IOPscience iopscience.iop.org

This content has been downloaded from IOPscience. Please scroll down to see the full text.

Download details:

IP Address: 18.116.164.246
This content was downloaded on 04/05/2024 at 20:40

Please note that terms and conditions apply.

You may also like:

Nanocarbon Allotropes Beyond Graphene

Roadmap on emerging hardware and technology for machine learning
Karl Berggren, Qiangfei Xia, Konstantin K Likharev et al.

Fabrication of a dense array of tall nanostructures over a large sample area with sidewall profile and tip sharpness control
Chang-Hwan Choi and Chang-Jin Kim

Beyond-CMOS roadmap—from Boolean logic to neuro-inspired computing
An Chen

A Multitransition Methanol Survey toward a Large Sample of High-mass Star-forming Regions
J. Y. Zhao, J. S. Zhang, Y. X. Wang et al.

IOP Publishing

Nanoelectronics
Physics, technology and applications
Rutu Parekh and Rasika Dhavse

Chapter 1

Physical and technological limitations of nano-CMOS devices to the end of the roadmap and beyond

Usually, discussion about nanodevices begins by outlining the limitations of metal-oxide semiconductor field-effect transistors (or MOSFETs). So, here we first introduce the two classes and working principle of MOSFETs. In the next section, we will speak about Moore's law, which states that due to technological advancements, the numbers of transistors placed on a single chip will double every 18–24 months [1]. Further, we give a qualitative account of MOSFET scaling methods and several constraints and impacts due to limitations of the technology (which are acknowledged as short-channel effects), setting back the further scaling process. We walk readers through the timeline of CMOS technology, right from this present possible failure of Moore's law due to extensive scaling issues, to finding technological solutions to elongate the timeline of CMOS scaling. We explore the underlying limitations of CMOS technology, the physical phenomena behind it, the need to overcome these limitations, and certain alternatives to work around these limitations for certain applications. We present an in-depth insight into the impact of these technological options on CMOS devices and their applications. Finally, the International Technology Roadmap for Semiconductors (ITRS) is introduced.

1.1 MOSFETs and their scaling

In the field of electronics, Julius Edgar Lilienfeld first proposed the concept of a field-effect transistor in 1925 [2]. In 1948, Walter Brattain, William Shockley and John Bardeen started a revolution by inventing the transistor: a device that can switch or amplify signals and electrical power [2]. Each and every modern electronic device uses these transistors as the primary building blocks of its circuits.

doi:10.1088/978-0-7503-4811-9ch1

© IOP Publishing Ltd 2023

A MOSFET is one type of transistor, defined as a metal-oxide semiconductor field-effect transistor. The first MOSFET was invented by Mohamed M Atalla and Dawon Kahng at Bell Labs in 1959, and first presented in 1960 [2]. A MOSFET is a voltage-controlled device with three terminals, named as the gate, drain and source. A fourth terminal substrate (or body) is primarily shorted with the source to avoid any body-bias effect on its threshold voltage. The main goal of a MOSFET is to control the flow of current in the channel region by changing the gate voltage, V_{GS}. The metal-oxide semiconductor capacitor plays a very important role in providing this functionality of MOSFET. The metallic gate and semiconductor substrate, separated by a thin oxide layer, together works as the M-O-S capacitor. The source and drain terminals are located at two opposite ends of the semiconductor surface. The gate voltage controls the conductivity between the drain and the source by enhancing a continuous path between them (i.e. a channel). The conduction current is either generated by the flow of electrons or that of holes; hence, it is a unipolar device. MOSFETs can be divided into two classes based on their types of doping elements, n-channel MOSFETs and p-channel MOSFETs.

1.1.1 n-Channel MOSFETs

Typically, n-channel MOSFETs (nMOSs) opt for a lightly doped p-type silicon substrate with two heavily doped n-type regions, i.e. a source and a drain. A thin silicon dioxide layer isolates the MOSFET channel from the gate layer. Initially, the channel between the two $n+$ regions does not exist. When the positive gate voltage, V_{GS}, is applied, 'N' channel(s) will form at the surface, which will carry the electrons from source to drain upon application of a potential difference across the drain and source. Figure 1.1(a) illustrates the structure of a nMOS [1–3].

1.1.2 p-Channel MOSFET

The structure of p-channel MOSFETs (pMOSs) is somewhat complementary to that of nMOSs. The substrate is a lightly doped n-type silicon. The source and the drain, also known as wells, are heavily doped with p-type dopant. Initially, the channel between the two p+ regions does not exist. When a negative gate voltage, V_{GS}, is applied, 'P' channel(s) will form at the surface, which will carry the holes from

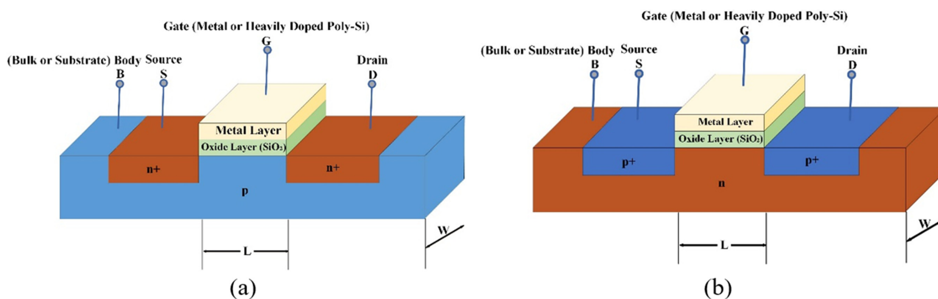

Figure 1.1. MOSFET structures: (a) n-channel, (b) p-channel.

source to drain upon application of a potential difference across the drain and source. Figure 1.1(b) illustrates the structure of a pMOS [1–3].

1.1.3 Working principle of MOSFETs

The applied voltage at the gate is either positive or negative for nMOSs and pMOSs, respectively. For an nMOS, application of positive V_{GS} repels the holes in the p-substrate downwards and generates a negative depletion region charge, which is dependent on the acceptor atoms' concentration. This also attracts and accumulates the electrons from $n+$ source and drain, creating a negatively charged channel between source and drain. At a threshold voltage V_T, the surface electron density becomes equal to the bulk hole density, and the surface is said to be completely inverted. At this point, the device is turned ON and current can flow freely in the channel when the potential difference is applied between source and drain. Similarly, in a pMOS, if a negative V_{GS} is applied to the gate, then the positively charged holes will form a channel. The transconductance and output characteristics of an nMOS are shown in figure 1.2 [1–3].

In figure 1.2(a), one can observe that $I_{DS} = 0$ at any value of V_{DS} when the gate voltage is below the threshold voltage, because the nMOS is in cut-off mode. The drain current I_{DS} will increase with increasing V_{DS} only if $V_{GS} > V_T$. In figure 1.2(b), one can observe that, if $V_{GS} > V_T$, I_{DS} will increase with increasing V_{DS}, at the beginning exhibiting linear or voltage-controlled resistor-like behaviour of the nMOS; and, after the pinch = off point, $V_{DS} = V_{GS} - V_T$, I_{DS} will remain the same for higher values of V_{DS}, causing saturation of the MOSFET. Here, the nMOS works as a constant current source. MOSFETs can be modelled by the set of equations as shown in table 1.1 [3].

