Appendix 1.

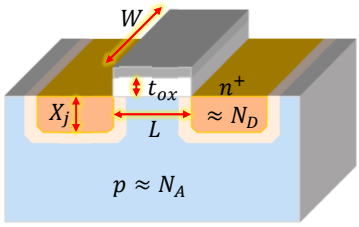
MOSFET Scaling Rule and Issues

Introduction

- Scaling of MOSFETs ➔ mostly down-scaling
- Down-scaling ➔ generally channel length (L) shrinks (shorter channel)
- But only reduction of L gives punch-through where the depletion regions at source and drain are merged ➔ abnormal MOSFET operation, and many other issues (e.g. SCE like the DIBL)
- So, we need a rule for down-scaling MOSFET while maintaining the MOSFET's normal operation.
- **What and How?**

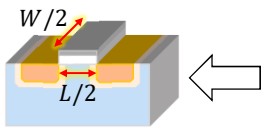
Part 1.
Scaling Rule
with constant power-density strategy
and constant Transconductance
(= keeping the MOSFET operation the same)

Scaling Rule with the ratio of S (>1) (example: n-MOSFET)



Geometry	Doping density	Signal level
$L \rightarrow L/S$	$N_D \rightarrow N_D \times S$	$V_{DD} \rightarrow V_{DD}/S$
$W \rightarrow W/S$	$N_A \rightarrow N_A \times S$	$V_{th} \rightarrow V_{th}/S$
$X_j \rightarrow X_j/S$		
$t_{ox} \rightarrow t_{ox}/S$	$\Rightarrow C_{ox} \rightarrow C_{ox} \times S$	

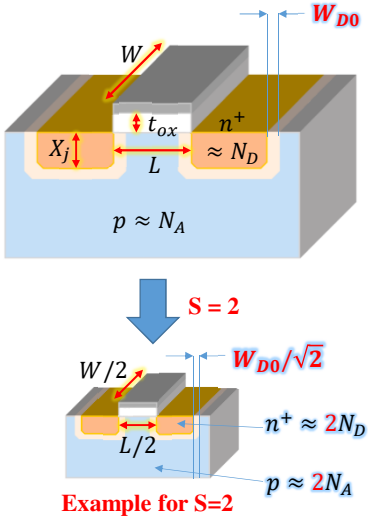
S = 2



Example for S=2

Geometry	Doping density	Signal level
$L \rightarrow L/2$	$N_D \rightarrow N_D \times 2$	$V_{DD} \rightarrow V_{DD}/2$
$W \rightarrow W/2$	$N_A \rightarrow N_A \times 2$	$V_{th} \rightarrow V_{th}/2$
$X_j \rightarrow X_j/2$		
$t_{ox} \rightarrow t_{ox}/2$	$\Rightarrow C_{ox} \rightarrow C_{ox} \times 2$	

Why scaling-up of doping density (1)



Example for S=2

$$pn^+ \approx N_A^- N_D^+ = n_i^2 exp\left(\frac{qV_{bi}}{kT}\right)$$

$$W_{D0} = \sqrt{\frac{2\varepsilon}{q}\left(\frac{1}{N_A^-} + \frac{1}{N_D^+}\right)V_{bi}} \approx \sqrt{\frac{2\varepsilon}{q}\left(\frac{1}{N_A^-}\right)V_{bi}}$$

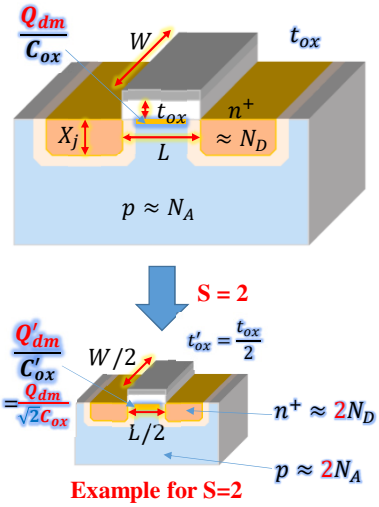
$$W'_{D0} = \sqrt{\frac{2\varepsilon}{q}\left(\frac{1}{sN_A^-}\right)V'_{bi}} = \frac{W_{D0}}{\sqrt{s}}$$

Assuming $V_{bi} \approx V'_{bi}$

← To reduce the depletion width of pn+ junctions.

