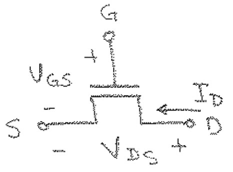
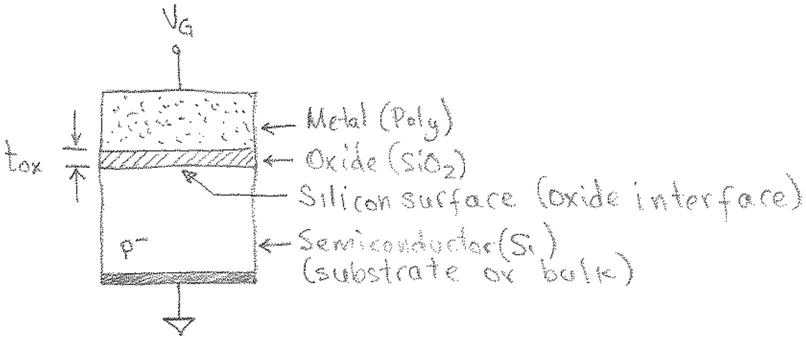
Electrical Characteristics of Mosfets

- MOSFET conduction: $I_D = f(V_{GS}, V_{DS})$
  - Depends on the operation of the underlying MOS system (MOS capacitor)





Metal: Actually poly-crystaline silicon
Oxide: Insulator
Semiconductor: Doped silicon crystal (NMOS: p-type , PMOS: n-type)

Oxide capacitance: $C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$

ε_{ox} = Silicon dioxide permittivity: $\varepsilon_{ox} = 3.9\,\varepsilon_0$, $\varepsilon_0 = 8.854 \times 10^{-14}$ F/cm
t_{ox} = Oxide thickness in cm
Very small. Modern processes $t_{ox} < 10nm$

C_{ox}: Determines the coupling level between gate and substrate (E_{ox} in V/cm) when V_{GS} is applied.
E_{ox} induces charge in the substrate Q_S: surface charge density

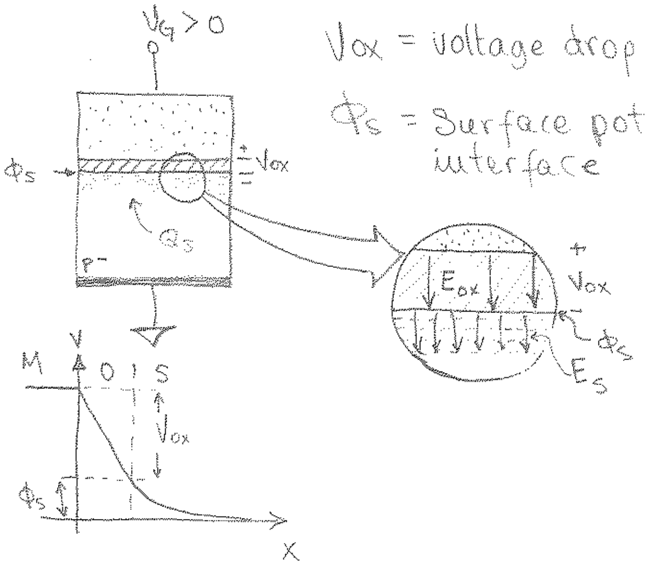
$$Q_S = -C_{ox} V_{GS}$$

A positive V_{GS} induces negative surface charge (NMOS).
For inducing Q_S, V_G must be at least equal to the threshold voltage V_T

$$V_T = V_{ox} + \phi_S$$

V_{ox} = voltage drop across the oxide layer

ϕ_S = Surface potential: Voltage at the silicon-oxide interface



Q_S is controlled by E_S

E_S = Surface electric field

E_s exerts a force F_h over the channel charged particles (holes)

$$F_h = +q \cdot E_s$$ ← Repels holes away from interface

$+q$ = hole charge ($-q$ for electrons) $= 1.6 \times 10^{-19}$ C

At the same time electrons are exerted a force $F_e = -qE_s$ in the opposite direction.