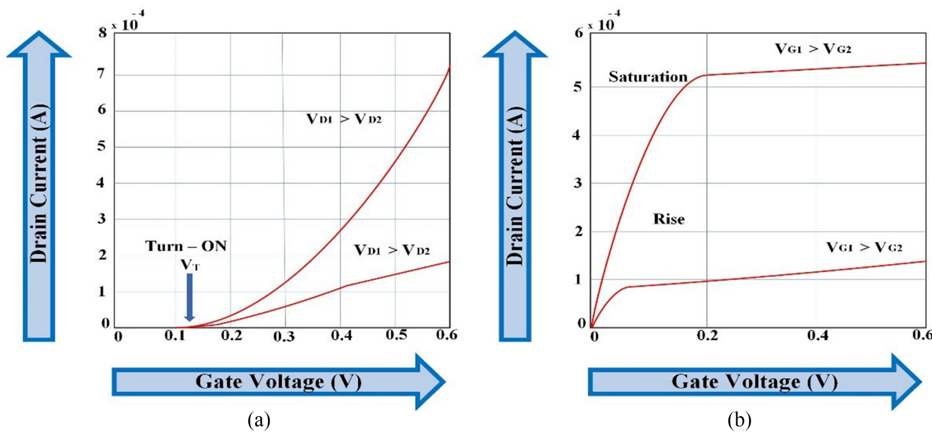

Figure 1.2. (a) Transconductance and (b) output characteristics of an nMOS MOSFET [3].

Table 1.1. Current–voltage equations of MOSFETs [3].

NMOS	Threshold voltage	$V_{TN} = \emptyset_{GC} - 2\emptyset_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_I}{C_{ox}}$ (1.1) $V_{TN0} = \emptyset_{GC} - 2\emptyset_F - \frac{Q_{B0}}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_I}{C_{ox}}$ (1.2) $V_{TN} = V_{TN0} + \gamma\left(\sqrt{\lvert -2\emptyset_F + V_{SB}\rvert} - \sqrt{2\emptyset_F}\right)$ (1.3) where, $\emptyset_{GC}$ is work function difference between gate and channel, $2\emptyset_F$ is surface potential (−ve), $\frac{Q_B}{C_{ox}}$ is depletion layer charge (−ve), $\frac{Q_{ox}}{C_{ox}}$ is surface charge, $\frac{Q_I}{C_{ox}}$ is implant charge, γ is the body effect coefficient (+ve), V_{SB} is source to body bias (+ve) and suffix 0 indicates respective parameter value at zero body bias.
	Cut-off region	$V_{GS} < V_{TN},\ I_{DS} = 0$ (1.4)
	Linear/triode region	$V_{GS} \geq V_{TN}, V_{DS} < V_{GS} - V_{TN},$ $I_{DS} = \frac{\mu_n C_{ox}}{2}\frac{W}{L}\left(2(V_{GS} - V_{TON})V_{DS} - V_{DS}^2\right)$ (1.5)
	Saturation region	$V_{GS} \geq V_{TN},\ V_{DS} > V_{GS} - V_{TN},$ $I_{DS} = \frac{\mu_n C_{ox}}{2}\frac{W}{L}(V_{GS} - V_{TON})^2(1 + \gamma V_{DS})$ (1.6) where, μ_n is surface electron mobility of NMOS.
PMOS	Threshold voltage	$V_{TP} = \emptyset_{GC} - 2\emptyset_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_I}{C_{ox}}$ (1.7) $V_{TP0} = \emptyset_{GC} - 2\emptyset_F - \frac{Q_{B0}}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_I}{C_{ox}}$ (1.8) $V_{TP} = V_{TP0} + \gamma\left(\sqrt{\lvert -2\emptyset_F + V_{SB}\rvert} - \sqrt{2\emptyset_F}\right)$ (1.9) where, $2\emptyset_F$ is surface potential (+ve), $\frac{Q_B}{C_{ox}}$ is depletion layer charge (+ve), γ is the body effect coefficient (−ve) and V_{SB} is source to body bias (−ve).
	Cut-off region	$V_{GS} > V_{TP},\ I_{DS} = 0$ (1.10)
	Linear/triode region	$V_{GS} \leq V_{TP},\ V_{DS} > V_{GS} - V_{TP},$ $I_{DS} = \frac{\mu_p C_{ox}}{2}\frac{W}{L}\left(2(V_{GS} - V_{TOP})V_{DS} - V_{DS}^2\right)$ (1.11) Where, μ_p is surface hole mobility of PMOS.
	Saturation region	$V_{GS} \leq V_{TP},\ V_{DS} < V_{GS} - V_{TP},$ $I_{DS} = \frac{\mu_p C_{ox}}{2}\frac{W}{L}(V_{GS} - V_{TOP})^2(1 + \gamma V_{DS})$ (1.12)

1.1.4 Introduction and failure of Moore's law

In 1965, Gordon Moore, in his paper 'Cramming more components onto Integrated Circuits', explained that the transistor density on an integrated circuit (IC) would grow exponentially. This speculation became popularly known as 'Moore's law', and its scaling is depicted in figure 1.3.

Moore law's states that the number of transistors on ICs will double roughly every two years. To stick to this pace of technological advancement, the size of

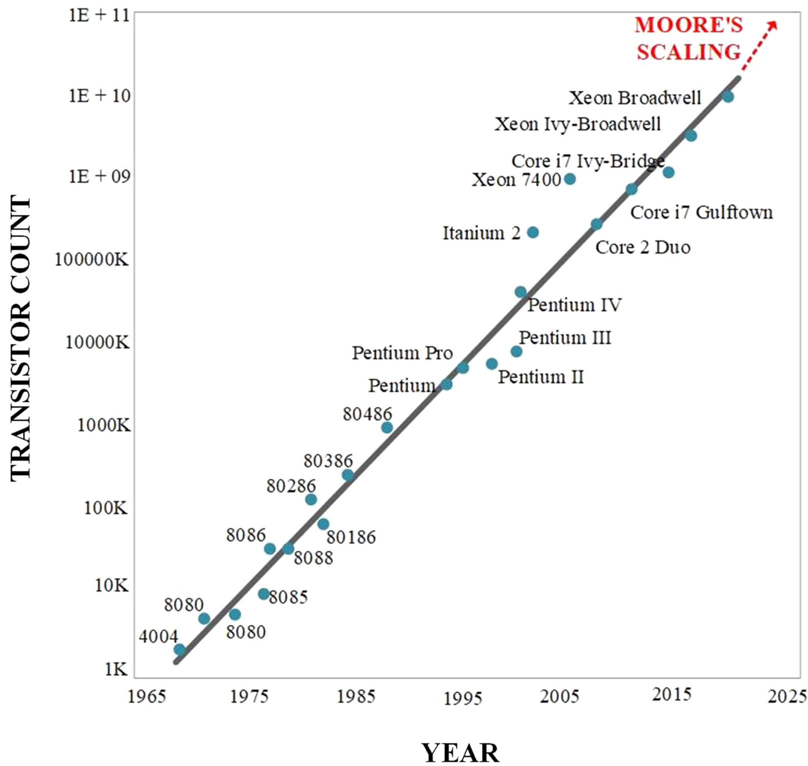

Figure 1.3. MOSFET transistor count for microprocessors. Adapted from an image by Max Roser and Hannah Ritchie/OurWorldinData (https://ourworldindata.org/uploads/2020/11/Transistor-Count-over-time.png). CC BY 4.0.

individual transistors should be reduced and their packing density should double every 18–24 months. For this, we need to miniaturise the dimensions of MOSFETs—a process known as scaling. This scaling of the MOSFET is key to continue following Moore's law. This law predicted the continuing evolution of the semiconductor industry, in which the realisation of increasingly complex devices and systems became possible. Gradual shrinking of the dimensions of the transistors below 100 nm enabled hundreds of millions of transistors to be placed on a single chip. Following Moore's law in semiconductor electronics, performance indexes such as processing speed, memory retention and efficiency were considerably enhanced [3]. Low power dissipation, smaller chip size, low cost and increased package density are additional advantages of scaling. The concepts of 'More Moore', which discusses hyper-miniaturisation, and 'More than Moore', which addresses diversification, were introduced in 2005 when the ITRS published its first white paper; these concepts are elaborated upon in figure 1.4 [4].

