Sub-Threshold Swing

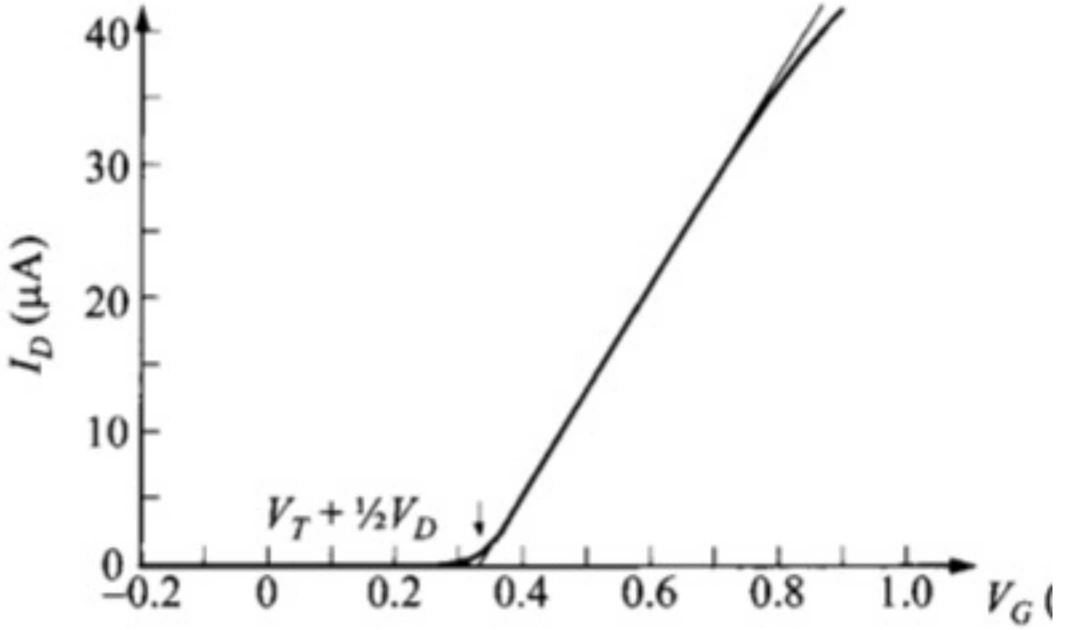
- The measure of how sharply the transistor can be turned on/off
- $\ln(I_D) = K(\beta\psi_s)$
- We define the sub-threshold slope

$$S = \ln(10)\frac{dV_G}{d(lnI_D)}$$

Since, $V_G \; - V_{FB} = \psi_s + \sqrt{2\epsilon\psi_S q N_A}/C_{ox}$

$$\frac{dV_G}{d\psi_s} = 1 + \frac{1}{C_{ox}}\left[\sqrt{\frac{\epsilon_s q N_A}{2\psi_s}}\right] = 1 + \frac{C_D}{C_{ox}}$$

$$S = \ln 10 \frac{1}{\beta}\frac{dV_G}{d\psi_s} = \ln 10 \; \frac{kT}{q}\left(1 + \frac{C_D}{C_{ox}}\right)$$



Steeper slope is obtained, when S is small. Hence, you want to work at low temperature and have small oxide thickness

Effect of Temperature

- Remember the three regions of transport in a MOSFET
  - Linear $I_D = \frac{W}{L}\mu_n C_{ox}\left(V_G \ - V_T \ - \frac{V_D}{2}\right)V_D$
  - Saturation $I_D \ = \frac{W}{2ML}\mu_n C_{ox}(V_G - V_T)^2$
  - Non-linear $I_D \ = \frac{W}{L}\mu_n C_{ox}\left(V_G - V_T \ - \frac{MV_D}{2}\right)V_D$
  - Sub-threshold $I_D \ = \frac{W}{L}\frac{\mu_n}{\beta^2}\sqrt{\frac{q\epsilon_S N_A}{2\psi_S}}\left(\frac{n_i}{N_A}\right)^2 e^{\beta\psi_S}$

Temperature affects the electron distribution in the material and hence the threshold for the strong inversion.

Effect of Temperature on Threshold

$$V_T = \phi_{ms} - \frac{Q_f}{C_{ox}} + 2\psi_p + \frac{\sqrt{4\epsilon_s q N_A \psi_p}}{C_{ox}}$$

We can assume the work function and the density of fixed charges are temperature invariant

The resulting equation is mostly dependent on ψ_p and so if we know $\frac{d\psi_s}{dT}$

Then $\frac{dV_T}{dT} = \frac{d\psi_p}{dT}\left(\frac{dV_T}{d\psi_p}\right) = \frac{d\psi_p}{dt}\left[2 + \frac{1}{C_{ox}}\sqrt{\frac{\epsilon_s q N_A}{\psi_p}}\right]$

