Scaling in Mechanical, Electronic, Magnetic, Ferroelectric, Optical and Electromagnetic devices

Charbel Tannous

Scaling in Mechanical, Electronic, Magnetic, Ferroelectric, Optical and Electromagnetic devices

C. Tannous

Université de Brest, Lab-STICC, CNRS-UMR 6285, F-29200 Brest, FRANCE

(Dated: January 30, 2023)

Miniaturization of Mechanical, Electronic, Optical, Electromagnetic devices and Storage media (Magnetic, Flash, Optical, Ferroelectric) is enhancing performance and portability of Computers, Multimedia, Information Storage and Telecommunication devices... Scaling, the most important notion underlying miniaturization, allows to understand how different building-block elements behave and how one should perform efficiently and safely this process.

PACS numbers: 07.10.-h, 42.79.Vb,84.40.-x, 85.30.-z,85.50.Gk
Keywords: Mechanical instruments,Optical storage devices,Radiowave technology,Semiconductor devices,Ferroelectric memories

Contents

I. INTRODUCTION

Mechanical, Electronic, Magnetic, Ferroelectric, Optical devices used in current devices such as Smartphones, e-tablets, Cameras, Telecommunication hardware... are constantly being miniaturized in order to make them more computationally and energy efficient while enhancing their portability to improve their handling and transportation...

Moore's law that specifies doubling of performance and cost decrease of electronic devices every eighteen months (corrected afterwards to two years) relies heavily on scaling since device characteristic size dictates its speed, energy consumption, response time...

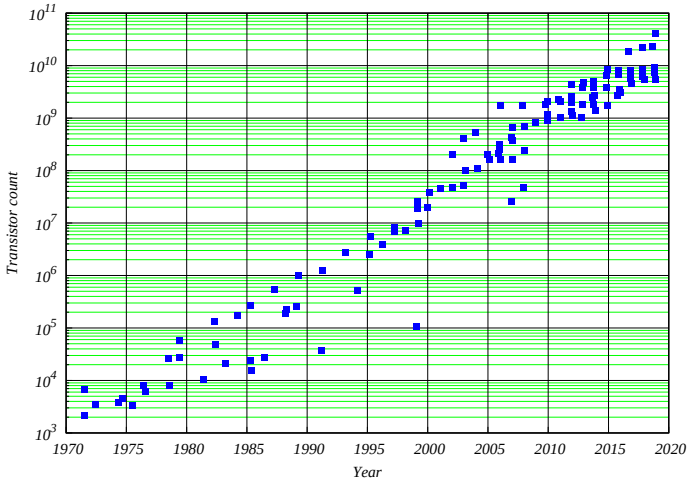

Fig.1: Moore's law illustrated with transistor count in a CPU chip from the beginning of circuit integration circa 1970 to present times. The chips originate from Intel, Apple, AMD, Motorola, ARM and SUN Microsystems. Data gathered by Hannah Ritchie and Max Roser under CC-BY Licence.

A practical measure of miniaturization is the minimum feature F (also called process node). It is the metric scale used by the foundry to control all sizes such as length, width and depth of various properties (gate, gate-oxide, contacts, current transport channels...) related to an individual device (transistor) fabrication or in making contacts between different devices (metal wire width and thickness) ... Its evolution with time is depicted in Table. 1.

10 μm	1971	6 μm	1974	3 μm	1977	1.5 μm	1981	1 μm	1984
800 nm	1987	600 nm	1990	350 nm	1993	250 nm	1996	180 nm	1999
130 nm	2001	90 nm	2003	65 nm	2005	45 nm	2007	32 nm	2009
22 nm	2012	14 nm	2014	10 nm	2016	7 nm	2018	5 nm	2020
3 nm	$\approx$ 2022	2 nm	$\approx$ 2024						

Table 1: Process node evolution and projection toward future (Intel and TSMC projections for 2022 and 2024).

Scaling allows to study precisely of the effect of miniaturization on the behavior of these devices that are made from solid-state components [1] while being metallic, insulating, semiconducting, electro-mechanical, ferroelectric, magnetoelastic... will behave differently under miniaturization.

In Electrical and Civil engineering, different effects on buildings, bridges, hydroelectric power plants... like seisms, vehicular traffic, vibrations... are studied with scale models due to the complexity of the underlying non-linear interaction mechanisms.

In this work we tackle scaling analytically from general first-principles unlike engineers in order to understand basic parameters involved in device miniaturization.
We have several types of scaling:

1. Mechanical scaling
2. Electronic scaling
    - Lumped element scaling

- Transistor (MOSFET) scaling
- Memory scaling
- SOC scaling

3. Optical scaling
4. Storage (Magnetic, Optical, Flash, Ferroelectric) device scaling

Scaling is affected by device type, functionality and architecture. For instance, CPU architecture is complex because of the variety of devices constituting it.

A. Logic families

The understanding of scaling, device type, functionality and architecture is based on logic families.
There are essentially three logic families:

1. Combinational: This family concerns: AND, NAND, OR, NOR, XOR... gates, Half-Adders, Full Adders, Encoders, Multiplexers... It is time-independent, based on Boolean arithmetic and the device output depends on the current input.
2. Sequential: This family can be viewed as combinational with memory and concerns: Flip-flops, Registers, Latches, RAM... (see fig. 2). Memory is created by introducing feedback in a combinational circuit inducing bistability. The device state is a function of time and its output depends on current and past inputs requiring storage of information. Registers, latches (used in CPU cache memory such as L1, L2, L3) rely on logical states to store information whereas DRAM rely on charge retained in capacitors (that are refreshed periodically in order to save energy consumption) to store information.
3. Conversion: This family concerns: Series to Parallel data converters, Samplers, Comparators, Quantizers, Bit-rate converters, Analog to Digital converters...

