Analog Integrated Circuits Design using the Inversion Coefficient

Preview Edition

Christian Enz (christian.enz@epfl.ch)

23.06.2026

Table of contents

Licensing

1 Technology roadmap

1.1 Introduction

> **Warning**
>
> A technology roadmap is by definition already obsolete as you write it. For this reason I hesitated to publish this chapter in its current status. However, the sections about *energy efficiency and power consumption* and *voltage scaling* are more fundamental and will not change drastically over the years. This made me decide to still publish this chapter in its current state. It will be the last chapter that I will update before finishing writing the book.

Integrated circuits (ICs) or chips have changed our daily life. We find them everywhere in almost every device starting with the PC which has been the driver for many IC innovations. Numerous chips are also found in data centers and in the cloud for artificial intelligence (AI) computations, servers, storage and networking setting big challenges in terms of power consumptions, heath dissipation and storage capacity as well as communication bandwidth. The reduction of power consumption has made it possible to have more and more powerful mobile devices among which the smartphones has now become essential for our daily life. The miniaturization and further reduction of power consumption enables new internet-of-things (IoT) applications to monitor cities, enable smart agriculture and increase manufacturing productivity with smart factories through low power wireless channels.

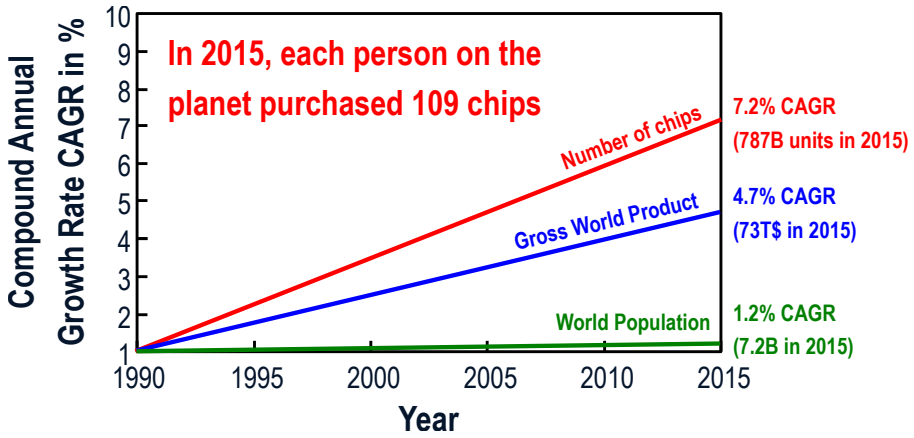


Figure 1.1: Semiconductor unit growth [1] [2] (source: [2]).

As illustrated in Figure 1.1, the proliferation of electronics devices contributes to the steady growth of the semiconductor market. The compound annual growth rate (CAGR) of the number of semiconductor units is about 7.2%, which is compared to the CAGR of the gross world product which is estimated to 4.7% and to the CAGR of the world population (about 1.2%). These numbers mean that, in 2015, each person on the planet purchased 109 chips [1] [2] !

As illustrated in Figure 1.2, after the initial fast growth period around the 1990s, the worldwide growth rate of the semiconductor revenue parallels that of the gross world product (GWP) for the past 20 years [3]. The global semiconductor market is estimated at 450~billion USD in revenue for 2020 while the products using these semiconductors represent a global revenue of 2 trillion USD which corresponds to about 3.5% of the global gross domestic product (GDP). Over the last four decades (1980 to 2020),

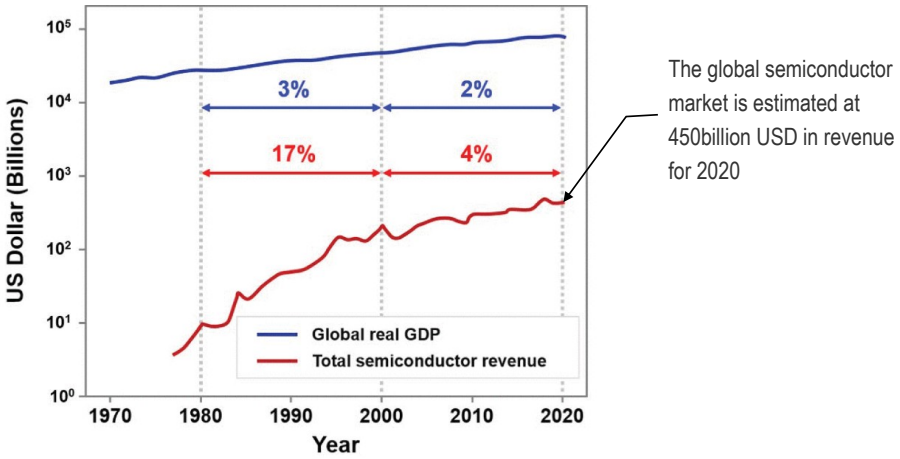


Figure 1.2: Growth of the semiconductor business (source: [3]).

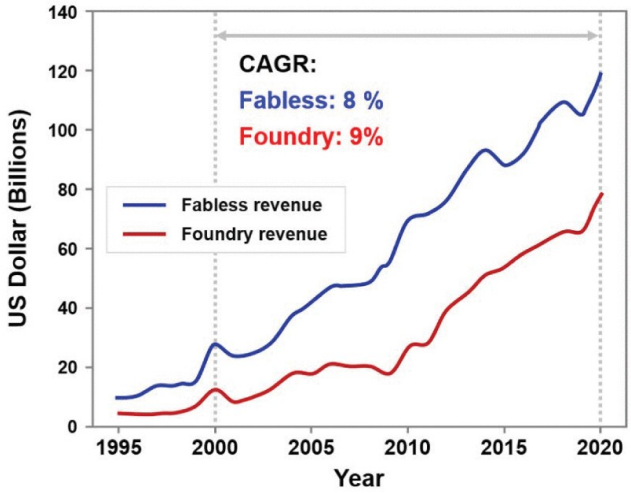


Figure 1.3: Growth of the fabless and foundry business (source: [3]).

the total semiconductor revenue has grown at a 10% CAGR compared to a steady global GDP growth of 3% CAGR.

Initially the semiconductor business was exclusively driven by *integrated device manufacturers* (IDM) who needed fabs to build chips used in their products. However, the global IC supply chain has been profoundly transformed by the *foundry business model*. The latter was established more than three decades ago to capture the overflow demand from the IDMs but has evolved into design platforms that can lead to system-level breakthrough [3]. Today, fabless companies are manufacturing very advanced chips including field programmable gate arrays (FPGAs), GPUs, application processors, communication chipsets, networking chips, and more recently CPUs [3]. This fabless business model enable fabless companies to have access to the world's most-advanced technologies without the need to invest in the significant capital required to develop the semiconductor technologies and build the foundries. The growth of the foundry business resulted in 4 of the top 10 semiconductor companies in the world being either fabless or foundry companies as of 2019 [3].

1.2 CMOS technology scaling

In this section, we will review the fantastic journey technology scaling from the intial idea of Gordon Moore to today's extraordinary evolution of CMOS technology.

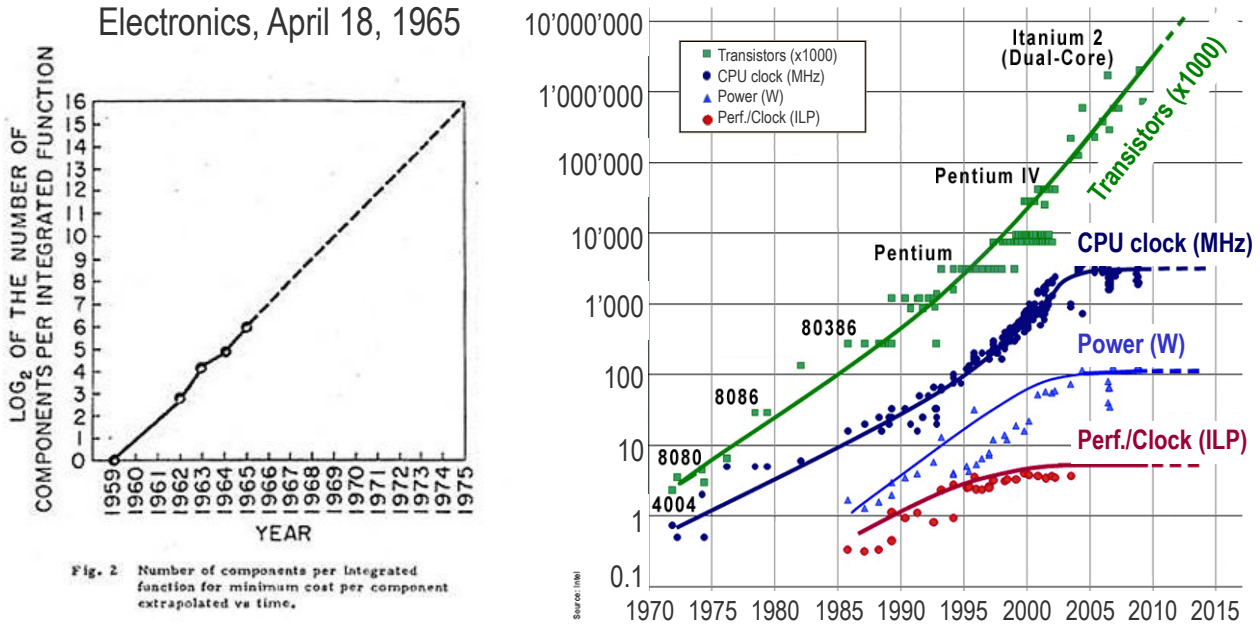


Figure 1.4: Left figure: original paper published by G. Moore in 1965 [4]. Right figure: extension with data until 2010.

In his seminal paper published in 1965 [4], Gordon Moore, who is one of the founders of Intel, predicted that the number of transistors on a silicon chip would double every two years [4] [6]. This prediction, which is illustrated on the left figure of Figure 1.4, turned out to become reality. This is shown by the green squares in the right figure of Figure 1.4 or the orange triangles in Figure 1.5, which represent the number of transistors on a single die for microprocessors starting at a few thousands for the first 4004 microprocessor and reaching 100 billions and hence corresponding to an increase by a factor of almost 100 million.

Figure 1.6 shows the largest single computer chip ever built as a wafer scale engine dedicated to AI that embeds 4 trillion transistors in a 5 nm CMOS technology [7] ! Although not really representative of the most common commercial microprocessor chips, it however shows the ultimate chip complexity that can be handled today.

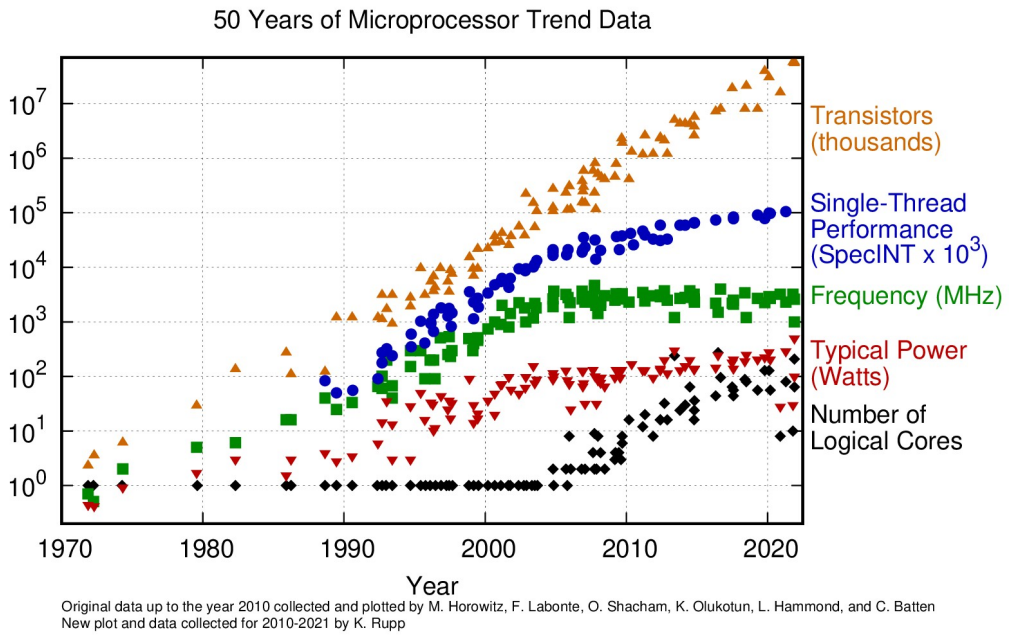


Figure 1.5: Moore's law with updated data until 2020 (source: [5]).