Since, ψ_p is the separation from the intrinsic energy level,

$$\psi_p = \beta \ln \frac{N_A}{n_i^2}$$

Remember, $n_i^2 = T^3 e^{-\frac{E_G}{KT}}$

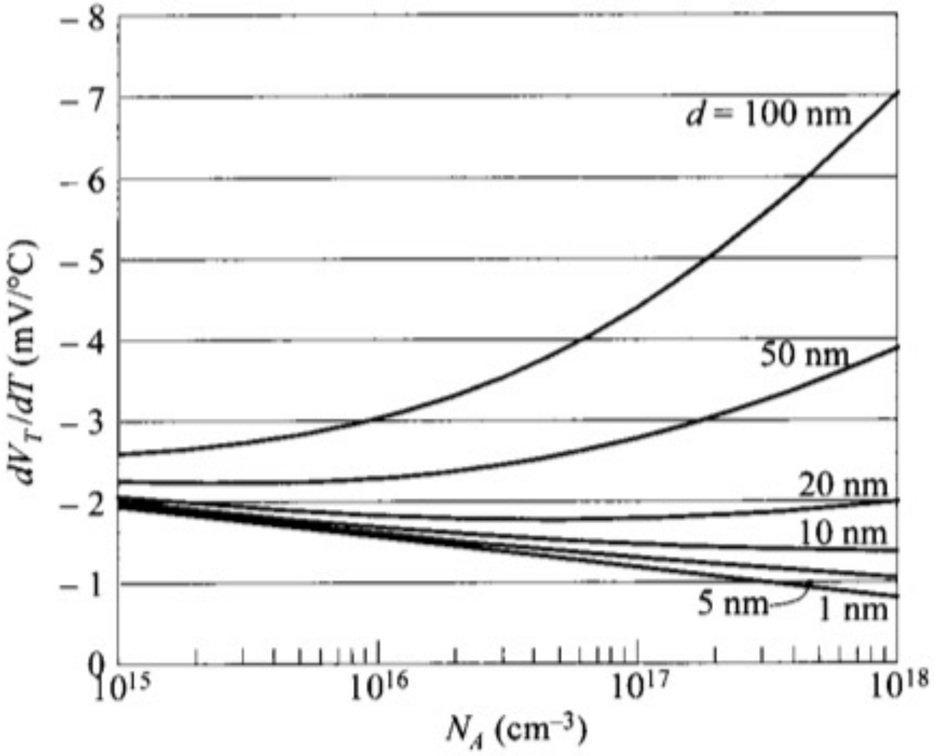
If we substitute everything $\frac{d\psi_p}{dT} = \frac{1}{T}\left[\psi_p - \frac{E_G}{2q}\right]$

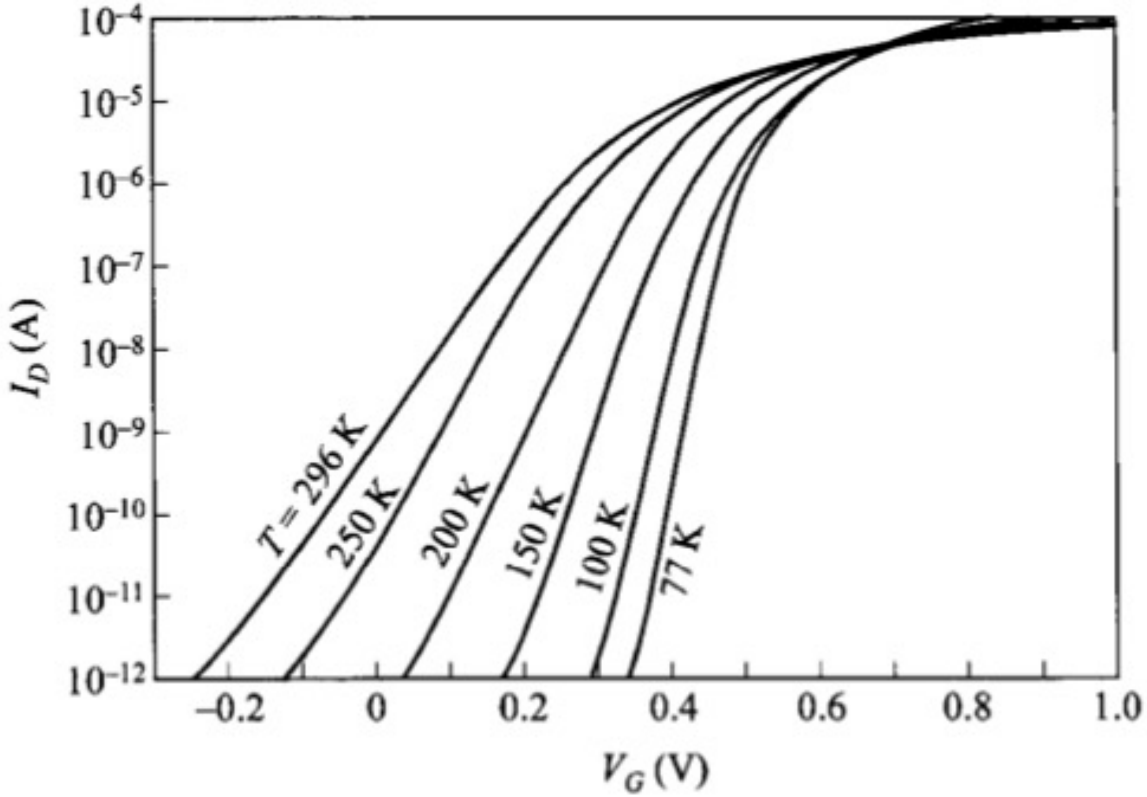
Since $\psi_p < \frac{E_{G0}}{2}$, the function $\frac{d\psi_p}{dt}$ is always negative and hence, increase in temperature reduces the threshold.

Or decreasing the temperature increases the threshold voltage. This improves the MOSFET characteristics, by increasing the noise margin.

Decreasing the temperature also reduces S and provides steep transfer characteristics.

Improvement in sub-threshold swing at 77K is by about a factor of 4





Control over threshold

The expression for threshold potential $V_T = V_{FB} - \frac{Q_D}{C_{ox}} + 2\phi_p$

The depletion charge was initially $Q_D = -\sqrt{2\epsilon_s q N_A\left(2\phi_p\right)}$

If the body of the silicon substrate is set to a potential V_B, the source-body depletion and the source-drain depletion width are changed. This creates another field countering the gate field.

This will make the gate field to work even harder, causing the threshold voltage to increase.

