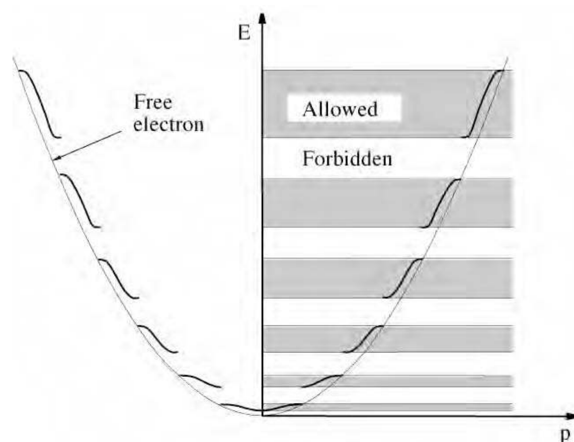
1. The MOS Transistor

Analog Design for CMOS VLSI Systems

Franco Maloberti

Electrical Conduction in Solids

- ➢The band diagram describes the energy levels for electron in solids.
- ➢The lower filled band is named Valence Band.
- ➢The upper vacant band is named conduction band.
- ➢The distance between valence and Conduction band is the energy gap.



Energy Gap in Solids

Material	Energy Gap
Metal	none
Semiconductor	0.5-3 eV
Insulator	> 3 eV

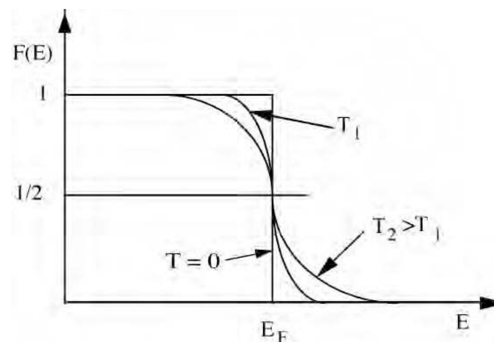
Fermi-Dirac Statisitcs

Gives the probability of occupation of energy levels:

$$F(E) = \frac{1}{1 + e^{(E - E_F)/kT}}$$

E_F is the Fermi energy level

Fermi-Dirac distribution at different temperatures:



At the Fermi level $F(E_F) = 1/2$.

Let Z(E) be the energy level distribution; the number of electrons in the energy interval E, E + dE is given by:

$$N(E)dE = Z(E)F(E)dE$$

The number of electrons in the conduction band is:

$$n = \int_{E_c}^{E_t} N_e(E)dE = \int_{E_c}^{\infty} N_e(E)dE = \int_{E_c}^{E_t} Z(E)\frac{dE}{1+e^{(E-E_F)/kT}}$$

similarly the number of holes, p, in the valence band:

$$p = \int_{E_b}^{E_v} N_h(E)dE = \int_{-\infty}^{E_v} N_h(E)dE = \int_{-\infty}^{E_v} Z(E)\frac{e^{(E-E_F)/kT}dE}{1+e^{(E-E_F)/kT}}$$

If the Fermi level is in the middle of the energy gap, the material is referred to as intrinsic, and we have:

$$n = p = n_i$$

n_i is strongly dependent on the temperature. For the silicon (empirical relationship):

$$n_i = 3.954 \cdot 10^{16} T^{3/2} e^{-1.21q/kT}$$

at room temperature $n_i = 1.42 \cdot 10^{10}$ cm^{-3}

If donor or acceptor impurities are added to the semiconductor, localized energy levels are set in the forbidden gap.
The activation energy are:

III Group	Activation Energy	V Group	Activation Energy
B	0.045 eV	P	0.045 eV
Al	0.067 eV	As	0.054 eV
Ga	0.072 eV	Sb	0.039 eV
In	0.160 eV		

Because of extremely low activation energy, even a low temperature one kT is enough to ionize the donor or the acceptor atoms (kT = 0.025 eV at 300 K).

The electrons (or holes) concentration increases in the conduction or valence band.

At room temperature:

for n-type $n \sim N_D$

for p-type $p \sim N_A$

The Fermi level is shifted with respect to the intrinsic level of the amount:

$\Phi_F = -\frac{kT}{q}\ln\left(\frac{n_i}{N_A}\right)$ for p-doping [V]; $\Phi_F = -\frac{kT}{q}\ln\left(\frac{N_D}{n_i}\right)$ for n-doping [V];

Properties of Silicon

Property	Value	Dimensions
Atomic density	$5 \cdot 10^{22}$	Atoms/cm^3
Density	2.33	g/cm^3
Atomic weight	28.1	g/mole
Reticular constant	0.543	nm
Thermal conductivity	1.41	Ω/cm ºC
Intrinsic resistivity	$2.5 \cdot 10^{5}$	Ω·cm
Relative dielectric constant	11.9	_
Absolute dielectric constant	$8.86 \cdot 10^{14}$	F/cm

Conductivity

$$\sigma = \sigma_n + \sigma_p = q(n\mu_n + p\mu_p)$$

for a doped material we have

$n \sim N_D$ for n-doping

$p \sim N_A$ for p-doping

$$n \cdot p = n_i^2$$

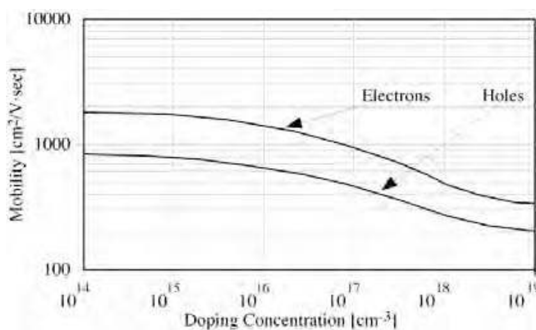
hence:

