**Constant Field Scaling (Lg) gives constant power density**

Device scaling: Simplified design goals/guidelines for shrinking device dimensions to achieve density and performance gains, and power reduction in VLSI.

**Issues:** Short-channel effect, Power density, Switching delay, Reliability.

![Diagram](image)

**FIGURE 4.1.** Principles of MOSFET constant-electric-field scaling. (After Dennard, 1986).

The principle of constant-field scaling lies in scaling the device voltage and the device dimensions (both horizontal and vertical) by the same factor, $\kappa$ ($\frac{L}{\kappa}$), such that the electric field remains unchanged.
Figure 7.1 Cross section of (a) original NMOS transistor and (b) scaled NMOS transistor.

\[ x_D = \sqrt{\frac{2\varepsilon (V_{bi} + V_D)}{eN_a}} \]  \hspace{1cm} (7.1)

\[ \frac{I_D}{W} = \frac{\mu_n \varepsilon_{ox}}{2t_{ox} L} (V_G - V_T)^2 \rightarrow \frac{\mu_n \varepsilon_{ox}}{2(kt_{ox})(kL)} (kV_G - V_T)^2 \approx \text{constant} \]  \hspace{1cm} (7.2)
Classical CMOS Scaling is Dead

Scaled Device

Voltage, \( V / \alpha \)  
Oxide, \( t_{ox} / a \)  
Wire width, \( W / a \)  
Gate width, \( L / a \)  
Diffusion, \( x_d / a \)  
Substrate, \( a \times NA \)

RESULTS:
Voltage: \( V / a \)  
Oxide: \( t_{ox} / a \)  
Wire width: \( W / a \)  
Gate width: \( L / a \)  
Diffusion: \( x_d / a \)  
Substrate: \( a \times NA \)

Higher Density: \( \sim a^2 \)  
Higher Speed: \( \sim a \)  
Power (\( V_{th} \)): \( \sim 1 / a^2 \)  
Power Density: \( \sim \text{Constant} \)

Why deviate from "ideal" scaling?
- unacceptable gate leakage/reliability
- additional performance at higher voltages

What's the consequence of this deviation?
- a dramatic rise in power density
**Rules of Constant Field Scaling**

<table>
<thead>
<tr>
<th>Scaling assumptions</th>
<th>MOSFET Device and Circuit Parameters</th>
<th>Multiplicative Factor ($\kappa &gt; 1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Device dimensions ($t_{ox}, L, W, x_i$)</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td></td>
<td>Doping concentration ($N_a, N_d$)</td>
<td>$\kappa$</td>
</tr>
<tr>
<td></td>
<td>Voltage ($V$)</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td>Derived scaling behavior of device parameters</td>
<td>Electric field ($E$)</td>
<td>$1$</td>
</tr>
<tr>
<td></td>
<td>Carrier velocity ($v$)</td>
<td>$1$</td>
</tr>
<tr>
<td></td>
<td>Depletion layer width ($W_d$)</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td></td>
<td>Capacitance ($C = \varepsilon A/l$)</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td></td>
<td>Inversion layer charge density ($Q_i$)</td>
<td>$1$</td>
</tr>
<tr>
<td></td>
<td>Current, drift ($I$)</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td></td>
<td>Channel resistance ($R_{ch}$)</td>
<td>$1$</td>
</tr>
<tr>
<td>Derived scaling behavior of circuit parameters</td>
<td>Circuit delay time ($\tau \sim CV/I$)</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td></td>
<td>Power dissipation per circuit ($P \sim VI$)</td>
<td>$1/\kappa^2$</td>
</tr>
<tr>
<td></td>
<td>Power-delay product per circuit ($P \times \tau$)</td>
<td>$1/\kappa^3$</td>
</tr>
<tr>
<td></td>
<td>Circuit density ($\propto 1/A$)</td>
<td>$\kappa^2$</td>
</tr>
</tbody>
</table>

1 (in transistor)
Problem 4.1

4.1. Apply constant-field scaling rules to the long-channel currents, Eq. (3.19) for the linear region, and Eq. (3.23) for the saturation region, and show that they behave as indicated in Table 4.1.