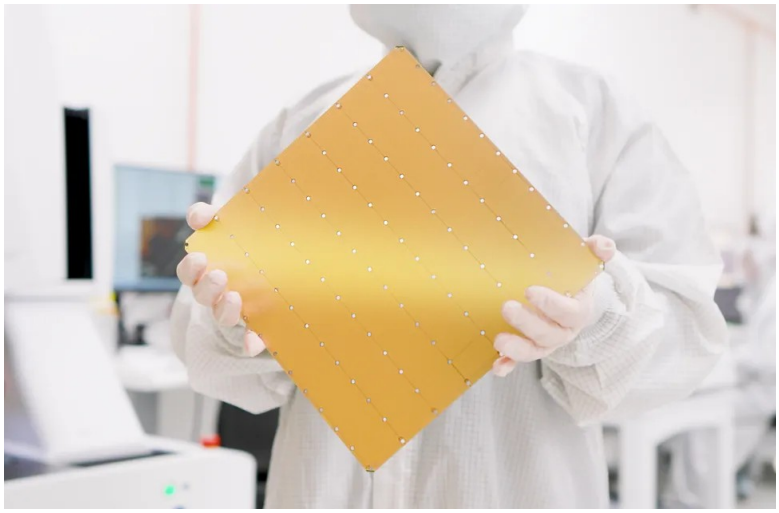


Figure 1.6: Cerebras' 4 trillion transistors waferscale AI chip (source: [7]).

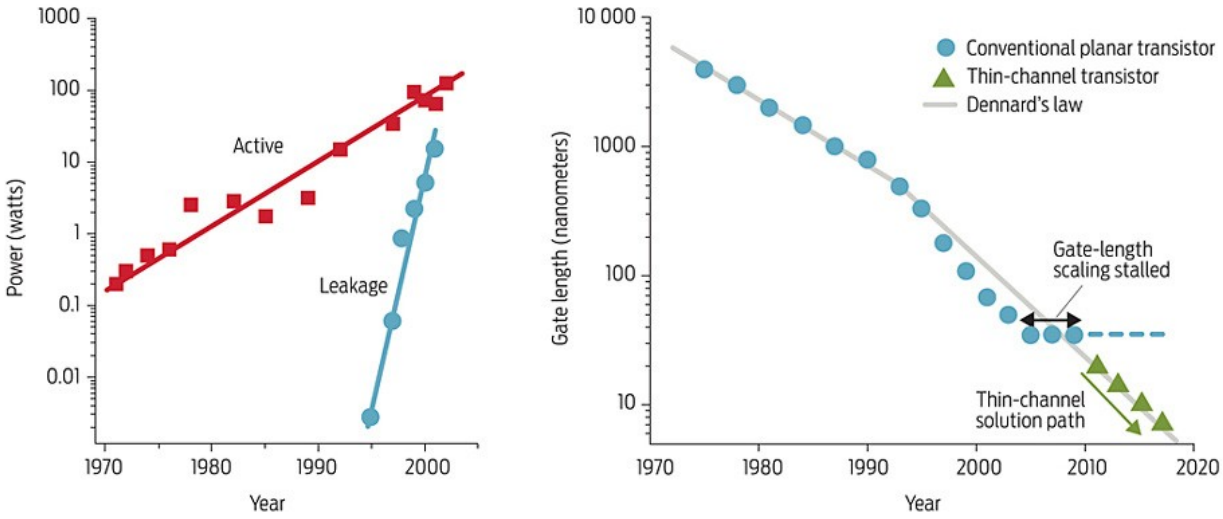


Figure 1.7: CMOS Technology Scaling (source: [8]).

This exponential increase was enabled by the reduction of the transistor dimensions initially following the Dennard scaling law [9] [10]. In particular, the reduction of the transistor length, as illustrated in the right plot of Figure 1.7, enabled a spectacular increase of the clock frequency reaching several GHz. However, this frequency increase led to an unsustainable increase of the power consumption ultimately limited by the maximum power that can be dissipated on a single chip. In order to limit this power dissipation to about 100 W (as shown by the light blue triangles), the clock frequency has been stalled at a few GHz (as shown by the dark blue circles). More computing power was obtained by introducing parallelism, increasing the number of processor cores taking advantage of the transistor density increase. This has enabled to continue to increase the computing power while stopping the increase of the power consumption and maintaining it below 100 W.

As shown in the left plot of Figure 1.7, the scaling of CMOS technology came with an increase of the active power consumption shown by the red square in the left plot [8]. The latter is made of the $f \cdot C \cdot V_{DD}^2$ active power consumption and the leakage power which grew exponentially, reaching the same level than the active power and becoming a major contribution to the overall power around year 2000. This leakage power increase eventually put a halt to the transistor scaling (right figure), a progression called Dennard's law [9]. Switching to alternate transistor architectures allowed chipmakers to shrink transistors further, extending Moore's law by boosting transistor density and performance [8].

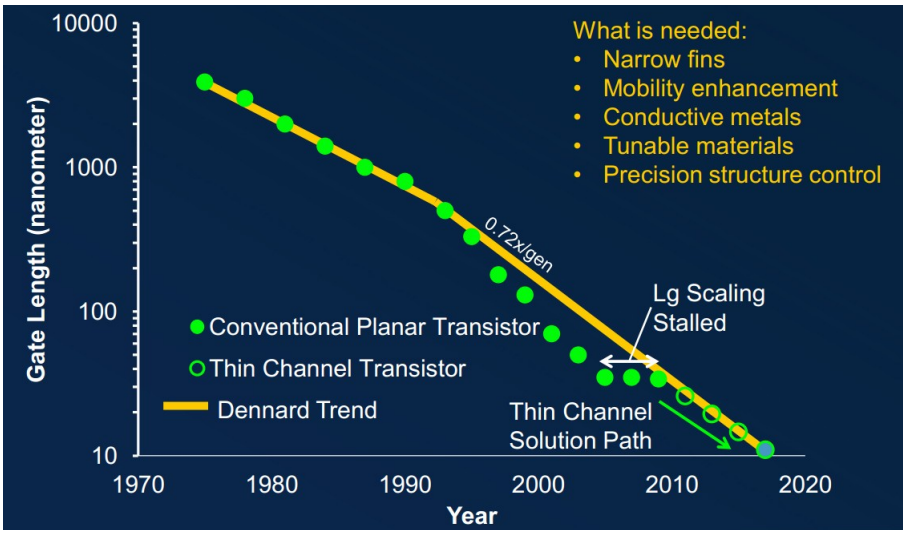


Figure 1.8: CMOS technology scaling (source: [11]).

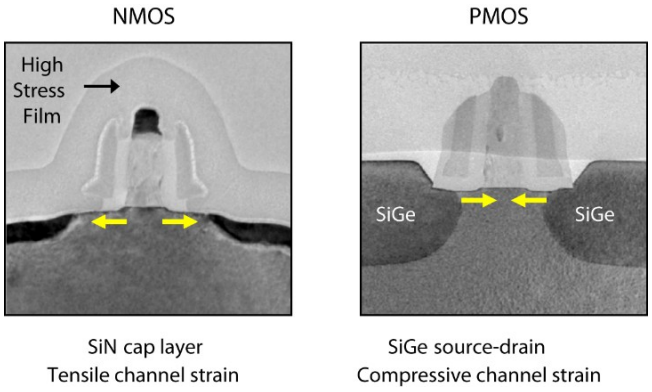


Figure 1.9: 90 nm Strained silicon transistors (source: [12]).

The reduction of transistor gate length is plotted versus the year in Figure 1.8. New transistor architectures with a thin silicon channel were introduced to enable further reduction of the gate length. Many technology innovations had to be introduced to continue the transistor scaling [13]. Among these innovations, Figure 1.9 shows the cross-sections of NMOS and PMOS transistors from a 90

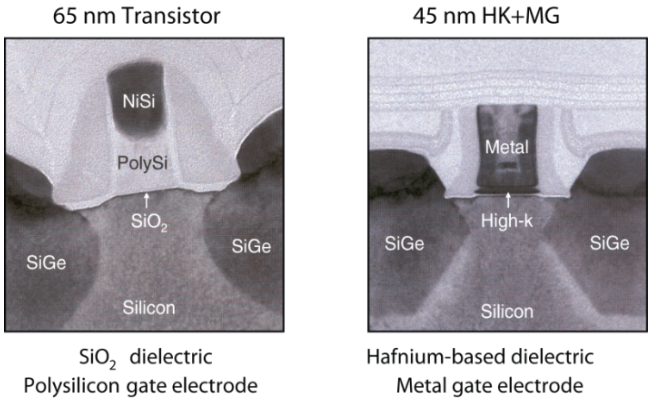


Figure 1.10: High-k and metal gate transistors (source: [12]).

nm technology strained silicon has been introduced to boost the mobility [12]. The left picture of Figure 1.10 shows the cross-sections of a 65 nm transistor with SiGe source and drain diffusions and NiSi gate stack. The right picture shows the cross-section of a 45 nm transistor showing the introduction of high-K dielectric combined with a metal gate [12].

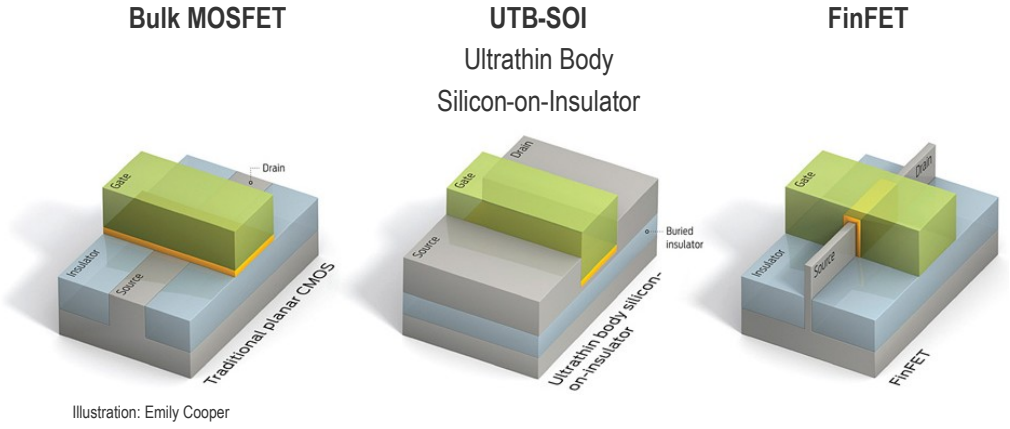


Figure 1.11: From bulk to UTB-SOI and FinFET (source: [8]).

In addition to these enhancements brought to bulk CMOS technologies, as shown in Figure 1.11, new device architectures were introduced replacing the bulk with an ultra-thin body silicon on insulator (UTB-SOI) and finally the FinFET where the thin silicon layer is rotated vertically and the gate is surrounding the fin on three sides offering much better electrostatic control on the channel region [8].

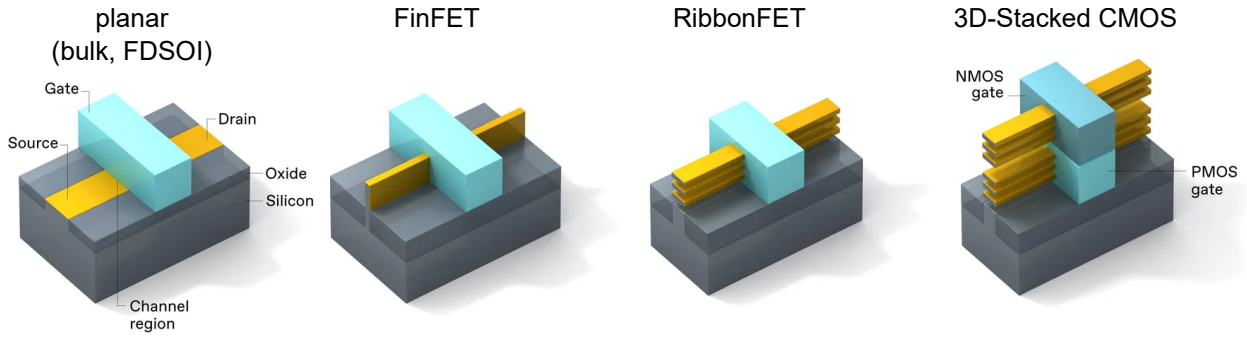


Figure 1.12: The transistor evolution, from planar to 3D-stacked CMOS (source: [14]).