In order to improve the conductivity of a MOSFET by a factor (say) S, where $S > 1$, we should decrease the dimensions by S. Here, S is identified as the scaling factor. Two distinct types of scaling are (i) constant field scaling, and (ii) constant

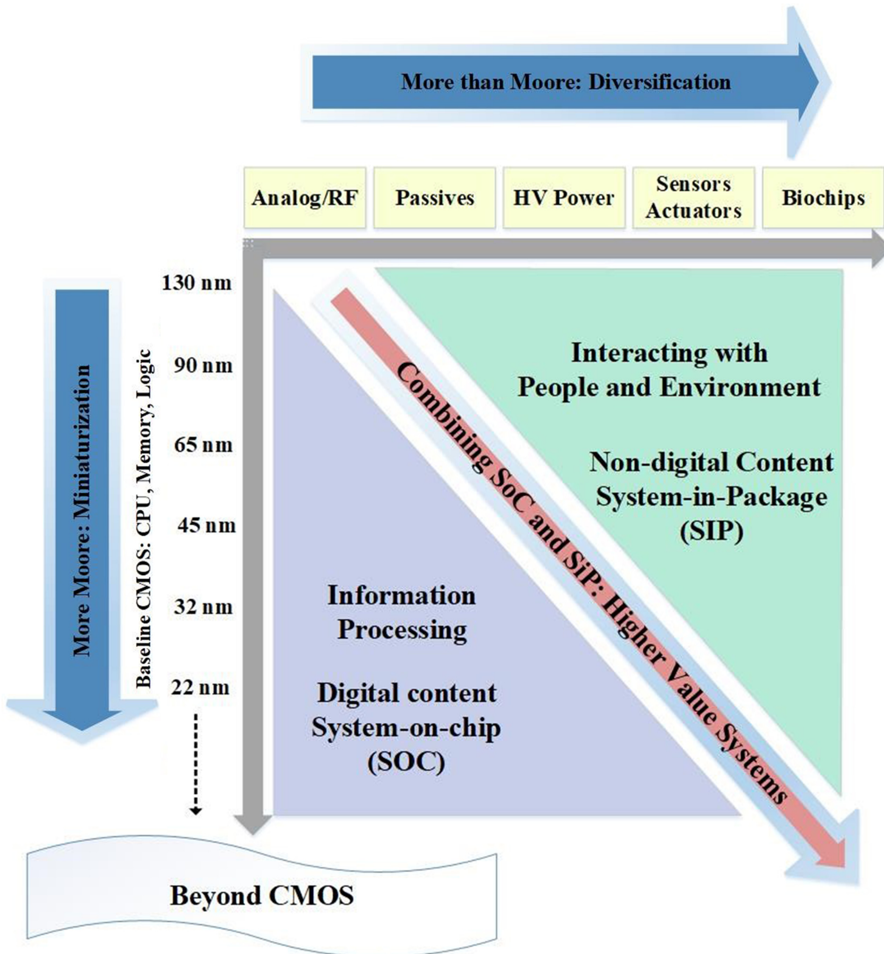

Figure 1.4. More Moore—beyond CMOS. Adapted from [4] with permission.

voltage scaling. Constant field scaling (or full scaling) retains the magnitude of the electric field inside the MOSFET, whereas the dimensions are reduced by a factor S. In this scaling type, the charge densities are gained by α because the magnitude of the electric field is kept the same. Let us say the scaling factor is $S = \alpha$. Before scaling, the power dissipated is $P = I_d \times V_{ds}$. As the power P is affected by both the drain current and source drain voltage, it will increase by α^2 [6].

In constant voltage scaling, the terminal voltage remains unchanged. Just as in full scaling, the MOSFET dimensions are reduced by S (=β). The power dissipation and power density will increase by β and β^3, respectively. The drain current is also increased by β.

The scaling-dependent changes in device dimensions and doping densities, and their effects on device parameters are listed in table 1.2 [3].

Table 1.2. Effect of scaling [3].

S. No.	MOSFET parameter	Before scaling	After scaling: Full scaling	After scaling: Constant voltage scaling
1	Channel length	L	$L' = L/\alpha$	$L' = L/\beta$
2	Channel width	W	$W' = W/\alpha$	$W' = W/\beta$
3	Channel area	A	$A' = A/\alpha^2$	$A' = A/\beta^2$
4	Gate oxide thickness	t_{ox}	$t'_{ox} = t_{ox}/\alpha$	$t'_{ox} = t_{ox}/\beta$
5	Junction depth	X_j	$X_j' = X_j/\alpha$	$X_j' = X_j/\beta$
6	Supply voltage	V_{DD}	$V'_{DD} = V_{DD}/\alpha$	$V'_{DD} = V_{DD}$
7	Gate to source voltage	V_{GS}	$V'_{GS} = V_{GS}/\alpha$	$V'_{GS} = V_{GS}$
8	Drain to source voltage	V_{DS}	$V'_{DS} = V_{DS}/\alpha$	$V'_{DS} = V_{DS}$
9	Body to source voltage	V_{BS}	$V'_{BS} = V_{BS}/\alpha$	$V'_{BS} = V_{BS}$
10.	Threshold voltage	V_{TH}	$V'_{TH} = V_{TH}/\alpha$	$V'_{TH} = V_{TH}$
11.	Doping densities	N_A N_D	$N'_A = \alpha.\ N_A$ $N'_D = \alpha.\ N_D$	$N_A = \beta^2.\ N_A$ $N_D = \beta^2.\ N_D$
12.	Electric field	$E = \frac{V_{GS}}{t_{ox}}$	$E' = \frac{V_{GS}}{t_{ox}} = E$	$E' = \beta\frac{V_{GS}}{t_{ox}} = \beta E$
13.	Oxide capacitance	$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$	$C_{ox}' = \alpha\frac{\varepsilon_{ox}}{t_{ox}} = \alpha C_{ox}$	$C_{ox}' = \beta\frac{\varepsilon_{ox}}{t_{ox}} = \beta C_{ox}$
14.	Drain current	I_{DS}	$I_{DS}' = \frac{I_{DS}}{\alpha}$	$I_{DS}' = \beta I_{DS}$
15.	Power dissipation	$P = I_{DS}V_{DS}$	$P' = \frac{I_{DS}}{\alpha}\frac{V_{DS}}{\alpha} = \frac{P}{\alpha^2}$	$P' = \beta(I_{DS}V_{DS}) = \beta P$
16.	Power density	$P_d = P/A$	$P_d' = \frac{P'}{A'} = \frac{(P/\alpha^2)}{(A/\alpha^2)} = P_d$	$P_d' = \frac{P'}{A'} = \frac{(\beta P)}{\frac{A}{\beta^2}} = \beta^3 P_d$

The scaling of voltages is not preferred due to the need for complicated level-shifter arrangements. As such, constant field scaling is used only for low-power applications. Constant voltage scaling is used in high-performance applications, but an increase in power density by a factor of β^3 may cause serious reliability issues.

