1 CMOS technology scaling and its implications

Tetsuya Iizuka

In these five decades after Gordon Moore propounded Moore's Law in 1965 [1], the semiconductor industry has been continuously growing in accordance with his expectations, and we are now facing sub-10-nm feature size transistors. Thanks to the immense intelligence devoted to pushing this exponential technology scaling, the transistor count on a single microprocessor chip almost doubles every 2 years, as shown in Figure 1.1, and this trend has even been accelerated in recent years. As a consequence, after the Cray-1 was marketed as the world's fastest computer in 1976, we now have almost 1000-times improved performance on only a single 300 mm^2 chip with a billion of integrated transistors operating with a 50-times faster clock [2, 3].

Besides the tremendous benefits of transistor technology scaling, we have been facing a lot of circuit design implications and problems with these scaled transistors. Due to a lot of imperfections in both the devices and the fabrication processes, the difficulties of circuit design are ever-increasing, and it is almost impossible to build highly sophisticated VLSI systems without a set of calibration and digital/analog assisting techniques.

This chapter first briefly looks over the fundamentals of scaling. Then we revisit several traditional scaling implications such as short-channel effects, which is followed

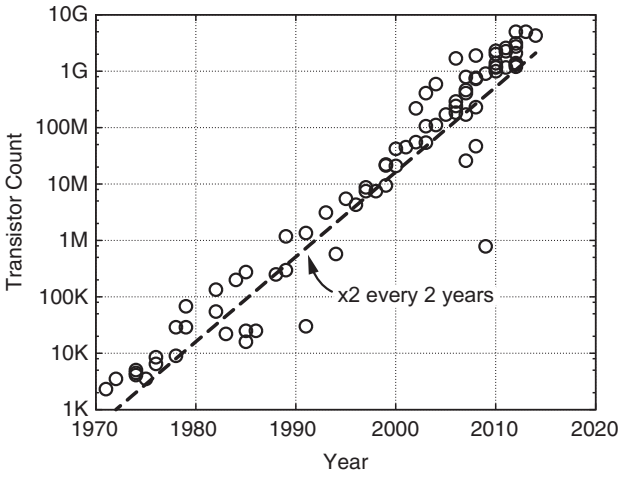


Figure 1.1 Transistor count on a single microprocessor chip (data obtained from [4]).

Table 1.1 Scaling in device dimensions and voltages.

Device or circuit parameter	Symbol	Scaling factor
Gate length	L	$1/\kappa$
Gate width	W	$1/\kappa$
Gate oxide thickness	t_{ox}	$1/\kappa$
Supply voltage	V_{DD}	$1/\kappa$
Gate-source voltage	V_{GS}	$1/\kappa$
Drain-source voltage	V_{DS}	$1/\kappa$
Threshold voltage	V_{TH}	$1/\kappa$
Doping concentration	N_A, N_D	κ

Table 1.2 Scaling results for device characteristics.

Performance of device	Symbol	Expression	Scaling factor
Number of devices per unit area	N_{tr}	$\alpha 1/(L{\cdot}W)$	κ^2
Gate oxide capacitance per unit area	C_{ox}	$\alpha 1/t_{ox}$	κ
Gate oxide capacitance	C_{gate}	$C_{ox}{\cdot}L{\cdot}W$	$1/\kappa$
Drain saturation current	I_D	$\frac{1}{2}\mu C_{ox}{\cdot}\frac{W}{L}(V_{GS}-V_{TH})^2$	$1/\kappa$
Intrinsic delay per device	τ	$C_{gate}{\cdot}V_{DD}/I_D$	$1/\kappa$
Power dissipation per device	P	$I_D{\cdot}V_{DD}$	$1/\kappa^2$
Power density	P_{dens}	$I_D{\cdot}V_{DD}{\cdot}N_{tr}$	1

by discussion of the crucial impacts of process variation and parasitic elements. This chapter also introduces several design issues specific to the recent nano-scale transistors, which include well proximity/STI stress-induced performance variations and aging effects due to hot carrier injection (HCI), time-dependent dielectric breakdown (TDDB), and negative bias temperature instability (NBTI).

