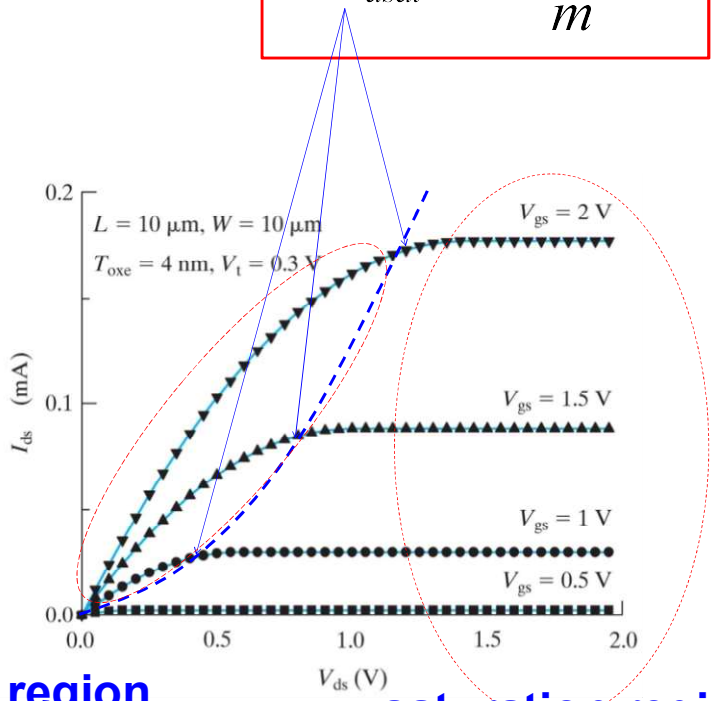
By differentiating the *I-V* equation with respect to V_{ds}, it can be shown that dI_{ds}/dV_{ds} becomes zero at certain V_{ds} (generally, this is the minimum value of dI_{ds}/dV_{ds} , because it can not be negative in normal case).

$$\frac{dI_{ds}}{dV_{ds}} = 0 = \frac{W}{L} C_{oxe} \mu_{ns} (V_{gs} - V_t - mV_{ds}) \quad at\ V_{ds} = V_{dsat}$$



$$\therefore V_{dsat} = \frac{V_{gs} - V_t}{m}$$

called the **drain saturation voltage**



linear region (ohmic region)

saturation region (active region)

The saturation current, I_{dsat}, can be obtained by substituting V_{dsat} for V_{ds}.

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

V_t also can be measured by extrapolating the $(\sqrt{I_{dsat}}\ vs.\ V_{gs})$ curve to $\sqrt{I_{dsat}} = 0$.

Pinch-off

What happens at $V_{ds} = V_{dsat}$ and why does I_{ds} stay constant beyond V_{dsat}?

From, $Q_{inv}(x) = -C_{oxe}(V_{gs} - mV_{cs}(x) - V_t)$

$$V_{cs}(x) = \frac{V_{gs} - V_t}{m}\left(1 - \sqrt{1 - \frac{x}{L}}\right)$$ (see Problem 6.9)

$$Q_{inv}(L) = -C_{oxe}(V_{gs} - mV_{cs}(L) - V_t)$$
$$= -C_{oxe}(V_{gs} - mV_{ds} - V_t), \; at \; drain \; end$$
$$= -C_{oxe}\left(V_{gs} - m\frac{V_{gs} - V_t}{m} - V_t\right) = 0, \; at \; V_{ds} = V_{dsat}$$

(onset of saturation)

This disappearance of the inversion layer at drain end is called channel **pinch-off**.

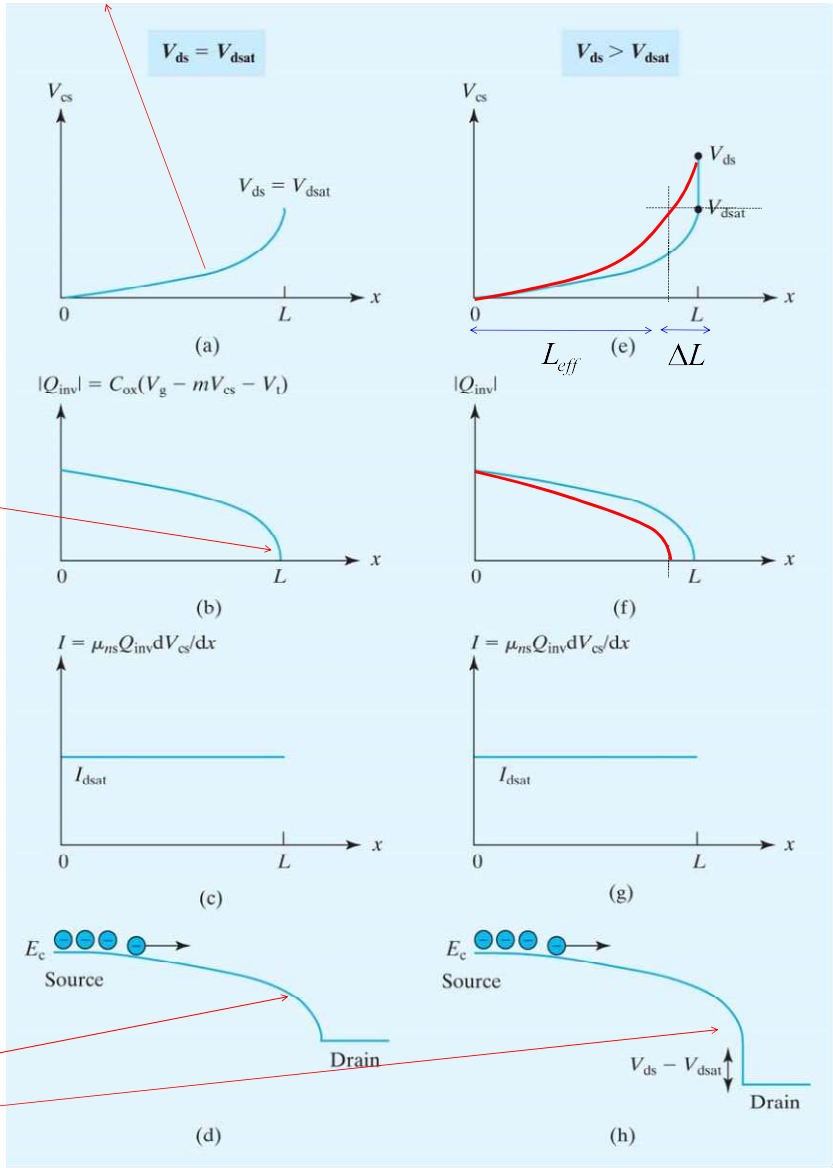
Region within L_{eff} ($V_{ds} < V_{dsat}$): low field region

Region within ΔL ($V_{ds} > V_{dsat}$): high field region

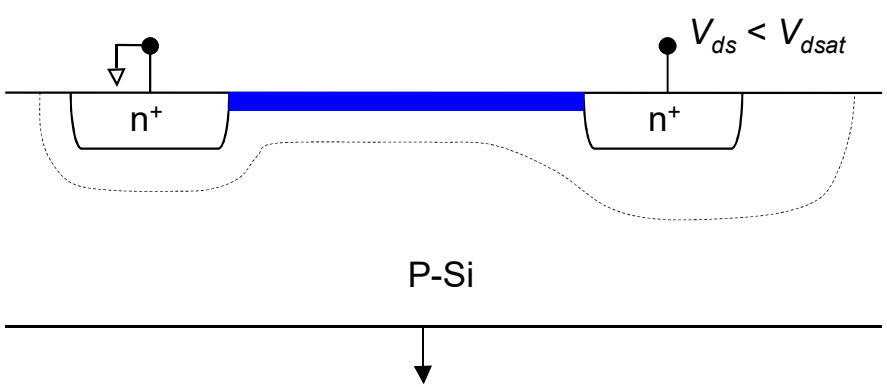
If $L >> \Delta L$, I_{ds} *does not change with* V_{ds} beyond V_{dsat}.

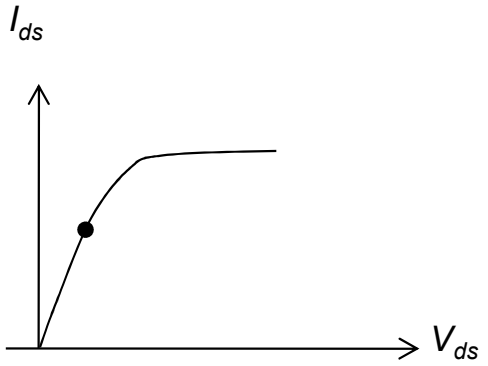
How can a current flow through the pinch-off region, which is similar to a depletion region?

When electrons reach the pinch-off region, they swept down the steep potential drop, but the current flow remains unchanged, as long as the number of supply electrons are equal. (**water fall analogy**)

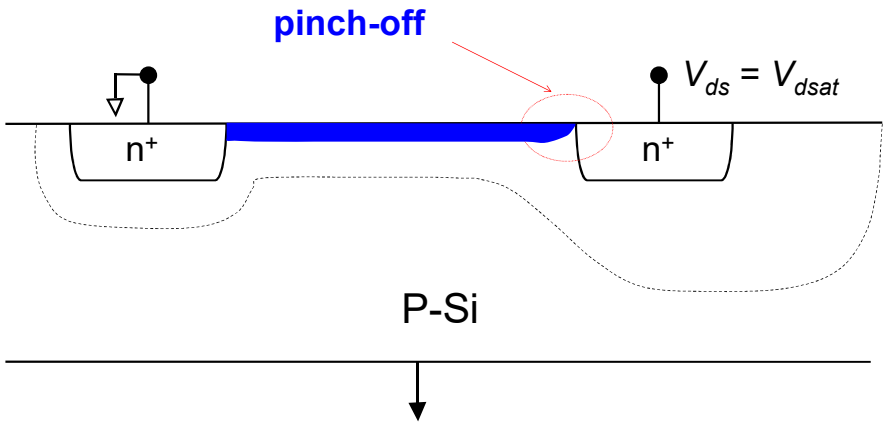


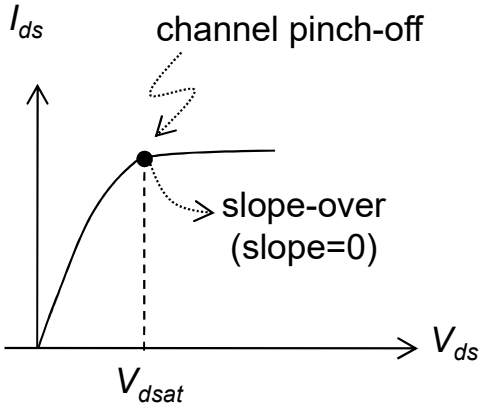
Linear region for $V_{gs} > V_t$ and $V_{ds} < (V_{gs}-V_t)/m$



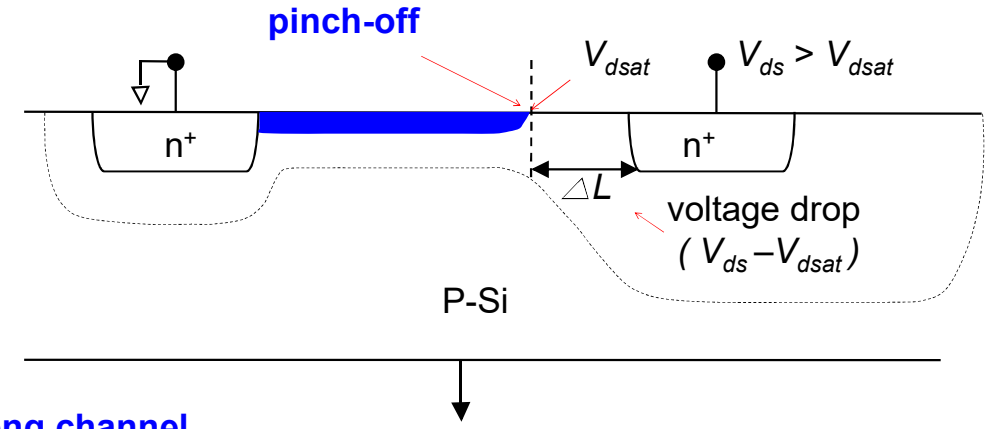


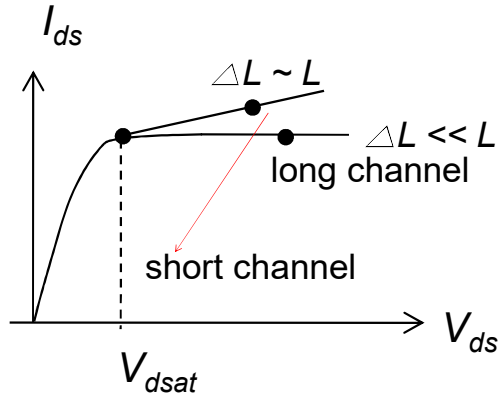
Onset of saturation at pinch-off, $V_{gs} > V_t$ and $V_{ds} = (V_{gs}-V_t)/m$





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





I_{ds} remains constant for long channel.
(the shape of a conductive region and potential applied across the region do not change.

channel length modulation for short channel

Transconductance, g_m

A measure of a transistor's sensitivity to the input voltage.

$$g_m \equiv \left.\frac{dI_{ds}}{dV_{gs}}\right|_{V_{ds}}$$

In saturation region,

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

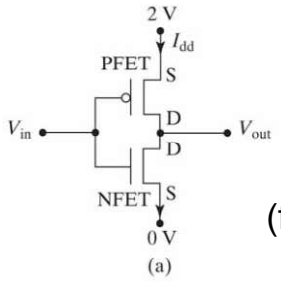
$$\therefore \; g_{msat} \equiv \frac{dI_{dsat}}{dV_{gs}} = \frac{W}{mL} C_{oxe} \mu_{ns} (V_{gs} - V_t)$$

CMOS Inverter-A Circuit Example

Little power consumption

Regenerating or cleaning up the digital signal

Voltage Transfer Curve(VTC) provides the important noise margin of the digital circuits.