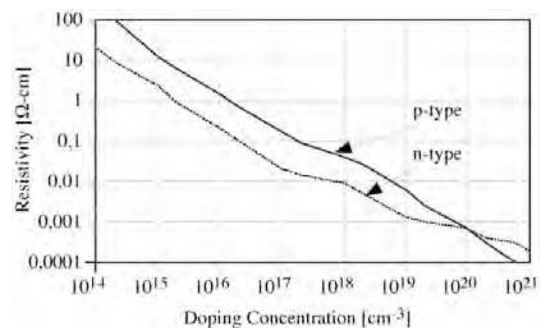
$\sigma = qN_D\mu_n$ for n-doping

$\sigma = q\ _{N}A\mu_p$ for p-doping

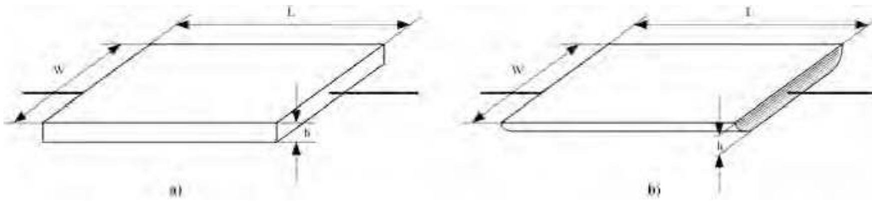
Mobility

The following figures show the surface mobility of electrons and holes as a function of the doping (at room temperature) and a resistivity as a function of the doping (at room temperature).





Resistance of thin layers

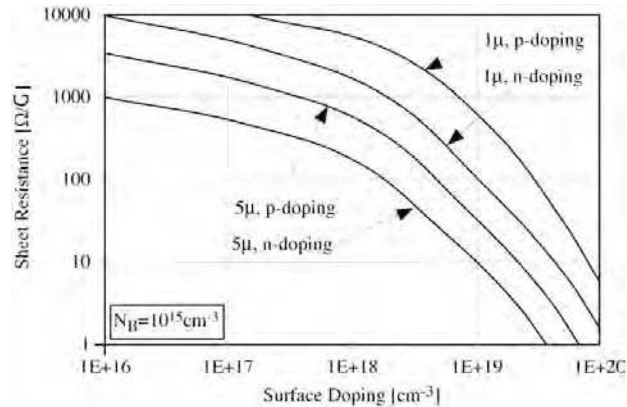


Homogeneous material

$$R = \frac{\rho L}{hW} = R_{\square}\frac{L}{W}$$

Diffused layer

$$G = \frac{1}{R} = \frac{W}{L}\int_0^h \sigma(z)dz = \frac{1}{R_{\square}}\cdot\frac{W}{L}$$



Polysilicon

Grown from pyrolytic decomposition of silane (SiH_4) at about 600°C.

The polycrystalline structure is made of monocrystal grains size in the range of 0.1 - 1 μm.

The typical layer are 200 - 600 nm thick with long term standard deviation in the 2% range.

The mobility is low because of the grain border resistance (30-40 cm^2/Vs).

In order to have a low sheet resistance the polysilicon must be strongly doped (10^{20}-10^{21} cm^{-3}). Part of the doping saturates the localized levels due to the grain border. The sheet resistance is in the range 20 - 40 Ω/☐.

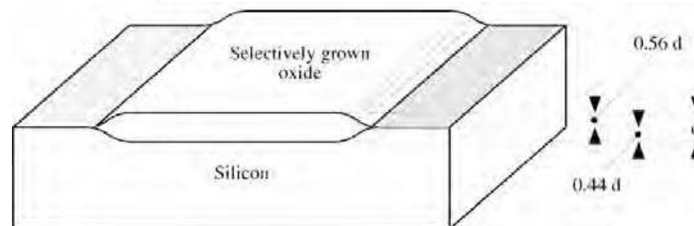
The sheet resistance can be reduced by using sandwich layers (polysilicide) made of 200 nm of polysilicon covered with a film of refractory metal silicide (WSi_2, $MoSi_2$, $TiSi_2$). The sheet resistance is reduced to 1 - 5 Ω/☐.

Silicon dioxide

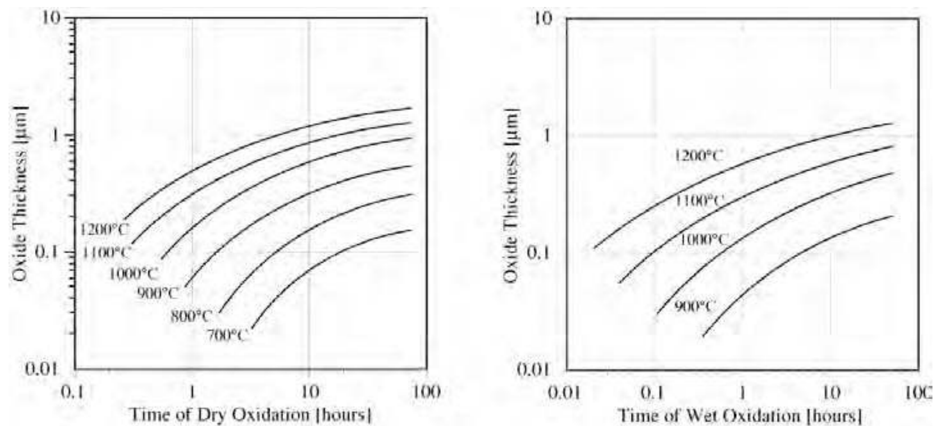
Thermally grown from silicon in dry or wet conditions at 800 - 1100°C.

Property	Value	Dimension
density	2.22	g/cm^3
dielectric strength	$2 - 8 \cdot 10^6$	V/cm
resistivity (at 300°K)	$10^{15} - 10^{17}$	Ω cm
relative dielectric constant	3.4 - 4.2	—

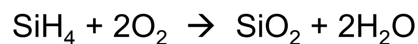
The silicon dioxide grown determine a silicon consumption: if *d* is the thickness of grown oxide, 0.44·*d* of silicon is consumed.



Growth speed:



Silicon dioxide can also be grown from chemical vapour deposition (CVD):



$$SiH_4 + 2O_2 \rightarrow SiO_2 + 2H_2O$$

by pyrolytic decomposition of silane in the presence of oxygen,
at atmospheric pressure (AP-CVD) or at low pressure (LP-CVD).