The final threshold voltage $V_T = V_{FB} + 2\phi_p + \frac{\sqrt{2\epsilon_s N_A\left(2\phi_p - V_{BS}\right)}}{C_{ox}}$

$$\Delta V_T = \frac{\sqrt{2\epsilon_S q N_A}}{C_{ox}}\left[\sqrt{2\phi_p - V_{BS}} - \sqrt{2\phi_p}\right]$$

Remember for a N-channel, VBS is negative to
increase the S-B depletion

Short Channel Effects

Scale the Device

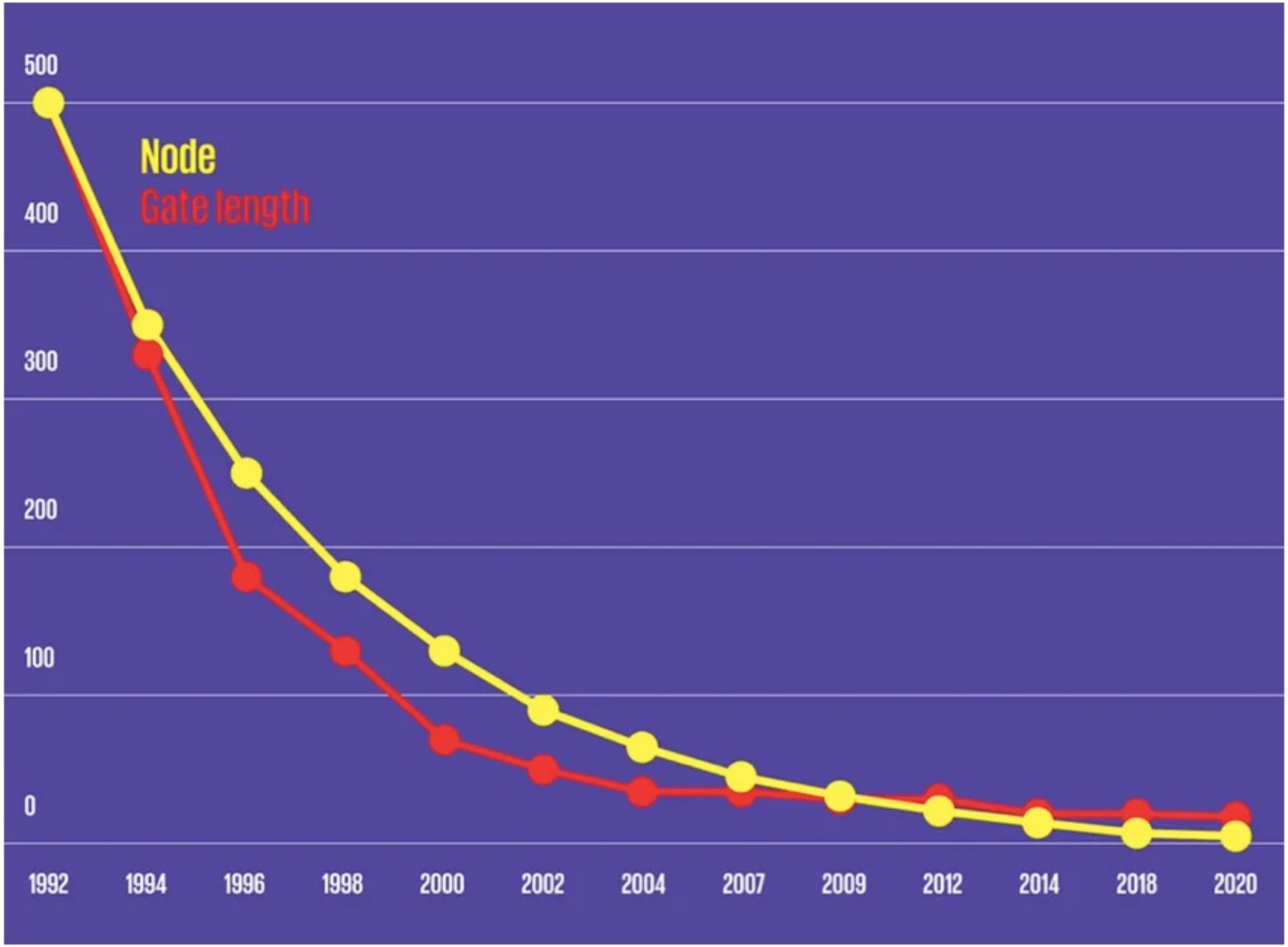
Change the dimensions of the device: Why ?

- Dennard (IBM) in 1959 suggested, if you shrink the MOSFET dimensions with an appropriate change in the other device dimensions, an improvement in packing density, speed and power dissipation can be obtained.

- If the length and width of the MOSFET channel (typically the gate dimension) is reduced by K, then the vertical dimensions of source drain junction depths, gate insulator thickness and other dimensions should be suitably scaled along with the power supply voltage as well

- However, voltage cannot be scaled independently as the transistor in a processor needs to work with other devices on the mother board as well.

- This causes enormous issues and has been the area of research in semiconductor device physics for more than 50 years now!!!



SOURCES: STANFORD NANOELECTRONICS LAB, WIKICHIP, IEEE INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS 2020

Deviation from the long-channel approximation [gradual-channel approximation]

When the devices are scaled (shrunk), if the electric field in the channel (S-D)

$$\mathcal{E}_y = \frac{V_{DS}}{L}$$

Can increase such that $\mathcal{E}_y$ becomes comparable to $\mathcal{E}_x$

Thus the two-dimensional potential at the surface deviates from GCA and the simplified expressions we obtained.

We therefore have the famous short channel effect

So what happens when the electric field gets increased ?

Mobility is the ease of an electron to respond to the electric field. For small enough fields

$$v = \mu\mathcal{E}$$

Ie, the velocity increases linearly with the applied electric field.

However, there are many other processes in the semiconductor. The electron scatters off defects, phonons and other impurities