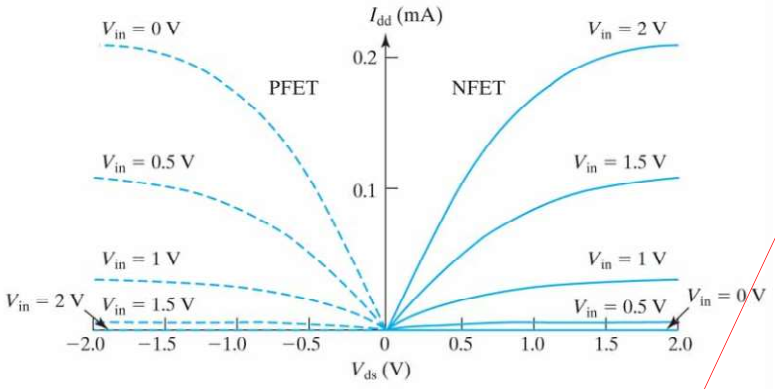
(a)

Assume that the PFET has identical(symmetric) *I-V*.

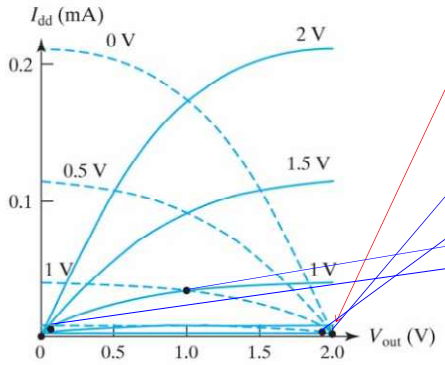
$V_{dsN} = V_{out}$

$V_{dsP} = V_{out} - 2\ V$

(for example,at $V_{out} = 2\ V$, $V_{dsN} = 2\ V$ and $V_{dsP} = 0\ V$.)



(b)

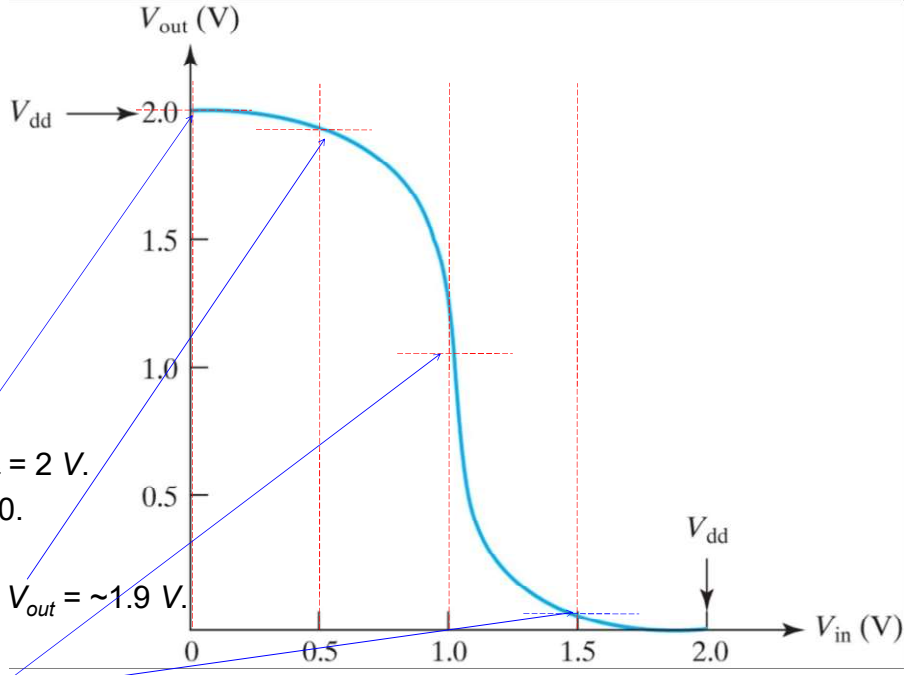


(c)

The two $V_{in} = 0$ curves intersect at $V_{out} = 2\ V$.
This means that $V_{out} = 2\ V$ when $V_{in} = 0$.

The two $V_{in} = 0.5\ V$ curves intersect at $V_{out} = \sim 1.9\ V$.

The two $V_{in} = 1\ V$ curves intersect at $V_{out} = \sim 1\ V$.



Voltage transfer curve(VTC) or **voltage transfer characteristics** of a CMOS inverter.

A **VTC** with a narrow and sharp transition near $V_{in} = V_{dd}/2$ region would maximize the noise tolerance.

V_{in} may be anywhere between 0 *V* and the NFET V_t and still produce a perfect $V_{out} = V_{dd}$.

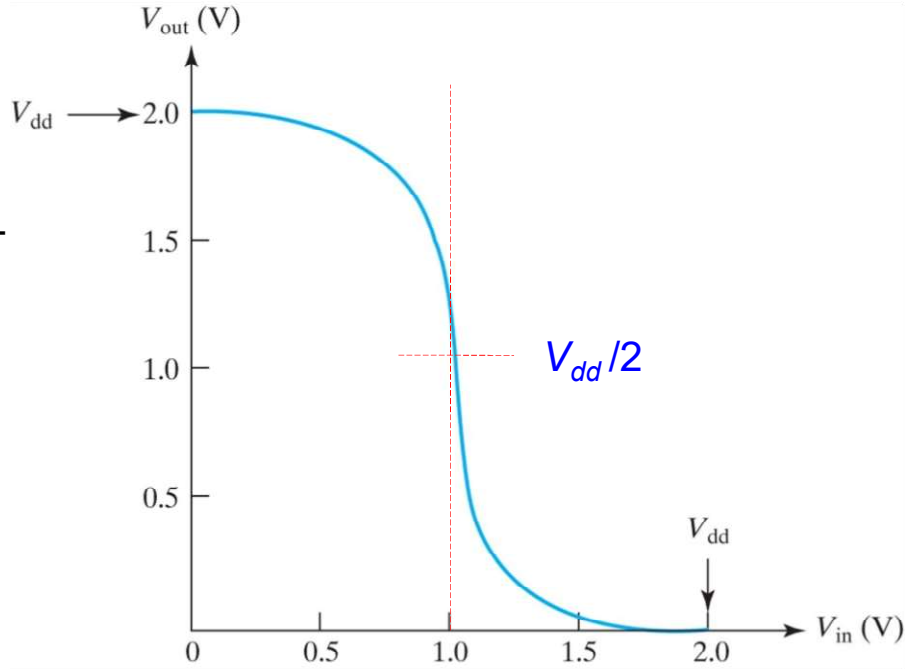
V_{in} may be anywhere between 2 *V* and 2 *V* plus the PFET V_t and still produce a perfect $V_{out} = 0$ *V*.

Therefore, perfect "0" and "1" outputs cane be produced by somewhat corrupted inputs. This regenerative property allows complex logic circuits to function properly in the face of inductive and capacitive noise and IR drops in the signal lines

Device characteristics include large g_m, low leakage in the off state, and a small $\partial I_{ds} / \partial V_{ds}$ in the saturation.

To achieve symmetry, the *I-V* curves of NFET and PFET need to be closely matched (symmetric) by choosing a larger W for the PFET than the NFET.

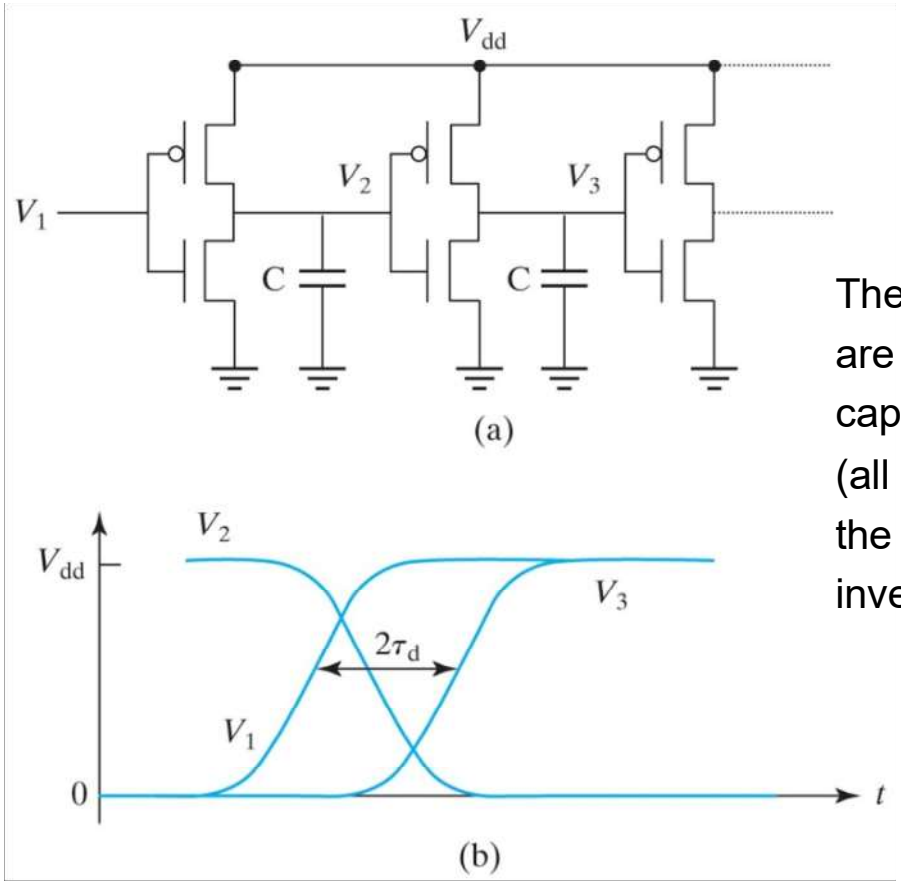
The W_P/W_N ratio is usually around two to compensate for the fact that μ_{ps} is smaller than μ_{ns} .



Inverter Speed-The Importance of I_{on}

Propagation delay, τ_d, is the time delay for a signal to propagate from one gate to the next in a chain of identical gates.

τ_d is the average of the delays of pull-down(rising V_1 pulling down the output, V_2) and pull-up(falling V_2 pulling up the output, V_3).



$$pull-down\ delay \approx \frac{CV_{dd}}{2I_{onN}},\quad pull-up\ delay \approx \frac{CV_{dd}}{2I_{onP}}$$

$$\tau_d = \frac{1}{2}(pull-down\ delay + pull-up\ delay)$$

The capacitance *C* represent the sum of all the capacitances that are connected to the output node of the inverter. They are the input capacitance of the next inverter in the chain.
(all the parasitic capacitances of the drain, and the capacitance of the metal interconnect that feeds the output voltage to the next inverter).

$$\therefore \tau_d \approx \frac{CV_{dd}}{4}\left(\frac{1}{I_{onN}} + \frac{1}{I_{onP}}\right)$$

On-state current, I_{on}

$$I_{on} \equiv I_{dsat}\Big|_{maximum|V_{gs}|} \Rightarrow \begin{cases} I_{onN} & taken\ at\ V_{gs} = V_{dd} \\ I_{onP} & taken\ at\ V_{gs} = -V_{dd} \end{cases}$$

The delay is the time for the on-state transistor supplying a current, I_{on}, to change the output by $V_{dd}/2$ (not V_{dd}).

The charge drained from (or supplied to) C by the FET during the delay is $CV_{dd}/2$.

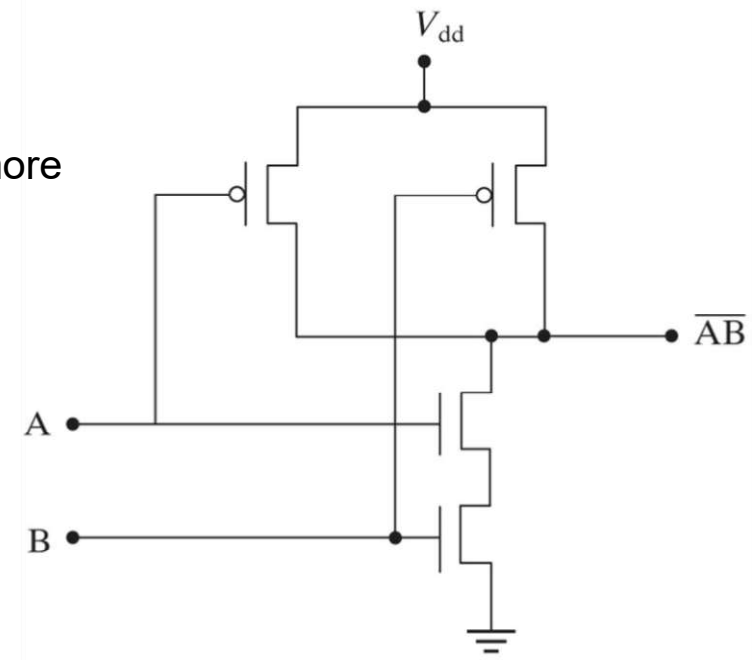
$$\therefore delay = \frac{Q}{I} = \frac{CV_{dd}}{2I_{on}} = RC, where\ R = \frac{V_{dd}}{2I_{on}} : switching\ resistance$$

In order to maximize circuit speed it is clearly important to maximize I_{on}.

Although the inverter is a very simple circuit, it is the basis of other more complex logic gate and memory cell.

For example, two-input NAND gate.

circuit with two series transistor in the pull-down path and two parallel transistor in the pull-up path.



Power Consumption

Dynamic power consumption

In each switching cycle, a charge CV_{dd} is transferred from the power supply to the load ,**C**.