← Why not S² to get $\frac{W_{D0}}{S}$ → Zener breakdown.

Why scaling-up of doping density (2)



Example for S=2

$$V_{th} = \phi_{ms} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{dm}}{C_{ox}} + 2\varphi_F$$

$Q_{dm} \approx -qpW_{dm}$

$p = n_i \exp\left(\frac{q\varphi_F}{kT}\right)$

$\Longrightarrow \varphi_F = \frac{kT}{q} ln\left(\frac{p}{n_i}\right)$

← saturated

$$W_{dm} = \sqrt{\frac{2\varepsilon_s}{q}\frac{1}{p}2\varphi_F}$$

$$W'_{dm} = \sqrt{\frac{2\varepsilon_s}{q}\frac{1}{sp}2\varphi_F} = \frac{W_{dm}}{\sqrt{s}}$$

$$Q'_{dm} \approx -qspW'_{dm} \approx -qsp\frac{W_{dm}}{\sqrt{s}} = \sqrt{s}Q_{dm}$$

$$\frac{Q'_{dm}}{C'_{ox}} \approx \frac{Q_{dm}}{\sqrt{s}C_{ox}} \quad \rightarrow C'_{ox} = sC_{ox}$$

← To reduce the depletion voltage ($-\frac{Q_{dm}}{C_{ox}}$)

← Why not S² to get $\frac{W_{D0}}{S}$ → Zener breakdown.

→ A reason why V_{th} is tardy to be scaled down.

Scaling rule by the factor of "s" for a constant DC-power density

Current

$$I_D = \frac{1}{2}\mu_n \frac{W}{L} C_{ox}(V_G - V_{th})^2 \Rightarrow \mu_n \frac{W}{L}\frac{\varepsilon_{ox}}{t_{ox}} V_{DD}^2 \Rightarrow I_D' = \frac{1}{2}\mu_n \frac{\frac{W}{s}}{\frac{L}{s}} \frac{\varepsilon_{ox}}{\frac{t_{ox}}{s}} \left(\frac{V_{DD}}{s}\right)^2 = \frac{I_D}{s}$$

$s > 1$

Voltage

$$V_{DD} \Rightarrow V_{DD}' = \frac{V_{DD}}{s}$$

$s = s_i = s_v$

DC-Power consumption

$$P_{dc} = V_{DD} I_D \Rightarrow P_{dc}' = \frac{V_{DD} I_D}{s^2} = \frac{P_{dc}}{s^2}$$

DC-Power density:

$$PD = \frac{P_{dc}}{A} = \frac{V_{DD} I_D}{LW} \Rightarrow PD' = \frac{P_{dc}'}{A'} = \frac{P_{dc}/s^2}{LW/s^2} = PD$$

← **Constant !**

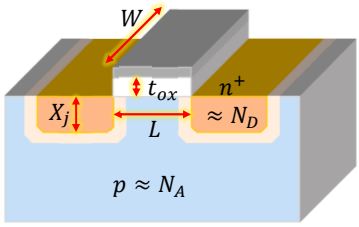
Transconductance:

$$g_m = \mu_n \frac{W}{L}\frac{\varepsilon_{ox}}{t_{ox}} V_{DD} \Rightarrow g_m' = \mu_n \frac{W/s}{L/s}\frac{\varepsilon_{ox}}{t_{ox}/s} V_{DD}/s = g_m$$

