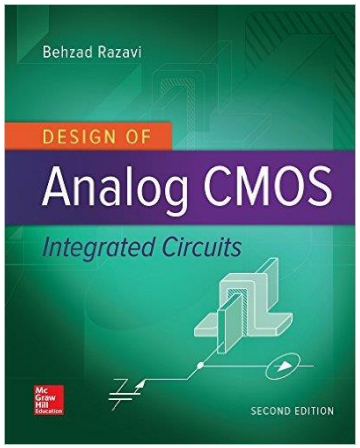
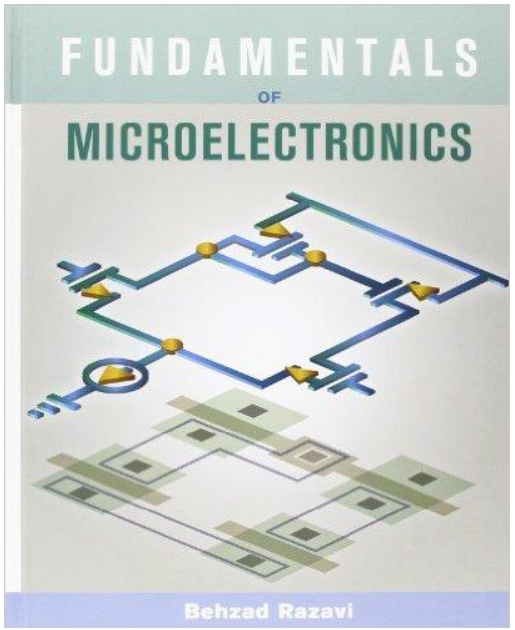
Metal-Oxide-Semiconductor (MOS) Transistor

- **Large signal characteristics**
- **Small signal model**
- **Small channel effects**
- **Technology progress**

Recommended books:

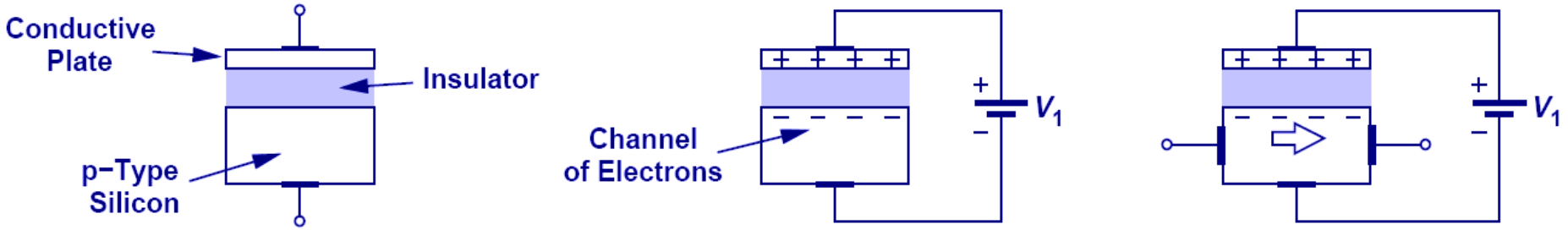


B. Razavi, McGraw-Hill Higher Education;
2 edition (January 22, 2016)



Large signal characteristics

Metal-Oxide-Semiconductor (MOS) Capacitor



p – type silicon:

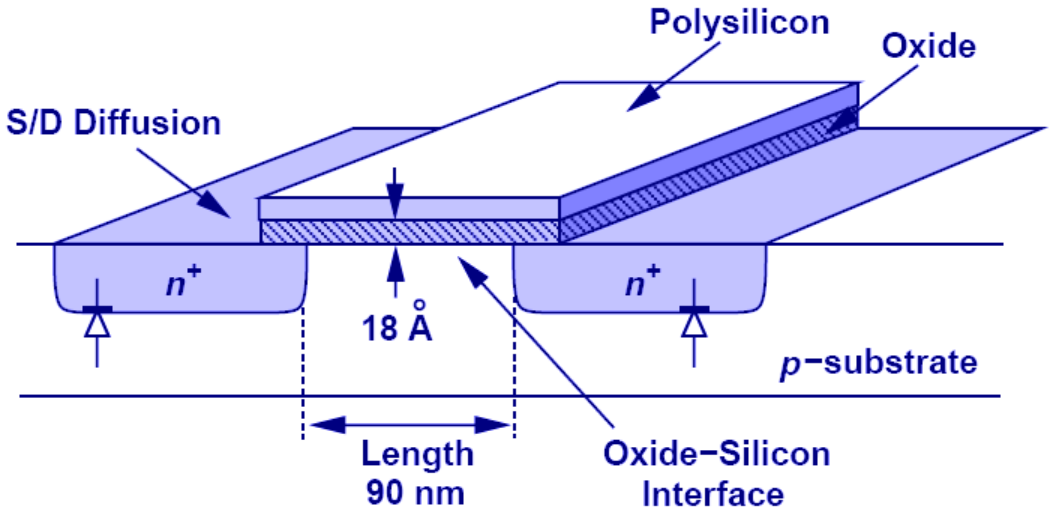
Majority carriers in p-type semiconductor: $p_p \approx N_A$

Minority carriers in p-type semiconductor: $n_p \approx \frac{n_i^2}{N_A}$

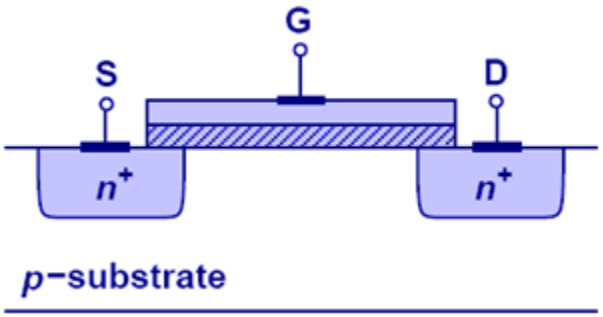
$$np = n_i^2$$

The MOS structure can be thought of as a parallel-plate capacitor, with the top plate being the positive plate, oxide being the dielectric, and Si substrate being the negative plate.

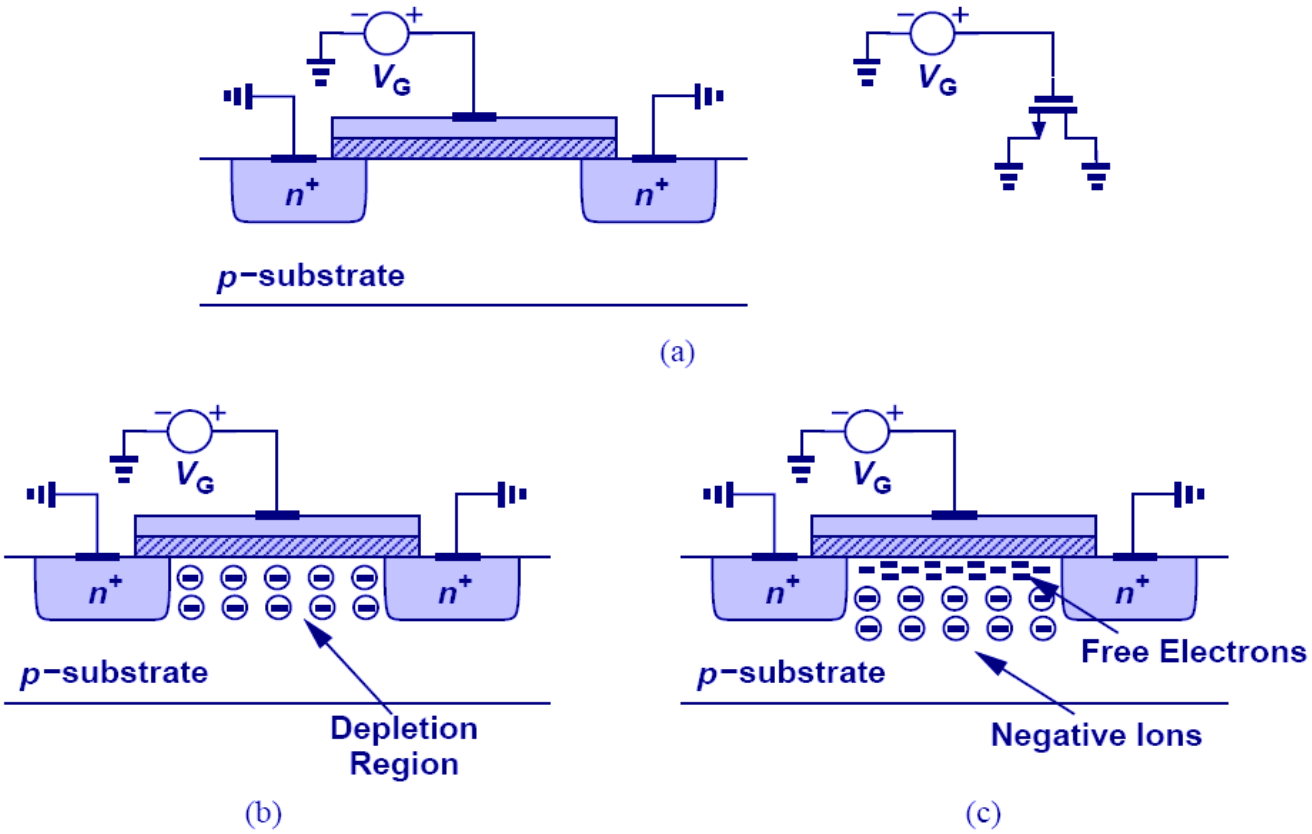
MOSFET Structure



The gate is formed by polysilicon, and the insulator by silicon dioxide.