To further increase the density of gates, the only way is to explore the 3rd-dimension and move from planar technologies (like in bulk and FDSOI) to FinFET as shown in Figure 1.11. Current research is now investigating how to stack several devices on top of each other to form the RibbonFET as shown

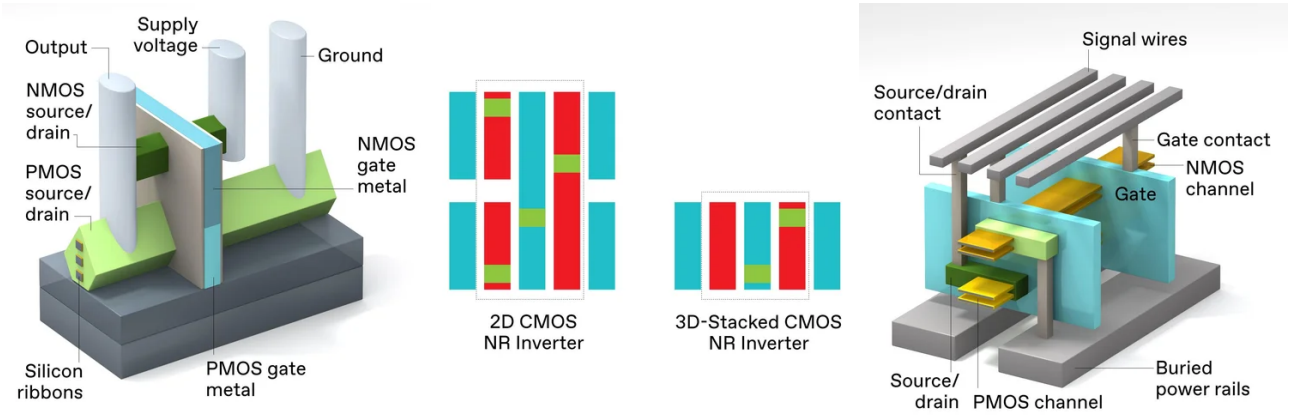


Figure 1.13: 3D stacked inverter (source: [14]).

in Figure 1.12. Stacking an NMOS on top of a PMOS as shown in Figure 1.12 will further increase the density. As shown in Figure 1.13, this introduces serious challenges in terms of interconnection.

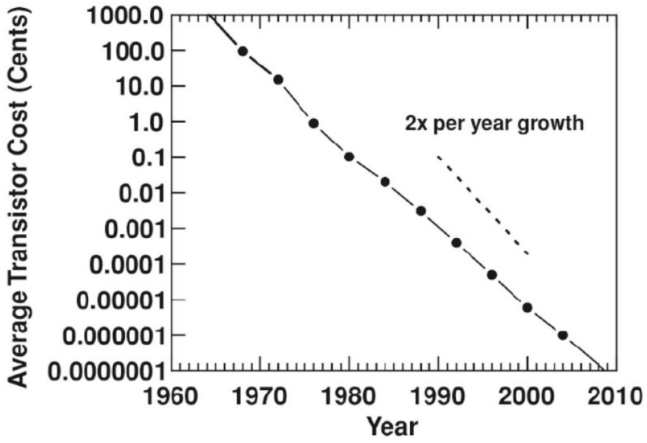


Figure 1.14: It's all about economics (source: Intel Corporation).

Of course, the main motivation behind Moore's law is mostly economic. Indeed, as shown in Figure 1.14, shrinking the size of a transistor has enabled to reduce its unit price steadily over the years by a factor of 10^9 in less than 5 decades! However the capital investment has become so high that less and less global players could afford investing in the development of new technologies. As shown in Figure 1.15, the number of global semiconductor players active in the development of new processes has shrunk as we have headed towards the 16/14-nm technology, leaving only 4 major companies with advanced fabs [1].

The technologies have also become much more complex leading to an exponential increase of the design cost as illustrated in Figure 1.16 [1]. A bigger part of this cost is actually taken for verification since the probability of making an error increases due to the complexity leading to prohibitive cost of a silicon re-spin. This has reduced the number of applications that can justify a large production volume and afford the latest technology node.

1.3 Energy efficiency and power consumption

1.3.1 Energy efficiency of computing

In this section we will have a broader look at the energy efficiency of computing. Figure 1.17 shows the evolution of the computation power and cost (measured as calculations per second and per 1'000

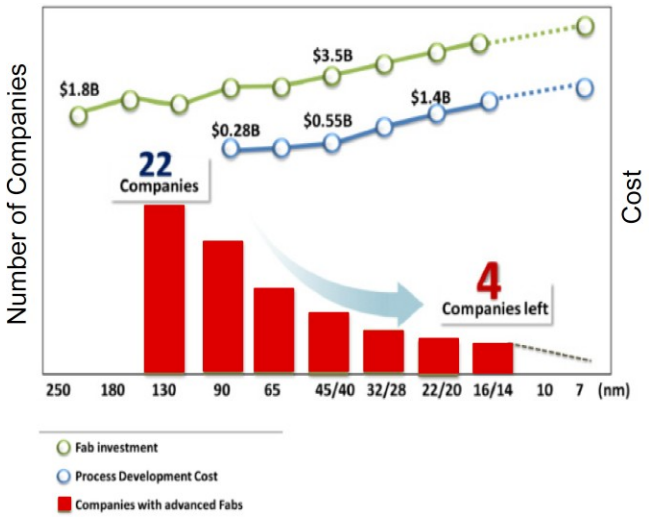


Figure 1.15: Smaller number of players for leading edge nodes [1] (source: Samsung foundry data).

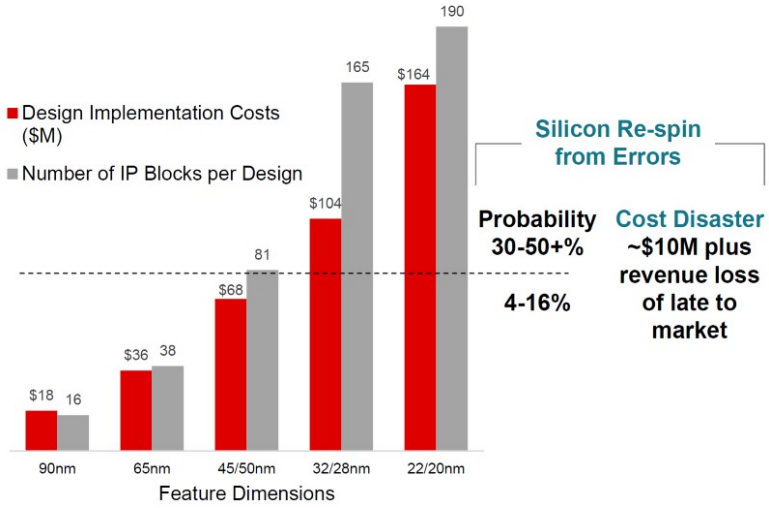


Figure 1.16: Rising design cost and complexity [1] (source: International business strategies).

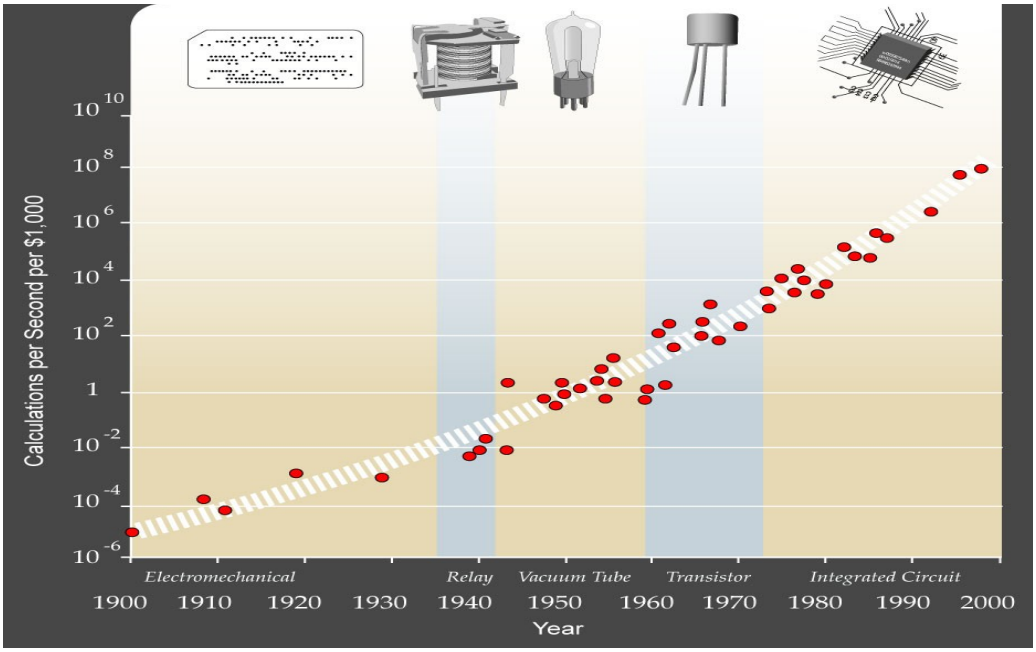
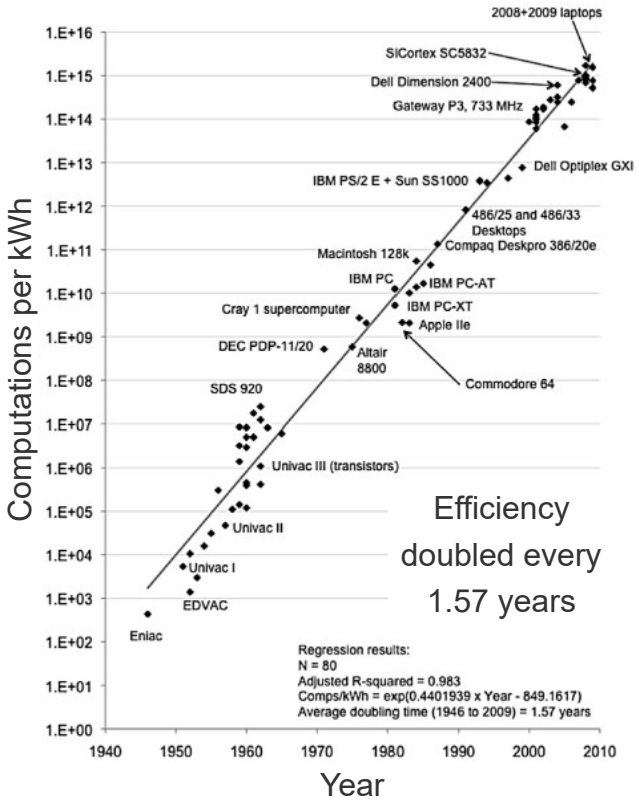
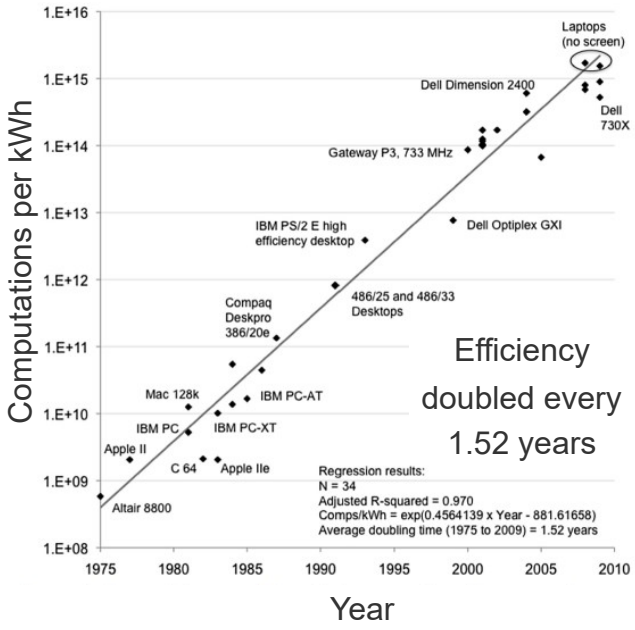


Figure 1.17: Computation power evolution [15] (source: Moore's law).

USD) versus the years [15]. The computation actually started as early as 1900 with electromechanical machines which were obviously very inefficient, bulky and costly. Things improved with the invention of the transistor and then the integrated circuit taking advantage of Moore's law [15]. The plot shows that, thanks to all these innovations, the computation efficiency could be improved by 13 orders of magnitude [15] !



(a) For all computing devices starting in 1946 with ENIAC (source: [16]).



(b) For personal computers only (source: [16]).

Figure 1.18: The Koomey law for computation efficiency [16].

Figure 1.18a shows the computation efficiency measured in computations per kWh which has followed an exponential growth with an average doubling time of about 1.57 years from 1946 (ENIAC) to 2009 [16]. Figure 1.18b shows the computation efficiency for personal computers alone which has doubled every 1.52 years from 1975 to 2009 [16].

As shown in Figure 1.19, it was already identified that to further increase the computing power efficiency would require to exploit the 3rd-dimension. It was proposed in [17] to use the combination of Silicon on Thin-buried-Oxide (SOTB which is a variation of the fully-depleted (FD) SOI MOSFET with thin buried oxide layer), NV memory and 3D stacking [17].

The computing efficiency has also increased tremendously for DSP as shown in Figure 1.20 which plots the power per million of MAC operations versus the years [18]. This law is called Gene's law after the first name of the author of the paper [18] who was part of the team that brought the first mainstream use of digital signal processing technology to the masses [19].