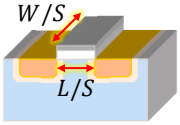
← **Constant !**

Department of Electronics Engineering

Summary of Scaling Rule and Its Electrical Aspects



Scaling



Scaling Rule:

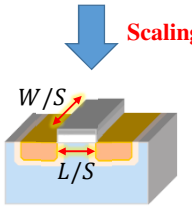
Geometry	Doping density	Signal level
$L \rightarrow L/S$	$N_D \rightarrow N_D \times S$	$V_{DD} \rightarrow V_{DD}/S$
$W \rightarrow W/S$	$N_A \rightarrow N_A \times S$	$V_{th} \rightarrow V_{th}/S$
$X_j \rightarrow X_j/S$		
$t_{ox} \rightarrow t_{ox}/S \Rightarrow C_{ox} \rightarrow C_{ox} \times S$		

Electrical aspects:

- Current (I) ➔ 1/S
- Voltage (V) ➔ 1/S
- Circuit delay (= g_m C LW) ➔ 1/S
  - g_m ➔ kept
  - C ➔ S
  - LW ➔ $1/S_2$
- DC-Power consumption (P_{DC}) ➔ $1/S^2$
- **DC-Power density (PD) and g_m ➔ kept**

Department of Electronics Engineering

Basic purpose of Scaling → Increase of Frequency capability of MOSFETs



Electrical aspects:

- Current (I) → 1/S
- Voltage (V) → 1/S
- Circuit delay (= g_m C LW) → 1/S
  - g_m → kept
  - C → S
  - LW → $1/S_2$
- DC-Power consumption (P_{DC}) → $1/S^2$
- DC-Power density (PD) and g_m → kept

$$f_T = \frac{g_m}{2\pi(C_{gs}+C_{gd})} = \frac{g_m}{2\pi(c_{gs}+c_{gd})LW}$$

Operating frequency (BW) is increased by s-times

$$f_T' = \frac{g_m'}{2\pi(sc_{gs}+sc_{gd})LW/s^2} = \frac{sg_m}{2\pi(c_{gs}+c_{gd})LW}$$

However, there are several issues on this rule of scaling

- **Threshold voltage (V_{th})** is tardy to be scaled-down, so, **the supply voltage (V_{DD})** is not scaled-down (tardy). So the DC-power consumption is not scaled-down while the area is scaled-down, thus higher power density → hot plate
  And this is added with the AC-power consumption (dynamic power consumption Which is proportional to the operating frequency), leading to further higher power density → very hot plate → very high frequency operation is to be avoided.
  → This is one of difficulties for a RF circuit design for the 5-G application (28 GHz).
- **Sub-threshold slope (SS)** based on diffusion current mechanism is not scalable.
- **Off-current ($I_{OFF} = I_{DS}(V_G=0)$)** is increased, so higher stand-by power consumption.
- **Gate oxide thickness (t_{ox})** is not easy to be very thin due to a dielectric breakdown.
- **And many other issues,** such as a limitation of lithography for a geometrical scaling while getting a failure of meeting the Moore's Law.

Example: Tardy Scaling-Down of Voltage Level

Current

$$I_D = \frac{1}{2}\mu_n \frac{W}{L} C_{ox}(V_G - V_{th})^2 \Rightarrow \mu_n \frac{W}{L}\frac{\varepsilon_{ox}}{t_{ox}} V_{DD}^2 \Rightarrow I_D' = \frac{1}{2}\mu_n \frac{\frac{W}{S}}{\frac{L}{S}} \frac{\varepsilon_{ox}}{\frac{t_{ox}}{S}} \left(\frac{V_{DD}}{S_v}\right)^2 = \frac{S}{S_v^2} I_D$$

Voltage

$$V_{DD} \Rightarrow V_{DD}' = \frac{V_{DD}}{S_v}$$

Issue: $S > S_v > 1$

DC-Power consumption

$$P_{dc} = V_{DD} I_D \Rightarrow P_{dc}' = \frac{S V_{DD} I_D}{S_v^3} = \frac{S}{S_v^3} P_{dc} \gg P_{dc}$$