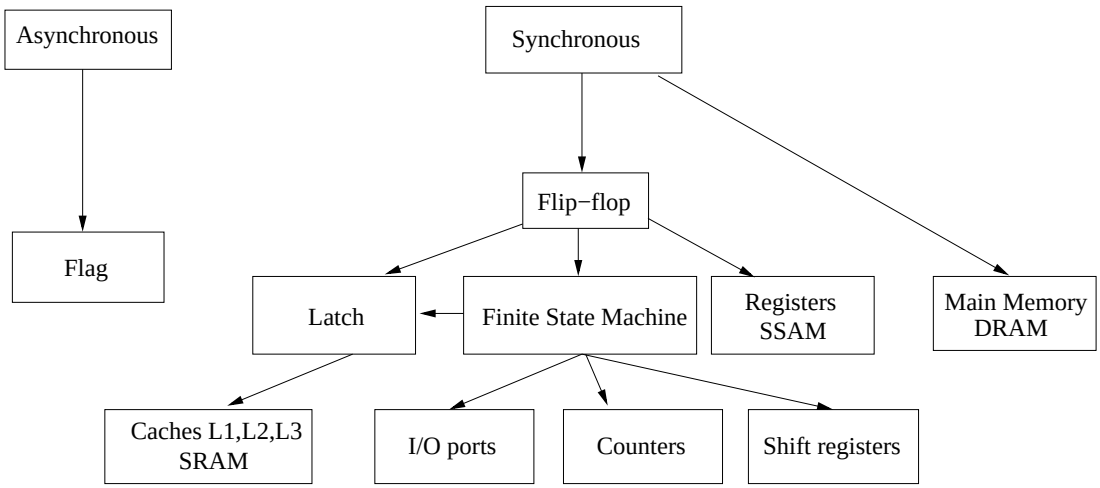

Fig.2: Sequential logic family tree. SRAM is Static Random Access Memory whereas SSAM is Static Sequential Access Memory. DRAM is Dynamic RAM (originating from the capacitor refreshment process). Memory types are Static (logic based) or Dynamic (Transistor and Capacitor based).

A CPU contains registers (called L_0 device where L is related to Latch device notion) accessed sequentially and made of flip-flops. On the other hand, a CPU contains a series of Cache memories (called L_1, L_2, L_3 where n is the level of accessibility of memory L_n.) that are accessed randomly. In addition, a CPU possesses an ALU (Arithmetic and Logic Unit) made of combinational logic circuits and an instruction decoder, controllers, program counter and stack pointer.

A CPU may have a CISC (Complex Instruction Set Computer) or a RISC (Reduced Instruction Set Computer) that is the backbone of ARM (Advanced RISC machine) used in SOC (System On Chip) the main processors of Smartphones.

In sharp contrast, a memory such as a DRAM (dynamic random access memory) situated outside the CPU die has a generic matrix architecture and consequently the simplicity of its structure speeds up its evolution as observed in Fig. 3. DRAM total volume depends on the number of addresses that the CPU can handle (cf. Memory scaling section).

A DSP (Digital Signal Processor) or an FPGA (Field Programmable Gate Array) belong to real-time devices such as controllers, communication ports, actuators, sensors... Their architecture is of the Harvard type with two buses instead of one carrying addresses, instructions and data similarly to Von Neumann type architecture used for CPU design. Since real-time processing needs speed, two buses are more efficient than a single one. Moreover the Harvard architecture possesses additional distinguishing features providing them with larger speed than Von Neumann machines.

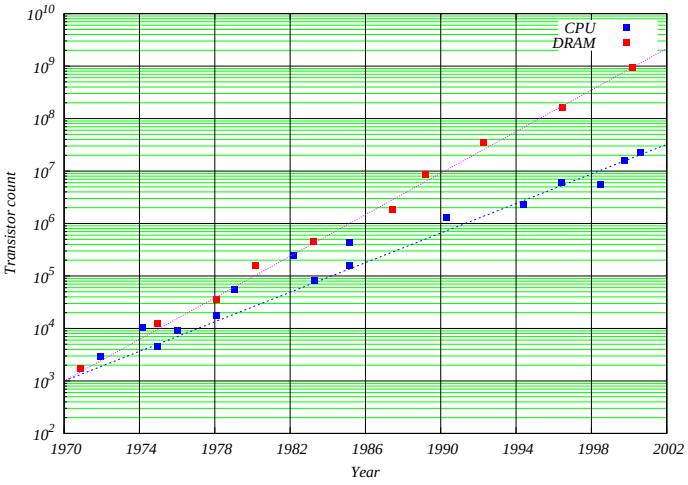

Fig.3: (Color on line) Evolution of transistor count for electronic CPU, and DRAM in Intel devices. The simplicity of DRAM architecture results in a number of transistors almost a hundred-fold higher in DRAM comparing to CPU. Adapted from Kilby [2].

II. MECHANICAL SCALING

In magnetic Hard Disk Drives (HDD), there are mechanical devices in their intrinsic sensors and actuators such as read-write heads and micro-motors acting on magnetic platter rotation. In Smartphones, pressure, accelerometers and Gyroscopes are important in many applications related to altitude, motion and orientation sensors...

Basic mechanical elements are the mass-spring system for translational motion, the pendulum for rotational motion and the cantilever whose deflection and resonance properties can be used in motion detection, accelerometers, Atomic Force Microscopy (AFM) [3], Magnetic Resonance Force Microscopy (MRFM) [3]...

It is possible to perform a scaling analysis for these elements in the same way it is done for RLC elements.

For a mass-spring system with mass m and spring constant k, mechanical performances i.e. the resonance angular frequency $\omega_0/2\pi$, Quality factor, signal-to-noise factor ... are functions of m, k whose values range from macroscopic to MEMS, NEMS and finally atomic.

While m is in the gram, kilogram values for macroscopic mass-spring systems, it is about 10^{-30} kg for an electron.

The spring constant k value ranges from more than several thousand N/m in macroscopic objects to less than 10^{-6} N/m in micro and nano objects. For a compression spring $k = YS/L$ where Y is Young modulus S, L are respectively the spring cross-section and length. Using scaling arguments, $L \sim \ell, k \sim \ell$ whereas $m \sim \ell^3$. This implies that resonance angular frequency $\omega_0 = \sqrt{\frac{k}{m}}$ scales as: $\omega_0 \sim 1/\ell$.

Typically materials have their elastic constants such as Young modulus, shear modulus as well as other elastic moduli given by ion pair interaction energy over volume occupied by the ion pair. Ion pair interaction energy is typically in the eV range and the volume is $1\mathring{A}^3$ yielding moduli in the 100 Giga-Pascal range or 10^{11} N/m^2.

Considering a macroscopic spring with $Y = 10^9$ N/m^2, S= 1 mm^2 and L=5 cm, we get: $k = 2.10^4$ N/m. In the case of MEMS and NEMS, the value of k can be made smaller than 10^{-6} N/m. As an example, a cantilever with L, b, h as length, width and thickness [4] (Fig. 4), has an effective spring constant [5] given by $k = 0.2575Yb\left(\frac{h}{L}\right)^3$ whereas fundamental angular frequency is: $\omega_0 = 1.03\frac{h}{L^2}\sqrt{\frac{Y}{\rho}}$. Y, ρ are Young modulus and mass density respectively. Thus it is possible by selecting geometrical parameters L, b, h and material parameters Y, ρ to achieve arbitrary small values of k and high angular frequency ω_0 since scaling laws [4] yield $k \sim \ell$ whereas $\omega_0 \sim 1/\ell$ as summarized in Table 2.