In 1988 Landauer published a paper exploring the ultimate limits to the switching energy dissipation in logic devices used in computing systems. The data plotted in Figure 1.21a was collected over many years by his IBM colleague Robert Keyes [20]. The data is starting from 1940 through the 1980s, a period that includes the replacement of vacuum tubes by bipolar transistors, the invention of the integrated circuit, and the early stages of the replacement of bipolar transistors by field-effect transistors (FETs). We see that the typical energy dissipated in a digital switching event dropped exponentially by over 10 orders of magnitude. Landauer extrapolated the data which should have reached the order

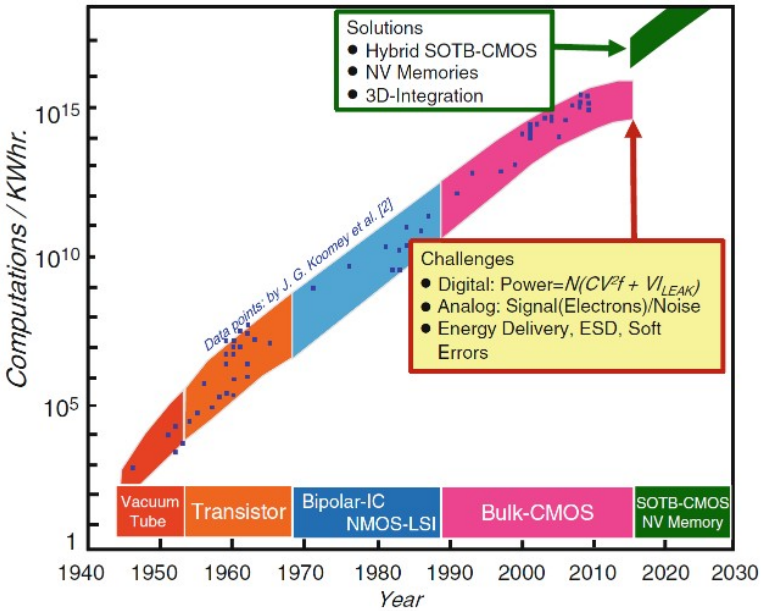


Figure 1.19: Extending the Koomey law by 3D integration (source: [17]).

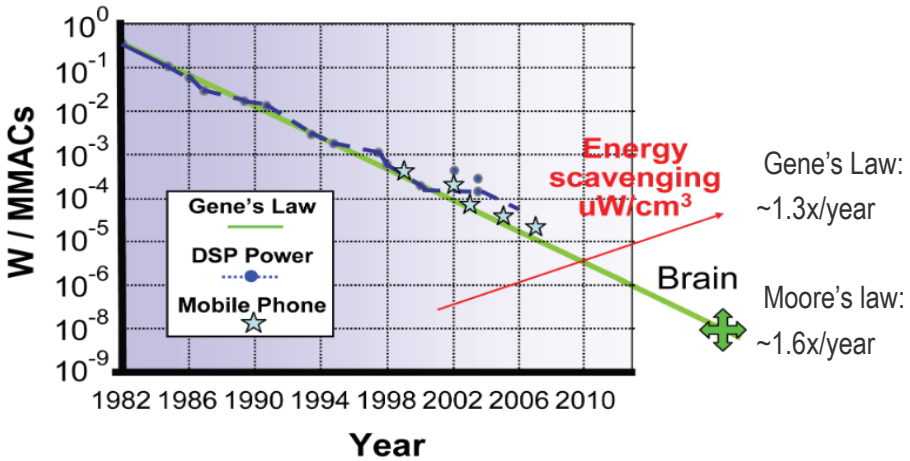
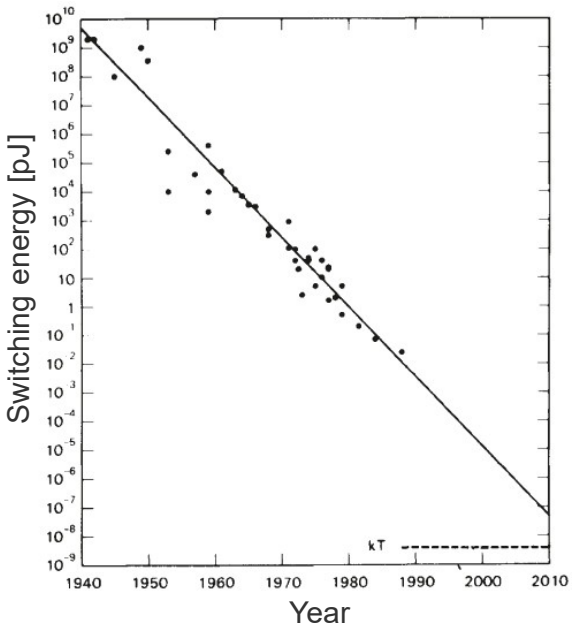
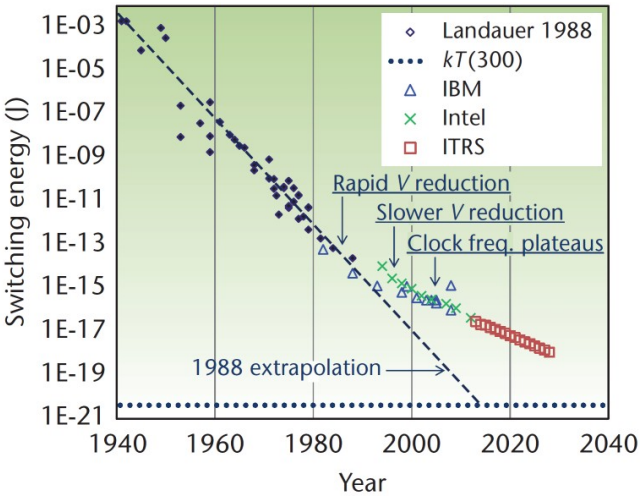


Figure 1.20: Gene's law (source: [18]).



(a) Switching energy from (source: [20]).



(b) Switching energy from (source: [21]).

Figure 1.21: Switching energy [20] [21].

of the thermal fluctuation energy, $k_B\,T$, evaluated at $T = 300\ K$ around 2015. Obviously this limit was not reached. There are many practical reasons for which this limit has not been reached, some of which are discussed by Theis *et al.* in [21]. In this paper they added some data collected from IBM, Intel and ITRS and added it to the original plot of Landauer and Keyes as shown in Figure 1.21b [21]. Despite the formidable performance improvements brought by the down-scaling of technology, the data clearly deviates from the asymptote as shown in Figure 1.21b [21].

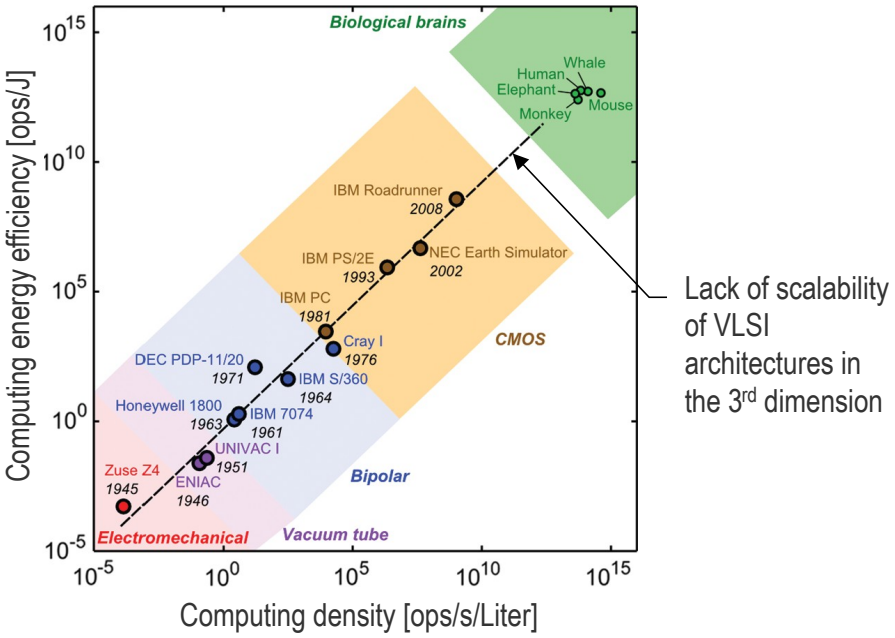


Figure 1.22: Computing efficiency versus computing density (source: [22]).

Another way to look at the improvement of computing energy efficiency is to also include the integration density as proposed by Ruch in [22]. In this paper Ruch plots the computing efficiency measured in operations per Joule versus the computing density measured in operations per second and per volume [22]. Both metrics have improved over many orders of magnitudes as shown in Figure 1.22

[22]. Improved computational efficiency was enabled thanks to new device technology and scaling. As shown in Figure 1.22, ten orders of magnitude efficiency improvement have been achieved, but it is unclear how this will continue upwards trying to reach the efficiency of biological systems (green region) [22]. It is striking that six different technologies (i.e., electromechanical switches, vacuum tubes, discrete transistor, transistor logic, emitter-coupled logic ICs, very large scale integrated (VLSI) CMOS, and biological brains) all fall onto one line on the log–log plot of operations per joule versus operations per second per liter, and changes in device technology did not lead to discontinuities. For microprocessor-based computers, this historic evolution is attributable to device shrinkage, which was motivated by a reduced cost per transistor and resulted in improved performance and power use per transistor. The main driver for efficiency improvement, therefore, has been integration density. Another way of expressing this trend is by stating that more efficient systems can be built more compact.

In all technologies, the volume fraction occupied by the devices was less than 0.1%. This ratio now has become extremely small with the current transistor generation just occupying 1 ppm of computer volume. Further CMOS device shrinkage is challenging due to the rising passive power fraction. Device technology needs to be efficient and small enough with low leakage currents to support integration density. However, the computational efficiencies and functional densities of VLSI architectures lag by orders of magnitude that of the human brain due to the lack of scalability of current VLSI architecture in the third dimension [22].

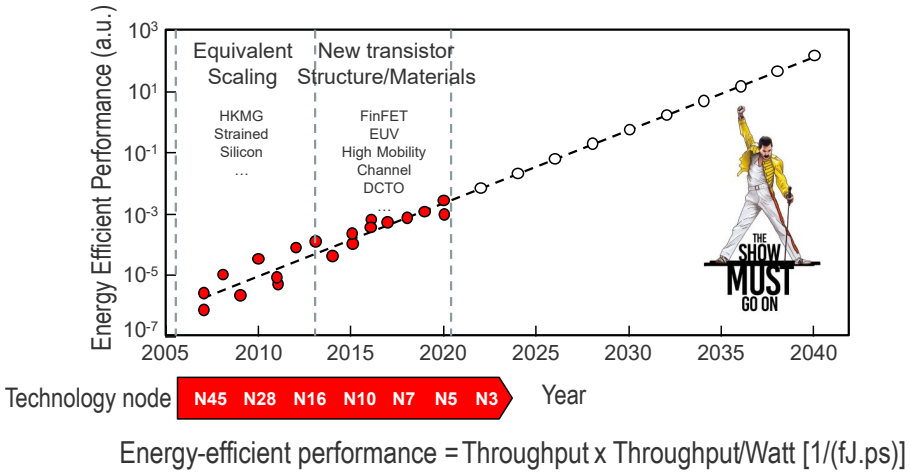


Figure 1.23: Energy efficient performance (source: [3]).

Figure 1.23 shows that the energy-efficient performance of real computing systems has improved at roughly a rate of 2× every 2 years over the last decade and a half [3]. This metrics is defined by Liu as the throughput times the throughput per watt measured in 1/(fJ.ps). Liu believes that this trend shall continue well into the future [3]. This energy-efficient performance is gained through architecture and design innovations in conjunction with, and enabled by, advancements in the underpinning transistor and integration technologies. Computation throughput is directly correlated with the number of transistors, while continued 2D scaling, design-technology co-optimization (DTCO), and 3D device integration will continue the trend of increased transistor count. Computation energy efficiency (throughput per Watt) will be improved by $C \cdot V^2$ reduction, where C is the switching capacitance of the transistors and the wires, and V is the power supply voltage. Wire capacitance can be reduced by further 2D scaling and 3D integration. The power-supply voltage may be reduced by improving transistor electrostatics and carrier transport that provides high on-state current while maintaining low off-state leakage. Domain-specific technology (DST) used in conjunction with domain-specific architecture (DSA) provides further opportunity for advancing computation energy efficiency [3].

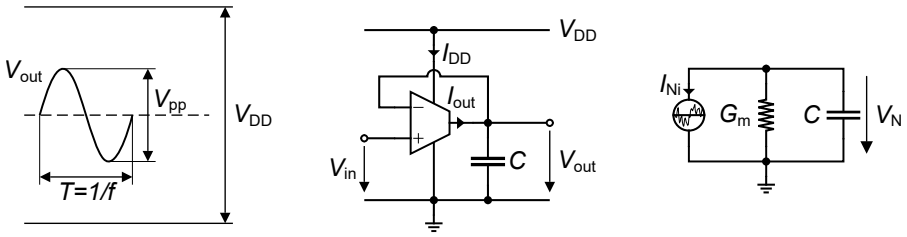


Figure 1.24: Lower limit power consumption.