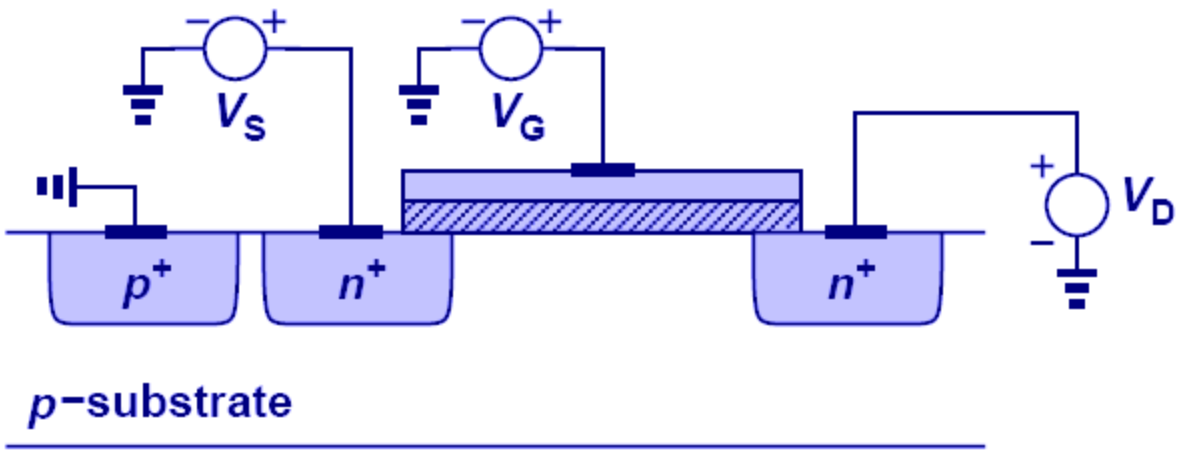


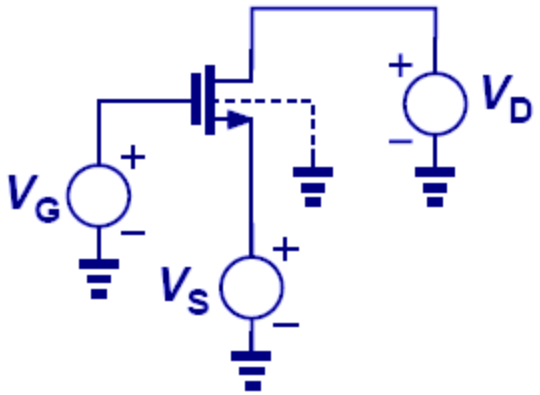
Formation of Channel



- First, the holes are repelled by the positive gate voltage, leaving behind negative ions and forming a depletion region. Next, electrons are attracted to the interface, creating a channel ("inversion layer").

Threshold voltage & body effect





$$\Phi_F = (kT/q)\ln(N_{sub}/n_i)$$

$$V_{TH0} = \Phi_{MS} + 2\Phi_F + \frac{Q_{dep}}{C_{ox}} \qquad Q_{dep} = \sqrt{4q\epsilon_{si}|\Phi_F|N_{sub}}$$

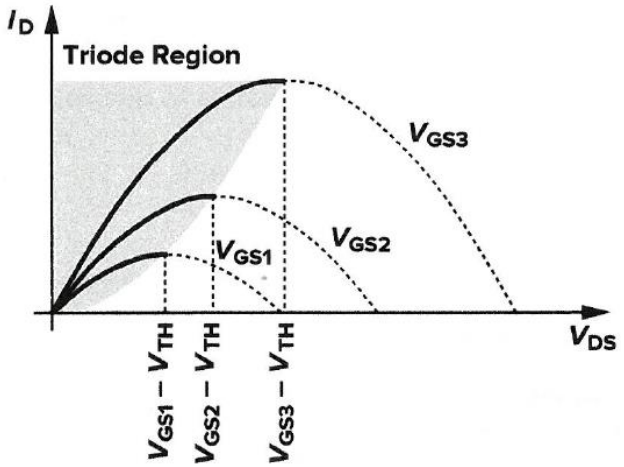
As the source potential departs from the bulk potential ($V_{SB} \neq 0$), the threshold voltage changes.

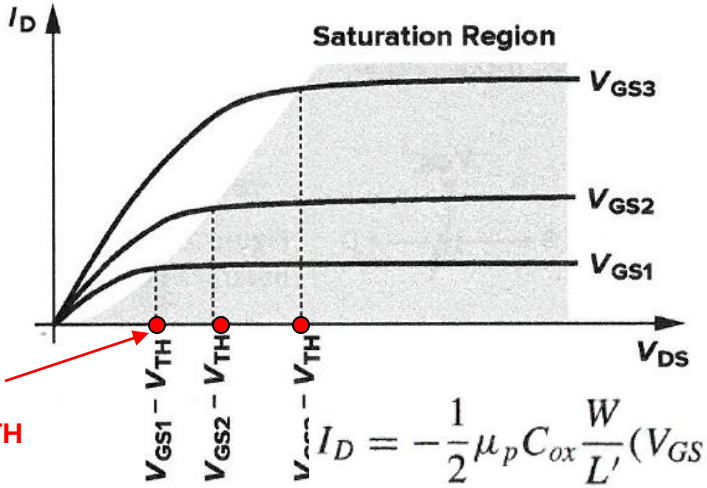
$$V_{TH} = V_{TH0} + \gamma\left(\sqrt{2\Phi_F + V_{SB}} - \sqrt{|2\Phi_F|}\right)$$

The value of γ typically lies in the range of 0.3 to 0.4 $V^{1/2}$.

Derivation of I/V Characteristics - saturation region

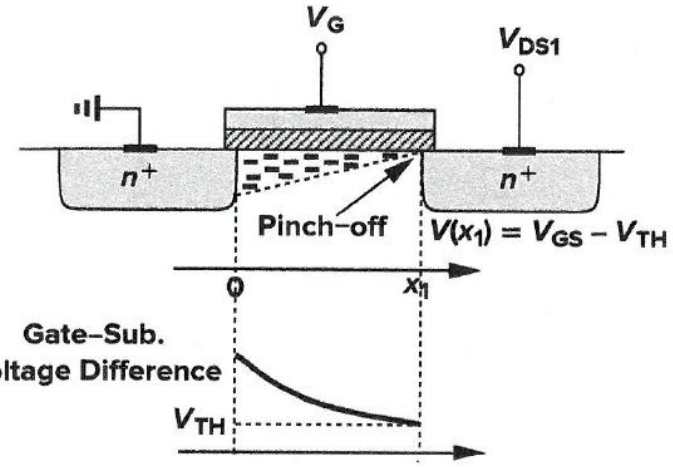
What happens if the drain-source voltage exceeds V_{GS}-V_{TH}?

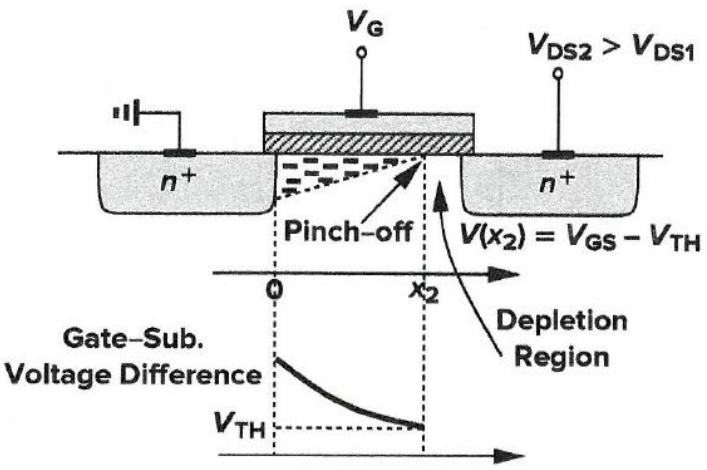


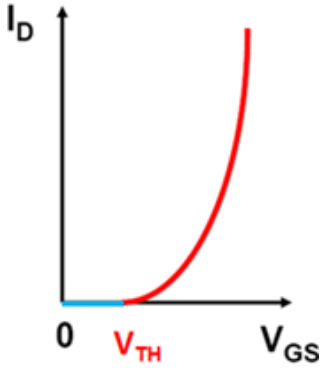


$V_{DSsat} = V_{GS}$-V_{TH}

$$I_D = -\frac{1}{2}\mu_p C_{ox}\frac{W}{L'}(V_{GS} - V_{TH})^2$$

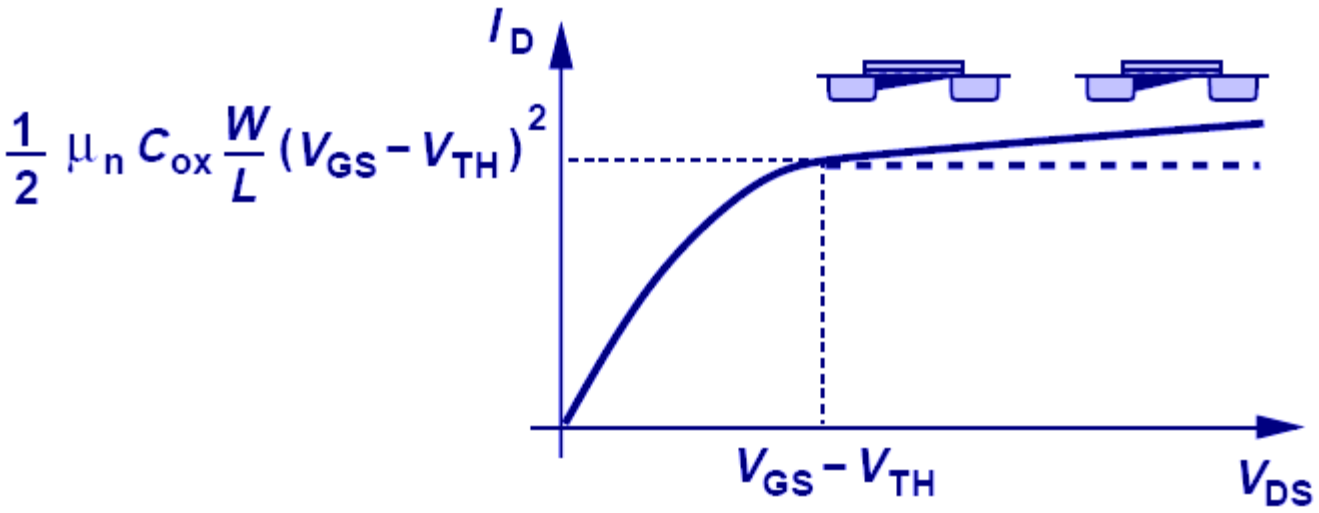






As the electrons approach the pinch-off point (where $Q_d \to 0$), their velocity rises tremendously ($v = I/Q_d$) the electrons simply shoot through the depletion region near the drain junction

Channel-Length Modulation



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

- The original observation that the current is constant in the saturation region is not quite correct. The end point of the channel actually moves toward the source as V_D increases, increasing I_D. Therefore, the current in the saturation region is a weak function of the drain voltage.