Charge taken from the power supply in each second (average current provided by the power supply)

$= kCV_{dd} f$ where **f** is the clock frequency and **k** (<1) is an **active factor** that represent the fact that a particular gate in a given circuit is not switched every clock cycle all the time.

$$\therefore P_{dynamic} = V_{dd} \times (average\ current) = kCV_{dd}^2 f$$

*Power consumption can be reduced by **lowering V_{dd}** and by **minimizing all capacitance** in the circuit as well as by reducing k.* It is interesting to note that *making I_{on} large by using a small L or improving the carrier mobility does not increase $P_{dynamic}$.*

Large I_{on}: to reduce circuit switching delay.
Low V_{dd}: to reduce circuit power consumption.

Reducing the transistor **L** and **W** (smaller device) would lower **C** (gate capacitance, source-drain junction capacitance, and interconnect capacitance)

Static power consumption

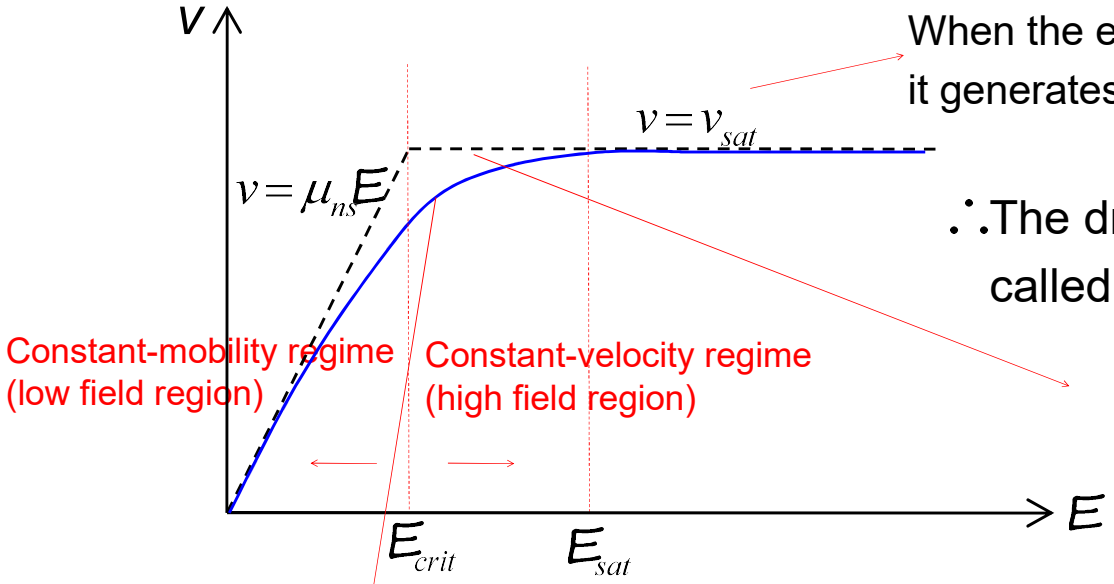
Another component of power consumption, is the **static power** or **leakage power** or **stand-by power** that is consumed when the inverter is static.

$$P_{static} = V_{dd} I_{off}$$

$$Total\ \ power\ \ consumption, P_{total} = P_{static} + P_{dynamic}$$

Velocity Saturation

A major weakness of the basic MOSFET *I-V* model is that a finite current flows through the pinch-off region, where $Q_{inv} = 0$. This requires the carrier velocity to be infinite, a physical impossibility.



When the energy of a carrier exceeds the optical phonon energy, it generates an optical phonon and loses much of its velocity

∴The drift velocity cannot exceeds a certain value, called the **saturation velocity, v_{sat}.**

piece-wise model

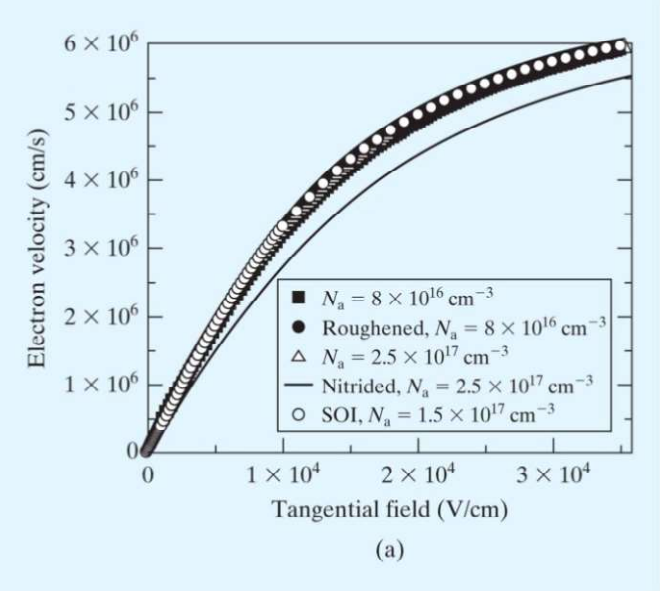
$$v = \frac{\mu_{ns} E}{1 + \mu_{ns} E / \mu_{ns} E_{sat}} = \frac{\mu_{ns} E}{1 + E/E_{sat}} = \frac{\mu_{ns} E}{1 + \mu_{ns} E / v_{sat}} \Rightarrow \begin{cases} \mu_{ns} E, & E << E_{sat} \ (at\ low\ field) \\ v_{sat}, & E >> E_{sat} \ (at\ high\ field) \end{cases}$$

Velocity saturation has a large and deleterious effect on the I_{on} of MOSFET.

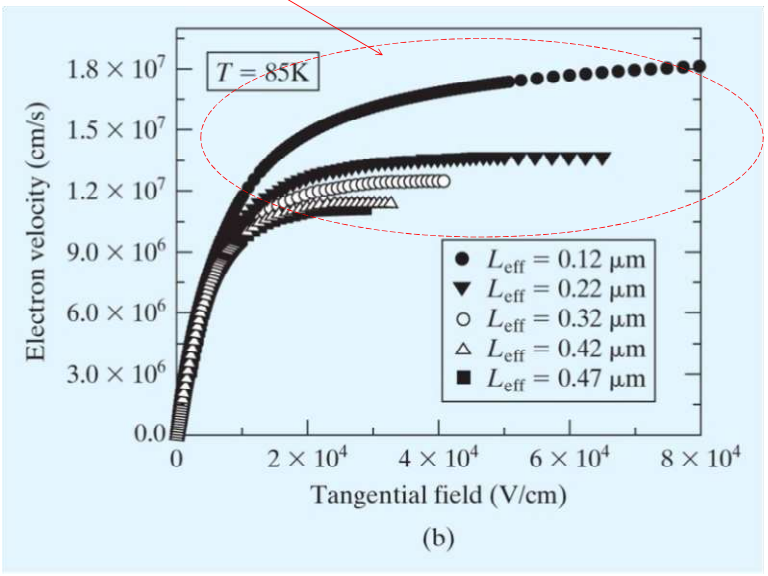
Velocity Overshoot

When *L* is sufficiently small, electrons may pass through the channel in too short a time for all the energetic carriers to lose energy by emitting optical phonons. As a result, the carriers can attain somewhat higher velocities in very small devices.

↳ This phenomenon is called **velocity overshoot**.
(In the basic velocity-saturation model, V_{sat} is independent of *L*.)



The inversion-layer electron velocity saturates at high field regardless of the body doping concentration and surface treatment. (From [8]. © 1997 IEEE.)



Velocity saturation is more prominent at low temperature. Velocity overshoot is also evident. (From [8]. © 1997 IEEE.)

MOSFET I-V Model with Velocity Saturation

$$I_{ds} = -W \cdot Q_{inv}(x) \cdot v = -W \cdot Q_{inv}(x) \frac{\mu_{ns} E}{1 + E / E_{sat}} = W C_{oxe} (V_{gs} - m V_{cs} - V_t) \frac{\mu_{ns} dV_{cs} / dx}{1 + \frac{dV_{cs}}{dx} / E_{sat}}$$

$$\Rightarrow \int_0^L I_{ds} dx \left(1 + \frac{dV_{cs}}{dx} / E_{sat} \right) = \int_0^{V_{ds}} W C_{oxe} \mu_{ns} (V_{gs} - m V_{cs} - V_t) dV_{cs}$$

$$\Rightarrow \int_0^L I_{ds} dx + \int_0^{V_{ds}} I_{ds} / E_{sat} dV_{cs} = \int_0^{V_{ds}} W C_{oxe} \mu_{ns} (V_{gs} - m V_{cs} - V_t) dV_{cs}$$

$$\Rightarrow \int_0^L I_{ds} dx = \int_0^{V_{ds}} [W C_{oxe} \mu_{ns} (V_{gs} - m V_{cs} - V_t) - I_{ds} / E_{sat}] dV_{cs}$$

$$\Rightarrow I_{ds} L + \frac{I_{ds} V_{ds}}{E_{sat}} = W C_{oxe} \mu_{ns} (V_{gs} - V_t - \frac{m}{2} V_{ds}) V_{ds}$$

$$\therefore I_{ds} = \frac{\frac{W}{L} C_{oxe} \mu_{ns} (V_{gs} - V_t - \frac{m}{2} V_{ds}) V_{ds}}{1 + \frac{V_{ds}}{E_{sat} L}} = \frac{long - channel \ I_{ds}}{1 + \frac{V_{ds}}{E_{sat} L}}$$

When L is large, denominator ~1 and the current reduces to long channel expression.

The effect of velocity saturation is to reduce I_{ds} by a factor of $1 + V_{ds} / E_{sat} L$.

The factor $1 + V_{ds} / E_{sat} L$ may be interpreted as $1 + E_{ave} / E_{sat}$, where $E_{ave} \equiv V_{ds} / L$ is the average field in the channel.

The saturation voltage, V_{dsat}, can be found by solving $dI_{ds}/dV_{ds} = 0$:

$$V_{dsat} = \frac{2(V_{gs} - V_t)/m}{1 + \sqrt{1 + 2(V_{gs} - V_t)/mE_{sat}L}}$$

A simpler and even more accurate V_{dsat} model may be derived from a piece-wise model:

$$v = \frac{\mu_{ns}E}{1 + E/E_{sat}} \quad for\ E \le E_{sat} \Rightarrow$$

leads to the previous I_{ds} equation, valid when the carrier speed is less than v_{sat}, *i.e.*, $V_{ds} \le V_{dsat}$.

$$v = v_{sat} \quad for\ E \ge E_{sat} \Rightarrow$$

leads to the following equation, describing the current at the drain end of the channel at the onset of velocity saturation (*i.e.*, $V_{ds} = V_{dsat}$):

$$(E_{sat} = 2v_{sat}/\mu_{ns},\ at\ E = E_{sat})$$

.

$$I_{ds} = -W \cdot Q_{inv}(L) \cdot v = WC_{oxe}(V_{gs} - mV_{cs}(L) - V_t)v_{sat}$$
$$= WC_{oxe}(V_{gs} - V_t - mV_{ds})v_{sat}$$

Equating two equation,

$$\frac{1}{V_{dsat}} = \frac{m}{V_{gs} - V_t} + \frac{1}{E_{sat}L}$$

$$I_{ds} = \frac{\frac{W}{L}C_{oxe}\mu_{ns}(V_{gs} - V_t - \frac{m}{2}V_{ds})V_{ds}}{1 + \frac{V_{ds}}{E_{sat}L}}$$

Substituting $\frac{1}{V_{dsat}} = \frac{m}{V_{gs} - V_t} + \frac{1}{\mathcal{E}_{sat} L}$ for V_{ds} in $I_{ds} = \frac{\frac{W}{L} C_{oxe} \mu_{ns} (V_{gs} - V_t - \frac{m}{2} V_{ds}) V_{ds}}{1 + \frac{V_{ds}}{\mathcal{E}_{sat} L}}$

$$I_{dsat} = \frac{W}{2mL} C_{oxe} \mu_{ns} \frac{(V_{gs} - V_t)^2}{1 + \frac{V_{gs} - V_t}{m\mathcal{E}_{sat} L}} = \frac{long - channel\ I_{dsat}}{1 + \frac{V_{gs} - V_t}{m\mathcal{E}_{sat} L}}$$

1. Long-channel or low V_{gs} case, $\mathcal{E}_{sat} L >> V_{gs} - V_t$

$$V_{dsat} = (V_{gs} - V_t) / m$$

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

identical to the long-channel model

2. Very short-channel case, $\mathcal{E}_{sat} L << V_{gs} - V_t$

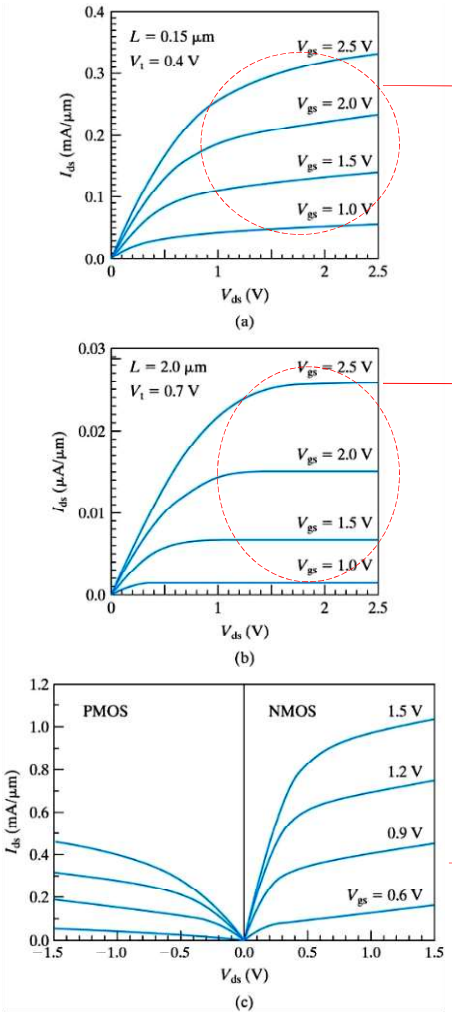
$$V_{dsat} \approx \mathcal{E}_{sat} L < (V_{gs} - V_t) / m$$

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

I_{ds} is proportional to W.