1.2 Limitations and showstoppers arising from CMOS scaling, and technological options for MOSFET optimisation

For a good four decades, the industry has kept pace with growth rate speculations, although the end of CMOS technology had been predicted since as early as the 1970s. The rapid growth rate seemingly approached an infinite number of transistors in an IC. In 2003, Gordon Moore revised his propositions in his paper 'No exponential is forever, but forever can be delayed!', which suggested that CMOS scaling will reach a limit but that the limit can be practically extended. Thus, it became important to seek out the material and structural design limitations that

can limit CMOS scaling, hence enabling the continuation of CMOS scaling in the near term.

Scaling a transistor beyond the deep-submicron regime gives rise to many unwanted physical mechanisms, leading to the failure of their classical behaviour. Due to these mechanisms, the industry is now reaching a number of hard limits that no amount of research can overcome. These limitations are categorised into physical challenges, material challenges, thermal challenges, technological limitations and economic challenges [5–7].

(i) Physical challenges

As we go on scaling MOSFET devices, the channel length becoming ever shorter eventually leads to it being comparable to the depletion region, which contributes to short-channel effects (SCEs), including but not limited to drain-induced barrier lowering (DIBL) and sub-threshold channel currents, gate tunnel currents, gate-induced drain leakage (GIDL), junction leakage, velocity saturation, hot carrier degradation, etc. Some of these are explained below.

(a) Drain-induced barrier lowering

With increasing positive drain potential, the drain depletion region starts expanding. Normally, this would not affect the energy bands near the source region. However, in short-channel devices, it extends towards the source region, lowers the barrier between source and channel, and causes current flow between the source drain even when the transistor is in the OFF state. This is known as DIBL, and is illustrated in figure 1.5. For large drain bias voltage, the depletion region of the drain extends

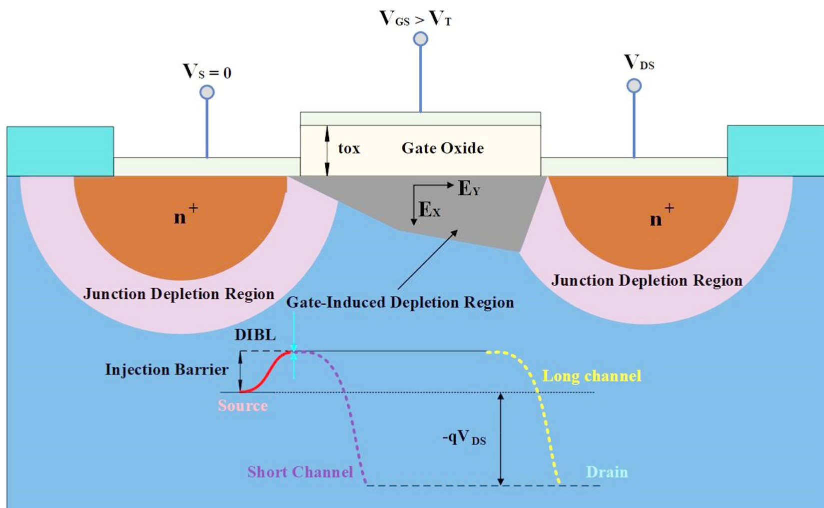

Figure 1.5. Drain-induced barrier lowering. Adapted from [3]. © The Authors. Published by McGraw-Hill Education.

towards the source and merges. This is known as 'punch through'. Punch through can be minimised by using thinner oxide, using larger substrate doping, shallower junctions and longer channels [3, 5–7].

(b) Gate tunnel currents

In MOSFET scaling, the gate oxide should be made as thin as possible so as to increase the channel conductivity and performance when the transistor is in an ON state and to reduce sub-threshold leakage when the transistor is in an OFF state. However, with very thin gate oxides the quantum-mechanical phenomenon of electron tunnelling occurs between the gate and channel, leading to increased power consumption. Insulators that have a larger dielectric constant than silicon dioxide (referred to as high-κ dielectrics), such as group IV-B metal silicates, e.g. hafnium and zirconium silicates and oxides, are being used to reduce gate leakage from the 45 nm technology node onwards [3, 5–7].

(c) Off-state leakage

The presence of a power supply when a short-channel transistor is OFF causes small drain leakage, as shown in figure 1.6. As the gate length decreases, the leakage current grows exponentially [3, 5–7]. Such off-state or sub-threshold current lead to static power dissipation, making a chip power hungry [3, 5–7].

(d) Gate-induced drain leakage

In a classical planar long-channel MOSFET, the substrate and gate are electrostatically shielded from the drain. Thus, the threshold voltage

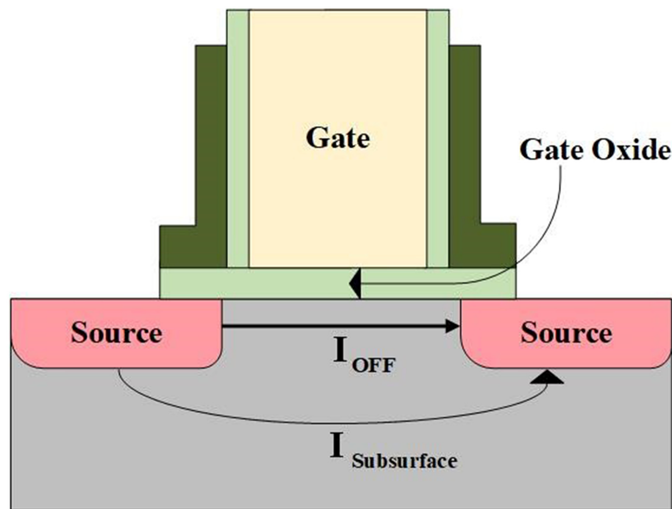

Figure 1.6. Off-state leakage.

is not dependent on drain supply. However, in the short-channel case, the drain is close enough to the gate. With a high drain supply, the *n*+ drain region under the gate becomes depleted and even inverted. This causes field crowding, resulting in avalanche multiplication and band-to-band tunnelling. Thus, minority carriers are emitted in the drain region underneath the gate and leakage current flows through the substrate [3, 5–7]. This is known as GIDL, and is illustrated in figure 1.7.

(ii) Material challenges

The failure of dielectric and wiring materials to continue providing dependable insulation and conduction, as we scale down, constitute the material challenges. The materials used in making ICs, such as silicon, silicon dioxide, aluminium, copper, etc., reach the limit of their physical capabilities like dielectric constant, carrier mobility, breakdown field strength, conductivity, etc. With the continued use of these materials, present and future scaled devices would not be able to keep up their performance [4–7].

(iii) Thermal challenges

Because the number of transistors per unit area on the chip is increasing, the total power consumption and thermal dissipation are also increasing. Because the supply voltage is not being scaled relative to the pace of the channel length, the power density is growing [4–7].