Eq. (3.19) : \( I_{ds} = \mu_{eff} C_{ox} \frac{W}{L} \left( V_g - V_t \right) V_{ds} \)

Eq. (3.23) : \( I_{ds} = I_{dsat} = \mu_{eff} C'_{ox} \frac{W}{L} \left( \frac{V_g - V_t}{2m} \right)^2 \)

Under the scaling transformation, \( W \rightarrow W / \kappa, L \rightarrow L / \kappa, t_{ox} \rightarrow t_{ox} / \kappa, V_{ds} \rightarrow V_{ds} / \kappa, V_g \rightarrow V_g / \kappa, \) and \( V_t \rightarrow V_t / \kappa, \) Eq. (3.19) becomes

\[
I_{ds} \rightarrow \mu_{eff} (\kappa C_{ox}) \frac{W}{L} \left( \frac{V_g}{\kappa} - \frac{V_t}{\kappa} \right) \frac{V_{ds}}{\kappa} = I_{ds};
\]

and Eq. (3.23) becomes

\[
I_{ds} \rightarrow \mu_{eff} (\kappa C_{ox}) \frac{W}{L} \frac{1}{2m} \left( \frac{V_g}{\kappa} - \frac{V_t}{\kappa} \right)^2 = I_{ds}.
\]

Note that both \( m = 1 + 3t_{ox} / W_{dm} \) and \( \mu_{eff} \), which is a function of \( E_{eff} \), given by Eq. (3.49) are nearly invariant under constant field scaling.
\( I_d = W Q_i V \)
\( v = V_s \) (velocity saturation)
\( = \mu E_{\text{parallel}} \) (mobility dominated)
\( Q_i = \varepsilon_s E_{\text{vertical}} \)

For constant E field scaling, \( I_d = 1/\kappa \)
Problem 4.3

4.3. Apply constant-field scaling rules to the saturation currents from the $n=1$ velocity saturation model, Eq. (3.78), and the fully saturation-velocity limited current, Eq. (3.80), and show that they behave as indicated in Table 4.1.

\[
\text{Eq. (3.78)}: \quad I_{ds} = C_{ox} W V_{sat}(V_g - V_t) \sqrt{1 + \frac{2 \mu_{\text{eff}} (V_g - V_t)}{(m v_{sat} L)} - 1} \sqrt{1 + \frac{2 \mu_{\text{eff}} (V_g - V_t)}{(m v_{sat} L)} + 1}
\]

\[
\text{Eq. (3.80)}: \quad I_{dsat} = C_{ox} W V_{sat}(V_g - V_t)
\]

Since the factor $\mu_{\text{eff}} (V_g - V_t) / (m v_{sat} L)$ is invariant under the scaling transformation, $W \rightarrow W / \kappa, L \rightarrow L / \kappa, t_{ox} \rightarrow t_{ox} / \kappa, V_{ds} \rightarrow V_{ds} / \kappa, V_g \rightarrow V_g / \kappa$, and $V_t \rightarrow V_t / \kappa$, the square-root expression and therefore the fraction in Eq. (3.78) is unchanged after scaling. The saturation current $I_{dsat}$ of Eq. (3.78) then scales the same way as the fully saturation-velocity limited current, Eq. (3.80), i.e.,

\[
I_{dsat} \rightarrow (kC_{ox}) \frac{W}{\kappa} V_{sat} \left( \frac{V_g}{\kappa} - \frac{V_t}{\kappa} \right) \propto \frac{I_{dsat}}{\kappa}
\]
doping \( \sim k \), \( 1k2 \): pn leakage, mobility degradation

Maximum drain depletion width:

\[
W_D = \sqrt{\frac{2\varepsilon_{si} (\psi_{bi} + V_{dd})}{qN_a}}
\]

For \( N_a \rightarrow \kappa N_a \) and \( V_{dd} \rightarrow V_{dd} / \kappa \),

\( W_D \rightarrow W_D / \kappa \) if \( V_{dd} \gg \psi_{bi} \)

However, the source depletion width,

\[
W_S = \sqrt{\frac{2 \varepsilon_{si} \psi_{bi}}{qN_a}}
\]

is indep. of \( V_{dd} \) and only scales as \( W_S \rightarrow W_S / \sqrt{\kappa} \).

Furthermore, the maximum gate depletion width,

\[
W_{dm}^0 = \sqrt{\frac{4\varepsilon_{si} kT \ln(N_a / n_i)}{q^2 N_a}}
\]

scales even less than \( 1 / \sqrt{\kappa} \)
Non-scaling of Subthreshold Current

4.2. Apply constant-field scaling rules to the subthreshold currents, Eq. (3.36), and show that instead of decreasing with scaling \((1/\kappa)\), it actually increases with scaling (note that \(V_g < V_t\) in subthreshold). What if temperature also scales down by the same factor \((T \rightarrow T/\kappa)\)?