Nevertheless, it is surprising to find out that for an "atomic spring", when we make an analogy [3] between resonance angular frequency $\omega_0 = \sqrt{\frac{k}{m}}$ and typical atomic emission frequency such as 10^{13} rad/s with m about 10^{-25} kg, we get an order of magnitude for the "atomic spring constant" as $k = 10$ N/m.

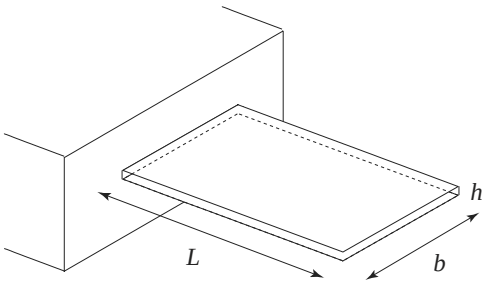

Fig.4: Cantilever with L, b, h characteristic lengths. Adapted from Lachut *et al.* [4].

ℓ	$\rightarrow$	ℓ/κ
	Mass-Spring system	
m	$\rightarrow$	$m/(\kappa)^3$
k	$\rightarrow$	k/κ
$\omega_0 = \sqrt{\frac{k}{m}}$	$\rightarrow$	$\omega_0\kappa$
	Pendulum	
L	$\rightarrow$	L/κ
$\omega_0 = \sqrt{\frac{g}{L}}$	$\rightarrow$	$\omega_0\sqrt{\kappa}$
	Cantilever	
$k = 0.2575b\left(\frac{h}{L}\right)^3 Y$	$\rightarrow$	k/κ
$\omega_0 = 1.03\frac{h}{L^2}\sqrt{\frac{Y}{\rho}}$	$\rightarrow$	$\omega_0\kappa$

Table 2: Summary of mechanical element scaling [4]. Notice that the mass-spring and the cantilever scale in the same way. $\kappa > 1$ is the length reduction factor. It is quite surprising that angular frequency scales as $\omega_0\kappa$ for the mass-spring system whereas $\omega_0\sqrt{\kappa}$ for the pendulum.

III. ELECTRONIC SCALING

Electronic families are divided in four categories: zero-polar, unipolar, bipolar and hybrid.

BBD (Bucket Brigade Devices) preceded CCD (Charge Coupled Devices) and both are considered as zero-polar, whereas MOS (Metal Oxide Semiconductor) and low power CMOS (Complementary MOS) devices are considered as unipolar since they are based on one type of semiconductors (n or p). A MOSFET is a Field-Effect Transistor (FET) based on the MOS structure with the metal (M) part plays the role of a grid separated from the semiconductor (S) by an insulating oxide (O) controlling the source-drain current in the transistor.
Bipolar family contains the BJT (bipolar junction transistor) based on the presence of both types of semiconductors (p-n-p or n-p-n...).
Hybrid families cover many types of devices, such as a mixture of unipolar and bipolar types such as Bi-CMOS to exploit large signals with bipolar devices and low consumption with CMOS. It could also be a mixture of analog and

digital devices, or a mixture of RF and low frequency devices...

A. Lumped element scaling

An electronic element such as R, L, C is considered as lumped when the frequency is so low that the applied Electromagnetic Field (EMF) wavelength λ applied to the element is such that $\lambda \gg \ell$ where ℓ is the typical length of the element.

In microwave circuits, electronic elements become distributed since we are in the regime $\lambda \lesssim \ell$ and $R(\boldsymbol{r}), L(\boldsymbol{r}), C(\boldsymbol{r})$ become space dependent.

Consider a typical length (such as the minimum feature or process node in electronics) ℓ and reduce it by a factor $\kappa > 1$. Since the resistance $R = \rho\ell/S$ we have a length scaling behavior $R \sim 1/\ell$ since the surface scales as ℓ^2.

Going from ℓ to ℓ/κ leads to $R \sim \kappa R$.

For a capacitance we have $C = \epsilon S/\ell$ thus $C \sim \ell$ and consequently C goes to $C \sim C/\kappa$ when ℓ changes to ℓ/κ.

Similarly for an inductance $L = n\mu\ell F(\ell/d)$ [6] giving a scaling behavior as $L \sim \ell$ like C. Hence when ℓ changes to ℓ/κ we have L going to $L \sim L/\kappa$.

As a consequence, the quality factor in electronic circuits $Q = RC/\omega$ transforms as $Q \to Q$ when ℓ goes to ℓ/κ. In contrast, when we are dealing with RF circuits containing inductances, $Q = L\omega/R$ and when ℓ goes to ℓ/κ, $Q \to Q/\kappa^2$ leading to strong reduction of the quality factor after miniaturization. This is why development of RFIC (smartphones, e-tablets...) took some time to overcome this problem. In analog radio systems $Q \in [10, 100]$ with an optimal value $Q = 50$.

In addition high-Q RF passive devices were located outside the main chip such as ceramic and SAW RF filters, SAW IF filters, Quartz Crystal resonators and other passive elements hampering seriously miniaturization.

To summarize we have the following table:

ℓ	$\to \ell/\kappa$
R	$\to \kappa R$
C	$\to C/\kappa$
L	$\to L/\kappa$
$Q = RC/\omega$	$\to Q$
$Q = L\omega/R$	$\to Q/\kappa^2$

Table 3: Summary of lumped element scaling.

In the first miniaturization attempts of RFIC devices Q was about 10^{-5}. The flat spiral suspended design of inductances allowed to reach a reasonable value of $Q > 5$ (cf. Fig. 5) that was enhanced additionally with digital techniques.

Inductor miniaturization in RFIC devices allowed shrinking Telecommunication devices dramatically as depicted in Fig. 6.

The first hand held cellphone, the DynaTAC 8000x was commercialized in March 1984 by Motorola. The phone dimensions were 13 x 1.75 x 3.5 inches, with a weight of almost 2 pounds with a price of of $ 4,000 equivalent to $ 9,000 today. The phone was nicknamed "The Brick" because of its size and weight.

B. Transistor scaling

Evolution with time of the Transistor process node is depicted in Fig. 8 along with its relation to resulting transistor density.

