2.4. Nonideal MOS Characteristics

Goal;

- To understand the nonideal MOS characteristics as the transistor size is reduced (or scaled). Also scaling theories as a device design guideline will be introduced.
- Issues related with nonstatic transport

Ref. 1. R. Dennard et.al "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions", JSSC, pp.256-267, 1976.
2. Y.Taur and E.Nowak,"CMOS Devices below 0.1㎛ : How will Performance go?," IEDM, 1997, pp.215-218

A. Constant Field Scaling

- As L is reduced to increase the density and speed of a chip, the MOSFET characteristics are changed from the ideal long channel characteristics(as introduced in §2.1~2.3). However, "electrostatics" and "electrodynamics" will change the "ideal" MOS characteristics. Dennard (ref. 1) proposed "device scaling law" in order to guarantee the ideal channel characteristics.

- Dennard's CE(Constant) scaling theory is on the "depletion" region width scaling. Consider the fig. 1 of ref. 1. As L is reduced to L'(=L/s; s is scaling factor), the depletion width of the drain side will touch the source depletion width and change the surface potential in the source side(so the V_T; called the DIBL effect to be explained later).

  If W_d is scaled in proportion to L, the device will look like the long channel counterpart.

$$W_d = \sqrt{\frac{2\varepsilon_s}{N_a}(2\Phi_f + V_{DS})}$$

In order for W_d' is scaled as L',

$$N_a' = sN_a$$

and

$$2\phi_f' + V_{DS}' = \frac{2\phi_f + V_{DS}}{s}.$$

(If $V_{DS} \gg 2\phi_f$, then $V_{DS}' = \frac{V_{DS}}{s}$).

The net result is the channel doping is increased by s factor whereas V_{DS} is reduced by s factor. Since L is reduced by same factor,

$$E' = \frac{V_{DS}'}{L'} = \frac{V_{DS}}{L},$$

so that the lateral field is same. So the theory is called "the constant Field scaling theory."

– V_T

Now consider V_T as,

$$V_T = V_{FB} + \gamma\sqrt{2\phi_f - V_B} + 2\phi_f$$

Usually, $V_{FB} \cong -2\phi_f$ (for N^+ poly silicon gate), and in order for V_T' scaled by the same factor s, $\gamma' = \frac{\sqrt{2\varepsilon_s q N_a'}}{C_{ox}'} = \frac{\gamma}{s}$

$$T_{ox}' = \frac{T_{ox}}{s}$$

and

$$(2\phi_f - V_B)' = \frac{2\phi_f - V_B}{s}.$$

Now the following circuit characteristics are also scaled according to the change in L and W (refer to table).

- $I_{DS}' = I_{DS}/s$
- Power/circuit : $P' = P/s^2$
- Power Density: $p' = p$.

– Factors which are not scaled:

$2\Phi_{f,p}$: In the above, $2\Phi_{f,p}$ which determines the depletion width (W_d) and the V_T is not scaled but rather slightly increases as N_a increases by s factor.

Another factor which is not scaled is the "subthreshold slope". Consider

$$\eta' = 1 + \frac{C_d' + C_{ss}'}{C_{ox}'}$$

Since $C_d' = C_d \cdot s$ and $C_{ox}' = C_{ox} \cdot s$ and if C_{ss} is neglected, $\eta' = \eta$, which means that η is not scaled. The net result of this is that the off current leakage increases in an exponential manner since V_T is scaled whereas η is not scaled;

$$\begin{aligned} I_{DS}(V_G = 0)' &= \frac{W'}{L'} C_d' V_t' D_n' e^{-\frac{V_T'}{\eta V_t}} \\ &= \frac{W}{L} s C_d V_t D_n \; e^{-\frac{V_T}{s\eta V_t}} \\ &= I_{DS}(V_{G=0}) s \; e^{\frac{V_T}{\eta V_t}(1-\frac{1}{s})} \end{aligned}$$

which means that the off leakage current increases by the factor of $e^{\frac{V_T}{\eta V_t}(1-\frac{1}{s})}$. This is very a important result since the off-current leakage is the limiting factor to increase the chip density.