Small signal model

Drain current in saturation

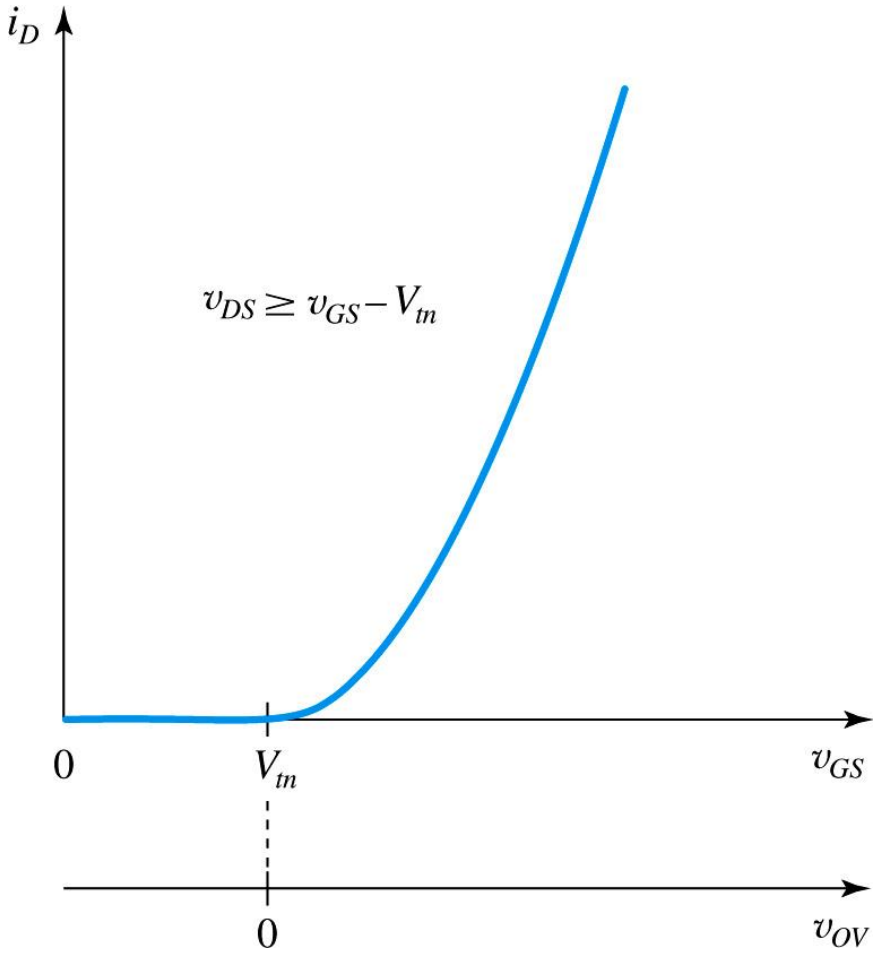
$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2$$

$$g_m = \left.\frac{\partial I_D}{\partial V_{GS}}\right|_{VDS\ \text{const.}}$$

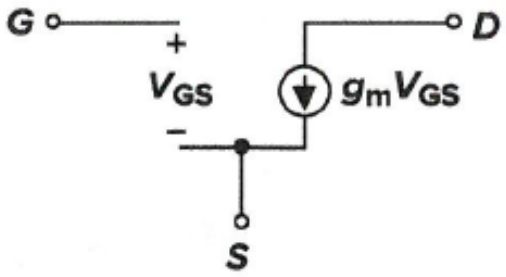
$$= \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})$$

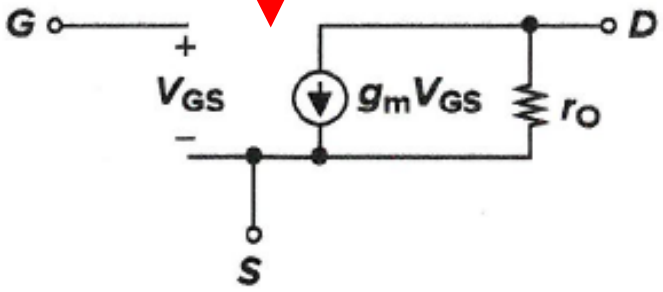
$$g_m = \sqrt{2\mu_n C_{ox}\frac{W}{L}I_D}$$

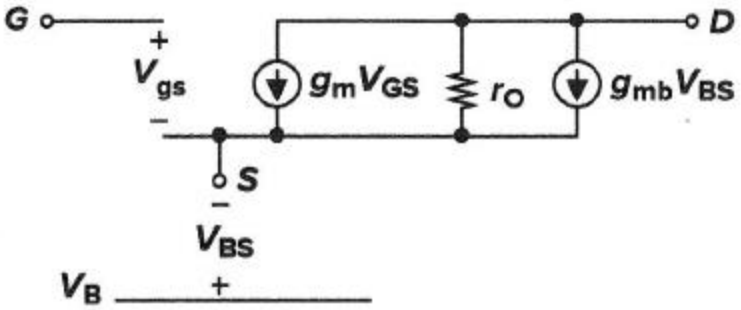
$$= \frac{2I_D}{V_{GS} - V_{TH}}$$



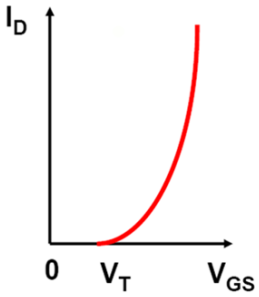
MOS Small-Signal Model (without capacitance)



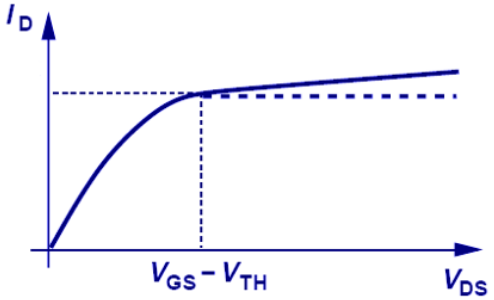




$$g_m = \sqrt{2\mu_n C_{ox}\frac{W}{L} I_D} = \frac{2I_D}{V_{GS} - V_{TH}}$$



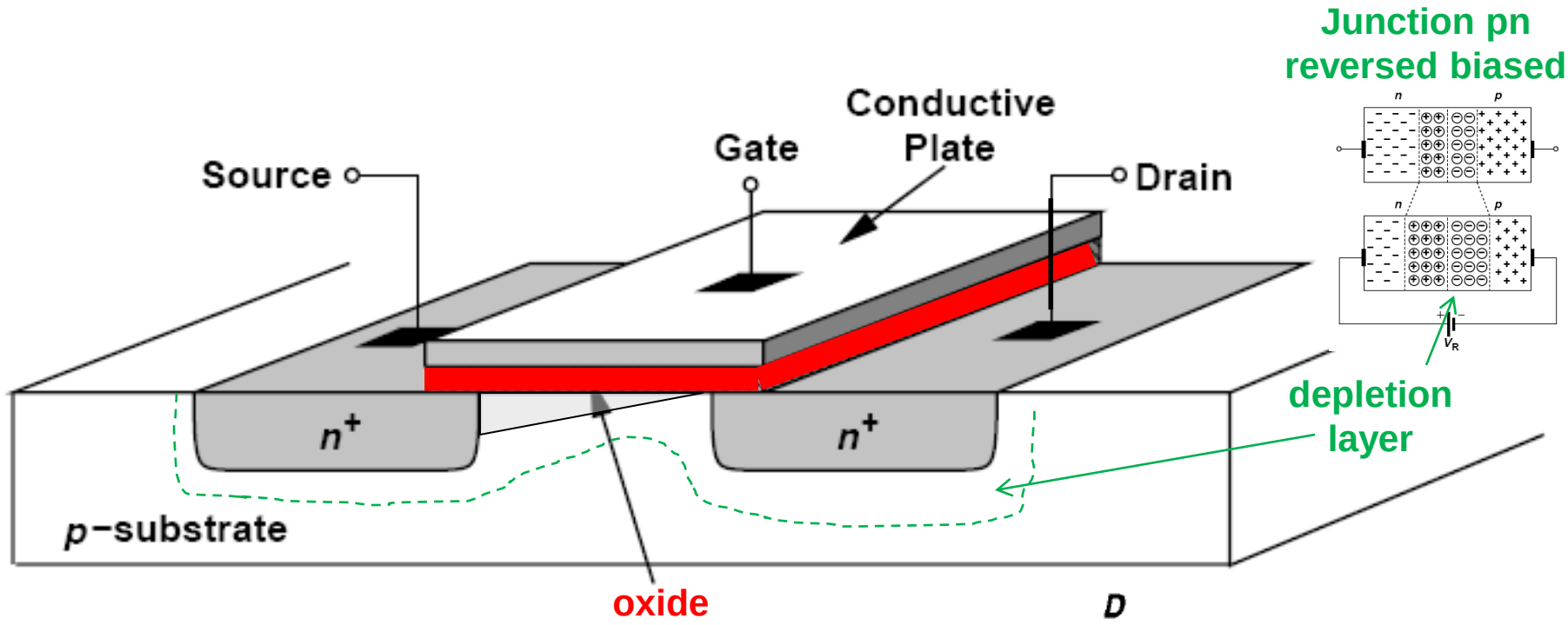
$$r_O \approx \frac{1}{\lambda I_D}$$



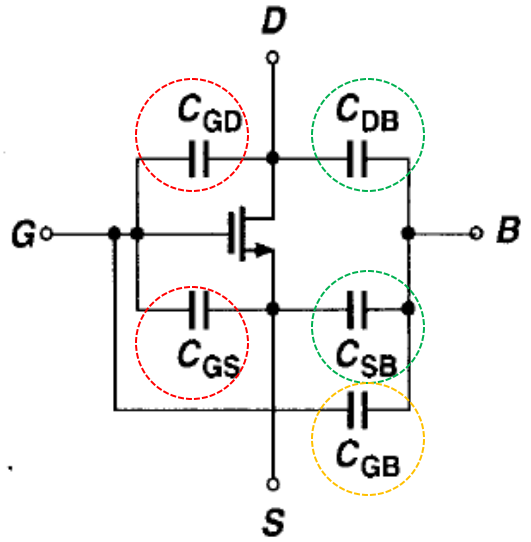
$$g_{mb} = \eta g_m$$

$$V_{TH} = V_{TH0} + \gamma\left(\sqrt{2\Phi_F + V_{SB}} - \sqrt{|2\Phi_F|}\right)$$

MOS Intrinsic Capacitances



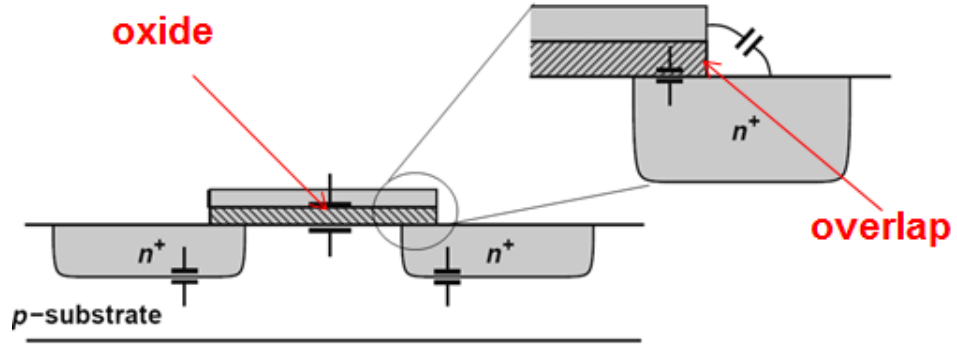
For a MOS, there exist oxide capacitance from gate to channel, junction capacitances from source/drain to substrate, and overlap capacitance from gate to source/drain.