1.3.2 Power Consumption of Analog Signal Processing

1.3.2.1 Fundamental limit

Most of the above trends in energy efficiency are based on digital systems and mostly computers. Things are quite different when looking at the power consumption of analog circuits [23] [24] [25] [26]. To get a better understanding of what dictates the power consumption of analog circuits, let's calculate the minimum power that is required to implement a 1st-order low-pass filter (or the power required to implement a single pole). This can be done as shown by the schematic shown in Figure 1.24 using a differential transconductor having a transconductance G_m loaded by a capacitor of capacitance C. It is easy to show that the circuit shown in Figure 1.24 implements a 1st-order low-pass filter with a cut-off frequency equal to G_m/C and a unity DC gain. The transconductor is considered as ideal, meaning that it is linear, can operate rail-to-rail with a 100% efficiency (meaning all the current drawn from the supply is used to charge the load capacitor. Additionally, we will assume that the differential transconductor has no offset and is only subject to thermal noise (no 1/f noise). Of course all the above assumptions are strong and quite far from a real transconductor. However it will give us as estimation of what is the ultimate minimum power that is needed for implementing a single pole with a certain signal-to-noise (SNR).

The average value of the output current that is needed to charge and discharge the load capacitor, assuming a sinusoidal input voltage of frequency f and having a peak-to-peak voltage V_{pp}, is given by

$$\overline{I_{out}} = f \cdot C \cdot V_{pp}. \tag{1.1}$$

The average power consumption is then simply derived as

$$P = V_{DD} \cdot \overline{I_{out}} = V_{DD} \cdot f \cdot C \cdot V_{pp} = \frac{V_{DD}}{V_{pp}} \cdot f \cdot C \cdot V_{pp}^2. \tag{1.2}$$

As shown in Figure 1.24, the noise of the OTA can be modelled by a current source at its output having a power spectral density (PSD)

$$S_{Ni} = 4k_B\,T \cdot \gamma_n\,G_m, \tag{1.3}$$

where γ_n is the OTA thermal noise excess factor which will be assumed to be unity. Because the transconductance G_m is setting at the same time the level of the noise PSD and the bandwidth, the total noise power on capacitor C is independent of G_m and given by

$$V_N^2 = \frac{\gamma_n\,k_B\,T}{C}. \tag{1.4}$$

The SNR is then defined as

$$SNR = \frac{V_{pp}^2/8}{\gamma_n\,k_B\,T/C} \tag{1.5}$$

from which we get the signal peak-to-peak voltage required for reaching a certain SNR

$$V_{pp}^2 = 8\,\frac{\gamma_n\,k_B\,T}{C} \cdot SNR \cong \frac{8\,k_B\,T}{C} \cdot SNR. \tag{1.6}$$

Substituting (1.6) into (1.2) leads to

$$P = 8\frac{V_{DD}}{V_{pp}} \cdot k_B T \cdot f \cdot SNR. \tag{1.7}$$

From (1.7), we see that the power can be minimized by maximizing the peak-to-peak signal with rail-to-rail operation $V_{pp} = V_{DD}$, resulting in

$$P_{min} = 8k_B T \cdot f \cdot SNR \tag{1.8}$$

Equations (1.7) and (1.8) show that the power P is proportional to the required SNR and to the frequency f, which actually corresponds to the bandwidth B of the low-pass filter. Equation (1.8) is plotted versus the SNR in Figure 1.25 by the red line. As a result of this linear relation, increasing the requirement on SNR by 10 dB results in a ten-fold increase of the minimum necessary power consumption.

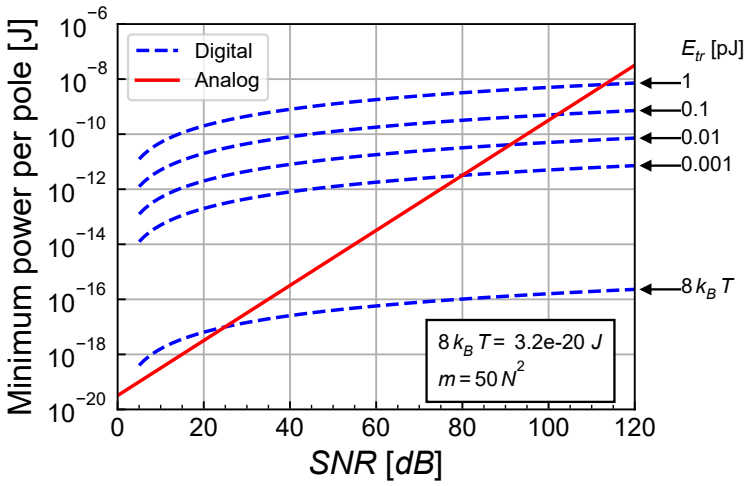


Figure 1.25: Minimum power consumption versus SNR [23] [24] [25] [26].

The minimum power for an analog system can be compared to that of a digital system, in which each elementary operation requires a certain number m of binary gate transition cycles, each of which dissipates an amount of energy E_{tr}. The minimum power is then simply given by $P_{min} = m \cdot f \cdot E_{tr}$. The number m of transitions is only proportional to some power a of the number of bits N, and therefore power consumption is only weakly dependent on SNR (essentially logarithmically)

$$m \cong N^a [\log(SNR)]^a. \tag{1.9}$$

The comparison with analog is then obtained by estimating the number of gate transitions that are required to compute each period of the signal, which for a single pole digital filter can be estimated to be approximately

$$m \cong 50 \cdot N^2. \tag{1.10}$$

Immunity to thermal noise imposes an absolute minimum energy per transition $E_{tr,min}$ estimated by Keyes to $8k_B T$, which provides the absolute minimum power limit [27] [28].

Note

Note that the minimum energy dissipation is directly linked to the error probability which for $E_{tr,min} = 8k_B T$ would be roughly 1.8% as shown in [29]. A 10^{-3} probability error would result in a minimum energy of about $14k_B T \cong 55 \times 10^{-21} J$ at room temperature.

However, in practice $E_{tr} = C \cdot V_{DD}^2$ is forced to a much higher value (10^{-15} to 10^{-12} Joules) by the need to recharge the equivalent capacitance C of each gate to the supply voltage V_{DD}. As shown in Figure 1.25, the minimum power for digital is therefore much higher than the absolute limit at room temperature. The minimum gate capacitance is strongly dependent on the process feature size and the supply voltage is imposed by the need to achieve the required delay time and by established standards. Furthermore, if the activation rate of the circuit is very low (very small percentage of the available gates in transition on average), then the leakage current of each of the gates may contribute to a non negligible additional static power consumption.

The comparison of analog and digital minimum power consumption plotted in Figure 1.25 clearly show that analog systems may consume much less power than their digital counterpart, provided a small SNR is acceptable. But for systems requiring large SNR, analog becomes very power inefficient. It is worth mentioning that a comparison of chip area basically leads to the same qualitative conclusion.

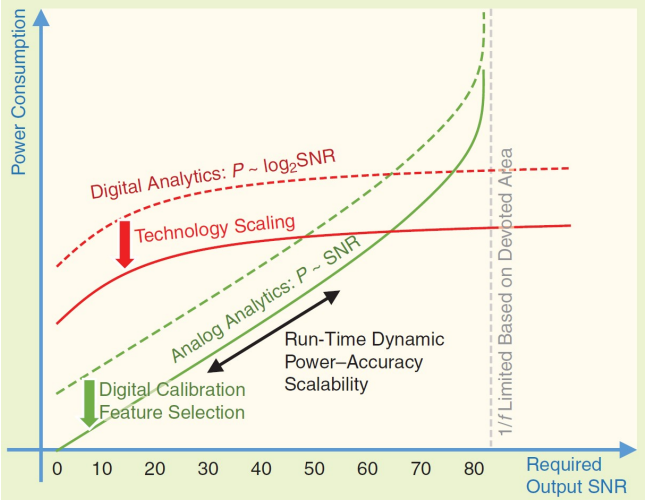


Figure 1.26: Minimum power consumption versus SNR (source: [26]).

The limits discussed so far are fundamental since they do not depend on the technology nor on the choice of power supply voltage. However, a number of obstacles or technological limitations are on the way to approach these limits in practical circuits [24]. One of them is the presence of additional noise sources such as 1/f noise in the devices. Another is the requirements on offset voltage due to device mismatch leading to a constraint on the device area. Both of them will make the power consumption increase significantly at larger SNR as shown in Figure 1.26 [23] [24] [25] [26]. Other practical limitations include:

- The above derivation assumed a 100% current efficient (non-ideal class B operation), linear and rail-to-rail transconductor.
- In addition the transconductor was also assumed to be perfectly linear, even with a rail-to-rail operation.
- Capacitors increase the power necessary to achieve a given bandwidth. They are only acceptable if their presence reduces the noise power by the same amount (by reducing the noise bandwidth). Therefore, ill-placed parasitic capacitors very often increase power consumption.
- The power spent in bias circuitry is wasted and should in principle be minimized. However, inadequate bias schemes may increase the noise and therefore require a proportional increase in power. For example, a bias current is more noisy if it is obtained by multiplying a smaller current.
- According to (1.2), power is increased if the signal at any node corresponding to a functional pole (pole within the bandwidth, or state variable) has a peak-to-peak voltage amplitude smaller than the supply voltage V_{DD}. Thus, care must be taken to amplify the signal as early as possible to its maximum possible voltage value, and to maintain this level all along the processing path. Using current-mode circuits with limited voltage swings is therefore not a good approach to reduce

power, as long as the energy is supplied by a voltage source. It only becomes attractive if voltage companding techniques can be used [30].
- The presence of additional sources of noise implies an increase in power consumption. These include 1/f noise in the devices, and noise coming from the power supply or generated on chip by other blocks of the circuit.
- When capacitive loads are imposed (for example by parasitic capacitors), the current I necessary to obtain a given bandwidth is inversely proportional to the transconductance-to-current ratio G_mID{} of the active device. The small value of G_mID{} inherent to MOS transistors operated in strong inversion may therefore cause an increase in power consumption.
- The need for precision usually leads to the use of larger dimensions for active and passive components, with a resulting increase in parasitic capacitors and power.
- All switched capacitors must be clocked at a frequency higher than twice the signal frequency. The power consumed by the clock itself may be dominant in some applications.

Ways to reduce the effect of these various limitations can be found at all levels of analog design ranging from device to system.

Technology scaling has significantly reduced the switching energy which has reduced the power and moved the dashed curve down. On the other hand, analog circuits don't take direct advantage of technology scaling. However, as mentioned in [26], some class of analog circuits can benefit from digital calibration to reduce their power consumption. In this case, the SNR at which the analog (red curves) and digital (green curves) curves in Figure 1.26 cross does not shift too much in more advanced technologies [26].

Unlike digital circuits, where the dynamic power decreases with the square of the supply voltage, according to (1.2), reducing the supply voltage of analog circuits while preserving the same bandwidth B and SNR, has no fundamental effect on their minimum power consumption. However, this absolute limit was obtained by neglecting the possible limitation of bandwidth B due to the limited transconductance G_m of the active device. The maximum value of B is proportional to G_m/C. Replacing the capacitor value C by G_m/B in (1.5) and expressing the product of the SNR times the bandwidth B yields

$$SNR \cdot B = \frac{V_{pp}^2 \cdot G_m}{8 k_B T}. \tag{1.11}$$

In most cases, scaling the supply voltage V_{DD} by a factor κ requires a proportional reduction of the signal swing V_{pp}. Maintaining the bandwidth and the SNR is therefore only possible if the transconductance G_m is increased by a factor κ^2. If the active device is a bipolar transistor (or a MOS transistor biased in weak inversion), its transconductance can only be increased by increasing the bias current I by the same factor κ^2; power is therefore increased by κ. The situation is different if the active device is a MOS transistor biased in strong inversion. Its transconductance can be shown to be proportional to I/V_P, where V_P is the pinch-off or saturation voltage of the device. Since this saturation voltage has to be reduced proportionally with V_{DD}, increasing G_m by κ^2 only requires an increase of current I by a factor κ and hence the power remains unchanged.

However, the maximum frequency of operation may be affected by the value of the supply voltage. For a MOS transistor in strong inversion, the transit frequency f_t for which the current gain falls to unity is approximately given by (assuming no velocity saturation)

$$f_t \cong \frac{\mu \cdot V_P}{L^2} \tag{1.12}$$

When the supply voltage is scaled by a factor κ (meaning that $V_{DD} \to V_{DD}/\kappa$) due to the scaling of the process, the length is also scaled by the same factor κ ($L \to L/\kappa$) and the transit frequency is therefore increased by κ.