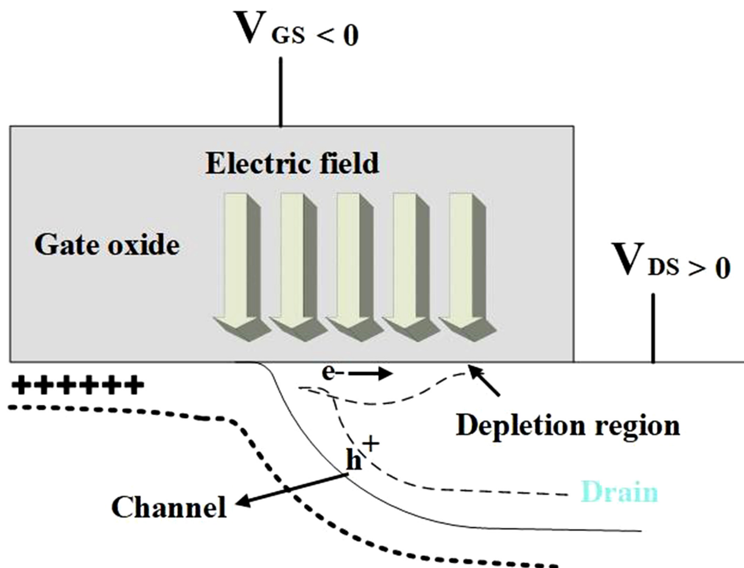

Figure 1.7. Gate-induced drain leakage [3].

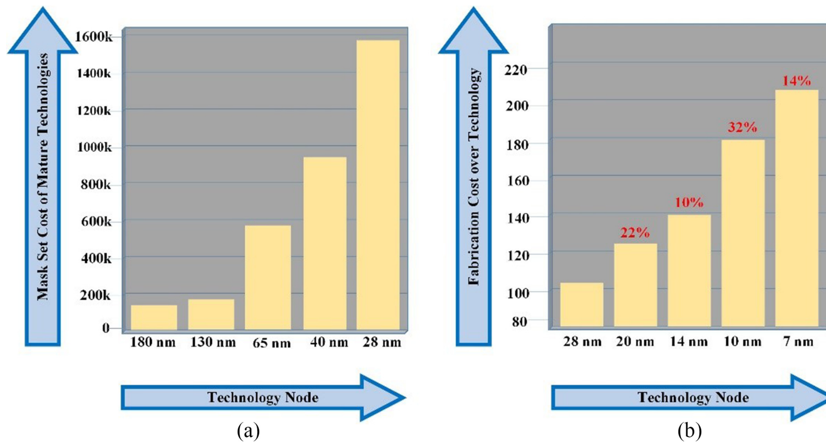

Figure 1.8. Cost vs technology nodes: (a) mask set, (b) fabrication. Adapted from [8].

(iv) Technological challenges—lithographic issues

This is considered to be the primary limitation of chip scaling. Lithographic issues pose a limit to the dimensions which can be fabricated on a chip. Research up to now has enabled lithographic technology to use the UV wavelength (193 nm), and current research is being undertaken to bring this down to the extreme-UV spectrum (13 nm). However, the dimensions of the transistor will face a hard limit: the size of the atom and molecule. The cost of fabrication has also increased and grown exponentially, which limits the profitability of increased scaling [4].

(v) Economical challenges

As aforementioned, production costs and testing costs are rising exponentially with time, as indicated in figure 1.8 and the figures given. Whilst we are advancing the technology and scaling down our semiconductor devices, we also need to simulate, test and mass produce our devices, which is what contributes the most to the cost [4–7].

Thus, various limits to the scaling of MOSFETs compel the semiconductor industry to look to novel devices and circuit design techniques.

1.2.1 ITRS: The International Technology Roadmap for Semiconductors

The ITRS was released to the public domain with the sole intention that it be used as the main common reference in the industry. It is devised to help consortia, students at institutes, and industry researchers to spur discovery in the field [4]. The ITRS consolidates a set of documents delivered by semiconductor industry professionals. These documents comprise a set of articles and reports delivered by the International Roadmap Committee (IRC), which is constituted by professional experts and representatives of semiconductor industry organisations across Europe, Taiwan, the United States, South Korea, China, and Japan between 1998 and 2015 [9]. The documents are generated with the disclaimer:

The ITRS is devised and meant for technology evaluation only and is without consideration to any commercial concern about individual commodities or devices [9].

The objectives of the ITRS include the following:

(i) To ensure cost-effective improvements in the performance of ICs, superior commodities, and applications that engage the beforementioned devices, thereby sustaining the strength and prosperity of the aforementioned organisations.

(ii) To help speed up the pace of research in the future in these domains of technology.

The documents represent the best collective opinion on the directions of research and proposed timelines extending up to about 15 years in the future for the following areas of technology:

(a) Yield enhancement: Yield is represented by the functionality and reliability of the ICs produced on wafer surfaces in the semiconductor industry. In most industries, yield has been defined as the number of products that can be sold divided by the number of products manufactured, and is an area that can potentially be made more efficient. During the manufacturing of ICs yield loss is caused by defects, faults, process variations, and design.

(b) Front-end processes: Identical ICs, known as die, are made on each wafer in a multi-step batch process. Every step adds a new layer to the wafer or modifies the existing one. These layers form the elements of the individual electronic circuits. This is known as wafer fabrication or the front-end process.

(c) Factory integration: This aspect of the ITRS focuses on integration of all the factory components needed to produce the required products efficiently in the right volumes on schedule while meeting cost targets.

(d) Test and test equipment: Semiconductor testing equipment (IC tester) and automated testing equipment (ATE) is a system for delivering electrical signals to a semiconductor device to compare output signals against expected values in its design specifications constraints.

(e) Process integration, devices and structures: This section deals with the main IC devices and structures, with overall IC process flow integration, and also with the reliability trade-offs associated with new options.

(f) Emerging research materials: Nanomaterials describe, in principle, materials of which a single unit is sized (in at least one dimension) between 1 and 1000 nm (but usually is 1–100 nm).

(g) Photolithography: Photolithography has become the most successful technique capable of producing sub-100 nm patterns.

(h) Radio frequency and analogue/mixed-signal technologies: Radio frequency (RF) and analogue/mixed-signal (AMS) technologies serve the rapidly growing advanced communications and 'More than Moore' markets. They also compose essential and critical technologies for the success of many semiconductor manufacturers, as well as the ultimate success of the future Internet of Things (IoT).

(i) IC interconnects: In ICs, interconnects are structures that connect two or more circuit elements (such as transistors) together electrically. The design and layout of interconnects on the IC is vital to its proper functioning, performance, power efficiency, reliability and fabrication yield.
(j) Assembly and packaging: The case, known as a 'package', supports the electrical contacts that connect the device to a circuit board. This process is often referred to as packaging in the IC industry; it is otherwise known as semiconductor device assembly, encapsulation or sealing.
(k) System drivers/design: The design and engineering of advanced solid-state nanoporous materials could, for example, allow for the development of novel gene-sequencing technologies that enable single-molecule detection at low cost and high speed with minimal sample preparation and instrumentation.
(l) Modelling and simulation: Semiconductor device modelling creates models for the behaviour of electrical devices based on fundamental physics, such as the doping profiles of the devices. Semiconductor process simulation is the modelling of the fabrication of semiconductor devices such as transistors.
(m) Emerging research devices: This section is motivated by the increasing difficulty of meeting all expected requirements for the rigorously scaled technologies projected for later technology nodes.
(n) Metrology: Generally, metrology denotes the methods of measuring numbers and volumes, primarily by using metrological equipment. The number of measurement points varies by semiconductor manufacturer or device.
(o) Microelectromechanical systems (MEMS): MEMS are a specialised field referring to technologies that are capable of miniaturising existing sensor, actuator or system products. Nanotechnology is a growing field that uses the unique properties of ultra-small-scale materials to an advantage.
(p) Environment, health and safety: This section addresses how nanotechnology can help in these three research areas.