Carriers travel at the saturation velocity at the drain end of the channel where $Q_{inv} = C_{oxe}(V_{gs} - V_t - mV_{dsat})$ and V_{dsat} is $\mathcal{E}_{sat} L$.

$$
\begin{aligned}
I_{dsat} &= \frac{W}{2mL} C_{oxe} \mu_{ns} \frac{(V_{gs} - V_t)^2}{1 + \dfrac{V_{gs} - V_t}{m\mathcal{E}_{sat} L}} \\
&= \frac{W}{2mL} C_{oxe} \mu_{ns} \frac{(V_{gs} - V_t)^2}{\dfrac{V_{gs} - V_t}{m\mathcal{E}_{sat} L}\left(1 + \dfrac{m\mathcal{E}_{sat} L}{V_{gs} - V_t}\right)} = \frac{W}{2} C_{oxe} \mu_{ns} \frac{(V_{gs} - V_t)}{\dfrac{1}{\mathcal{E}_{sat}}}\left(1 + \frac{m\mathcal{E}_{sat} L}{V_{gs} - V_t}\right)^{-1} \\
&\approx \frac{W}{2} C_{oxe} \mu_{ns} \mathcal{E}_{sat} (V_{gs} - V_t)\left(1 - \frac{m\mathcal{E}_{sat} L}{V_{gs} - V_t}\right) = W C_{oxe} v_{sat} (V_{gs} - V_t)\left(1 - \frac{m\mathcal{E}_{sat} L}{V_{gs} - V_t}\right) \\
&= W C_{oxe} v_{sat} (V_{gs} - V_t - m\mathcal{E}_{sat} L)
\end{aligned}
$$



$I_{dsat} \propto (V_{gs} - V_t)$ for short channel MOSFET with finite output conductance(g_{ds}) at saturation (0.15 μm channel device with V_t = 0.4 V)

& less sensitive to L

V_{dsat} is significantly less than $(V_{gs} - V_t)$/m.

$I_{dsat} \propto (V_{gs} - V_t)^2$ for long channel MOSFET with zero output conductance at saturation (2 μm channel device with V_t = 0.7 V)

To raise I_{dsat} we must increase C_{oxe} $(V_{gs} - V_t)$:

- reduce T_{oxe} (limit of T_{oxe} is set by oxide tunneling leakage and reliability)
- minimize V_t (lower limit of V_t (is set by MOSFET leakage in the off state)
- use high V_{gs} (maximum V_{gs} is the power supply voltage, V_{dd}, limited by concerns over circuit power consumption and device reliability)

I-V characteristics of PFET and NFET with T_{oxe} = 3 nm and $L \approx$ 100 nm. Both exhibit a linear I_{dsat}-V_{dsat} relationship. I_P is half of I_N. The holes' mobility is three times smaller and their saturation velocity is 30 % smaller than that of the electrons.

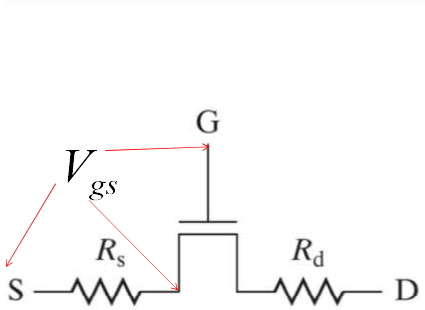
Velocity Saturation vs. Pinch-off

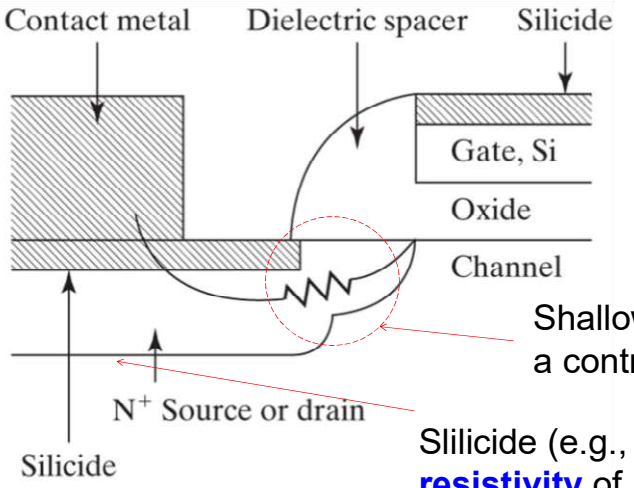
Concept of pinch-off : I_{ds} saturates when Q_{inv} becomes zero at the drain end of the channel.

Velocity saturation (more accurate description): I_{ds} saturates when the carrier velocity has reached v_{sat} at the drain. Instead of the pinch-off region, there is a velocity saturation region next to the drain where Q_{inv} becomes constant (I_{dst}/Wv_{sat}) at the drain end of the channel.

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 $R_s \neq 0$,

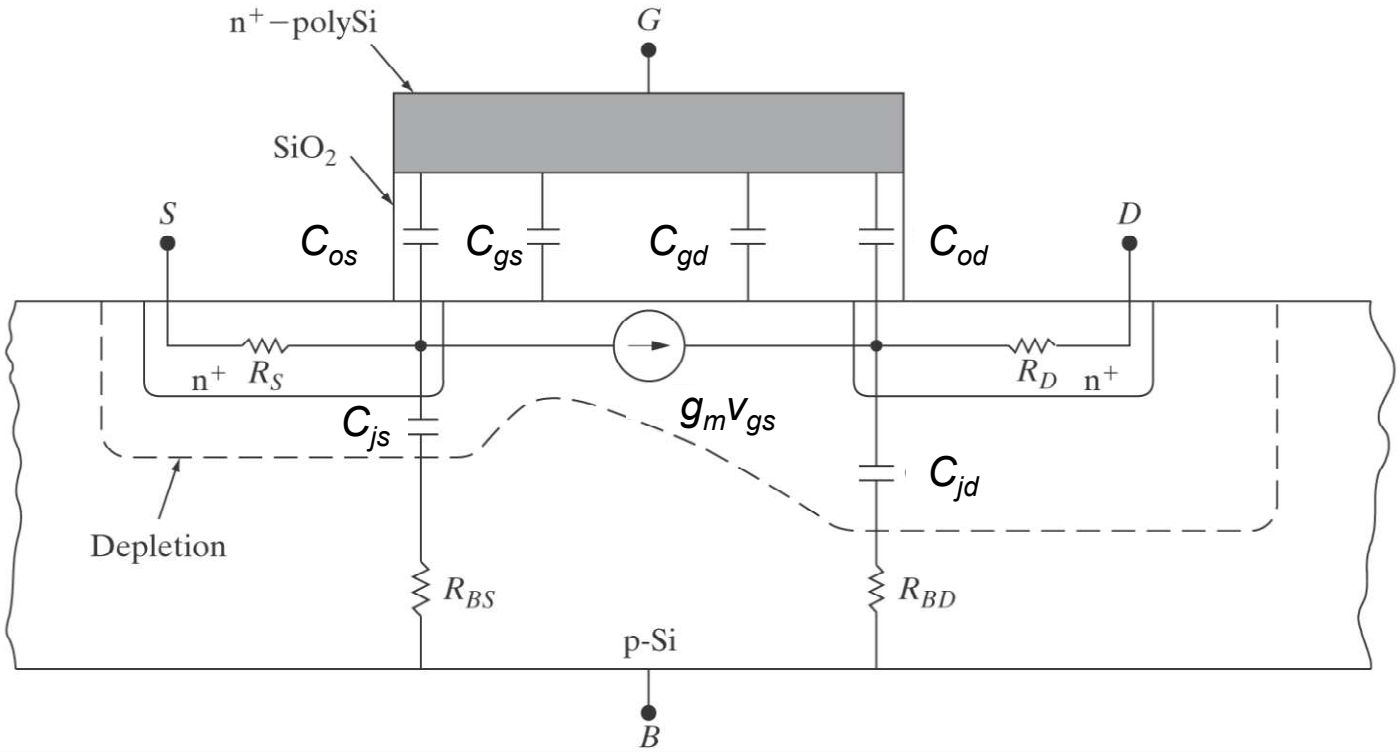
$$I_{dsat} \approx WC_{oxe} v_{sat} (V_{gs} - I_{dsat} R_s - V_t - m\mathcal{E}_{sat} L) = I_{dsat0} - WC_{oxe} v_{sat} I_{dsat} R_s \Rightarrow I_{dsat} (1 + WC_{oxe} v_{sat} R_s) = I_{dsat0}$$

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

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

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

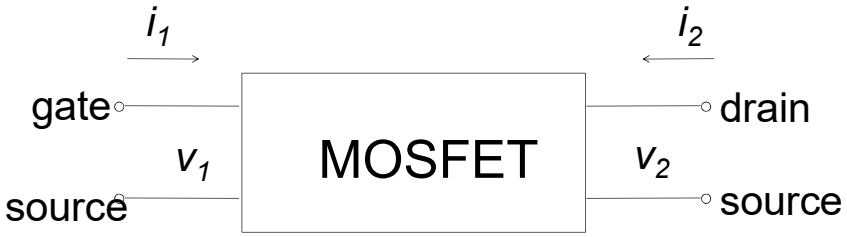
Equivalent Circuit for MOSFET



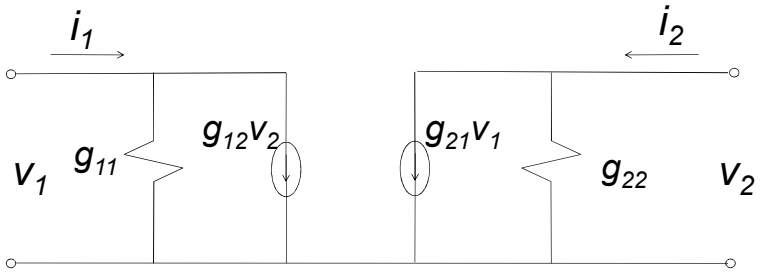
Miller overlap capacitance (C_{od}) due to the overlap between the gate and the drain region → can be reduced by self-aligned gate

Small Signal Equivalent Circuit

Consider a two-port network.



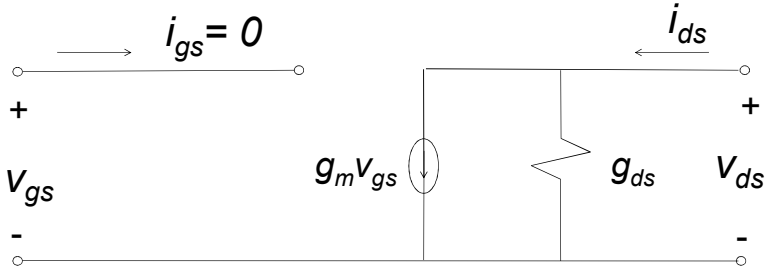
$$\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$



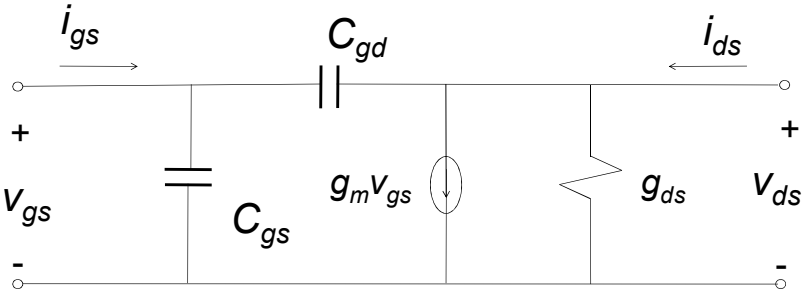
The input (gate) of MOSFET behaves like an open circuit.

$$i_1 = i_{gs} \approx 0, i_2 = i_{ds}, v_1 = v_{gs}, v_2 = v_{ds}, g_{11} \approx 0, g_{12} \approx 0, g_{21} = g_m, g_{22} = g_{ds}$$

At low frequency,

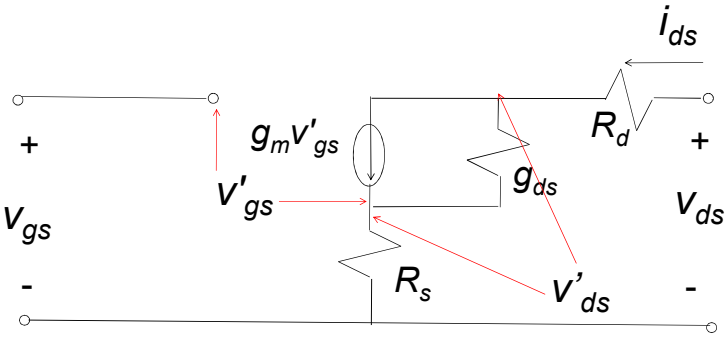


At high frequency,



Effect of R_s and R_d

With low frequency equivalent circuit,



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

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

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

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

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

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

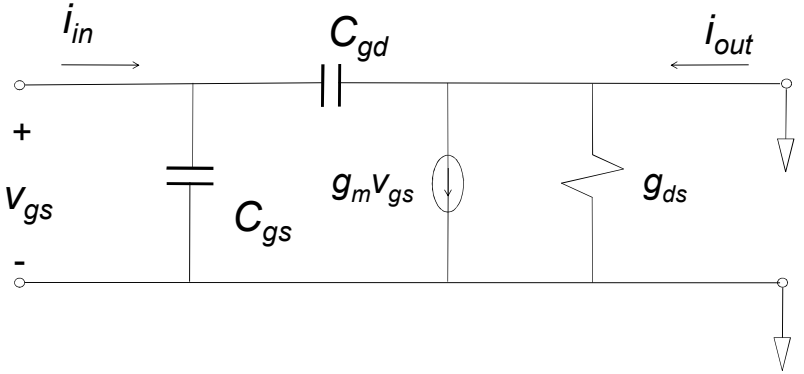
R_s and R_d affect g'_m and g'_{ds}.