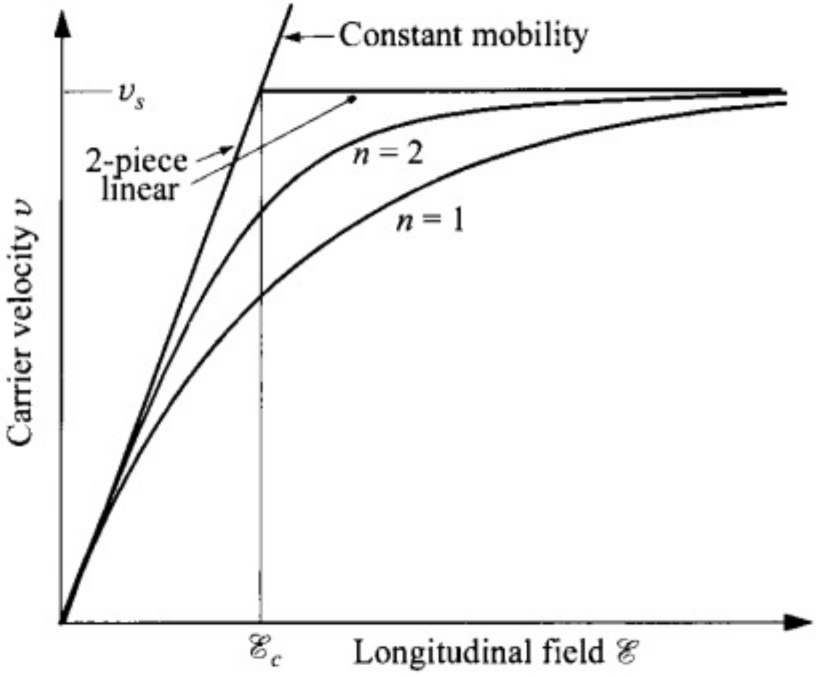
This leads to issues with velocity scaling linearly with the electric field.

After a certain point, the electron velocity cannot increase further. The scattering rate increases with the electron velocity.

This causes the velocity to saturate called as the saturation velocity v_s

In-between the linear mobility and the saturation regime, the velocity is described by the relation

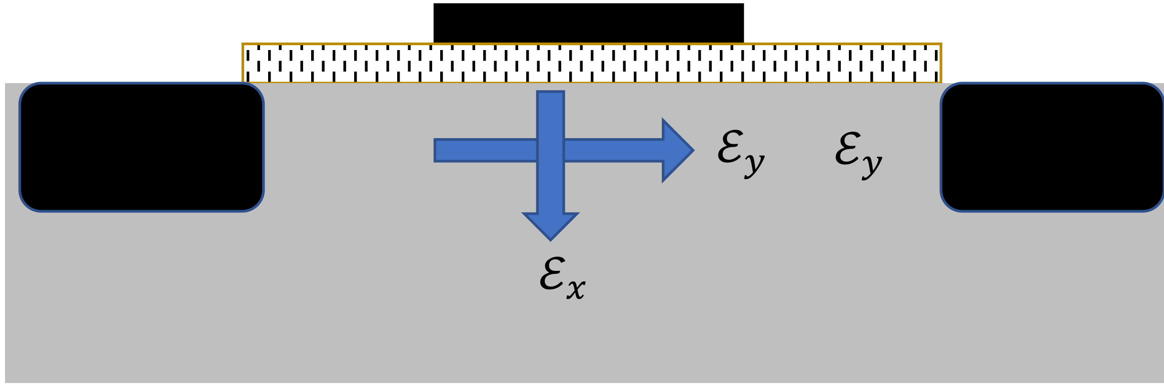
$$v(\mathcal{E}) = \frac{\mu\mathcal{E}}{\left[1 + \left(\frac{\mu\mathcal{E}}{v_s}\right)^n\right]^{\frac{1}{n}}}$$



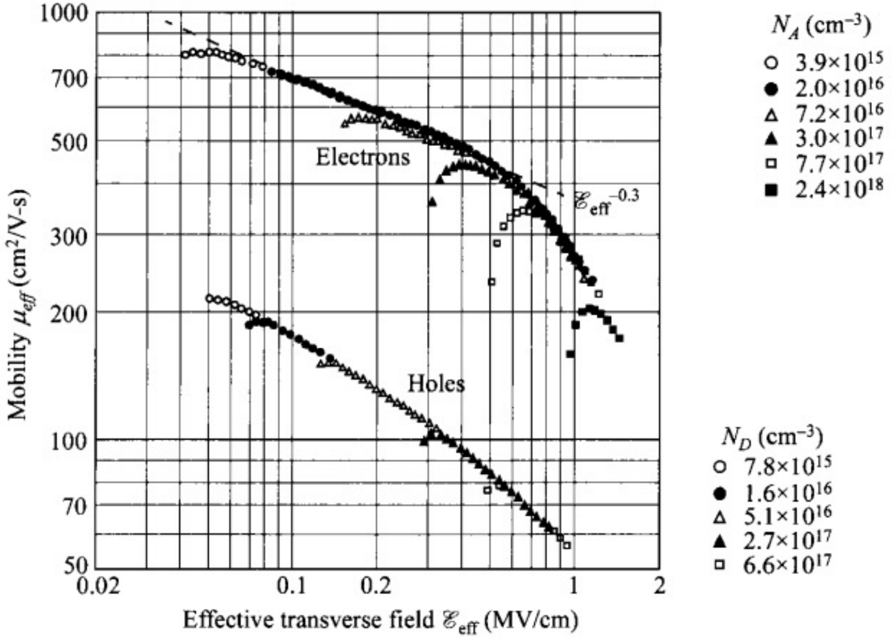
Fields in the MOSFET channel

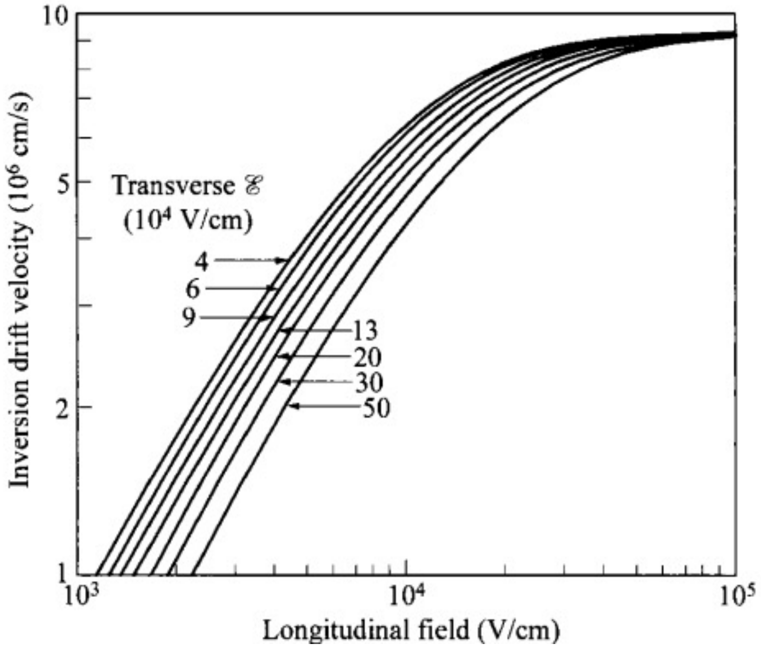
Longitudinal field $\mathcal{E}_y$ parallel to the surface