Under the scaling transformation, \(W \rightarrow W/\kappa, L \rightarrow L/\kappa, t_{ox} \rightarrow t_{ox}/\kappa, V_{ds} \rightarrow V_{ds}/\kappa, V_g \rightarrow V_g/\kappa, V_t \rightarrow V_t/\kappa\), Eq. (3.36) becomes

\[
I_{ds} \rightarrow \mu_{\text{eff}} \left(\kappa C_{ox}\right) \frac{W}{L} \left(\frac{kT}{q}\right)^2 \exp \left(\frac{m-1}{2} \frac{kT}{q} \frac{V_g - V_t}{\kappa m k T}ight).
\]

The \(\exp (-q V_{ds}/kT)\) term has been neglected since typically \(V_{ds} >> kT/q\). In subthreshold, \(V_g < V_t\) and \(\exp[q(V_g - V_t)/\kappa m k T] > \exp[q(V_g - V_t)/m k T]\) (note that \(\kappa > 1\)), therefore, the subthreshold current increases with scaling faster than \(\kappa I_{ds}\).

If the temperature also scales down by the same factor, i.e., \(T \rightarrow T/\kappa\), then \(I_{ds} \rightarrow I_{ds}/\kappa\), same as the drift current in Exercise 4.1.
Generalized Scaling

Allows electric field to scale up by $\alpha (E \to \alpha E)$ while the device dimensions scale down by $\kappa$, i.e., Voltage scales by $\alpha / \kappa (V \to (\alpha / \kappa) V)$.

More flexible than constant-field scaling,
but has reliability and power concerns.

To keep Poisson's equation invariant under the transformation,

$$(x, y) \to (x, y) / \kappa \text{ and } \psi \to \psi / (\kappa / \alpha)$$

within the depletion region:

$$\frac{\partial^2 \left( \frac{\alpha \psi}{\kappa} \right)}{\partial \left( \frac{x}{\kappa} \right)^2} + \frac{\partial^2 \left( \frac{\alpha \psi}{\kappa} \right)}{\partial \left( \frac{y}{\kappa} \right)^2} = qN_a \varepsilon_{si}$$

$N_a$ should be scaled to $(\alpha \kappa)N_a$. 

207
Generalized scaling rule

electrical field \sim \alpha

(\alpha > 1)
## Rules of Generalized Scaling

<table>
<thead>
<tr>
<th>Scaling assumptions</th>
<th>MOSFET Device and Circuit Parameters</th>
<th>Multiplicative Factor ($\kappa &gt; 1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Device dimensions ($t_{ox}, L, W, x_j$)</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td></td>
<td>Doping concentration ($N_a, N_d$)</td>
<td>$\alpha \kappa$</td>
</tr>
<tr>
<td></td>
<td>Voltage (V)</td>
<td>$\alpha / \kappa$</td>
</tr>
<tr>
<td>Derived scaling behavior of device parameters</td>
<td>Electric field (E)</td>
<td>$\alpha$</td>
</tr>
<tr>
<td></td>
<td>Depletion layer width ($W_d$)</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td></td>
<td>Capacitance ($C = \varepsilon A/t$)</td>
<td>$1/\kappa$</td>
</tr>
<tr>
<td></td>
<td>Inversion layer charge density ($Q_i$)</td>
<td>$\alpha$</td>
</tr>
<tr>
<td></td>
<td>Carrier velocity (V)</td>
<td>$\alpha$</td>
</tr>
<tr>
<td></td>
<td>Current, drift (I)</td>
<td>$\alpha^2 / \kappa$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha / \kappa$</td>
</tr>
<tr>
<td>Derived scaling behavior of circuit parameters</td>
<td>Circuit delay time ($\tau \sim CV/I$)</td>
<td>$1/\alpha \kappa$</td>
</tr>
<tr>
<td></td>
<td>Power dissipation per circuit ($P \sim VI$)</td>
<td>$\alpha^3 / \kappa^2$</td>
</tr>
<tr>
<td></td>
<td>Power-delay product per circuit ($P \times \tau$)</td>
<td>$\alpha^2 / \kappa^3$</td>
</tr>
<tr>
<td></td>
<td>Circuit density ($\propto 1/A$)</td>
<td>$\kappa^2$</td>
</tr>
<tr>
<td></td>
<td>Power density ($P/A$)</td>
<td>$\alpha^3$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha^2$</td>
</tr>
</tbody>
</table>
NON ideal Factor

\[ V_T = V_{FB} + 2|\phi_{FP}| + \frac{\sqrt{2\epsilon eN_a (2|\phi_{FP}|)}}{C_{ox}} \] (7.3)

\[ \sim (k)^{1/2} \]
Figure 7.2 Comparison of ideal and experimental plots of $\sqrt{I_D}$ versus $V_{GS}$.