Cutoff Frequency, f_T (unity current gain frequency)

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

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

With high frequency equivalent circuit,



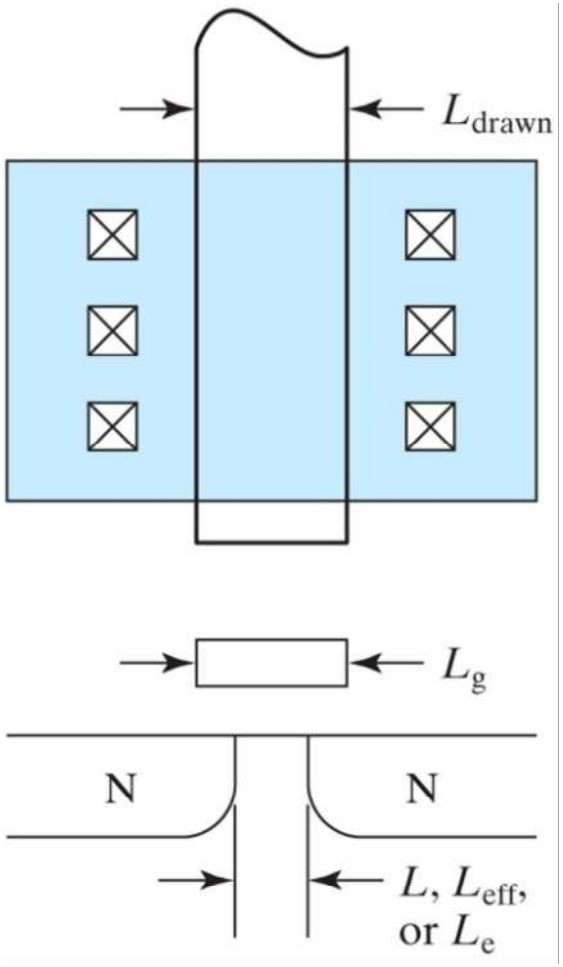
$$i_{in} = j\omega (C_{gs} + C_{gd}) W L_g v_{gs} \approx j(2\pi f) C_{ox} W L_g v_{gs}$$

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

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

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

Extraction of the Series Resistance and the Effective Channel length



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

$\approx$ L_g (patterned physical gate length)

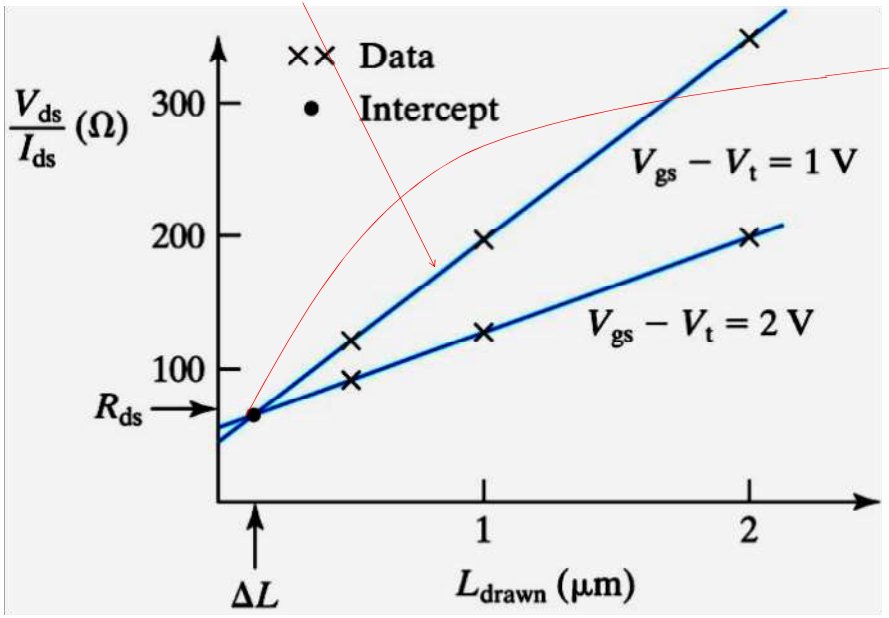
L (channel length),
L_{eff} (effective channel length), or
L_e (electrical channel length)
$\neq$ L_{drawn}

$$L = L_{drawn} - \Delta L,$$ where ΔL: the difference between L_{drawn} and L independent of L_{drawn}

For small V_{ds}, $I_{ds} = qnv = W \cdot Q_{inv} \cdot v = WC_{oxe}(V_{gs} - V_t)\mu_{ns} V_{ds} / L \quad \Rightarrow V_{ds} = \dfrac{I_{ds}(L_{drawn} - \Delta L)}{WC_{oxe}(V_{gs} - V_t)\mu_{ns}}$

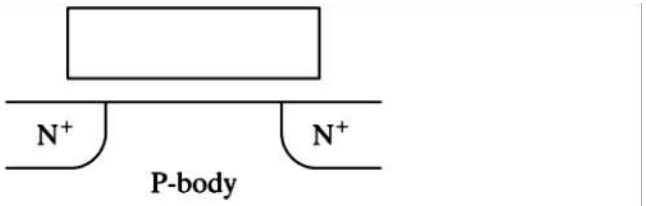
Including series resistance, $R_{ds} \equiv R_d + R_s, \Rightarrow V_{ds} = I_{ds} R_{ds} + \dfrac{I_{ds}(L_{drawn} - \Delta L)}{WC_{oxe}(V_{gs} - V_t)\mu_{ns}}$

$$\frac{V_{ds}}{I_{ds}} (= R_{ds} + channel\ \ resistance) = R_{ds} + \frac{(L_{drawn} - \Delta L)}{WC_{oxe}(V_{gs} - V_t)\mu_{ns}}$$



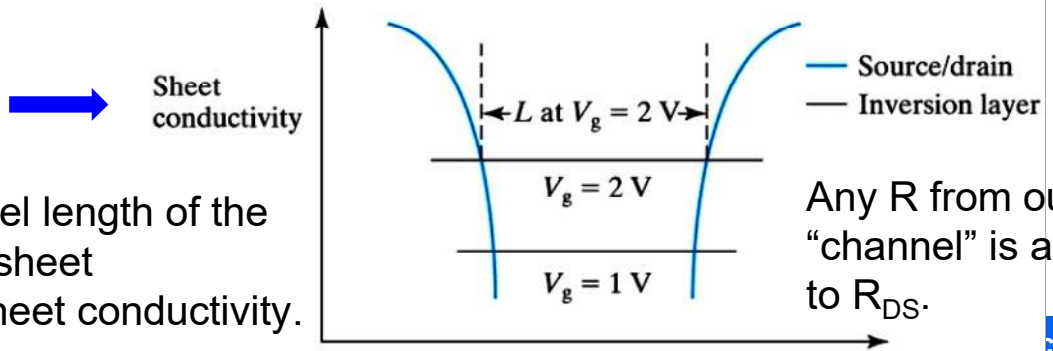
The two straight lines intersect at a point where V_{ds}/I_{ds} is independent of V_{ds} - V_t, *i.e.*, where $L_{drawn} = \Delta L$ and $V_{ds}/I_{ds} = R_{ds}$.

Once ΔL is known, L can be calculated from the equation; $L = L_{drawn} - \Delta L$,



Detailed measurements indicate;

R_{ds} tends to decrease and
L increases as V_g increases.



The channel can be interpreted as the channel length of the part if the channel where the inversion-layer sheet conductivity is larger than the source/drain sheet conductivity.

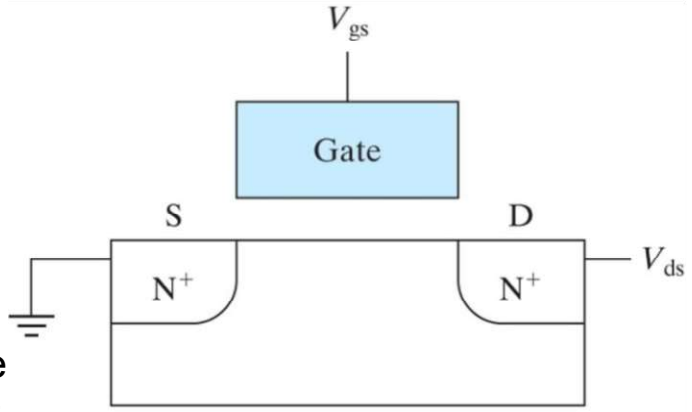
Any R from outside the "channel" is attributed to R_{DS}.

Velocity Overshoot and Source Velocity Limit

the carrier velocity at the drain end of the channel is limited by the saturation velocity which determines I_{dsat}.

$$I_{dsat} = WC_{oxe}v_{sat}(V_{gs} - V_t - mE_{sat}L)$$

$$v_{sat} = \begin{cases} 8\times10^6\ cm/s, & for\ electrons \\ 6\times10^6\ cm/s, & for\ holes \end{cases}$$



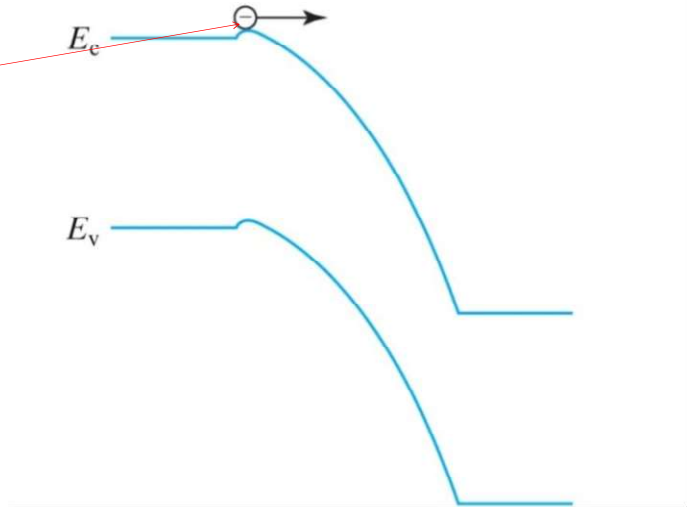
However, when the channel length is reduced much below 100 nm (when the channel length is comparable to or smaller than the mean free path), the saturation velocity may be greatly raised by velocity overshoot.

⬇

In that case, the velocity is limited by the thermal velocity, with which the carriers enter the channel from the source. This is known as the **source injection velocity** limit.

$$I_{dsat} = WBv_{thx}Q_{inv} = WBv_{thx}C_{oxe}(V_{gs} - V_t)$$

$$v_{thx} = \begin{cases} 1.6\times10^7\ cm/s, & for\ electrons \\ 1.0\times10^7\ cm/s, & for\ holes \end{cases}$$



B is the fraction of carriers captured by the drain. The rest of the injected carriers scattered back toward the source. (***B*** is ~ 0.5 from Monte Carlo simulation)

Considering the ***B*** value with v_{thx} and v_{sat}, above two equations predict similar I_{dsat}.

In the limit of no scattering in a very short channel, carriers are injected from the source into the channel at the thermal velocity and travel ballistically to the drain.

Output Conductance

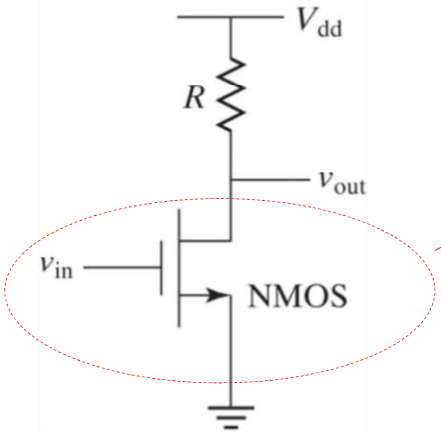
The slope of the *I-V* curve, $g_{ds} \equiv \frac{dI_{dsat}}{dV_{ds}}$

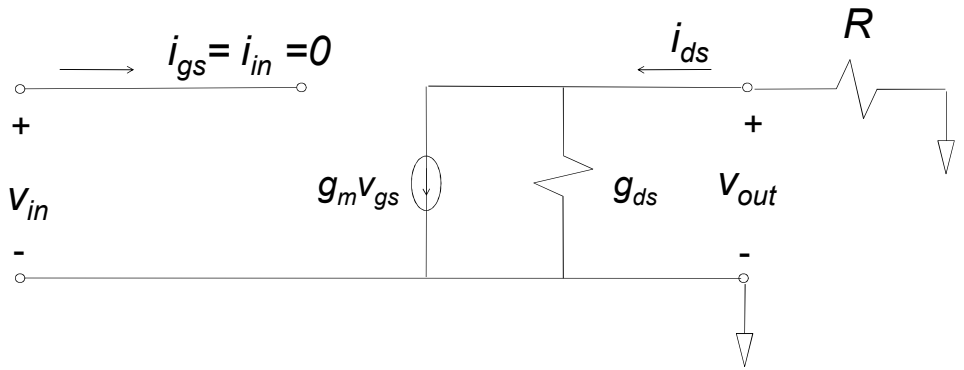
The physical causes of the output conductance:

1) influence of V_{ds} on V_t.
2) channel length modulation.

A clear saturation of I_{ds}, *i.e.*, a small g_{ds} is desirable. Why?