1.1 Scaling theory and technology roadmap

CMOS technology advance relies on scaling theory, which was first formulated by Dennard *et al.* in 1974 [5]. Tables 1.1 and 1.2 summarize the changes in device sizes and performance, which follow the scaling by a factor of κ ($\kappa > 1$). Ideal scaling reduces all lateral and vertical dimensions by κ and all nodal voltages and the supply voltage are reduced simultaneously by κ. As also illustrated in Figure 1.2, all the doping concentrations are increased by κ to scale the width of each depletion region at the same rate. Since the dimensions and voltages scale together at the same rate κ, the electric field strength at any corresponding point is unchanged, hence the name "constant-field scaling" is often used [6]. As a consequence, all the primary performance metrics of MOS devices are improved, as summarized in Table 1.2. Note that even though the device density is increased by κ^2, the power density remains constant due to the reduced

Table 1.3 Scaling results for analog performance.

Analog performance	Symbol	Expression	Scaling factor
Transconductance	g_m	$\mu C_{ox} \cdot \frac{W}{L}(V_{GS} - V_{TH})$	1
Thermal noise of transistors (input referred)	$< v_n^2 >$	$4kT \cdot \frac{2}{3} \cdot \frac{1}{g_m}$	1
Dynamic range	DR	$\sim \frac{V_{DD}}{< v_n >}$	$1/\kappa$
Cut-off frequency	f_T	$\sim \frac{g_m}{C_{gate}}$	κ

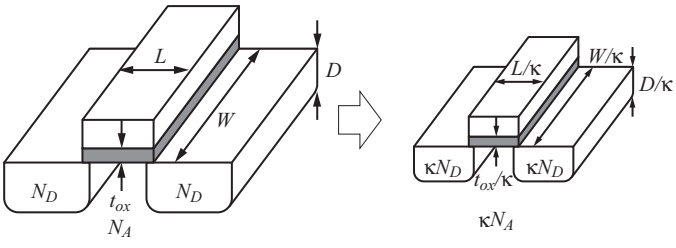


Figure 1.2 Ideal scaling of MOS transistors.

power dissipation per device by κ^2. Thus, the requirements for cooling equipment are essentially unchanged with scaling [5].

These scaling results provide significant benefits especially for digital systems. Once we look at this from the analog viewpoint, the scaling gives a different perspective. As shown in Table 1.3, the transconductance g_m of a transistor remains constant with scaling. Therefore, the thermal noise from the scaled transistor also stays the same and the analog voltage dynamic range, which is usually defined as the ratio of the maximum allowable voltage swing and the noise level, is decreased by $1/\kappa$. To keep the same dynamic range with scaling, we have to increase the transistor width W by κ, thus increasing the drain current I_D by the same ratio. This makes the power dissipation $I_D \cdot V_{DD}$ constant for the same dynamic range requirement. Therefore, technology scaling does not provide a power scaling advantage for analog designers. Furthermore, the scaled supply voltage usually introduces a lot of analog design difficulties, especially in the case of stacked transistors. On the other hand, one of the big impacts of scaling on analog designs is the improvement of the device cut-off frequency f_T, where a transistor provides a unity current gain. As listed in Table 1.3, the transistor cut-off frequency is often approximated by the ratio between g_m and C_{gate}, hence it increases along with technology scaling; Figure 1.3 plots several measured results from the literature [7–12]. This f_T is an important performance criterion especially for high-frequency analog design, and its improvement due to scaling has enabled CMOS RF/THz circuit applications.

Although we have enjoyed significant performance improvement through scaling so far, we have reached practical limitations, and it is hard to keep on track with ideal

Table 1.4 Some MOS transistor parameters from the ITRS roadmap.