DC-Power density:

$$PD = \frac{P_{dc}}{A} = \frac{V_{DD} I_D}{LW} \Rightarrow PD' = \frac{P_{dc}'}{A'} = \frac{\frac{S}{S_v^3} P_{dc}}{LW/S^2} = \frac{S^3}{S_v^3} PD \gg PD$$

Transconductance:

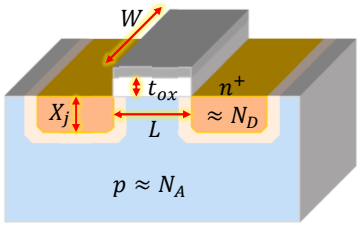
$$g_m = \mu_n \frac{W}{L}\frac{\varepsilon_{ox}}{t_{ox}} V_{DD} \Rightarrow g_m' = \mu_n \frac{\frac{W}{S}}{\frac{L}{S}} \frac{\varepsilon_{ox}}{\frac{t_{ox}}{S}} V_{DD}/S_v = \frac{S}{S_v} g_m \gg g_m$$

← Increased !

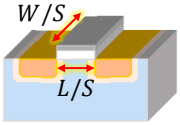
← Increased !

PD is not kept but increased → Hot! → Burned Out !

Example: Summary of Tardy Scaling-Down of Voltage Level and its effect on Electrical Aspects



Scaling



$S > S_v > 1$

Scaling Rule:

Geometry	Doping density	Signal level
$L \rightarrow L/S$	$N_D \rightarrow N_D \times S$	$V_{DD} \rightarrow V_{DD}/S_v$
$W \rightarrow W/S$	$N_A \rightarrow N_A \times S$	$V_{th} \rightarrow V_{th}/S_v$
$X_j \rightarrow X_j/S$		
$t_{ox} \rightarrow t_{ox}/S \Rightarrow C_{ox} \rightarrow C_{ox} \times S$		

Electrical aspects:

- Current (I) → S/S_v^2
- Circuit delay (= g_m C LW) → $1/S_v$
  - $g_m \rightarrow \frac{S}{S_v}$
  - C → S
  - LW → $1/S_2$
- Voltage (V) → $1/S_v$
- DC-Power consumption (P_{DC}) → $\frac{S}{S_v^3}$
- DC-Power density (PD) → $\frac{S^3}{S_v^3} > 1$ and g_m → $\frac{S}{S_v} > 1$

Part 2.
Issues and Possible Solutions

Issue 1. Difficulty of Down-Scaling of V_{th} and Special MOSFETs to overcome: SOI-MOSFET

Typical MOSFET (bulk)

WF design is limited by nature to be negative value for cancelling $-\frac{Q_{dm}}{C_{ox}}$

$-\frac{Q_{ox}}{C_{ox}}$ **is constant**

$-\frac{Q_{dm}}{C_{ox}}$ **is tardy to be scaled down**

$$W_{dm} = \sqrt{\frac{2\varepsilon_s}{q}\frac{1}{p}2\varphi_F}$$

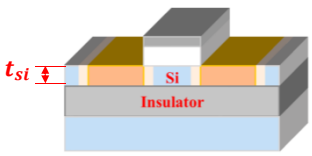
$$\varphi_F = \frac{kT}{q} ln\left(\frac{p}{n_i}\right)$$ **is saturated (→ constant)**

$$V_{th} = \phi_{ms} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{dm}}{C_{ox}} + 2\varphi_F$$

SOI-MOSFET (Fully depleted=FD)

$Q_{dm}^{si} < Q_{dm}$

$t_{si} < W_{dm} \Rightarrow\Rightarrow -qpt_{si} < -qpW_{dm}$



Depletion width and charge is limited by the thickness of Silicon on Insulator (SOI).

V_{th} *can be reduced slightly* ...