Consider an simple amplifier circuit.





$$i_{ds} = g_{msat} \cdot v_{gs} + g_{ds} v_{ds} = g_{msat} \cdot v_{in} + g_{ds} \cdot v_{out}$$

$$v_{out} = -R \times i_{ds}$$

The **intrinsic voltage gain**, or

$$Maximum\ Voltage\ Gain = \frac{g_{msat}}{g_{ds}}$$

$$\therefore v_{out} = \frac{-g_{msat}}{g_{ds} + 1/R} \times v_{in}$$

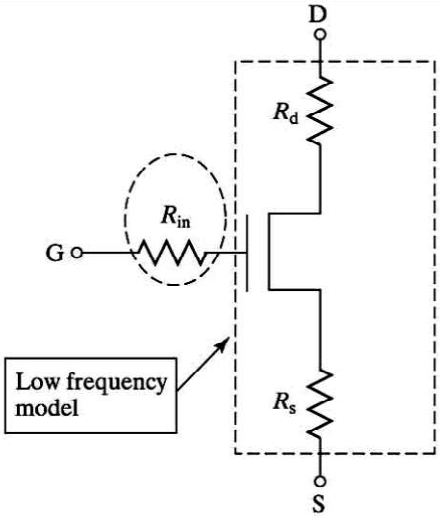
gain factor (can be increased by using a large R.)

g_{ds} must be kept much lower than g_{msat} for large gain.

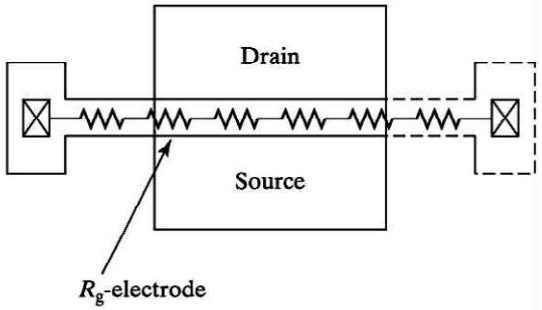
In order to achieve a small g_{ds} and a large voltage gain, L should be large and/or T_{ox}, W_{dep}, and X_j should be small.

1) Large gain is obviously beneficial to analog circuit application.
2) A reasonable large gain is also needed to obtain a steep transition in the VTC for digital applications to enhance noise immunity.

High-Frequency Performance



(a)

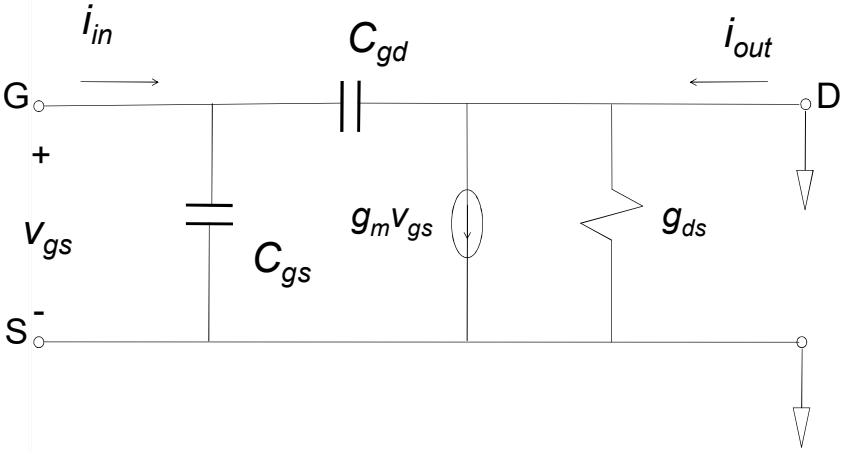


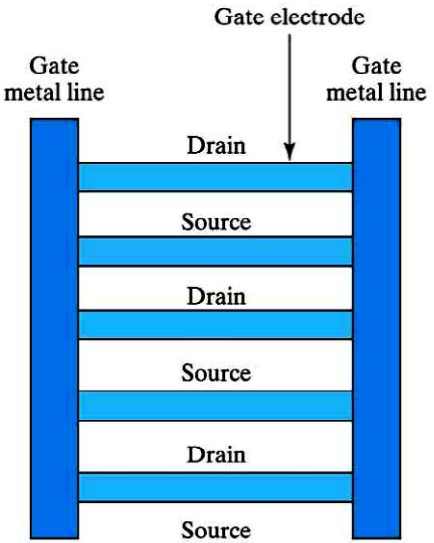
(b)

The high-frequency performance of the MOSFET is limited by the input RC time constant.

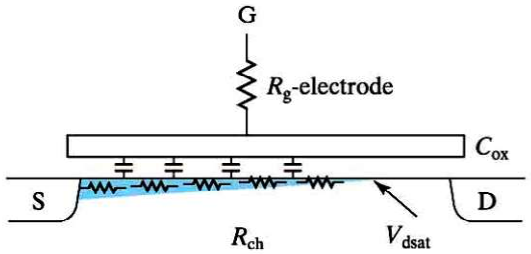
$C_{in} = C_{ox} W L_g$: gate capacitance

$R_{in} = R_{g\text{-}electrode} + R_{ii}$ (intrinsic input resistance)





(c)



(d)

(a) The input resistance together with the input capacitance sets the high-frequency limit.
(b) One component of R_{in} is the gate-electrode resistance.
(c) The multi-finger layout dramatically reduces the gate-electrode resistance.
(d) The more fundamental and important component of R_{in} is the channel resistance, which is also in series with the gate capacitor.

As frequency ↑, the gate capacitive impedance ($1/2\pi fC$) ↓, the gate AC current ↑, the gate signal voltage dropped across R_{in} ↑, the output current ↓.

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

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

At some frequency, $i_{out} = i_{in}$

This **unit current-gain** is called the **cutoff frequency, f_T.**

$$\left|\frac{i_{out}}{i_{in}}\right| = \frac{g_m v_{gs}}{2\pi C_{ox}WL_g v_{gs}}\Big|_{f=f_T} = 1 \quad \Rightarrow \quad \boxed{f_T = \frac{g_m}{2\pi C_{ox}WL_g}}$$

In narrow-band analog circuits operating at a particular high frequency, the gate capacitance may be compensated with an on-chip inductor at that frequency to overcome the f_T limit.

In that case, R_{in} still consumes power and at some frequency typically somewhat higher than f_T, the power gain drops to unity.

↳ This frequency is called the **maximum oscillation frequency, f_{max}.**

In either case, it is important to minimize R_{in}.

R_{in} consists of two components,
the gate-electrode resistance and the intrinsic input resistance.

$$\boxed{R_{in} = R_{g-electrode} + R_{ii}}$$

Gate-electrode Resistance

$$R_{g-electrode} = \frac{\rho W}{12 T_g L_g N_f^2}$$

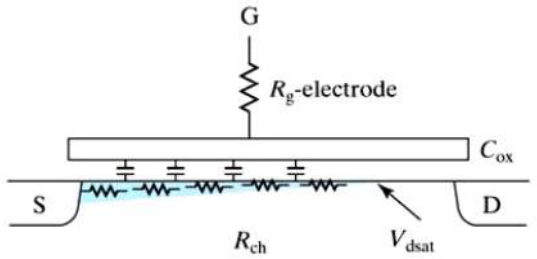
ρ : resistivity of the gate material
W : total channel width
T_g : gate thickness
L_g : gate length
N_f : number of fingers

➡ Using multi finger layout, the gate-electrode resistance can be quite low (if the gate material is silicided poly-silicon).

The factor of 12 comes from two sources:

1) A factor of 3 comes from the fact that the gate current is distributed over the finger width and all the gate capacitor current does not flow the entire finger resistor.
2) The remaining factor of 4 arises from contacting the gate fingers at both the left and the right ends of the fingers. Each finger is further divided into two at the middle of the finger.

Intrinsic Input Resistance



The more important, fundamental, and interesting component.

The gate capacitor current flows through the channel resistance, R_{ch}, to the source, then through the input signal source back to the gate to complete the current loop. R_{ii} is a resistance in the path of the gate current.

$$R_{ii} = \kappa \int dR_{ch} = \kappa \frac{V_{ds}}{I_{ds}}, \quad \kappa < 1$$

, because due to distributed nature of the *RC* network, the capacitance current does not flow through the entire channel resistance.

New generation of MOSFET technology:

L_g↓, R_{ii} ↓(due to larger I_{ds} and smaller V_{ds}) ➡ f_T and f_{max} > 200 GHz(45 nm technology node)
Input capacitance ($C_{ox}WL_g$)↓

MOSFET Noises

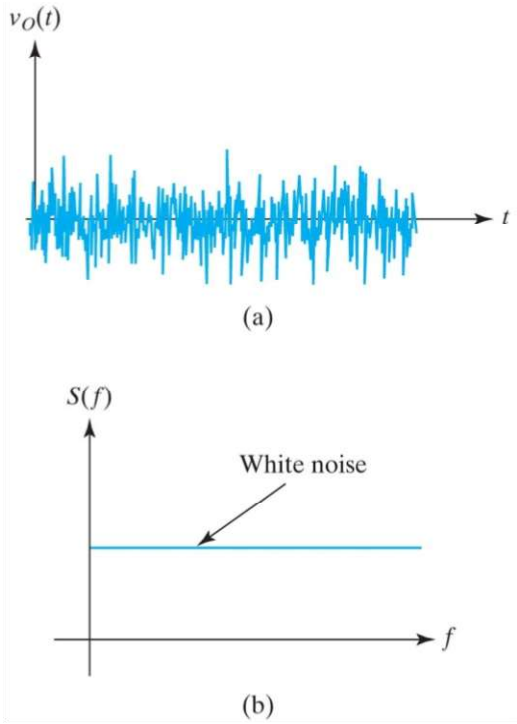
Cross talk: the inductive and capacitive interference created by the interconnect network.
(can be reduced by shielding and isolation by the circuit designer)

Device noise: inherent to the electronic devices
(due to the random behavior of the electric carriers)

Thermal Noise of a Resistor

The origin of the noise is the random thermal motion.

If a resistor is connected to the input of an oscilloscope, the noise voltage across the resistor can be observed as shown below.



(a) The thermal noise voltage across a resistor
(b) The spectral density of white noise.

The root-mean-square value of the noise voltage in the frequency band,

$$\overline{v_n^2} = S_{v_n}\Delta f = 4kT\Delta fR$$

Where R: resistance
Δf: bandwidth
S: noise power density

The noise current that would flow if the resistor's terminals were short-circuited.

$$\overline{i_n^2} = S_{i_n}\Delta f = 4kT\Delta f / R$$

$\overline{v_n^2}$ *and* $\overline{i_n^2}$ $\propto \Delta f$
but independent of f

This characteristic is called **white noise**.

MOSFET Thermal Noise ; intrinsic thermal noise originates from the channel resistance.

$$\overline{v_n^2} = S_{v_n}\Delta f = 4kT\Delta f R \Rightarrow \overline{v_{ds}^2} = 4\gamma kT\Delta f / g_{ds}$$

$$\overline{i_n^2} = S_{i_n}\Delta f = 4kT\Delta f / R \Rightarrow \overline{i_{ds}^2} = 4\gamma kT\Delta f g_{ds}$$

$g_{ds} = I_{ds}/V_{ds}$ in the linear region.
γ is fitting parameter and function of V_{ds} and V_{gs} ($\gamma \rightarrow 2/3$ at $V_{ds} > V_{dsat}$)

The noise current is added to the normal MOSFET current as a parallel current source.

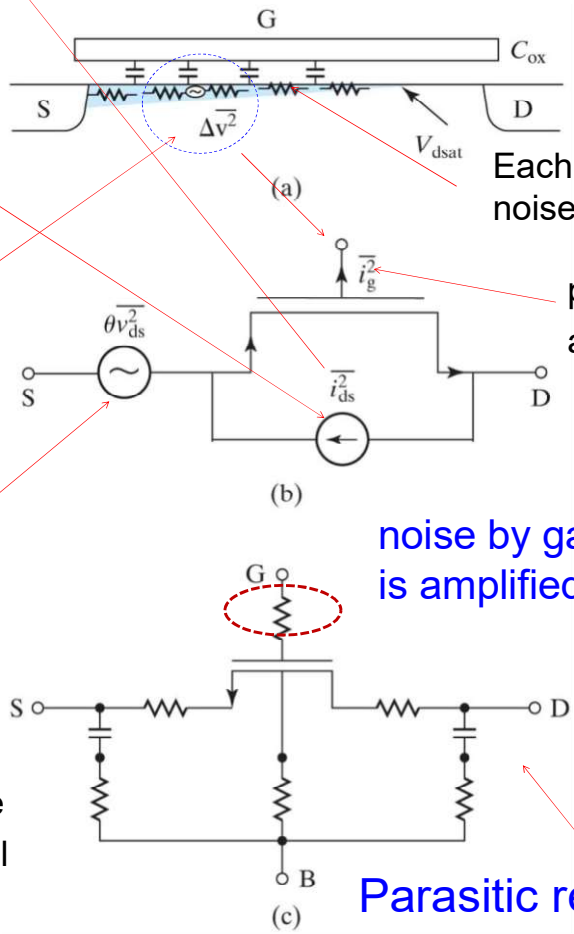
The noise voltage is multiplied by the transconductance into another component of noise current.

The channel noise voltage also induces a gate current through the gate capacitance

The gate noise current multiplied by the impedance of the gate input network and the transconductance produces a second noise current at the output..

approximately modeled by lumping the channel noise voltage at the source.

θ is a function of L and V_{gs} and accounts for the fact that the noise voltage is actually distributed throughout the channel rather than lumped at the source.



Each segment contributes thermal noise.

partially correlated noise source appearing at the gate terminal.