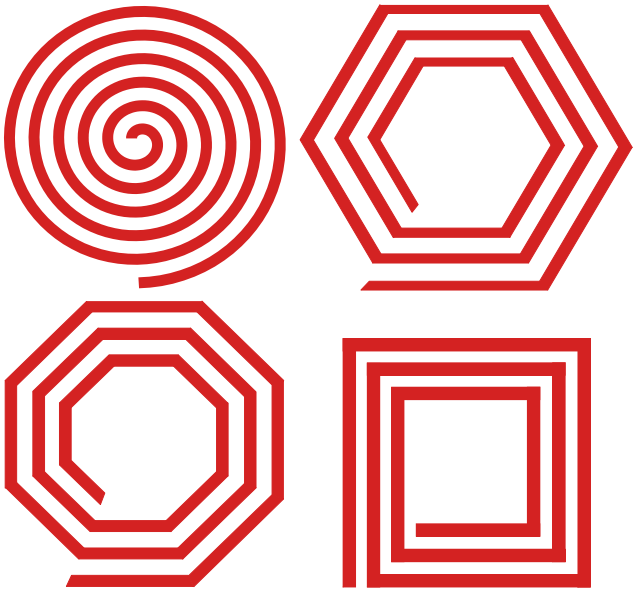

Fig.5: Flat inductors [7] with circular, hexagonal, octagonal and square shapes are used to increase Q when lying over an insulator substrate or simply suspended over it in order to decrease losses.

1. Gate-Oxide thickness

Silicon dioxide has been used traditionally as a gate oxide material for electric insulation. Shrinking the transistor implied reducing the dioxide gate dielectric thickness (cf. fig. 7) in order to increase the gate capacitance and thereby drive current and device performance. When gate oxide thickness is thinner than 2 nm, quantum tunneling induces leakage currents across the oxide decreasing device performance. If the silicon dioxide is replaced by a high-dielectric constant (called kappa by electronic engineers) material gate capacitance increases without current leakage.

The dielectrics used by several electronic foundries belong to a family of Hafnium silicates such as HfO_2 and HfSiO. Unfortunately, Hafnium silicates are susceptible to trap-related leakage currents and when Hafnium concentration is increased, performance decreases.

2. Multigate devices

These devices belong to the Multigate class tailored to improve scaling and reduce leakage between source and drain (This occurs when the gate becomes too thin) with fins, nanowires or nanoribbons arranged perpendicularly to the gate (cf. fig. 7). While fins are perpendicular to both gate and gate-oxide, nanowires and nanoribbons are perpendicular to gate but parallel to gate-oxide. Presently the MOSFET is replaced by the FinFET and in the future it will be the Ribbon FET with a double digit naming by Intel such as 20A meaning the process node is 20Å. Ribbon FET is, in fact, a version of the GAA (Gate-All-Around) transistor where nanoribbons replace the nanowires. Samsung version of the Ribbon FET is called the MBCFET (Muti-Bridge Channel) (cf. fig. 7).

Comparing process node (minimum feature) evolution with time to transistor density is depicted in fig. 8.

C. Memory scaling

The transition from 32-bit to 64-bit CPU enhanced tremendously CPU processing. For instance a 32-bit CPU can handle 2^{32} addresses or roughly 4GB memory volume assuming that for every address we have 1 Byte is available. Since currently we have 64-bit CPU's we get a gigantic value equal to 2^{64} addresses (about 16 $\times 10^{18}$ or 16 billion[2]). With such a volume, one can load the operating system (typically several GBytes), applications and files in DRAM and work a million times faster since CPU-DRAM access time is about a million times shorter than any other one related to CPU access to HDD (Hard Disk Drive), SSD (Flash Solid State Disk) or to other storage media types.