The temperature ranges from 300 to 500°C.

Growth speed, about one order of magnitude larger than the one of thermal dioxide.

Charge voltage hysteresis effect when deposited on silicon (not suitable for capacitors).

For surface protection p-doped to compensate the sodium ions action.

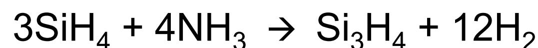
Property	Value	Dimension
density	2.22	g/cm^3
dielectric strength	$2 - 8 \cdot 10^6$	V/cm
resistivity (at 300°K)	$10^{15} - 10^{17}$	Ω cm
relative dielectric constant	2.7 - 4.2	—

Long term standard deviation 5 - 6%

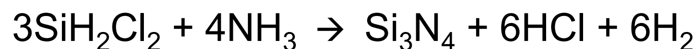
Silicon nitride

Its major use is to protect surface.

It is grown by decomposition of silane or dichlorosilane and ammonia at 700 - 800°C.



$$3SiH_4 + 4NH_3 \rightarrow Si_3H_4 + 12H_2$$



$$3SiH_2Cl_2 + 4NH_3 \rightarrow Si_3N_4 + 6HCl + 6H_2$$

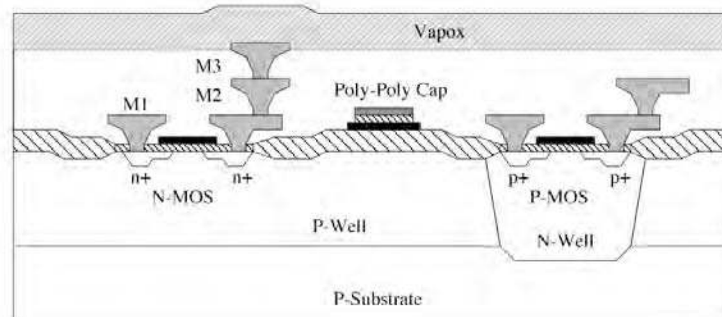
Growth speed: 10 - 20 nm/min

Resistivity: $10^{14} - 10^{16}$ Ω/cm

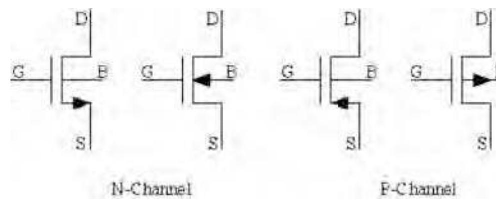
Dielectric strength: 5 - 10 MV/cm

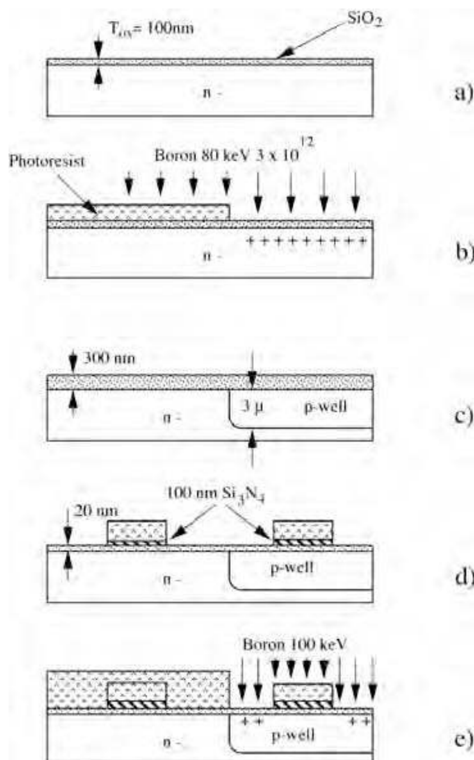
Long term standard deviation: 3 - 4%

CMOS technology



Symbols of the MOS transistors





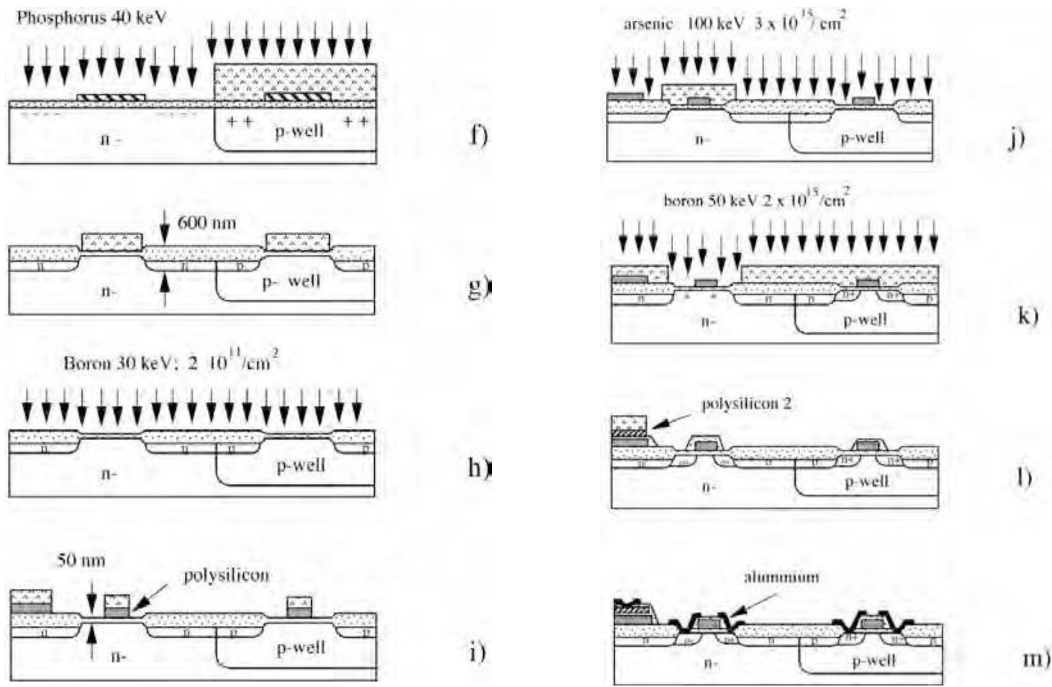
Typical CMOS process

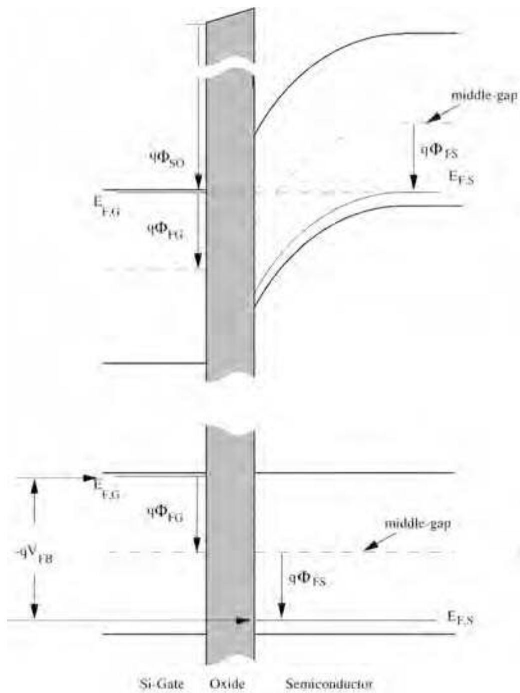
MOS technology integrates both n-channel and p-channel transistors on the same chip.

If the substrate of the circuit is n-doped, the p-channel transistors sit directly on the substrate, whereas the n-channel devices need a well.

Modern technologies use *twin-well* to make the two type of transistors inside wells regardless of substrate doping.

This approach optimize the electrical behavior at the expense of additional step.





The MOS threshold voltage

The threshold voltage is the voltage required at the gate to generate a conductive channel between source and drain.
A conductive channel is generated when the oxide-semiconductor interface is in strong inversion (bandbending = -2ϕ_{FS}).