Year	Unit	2001	2005	2010	2015	2020
Physical gate length	nm	65	32	27	16.7	10.6
Equivalent gate oxide thickness	nm	2.3	1.2	0.95	0.73	0.59
Power supply voltage	V	1.2	1.1	0.97	0.83	0.75
Threshold voltage	V	–	0.195	0.289	0.206	–
NMOS saturation current	μA/μm	900	1020	1200	1340	–
OFF current	nA/μm	10	60	100	100	100
Total gate capacitance	fF/μm	–	0.573	0.97	1.07	0.95
NMOS intrinsic delay (*CV/I*)	ps	1.6	0.87	0.78	0.666	–

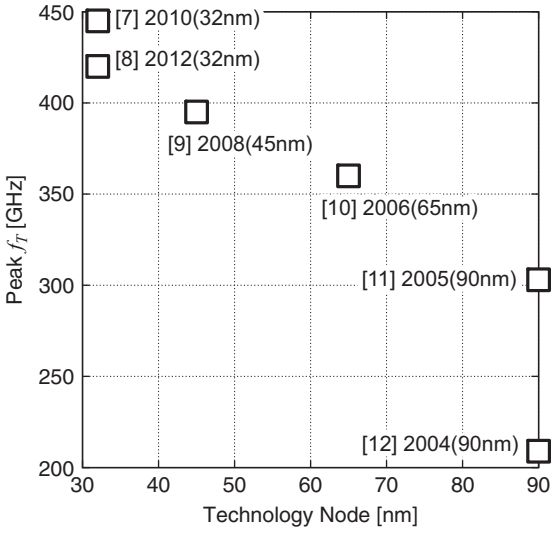


Figure 1.3 Measurement results of cut-off frequencies.

scaling, especially with respect to the "constant-field" perspective. Table 1.4 presents some primary parameters of the MOS transistor from the International Technology Roadmap for Semiconductors (ITRS) [13] after 2001. From 2001 to 2015, the physical gate length scales by four times, whereas the oxide thickness scales only by three times. Since the gate oxide thickness has been approaching atomic dimensions and now consists of only a few numbers of atomic layers, further oxide thickness scaling, while maintaining low gate direct tunneling current and reliability, becomes more and more challenging as it approaches the limit of one atomic layer thickness [14].

When we take a look at the power supply voltage, it decreases by only 2/3 while the threshold voltage has even increased from 65 nm to 16 nm. The scale-down of the threshold voltage is mainly limited by the exponential increase of subthreshold leakage

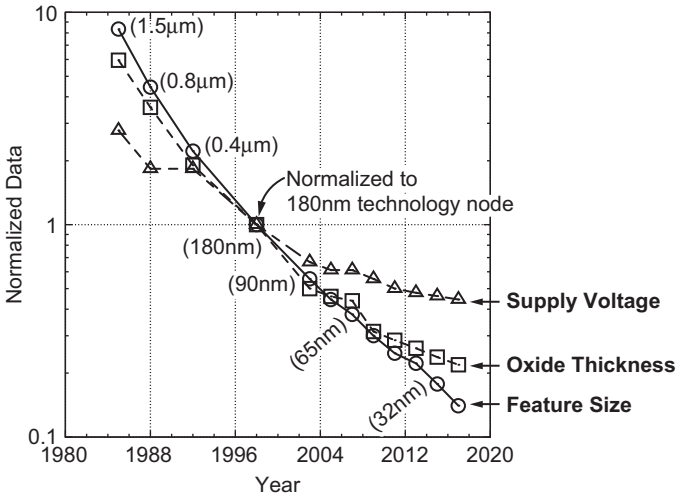


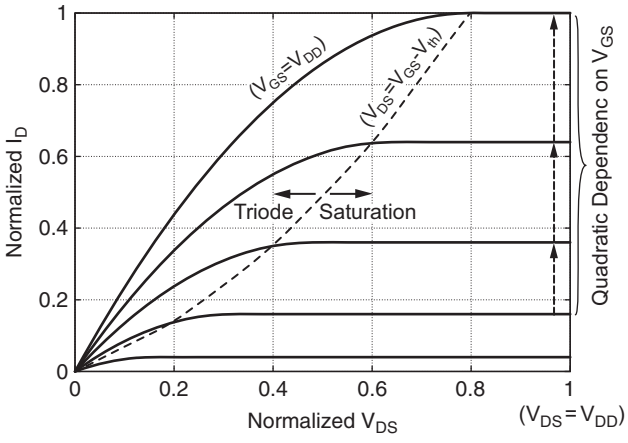
Figure 1.4 Scaling trends of technology node, oxide thickness, and supply voltage.

current (I_{off}), hence it also limits the power-supply scaling [15]. In addition, the random dopant fluctuation within the scaled channel region becomes a more and more dominant cause of threshold voltage variation.