Figure 7.3 Energy-band diagram when $|\phi_{fp}| < \phi_s < 2|\phi_{fp}|$. 
Figure 7.4 (a) Cross section along channel length of n-channel MOSFET. Energy-band diagrams along channel length at (b) accumulation, (c) weak inversion, and (d) inversion.

\[
I_D(\text{sub}) \propto \left[ \exp\left( \frac{V_G - V_T}{V_t} \right) \right] \left[ 1 - \exp\left( \frac{-V_{DS}}{V_t} \right) \right] \tag{7.4}
\]
Figure 7.5 Subthreshold current-voltage characteristics for several values of substrate voltage (the threshold voltage is indicated on each curve). (From Schroder [16].)
Figure 7.6 Cross section of an n-channel MOSFET showing the channel length modulation effect.
Channel length modulation

\[ x_p = \sqrt{\frac{2\varepsilon_s |\phi_{Fp}|}{eN_a}} \]  \hspace{1cm} (7.5)

\[ x_p = \sqrt{\frac{2\varepsilon_s (|\phi_{Fp}| + V_{DS})}{eN_a}} \]  \hspace{1cm} (7.6)

\[ \Delta L = \sqrt{\frac{2\varepsilon_s}{eN_a}} \left[ \sqrt{|\phi_{Fp}| + V_{DS\text{ (sat)}} + \Delta V_{DS}} - \sqrt{|\phi_{Fp}| + V_{DS\text{ (sat)}}} \right] \] \hspace{1cm} (7.7)

\[ \Delta V_{DS} = V_{DS} - V_{DS\text{ (sat)}} \] \hspace{1cm} (7.8)

\[ I'_D = \left( \frac{L}{L - \Delta L} \right) I_D \] \hspace{1cm} (7.9)
Figure 7.7 Current-voltage characteristics of a MOSFET showing short-channel effects. (*From Sze [21].*)
Figure 7.8 Vertical electric field in an n-channel MOSFET.

Figure 7.9 Schematic of carrier surface scattering effects.
Figure 7.10 Measured inversion layer electron mobility versus electric field at the inversion layer. 
(from Yang [26].)

Figure 7.11 Comparison of $I_D$ versus $V_D$ characteristics for constant mobility (dashed curves) and for field-dependent mobility and velocity saturation effects (solid curves). 
(from Sze [21].)
Universal mobility
\[ \mu = f(\text{electrical field}) \]

\[
\frac{1}{\mu_{\text{uni}}} = \frac{1}{\mu_{\text{imp}}} + \frac{1}{\mu_{\text{sr}}} + \frac{1}{\mu_{\text{ph}}}
\]
\[ V_{DS} = V_{DS}^{(sat)} = V_{GS} - V_T \quad (7.12) \]

\[ I_D^{(sat)} = WC_{ox} (V_{GS} - V_T) v_{sat} \quad (7.13) \]

\[ \left. \frac{I_D}{I_D^{ideal}} \right|_{V, sat} = \frac{WC_{ox} (V_{GS} - V_T) v_{sat}}{W \mu_n C_{ox} (V_{GS} - V_T)^2} = \frac{2L}{\mu_n} \frac{v_{sat}}{(V_{GS} - V_T)} \quad (7.14) \]

\[ g_{ms} = \frac{\partial I_D^{(sat)}}{\partial V_{GS}} = WC_{ox} v_{sat} \quad (7.15) \]

\[ f_T = \frac{g_{ms}}{2\pi C_G} = \frac{WC_{ox} v_{sat}}{2\pi (C_{ox} WL)} = \frac{v_{sat}}{2\pi L} \quad (7.16) \]
Figure 7.10 Cross section of a long n-channel MOSFET (a) at flat band and (b) at inversion.