In order to evaluate V_{Th}, the following points must be taken into account:

- Contact potential of the MOS structure.
- The energy gap E_g.
- Fixed charge trapped at the oxide-semiconductor interface.
- For an ideal MOS structure (without interface charge) the contact potential is neutralized by the so called flat band voltage V_{FB} (the band diagram in the semiconductor is flat).

$$V_{FB} = \phi_{FG} + \phi_{FS} \cong \frac{E_g}{2q} + \phi_{FS}$$

- In a real MOS structure, within a thin oxide layer at the semiconductor oxide interface, a charge Q_{SS} is trapped

$Q_{SS} = 2 \cdot 10^{-8}$ C/cm^2 for <111>
$Q_{SS} = 4 \cdot 10^{-9}$ C/cm^2 for <100>

The flat band voltage becomes:

$$V_{FB,\ real} = \frac{E_g}{2q} + \phi_{FS} - \frac{Q_{SS}}{C_{ox}}$$

The bending of the bands is obtained by depleting the semiconductor:

$$Q_{gate} = Q_{depletion}$$

$$\left(V_{Th} - V_{FB} + 2\phi_{FS}\right) C_{ox} = qN_A x_d$$

$$x_d = \sqrt{\frac{2\varepsilon}{qN_A}\left|V_{SB} - 2\phi_{FS}\right|}$$

$$V_{Th} = \frac{E_g}{2q} - \phi_{FS} + \frac{(Q_{SS} + Q_{imp})}{C_{ox}} + \gamma\sqrt{|V_{SB} - 2\phi_{FS}|}$$

$$\gamma = \sqrt{\frac{2q\varepsilon N_A}{C_{ox}}}$$

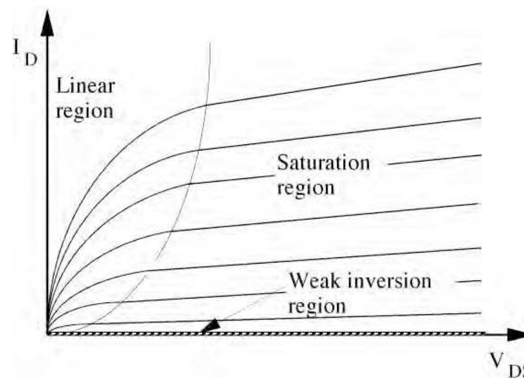
where γ is the body effect coefficient. If $V_{SB} = 0$

$$V_{Th,0} = V_{FB,\ real} - 2\phi_{FS} + \gamma\sqrt{2\phi_{FS}}$$

The threshold voltage can be expressed as:

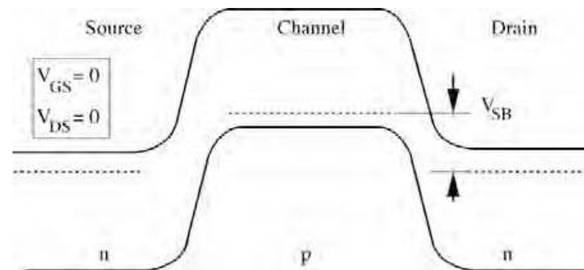
$$V_{Th} = V_{Th,0} + \gamma\left(\sqrt{|V_{SB} - 2\phi_{FS}|} - \sqrt{2\phi_{FS}}\right)$$

I-V characteristics



- Weak inversion region $V_{GS} < V_{Th}$
- Linear (or Triode) $V_{Th} < V_{GS} > V_{DS} + V_{Th}$
- Saturation region $V_{Th} < V_{GS} < V_{DS} + V_{Th}$

Weak inversion region

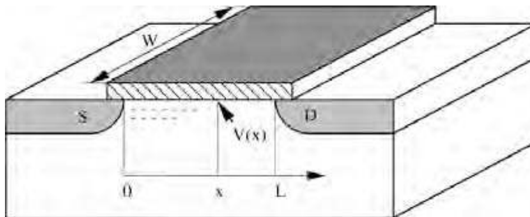


The band diagram indicates that the structure in equivalent to two back to back p-n diodes where the saturation current depends on the barrier height.