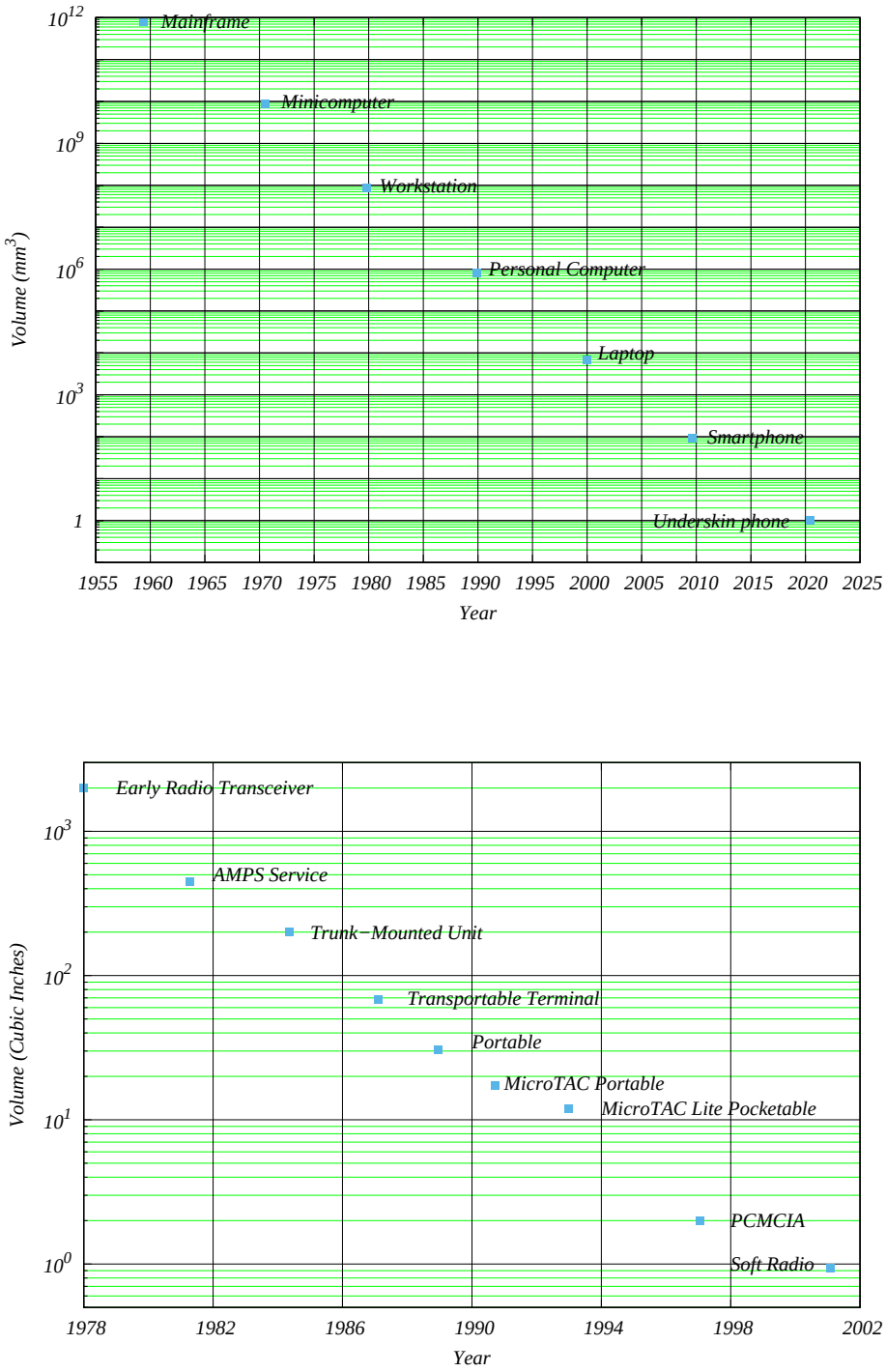

Fig.6: (Color on line) Above, comparison of computer and cellphone evolutions with the ubiquitous underskin phone. Below, projected size decrease of cellphones [8] from the 1970's when a radio transceiver had the size of a large building while AMPS services occupied volumes equivalent to small buildings. The portable telephone started around 1984 with Motorola DynaTAC 8000x.

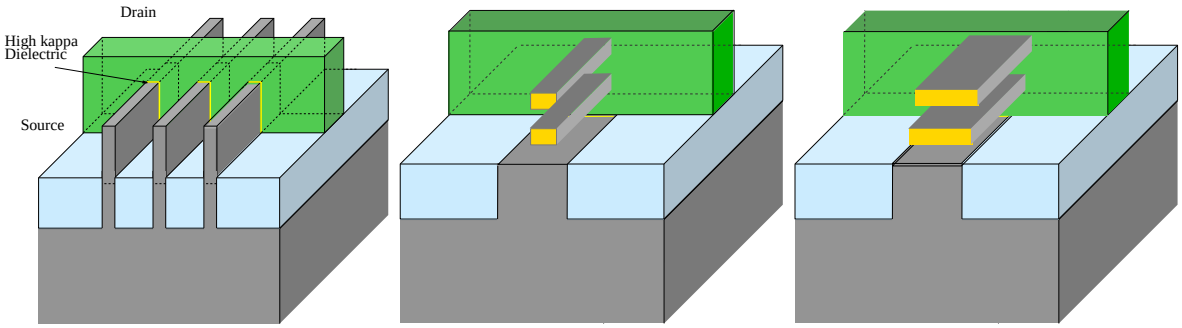

Fig.7: (Color on line) Architecture of the 3D FinFET, GAAFET (Gate-All-Around) and MBCFET (Muti-Bridge Channel) devices. Three fins linking source to drain are displayed in the FinFET, whereas two nanowires are shown in the GAAFET and two nanosheets (ribbons) are used in the MBCFET. Adapted from Tech-Brief [9].

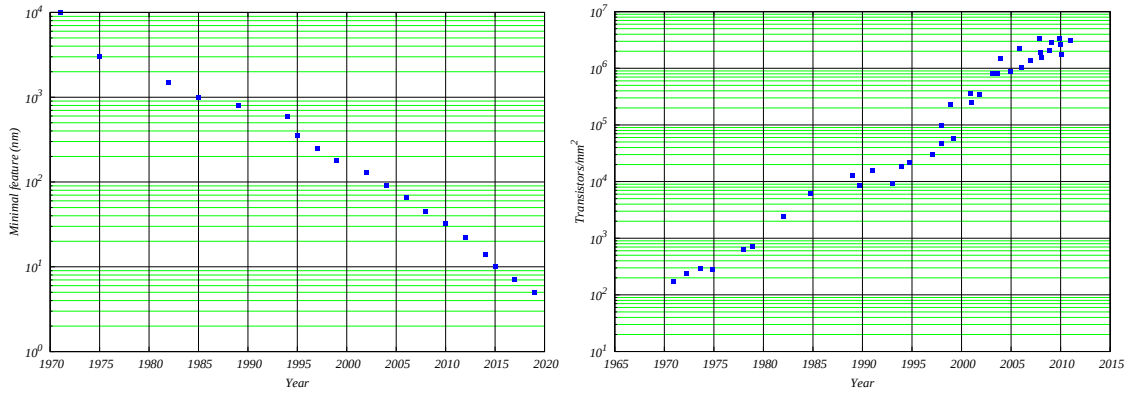

Fig.8: (Color on line) Variation of process node (minimum feature) with time and comparison to transistor density evolution from several chip foundries (Intel, AMD, IBM and Motorola).

Parameters selected for Scaling	Constant E Scaling	Generalised Scaling
t_{ox}, L, W, X_j, W_d	$\frac{1}{\kappa}$	$\frac{1}{\kappa}$
N_a, N_d (ions/cm^3)	κ	$\alpha\kappa$
Power supply V_{dd}	$\frac{1}{\kappa}$	$\frac{\alpha}{\kappa}$
Electric field in device E	1	α
Capacitance C	$\frac{1}{\kappa}$	$\frac{1}{\kappa}$
Inversion charge density Q	1	α
Circuit delay time: $\tau = \frac{C_{gate}V_{dd}}{I_{dsat}}$	$\frac{1}{\kappa}$	$\frac{1}{\kappa}$
Chip Area A	$\frac{1}{\kappa^2}$	$\frac{1}{\kappa^2}$
Power dissipation P	$\frac{1}{\kappa^2}$	$\frac{\alpha^2}{\kappa^2}$
Power density ($\sim P/A$)	1	α^2
Circuit density	κ^2	κ^2
Drift current I	$\frac{1}{\kappa}$	$\frac{1}{\kappa}$

Table 4: Summary of the constant field scaling and the generalised scaling rules. α is the scaling factor for field and potential. C_{gate} is transistor gate capacitance, V_{dd} is power supply applied to drain and I_{dsat} is drain saturation current. Adapted from Iwai [10].

Memory types are Static (logic based such as in latch devices L_n within CPU), Dynamic (Transistor and Capacitor based such as in DRAM outside CPU) and Mass (storage devices that are different from standard memory usually considered as "live").