This explains the formation of the surface charge (negative in this case)

Q_s depends on V_G:

- Smal V_G: Creates small charge Q_B (bulk charge) due to the ionization of dopant. In the case bulk is p^-, dopant is boron, so $Q_B < 0$

  $$Q_B = -\sqrt{2q\mathcal{E}_{si} N_a \Phi_s}$$, where

  $\mathcal{E}_{si}$ = Silicon permittivity $= 11.8\mathcal{E}_0$
  N_a = Dopant concentration

  Note that under this condition, the channel region is depleted from charge carries (Q_B is static). This is called the DEPLETION MODE of the MOS system

  The depth of the depletion region (x_d) is a function of V_G. There is no current circulation in this mode.

- $V_G > V_T$: While $V_G < V_T$, $Q_s = Q_B$ → inmobile charge
  When $V_G = V_T$

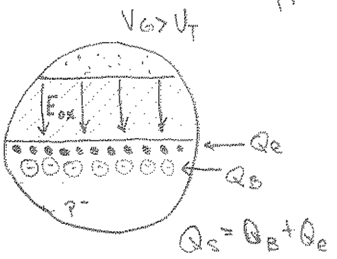
  $$Q_s = Q_B + Q_e < 0$$, where

  Q_e = electron charge $\qquad Q_e = -C_{ox}(V_G - V_T)$

  * Q_e is created by an electron accumolation in the oxide-bulk interface.

  * Q_e begins to form when $\Phi_s = 2|\Phi_F|$, where

  Φ_F = Fermi Potential



$$|\Phi_F| = \frac{KT}{q} \ln\left(\frac{N_a}{n_i}\right)$$, where

K = Boltzmann's constant
T = Temperature in °K

$\frac{KT}{q}$ = thermal voltage = 26mV @ 300°K

$$V_{Tn} = V_{ox}\Big|_{\phi_s = 2|\phi_F|} + 2|\phi_F|$$

So

$$V_{Tn} = \frac{1}{C_{ox}}\sqrt{2q\varepsilon_{Si}N_a(2|\phi_F|)} + 2|\phi_F| \leftarrow$$

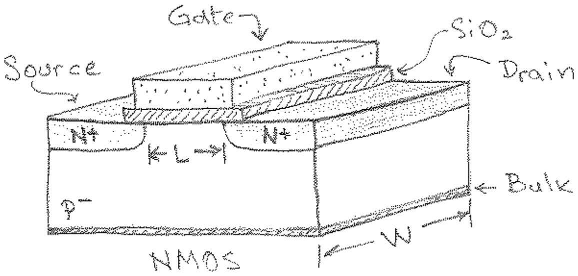
Threshold voltage of ideal MOS capacitor

More realistic situation compensates for flatband voltage (caused by stray charges w-the oxide)

$$V_{Tn} = \frac{1}{C_{ox}}\sqrt{2q\,\varepsilon_{Si}\,N_a\,(2|\phi_F|)} + 2|\phi_F| + V_{FB} + \frac{qD_I}{C_{ox}}$$

where D_I is an implant dose (Boron) to compensate V_{FB} ($V_{FB} < 0$)

NMOS Current

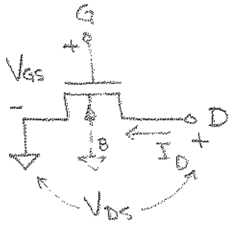


L = Channel length = minimum feature size

W = Channel width

$\frac{W}{L}$ = MOSFET aspect ratio

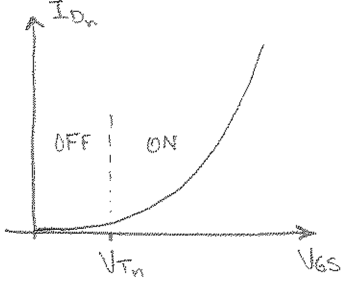
W & L refer to effective dimensions.