12

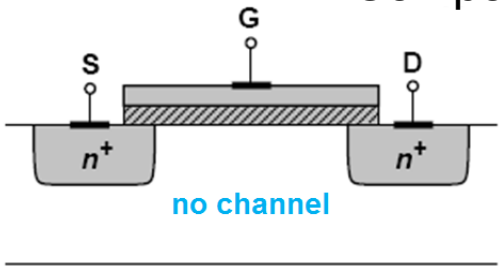
Variation of CGS and CGD versus VGS [1].

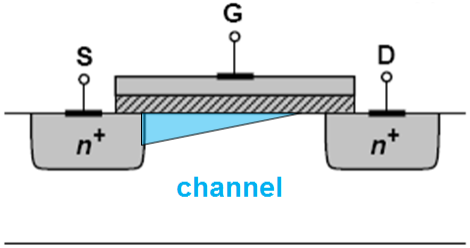
Two components of CGS and CGD: channel + overlap

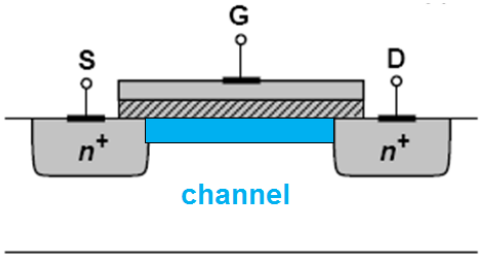


Example - CMOS 130nm
tox = 2.7 nm
Cox=12.8 fF/μm²
Cov=0.35 fF/μm

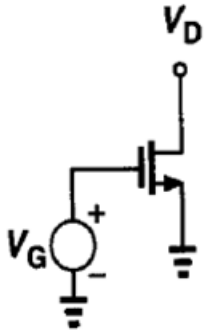
Component gate-channel depends on VGS

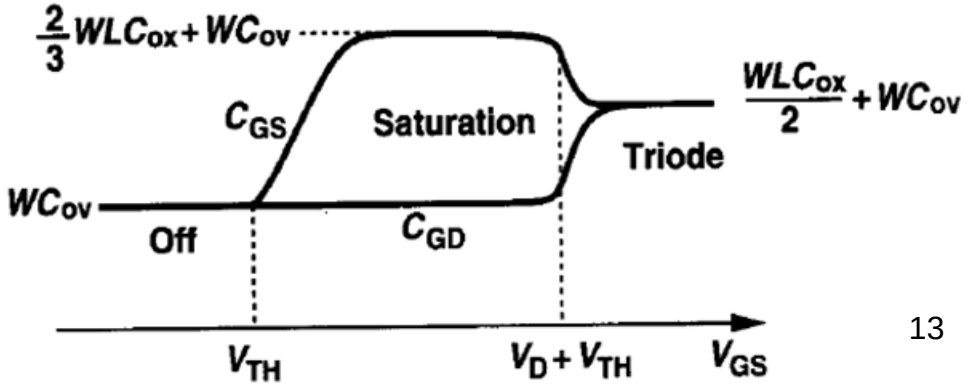




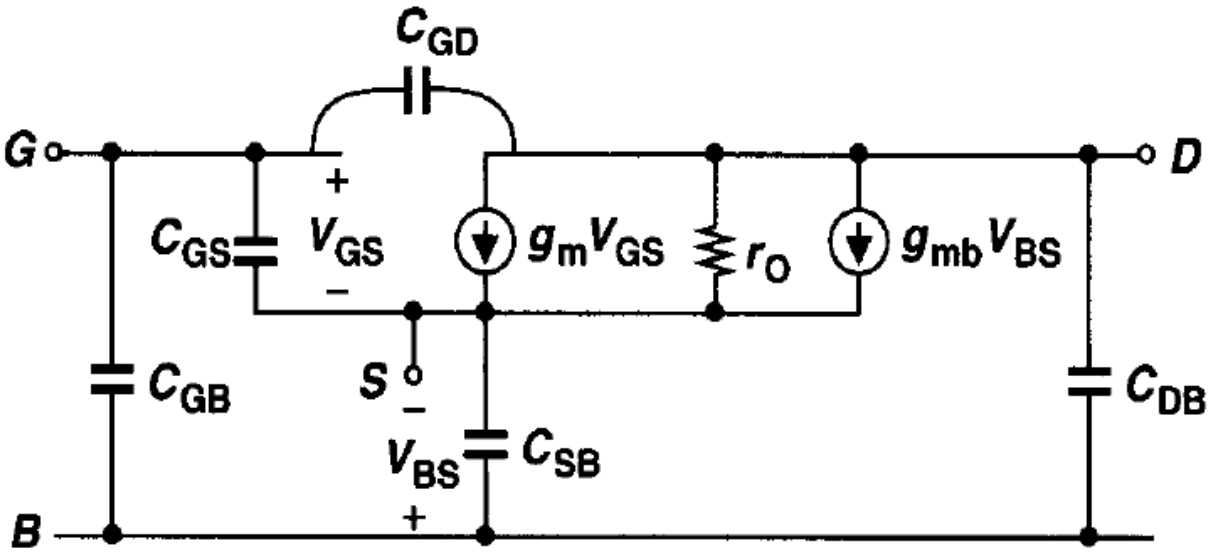


CGS = f(VGS)
CGD = f(VGS)





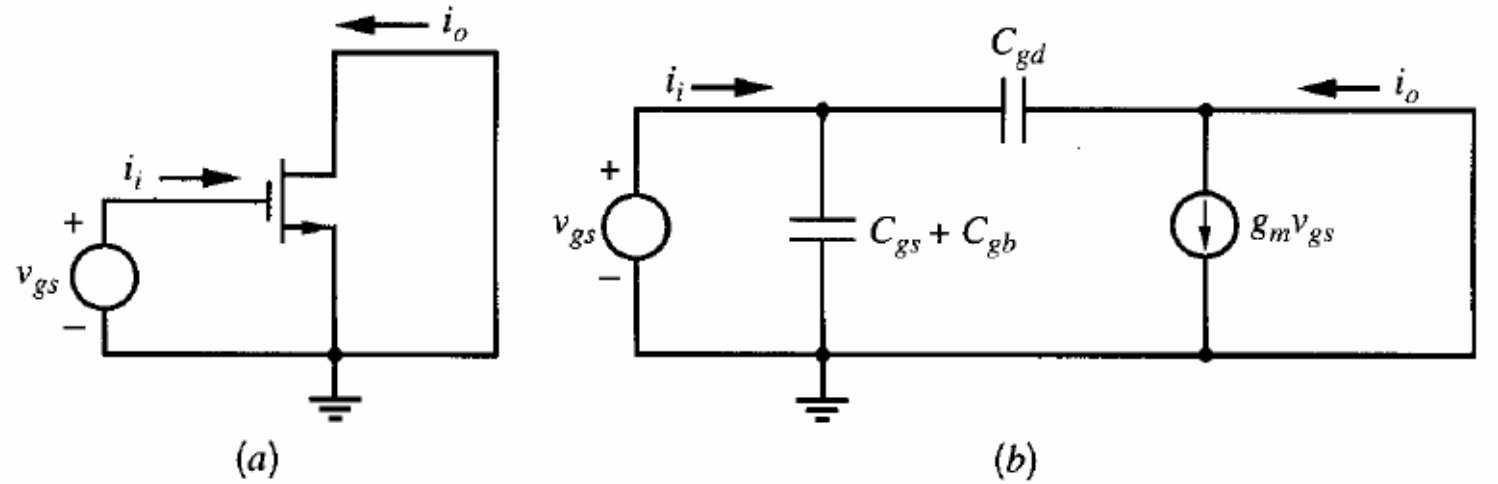
Complete MOS small-signal model [1]



Technology	Drain current (VGS=VDS)	W/L [μm/μm]	CGS [F]	CGD [F]	CSB [F]	CDB [F]	CGB [F]
130 nm	100 μA 10 μA	100 /0.2	55 f 45 f	44 f 44 f	2.1 f 0.16 f	103 f 106 f	32 f 25 f
45 nm	100 μA 10 μA	100 /0.2	77 f 34 f	15 f 15 f	8.9 f 2.9 f	0.048 f 0.026 f	19 f 25 f
130 nm	10 μA 1 μA	0.15 /0.13 (min)	0.181 f 0.103 f	0.071 f 0.072 f	0.637 f 0.627 f	0.540 f 0.565 f	0.008 f 0.018 f
45 nm	10 μA 1 μA	0.12 /0.04 (min)	0.042 f 0.028 f	0.002 f 0.019 f	- -	- -	- -

MOS transistor frequency response

Transition frequency *fT* is definied as the frequency where the magnitude of the short-circuit, commom source current gain falls to unity.



Circuit for calculating *fT:* a) AC schematic, b) small signal equivalent [Gray, Wiley, 2010].