Low-voltage limitations are not restricted to power or frequency problems. Reducing V_P increases the transconductance-to-current ratio of MOS transistors which in turn increases the noise content of

current sources and drastically degrades their precision. Conductance in analog switches is difficult to ensure when the supply voltage falls below approximately the sum of the p- and n-channel transistor threshold voltages. For a given value of time constant, charge injection in a switch does not depend on V_{DD} in absolute value, but it increases in relative value if V_{DD} and V_{pp} are decreased. The same is true for any constant voltage overhead such as the base-emitter voltage in bipolar transistors or the threshold voltage in MOS transistors.

A factor of merit (actually demerit) can be defined to evaluate how far a given circuit is from the ideal case [31]

$$K \triangleq \frac{P}{k_B T \cdot B \cdot SNR} = 8\frac{V_{DD}}{V_{pp}} \tag{1.13}$$

which is minimum for a rail-to-rail operation ($V_{pp} = V_{DD}$)

$$K_{min} = K|_{V_{pp}=V_{DD}} = 8 \tag{1.14}$$

As mentioned above, K_{min} constitute an absolute minimum not accounting for many non-idealities that can seriously degrade (increase) the factor K far beyond K_{min} in practical analog circuits.

1.3.2.2 Minimum power of a transconductance amplifier

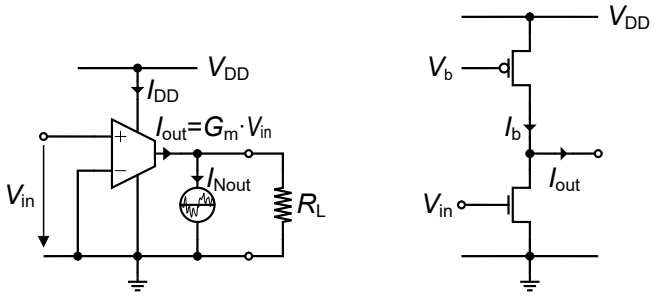


Figure 1.27: Transonductance amplifier.

In analog circuits, signals are frequently converted from voltage to current and vice versa, in order to best exploit the respective features of theses two modes of representation. Such conversions are carried out by transconductors having a transconductance G_m as illustrated in the left figure of Figure 1.27. The output signal and thermal noise power are given by

$$S = (G_m \cdot V_{in,rms})^2, \tag{1.15}$$

$$N = 4k_B T \cdot \gamma_{neq} \cdot G_m \cdot B_n, \tag{1.16}$$

where B_n is the noise bandwidth which is larger than the actual signal bandwidth $B = G_m/C$ and $\gamma_{neq} = G_m \cdot R_{neq}$ is the equivalent thermal noise excess factor with R_{neq} being the equivalent input-referred thermal noise resistance. The SNR is then given by

$$SNR = \frac{G_m \cdot V_{in,rms}^2}{4k_B T \cdot \gamma_{neq} \cdot B_n}. \tag{1.17}$$

The power consumption is simply given by

$$P = V_{DD} \cdot I_{DD} \tag{1.18}$$

We can now calculate the K factor as [31]

$$K \triangleq \frac{P}{k_B T \cdot B \cdot SNR} \geq 4\gamma_{neq}\frac{V_{DD}}{V_{in,rms}^2} \cdot \frac{I_{DD}}{G_m}. \tag{1.19}$$

From (1.19), we see that the K factor can be minimized by maximizing the G_m/I_{DD} and $V_{in,rms}/V_{DD}$ ratios. Maximizing the G_m/I_{DD} is achieved by biasing the transconductor (differential pair) in weak inversion and maximizing the $V_{in,rms}/V_{DD}$ ratio is obtained by having a rail-to-rail operation. However, maximizing the transconductor G_m/I_{DD} reduces at the same time its linear range and the maximum input voltage for a given distortion.

The simplest example of a transconductor is shown in the right schematic of Figure 1.27 consisting of a single NMOS common source transistor biased by a PMOS current source. The NMOS and PMOS are biased in strong inversion for better linearity for the NMOS transconductor and better matching for the PMOS. The transconductances of the NMOS and PMOS devices are then related to the bias current I_b{} and their saturation voltages V_{DSsatn} and V_{DSsatp} according to

$$G_{mn} = \frac{2I_b}{nV_{DSsatn}}, \tag{1.20}$$

$$G_{mp} = \frac{2I_b}{nV_{DSsatp}}, \tag{1.21}$$

where we have assumed that both transistors have the same slope factor n. The output thermal noise current PSD is given by

$$S_{nout} = 4k_B\,T \cdot G_{nout}, \tag{1.22}$$

where

$$G_{nout} = \gamma_n \cdot G_{mn} + \gamma_p \cdot G_{mp} = \gamma G_m \cdot \left(1 + \frac{G_{mp}}{G_{mn}}\right) = \gamma G_m \cdot \left(1 + \frac{V_{DSsatn}}{V_{DSsatp}}\right), \tag{1.23}$$

where $\gamma = \gamma_n = \gamma_p = 2n/3$ and $G_m = G_{mn}$ is the transconductor transconductance. The input noise resistance is given by

$$R_{neq} = \frac{G_{nout}}{G_m^2} = \frac{\gamma}{G_m} \cdot \left(1 + \frac{V_{DSsatn}}{V_{DSsatp}}\right). \tag{1.24}$$

The equivalent noise excess factor then reads

$$\gamma_{neq} \triangleq G_m \cdot R_{neq} = \gamma \cdot \left(1 + \frac{V_{DSsatn}}{V_{DSsatp}}\right). \tag{1.25}$$

The supply voltage needs to be larger than $V_{DSsatn} + V_{DSsatp}$ in order to maintain both transistors saturated, assuming that the output voltage remains constant

$$V_{DD} > V_{DSsatn} + V_{DSsatp}. \tag{1.26}$$

The power is then given by

$$P = I_b \cdot V_{DD} > I_b \cdot (V_{DSsatn} + V_{DSsatp}). \tag{1.27}$$

Finally the K-factor can be derived as [31]

$$\begin{aligned} K &= 4\gamma_{neq} \cdot \frac{V_{DD} \cdot I_b}{G_m \cdot V_{in,rms}^2} > 4\gamma_{neq} \cdot \frac{I_b \cdot (V_{DSsatn} + V_{DSsatp})}{G_m \cdot V_{in,rms}^2} \\ &= 2\gamma_{neq} \cdot n \cdot \left(\frac{V_{DSsatn}}{V_{in,rms}}\right)^2 \cdot \left(1 + \frac{V_{DSsatp}}{V_{DSsatn}}\right). \end{aligned} \tag{1.28}$$

Replacing γ_{neq} by (1.25) results in [31]

$$K > \frac{4n^2}{3} \cdot \left(1 + \frac{V_{DSsatn}}{V_{DSsatp}}\right) \cdot \left(1 + \frac{V_{DSsatp}}{V_{DSsatn}}\right) \cdot \left(\frac{V_{DSsatn}}{V_{in,rms}}\right)^2. \tag{1.29}$$

(1.29) is minimum for equal saturation voltages of the NMOS and PMOS $V_{DSsatn} = V_{DSsatp} = V_{DSsat}$ resulting in [31]

$$K > \frac{16n^2}{3} \cdot \left(\frac{V_{DSsat}}{V_{in,rms}}\right)^2. \tag{1.30}$$

The K-factor can be further educed by decreasing the the common saturation voltage enabling a lower supply voltage. However, decreasing V_{DSsat} unfortunately increases the total harmonic distortion THD which for a square-law characteristic is given by [31]

$$THD = \frac{V_{in}}{4nV_{DSsat}} = \frac{\sqrt{2}V_{in,rms}}{4nV_{DSsat}} \tag{1.31}$$

which can be used to express the ratio

$$\left(\frac{V_{DSsat}}{V_{in,rms}}\right)^2 = \frac{1}{8n^2THD^2} \tag{1.32}$$

leading to

$$K > \frac{\gamma}{n \cdot THD^2} = \frac{2}{3THD^2}. \tag{1.33}$$

The K-factor is therefore strongly degraded by the requirements on linearity since it is inversely proportional to the square of the THD. For example, a 1% THD requirement leads to K larger than 6700 which corresponds to K being 640 times larger than its minimum value $K_{min} = 8$ [31] !

1.3.2.3 Power consumption trends in ADCs

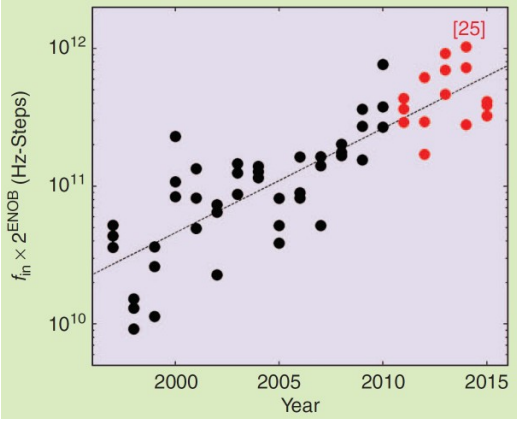


Figure 1.28: Evolution of ADCs speed-resolution product (source: [32]).

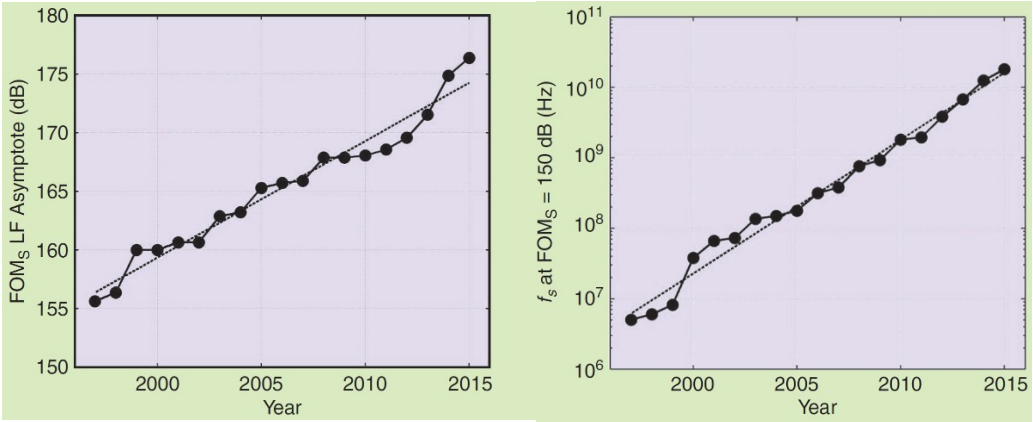


Figure 1.29: ADCs FoMs trends (source: [32]).

Figure 1.28 plots the speed times the resolution product versus the year which is approximately doubling every four years. This progress rate is smaller than other figures such as the ADC energy efficiency.

The left figure in Figure 1.29 plots the figure-of-merit (FoM) proposed by Schreier

$$FoM_S \triangleq SNDR(dB) + 10\log\left(\frac{f_s/2}{P}\right) \tag{1.34}$$

versus the year for ADCs operating at rather low conversion rate. The linear fit corresponds to an improvement of 1dB per year [32]. The right figure in Figure 1.29 plots the high frequency asymptote of FoM_S. The linear fit corresponds to a doubling of the conversion rate for every 1.8 years [32].

1.4 Voltage Scaling

1.4.1 Voltage scaling roadmap

There are mainly two main motivations for the voltage to be decreased as CMOS technology has scaled down over the years. The principal reason was related to the dynamic power consumption of digital circuits which is proportional to the square of the supply voltage

$$P_{dyn} = f \cdot C \cdot V_{DD}^2. \tag{1.35}$$

So there is obviously a great incentive to decrease the supply voltage in order to reduce the dynamic power consumption.

Another reason is related to the electric field within the MOSFET. Indeed, the ideal constant field scaling would require the voltage to be scaled down proportionally to the same scaling factor κ used for scaling the transistor dimensions (mostly the longitudinal direction along the channel) [9].

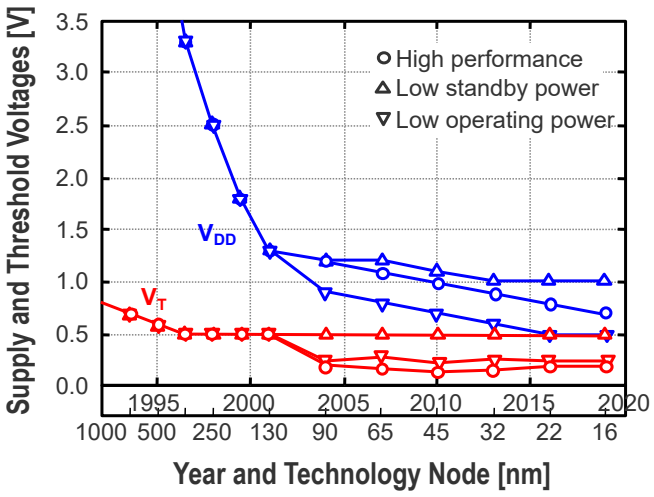


Figure 1.30: Supply voltage scaling according to the 2005 ITRS roadmap [33] [34].