$$I_S = I_{D0}\, e^{qV_G/nkT}\, e^{-qV_B/nkT}$$

$$I_D = I_{D0}\, e^{qV_G/nkT}\, e^{-qV_B/nkT}\left(1-e^{-qV_{DS}/kT}\right)$$

Linear (or Triode) region



The voltage exceeding the threshold determines an accumulation of mobile charge on the channel (inversion region).

$$Q_{inv}(x) = C_{ox}\left(V_{GS} - V(x) - V_{TH}(x)\right)$$

$V(x)$ is the drop voltage from source to x.
The resistance of an incremental element x, $x + dx$ in the channel is:

$$dR = \frac{dx}{\sigma A} = \frac{dx}{Q_{inv}(x)\mu W}$$

The drop voltage across the element is:

$$dV = I_D dR = \frac{I_D dx}{Q_{inv}(x)\mu W}$$

V_{Th} changes along the channel due to the body effect:

$$V_{Th} = V_{Th,0} + \gamma\left(\sqrt{\left|V_{SB} - 2\phi_{FS} - V(x)\right|} - \sqrt{2\phi_{FS}}\right)$$

$$V_{DS} = \int_0^L \Delta V \, dx$$

We get:

$$I_D = \mu C_{ox}\frac{W}{L}\left[\left(V_{GS} - V_{Th,0} - \gamma\sqrt{2\phi_F}\right)V_{DS} - \frac{1}{2}V_{DS}^2 + \frac{2}{3}\left(\left|V_{SB} - 2\phi_F\right|^{3/2} - \left|V_{DB} - 2\phi_F\right|^{3/2}\right)\right]$$

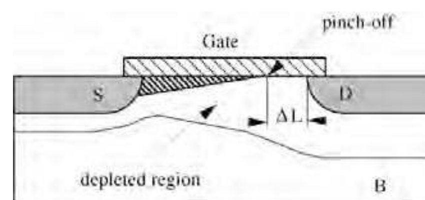
if the last term can be neglected

$$I_D = \mu C_{ox}\frac{W}{L}\left[\left(V_{GS} - V_{Th,0} - \gamma\sqrt{2\phi_F}\right)V_{DS} - \frac{1}{2}V_{DS}^2\right]$$

Saturation region

As $V(x)$ increases $Q_{inv}(x)$ decreases. Its minimum is at the drain is

$$Q_{inv}(L) = C_{ox}\left(V_{GS} - V_{TH} - V_{DS}\right)$$



if

$$V_{DS} = V_{sat} = V_{GS} - V_{TH}$$

$$Q_{inv}(L) = 0$$

the drain is in the pinch-off condition.

If $V_{DS} > V_{sat}$, the pinch-off point moves toward the source; the part of the V_{DS} voltage exceeding V_{sat} drops along the depleted region, ΔL, extending from the pinch-off to the drain.

$$\Delta L = \sqrt{\frac{2\varepsilon}{qN_A}\left(V_{DS} - V_{sat}\right)}$$

The structure can be assumed equivalent to a transistor with the pinch-off at the drain but with length reduced of ΔL. It results:

$$I_D = \frac{1}{2}\mu C_{ox}\frac{W}{L-\Delta L}\left(V_{GS}-V_{Th}\right)^2 = \frac{1}{2}\mu C_{ox}\frac{W}{L}\left(V_{GS}-V_{Th}\right)^2\frac{L}{L-\Delta L}$$

$$\frac{L}{L-\Delta L} = \frac{1}{1-\sqrt{\frac{2\varepsilon}{qN_A L^2}\left(V_{DS}-V_{sat}\right)}} = 1+\sqrt{\frac{\varepsilon}{qN_A L^2}}\sqrt{V_{DS}-V_{sat}} \cong 1+\lambda V_{DS}$$

having neglected V_{sat} with respect to V_{DS}

$$\lambda = \sqrt{\frac{\varepsilon}{qN_A L^2}} \cong \frac{10^7}{L\sqrt{N_A}}$$

λ = channel length modulation parameter ≈ $5 \cdot 10^{-2}\ V^{-1}$

hence in saturation:

$$I_D = \frac{1}{2}\mu C_{ox}\frac{W}{L}\left(V_{GS}-V_{Th}\right)^2\left(1+\lambda V_{DS}\right)$$

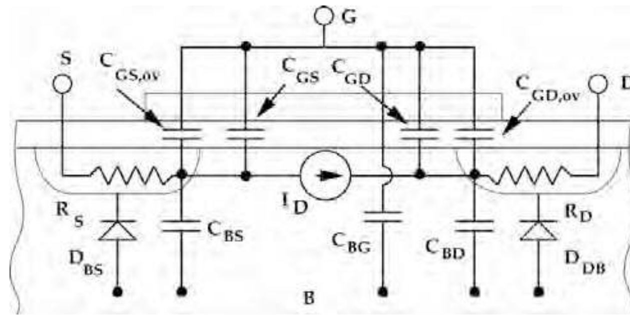
μC_{ox} is often represented by the symbol k_n (k_p) that is called the process transconductance parameter. For a given CMOS technologies we can use the following figures t_{ox} = 15 nm, μ_n = 520 cm^2/V^2s and μ_p = 180 cm^2/V^2s.

Therefore we have:

$k_n = \mu_n C_{ox}$ = 108 $\mu A/V^2$

$k_p = \mu_p C_{ox}$ = 38 $\mu A/V^2$

Large signal equivalent circuit



Non linear

- current source
- diodes
- C_{GS}, C_{BG}, C_{GD}, C_{BS}, C_{BD}

Linear (1st approximation)

- resistors
- $C_{GS,ov}$, $C_{GD,ov}$

Typically: $R_D \approx R_S \approx 10 - 50\ \Omega$

$C_{GS,ov} = C_{GD,ov} = W x_{ov} C_{ox}$

Diodes reversely biased; the reverse current is dominated by generation recombination term.