Small signal input current

$$i_i = s\left(C_{GS} + C_{GB} + C_{GD}\right)v_{gs}$$

$$\frac{i_o}{i_i} \approx \frac{g_m}{s\left(C_{GS} + C_{GB} + C_{GD}\right)}$$

If the current fed forward through CGD is neglected

$$i_o \approx g_m v_{gs}$$

$$f_T \approx \frac{g_m}{2\pi\left(C_{GS} + C_{GB} + C_{GD}\right)} \approx \frac{g_m}{2\pi\ C_{GS}}$$

Maximum f_T versus channel length L

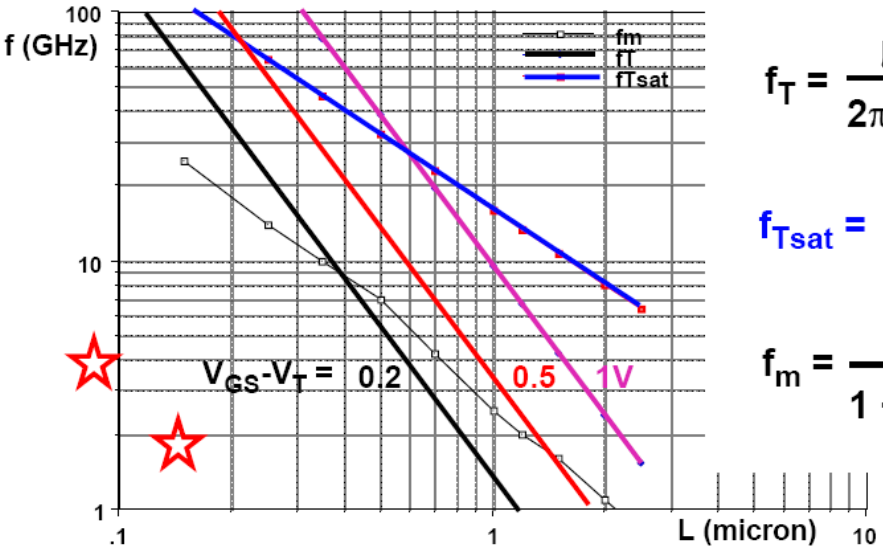
Sansen, Springer, 2008

$$C_{GS} = \frac{2}{3} WLC_{ox} \qquad g_m = 2K' \frac{W}{L} (V_{GS}-V_T) \qquad K' = \frac{\mu C_{ox}}{2n}$$

$$f_T = \frac{g_m}{2\pi C_{GS}} = \frac{1}{2\pi} \frac{3}{2n} \frac{\mu}{L^2} (V_{GS}-V_T) \quad \text{or} \quad \approx \frac{v_{sat}}{2\pi L}$$

This frequency is proportional to $V_{GS}-V_T$ and inversely proportional to L^2. Decreasing the channel length allows a higher frequency performance. In velocity saturation however, the time that the electrons need to cross the channel length is L/v_{sat}. The frequency f_T in velocity saturation is, to put it simply $v_{sat}/2\pi L$. This is the highest frequency that can be obtained with a MOS transistor.

1. For process with L≈100nm f_T of 100 GHz is available.
2. Experimental upper frequencies of VCO and LNA marked as fm is about 1/5 of f_T.
3. Clock frequency of processors is about 1/100 of f_T.



Processors

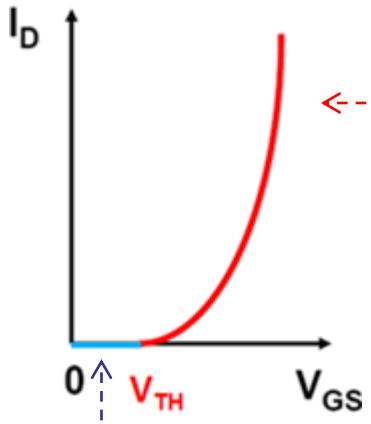
$$f_T = \frac{\mu}{2\pi L^2} \underbrace{(V_{GS}-V_T)}_{0.2 \ldots 1\ V}$$

$$f_{Tsat} = \frac{v_{sat}}{2\pi L}$$

$$f_m = \frac{f_T}{1 + \alpha_{BD}} \qquad \alpha_{BD} \approx \frac{C_{BD}}{C_{ox}}$$

Subthreshold operation

Weak inversion – strong inversion transfer characteristic

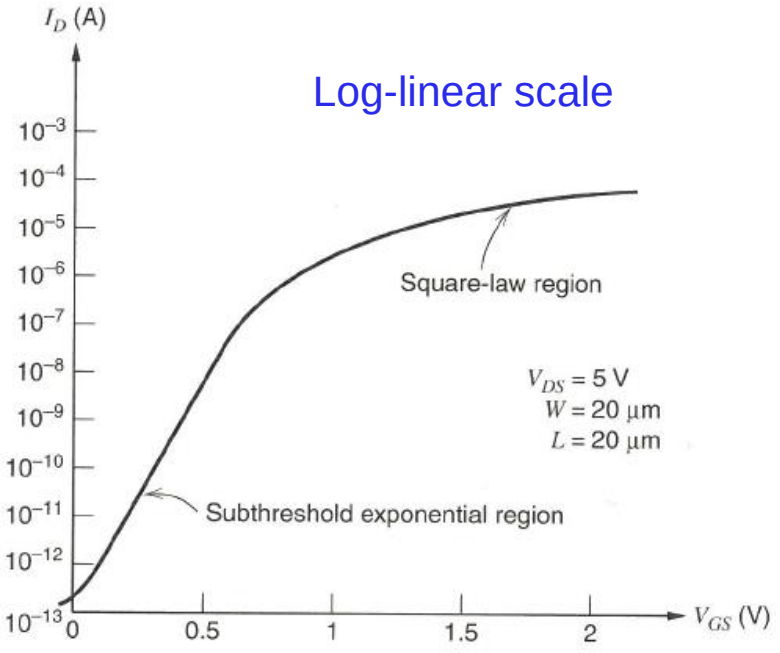


$$I_D \approx \frac{1}{2}\mu_n C_{ox}\frac{W}{L}\left(V_{GS}-V_{TH}\right)^2$$

Weak Inversion
Subthreshold region

$$I_D \approx \frac{W}{L} I_t \exp\left(\frac{V_{GS}-V_{TH}}{nV_T}\right)$$

Log-linear scale



Gray, Wiley, 2010

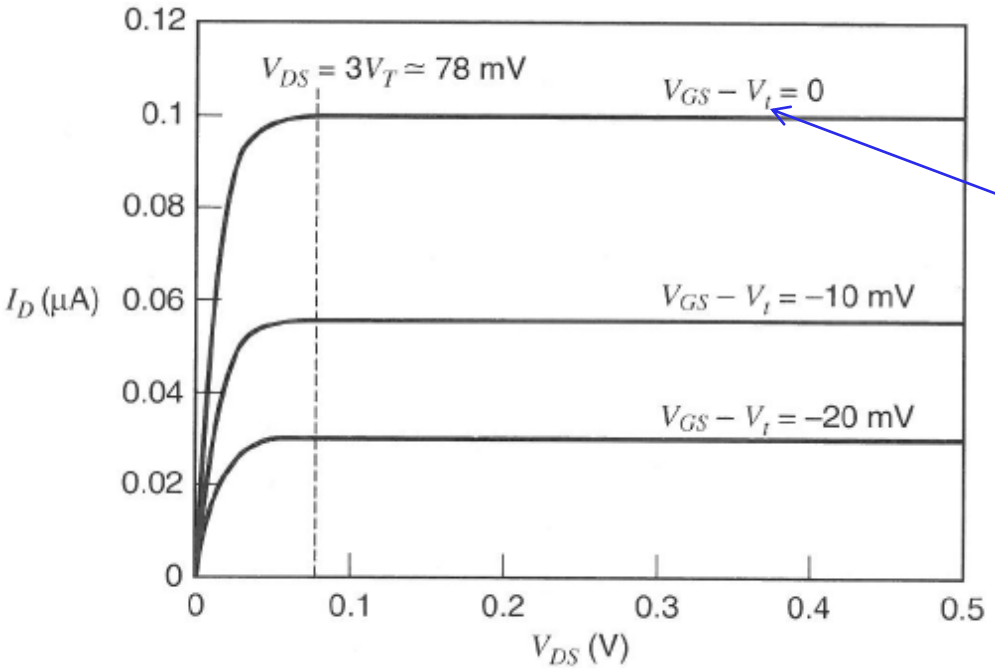
Weak inversion in MOS – drain current,

$$I_D = \frac{W}{L} I_t \exp\left(\frac{V_{GS} - V_{TH}}{nV_T}\right)\left[1 - \exp\left(-\frac{V_{DS}}{V_T}\right)\right]$$

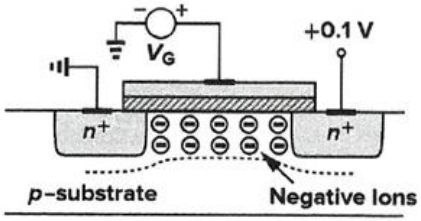
$V_T \approx 26\text{mV at } T = 300\text{K}$

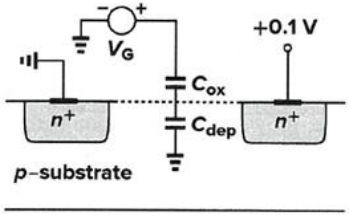
$$n = \frac{C_{dep} + C_{ox}}{C_{ox}} \approx 1.5$$

$C_{dep}(= C_{js})$ depletion-region cap.



Gray uses V_t as threshold voltage





Drain current versus drain-source voltage in weak inversion.

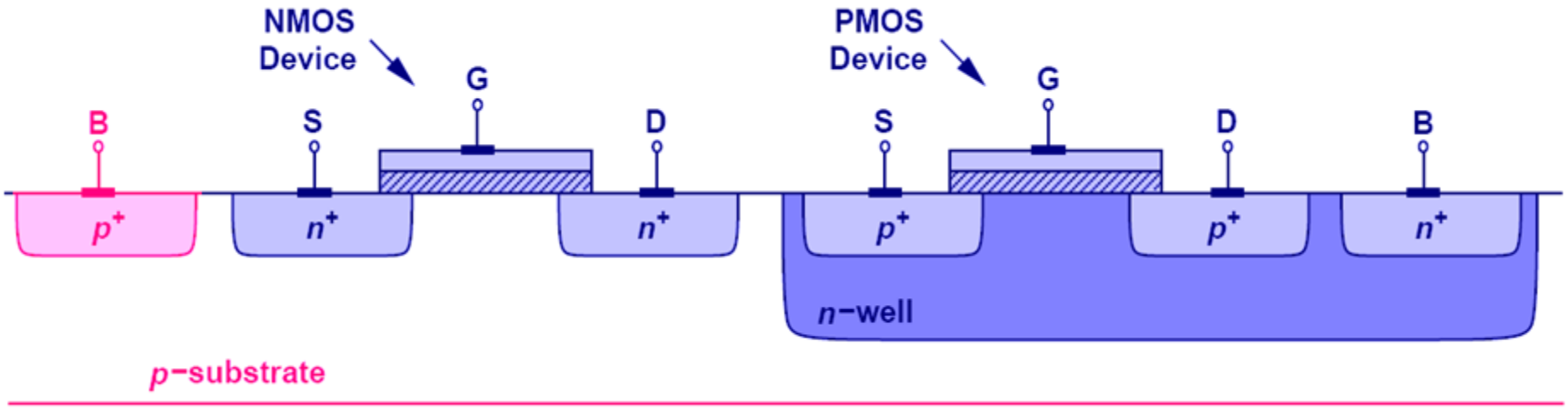
$W = 20\ \mu\text{m}$, $L = 20\ \mu\text{m}$, $n = 1.5$, and $I_t = 0.1\ \mu\text{A}$.