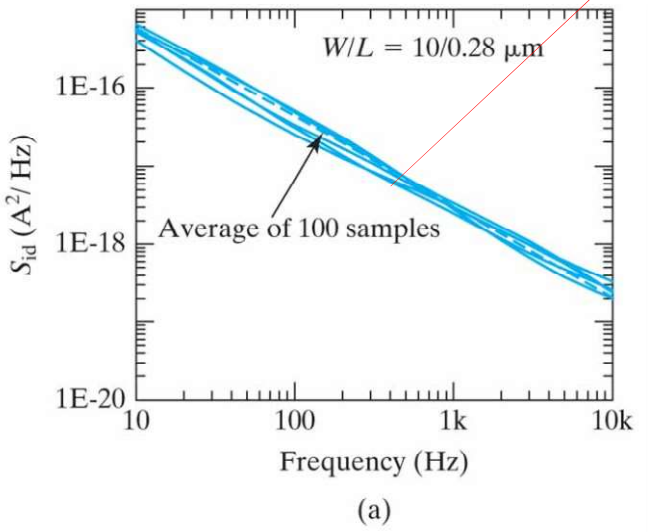
noise by gate electrode resistance is amplified by g_{msat} into I_{ds} noise

Minimized by multifinger layout

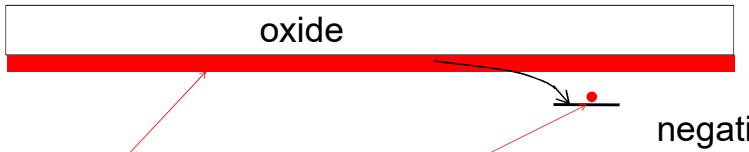
Parasitic resistances also contribute to the thermal noises.

MOSFET Flicker Noise (1/f Noise)

Flicker noise is also known as 1/f noise
because the noise power density is proportional to 1/frequency.

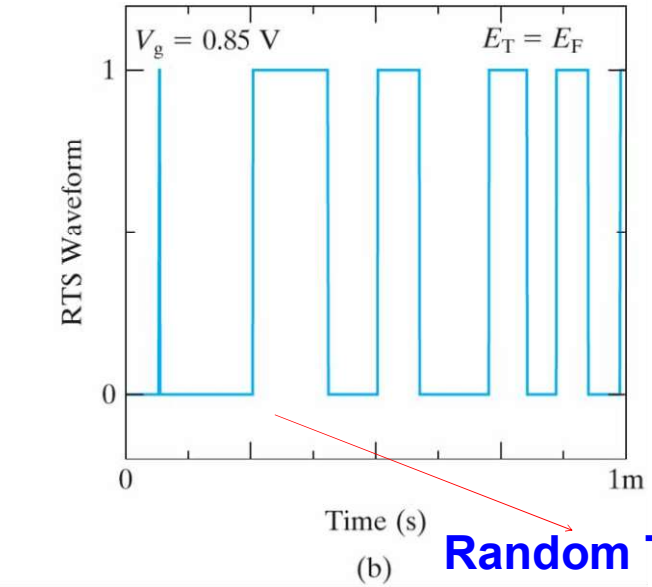


The mechanism for flicker noise is the random capture and release of electrons by traps located in the gate dielectric.



inversion channel with one less electron to conduct current

reduces the channel carrier mobility due to Coulombic scattering.

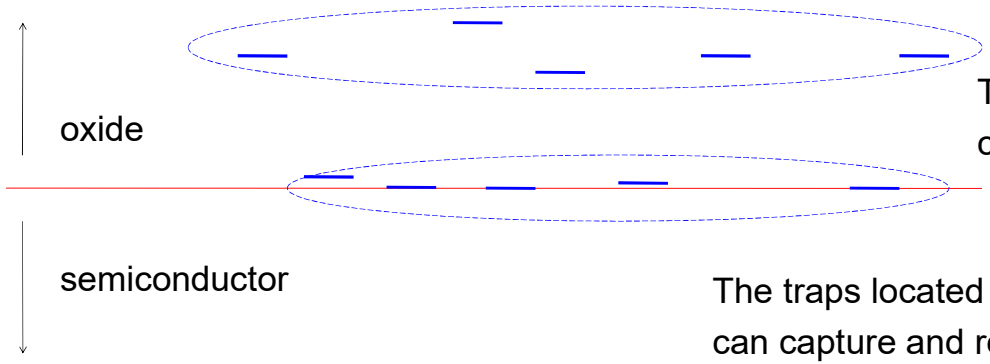


Both the carrier number and the mobility fluctuate due to charge trapping and detrapping.

Random Telegraph Noise.

In a MOSFET with very small W and L,
there may be only one operative trap and I_{ds} fluctuates between two levels.

In a large area ($W \times L$) MOSFET, there are many traps.



Traps located far from the interface have long time constants and contribute mostly low-frequency noises.

The traps located at or near the oxide-semiconductor interface can capture and release electrons with short tome constant and they contribute mostly high-frequency noises.

Assuming a uniform distribution of traps in the oxide leads to the 1/f noise spectrum,

$$\overline{i_{ds}^2} = \frac{KF \cdot W}{fL^2 C_{ox}} \left(\frac{I_{ds}}{W} \right)^{AF} \cdot kT\Delta f$$

KF is proportional to the oxide trap density
AF depends on the importance of Coulombic scattering to carrier mobility (1 ~ 2)

The flicker noise is the dominant noise at low frequency. At frequencies above 100 MHz, one can safely ignore the flicker noise as it is much smaller than the thermal noise.

Signal to Noise Ratio, Noise Factor, Noise Figure

Signal to Noise Ratio (SNR) a measure of the detectability of the signal in the presence of noise.

The SNR at the output of a linear device or circuit is smaller than the SNR at the input.

Noise Factor

$$F = \frac{S_i / N_i}{S_0 / N_0}$$

Noise can be minimized with an optimum gate network impedance.
Achieving this $N_{F\text{-}min}$ is an important goal of low-noise circuit design.

Noise Figure

$$N_F = 10 \times \log F \quad (dB)$$

The MOSFET noise is more relevant to analog circuits than digital circuits.

1) A linear circuit such as a linear amplifier;
must faithfully preserve the input waveform while amplifying its magnitude.
The SNR at the output is at best the same as the input.

2) A digital circuit such as an inverter;
eliminates the small noise at the input with its nonlinear VTC.
A digital circuit has no gain for the small-amplitude noise at the input and has gain only for the larger real digital signal.

SRAM, DRAM, Nonvolatile (Flash) Memory Devices

Three types of semiconductor memories:

Static RAM (SRAM): completely compatible with basic CMOS technology.

Dynamic RAM (DRAM): smaller than SRAM but requires some special fabrication steps.

Non volatile memory (with flash memory): employs one of a variety of physical mechanisms to perform nonvolatile storage and has even smaller size than DRAM.

RAM stands for random access memory meaning every data byte is accessible any time unlike hard disk memory, which has to move the read head and the disk to fetch new data with a significant delay.

Nonvolatile means that data will not be lost when the memory is disconnected from electrical power source.

TABLE 6–1 • The differences among three types of memories.

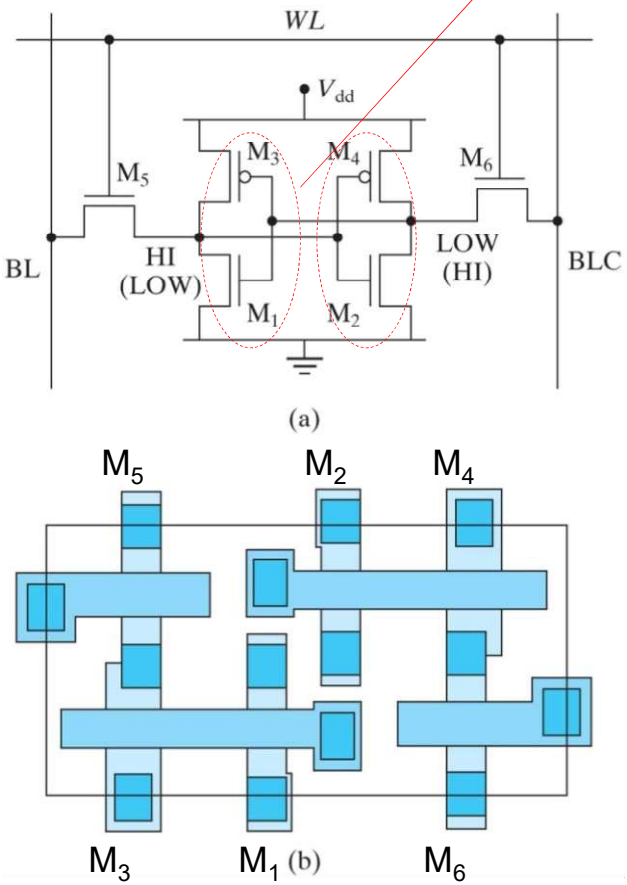
	Keep Data Without Power?	Cell Size and Cost/bit	Rewrite Cycles	Write-One-byte Speed	Compatible with Basic CMOS Manufacturing	Main Applications
SRAM	No	Large	Unlimited	Fast	Totally	Embedded in logic chips
DRAM	No	Small	Unlimited	Fast	Need modifications	Stand-alone chips and embedded
Flash memory	Yes	Smallest	Limited	Slow	Need extensive modifications	Nonvolatile storage stand-alone

SRAM

A basic SRAM cell uses six transistors to store one bit of data.

Two inverters and two pass transistors (M_5 and M_6).

The output of the left inverter is connected to the input of the right inverter and vice versa.



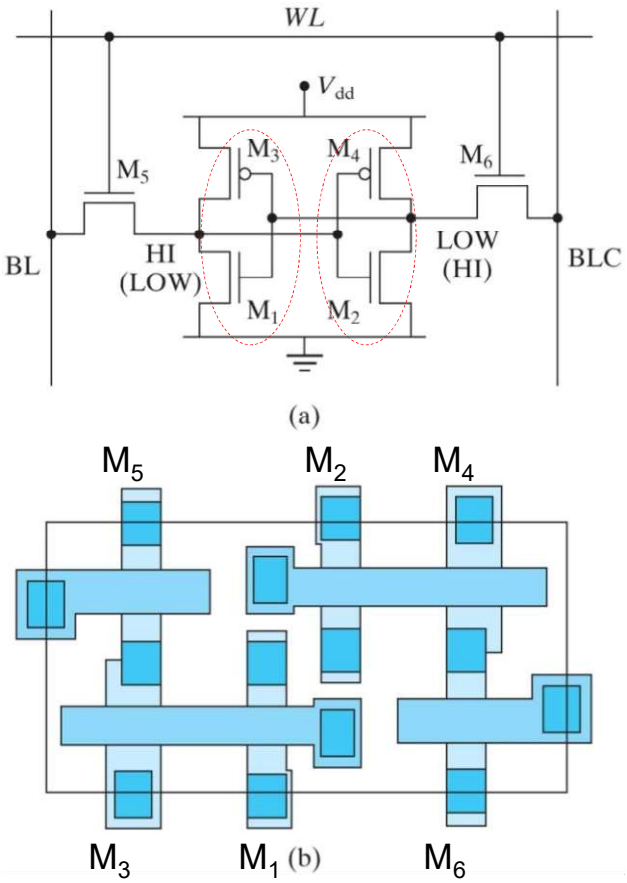
If the left-inverter output (input of the right inverter) is *high,* The right-inverter output would be *low*. This low output in turn makes the left-inverter out *high*.

↳ The positive feedback ensures that this state is stored and stable.

If we change the left-inverter output to *low* and the right-inverter output to *high*, that would be a second stable state.

This cell has two stable states, "0" and "1", and can store one bit of data.

(a) Schematic of an SRAM cell.
(b) Layout of a 32 nm technology SRAM. (From [16]. © 2007 IEEE.) The dark rectangles are the contacts. The four horizontal pieces are the gate electrodes and the two PFETs have larger Ws than the four NFETs. Metal interconnects (not shown) cross couple the two inverters.



In order to read the stored data (determine the inverter state), the selected cell's WL is raised to turn on the pass transistor.

A sensitive **sense amplifier** circuit compares the voltage on BL and BLC to determine the stored state.

In order to write the left-low state into the cell, BL is set to low and BLC is set to high. Next, the word-line voltage is raised and the inverter will be forced to this (new) state.

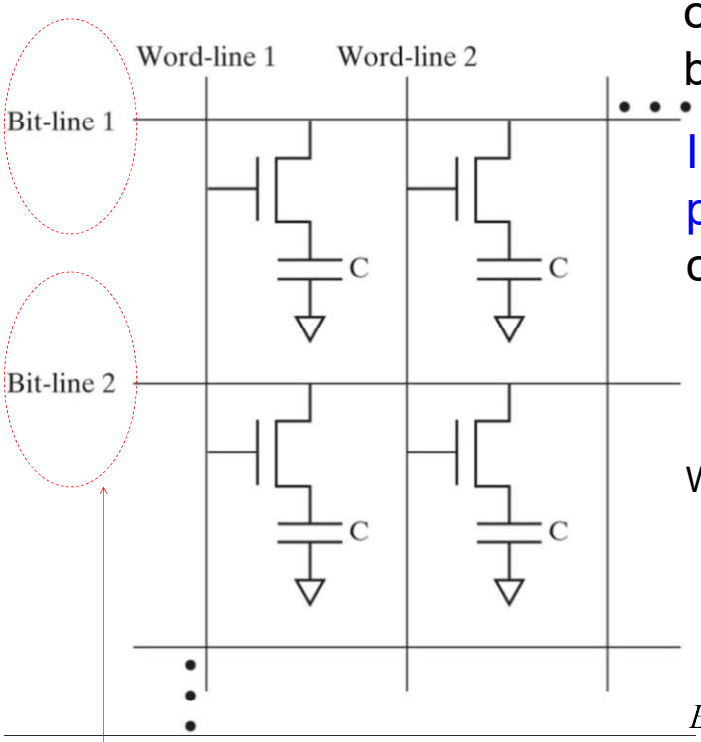
SRAM cells provides the fastest operation among all memories. But since it requires six transistors to store one bit of data.

↳ used as cache memory in a processing unit where speed is critical.