Transverse field $\mathcal{E}_x$ parallel to the gate



Experimental measurements show that the low field mobility is a unique function of $\mathcal{E}_x$





As the transverse field is increased, the velocity saturation behavior is altered significantly.
This causes strong deviations from the long-channel approximation

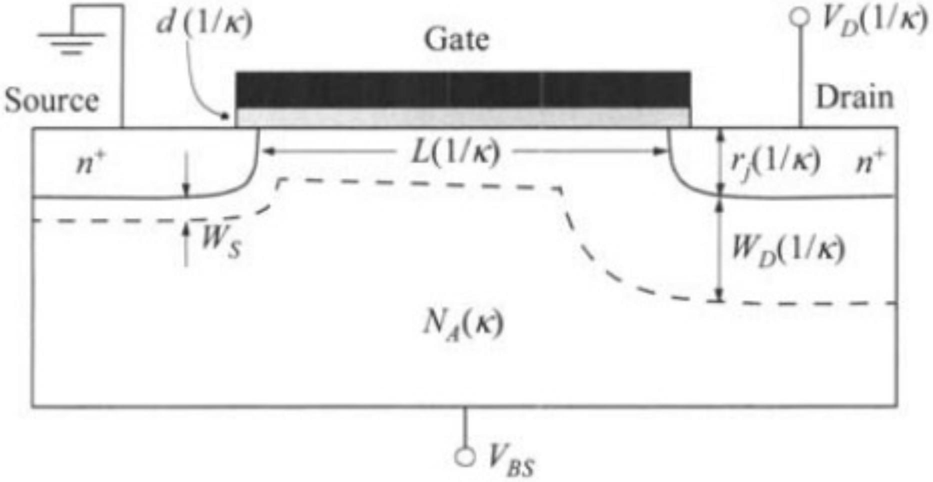
Ideal scaling rule

Ideal scaling is obtained, when you maintain the electric fields constant before and after scaling

If the channel length and the width is reduced by K

- Oxide thickness scaled by 1/K
- V_T scaled by 1/K
- V_D scaled by 1/K
- N_A scaled by K

Keeps the $\mathcal{E}$ the same in the semiconductor.



Sub-threshold swing S remains unaffected since $1 + C_D/C_{ox}$ remains unaffected.

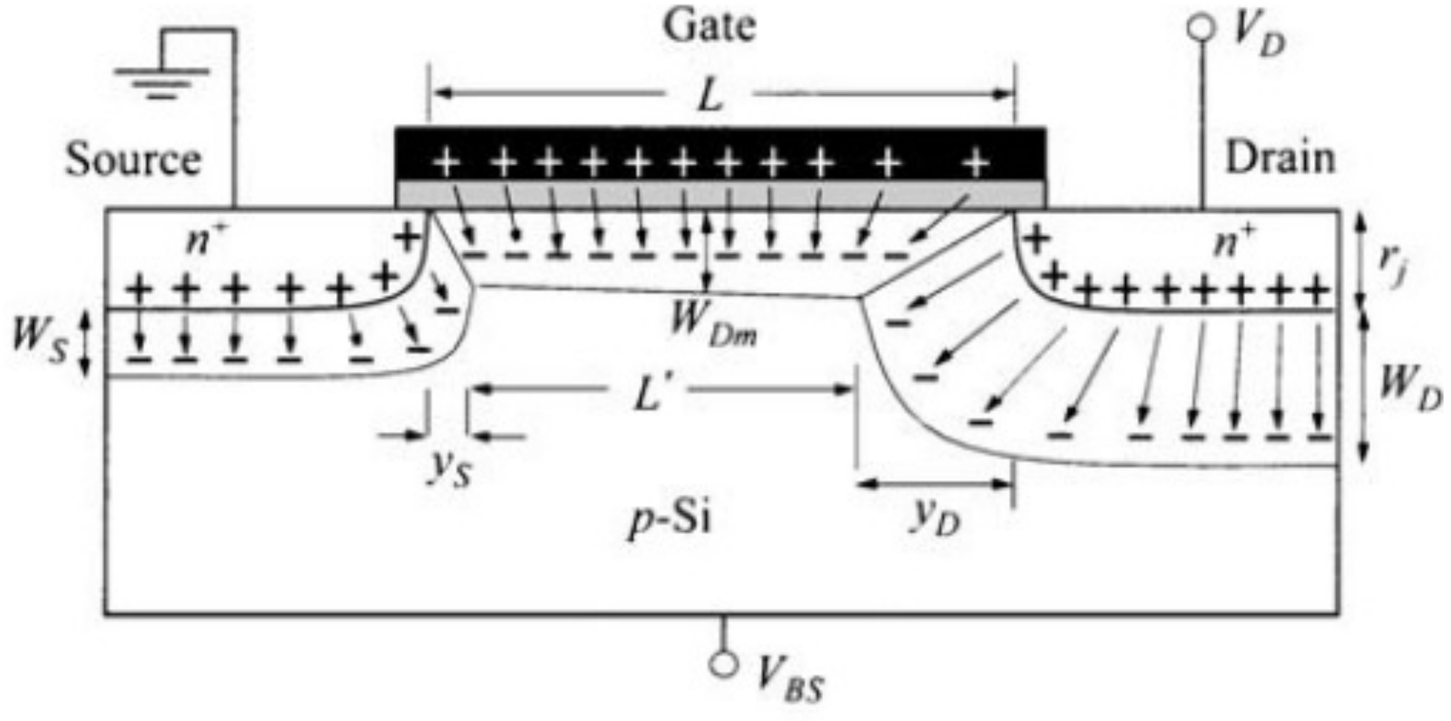
Though ideal, it all seems, the properties could not be scaled in practice.