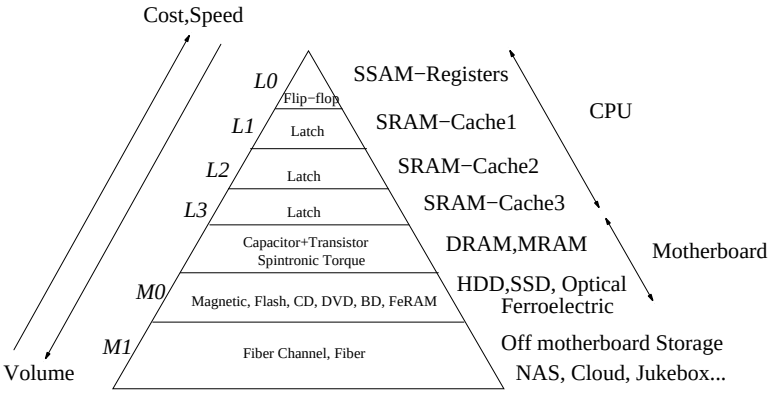

Fig.9: Memory hierarchy with flip-flop registers at the top with serial access. As one goes down the pyramid, storage volume increases and speed decreases.

In traditional DRAM design, $8F^2$ cell design and a folded bit-line array have characterized mainstream DRAM architecture. $6F^2$ cell design is also used, providing about a 25% improvement in DRAM cell area (see Table 6).

Following Choi [11], the following tables, make detailed comparisons between $6F^2$ and $8F^2$ architectures based on same process nodes (Table 5) or different ones (Table 6) in order to highlight the implications and stakes involved in their manufacture.

Cell structure	$6F^2$	$8F^2$
Density	512 Mbits	512 Mbits
Process node	80 nm	80 nm
Cell size	0.038 μm^2	0.05 μm^2
Chip size	49 mm^2	56 mm^2
Number of metal layers	3	3
Gross dice (12 in. wafer)	1266	1100

Table 5: Comparison of typical $6F^2$ and $8F^2$ DDR2 DRAM designs with same 80 nm process node [11].

Cell structure	$6F^2$	$8F^2$
Density	512 Mbits	512 Mbits
Process node	80 nm	90 nm
Cell size	0.038 μm^2	0.076 μm^2
Chip size	49 mm^2	72 mm^2
Wordlines/[array block]	320	512
Cells per bit-line	160	256
Gross dice (12 in. wafer)	1266	862

Table 6: Comparison of $6F^2$ and $8F^2$ DDR2 DRAM designs with different process nodes [11]: 80 nm and 90 nm.

The array block, a building block having cell array and bit-line sense has 320 wordlines per array block, a reduction from the 512 wordlines per block in 90 nm $8F^2$-based design.

Traditionally, array blocks provide total wordline as powers of 2 (128 (2^7), 256 (2^8) or 512 (2^9)). In sharp contrast, $6F^2$ design has 320 wordlines departing from a power of 2.

Table 5 shows dramatic improvement in terms of number of gross dice per wafer by comparing $6F^2$ and $8F^2$ designs, nevertheless device geometry scaling (90 nm to 80 nm) along with cell-array architecture change ($8F^2$ to $6F^2$) produce some non trivial results.

The comparison of cell size between the different designs demonstrates the trade-offs of $6F^2$ DRAM cells: a 24 percent reduction in cell size. But the impact on chip size (although other factors may affect the chip size as well, assuming the periphery designs are done similarly) is only about half the result due to $6F^2$ cell design. The effectiveness of $6F^2$-cell-based DRAM is reduced by the additional design challenges associated with an open bit-line architecture.

Careful selection of the array block is the key to optimum design of $6F^2$-cell-based DRAM products.

The overall gain of $6F^2$-cell-based designs in gross dice per wafer for a 12-inch production line is estimated to be around 15 percent.

Although the $6F^2$ cell's advantage of 24 percent less area is mitigated by a die-count increase of only 15 percent more dice per wafer, this increase in the number of gross dice is nevertheless essential to the manufacturing process.

D. SOC scaling

In e-tablets, smartphones... CPU has been replaced by SOC (System On Chip) that embed Memory, Logic, Photonic, Wireless and Wired Communication and Sensor circuits on the same chip.

Thus DSP circuits, Communication Circuits, tactile sensing, inertial devices (2D or 3D accelerometers and gyroscopes), magnetic sensing devices (Hall...) can be gathered on the same chip (2D integration) or piled up (3D integration).

SOC circuits are based on ARM architecture that is fast and reliable since a RISC processor uses a relatively small set of instructions reducing fetch and decoding of instructions.

IV. OPTICAL AND ELECTROMAGNETIC SCALING

Microwave circuits are being coupled to optical devices in many areas of Science and Engineering and particularly in Telecommunications.

As an example, LIDAR is replacing RADAR in autonomous (self-driving) vehicles since its wavelength is 1/1000 smaller and is being constantly improved to tackle longer detection ranges ($>$ 100 m).

Presently, there are at least two platforms: Si and III-V semiconductor [12] to deal with OIC (Optical Integrated Circuits) and OEIC (Opto-Electronic Integrated Circuits).

Silicon platforms comprise silica-on-silicon, silicon-on-insulator (SOI), and silicon nitride-on-silica. Waveguides, for instance, are made with silica-on-silicon and are easily interfaced with free-space optics and fibers to make quantum gates for quantum computation logic devices.

III-V platforms based on GaAs and InP offer solid-state on-chip laser integration of highly tunable pump sources with electrical injection. 3D heterostructures confinement of both electrons and holes make quantum dots (artificial atoms) behave like single-photon sources with near-unity internal efficiency.

Several issues or fields of interest are impacted by this hybrid technology such as:

- MOEMS: Micro Opto-Electro-Mechanical Systems

- Electromagnetic cavity design
- Magneto-Optical traps
- Nanophotonics
- Optical switches
- Wafer-scale phased-array antennas
- Wafer-scale phased-array Tx-Rx
- Wafer-scale phased-array Radars and Lidars

Wavelength (λ) is the main control parameter in Optical and Electromagnetic scaling since it determines not only the nature of wave propagation in comparison with to observation-propagation distance d ($\lambda \gg d$: Line of sight (straight propagation or geometrical optics) regime, $\lambda \lesssim d$: Spherical (radio propagation or physical optics) regime but also the diffraction limit. This is similar to the distinction we made previously between lumped and distributed electrical elements where Electromagnetic Field (EMF) wavelength λ is compared to physical size of the element.

Diffraction limit is given by $\frac{\lambda}{n \sin\theta}$, for coherent transmission and $\frac{\lambda}{2n \sin\theta}$ for incoherent transmission where n is the propagation medium refractive index and θ is half the angle subtended by the objective lens. The angle is taken between focal distance taken perpendicularly to lens mid-plane and lens radius taken along the latter. $n \sin\theta < 1$ is called the numerical aperture of the objective lens and its square is a measure of the light-gathering capabilities of the lens [13].