Figure 7.11 Cross section of a short n-channel MOSFET at flat band.
Figure 7.14 Charge sharing in the short-channel threshold voltage model. (From Yau [27].)
\[ x_s \approx x_d \approx x_{dT} \equiv x_{dT} \quad (7.18) \]

\[ |Q'_B| \cdot L = eN_a x_{dT} \left( \frac{L + L'}{2} \right) \quad (7.19) \]

\[ \frac{L + L'}{2L} = \left[ 1 - \frac{r_j}{L} \left( \sqrt{1 + \frac{2x_{dT}}{x_j}} - 1 \right) \right] \quad (7.20) \]

\[ |Q'_B| = eN_a x_{dT} \left[ 1 - \frac{r_j}{L} \left( \sqrt{1 + \frac{2x_{dT}}{x_j}} - 1 \right) \right] \quad (7.21) \]

\[ \Delta V_T = -\frac{eN_a x_{dT}}{C_{ox}} \left[ \frac{r_j}{L} \left( \sqrt{1 + \frac{2x_{dT}}{r_j}} - 1 \right) \right] \quad (7.22) \]
Figure 7.17 Cross section of an n-channel MOSFET showing the depletion region along the width of the device.
\[ Q_B = Q_{B0} + \Delta Q_B \quad (7.24) \]

\[ |Q_{B0}| = eN_a W L x_{dT} \quad (7.25) \]

\[ \Delta Q_B = eN_a W L x_{dT} (\xi x_{dT}) \quad (7.26) \]

\[ |Q_B| = |Q_{B0}| + |\Delta Q_B| = eN_a W L x_{dT} + eN_a W L x_{dT} (\xi x_{dT}) \]

\[ = eN_a W L x_{dT} \left(1 + \frac{\xi x_{dT}}{W}\right) \quad (7.27) \]

\[ \Delta V_T = \frac{eN_a x_{dT}}{C_{ox}} \left(\frac{\xi x_{dT}}{W}\right) \quad (7.28) \]
Body effect is weaker for short channel device.

Figure 7.20 Variation of $V_T$ with $V_{BS}$ for three different channel lengths and different values of $V_{DS}$. 

$L = 10 \mu m$
$V_{DS} = 10 V$

$L = 1.5 \mu m$
$V_{DS} = 2 V$

$L = 1.5 \mu m$
$V_{DS} = 8 V$

$L = 0.7 \mu m$
$V_{DS} = 1 V$

$L = 0.7 \mu m$
$V_{DS} = 2 V$

$\sqrt{V_{BS} + |\phi_{FP}|} (V^{1/2})$
Figure 7.21 Short-channel MOSFET showing effect of increasing $V_{DS}$. With an increase in $V_{DS}$, there is less channel charge that is controlled by the substrate voltage. Channel length modulation
Figure 7.22 (a) Equipotential plot along the surface of a long-channel MOSFET. (b) Equipotential plot along the surface of a short-channel MOSFET before and after punch-through.
Figure 7.23 Typical $I-V$ characteristics of a MOSFET exhibiting punch-through effects.
Figure 7.24 Hot carrier generation current components, and electron injection into the oxide.
Figure 7.25 (a) Ion-implanted profile approximated by a delta function. (b) Ion-implanted profile approximated by a step function, in which the depth $x_I$ is less than the space charge width $x_{dT}$. 
\[ V_{DS} = V_{DS} \text{(sat)} = V_{GS} - V_T \tag{7.12} \]

\[ \Delta V_T = + \frac{eD_I}{C_{ox}} \tag{7.29} \]

\[ x_{dT} = \sqrt{\frac{2 \varepsilon_s}{eN_a} \left[ 2|\phi_{Fp}| - \frac{ex_I^2}{2\varepsilon_s} (N_s - N_a) \right]}^{1/2} \tag{7.30} \]

\[ V_T = V_{T0} + \frac{eD_I}{C_{ox}} \tag{7.31} \]

\[ V_{T0} = V_{FB0} + 2|\phi_{Fp0}| + \frac{eN_a x_{dT0}}{C_{ox}} \tag{7.33} \]
Figure 7.36 Cross section of an example of a floating-gate nonvolatile memory.
Figure 7.31 Cross section of a double-diffused MOS (DMOS) transistor.

Figure 7.32 Cross section of a vertical channel MOS (VMOS) transistor.
Table 7.2 Characteristics of two power MOSFETs

<table>
<thead>
<tr>
<th>Parameter</th>
<th>2N6757</th>
<th>2N6792</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{DS}$ (max) (V)</td>
<td>150</td>
<td>400</td>
</tr>
<tr>
<td>$I_D$ (max) (A) (at $T = 25^\circ$C)</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>$P_D$ (W)</td>
<td>75</td>
<td>20</td>
</tr>
</tbody>
</table>

Figure 7.33 A HEXFET structure.
Figure 7.34 Typical drain-to-source resistance versus drain current characteristics of a MOSFET.

\[ R_{\text{on}} = R_S + R_{\text{CH}} + R_D \quad (7.34) \]

\[ R_{\text{CH}} = \frac{L}{W \mu_n C_{\text{ox}} (V_{GS} - V_T)} \quad (7.35) \]