- V_T cannot be scaled so easily by the doping density do to log dependence. 10% change is obtained by an order of magnitude change in the doping density
- Gate oxide cannot be made very thin due to practical difficulties, defects and tunneling through very thin films
- The source and drain contact resistances scale with the shrinking contacts
- Increasing the substrate doping density reduces the depletion width in the drain side, this causes P-N junction breakdown
- The voltage supply cannot be changed independently, it depends on several other factors.

Charge Sharing

Till now, the gate charge was assumed to be totally countered by the semiconductor depletion and inversion.

However, in practice, the gate field will also be terminated by source and drain depletion .



(a)

$V_T = V_{FB} + 2\phi_p + \frac{Q_D}{c_{ox}A}$ where A is the W*L the area of the transistor in the ideal case.

However, as discussed before, for the long channel, the source and drain depletion widths are much smaller in comparison with L

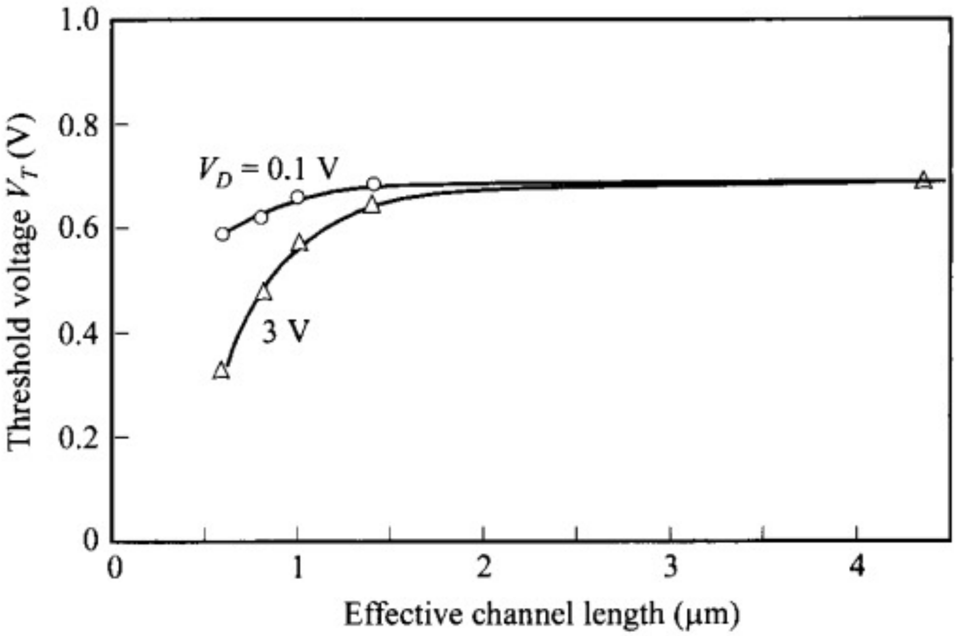
Hence $Q_D = qAN_AW_{Dm}$ where W_{DM} is the width of the maximum depletion

For a very simple first order analysis:

The shift in threshold voltage to the length decreasing to L' is given by

$$\Delta V_T = -\frac{qN_AW_{Dm}}{C_{ox}}\left[1 - \frac{(L+L')}{2L}\right]$$

Since, L + L' < 2L, the threshold always reduces with drain bias



Channel Length Modulation

The effective channel length is reduced due to the drain depletion.

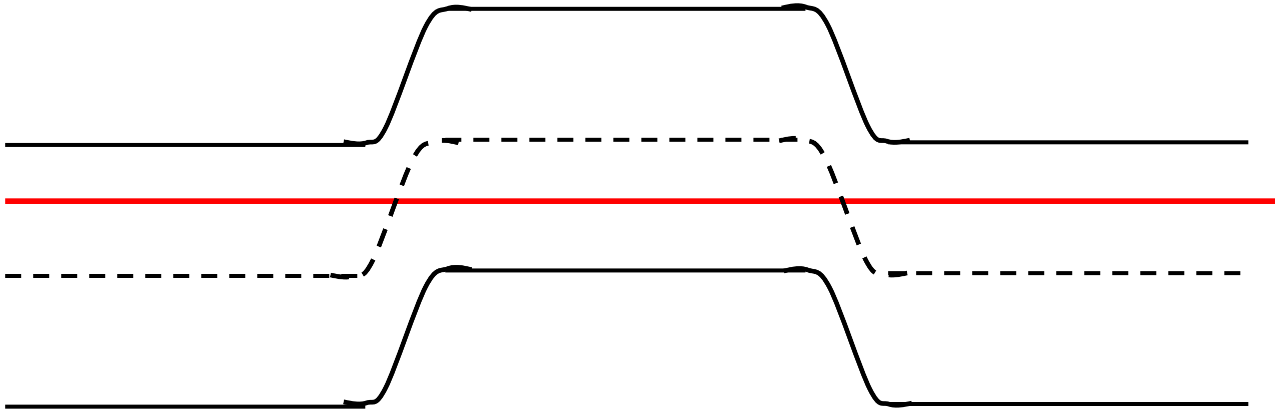
In a large channel $L \cong L'$ and hence the current remains the same.

However, in short channels, L is small and hence $L_{eff} = L' = L - y_s - y_D$

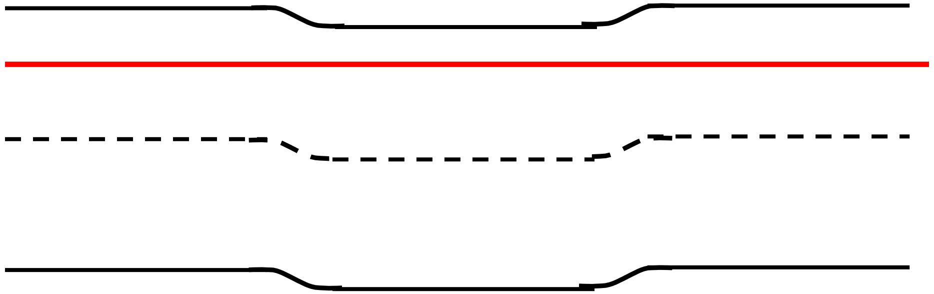
Since, there is drain dependence on the length, the current does not saturate!