B. Nonideal Scaling Due to V_{DD} Nonscaling

The CE theory introduced in the previous section is ideal. However the historical power supply voltage(V_{DD}) has not been scaled according to the CE theory due to the system requirement. Rather it has not been scaled down to L of 1μm. Also, in the submicron range, fig. 1 of Ref. 2 shows that V_{DD} is scaled in smaller rate than that of L as the slope in V_{DD} vs. L shows.

The net result was an increase in the Electric fields(both in the vertical and horizontal directions) which makes the device nonideal since the increase in the electric field changes the ideal device characteristics in many ways, which are the topics of this section.

B-1. Electrodynamics 1 : Current Saturation due to velocity saturation

The mobility decreases as the lateral field increases since carriers are heated up by the lateral field. As the carriers are heated up by the field, the momentum relaxation time in eq.(B-4) of §1.3.2 decreases so does the mobility (Hot carriers are more frequently collided with the lattice vibrations !). The net result is the saturation of velocity with the lateral field (See fig. 1).

Fig. 1. fig. 7-19 of YJP.

— Current saturation

Now consider the I_{DS} vs. V_{DS} for a long channel MOSFET,

$$I_{DS} = C_{ox}(V_{GS} - V_T - \frac{1}{2}V_{DS})\mu\frac{V_{DS}}{L}$$
$$= C_{ox}(V_{GS} - V_T - \frac{1}{2}V_{DS})v_d,$$

where $E_y|_{av} = \frac{V_{DS}}{L}$ is assumed.

As V_{DS} increases, I_{DS} increases with the slope of

$$C_{ox}(V_{GS} - V_T - \frac{1}{2}V_{DS})\frac{\mu}{L},$$

which tends to decrease as V_{DS} increases by two reasons;

$C_{ox}(V_{GS} - V_T - \frac{1}{2}V_{DS})$ decreases, which basically is due to $Q_n(y)$ decrease in the channel. The slope decreases until I_{DS} saturates. The saturation begins when

$$Q_n(L) = -C_{ox}(V_{GS} - V_T(L)) = 0$$

where $V_T(L) = V_{FB} + \gamma\sqrt{2\Psi_s - V_B} + 2\Phi_f + V_{DS}$.

Or, approximately,

$$Q_n(L) = -C_{ox}(V_{GS} - V_T - V_{DS}) = 0, \text{ so that}$$
$$V_{D,sat} \simeq V_G - V_T.$$

However, there is another reason for the slope decrease, which is due to the mobility decrease as V_{DS}(so the lateral field) increases. I_{DS} will eventually saturates as the drift velocity saturates.

Which comes first? It depends on L. If L is big, the pinch off comes first. If L is small, velocity saturation comes first. See fig. 2 for I_{DS}-V_{DS}'s of L=10μm and 0.5μm.

Fig. 2. fig. 7-20 of YJP.

– A phenomenological model for I_{DS} including both physics

ref. 1. C.G. Sodini et.al, "The effects of high fields on MOS devices and circuit performance," IEEE TED, vol. ED-31, pp.1386-1393. Oct. 1984.

ref. 2. Y.Cheng et.al, "A physically and scalable I-V model in BSIM3 for Analog/Digital Circuit Simulation," IEEE TED, vol.44, pp. 277-287, Feb., 1997

$$I_{DS,\,sat} = W v_{sat} C_{ox} \frac{(V_{GS} - V_T)^2}{E_{sat} L + (V_{GS} - V_T)}$$

$$I_{DS,\,lin} = \frac{W}{L} \frac{\mu_{eff}}{1 + \frac{V_{DS}}{E_{sat} L}} C_{ox} \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) V_{DS}$$

Even though the above equations are simple and crude equations, they do have essential features for the current saturation. For the evolution of BSIM type model for the short channel MOSFET, please refer to ref. 2.

B-2. Issues related with horizontal Field: Electrostatics

(Threshold voltage lowering as V_{DD} increases and L decreases)

Ref. 1. L.D. Yau, "A simple theory to predict the threshold voltage of short channel IGFET's," Solid State Electronics, vol. 17, pp 1059-1063, 1974.

2. R.R. Troutman, " VLSI limitation from Drain Induced Barrier Lowering," TED, pp.461-469, April, 1979.
3. M. Miura-Mattaush, "Analytical MOSFET model for quarter micron technologies," IEEE Tr. CAD of IC and Systems, pp.610-615, May 1994.
4. S. Takaki et. al, "On the universality of inversion layer mobility in N- and P- channel MOSFET's," IEDM, pp.398-401, 1988.