Drain current is almost constant when $V_{DS} > 3\ V_T$

Gray, Wiley, 2010

Selected short channel effects
&
Technology progress

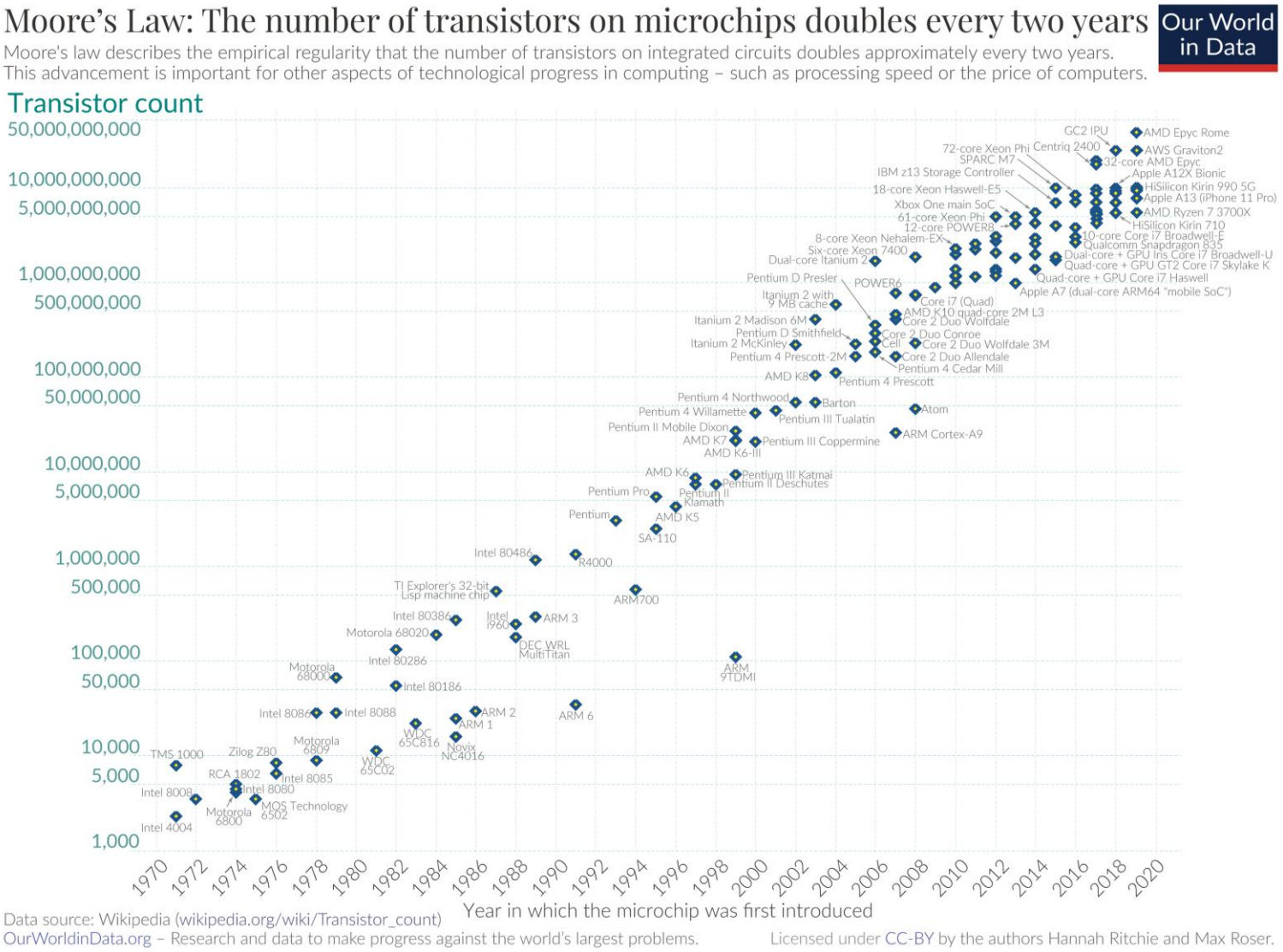
NMOS and PMOS device in CMOS process



- It possible to grow an n-well inside a p-substrate to create a technology where both NMOS and PMOS can coexist.
- It is known as CMOS, or “Complementary MOS”.

Technology progres – Moore’s law

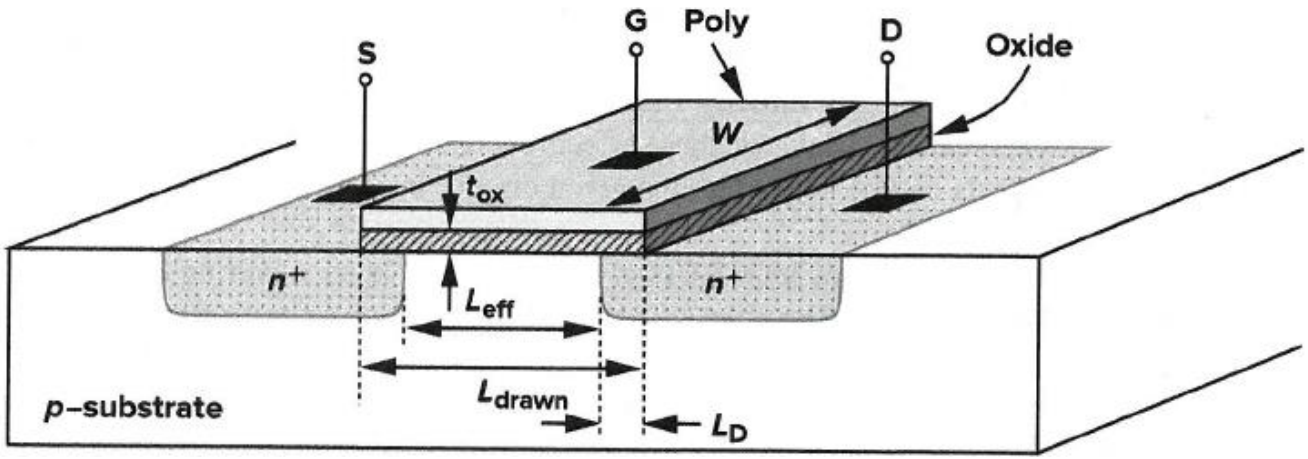
Gordon Moore co-founder of Intel Corporation, and the author of Moore's law



Moore's law is the observation that the number of transistors in a dense integrated circuit doubles approximately every two years.

Structure of MOS Devices

Razavi, McGrawHill, 2016



2001 r L = 150 nm

2003 r L = 90 nm

2009 r L= 45 nm

2013 r L = 22 nm

…….