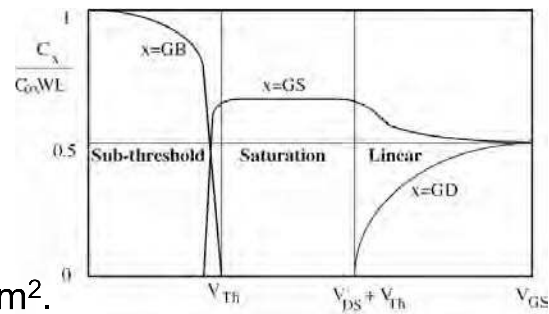
$$I_{GR} = A \frac{q n_i x_j}{2\tau_0}$$

A: area of the junction

x_j: depletion region width

τ_0: mean lifetime for minority carriers

I_{GR} doubles for an increase of 10 K. Typically at room temperature $I_{GR} / A = 10^{-16}$ A/μm^2.



Small signal equivalent circuit

Obtained by a linearization of the large signal equivalent circuit.

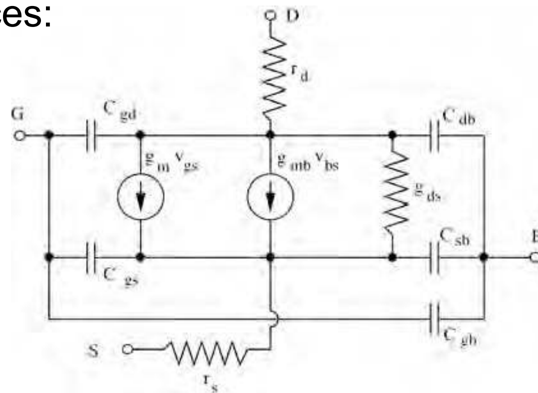
$$I_D = I_D\,(V_{GS}, V_{DS}, V_{BS})$$

The linearization of the current source generates three voltage controlled current sources:

$g_m = \partial I_D \,/\, \partial V_{GS}$
transconductance

$g_{ds} = \partial I_D \,/\, \partial V_{DS}$
drain output transconductance

$g_{mb} = \partial I_D \,/\, \partial V_{BS}$
substrate transconductance



Transconductance

- Subthreshold region (like a bipolar transistor):

$$g_m = -g_{mb} = \frac{I_D}{n\dfrac{kT}{q}}$$

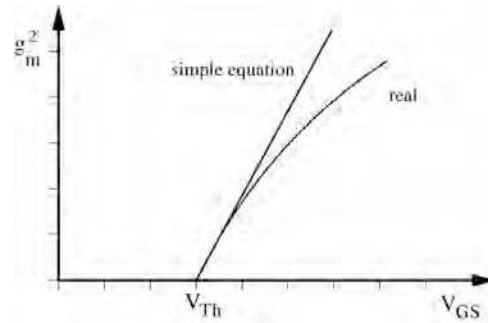
- Linear region:

$$g_m = \mu C_{ox}\frac{W}{L}V_{DS}$$

- Saturation region:

$$g_m = \mu C_{ox}\frac{W}{L}\left(V_{GS} - V_{Th}\right) = \frac{2I_D}{V_{GS} - V_{Th}} = \sqrt{2\mu C_{ox}\frac{W}{L}I_D}$$

In the real situation g_m is smaller than the value predicted by the given simple equation. Moreover g_m changes because of the dependence of μ from the temperature, the transversal and the mean lateral electric field.



$$\mu = \mu_0 \left(\frac{T}{T_0}\right)^{-\frac{3}{2}} \left(1 + \frac{E_y}{E_{crit}}\right)^{-m} \frac{1}{1 + \frac{V_{DS}}{LE_{sat}}}$$

SPICE uses the fitting equation:

$$\mu = \mu_0 \left(\frac{u_{crit}\varepsilon_{si}}{C_{ox}\left(V_{GS} - V_{on} - u_{tra} - V_{DS}\right)}\right)^{u_{exp}}$$

Drain output conductance (g_{ds})

- Linear region:

$$g_{ds} = \mu C_{ox} \frac{W}{L}\left(V_{GS} - V_{Th} - V_{DS}\right)$$

- Saturation region (first order):

$$g_{ds} = \lambda I_D$$

Second order effects

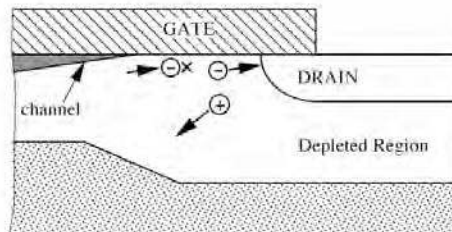
- The channel length reduction has been calculated taking into account only the lateral drop voltage. A more accurate analysis gives:

$$\frac{1}{\Delta L} \cong \frac{1}{\Delta L_{(1^{st}\ order)}} + \frac{C_{ox}}{\varepsilon_S}\left[\frac{\alpha\left(V_{DS} - V_{GS}\right) + \beta\left(V_{GS} - V_{sat}\right)}{V_{DS} - V_{sat}}\right]$$

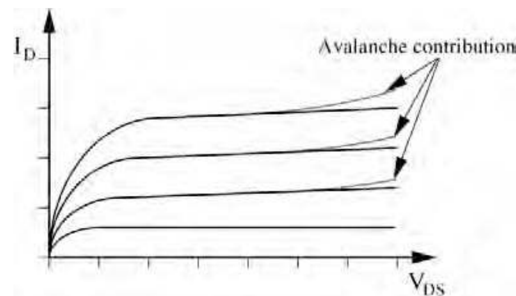
- The threshold voltage depends on the V_{DS} (short channel)
- The mobility depends on the lateral mean field V_{DS} / L
- The avalanche effect increases I_D

Avalanche current

Mobile charges, accelerated by electric field in the drain depleted region, creates, by impact ionization, electron-holes pair.

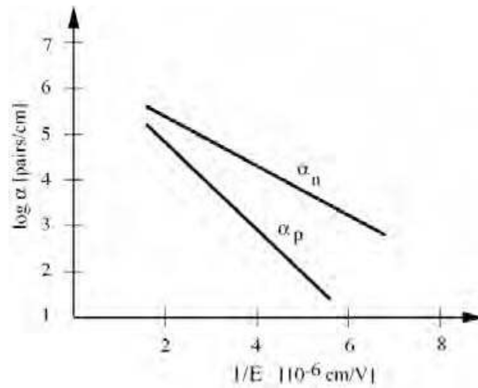


The drain current I_D and a substrate current I_B increase due to this contribution.