– As the lateral field increases, the lateral field originates from the drain voltage may affect the energy barrier as shown in Fig. 1. The net result of this phenomenon is the decrease of V_T. And more importantly, it may change the subthreshold slope as the $\Delta\psi_s$ changes with V_{DD}.

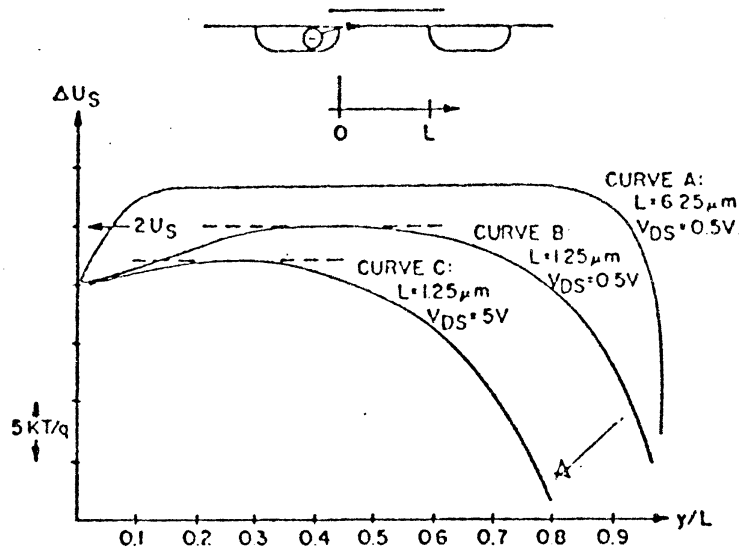

Fig. 1. Surface potential distribution for constant gate voltage (V_G = 1.8). Only channel length and drain voltage are varied.

– Charge sharing

One way of modeling V_T decrease as L reduces and V_{DD} decreases is "the charge sharing". The main idea has been proposed by Yau(ref. 1). See fig.2 for the geometric description . Total charges in the device can be written as

$$Q_G + Q_S + Q_D + Q_B = 0.$$

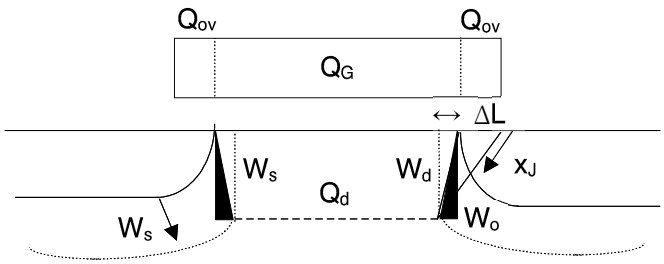

Fig. 2. Charge sharing relation in short channel MOSFET.

V_T is the voltage to maintain Q_G at the gate.

$$Q_G = Q_d + Q(\text{overlap}).$$

Q_d is the depletion charge to invert the surface. Total depletion charge is $Q_d WL$. However, as L decreases, some of the charges are from the source and drain field so that total charge in the gate to maintain the depletion region is $Q_d W(L - 2\Delta L)$.

ΔL is function of r_j, x_j, and V_{DD} as

$$(X_j + \Delta L)^2 = (X_j + W_o)^2 - W_d^2$$

$$\Delta L = X_j\left(\sqrt{1 + \frac{2W_d}{X_j}} - 1\right)$$

(if $W_d = W_o$)

So that

$$\frac{L-2\Delta L}{L} = 1 - \frac{2X_j}{L}\left(\sqrt{1+\frac{2W_d}{X_j}} - 1\right) \equiv \alpha$$

and

$$V_T = V_{FB} + \gamma_s\sqrt{2\Phi_{fp}} + 2\Phi_{fp}$$

where

$$\gamma_s = \gamma(1-2\alpha).$$

If W_s is different from W_d, then a_s and a_d should be used instead of 2α.

The charge sharing model implies that the surface potential are affected all along the channel region. The theory alone cannot explain many of the V_T (and subthreshold characteristics) dependence on V_{DD} and L.