A. Opto-electronic device scaling

While a CCD does not obey scaling laws, the equivalent CMOS device has known tremendous development in its performance and scaling properties specially in sensors, e-tablets, smartphones, laptops, and desktop computer cameras.

Sanguinetti *et al.* [14] used uniform illumination of smartphone camera image sensor by a LED to produce a number of photons generated per pixel.

The CCD image sensor has 16 bits per pixel (detection capability) and a photon flux producing 2×10^4 electrons per pixel, whereas the CMOS image sensor had a smaller photon flux since it has 10 bits per pixel with only 500 electrons per pixel (cf. Table 7).

Smartphone	ATIK 383L	Nokia N9
Noise, σ_n (e^-)	10	3.3
Saturation (e^-)	2 $\times 10^4$	500
Illumination (e^-)	1.5 $\times 10^4$	410
Quantum uncertainty, σ_q (e^-)	122	20
Offset (e^-)	144	-6
Output bits per pixel	16	10

Table 7: Experimental parameters for two smartphone cameras employing a CCD (ATIK 383L) or a CMOS (Nokia N9). All data except last line, are expressed in number of electrons (e^-). Adapted from Sanguinetti *et al.* [14].

From the scaling point of view, a CCD is considered as non scalable, belonging to zero-polar family of devices, whereas MOS and CMOS are the archetypical unipolar family of the Moore type scalable device.

V. STORAGE DEVICE SCALING

A. Magnetic scaling

The analogue of Moore law in magnetic storage is Kryder law and it is illustrated in Fig. 10 with areal density in Gb/in^2 per year. Remarkably the transition from LMR (Longitudinal Magnetic Recording) to PMR (Perpendicular

Magnetic Recording) occurs smoothly respecting Kryder law without any discontinuity.

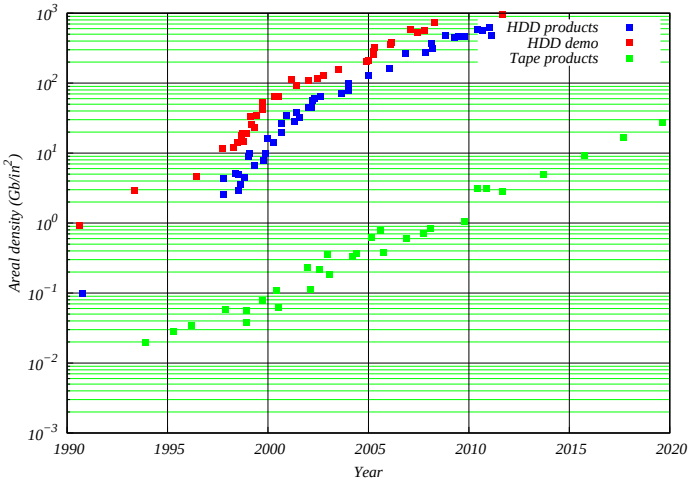

Fig.10: (Color on line) Areal density evolution for HDD (products and demos) as well as tapes. The transition from LMR to PMR in HDD demos occurred about 2002.

Areal density d controls scaling laws impacting several parameters of LMR and PMR. Marchon *et al.* [15] derived the following rules for areal densities d spanning an interval ranging from 0.1 Gbit/in^2 to beyond 1000 Gbit/in^2:

- Bit size (in nm) $\approx 167.01\, d^{-0.3715}$, d in Gbit/in^2
- Bit Aspect Ratio $\approx 23.296\, d^{-0.2598}$
- Track width (in nm) $\approx 3890.7\, d^{-0.6313}$
- Head Media Separation (HMS) (in nm) $\approx 71.5\, d^{-0.317}$

Additionally, we have an approximate relation [15] between HMS and bit size, expected to be valid even when the bit size is smaller than 10 nm [16]: HMS $\approx$ (bit size)/2 as long as $10 <$ bit size (nm) < 1000.

B. Optical scaling

From CD to DVD to Blu-Ray how do we scale: Laser spot, pit separation, track separation or pitch, Numerical Aperture of the objective lens?

In optical devices, the simplest scaling procedure pertains to the transition from CD to DVD, HD-DVD and finally Blu-Ray or BD by decreasing the laser wavelength.

C. Flash scaling

Flash technology is important for mobile electronic devices (e.g. digital camera, smartphone, portable SSD drives, key drives, etablets and notebooks).

A Flash transistor is a MOSFET transistor containing within its Oxide layer a thin metallic film called floating gate (hence the name flash) than can trap electrons by application of a positive voltage to the gate.

At the beginning of Flash technology, a single bit corresponded to a single Flash transistor (SLC or single level cell). Later development led to 2-bit/transistor (MLC or Multiple level cell), then 3-bit/transistor (TLC or triple

	CD	DVD	HD-DVD	BD
Laser wavelength	780 nm	650 nm	405 nm	405 nm
Laser spot diameter	1.6 μm	1.1 μm	620 nm	480 nm
Smallest pit length	800 nm	400 nm	200 nm	150 nm
Pit width	600 nm	320 nm	200 nm	130 nm
Pitch value	1.6 μm	740 nm	400 nm	320 nm
Storage capacity/layer	780 MB	4.7 GB	15 GB	25 GB
Numerical Aperture	0.45	0.6	0.6	0.85

Table 8: Laser, land and pit characteristics for CD, DVD, HD-DVD and BD. Pitch is the distance between neighbouring tracks. Numerical aperture square is a measure of the light-gathering capabilities of the lens [13] needed for laser focusing over optical disks.

	NOR	NAND
Cell size	8-12F^2	$4F^2$
Read Performance	Fast ($\sim 10\mu$sec)	Slow ($> 200\mu$sec)
Storage Information	Code	Data

Table 9: Major Differences between NOR and NAND Flash memory.

level cell) and finally 4-bit/transistor (QLC or quadruple level cell).

NAND Flash memory is the most aggressively scaled technology among electronic devices because the NAND Flash memory cell have relatively simple structure compared with others along with a very strong competition between foundries. Also, several technological innovations occurred (such as multi-level cell and double patterning lithography) leading to cell size shrinking faster than dictated by Moore's law. This dramatic cell size scaling resulted in a radical price drop of flash memory and growing interest in NAND technology.

More importantly, the price drop of storage media brought us new innovative electronics and NAND Flash memory market has grown tremendously to more than 40% per year.