(a) Schematic of an SRAM cell.
(b) Layout of a 32 nm technology SRAM. (From [16]. © 2007 IEEE.) The dark rectangles are the contacts. The four horizontal pieces are the gate electrodes and the two PFETs have larger Ws than the four NFETs. Metal interconnects (not shown) cross couple the two inverters.

DRAM

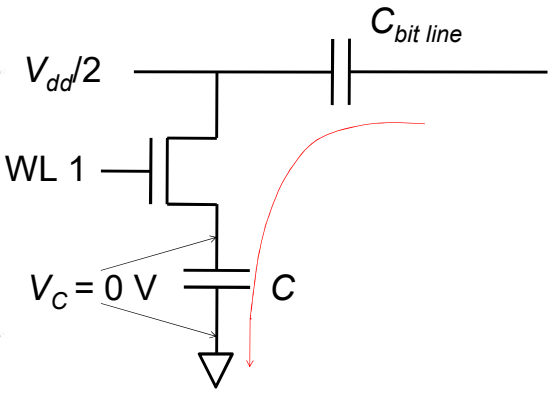
A DRAM cell consists of a transistor and a capacitor and can provide a large number of bits per area and lower cost per bit.



Each bit line has its own (unavoidable) capacitance, $C_{bit\ line}$.

To write data into the upper-left cell, WL 1 is raised high to turn on the transistor (connecting the capacitor to bit line 1) and bit line 1 is set to V_{dd} to write "1" or 0 V to write "0".

In order to read the stored data from the upper-left cell, bit line 1 is precharged to $V_{dd}/2$ and then left floating. WL 1 voltage is raised to connect the cell capacitor in parallel with the larger $C_{bit\ line}$.



Before WL 1 voltage is raised,

$$Q_{total} = \frac{V_{dd}}{2} \cdot C_{bit\ line}$$

After WL 1 voltage is raised,

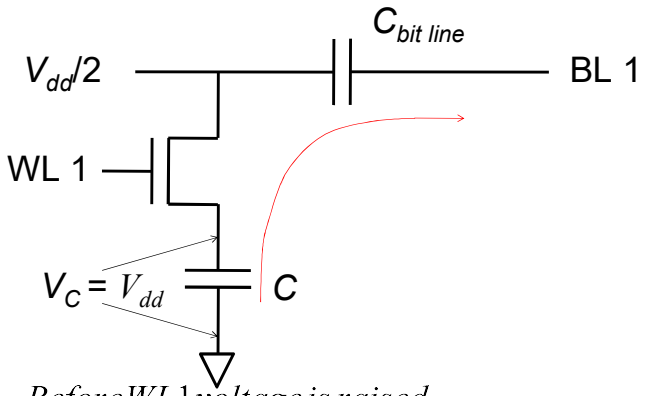
$$Q_C = \frac{V_{dd}}{2} \cdot C_{bit\ line} \frac{C}{C + C_{bit\ line}}$$

$$V_C = \frac{Q_C}{C} = \frac{V_{dd}}{2} \frac{C_{bit\ line}}{C + C_{bit\ line}}$$

∴ Bit line voltage is lowered by

$$\frac{V_{dd}}{2} - \frac{V_{dd}}{2} \frac{C_{bit\ line}}{C + C_{bit\ line}} = \boxed{\frac{V_{dd}}{2} \frac{C}{C + C_{bit\ line}}}$$

usually tens of milivolts.



Before WL 1 voltage is raised,

$$Q_C = C \cdot V_{dd}, \quad Q_{bit\ line} = C_{bit\ line} \cdot \frac{V_{dd}}{2}, \quad Q_t = Q_C + Q_{bit\ line}$$

After WL 1 voltage is raised,

$$Q_C = Q_t \frac{C}{C + C_{bit\ line}} = V_{dd}(C + \frac{C_{bit\ line}}{2}) \frac{C}{C + C_{bit\ line}}$$

$$V_C = \frac{Q_C}{C} = V_{dd}(C + \frac{C_{bit\ line}}{2}) \frac{1}{C + C_{bit\ line}}$$

∴ Bit line voltage is raised by

$$V_{dd}(C + \frac{C_{bit\ line}}{2}) \frac{1}{C + C_{bit\ line}} - \frac{V_{dd}}{2} = \boxed{\frac{V_{dd}}{2} \frac{C}{C + C_{bit\ line}}}$$

A sense amplifier circuit connected to the bit line monitors this voltage change to determine (read) the stored data.

The DRAM capacitor can only hold the data for a limited time because its charge gradually leaks through the capacitor dielectric,
the PN junction(S/D),
the transistor subthreshold leakage,…..

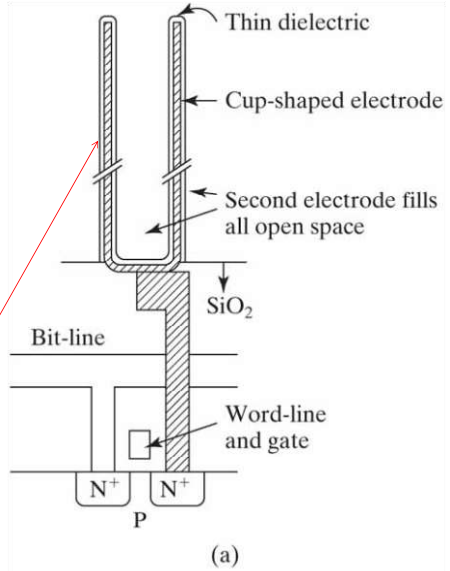
To prevent data loss, the charge must be refreshed (read and rewritten) many times each second.

(The D in DRAM refers to this dynamic **refresh** action. Refresh consumes stand-by power.)

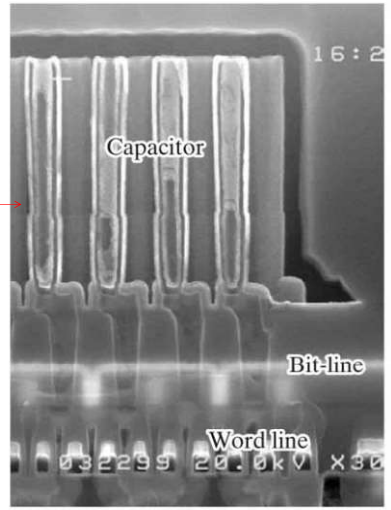
To increase the refresh interval,
the cell capacitance should be large so that more charge is stored.

A larger cell capacitance (not too much smaller than $C_{bit\ line}$) is also important for generating a large read signal for fast and reliable reading. However, it has become increasingly difficult to provide a large C while the cell area has been reduced to a few percent of 1 μm^2.

Use very thin capacitor dielectrics with complex three-dimensional capacitor structures.



(a)

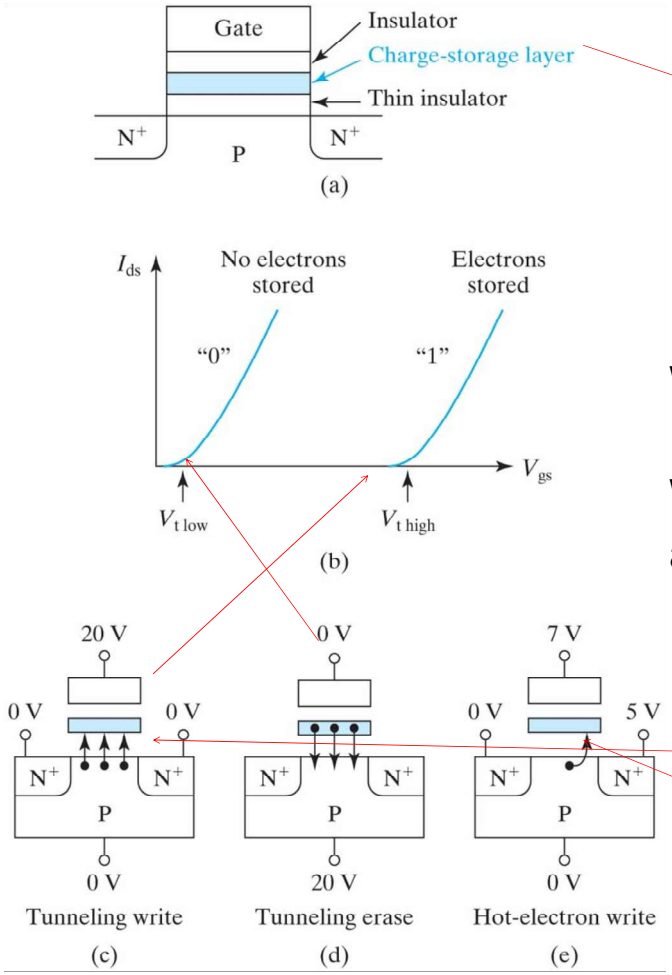


(b)

(a) Schematic drawing of a DRAM cell with a cup-shaped capacitor.
(b) Cross-sectional image of DRAM cells. The capacitors are on top and the transistors are near the bottom. (From [17]. © 2002 IEEE.)

Nonvolatile (Flash) Memory

Nonvolatile memory or **NVM** is a memory device that keeps its content without power for many years. NVMs are used for program **code storage** in cell phones and most microprocessor based systems, **data storage** medium (over hard disks and CDs) in portable applications.



silicon nitride, another insulator with high density of electron traps, or floating conductor (polycrystalline Si)

floating-gate memory

When the traps are empty or neutral, the transistor has a low V_t ("0").
When electrons are trapped in the insulator, the transistor has a high V_t ("1").

Writing
- electron injection by tunneling:
  slow(performed on hundreds of bytes at the same time)
- electron injection by hot-electron injection (HCI):
  fast, but more current and power required

Because the erase operation by tunneling is slow (taking milliseconds compared to nano-second for SRAM and DRAM), NVMs are erased in blocks of kilobytes rather than byte by byte.

↳ This electrical erase by large memory blocks us called **flash erase** and this type of memory is called **flash memory**.

Limitation of the flash memory

repeated write and erase cycling under high-electric field
can break chemical bonds in the insulator and
creates leakage paths with diameters of a few atoms and at random locations.

This sets an NVM endurance limit of less than 10^6 write/erase cycles.

Charge-trap NVM

If the floating gate is replaced with a dielectric film containing many isolated electron traps or isolated nanocrystals of metal or semiconductor, one leakage path can only discharge a fraction of the stored electrons in the cell. Endurance may be improved. They are called charge-trap NVM and the nano-crystal NVM.

Multilevel cell

The NVM cell is simple and small even in comparison with DRAM cell and, therefore, can store large numbers of bits per centimeter square than DRAM and SRAM.

It is possible to write and store more than two V_t values in a flash memory cell by controlling the number of stored electrons.

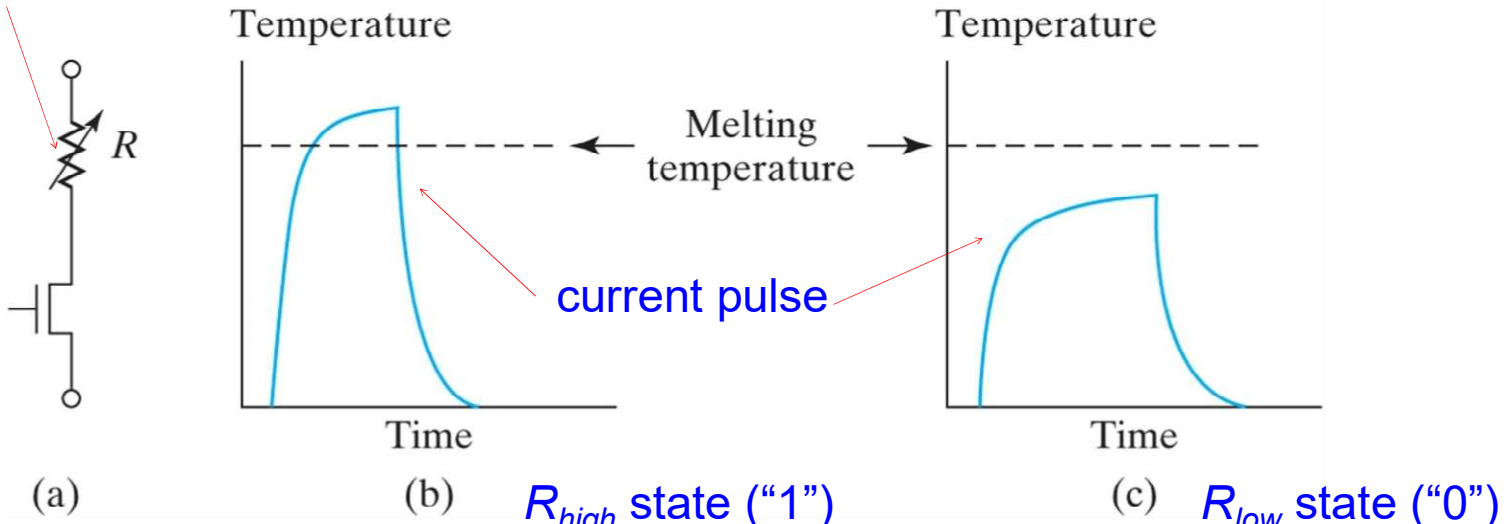
Two V_ts can code one bit of data. Four V_ts can code two bits of data (00, 01, 10, and 11). This technique is called the multilevel cell technology.

Resistance-change NVM or RRAM

Programmable resistor
- Phase change memory (PCM): alloy of Ge, Sb, and Te programmed by current pulse
- Metal migration memory: extremely thin filament of metal ion programmed by electrical field pulse

Concept of PCM

programmable resistor
(alloy of Ge, Sb, and Te)



R_{high} state ("1") R_{low} state ("0")

(a) Concept of a resistance-change memory such as a PCM.
(b) Program the PCM into high-resistance state by rapid solidification, producing a highly resistive amorphous phase.
(c) Program the PCM into low-resistance state by annealing, turning the amorphous material into a conductive crystalline phase.

PCM can be written and erased at SRAM speed and has much better endurance than the charge-storage memory.