Issue 1. Difficulty of Down-Scaling of V_{th} and Special MOSFETs to overcome: Fin-FET

Typical MOSFET (bulk)

WF design is limited by nature to be negative value for cancelling $-\frac{Q_{dm}}{C_{ox}}$

$-\frac{Q_{ox}}{C_{ox}}$ **is constant**

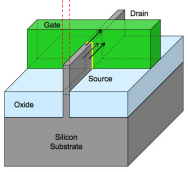
$-\frac{Q_{dm}}{C_{ox}}$ **is tardy to be scaled down**

$W_{dm} = \sqrt{\frac{2\varepsilon_s}{q}\frac{1}{p}2\varphi_F}$

$\varphi_F = \frac{kT}{q} ln\left(\frac{p}{n_i}\right)$ **is saturated (→ constant)**

$$V_{th} = \phi_{ms} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{dm}}{C_{ox}} + 2\varphi_F$$

Fin-FET



22nm Bulk Fin-FET

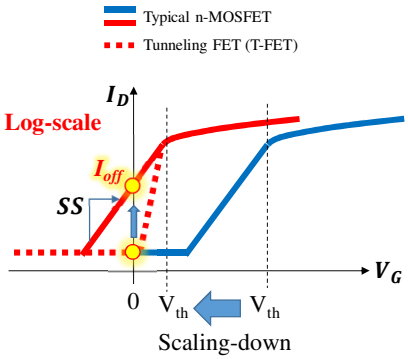
$|Q_{dm}^{si}| < |Q_{dm}|$

$t_{si} < W_{dm} \Rightarrow\Rightarrow |-qpt_{si}| < |-qpW_{dm}|$

Depletion width and charge is limited by the thickness of Silicon Fin (지느러미).

V_{th} *can be reduced slightly* ...

Issue 2. Difficulty of Down-Scaling of *SS and* I_{off} and Special MOSFETs to overcome: Tunneling FET (T-FET)

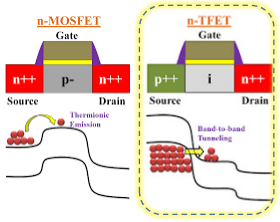


Although Vth is scaled down successfully, Off-current (I_{DS} at V_G=0) is increased due to SS which is not scaled-down below 60 mV/dec

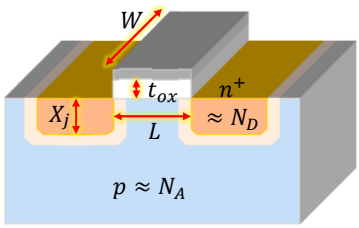
$$SS_{min} = 60\ mV/dec\ at\ T = 300K$$

This is limited by the diffusion mechanism in the Sub-threshold regime.

So, we need to use other mechanism rather than the diffusion.➔ It is **Tunneling**



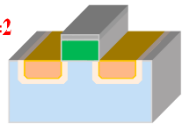
Issue 3. Difficulty of Down-Scaling of t_{ox} and Special MOSFETs to overcome: High-K insulator



Scaling by S=2

$t_{ox} \rightarrow t_{ox}$

$\Rightarrow C_{ox} \rightarrow C_{ox} \times 2$



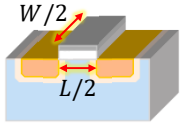
$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} = \frac{\varepsilon_r \varepsilon_0}{t_{ox}}$

$\Rightarrow \varepsilon_{ox} \rightarrow \varepsilon_{ox} \times 2$

Scaling by S=2

$t_{ox} \rightarrow t_{ox}/2$

$\Rightarrow C_{ox} \rightarrow C_{ox} \times 2$



Thinning the dielectric gives a breakdown
→ Limitation of scaling down
→ With the main purpose of scaling up Cox,
To find the high ε_r insulator is better.

High-K == High Dielectric Constant (ε_r)

For example, HfO_x, Al_2O_3, etc, which have a much higher value of ε_r than SiO_2 (ε_r =3.9).