The total current increases with the drain bias due to reducing effective channel length

Drain Induced Barrier Lowering



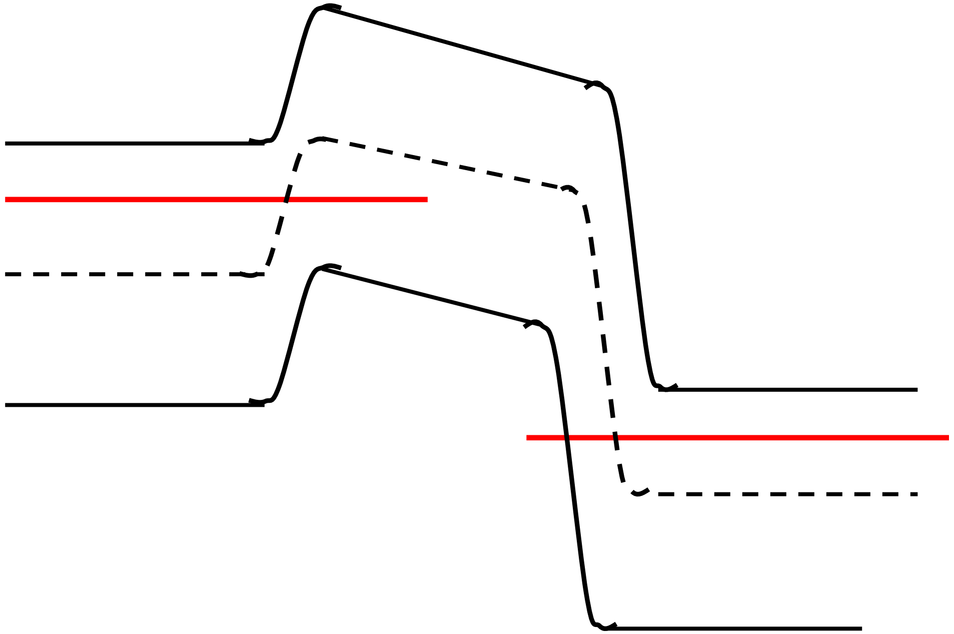
At Bulk

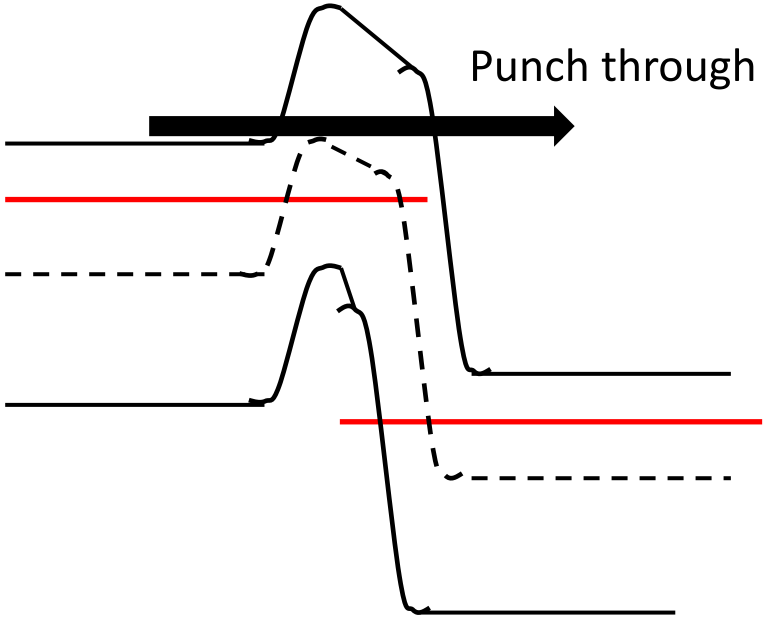


At surface

The bulk does not participate in the MOSFET conduction!

However, in the presence of drain bias, close to the drain diffusion





The barrier to source injection into the depletion width of the semiconductor is reduced due to drain bias

The punch though happens very close to the surface typically.

However, the substrate dopant concentration is sometimes observed to drop close to the edge of S/D diffusion

This causes the depletion width to the maximum around that region causing the punch though to happen in the bulk

Ways to reduce DIBL:

1. Source and drain diffusions should be scaled properly. Made shallow for shorter channels.
2. Channel doping must be made high to reduce the depletion width. However, this has other issues like changing the threshold voltage.
3. Sometimes, implants are made in localized regions near source/drain. These are called as Halo or pocket implants.

Generally,

$$I_D \ \alpha \frac{1}{L_{eff}} \ \alpha \frac{1}{L \ - \Delta L} = \frac{1}{L}\left(1 + \frac{\Delta L}{L}\right)$$