The scaling of the supply voltage for digital circuits (logic) is illustrated in Figure 1.30, Figure 1.31 and Figure 1.32. Figure 1.30 is taken from the 2005 edition of the International Technology Roadmap for Semiconductors (ITRS) [33] and [34]. It covers the period 1997 to 2020 corresponding to the CMOS technology nodes from 0.5-μm down to 16-nm. From Figure 1.30, we see that the supply voltage is scaled down starting at 3.3 V and reaching 0.5 V in 2020. At that time the ITRS envisioned three different supply scaling trend according to the target application: high performance ($\circ$ symbol), low standby power ($\triangle$ symbol) and low operating power ($\triangledown$ symbol) [34]. High-performance logic refers to chips of high complexity, high performance, and high power dissipation, such as microprocessor unit (MPU) chips for desktop PCs, servers, etc. Low-power logic refers to chips for mobile systems, where the allowable power dissipation and hence the allowable leakage currents are limited by battery life [34]. There are two major categories within low-power, *low operating power* and *low standby power* logic [34]. Low operating power chips are typically for relatively high-performance mobile applications,

such as notebook computers, where the battery is likely to be high capacity and the focus is on reduced operating (i.e., dynamic) power dissipation [34]. Low standby power chips are typically for lower performance, lower cost consumer type applications, such as internet-of-things (IoT) devices with lower battery capacity and an emphasis on the lowest possible static power dissipation, i.e., the lowest possible leakage current [34]. The transistors for high-performance ICs have both the highest performance and the highest leakage current of all, and hence the physical gate length and all the other transistor dimensions are most rapidly scaled for high-performance logic. The transistors for low operating power chips have somewhat lower performance and considerably lower leakage current, while the transistors for low standby power chips have both the lowest performance and the lowest leakage current of all. For low operating power logic, the gate length lags behind the high-performance transistor gate length by two years, reflecting historical trends and the need for low leakage current in mobile applications. For low standby power logic, the gate length lags that of high-performance logic by four years, reflecting the ultra-low leakage current required.

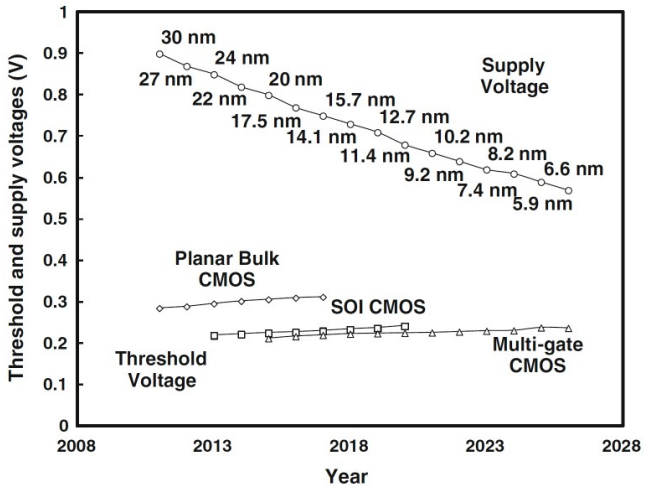


Figure 1.31: Supply voltage scaling according to the 2011 ITRS roadmap (source: [35]).

Figure 1.31 updates the earlier roadmap and extends them towards 2025 corresponding to the 6-nm technology node [35]. The supply voltage starts at 0.9 V for 30-nm and ends at 0.6 V for 6-nm. It also shows the threshold voltages for bulk, SOI (FDSOI) and multigate (FinFET) transistors which almost doesn't change over the years.

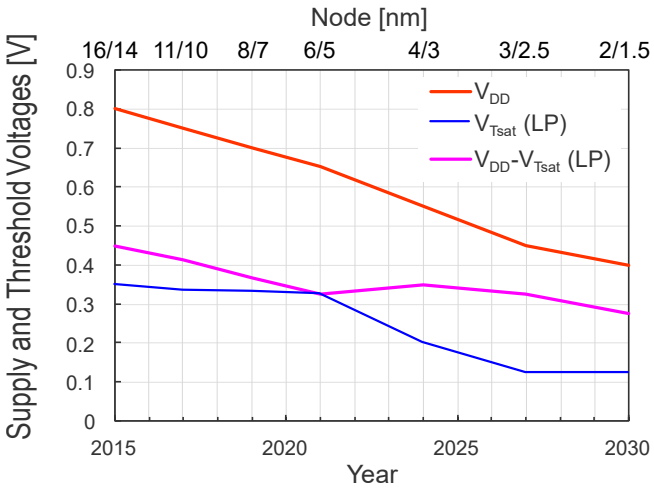


Figure 1.32: Supply voltage scaling according to the 2015 ITRS roadmap (source: [36]).

Finally, Figure 1.32 updates the earlier roadmaps and extends them towards 2030 corresponding to horizon of the 2-nm technology node [36]. From Figure 1.32, we see that the supply voltage continues to scale down reaching 0.4 V for the 2-nm node in 2030. The threshold voltage decreases slightly to

roughly 120 mV. For analog circuit both Figure 1.30 and Figure 1.32 show that the maximum available overdrive voltage $V_{DD} - V_T$ has been progressively reduced to reach typically $280\,mV$ in 2030. As already mention in Section??, this reduction leading to a decrease of the available overdrive voltage is actually pushing the operating points away from the traditional strong inversion region into the moderate and eventually even weak inversion region. As will be shown below, the reduction of the supply voltage does not help reducing the power consumption of analog circuits. Actually it may even lead to an increase of the power for realizing the same function and specification with a much lower supply voltage. With higher supply voltages it was easy to stack transistor using the same current branch and still having enough headroom for handling the signal. With a reduced supply voltage, stacking transistors biased in strong inversion becomes impossible and requires to duplicate the current branch. All what is gained for power by reducing the supply voltage is then lost by multiplying the current branches. The power might even increase because the additional current branches may require additional bias circuits.

1.4.2 Impact of voltage scaling for analog circuits

1.4.2.1 Impact of voltage scaling on power consumption

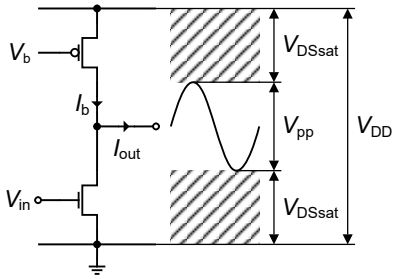


Figure 1.33: Common-source transconductance amplifier.

The analysis of the simple transconductance amplifier presented in has shown that the saturation voltages of the NMOS transconductor and PMOS current source of the simple transconductor circuit shown in Figure 1.33 should be taken equal to minimize the K-factor. Under this assumption, K can be further reduced by reducing V_{DSsat} and the supply voltage accordingly. However the saturation voltage cannot be reduced indefinitely. If it is assumed that it is incompressible, then the K-factor can be evaluated easily as follows. As shown in Figure 1.33, the minimum supply voltage for a given signal peak-to-peak voltage V_{pp} is

$$V_{DD} = V_{pp} + 2V_{DSsat}, \tag{1.36}$$

which allows to express $V_P = V_{DD} - 2V_{DSsat}$. Replacing V_{pp} in the power consumption expression leads to

$$P = K_{min} \cdot \frac{V_{DD}}{V_{pp}} \cdot k_B\,T \cdot B \cdot SNR = K_{min} \cdot \frac{V_{DD}}{V_{DD} - 2V_{DSsat}} \cdot k_B\,T \cdot B \cdot SNR \tag{1.37}$$

which allows to express the degradation of the K-factor with respect to its minimum value K_{min} as

$$\frac{P}{P_{min}} = \frac{K}{K_{min}} = \frac{V_{DD}}{V_{DD} - 2V_{DSsat}}. \tag{1.38}$$

Equation (1.38) is plotted versus V_{DD} in Figure 1.34 for $V_{DSsat} = 200\,mV$. We see that the K-factor is increasing as V_{DD} is reduced and starts to increase drastically when V_{DD} gets close to $2V_{DSsat}$. This is due to the fact that reducing V_{DD} without scaling further the saturations voltages V_{DSsat} forces to reduce the signal peak-to-peak voltage and hence the signal power. Maintaining the same SNR then requires to lower the noise by increasing the capacitance C forcing the transconductance to be increased in order to keep the same bandwidth until the peak-to-peak voltage gets to zero while the capacitance and the bandwidth tend to infinity.

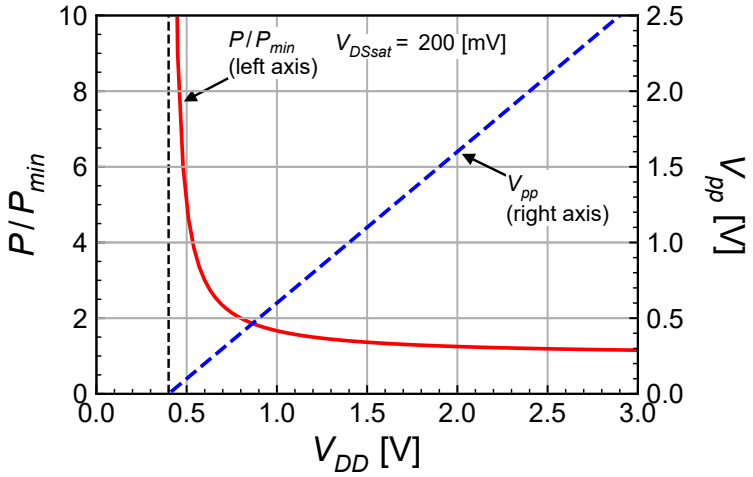


Figure 1.34: Power of common-source amplifier versus supply voltage.

1.4.2.2 Impact of voltage scaling on operating point

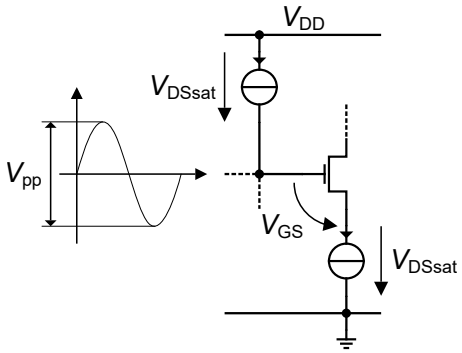


Figure 1.35: Impact of supply voltage scaling on voltage follower power consumption

The scaling of the supply voltage unavoidably pushes the operating point from the traditional strong inversion region towards subthreshold (i.e. moderate and weak inversion). This is illustrated by the simple circuit shown in Figure 1.35 which represents an NMOS voltage follower biased by a current source on the bottom and preceded by another stage with an additional PMOS current source. Assuming that the signal voltage at the gate of the voltage follower has a peak-to-peak voltage V_{pp}, the required supply voltage supply voltage is given by

$$V_{DD} = V_{GS} + 2V_{DSsat} + V_{pp}. \tag{1.39}$$

This circuit is a typical example where it might be useful to have an expression of the gate-to-source voltage in terms of the transistor inversion coefficient. In Section??, we will see that the gate-to-source voltage V_{GS} can be expressed as a function of the inversion coefficient IC according to

$$V_{GS}(IC) = V_{T0} + 2n\,U_T \ln\left(e^{\sqrt{IC}} - 1\right). \tag{1.40}$$

The supply voltage is plotted versus the inversion coefficient IC of the voltage follower transistor in Figure 1.36 for various threshold voltages ranging from 0.5 down to $0.2\,V$. We see that for a $0.4\,V$ threshold voltage, the transistor can no more be biased in strong inversion but only in the subthreshold region (i.e. $IC < 1$). For an even lower supply voltage of $0.5\,V$ and a threshold voltage of $0.3\,V$, the transistor is then confined into weak inversion! Under low supply voltage, it is therefore unavoidable for the designer to design analog circuits with most of the transistors biased in the moderate inversion.