For $V_{GS} < V_T$, $I_D \simeq 0 \Rightarrow$ cutoff

Assuming $V_{DS} = V_{DD}$ and sweeping V_{GS}

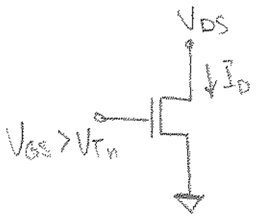
FET square-law model →

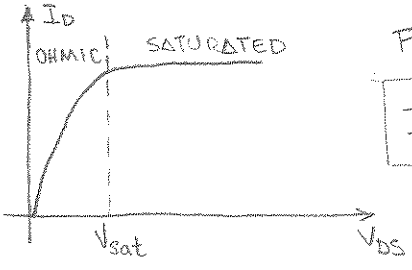


$$I_D = \frac{K_n}{2}(V_{GS} - V_{Tn})^2$$

$$K_n = \mu_n C_{ox}\frac{W}{L}$$

If we assume $V_{GS} > V_{Tn}$ constant and sweep V_{DS}





For $V_{DS} < V_{SAT}$

$$\boxed{I_{Dn} = \frac{K_n}{2}\left[2(V_{GS} - V_{Tn})V_{DS} - V_{DS}^2\right]}$$

The maximum value (to obtain V_{sat}) occurs at $\partial I_D / \partial V_{DS} = 0$

Solving for V_{DS}, we obtain $V_{DS}\Big|_{I_{D\,max}} = V_{sat} = V_{GS_n} - V_T$

For $V_{DS} > V_{sat} = V_{GS} - V_T$

$$I_D = \frac{k_n}{2}(V_{GS} - V_T)^2$$

If we consider the variation of I_D within the saturation region

$$\boxed{I_D = \frac{k_n}{2}(V_{GS} - V_T)^2 (1 + \lambda V_{DS})}$$

where λ = Channel-length modulation effect in $1/V$.

Body Effect: Develops when $V_{SB} \neq 0$. Note that when $V_{SB} \neq 0$ the charge density Q_B is given by

$$Q_B = -\sqrt{2q\varepsilon_{Si} N_a (2|\phi_F| + V_{SB})}$$

Thus, the actual threshold voltage for an NMOS can be written as

$$\boxed{V_{T_n} = V_{TO_n} + \gamma\left[\sqrt{2|\phi_F| + V_{SB_n}} - \sqrt{2|\phi_F|}\right]},$$

where γ is the body effect coefficient and $V_{TO_n} = V_{T_n}\Big|_{V_{SB}=0}$

γ can be expressed as $\gamma = \dfrac{\sqrt{2q\varepsilon_{Si} N_a}}{C_{ox}}$ in $\sqrt{V}$

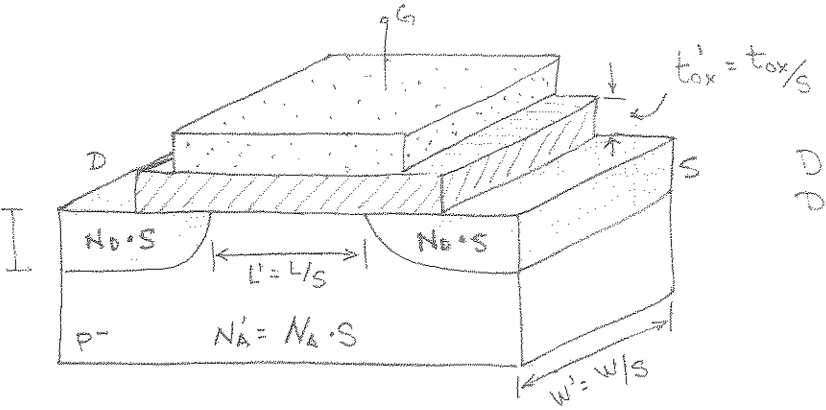
MOSFET Scaling: The operational characteristics of MOSFETs change as dimensions are reduced. Reductions enable higher packing densities.