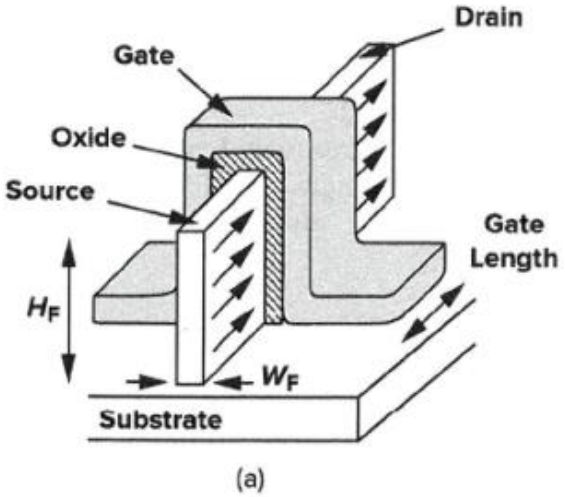
2016 r L = 10 nm

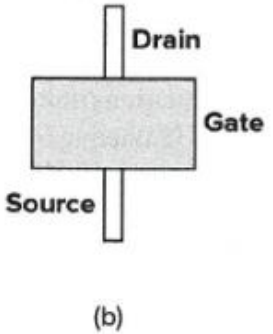
2018 L= 7nm

2020 L=5 nm

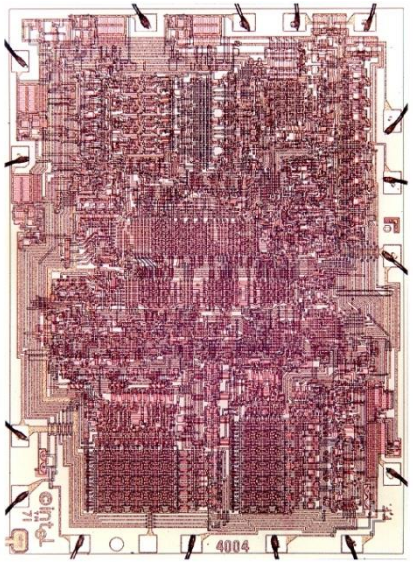
2022 L=3 nm

Starting from 16 nm node:
FINFET

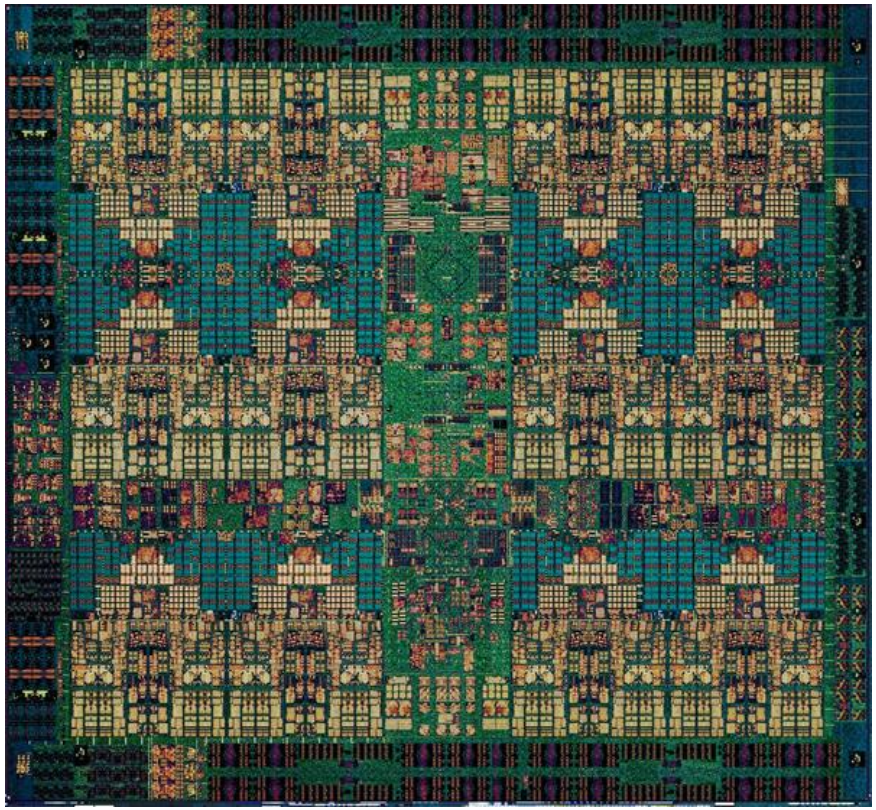




Technology progres – procesor example



Microprocesor Intel 4004 (1971)
Technology NMOS
~2.000 transistors, 1 MHz



https://en.wikichip.org/wiki/ibm/microarchitectures/power9

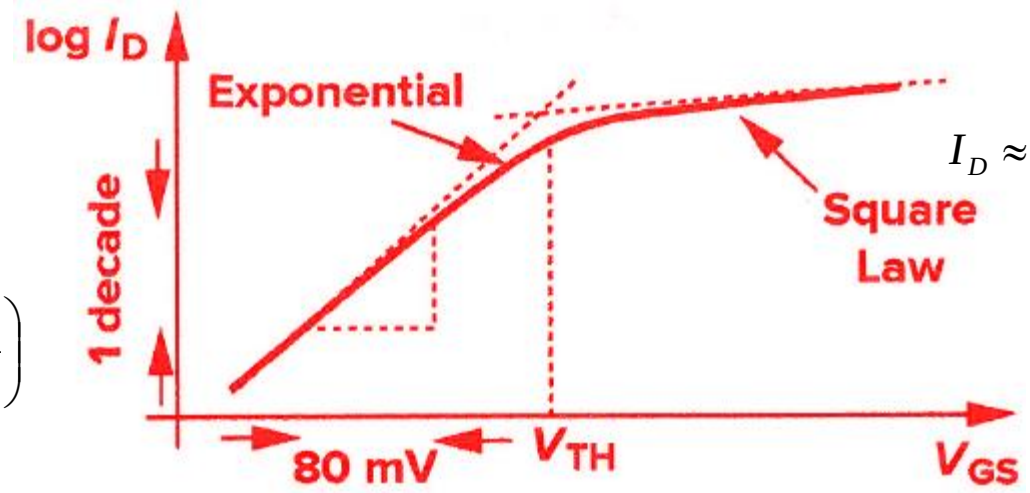
GlobalFoundries 14 nm FinFET on SOI Process
17-layer metal stack
8,000,000,000 transistors
15 miles of wire
693.37 mm² die size
25.228 mm x 27.48416 mm

Leakage current – Threshold voltage

Razavi, McGrawHill, 2016

The key point here is that as V_{GS} falls below V_{TH}, the drain current drops at a finite rate. With typical values of $n \approx 1.5$, at room temperature V_{GS} must decrease approximately 80 mV for ID to decrease by one decade.

$$I_D \approx \frac{W}{L} I_t \exp\left(\frac{V_{GS} - V_{TH}}{nV_T}\right)$$



$$I_D \approx \frac{1}{2}\mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{TH})^2$$

For example, if a threshold of 0.3V is chosen in a process to allow low-voltage operation, then when V_{GS} is reduced to zero, the drain current decreases only by a factor of:

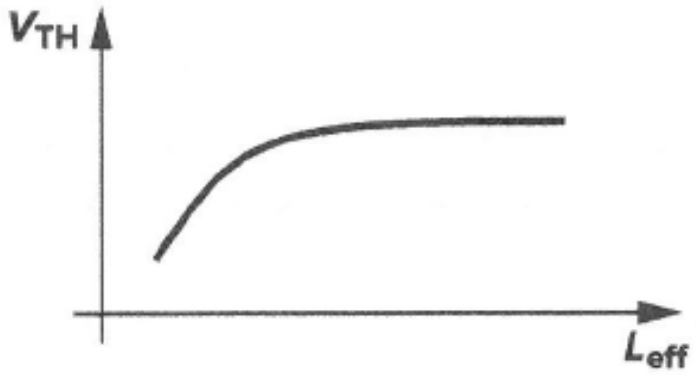
$$10^{0.3\ \text{V}/80\ \text{mV}} = 10^{3.75} \approx 5.62 \times 10^3$$

For example, if the transistors carries about 1μA for V_{GS}=V_{TH} and we have 100 million such devices, then they draw 18 mA when they are nominally off (e.g. problem in memories – significant power dissipation or loss of analog information.

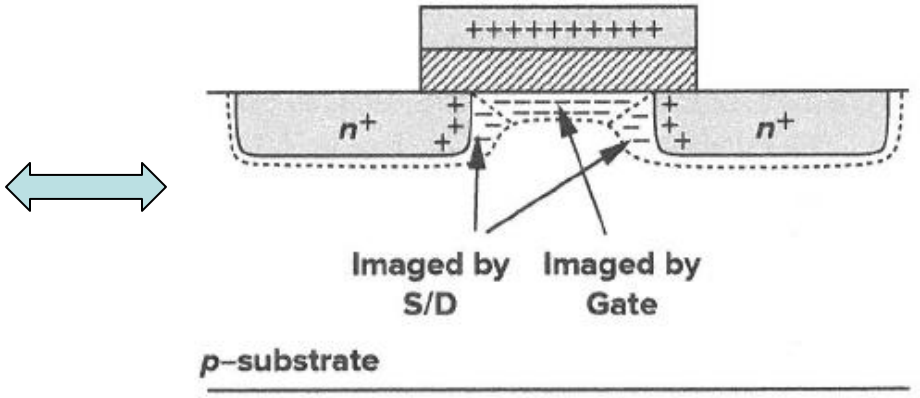
A subthreshold slope of 80 mV/dec imposes a lower bound of 400 mV for VTH if the „off curent" must be roughly five orders of magnitude lower than the „on current".

Short-Channel Effects – Threshold Voltage Variation

Transistors fabricated on the same wafer but with with different lengths yields lower V_{TH} as L decreases



Variation of threshold with channel length

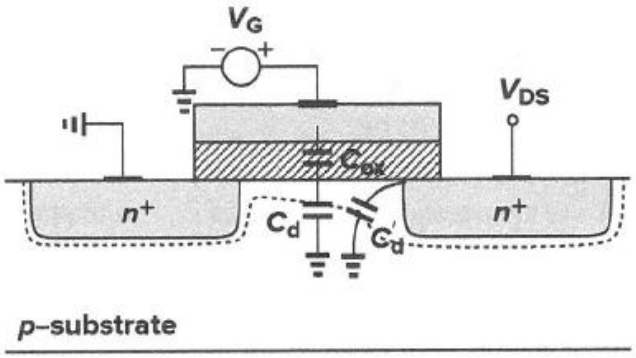


Charge sharing between source/drain depletion regions and the channel depletion region (problem with high output resistance)

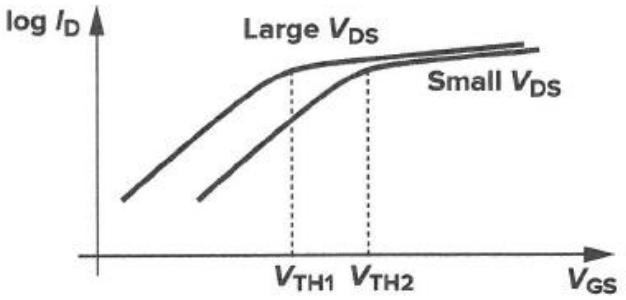
This is because the depletion region associated with source and drain junction protrude into the channel area considerably, thereby reducing the immobile charge that must be imaged by the charge on the gate.

Short-Channel Effects – Threshold Voltage Variation

Drain-induced barrier lowering or **DIBL** is a short-channel effect in MOSFETs referring originally to a reduction of threshold voltage of the transistor at higher drain voltages. Drain voltage makes the surface potential more positive by creating two-dimentional field in the depletion region. In essence, the drain introduces the capacitance *Cd'* (see fig. below) that rises the surface potential in a manner similar to *Cd* – as the results the the V_{TH} is reduced.



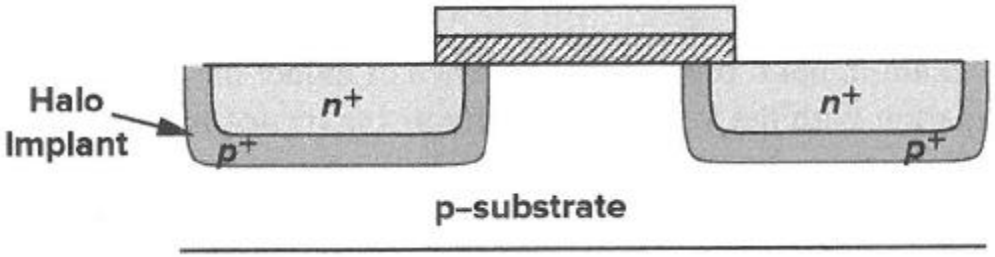
DIBL in a short-channel device



effect of DIBL on current characteristic

Short-Channel Effects – Threshold Voltage Variation

Reverse short channel effects – in namometer technologies, the threshold voltage decreases as the channel length increases. Nanometer technologies uses „halo" implant to reduce the penetration of the drain depletion region into the channel area.



MOS structure with halo implant.

Threshold voltage V_{TH} is a function of *Nsub: Nsub* ↑ *then* V_{TH} ↑

$$V_{TH} = \phi_{MS} + 2\phi_F + \frac{Q_{dep}}{C_{ox}}$$

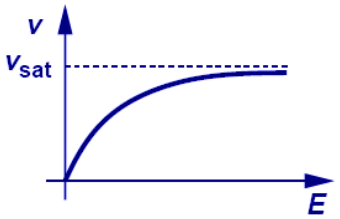
$$\phi_F = (kT/q)\ln(N_{sub}/n_i)$$

$$Q_{dep} = \sqrt{4q\epsilon_{si}|\phi_F|N_{sub}}$$

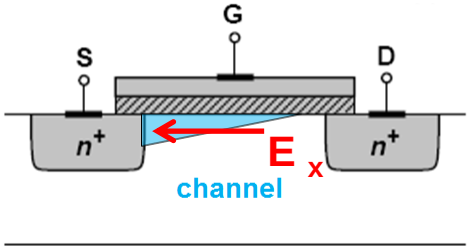
Due to non-uniform substrate doping the „local" threshold voltage also varies from source to drain. We can take the average along the channel to obtain an overal threshold voltage for a given structure. As the channel length increases, the average substrate doping decreases, and also V_{TH} is reduced.