- The substrate current may contribute to latch-up
- The device noise increases
- The output impedance decreases
- Carriers can be trapped on the oxide and V_{Th} changes (hot electron effect)

- Avalanche current worse in n-channel



More accurate expression of the output conductance:

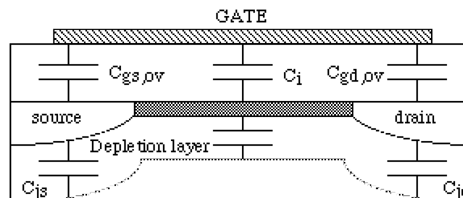
$$g_{ds} = \lambda I_D - g_m \frac{\partial V_{Th}}{\partial V_{DS}} + \frac{I_D}{\mu}\frac{\partial \mu}{\partial V_{DS}} + \frac{\partial I_S}{\partial V_{DS}}$$

(first order) (short channel) (velocity saturation) (avalanching)

Capacitances

- Linear region:

$$C_i = C_{ox} WL$$



$$C_{dep} = \frac{\varepsilon_{Si}}{x_{dep}} WL$$

$$C_{gs} = C_{gsov} + \frac{C_i}{2}$$

$$C_{gd} = C_{gdov} + \frac{C_i}{2} \qquad C_{sb} = C_{js} + \frac{C_{dep}}{2} \qquad C_{db} = C_{jd} + \frac{C_{dep}}{2}$$

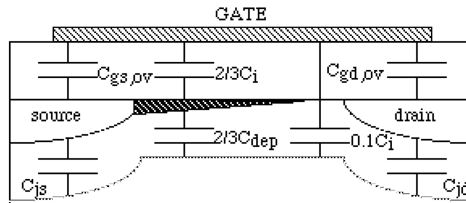
- In the saturation region

$$C_{gs} = C_{gsov} + \frac{2C_i}{3}$$

$$C_{gd} \cong C_{gdov}$$

$$C_{sb} = C_{js} + \frac{2C_{dep}}{3}$$

$$C_{db} = C_{jd}$$



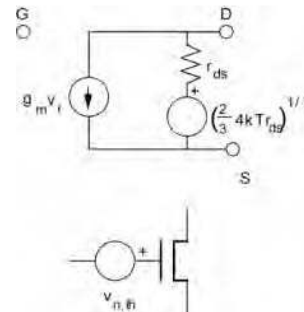
$$C_j = \frac{C_{j0}}{\sqrt{1 - \frac{V}{\phi_T}}} \qquad C_{gb} \cong \frac{1}{10} C_i \qquad \phi_T = \frac{KT}{q} \ln\left(\frac{N_D NA}{n_i^2}\right)$$

Noise

Thermal noise

Due to the finite output resistance:

$$\frac{\overline{v_{nth}^2}}{\Delta f} = 4\gamma kT \frac{1}{g_m}$$



γ = 2/3, for I_D = 50 μA, (V_{GS} - V_{Th}) = 300 mV, V_{nth} = 5.6 nV/√Hz

If the bandwidth BW of the system is f_2 - f_1, the input referred noise voltage is:

$$\overline{V_n^2} = \int_{f_1}^{f_2} \frac{\overline{v_{nth}^2}}{\Delta f} df = 4\frac{2}{3} kT \frac{1}{g_m} \text{BW}$$

Flicker noise

Due to the trapping and detrapping of carriers by surface states at different energy levels.

Modeled as:

$$\frac{\overline{i_{nf}^2}}{\Delta f} = \frac{2\,K_f\,I_D}{C_{ox}^{k_c}\,L^2\,f^{\alpha}}$$

$$\frac{\overline{v_{nf}^2}}{\Delta f} = \frac{\overline{i_{nf}^2}}{\Delta f\,g_m^2} = \frac{2\,K_f\,I_D}{C_{ox}^{k_c}\,L^2\,f^{\alpha}\,g_m^2} = \frac{K_f}{\mu\,C_{ox}^{k_c+1}\,W\,L\,f^{\alpha}}$$

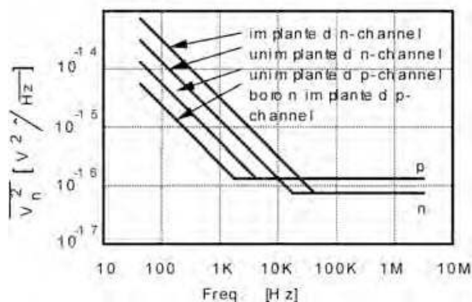
Typically $\alpha \cong 1$, $k_f \cong 1$, $k_c + 1 \cong 1$,
V_{nf} = 40 nV/√Hz at 1 kHz with $W \cdot L$ = 1000 μm².

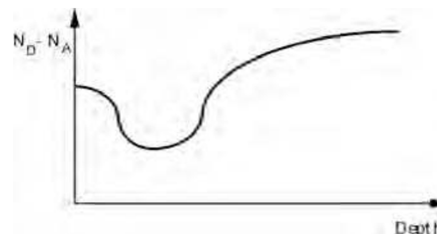
The power of the flicker noise is concentrated at low frequency.

$$\overline{V_{nf}^2} = \int_{f_1}^{f_2} \frac{\overline{v_{nf}^2}}{\Delta f}\,df = \frac{K_f}{\mu\,C_{ox}\,W\,L}\,\ln\left(\frac{f_2}{f_1}\right)$$

Noise spectra for n-channel and p-channel transistors.