– DIBL effect

Three curves are shown in fig. 1. Curve A is the surface potential for a long channel device. B is for short channel device with small V_{DS}. $\Delta\Psi_s$ can be found. This is due to the charge sharing. As V_{DS} is increased from B, the potential barrier between Source and Channel is reduced, inducing an anomalour current increase in the subthreshold region.

The electric field penetrates to the surface by way of "Bulk" or "Surface" depending upon the channel doping profile. If I_{DS} in the subthreshold can be shut off by reducing V_{GS}, the penetration is most likely through "surface" (see fig. 3).

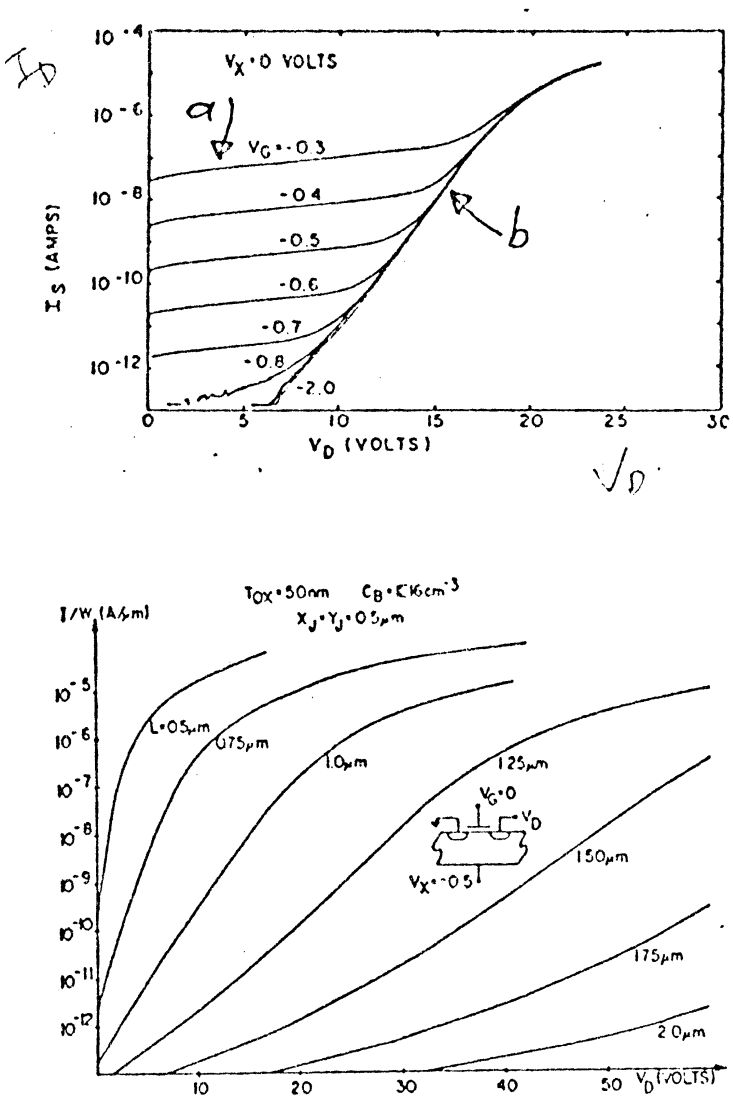

Fig. 3. I_s vs. V_{DS} for various V_{GS}'s(ref. 2)

– dE_y/dy at the source end of the channel (see fig. 4)

Another way of looking at this problem is reconsideration of the MOS equation. In the MOS equation, i.e.,

$$V_{GS} - V_{FB} = -\frac{Q_n + Q_d}{C_{ox}} + \Psi_s$$

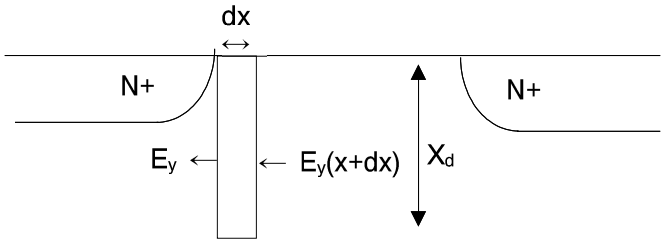

Fig. 4.