This shift of the operating point towards subthreshold is also illustrated in Figure 1.37 which shows the normalized current (which is actually the inversion coefficient) versus the gate voltage for short-channel

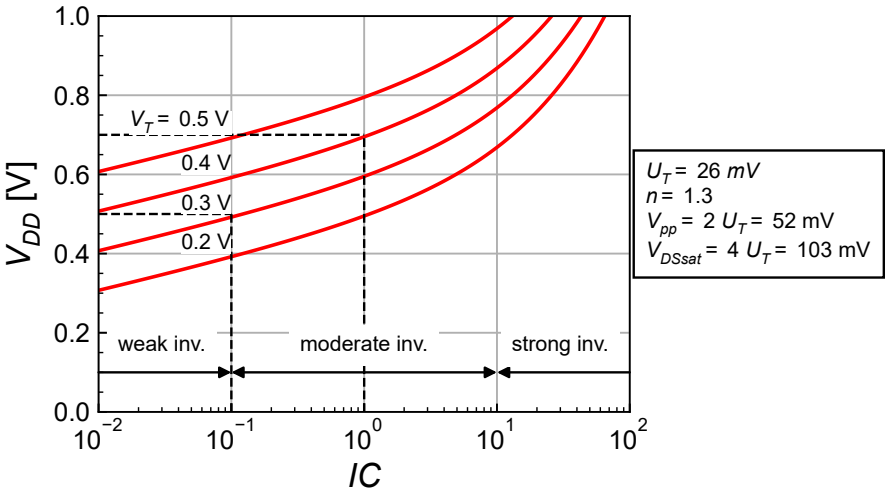


Figure 1.36: Shift of operating point towards weak inversion due to voltage scaling.

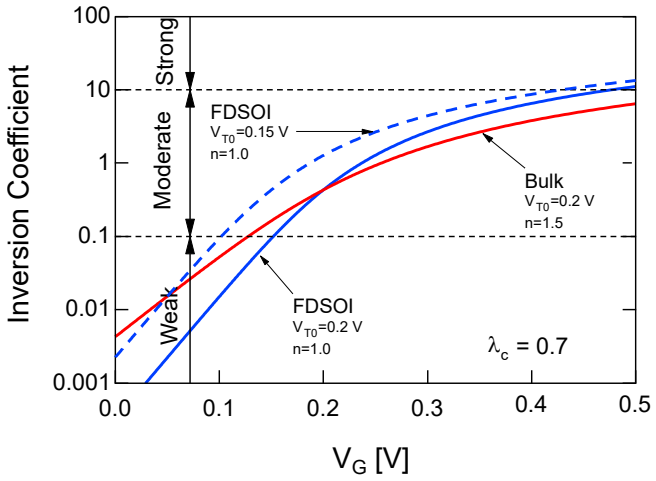


Figure 1.37: Strong inversion will disappear.

devices from a bulk and an FDSOI CMOS technologies. As we will see later, the inversion coefficient defines the level of inversion of the transistor: for $IC < 0.1$ the transistor operates in weak inversion, for $0.1 < IC < 10$ the transistor operates in moderate inversion, whereas for $10 < IC$ it operates in strong inversion. Looking first at the case of a short-channel transistor from a bulk technology with a slope factor $n = 1.5$ and a threshold voltage $V_{T0} = 200\,mV$, we get the red curve in Figure 1.37. Because of the low supply voltage and strong velocity saturation, the transistor does not reach strong inversion, even at the maximum gate voltage. Things improve a bit for a transistor from an FDSOI technology with $n = 1$ and with the same threshold voltage $V_{T0} = 200\,mV$ corresponding to the blue curve in Figure 1.37. This improvement is mostly due to the steeper slope ($n = 1$). However, note that the ideal slope factor of one is only obtained for non-minimum channel length. The advantage of FDSOI is that the threshold voltage can be changed and lowered to improve the current. If it is lowered enough to reach $V_{T0} = 150\,mV$, then the FDSOI transistor can finally reach strong inversion as shown by the dashed blue curve in Figure 1.37.

This shows that the strong inversion regime, which as been the preferred bias region for many years, will progressively disappear in more scaled technologies. The strong inversion model that is still taught to designers today can therefore no more be applied. We need a better simple model that also handles moderate and weak inversion consistently. Of course the designer can use the sophisticated compact models (CM) that are available in modern circuit simulators. Yet, most analog designers would still rely on their design intuition and do "hand calculations" before performing any simulation. This pre-simulation phase requires a stripped-down version of the CM which is simple enough for initial design guidance, yet accurate enough to minimize trial-and-error simulations.

1.4.2.3 Impact of voltage scaling on analog and RF performance

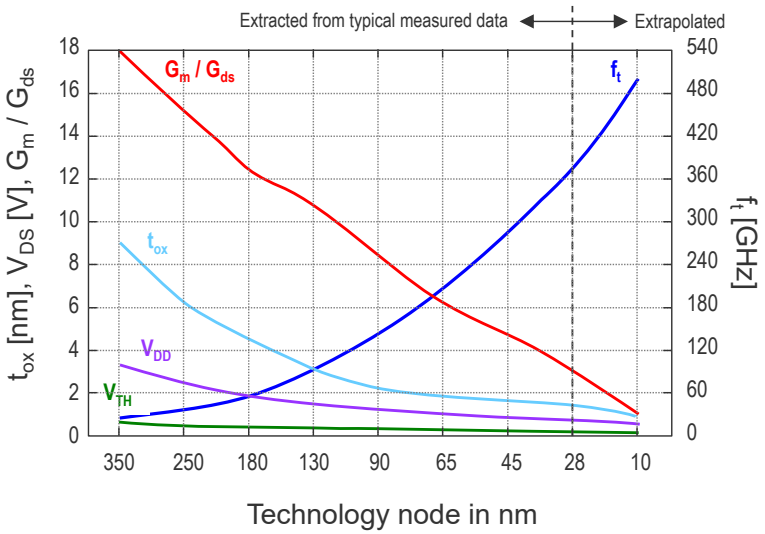


Figure 1.38: Impact of voltage scaling on analog and RF performance (source: [37]).

As shown in Figure 1.38, technology scaling brings a clear benefit to RF circuit design by increasing the transit frequency. However, it seriously degrades the self-gain G_m/G_{ds} [37]. This will degrade the performance of many analog circuits which ideal operation often requires a large voltage gain. However, the tremendous increase of transit frequency resulting from the reduction of the transistor gate length can be traded with voltage gain by using non-minimum gate length.

[1] A. Bahai, "Dynamics of exponentials in circuits and systems," in *2017 IEEE international solid-state circuits conference (ISSCC)*, 2017, pp. 14–20.

[2] A. Bahai, "Ultra-low energy systems: Analog to information," in *Proc. Of the european solid-state circ. Conf. (ESSCIRC)*, 2016, pp. 3–6.

[3] M. Liu, "Unleashing the Future of Innovation," in *2021 IEEE international solid- state circuits conference (ISSCC)*, 2021, pp. 9–16.

[4] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, 1965.

[5] K. Rupp, "42 Years of Microprocessor Trend Data." https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/, 2018.

[6] G. E. Moore, "No exponential is forever: but "Forever" can be delayed!" in *2003 IEEE international solid-state circuits conference, 2003. Digest of technical papers (ISSCC)*, 2003, pp. 20–23 vol.1.

[7] Moore, S. K., "Cerebras Unveils Its Next Waferscale AI Chip." https://spectrum.ieee.org/cerebras-chip-cs3, 2024.

[8] Ahmed, K. and K. Schuegraf, "Transistor Wars." https://spectrum.ieee.org/semiconductors/devices/transistor-wars, 2011.

[9] R. H. Dennard, F. H. Gaensslen, H. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, 1974.

[10] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and W. H.-S. Philip, "Device scaling limits of Si MOSFETs and their application dependencies," *Proceedings of the IEEE*, vol. 89, no. 3, pp. 259–288, 2001.

[11] Anthony, S., "7nm, 5nm, 3nm: The new materials and transistors that will take us to the limits of Moore's law." https://www.extremetech.com/computing/162376-7nm-5nm-3nm-the-new-materials-and-transistors-that-will-take-us-to-the-limits-of-moores-law, 2013.

[12] M. Bohr, "The new era of scaling in an SoC world," in *2009 IEEE international solid-state circuits conference - digest of technical papers*, 2009, pp. 23–28.

[13] Bohr, M., "Technology Leadership." https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/09/mark-bohr-on-intels-technology-leadership.pdf, 2017.

[14] M. Radosavljevic and J. Kavalieros, "Taking Moore's Law to New Heights: When transistors can't get any smaller, the only direction is up," *IEEE Spectrum*, vol. 59, no. 12, pp. 32–37, 2022, doi: 10.1109/MSPEC.2022.9976473.

[15] R. Kurzweil, *The Singularity Is Near*. Penguin Books, 2006.

[16] J. Koomey, S. Berard, M. Sanchez, and H. Wong, "Implications of Historical Trends in the Electrical Efficiency of Computing," *IEEE Annals of the History of Computing*, vol. 33, no. 3, pp. 46–54, 2011.

[17] T. Masuhara, "The Future of Low-Power Electronics," in *Chips 2020 vol. 2 - new vistas in nanoelectronics*, vol. 2, B. Hoefflinger, Ed., in The frontiers collection, vol. 2., Springer, 2016, pp. 21–50.

[18] G. Frantz, "Digital signal processor trends," *IEEE Micro*, vol. 20, no. 6, pp. 52–59, 2000.

[19] G. Frantz, "Signal Core: A Short History of the Digital Signal Processor," *IEEE Solid-State Circuits Magazine*, vol. 4, no. 2, pp. 16–20, 2012.

[20] R. W. Keyes, "Miniaturization of electronics and its limits," *IBM Journal of Research and Development*, vol. 32, no. 1, pp. 84–88, 1988.

[21] T. N. Theis and H. P. Wong, "The End of Moore's Law: A New Beginning for Information Technology," *Computing in Science & Engineering*, vol. 19, no. 2, pp. 41–50, 2017.

[22] P. Ruch, T. Brunschwiler, W. Escher, S. Paredes, and B. Michel, "Toward five-dimensional scaling: How density improves efficiency in future computers," *IBM Journal of Research and Development*, vol. 55, no. 5, pp. 15:1–15:13, 2011.

[23] E. A. Vittoz, "Future of analog in the VLSI environment," in *IEEE international symposium on circuits and systems (ISCAS)*, May 1990, pp. 1372–1375 vol.2.

[24] C. C. Enz and E. A. Vittoz, "CMOS Low-power Analog Circuit Design," in *Emerging technologies, tutorial for 1996 int. Symp. On circuits and systems*, R.Cavin and W.Liu, Eds., Piscataway: IEEE Service Center, 1996, pp. 79–133.

[25] R. Sarpeshkar, "Analog Versus Digital: Extrapolating from Electronics to Neurobiology," *Neural Computation*, vol. 10, no. 7, pp. 1601–1638, 1998.

[26] M. Verhelst and A. Bahai, "Where Analog Meets Digital: Analog-to-Information Conversion and Beyond," *IEEE Solid-State Circuits Magazine*, vol. 7, no. 3, pp. 67–80, 2015.

[27] R. W. Keyes, "Physical limits in digital electronics," *Proceedings of the IEEE*, vol. 63, no. 5, pp. 740–767, May 1975.

[28] R. W. Keyes, "Fundamental limits in digital information processing," *Proceedings of the IEEE*, vol. 69, no. 2, pp. 267–278, 1981.

[29] M. S. Fashami, J. Atulasimha, and S. Bandyopadhyay, "Energy dissipation and error probability in fault-tolerant binary switching," *Scientific Reports*, vol. 3, no. 1, p. 3204, 2013.

[30] C. Enz and M. Punzenberger, "1-V Log-Domain Filters," in *Advances in analog circuit design (AACD'98)*, W. Sansen, J. Huijsing, and R. v. d. Plassche, Eds., Kluwer Book, 1998.

[31] E. Vittoz and Y. Tsividis, "Frequency-dynamic Range-power," in *Trade-offs in analog circuit design, the designer's companion*, C. Toumazou, G. S. Moschytz, and B. Gilbert, Eds., Springer, 2002, pp. 283–313.

[32] B. Murmann, "The Race for the Extra Decibel: A Brief Review of Current ADC Performance Trajectories," *IEEE Solid-State Circuits Magazine*, vol. 7, no. 3, pp. 58–66, 2015.

[33] ITRS, "2005 International Technology Roadmap for Semiconductors - Process Integration, Devices, and Structures (PIDS)." https://www.semiconductors.org/wp-content/uploads/2018/08/2005PIDS.pdf, 2005.

[34] S. Chatterjee, K. P. Pun, S. Nebojša, Y. Tsividis, and P. Kinget, *Analog Circuit Design Techniques at 0.5V*. Springer, 2007.

[35] H. S. Ruiz and R. B. Pérez, *Linear CMOS RF Power Amplifiers - A Complete Design Workflow*, 1-st ed. Springer, 2014.

[36] ITRS, "2015 International Technology Roadmap for Semiconductors - Executive Report." https://www.semiconductors.org/wp-content/uploads/2018/06/0_2015-ITRS-2.0-Executive-Report-1.pdf, 2015.

[37] F. Ellinger, M. Claus, M. Schröter, and C. Carta, "Review of advanced and Beyond CMOS FET technologies for radio frequency circuit design," in *2011 SBMO/IEEE MTT-s international microwave and optoelectronics conference (IMOC 2011)*, 2011, pp. 347–351.