Boron implanted p-MOS has low 1/*f* noise (buried channel)





Avalanche noise

Due to the statistical fluctuation in the number of carriers of the avalanche current (shot noise). Modeled as:

$$\frac{\overline{i_{nav}^2}}{\Delta f} = 2\, q\, I_{av} \qquad \frac{\overline{v_{nav}^2}}{\Delta f} = \frac{\overline{i_{nav}^2}}{\Delta f\, g_m^2} = \frac{q\, I_{av}}{\mu\, C_{ox} \frac{W}{L}\, I_D}$$

If we compare thermal noise and avalanche noise, we have:

$$\frac{v_{nav}^2}{v_{nth}^2} = \frac{q\, I_{av}}{\mu\, C_{ox} \frac{W}{L}\, I_D} \frac{g_m}{4\, \gamma\, k\, T} = \frac{V_{GS} - V_{Th}}{4\, \gamma \frac{k\, T}{q}} \frac{I_{av}}{I_D}$$

we get comparable noise if I_D / I_{av} is of the same order of $(V_{GS} - V_{Th}) / (kT/q)$. The avalanche current at V_{DS} = 5 V can be of the order of 0.5 ÷ 1 μA.

To minimize noise

- Thermal noise:
  - use large g_m (large W / L)
  - use low series resistance (connection and gate resistance)
- Flicker noise:
  - use large device area ($W \cdot L$)
  - use thin oxide (high C_{ox})
  - use "clean" technology (low N_{SS})
  - try to get buried channel
  - use p-channel devices
- Avalanche noise:
  - reduce V_{DS}
  - use p-channel devices

Layout

Rules

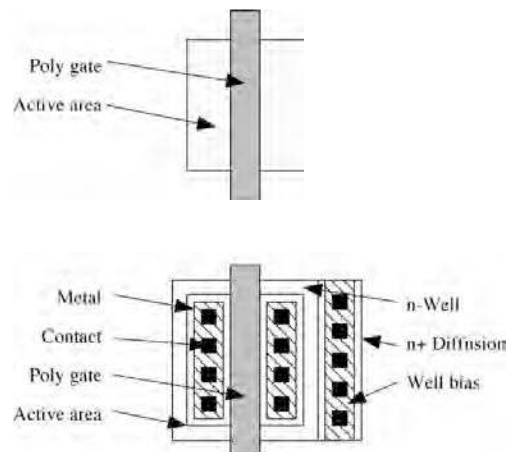
- Use poly connections only for signal, never for current because the offset $R\,I \approx 15$ mV.
- Minimize line length, especially for lines connecting high impedance nodes (if they are not the dominant node).
- Use matched structure. If necessary common centroid arrangement.
- Respect symmetries (even respect power devices).
- Only straight-line transistors.
- Separate (or shield) the input from the output line, to avoid feedback.
- Shielding of high impedance nodes to avoid noise injection from the power supply and the substrate.
- Regular shape.

Layout of transistors

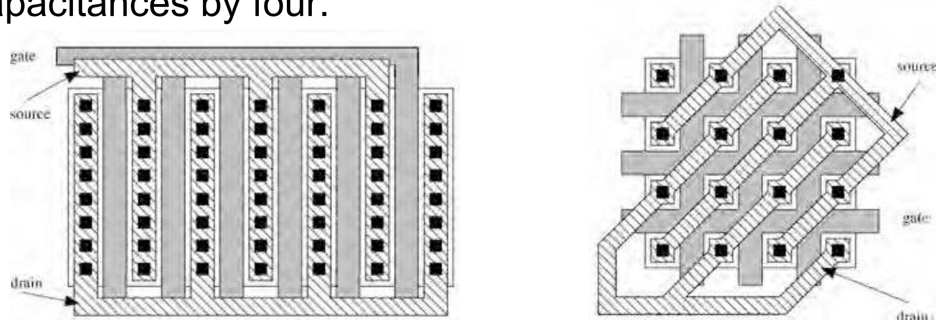
The MOS transistor is a overlap of two rectangles: active area (not protected, to originate the source and the drain) and polysilicon gate.

Key points:

- parasitic resistance at source and drain must kept as low as possible
- parasitic capacitances must be minimized
- matching between paired elements is very important

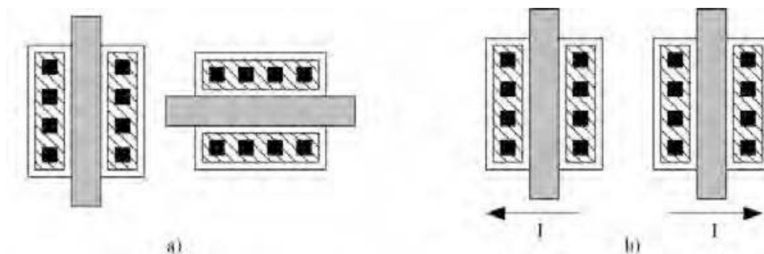


- Use multiple contacts. Many contacts placed close to each another make the surface of metal connection smoother, preventing micro-cracks in the metal.
- Splitting the transistor in a number of equal parts connected in parallel reduces the area of the transistor and its parasitic capacitances.
  Two layouts: The one on the left reduces parasitic capacitances by two; the one on the right reduces parasitic capacitances by four.

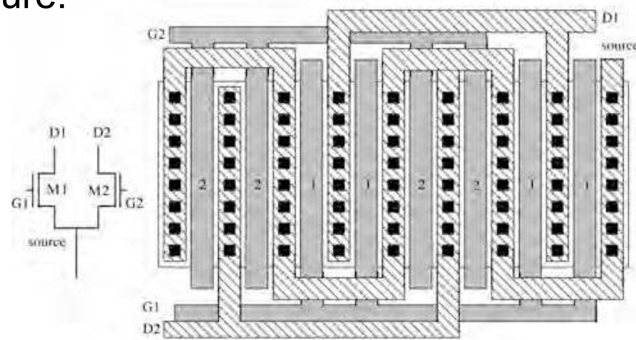


Matching is very important when we design current mirrors and differential pairs. Bad matching produces high offset.

- Transistors with different orientation match badly (left).
- Mismatch may occur if current flows in opposite directions (right).
- Physical and technological parameters may change in points of the chip that are relatively far away.



The best method of achieving good matching is shows in the following figure:



- Each transistor is split into four equal parts with a proper interleaving. For each pair of fingers of the same transistor currents flow in opposite directions.
- Any noisy signal affecting the substrate or the well should be sunk by the biasing and should not affect the circuit.