$\frac{Q_n + Q_d}{C_{ox}}$ represents the voltage drop in the oxide, and that is from the surface field($\frac{Q_s}{\varepsilon_s}$). If there exists dE_y/dy due to the drain field, the E_s should be modified as,

$$E_y(x+dx) - E_y(x)\,x_d + E_s dx = -\frac{(Q_d + Q_n)dx}{\varepsilon_s}$$

Or

$$E_s = -\frac{Q_s}{\varepsilon_x} - \frac{dE_y}{dy}\,x_d$$

where dE_y/dy is assumed uniform along the depletion width, x_d. Net result is decrease of E_s for same Q_s, meaning that less V_{GS} is required to induced the same Q_s.

In ref. 3, assuming the quadratic potential distribution in the channel region, dE_y/dy is modeled as

$$\frac{dE_y}{dy} = \frac{2(V_{bi} - \Psi_s(0))}{L_{eff}^2}$$

where V_{bi}, $\Psi_s(0)$ are the built-in potential and surface potential in the source side, respectively.

– dE_y/dy at the drain side of the channel

Read pp. 154-158 of Taur and Ning to find out the dE_y/dy in the velocity saturation region in the drain section of the channel.

B-3 Issues related with Vertical Field

– As the vertical field increases as the T_{ox} scales while V_{DD} is not scaled the vertical field increases. The following equations are the Electric field in each section of the device.

$$E_s = -\frac{Q_n + Q_d}{\varepsilon_s} = C_{ox}\frac{(V_{GS} - V_T) + \gamma\sqrt{2\phi_f - V_{sub}}}{\varepsilon_s}$$

$$E_{ox} = E_s\left(\frac{\varepsilon_s}{\varepsilon_{ox}}\right) = \frac{(V_{GS} - V_T) + \gamma\sqrt{2\phi_f - V_{sub}}}{T_{ox}}$$

$$E_s' = sC_{ox}\frac{V_{GS} - V_T}{s} + \frac{\gamma}{\sqrt{s}}\sqrt{\frac{2\phi_f - V_{sub}}{s}}/\varepsilon_s = E_s$$

$E_{ox}' = E_{ox}$, which means the vertical field is also constant in the CE scaling.

Ex) You may have different vertical field vs. VGS relation if Quantum mechanics is considered for inversion layer. Yet, the relationship between Qs vs. vertical field may be conserved.

– Mobility reduction and the "Universality"
It the mobility is plotted vs. the Effective mobility defined as,

$$E_{eff} = \frac{0.5Q_n + Q_d}{\varepsilon_s} \quad \text{for electrons and}$$

$$E_{eff} = \frac{\frac{1}{3}Q_p + Q_d}{\varepsilon_s} \quad \text{for holes,}$$

mobilities are on a universal curve. Three definite regions can be noticed in the fig. 5. First region is where the mobility increases with E_{eff}. In the region, the Coulomb scattering with the channel impurity atoms dominates the mobility. As E_{eff} increases, the inversion carriers increase and screen the impurities and reduce the scattering rates. In the 2nd region, the mobility reduces with $E_{eff}^{-1/3}$. Theoretically, the dependence comes from the lattice scattering rates. In the 3rd region, the mobility reduced with E_{eff}^{-2}, which is known due to the surface roughness scattering. As the E_{eff} increases as L is scaled to 0.1μm range, the device operates in the "surface roughness" dominant region.

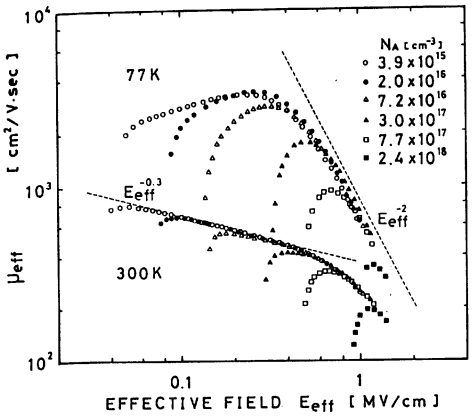

Fig.1 Effective electron mobility μ_{eff} at 300 K and 77 K in n-channel MOSFETs versus effective normal field E_{eff} as a parameter of substrate acceptor concentration. Here, E_{eff} is defined by $E_{eff} = q(Ndpl+Ns/2)/\varepsilon_{Si}$.