Scaling strategies:
- Full scaling (Constant field)
- Constant-voltage scaling

In both cases, device dimensions are divided by a scaling factor $S > 1$. The value of S is limited by the available technology.

Year	1985	1987	1989	1991	1993	1995	1997	1999	2001	2003
MFS	2.5	1.7	1.2	1.0	0.8	0.5	0.35	0.25	0.13	0.09

$\underbrace{0.35 \quad 0.25 \quad 0.13 \quad 0.09}_{\text{Deep sub-micron era}}$



Dimensions: divide by S
Doping: multiply by S

Full Scaling: Preserves the magnitude of internal electrical fields

Since $E = V/X \Rightarrow$ Voltages are scaled-down
By Poisson Eqn.: $E = \frac{1}{\varepsilon}Q \Rightarrow$ Dopant concentration scaled-up

$$\left.\begin{aligned} C'_{ox} &= \frac{\varepsilon_{ox}}{t'_{ox}} = S\frac{\varepsilon_{ox}}{t_{ox}} = SC_{ox} \\ \frac{W'}{L'} &= \frac{W}{L} \\ \mu'_n &= \mu_n \end{aligned}\right\} K'_n = SK_n$$

$$V'_{DD} = V_{DD}/S \qquad V'_{T0} = V_{T0}/S \quad \Rightarrow \quad I'_{D_{lin}} = \frac{I_{D_{lin}}}{S}, \quad I'_{D_{sat}} = \frac{I_{D_{sat}}}{S}$$

$$\boxed{A' = W' \times L' = A/S^2} \qquad \boxed{P' = I' \cdot V' = P/S^2}$$

- Very attractive power and area characteristics
- Delays get also improved $t'_{PD} = C'_L \frac{\Delta V'}{I'} = SC_{ox} \cdot \frac{A}{S^2} \frac{\Delta V/S}{I/S} = \frac{t_{PD}}{S}$
- However, the voltage scaling of this approach is not straightforward
  - Supply voltages cannot be arbitrarily modified: multiplicity of V_{DD}
  - Threshold voltages cannot be made arbitrarily small (Low threshold transistors are difficult to turn off)

Constant Voltage Scaling: All dimensions are reduced by S while preserving the voltage levels.
In this case, dopant concentrations must be increased by S^2 to preserve charge-field relations.

$$I'_{D_{lin}} = S \cdot I_{D_{lin}}, \quad I'_{D_{sat}} = S \cdot I_{D_{sat}}, \quad P' = SP$$

Current density $\sigma' = I'/A' = SI/(A/S^2) = S^3\sigma$ ← Riability concerns.

power density $= P'/A' = S^3 \, P/A$

Delay on the other hand is reduced by a factor of S^2

A commonly applied scaly strategy uses different scaling factors for dimensions and voltages (General Scaling)

- Voltages reduced by a factor U
- Dimensions scaled by a factor S

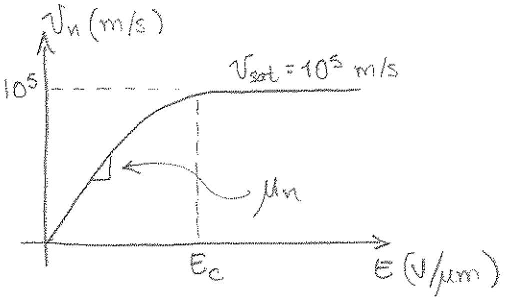
Short Channel Effects

A transistor is considered short channel when $L_{eff} \approx x_j$

L_{eff} = Effective channel length (L in Uyemura)
x_j = Depth of source-drain difussion

Under a constant voltage, scaling down L increases the transversal channel electric field. This leads to a deficiency named <u>velocity saturation</u>.

Although long channel devices are ok with $v = \mu E$, there is a value of E, namely E_c where v saturates.



E_c depends on dopant concentration. For electrons it can be between 1 & 5 V/μm

E_c is easily reached in short-channel devices.