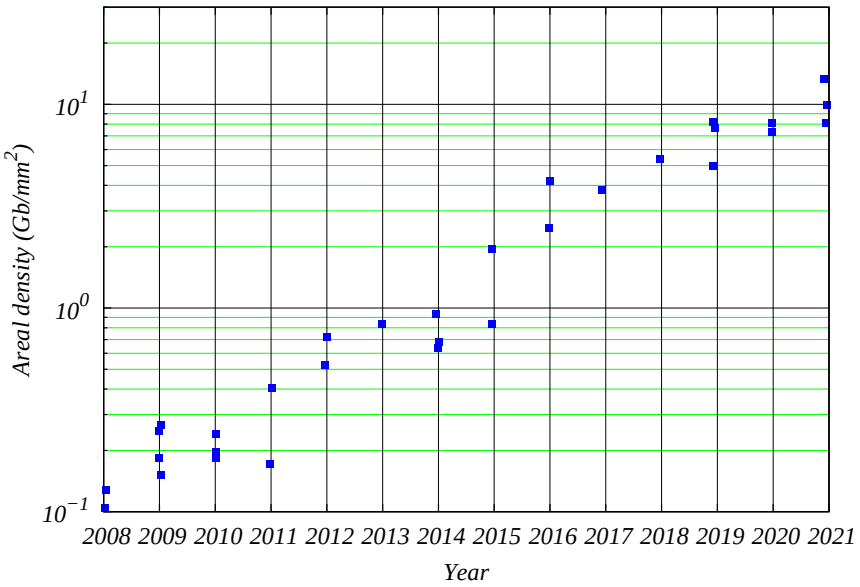

Fig.11: (Color on line) Evolution of Flash areal density over the period 2008-2021. The data contain 2D and 3D MLC, TLC and QLC Flash types. The transition from 2D to 3D started circa year 2014.

1. Limitations of conventional Flash memory technology

Conventional Flash memory is facing technological barriers as it scales to sub 20 nm nodes. These barriers are rather fundamental, in view of key cell features:

Firstly, additional channel length scaling (without gate-oxide scaling) causes poor I_{on}/I_{off} ratio of the cell, which is not enough for read operation.

Secondly, no physical space for the control gate below 20 nm technology is expected because NAND Flash cell using 25 nm technology has already about 10 nm space for the control-gate material. It also results in a significant cell-to-cell interference problem. More importantly, the number of stored electrons in the floating gate decrease as the floating gate size shrinks.

D. Ferroelectric scaling

In sharp contrast to magnetic and Flash, ferroelectric scaling does not behave as desired.
In magnetic scaling, the moment is not affected whereas an electric dipole moment $\boldsymbol{p}$ is directly reduced by scaling since $\boldsymbol{p} = q\ell$.

VI. EMERGING MASS MEMORY TECHNOLOGY

A. Cross-point Memory

A lot of alternative memory device designs and new materials have been proposed to overcome the scaling limits for conventional Flash memory technology. Cross-point memory architecture is among the most attractive successors to the current Flash technology because of following reasons: it allows for the most compact storage with $4F^2$ cell layout size (where F is the minimum half-pitch) and also can be fabricated using a relatively simple process that is more amenable to 3D integration.

B. Phase-Change Memory: Re-RAM

Programmable resistance devices such as phase-change memory and resistive RAM have been explored for cross-point memory applications. However, those emerging memory devices generally require a selector device within each memory cell to reduce unwanted leakage current through unselected cells during a read operation; otherwise, the size (number of rows/columns) of the array will be severely limited, resulting in poor memory-array area efficiency. However, the selector devices require additional process steps and can significantly reduce the cell current, resulting in slower read operation.

C. Phase-Change Memory: Topological

A novel possibility is opening with Mott Topological devices based on Metal-Insulator phase transition. Some of the issues to solve are the critical temperature value and the speed of the phase transition.

VII. CONCLUSION AND PERSPECTIVES

A major paradigm change occurred in the chip industry with the TCPA (Trusted Computing Platform Alliance) Palladium agreement abolishing the retro-compatibility principle of computer science and allowing coupling between hardware and software. Questions are asked about its sustainability since one is forced to constantly acquire new hardware in order to adapt to a software changing always at a faster rate.

Another paradigm change in Telecommunication is presently occurring with the deployment of the 5G standard. The Internet of Things is a powerful concept allowing to exchange information remotely between customers and their appliances, cars... Questions are about its extensive privacy invasion and sustainability given the massive amount of data generated by the 5G standard to process, communicate and store through data centers, requiring excessive

amount of energy and storage facilities.

Another paradigm change in Material Science is also currently happening with the swift development of Quantum and Topological materials opening opportunities in faster processing, phase-change storage, disspationless transport... requiring novel approaches, protocols and devices along with their scaling laws and physics.

References

[1] C. Kittel, *Introduction to Solid State Physics*, Wiley, New-York (1975).
[2] J.S. Kilby, *Nobel Lecture* (2000).
[3] D. Sarid, *Scanning Force Microscopy*, Oxford University Press (1994).
[4] M. J. Lachut and J. E. Sader, Phys. Rev. Lett. **99**, 206102 (2007).
[5] A. Suter, *The magnetic resonance force microscope*, Progress in Nuclear Magnetic Resonance Spectroscopy (Elsevier), **45**, 239, (2004).
[6] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media*, Pergamon, Oxford (1975).
[7] T.H. Lee, *The design of CMOS radio-frequency integrated circuits*, Cambridge University Press, Cambridge (1998).
[8] W. F. Brinkman and D. V. Lang, *Physics and the communications industry*, Reviews of Modern Physics, **71**, S480, Centenary 1999.
[9] N. Draeger, *Tech Brief: FinFET Fundamentals*, September 12, (2016).
[10] H. Iwai, *Roadmap for 22 nm and beyond*, Microelectronic Engineering (Elsevier), **86**, 1520 (2009).
[11] Y Choi, *Under the hood: DRAM architectures* $8F^2$ *vs* $6F^2$, EE-Times, February, 2008.
[12] S. Bogdanov, M. Y. Shalaginov, A. Boltasseva and V. M. Shalaev, *Material platforms for integrated quantum photonics*, Optical Materials Express, **7**, 111 (2017).
[13] E. Hecht, *Optics*, Fourth Edition, Pearson Education-Addison-Wesley, San Francisco (2002).
[14] B. Sanguinetti, A. Martin, H. Zbinden and N. Gisin, Phys. Rev. X **4**, 031056 (2014).
[15] B. Marchon and T. Olson, IEEE Trans. Magn. **45**, 3608 (2009).
[16] T. Wang, V. Mehta, Y. Ikeda, H. Do, K. Takano, S. Florez, B. D. Terris, B. Wu, C. Graves, M. Shu, R. Rick, A. Scherz, J. Stohr and O. Hellwig, Appl. Phys. Lett. **103**, 112403 (2013).