Since, the dependence is on VD, we write $\frac{\Delta L}{L} = \lambda V_D$

And the total current in saturation to be

$$I_D = \frac{W}{2L}\mu_n C_{ox}(V_G \ - V_T)^2(1 + \lambda V_D)$$

Multiplication and Oxide Reliability

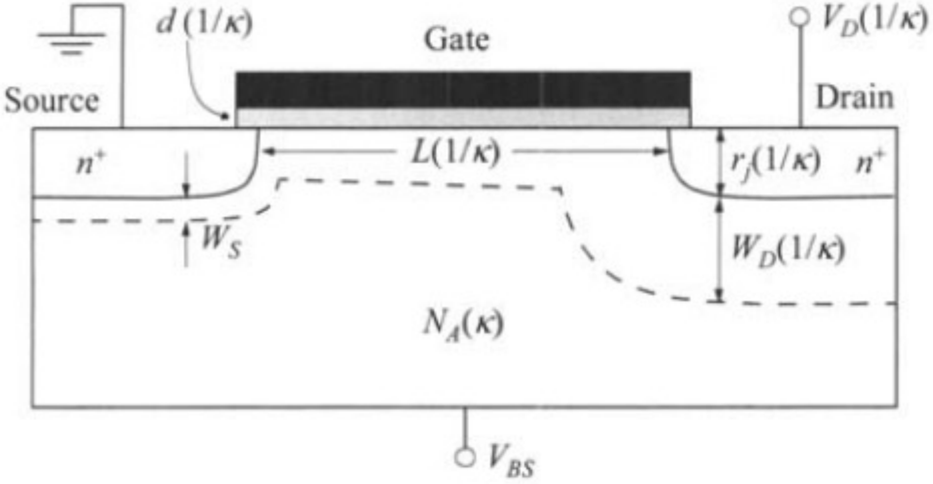
Due to nonideal device scaling, the electrons reach high field regions with excess energies (Hot carriers)

If an energy > 3.1 eV is obtained, then the electrons can cross the Si-SiO2 barrier

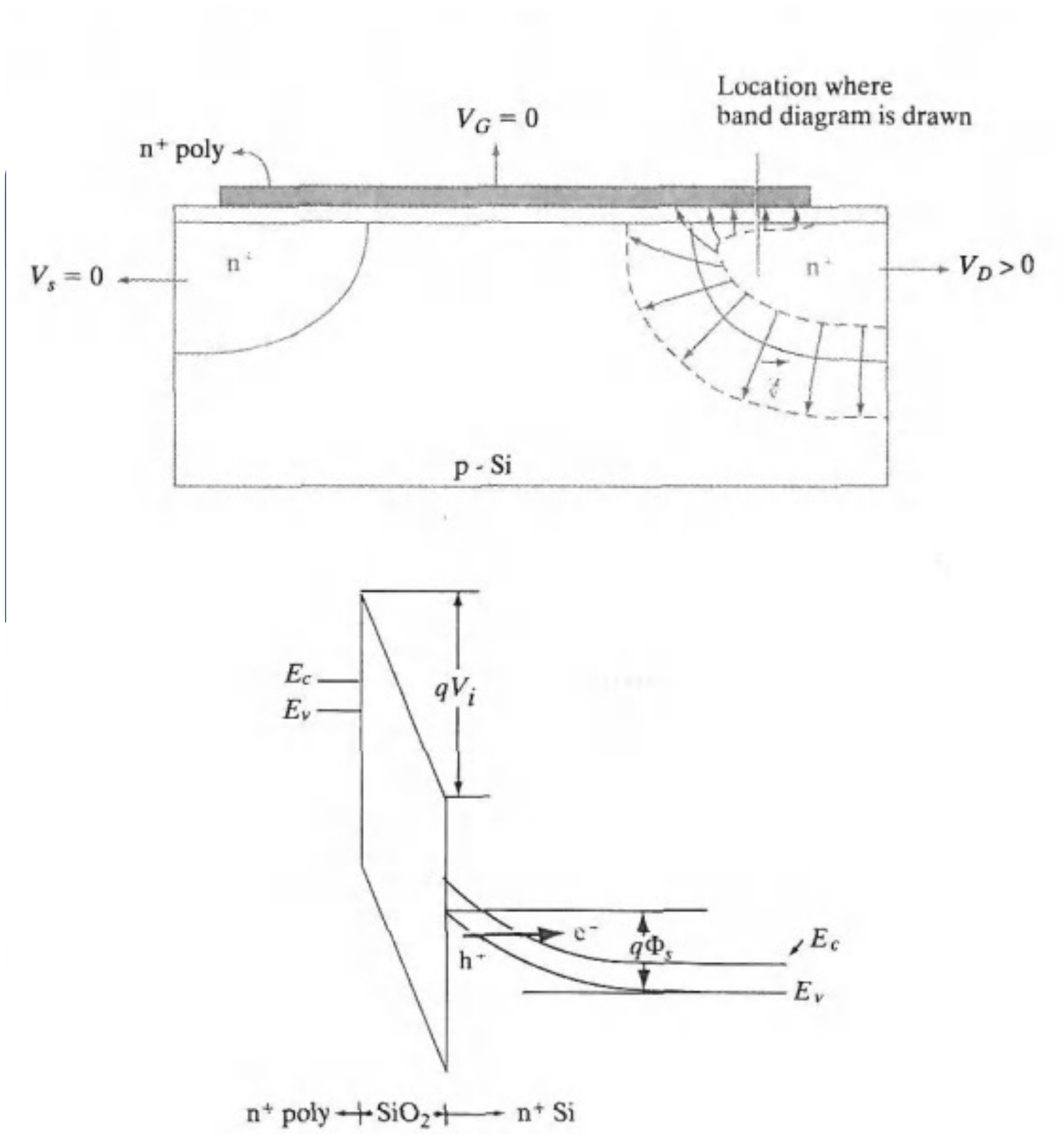
Extreme kinetic energy also leads to impact ionization – formation of e-h pairs

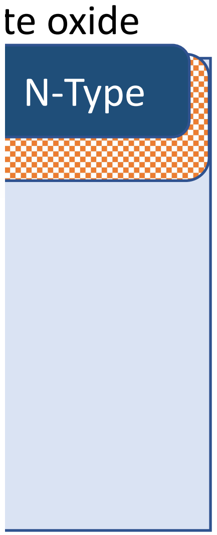
Electrons reach drain due to depletion field – contribute to current

Holes either reach the gate oxide/get sunk into the bulk giving rise to I_{BS}



Gate Induced Drain Leakage





Ideal, However, never like this.

There is always a gate – drain overlap

This causes increasing drain current as the gate potential is reduced.

This is because, the Gate-Drain field is now very large.

This causes, the bands to bend significantly near the drain
This causes extra current through Gate-Drain insulator interface.