Significant milestones for CMOS processor technology projected by the ITRS are shown in table 1.3.

1.2.2 Update beyond the end of the roadmap

To address the altered ecosystem of the microelectronics industry, the IRC wanted to reconstruct its constitution and goals in its annual meeting convened in Taiwan in December 2012 [9]. Accordingly, ITRS 2.0 was born in 2013–14 to fulfill the above task of the IRC. The process of reframing the ITRS was divided into 17 Technology Working Groups, which were further mapped to seven focus teams, known as IFTs [9]. These include the following:

(i) Beyond CMOS: Beyond CMOS addresses devices that are not CMOS based, but refers instead to the possible future digital electronic technologies beyond the scaling limits of CMOS. These new devices exhibit better speeds and lower densities than CMOS. Some examples of beyond CMOS devices are spin FET and spin MOSFET transistors, negative gate capacitance FETs,

Table 1.3. ITRS characteristics [9].

S. No.	Characteristics	2010	2012	2014	2016	2018	2020
1.	Metal pitch (nm)	45	32	24	18.6	15	11.9
2.	V_T (V)	0.289 EPbulk	0.291 EPbulk	0.221 UTB FD	0.202 MG	0.207 MG	0.219 MG
3.	V_{DD} (V)	0.97	0.9	0.84	0.78	0.73	0.68
4.	Power density (W/mm^2)	0.5	0.6	0.7	0.8	0.9	1
5.	Max pin count	4900	5300	5900	6500	7200	7900
6.	Performance: on-chip (GHz)	5.88	6.82	7.91	9.18	10.65	12.36
7.	Performance: chip-to-board (Gb/s)	10	14	17	30	40	50

NEMS switches, Mott FETs, piezotronic logic transduction devices, spin-wave devices, nanomagnetic logic devices, spin torque majority logic gates, spintronics, memristors and all spin logic devices.

(ii) Outside system connectivity: This focuses on technologies based on wireless communication, including defining the type of work needed and finding the best solution. It aims at identifying technology and device requirements for enhancing intersystem communications and the corresponding research needed.

(iii) Factory integration: This focuses on new tools and processes and on producing heterogeneous integration of all these things. The specific scope of factory integration is wafer fabrication, or manufacturing in the front-end and back-end.

(iv) Heterogeneous integration: This mainly focuses on the integration of components that are manufactured separately with different technologies into a new combined component which works better than they do separately. For instance, cameras and microphones are two components that are manufactured differently.

(v) More Moore: The More Moore focus team provides electrical, physical and reliability requirements for memory and logic technologies to assist More Moore scaling for cloud (IoT and server), mobility and big data applications. Shrinking of CMOS is the main objective.

(vi) Heterogeneous component: This chiefly focuses on the various components that create heterogeneous systems such as in power generation, sensing devices and MEMS. These components cannot certainly scale down as per Moore's law.

(vii) System integration: System integration is a topic that concentrates on the design of and knowledge pertaining to the integration of heterogeneous blocks.

The above IFTs include elements from the ITRS along with many **new** elements, like novel charge-based and non-charge-based devices, wireless connectivity, heterogeneous integration, etc [9]. It is needless to say that the purpose of ITRS 2.0 was not limited to CMOS devices and their technologies but served rather to address a newer approach to system integration, as well as the topics of traversing the means of communications from conductors to wireless and then to optical fiber, in addition to continuing to search for non-electron-based technologies [9]. The last version of ITRS was published in 2013, which as mentioned aimed to provide a roadmap for the next 15 years, that is, up to 2028. But, as we now know, Moore's law was by then already reaching its last days, so researchers decided to generate a new roadmap, named the International Roadmap for Devices and Systems (IRDS). The IRDS, which was introduced in 2016, is thus the effective successor of the ITRS. Its goals include the following [11]:

(a) To find new technology for devices and systems and provide a roadmap for the next 15 years.
(b) To determine needs and challenges, and to find new opportunities for change.
(c) To encourage people in these aims around the world by conducting seminars, workshops or through IEEE conferences.

1.2.3 The show must go on!

According to many predictions, if the trend continues as per IRDS guidelines, CMOS scaling will no longer be effective. Practical speculations restrict the technology to a node size of 14 nm; and, even before this restriction is reached, there are extreme barriers to overcome. So, what technological options might be used to optimise MOSFETS and elongate the CMOS scaling period? The next section shall provide some solutions to successfully (however theoretically) overcome these constraints and optimise MOSFET structures.

(i) Improvements and technological aspects of MOSFET optimisation

    The optimisation of MOSFET structures to extend the boundaries of CMOS scaling takes a two-faceted approach: (a) use of new materials in existing structures, and (b) use of new structures.

    (a) Use of new materials

        The problems faced by the chips can be handled by adjusting the relevant critical properties. These sensitive properties are low-resistivity conductors, low-*k* dielectrics and strained silicon. It has been a common practice to use tungsten and copper to minimise device and interconnect resistance. Additionally, metal gate electrodes offer many advantages over doped-polysilicon gates. The polysilicon depletion effect (PDE) of doped-polysilicon gates effectively increases gate oxide thickness at inversion by a couple of nanometres, resulting in degradation of the gate capacitance. For sub-50 nm CMOS nodes, the gate oxide thickness is typically less than 1.5 nm. The PDE is quite dominant in this regime.

Additionally, boron penetration in the thin oxide underneath the doped poly-gate degrades its quality; and further, there is a practical limit on oxide scaling due to plausible gate leakage. A metal gate offers no depletion, very low gate resistance, no boron penetration and compatibility with high-κ. Strained silicon crystal has been the star of technology scaling for high-performance requirements. In this, the silicon lattice is subjected to physical strain (mostly by adding a layer of another lattice having a larger lattice size on top of the silicon) to improve carrier mobility and cut down the device resistance as well as several other critical properties. The disadvantage of this technique is in the fabrication aspect, as it does not fit with many technologies that are new in the market. High-κ dielectrics address the issue of off-state power consumption and more particularly gate leakage due to tunnelling currents. As dielectric thickness ultimately restricts the gate length, the dielectric thickness will be the first to reach atomic dimensions. Practically, the length of the gate has to be 40 times in comparison to the dielectric thickness so that it can properly control the SCEs. The proposed solution is to find and use a material with higher dielectric constant than SiO_2 that will allow us to reduce the effective dielectric thickness without affecting tunnelling [9].

(b) Use of new structures

Proposals for changing the structures of the transistor have also been made. The two main proposed structural modifications are silicon-on-insulator (SOI) and double-gate (subsequently multi-gate and gate-all-around) complementary metal-oxide semiconductors (DGCMOSs). Normally, a transistor is fabricated by connecting its body to a substrate, but in SOI we first bury an oxide on top of the substrate and the transistor is fabricated on its pinnacle. This isolates the frame electrically from its environment and a positive body bias exists (for nMOS, of course), reducing the device threshold voltage and increasing performance. An insulating layer is used to achieve a highly resistant element with zero junction area capacitance, and no reverse body effect by stacked circuits. Schematics of a bulk MOSFET and a SOI MOSFET are shown in figure 1.9.