Issue 4. AC Power Density at Higher Frequency

← Solution = Stop increasing frequency → but Circuit optimization needed

$s > s_v > 1$

DC-Power consumption

$$P_{dc} = V_{DD} I_D \Rightarrow P'_{dc} = \frac{s V_{DD} I_D}{s_v^3} = \frac{s}{s_v^3} P_{dc} \gg P_{dc}$$

← Increased !

DC-Power density:

$$PD_{dc} = \frac{P_{dc}}{A} = \frac{V_{DD} I_D}{LW} \Rightarrow PD'_{dc} = \frac{P'_{dc}}{A'} = \frac{\frac{s}{s_v^3} P_{dc}}{LW/s^2} = \frac{s^3}{s_v^3} PD_{dc} \gg PD_{dc}$$

← Increased !

AC-Power consumption (Dynamic Power Consumption)

$C'_{tot} = s C_{ox} \frac{LW}{s^2} = \frac{C_{tot}}{s}$

$s > s_v > 1$

$$P_{ac} = \frac{1}{2} C_{tot} V_{DD}^2 f_{CLK} \Rightarrow P'_{ac} = \frac{1}{2} \frac{C_{tot}}{s} \frac{V_{DD}^2}{s_v^2} s f_{CLK} = \frac{1}{s_v^2} P_{ac} < P_{ac}$$

$C_{tot} = C_{ox} LW$

← Decreased !

AC-Power density:

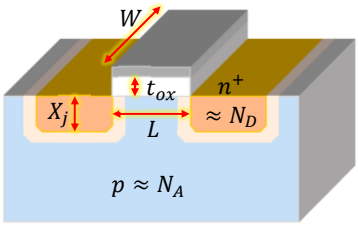
$$PD_{ac} = \frac{P_{ac}}{A} = \frac{V_{DD} I_D}{LW} \Rightarrow PD'_{ac} = \frac{P'_{ac}}{A'} = \frac{\frac{1}{s_v^2} P_{ac}}{LW/s^2} = \frac{s^2}{s_v^2} PD_{ac} \ggg PD_{ac}$$

← Increased !

← Very High Frequency Circuit would be burned out !

Issue 4. AC Power Density at Higher Frequency

← Solution = Revising of Scaling Rule → But Poor Performance



Scaling Rule:

Geometry	Doping density	Signal level
$L \rightarrow L/S$	$N_D \rightarrow N_D \times S$	$V_{DD} \rightarrow V_{DD}/S_v$
$W \rightarrow W/S$	$N_A \rightarrow N_A \times S$	$V_{th} \rightarrow V_{th}/S_v$
$X_j \rightarrow X_j/S$		$S > S_v > 1$
$t_{ox} \rightarrow t_{ox}$	$\Rightarrow C_{ox} \rightarrow C_{ox}$	

Non-scaling C_{ox}

Current ⇒ $I_D' \Rightarrow \frac{1}{2}\mu_n \frac{\frac{W}{S}}{\frac{L}{S}} \frac{\varepsilon_{ox}}{\frac{t_{ox}}{1}} \left(\frac{V_{DD}}{S_v}\right)^2 = \frac{1}{S_v^2} I_D$

Voltage ⇒ $V_{DD}' = \frac{V_{DD}}{S_v}$

Transconductance (not OK now):

$$g_m' = \mu_n \frac{\frac{W}{S}}{\frac{L}{S}} \frac{\varepsilon_{ox}}{\frac{t_{ox}}{1}} V_{DD}/S_v = \frac{1}{S_v} g_m \ll g_m$$

g_m decreased

➔ Poor performance of circuits

AC-Power density (now OK):

⇒ $PD_{ac}' = \frac{P_{ac}'}{A'} = \frac{\frac{1}{S_v^3} P_{ac}}{LW/S^2} = \frac{S^2}{S_v^3} PD_{ac} < PD_{ac}$