Short-Channel Effects: Velocity Saturation

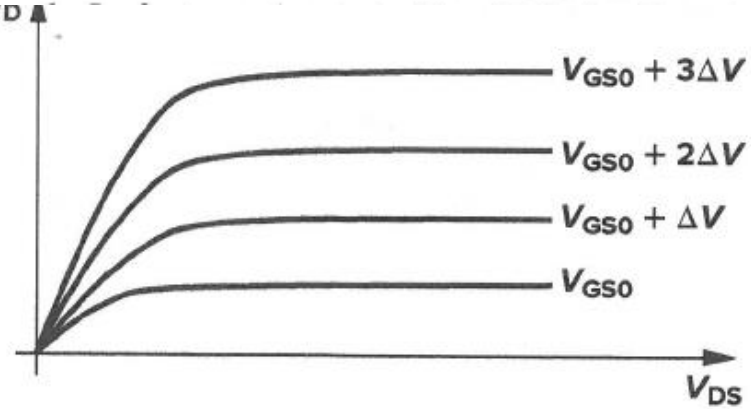
The mobility of carriers also depends on the *lateral* electric field in the channel

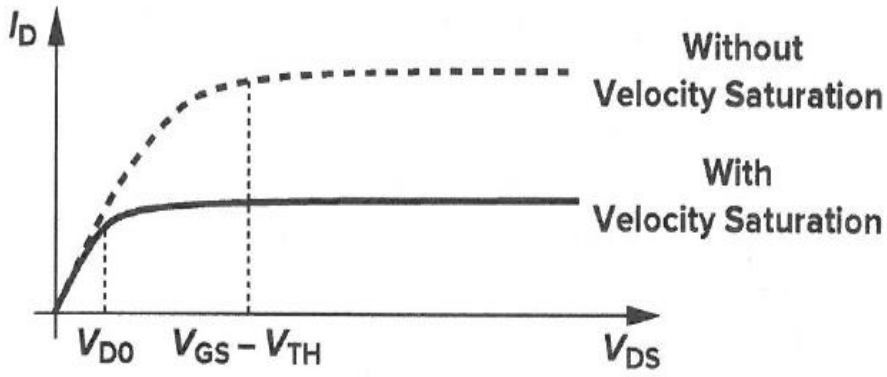


$$I_D = v_{sat} Q_d$$
$$= v_{sat} W C_{ox} (V_{GS} - V_{TH})$$



Interestingly, the current is *linearly* proportional to the overdrive voltage





Effect of velocity saturation on drain-current characteristics

A compact and versatile equation developed to represent velocity saturation (in the saturation region) is

$$I_D = W C_{ox} v_{sat} \frac{(V_{GS} - V_{TH})^2}{V_{GS} - V_{TH} + 2\dfrac{v_{sat} L}{\mu_{eff}}}$$

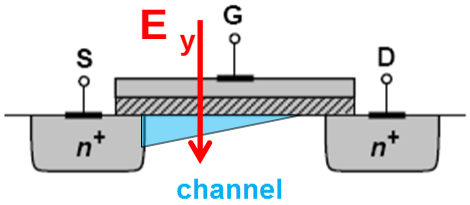
$$\mu_{eff} = \frac{\mu_0}{1 + \theta(V_{GS} - V_{TH})}$$

Short-Channel Effects: Mobility Degradation with Vertical Field

At large gate source voltages, the high electric field developed between the gate and the channel confines the charge carriers to a narrower region below the oxide silicon interface, leading th the more oxide scattering and hence lower mobility.

An empirical equation modeling

$$\mu_{eff} = \frac{\mu_0}{1 + \theta(V_{GS} - V_{TH})}$$



θ is a fitting parameter roughly equal to $(10^{-7}/t_{ox})\ \mathrm{V}^{-1}$

In addition to lowering the current capability and transconductance of MOSFET, mobility degradation causes the I/V characteristic to deviate from simple square-law behaviour. Specifally, whereas a square-law devices generates only even harmonics in its drain current in response to a sinusoidal gate-source voltage, the above equation predicts odd harmonics as well. In fact writing:

$$I_D = \frac{1}{2}\frac{\mu_0 C_{ox}}{1 + \theta(V_{GS} - V_{TH})}\frac{W}{L}(V_{GS} - V_{TH})^2$$

assuming that $\theta(V_{GS} - V_{TH}) \ll 1$, we obtain

$$I_D \approx \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}[1 - \theta(V_{GS} - V_{TH})](V_{GS} - V_{TH})^2$$

$$\approx \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}\left[(V_{GS} - V_{TH})^2 - \theta(V_{GS} - V_{TH})^3\right]$$

I_{ON} improvement by technologies

$$I_d = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{gs} - V_t)^2 \times \text{n (floor number)}$$

→ Nanosheet (N2)

FinFET (16nm) : $2H_{FIN}+W_{FIN}$

High mobility $\frac{\varepsilon_{ox}}{t_{ox}} \rightarrow HK(45nm\ node)$

Intel : 22nm

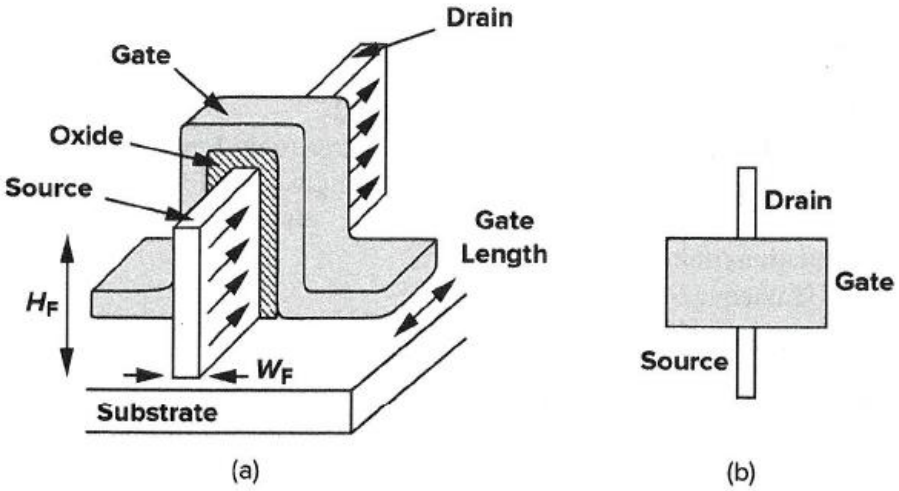
$$\frac{\varepsilon_{ox}}{t_{ox}} = \frac{\varepsilon_{HK}}{t_{HK}} = \frac{\varepsilon_{ox}}{EOT} \quad EOT = t_{HK}\frac{\varepsilon_{ox}}{\varepsilon_{HK}} \quad \varepsilon = k\varepsilon_0$$

t_{ox} >1nm. Otherwise, I_g is significant

1. Strained Si (90nm)
2. High mobility channel (5nm SiGe p-channel)

FinFET

This device exhibits superior performance as channel lengths fall below approximately 20 nm.



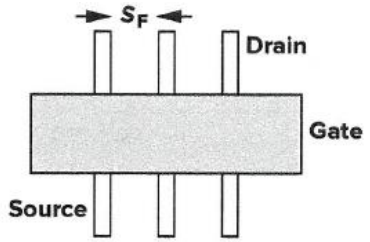
(a) FinFET structure, and (b) top view.

the gate length can be readily identified,

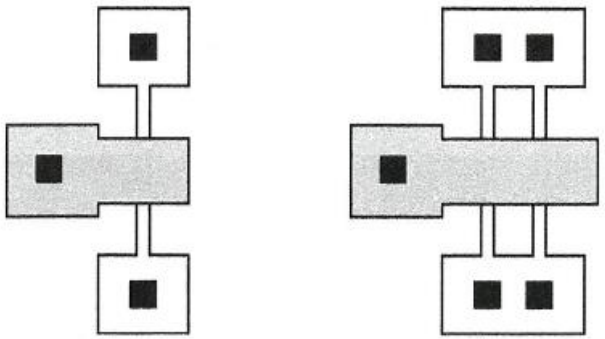
but how about the gate width?

$$W = W_F + 2H_F$$

Typically, $W_F \approx 6$ nm and $H_F \approx 50$ nm.



FinFET with multiple fins.

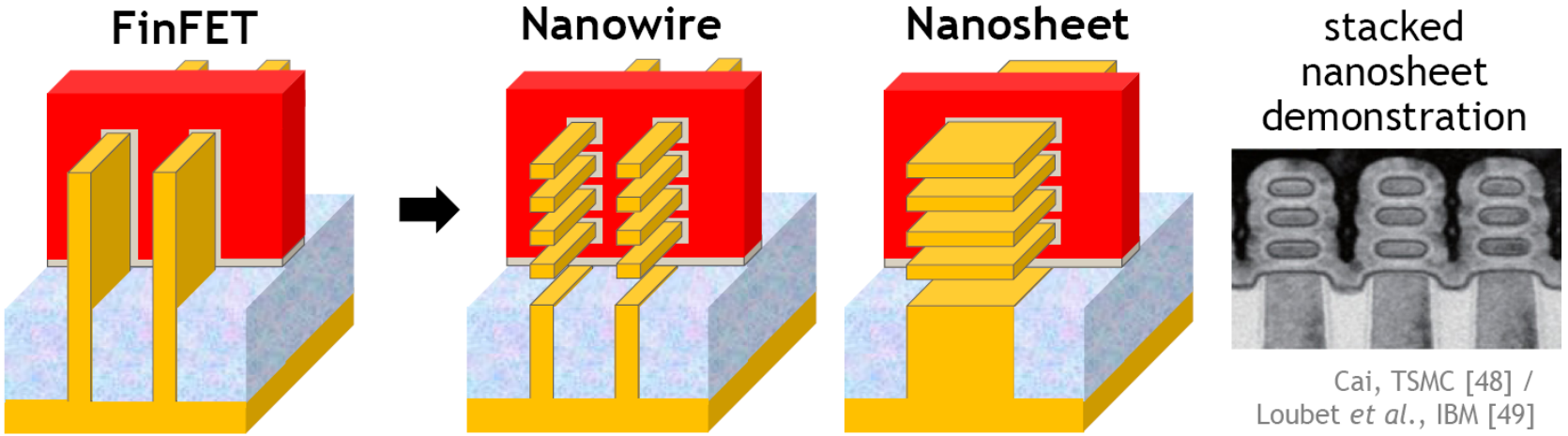


the gate and S/D contacts must be placed away from the core of the device

Nanowire and nanosheet

Migrating to Gate-All-Around (GAA)

- FinFET has poor short-channel control with further L_{gate} scaling
- Need better short-channel control & more W_{eff} per die area
- Stacked GAA nanowires & nanosheets are promising
- Nanowires offer better SCE, nanosheets offer better area scaling



Cai, TSMC [48] / Loubet *et al.*, IBM [49]