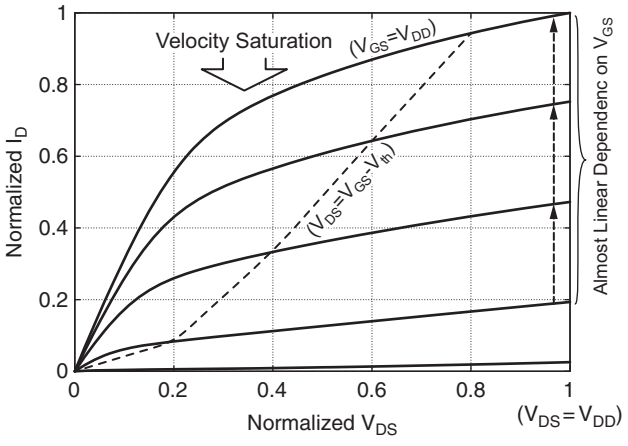
These scaling trends [13] are clearly depicted in Figure 1.4. This graph plots technology nodes, oxide thickness, and supply voltage, which are normalized to the values at the 180 nm technology node. This graph shows that the supply voltage scaling no longer follows the feature size scaling and is almost saturated after 180 nm, and also that the oxide thickness approaches the limit and deviates from the ideal scaling after 65 nm technology. Due to this non-ideal scaling, the performance improvement of the MOS transistor through technology scaling gradually diminishes. Furthermore, this "constant-voltage scaling" situation causes several adverse effects mainly due to the increasing internal electric field [16, 17].

1.2 Short-channel effects

By scaling the gate length of the transistors, we see several phenomena that impact the device performance and they have become apparent below approximately 3 μm gate length [6]. Figure 1.5 compares the NMOS transistor $V_{DS}-I_D$ characteristics of (a) long-channel 10 μm and (b) short-channel 65 nm technologies. In the case of the 10 μm long-channel device, a transistor acts as a perfect current source when it is in saturation, whereas in the 65 nm technology we can no longer see a clear boundary between triode and saturation regions, and it no longer looks like a current source. This characteristic is a consequence of several phenomena which accompany scaled short-channel devices. In this section, we briefly look into these phenomena, which are commonly known as *short-channel effects*.



(a) L_g = 10 μm



(b) L_g = 65 nm

Figure 1.5 $V_{DS}-I_D$ characteristics of NMOS transistors for 10 μm and 65 nm technologies.

1.2.1 Threshold voltage dependence on channel length

It is commonly known that the threshold voltage of the scaled transistor exhibits gate length dependence. As shown in Figure 1.6, the threshold voltage tends to decrease along with gate length scaling. This effect is explained in Figure 1.7 for the NMOS transistor case. The depletion regions extended from the source and drain regions intrude into the channel region, and some of the immobile charge beneath the channel couples with the charge in the source and drain regions. Therefore the total immobile

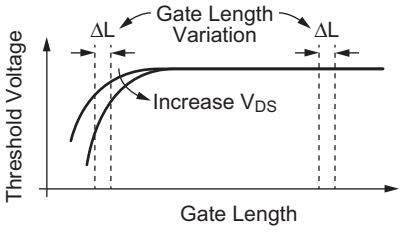


Figure 1.6 Threshold voltage dependence on gate length.

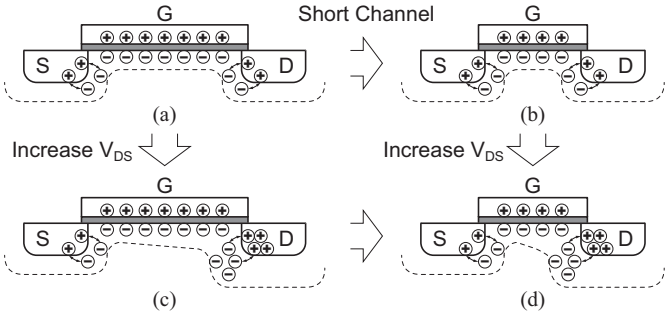


Figure 1.7 Source/drain depletion region affect the threshold voltage.

negative charge seen from the gate is reduced, and hence the total positive charge required to form an inversion layer decreases. Comparison between Figure 1.7(a) and (b) clearly explains that the impact of this phenomenon becomes non-negligible in the short-channel devices. In addition, when we apply $V_{DS} > 0$, the depletion region associated with the drain region is extended and this phenomenon becomes more considerable, as shown in Figure 1.7(c) and (d). Thus, the threshold voltage decreases further by applying V_{DS}, as depicted in Figure 1.6.