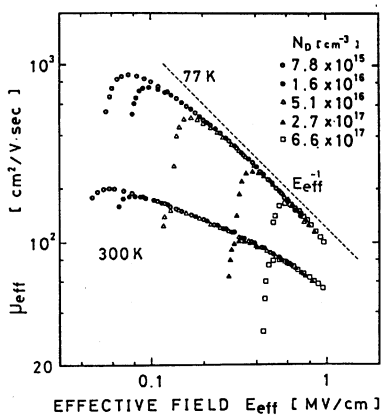

Fig.2 Effective hole mobility μ_{eff} at 300 K and 77 K in p-channel MOSFETs versus effective normal field E_{eff} as a parameter of substrate donor concentration. Here, E_{eff} is defined by $E_{eff} = q(Ndpl+Ns/3)/\varepsilon_{Si}$.

Fig. 4

– Effective C_{ox} decrease due to quantum and gate poly depletion effects. As we studied in the section 2-1, the effective Cox is decreased due to the vertical quantization of the surface and the poly depletion effects. In order to reduce the poly depletion effect, people tries to use the metal gate(such as W, Co).
The comprehensive coverage of the modern MOSFET technology may be found in good review papers found in IEDM.

C. Ballistic Transport

ref> 1. Tang et. al,
2. Lundstrom

As the channel length of MOSFET devices becomes extremely small (comparable to the average length between scattering, or the mean free length), or the spatial(or temporal) change in the electric field becomes abrupt, then the 'local mobility' concept breaks down. The carrier transport in this regime is called the 'nonstatic transport'.

The DD model is in one extreme and the 'ballistic' transport is in the other extreme. Most of the devices with the channel length is the few 10s of nanometer are in between two extremes. There are several approaches to handle the carrier transport in this regime;

- Hydrodynamic modeling
- Monte Carlo approach
- The transmission and reflection approach

-Hints from Takagi

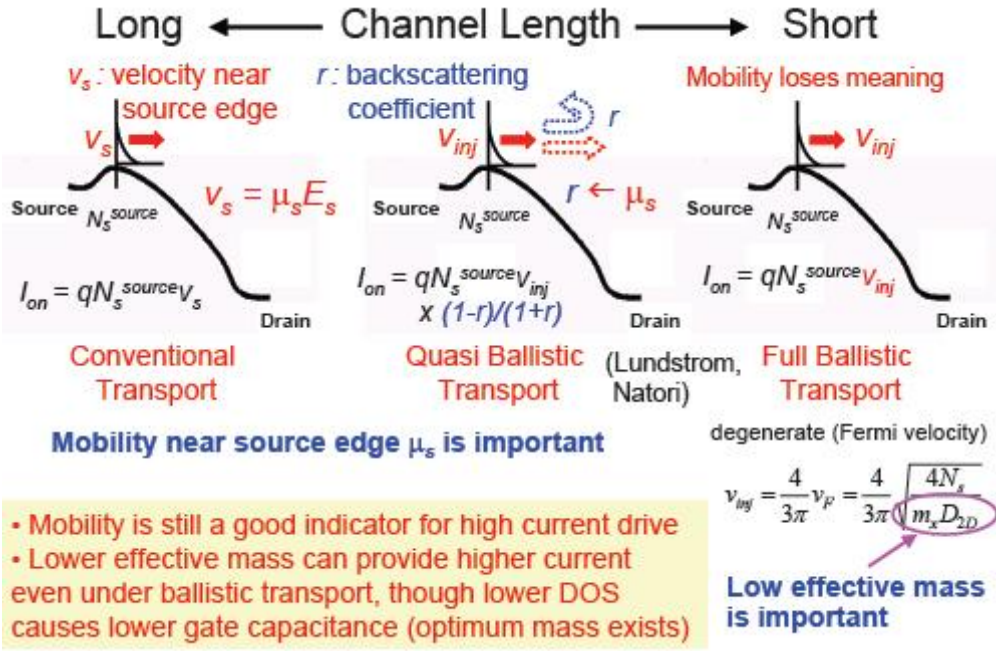

Fig. Three regimes in the carrier transport. Left: Long channel limt with DD model. Middle : quasi Ballistic transport regime, Right: Full transport regime. It should be noted that the mobility value is the determing factor in MOSFET current(From Takagi's lecture material)