$E_{c_p} > E_{c_n}$ due to lower hole mobility

$$v_n = \begin{cases} \dfrac{\mu_n E}{1 + E/E_c} & \text{for } E < E_c \\ v_{sat} & \text{for } E \geqslant E_c \end{cases}$$

Considering the $I_D = -v_n(x) Q_i(x) W$

$v_n \rightarrow$ Electron velocity
$Q_i \rightarrow$ Induced channel charge

with $Q_i = -C_{ox}\left[V_{GS} - V(x) - V_T\right]$. Integrating over L, the expression for I_{D_P} results in the ohmic region:

$$I_D = \frac{\mu_n C_{ox}}{1 + (V_{DS} / E_c \cdot L)} \left(\frac{W}{L}\right)\left[(V_{GS} - V_T)V_{DS} - V_{DS}^2/2\right]$$

So

$$\boxed{I_{D_{lin}} = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right] K(V_{DS})}$$

where

$$K(V) = \frac{1}{1 + \left(\frac{V}{E_c \cdot L}\right)}$$

$$K: \begin{cases} \sim 1 & \text{for long channel devices} \\ < 1 & \text{for short " "} \end{cases}$$

When V_{DS} increases and E reaches the critical value E_c and carriers become velocity saturated

V_{DSAT} = V_{DS} for velocity saturation

I_{DSAT} = I_D when velocity saturation occurs

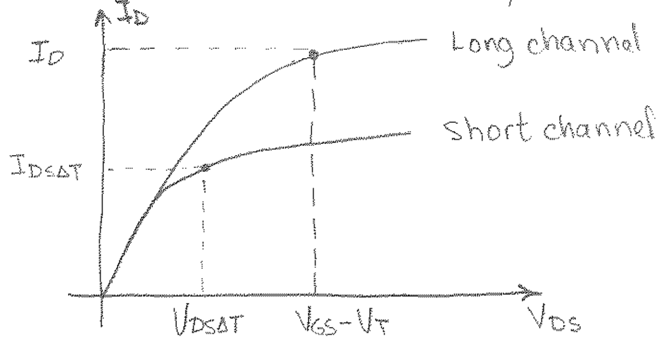
$$I_{DSAT} = v_{sat} C_{ox} W \cdot (V_{GT} - V_{DSAT}) = K(V_{DSAT}) \mu_n C_{ox} \frac{W}{L}\left[V_{GT} V_{DSAT} - \frac{V_{DSAT}^2}{2}\right]$$

where $V_{GT} = V_{GS} - V_T$

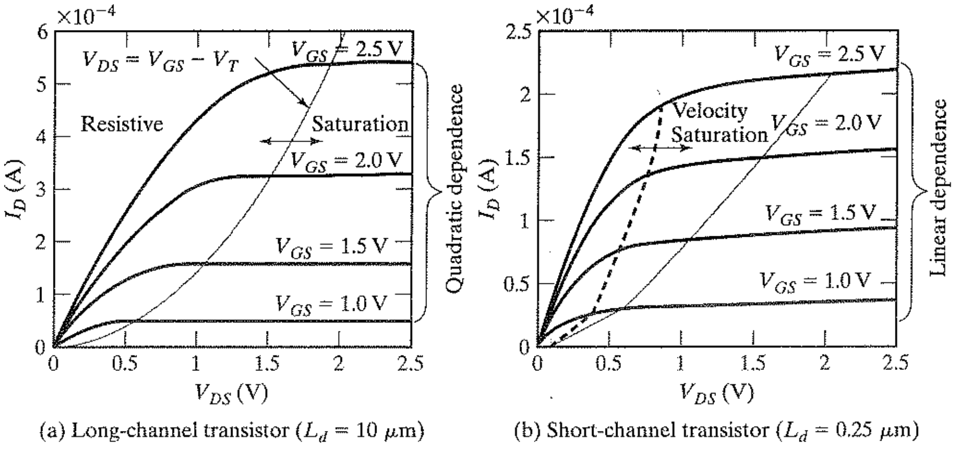
Solving for V_{DSAT}, $\quad V_{DSAT} = K(V_{GT}) V_{GT} = \dfrac{V_{GT}}{1 + \dfrac{V_{GT}}{E_c \cdot L}}$