This threshold voltage dependence on the gate length produces a practical problem, because the device gate length cannot be controlled accurately during fabrication, and there is always a certain amount of gate length variation. Even if the absolute gate length variation ΔL is the same, its impact on the threshold variation is exacerbated by this phenomenon, as shown in Figure 1.6.

1.2.2 Drain-induced barrier lowering (DIBL)

In an NMOS transistor device, the channel potential, i.e., the potential barrier for electrons, is controlled by the gate voltage, and this is conventionally independent of the drain voltage in long-channel devices. In short-channel devices, however, this potential barrier is also lowered by the drain voltage, because the drain region is now

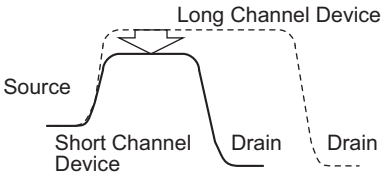


Figure 1.8 Drain-induced barrier lowering (DIBL).

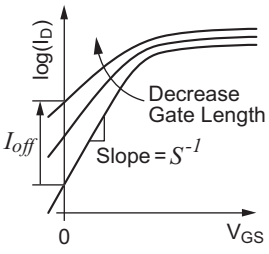


Figure 1.9 Subthreshold slope is degraded due to DIBL.

located close enough to have an impact on the potential barrier, as shown in Figure 1.8. Thus this phenomenon is called drain-induced barrier lowering (DIBL). This phenomenon also decreases the device threshold. Another outcome of DIBL which impacts circuit design is known as the subthreshold slope degradation. Since DIBL lowers the potential barrier even in the case of $V_{GS} < V_{th}$, it causes an increase of leakage current through the source and drain terminals. This effect is depicted in Figure 1.9. As the gate length shrinks, the slope of the $\log I_D$ vs. V_{GS} curve in the subthreshold region becomes lower, hence the current I_{off} at $V_{GS} = 0$ increases significantly. The inverse of this slope is commonly called the subthreshold slope $S = \frac{dV_{GS}}{d(\log I_D)}$ and is used as a measure of the controllability of the channel potential through the gate terminal. In bulk MOS transistor devices, the subthreshold slope is known to be limited to the minimum value of 60 mV/dec, and a typical bulk MOS device has an S of around 70 to 100 mV/dec [18]. To improve this subthreshold behavior, several novel device structures, e.g., FinFET, have been developed [19–21].

1.2.3 Velocity saturation

The velocity of the carrier v in the MOS channel is accelerated by the lateral electric field E within the channel proportionally to the electric field as $v = \mu E$, where μ is the carrier mobility. After the electric field reaches its critical value E_c, as Figure 1.10 illustrates for an NMOS case, the carrier velocity is saturated and limited to a constant velocity v_{sat}, mainly due to scattering effects. For example, the value of this critical electric field for an electron is known to be 1–3 V/cm. Thus in short-channel devices such as 65 nm gate length, a V_{DS} of 100–200 mV easily makes the carrier velocity saturate even before pinch-off occurs. This situation is explained in Figure 1.11 and also

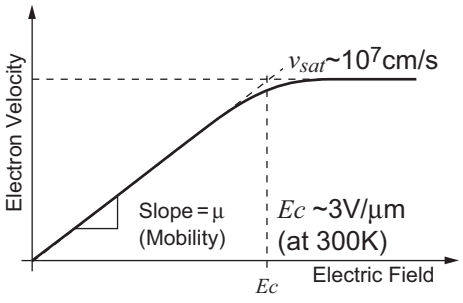


Figure 1.10 Charge velocity in a MOS device vs. electric field.