DGCMOSs utilise the fundamental concept of adding an extra gate to boost coupling among the gate and the channel. This scheme has been hailed as the 'perfect structure for scalability'. A cross section of a DG-SOI MOSFET is shown in figure 1.10 [9].

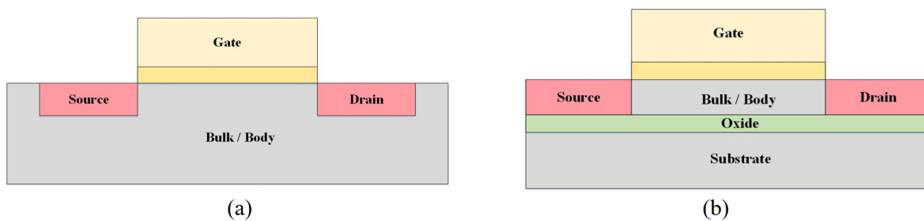

Figure 1.9. FET schematics: (a) bulk, (b) SOI.

The principal objectives of DG-SOI MOSFETs are to reduce SCEs while significantly maintaining good electrical characteristics. Less gate leakage and sub-threshold current and high ON current are the benefits conferred by DG-MOSFETs; however, realising their structure poses some difficulties. When using traditional fabrication techniques, adding a second gate below the device's body results in troublesome alignment issues. This demands a complex process, higher gate capacitance and higher source to drain series resistance. The abovementioned problems are addressed by using fin-shaped FETs (or FinFETs). Working beyond the 45 nm node, FinFETs have done wonders. These are categorically different from planar MOSFETs: they offer low leakage with high driving capability, supply voltage scaling and increased intrinsic gain. A three-dimensional profile of a FinFET is illustrated in figure 1.11.

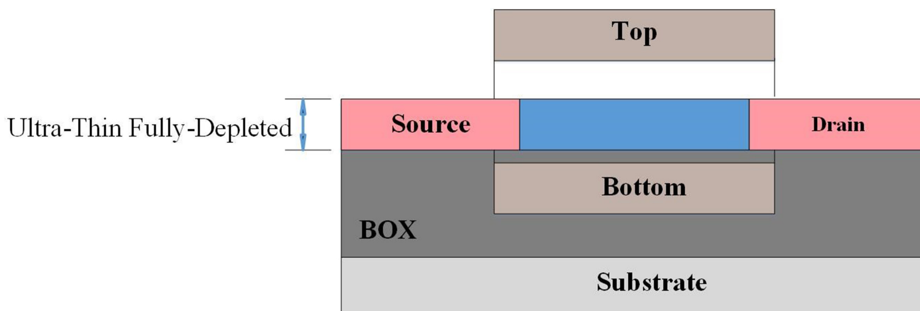

Figure 1.10. Double-gate SOI MOSFET.

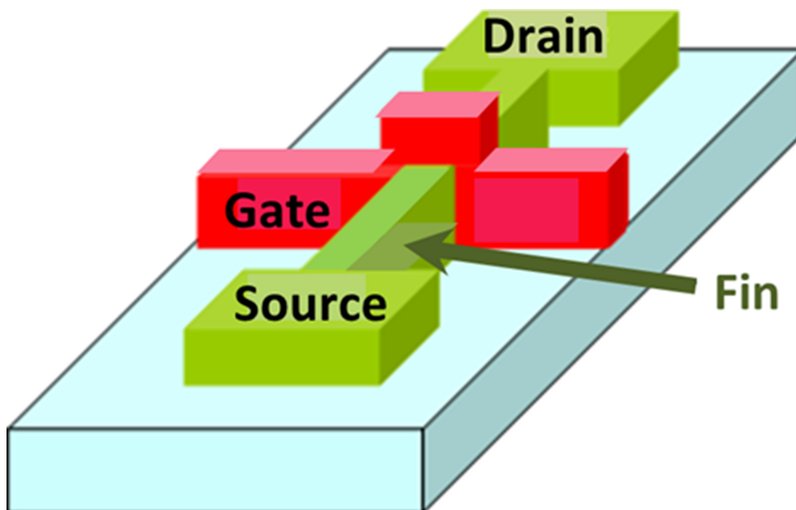

Figure 1.11. FinFET structure. (This Doublegate FinFET.PNG image has been obtained by the author(s) from the Wikimedia website where it was made available under a CC BY-SA 3.0 licence. It is included within this book on that basis. It is attributed to Irene Ringworm.)

FinFETs demonstrate high parasitic resistance and quantised device width. An important point to note is that SCEs can be controlled only if the width of the body is taken to be a quarter of the gate length. As the gate length is the smallest dimension fabricated, it is a difficult challenge to overcome. Although FinFETs have proved their worth, research in this area is still in progress. Better gate control can be obtained in further variants of FinFETs such as multi-gate or gate-all-around MOSFETs.

(ii) Novel devices

Novel and emerging logic devices that extend MOSFETs to the end of the roadmap, in addition to unconventional FETs and non-FETs are illustrated in figure 1.12 [11].

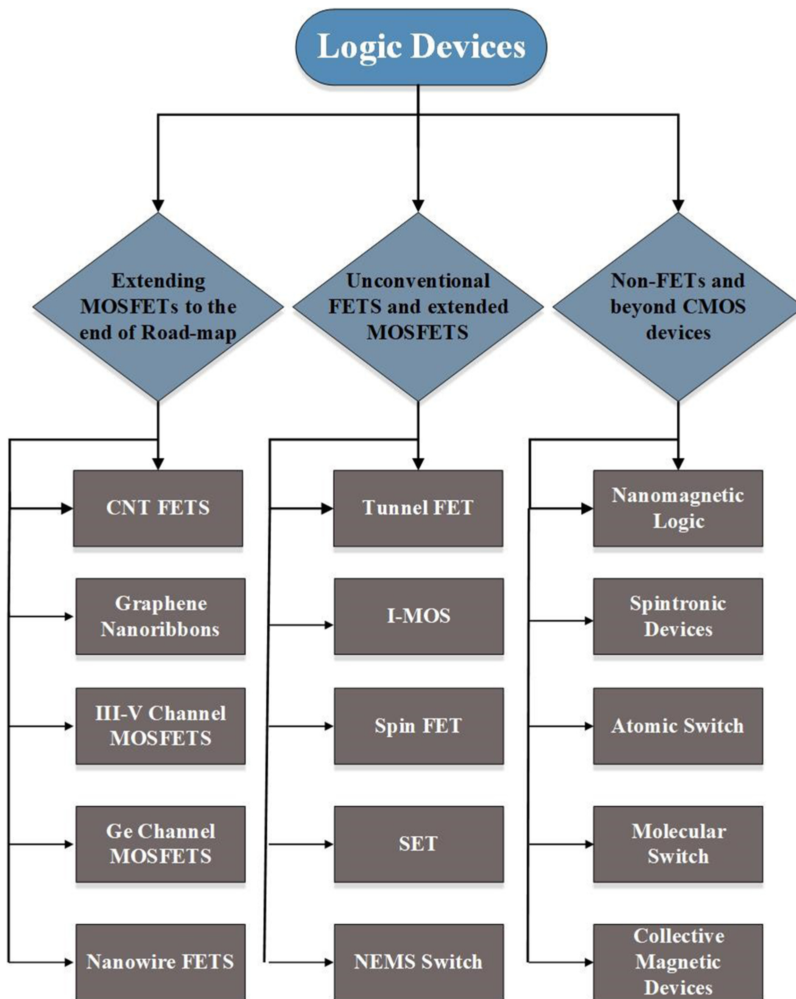

Figure 1.12. Logic devices. © 2021 IEEE. Adapted, with permission, from [11].