Therefore

- Large enough $V_{GT} = V_{GS} - V_T$, $K(V_{GT}) \ll 1$ and thus $V_{DSAT} < V_{GT}$
- Current saturation will occur before V_{DS} reaches $V_{GS} - V_T$



- The saturation current varies linearly with V_{GS}, reducing I_{DS} and the impact of voltage changes on I_D



(a) Long-channel transistor (L_d = 10 μm) (b) Short-channel transistor (L_d = 0.25 μm)

Figure 3-19 *I-V* characteristics of long- and a short-channel NMOS transistors in a 0.25 µm CMOS technology. The (*W*/*L*) ratio of both transistors is identical and equals 1.5. Observe the difference in the *y*-axis scale.

A MODEL FOR MANUAL ANALYSIS OF SHORT CHANNEL DEVICES [Rabaey 03] pp 94-103

Assumptions

1) v saturates abruptly at E_c

$$v = \begin{cases} \mu_n E & \text{for } E < E_c \\ v_{sat} = \mu_n E_c & \text{for } E \geq E_c \end{cases}$$

2) V_{DSAT} is constant and given by

$$V_{DSAT} = L E_c = \frac{L\, v_{sat}}{\mu_n}$$

These assumptions allow the usage of the same equations in the linear region as in long channel devices, and estimate I_{DSAT} as:

$$I_{DSAT} = v_{sat} C_{ox} W \left(V_{GS} - V_T - \frac{V_{DSAT}}{2}\right)$$

Note independence from L. v_{sat} is know, and V_{DSAT} given from process parameters (or computed from 2nd assumption)

Unifying the long and short channel equations

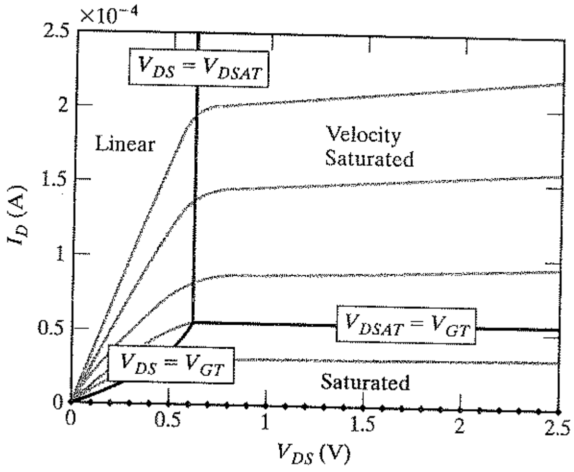
1) $I_D = 0$ for $V_{GT} \leq 0$

2) $I_D = \mu_n C_{ox} \frac{W}{L}\left(V_{GT} \cdot V_{min} - \frac{V_{min}^2}{2}\right)(1 + \lambda V_{DS})$ for $V_{GT} > 0$

where: $V_{min} = \min\{V_{GT}, V_{DS}, V_{DSAT}\}$

$V_{GT} = V_{GS} - V_T$

$V_T = V_{T0} + \gamma\left(\sqrt{2|\phi_F| + V_{SB}} - \sqrt{|2\phi_F|}\right)$



Provides for 3 different on regions

1) $V_{DS} < V_{GS} - V_T$ and $V_{DS} < V_{DSAT}$
$\Rightarrow$ Linear Operation

2) $V_{DS} > V_{GS} - V_T$ and $V_{GS} - V_T < V_{DSAT}$
$\Rightarrow$ Saturated Operation

3) $V_{DS} > V_{GS} - V_T$ and $V_{GS} - V_T > V_{DSAT}$
$\Rightarrow$ Velocity Saturated