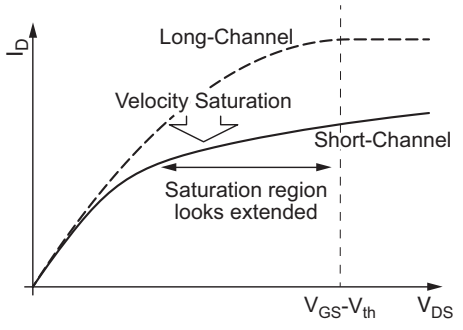


Figure 1.11 Velocity saturation limits the maximum drain current in short-channel devices.

noted in Figure 1.5(b). Due to the saturated velocity of the carrier, the maximum drain current is limited before $V_{DS} > V_{GS} - V_{th}$ in short-channel devices, and it looks like the saturation region is extended to a lower V_{DS}. In addition to this, I_D exhibits a linear dependence on V_{GS} in short-channel devices due to velocity saturation, whereas it has a square dependence in long-channel devices, as shown in Figure 1.5. This limits the maximum current swing controlled by the gate input voltage in scaled transistor devices.

1.3 Scaling impact on power consumption

As explained in section 1.1, while the feature size has been scaling in accordance with Moore's Law, ideal, constant-field scaling is no longer maintained. Although the power density is expected to be constant under ideal scaling, the non-ideal technology scaling due to the limitations of the power supply/threshold voltages and oxide thickness has led to an increase in power density. In addition to this active power density, an even worse situation has been happening in terms of the standby leakage power. Figure 1.12 summarizes this situation [22]. The active power density has been increasing about 30% per technology generation while the standby power

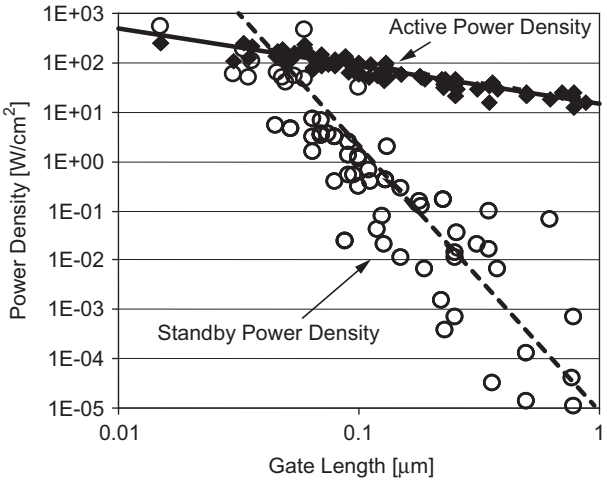


Figure 1.12 Active and standby power density trend [22] (data courtesy of IBM).

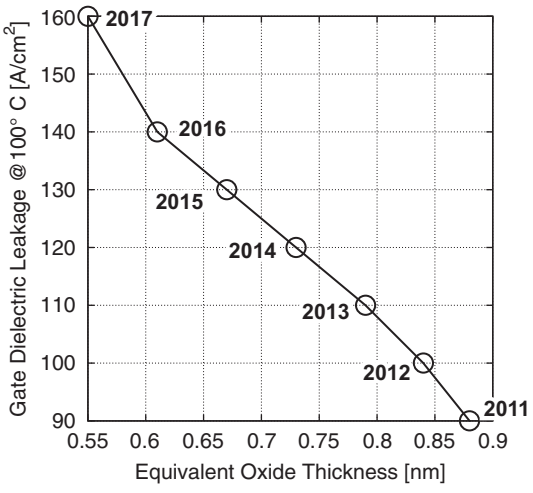


Figure 1.13 Gate leakage current versus equivalent oxide thickness (data obtained from ITRS2011 [13]).

density has grown 3- to 4-fold per generation. Therefore, the static leakage contribution to the total power consumption has become equal to or even worse than that of dynamic power. This standby power is dominated not only by the subthreshold leakage through the channel but also by the gate direct tunneling current, i.e., gate leakage. As shown in Figure 1.13, the gate leakage increases along with the scaling of the gate dielectric and is no longer negligible, although a lot of research on gate materials, including high-*k* gate dielectric materials and metal gates, has been devoted to preventing its steep growth [23].