On similar grounds, memory devices are categorised into three classes: baseline, prototypical and emerging. These are illustrated in figure 1.13 [11].

A comparison of the major emerging devices based on their traits is briefly outlined in table 1.4.

In summary, the emergence of novel nano-regime devices and advancements in fabrication technology down to such a tiny scale are saviours for the semiconductor

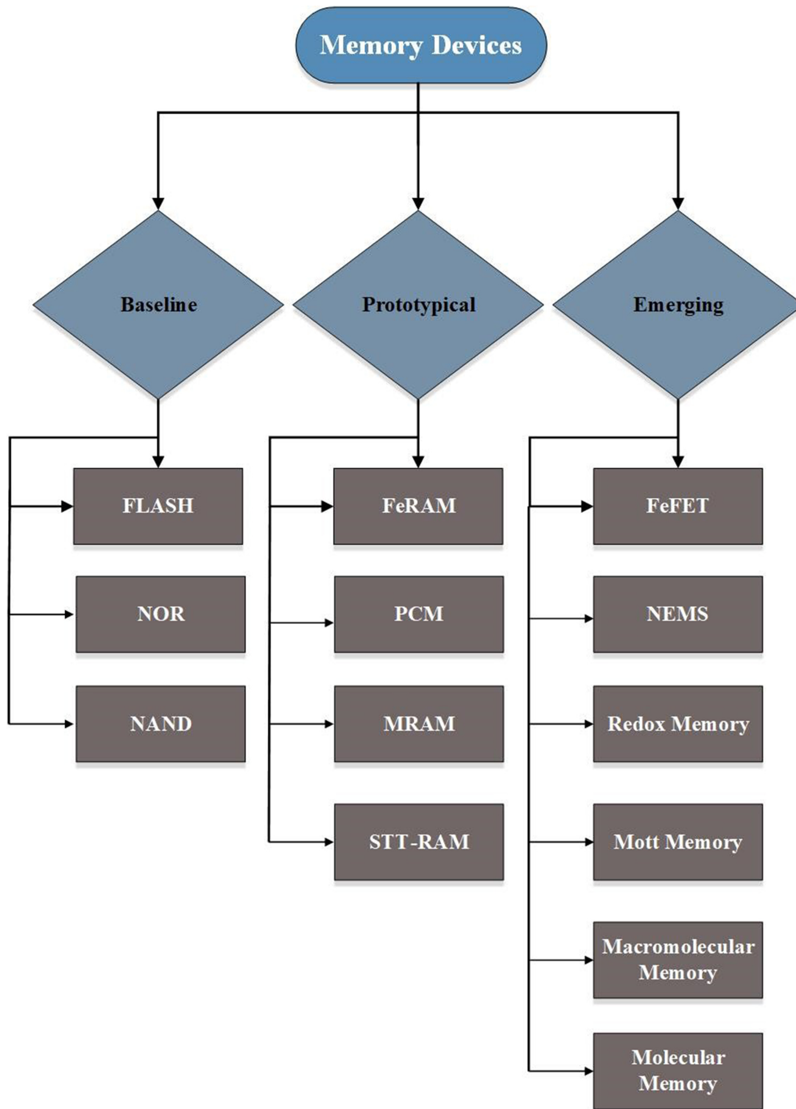

Figure 1.13. Memory devices. © 2021 IEEE. Adapted, with permission, from [11].

Table 1.4. Comparison of different device technologies [11].

Sr No.	Technology	Pluses	Minuses
1.	CNTs/graphene/ nanotubes	High conductivity	Fabrication issues
2.	Ge channel devices	High mobility, high performance	Carrier scattering
3.	Nanowire devices	High aspect ratio leading to better channel control, remarkable optical properties	Poor current drive
4.	RTD/tunnel FET	Apt for RF applications	Device matching across the wafer
5.	Spintronics	Scalability, low power, high performance	Perfection in spin control mechanism
6.	Single-electron transistors	Ultra-low power	Room-temperature operation, lower drive current
7.	Fe-RAM	Does not need charge pump, guaranteed data reliability	Limited performance
8.	PCM	Fast switching, scalability	Thermal sensitivity, fabrication
9.	MRAM	Refresh not required	Less packing density
10.	Fe FET	Low voltage operation	Material choice, leakage, charge injection
11.	NEMS	Efficient, small and cheaper systems	Reliability and long-term stability
12.	Mott memory	Fast switching, ultra-low power operation	Instability
13.	Molecular electronics	Can overcome long interconnect issues	Stability problem

industry. Though complete replacement of CMOS technology is a far-fetched proposal, newer technologies can coexist with the existing technology and the lifetime of CMOS can be extended. With newer materials, devices, instruments and processes, it is possible to commence on the development of exclusive, smart, portable, power/energy efficient, high-performance and cheaper products.

Questions

1. State Moore's law. Does it hold in today's age?
2. What is the ITRS? State its significance.
3. State the different types of scaling. Explain the physics behind them along with their applications.
4. What are the limitations of CMOS scaling? Explain briefly.
5. What is the future of microelectronics/very-large-scale integration (VLSI) technology?

References

[1] Sze S M 2002 *Semiconductor Devices, Physics and Technology* 2nd edn (New York: Wiley)

[2] Streetman B 2015 *Solid State Electronic Devices* (London: Pearson) 7th edn

[3] Kang and Leblebici 2003 *CMOS Digital Integrated Circuits: Analysis and Design* 3rd edn (New Delhi: Tata McGraw-Hill)

[4] *2005 International Technology Roadmap for Semiconductors (ITRS) by Semiconductor Industry Association* 2005 (https://semiconductors.org/resources/2005-international-technology-roadmap-for-semiconductors-itrs)

[5] Haron N Z and Hamdioui S 2008 Why is CMOS scaling coming to an END? *3rd Int. Design and Test Workshop, Monastir* pp 98–103

[6] Iwai H 2004 CMOS Scaling for sub-90 nm to sub-l0 nm *Proc. 17th Int. Conf. on VLSI Design (VLSID04)* pp 30–5

[7] Rairigh D 2005 *Limits of CMOS Technology Scaling and Technologies Beyond-CMOS* (Piscataway, NJ: IEEE)

[8] Ferguson J 2017 Assessing the true cost of node transitions (https://techdesignforums.com/practice/technique/assessing-the-true-cost-of-node-transitions)

[9] *Introduction to International Technology Roadmap for Semiconductor 2.0* 2015 (http://itrs2.net)

[10] Hiraki K 2012 *Speculative Aspects of High-Speed Processor Design* (Tokyo: The University of Tokyo)

[11] IEEE 2021 *International Roadmap for Devices and Systems (IRDS)* (Piscataway, NJ: IEEE) (irds.ieee.org)