



Project Number: FP7-611404

# D2.1 - Report on future technologies that may be used in future computer systems

**Authors**[1]

S. Ozdemir (INTEL), R. Canal (UPC), M. Kaliorakis (UoA), S. Tselonis (UoA), N. Foutris (UoA), D. Gizopoulos (UoA), A. Grasset (THALES), G. Rafiq (ABB), T. Loekstad (ABB)

Version 1.2 – 31/10/2014

| |
|---|
| **Lead contractor:** UPC |
| **Contact person:**<br>Ramon Canal<br>Dep. of Computer Architecture<br>Universitat Politècnica de Catalunya<br>Campus Nord UPC, Cr. Jordi Girona 1-3, 08034 Barcelona (ES)<br>E-mail: rcanal@ac.upc.edu |
| **Involved Partners**[2]**:** UoA, INTEL, THALES, YOGITECH, UPC, POLITO, ABB |
| **Work package: WP2** |
| **Affected tasks: T2.1** |

| | | | | |
|---|---|---|---|---|
| **Nature of deliverable**[3] | R | P | D | O |
| **Dissemination level**[4] | PU | PP | RE | CO |

[1] *Authors listed here only identify persons that contributed to the writing of the document.*

[2] *List of partners that contributed to the activities described in this deliverable.*

[3] ***R**: Report, **P**: Prototype, **D**: Demonstrator, **O**: Other*

# COPYRIGHT



[4] ***PU**: public, **PP:** Restricted to other program participants (including the commission services), **RE** Restricted to a group specified by the consortium (including the Commission services), **CO** Confidential, only for members of the consortium (Including the Commission services)*

# INDEX

# Scope of the document

This document is an outcome of task T2.1, "***Reliability failure mechanisms for future systems***", elaborated in the description of work (DoW) of the CLERECO project under the Work Package 2 (WP2).

Figure 1 graphically depicts the goal of this deliverable, its inputs and main results and which work packages will use its outputs.

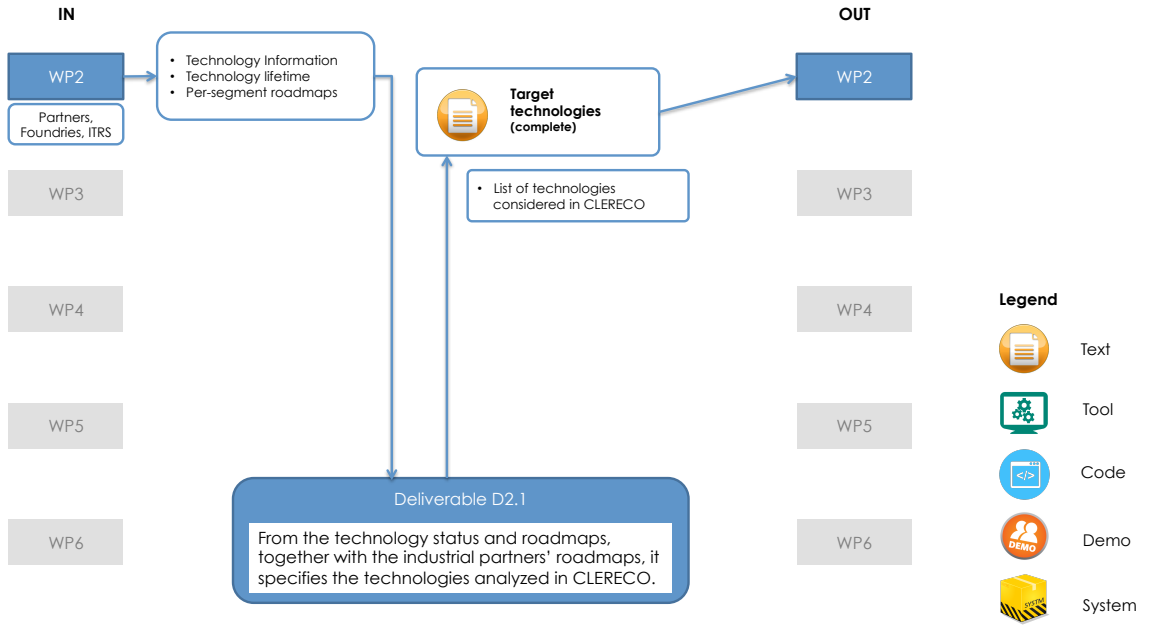



**Figure 1: Deliverable summary**

D2.1 aims at identifying the *technologies employed in future computer systems*. The activity described in this deliverable is a survey activity that is fundamental for the CLERECO project and represents a base for several other activities within the project. The technologies identified in this deliverable will be characterized in WP2 to identify their major failure mechanisms and to provide a knowledge base to exploit in WP3, WP4 and WP5 activities. This characterization is part of deliverables D2.2.1 and D2.2.2 (Characterization of failure mechanisms for future systems).

With the term *technologies,* we refer to different material processing techniques, leading to the development of transistors and other structures with unique characteristic properties. In this document, we describe the relevant foreseen technologies. They include the technologies that are of interest in the short-term ahead as well as proposals/technologies that are more revolutionary (or the emerging technologies that will remain in the market in the long-term). Based on the maturity of each technology, we select the ones within the scope of CLERECO.

The document is organized in the following sections:

- **Introduction**. This section sets the background for the document. The objectives of the document and the investigations made for its development are included.
- **Technology requirements**. This section identifies the technological needs of the different computing segments.

- **State of the art**. This section provides information about the technologies available today, as well as the advances in the near future that meet the CLERECO time frame of analysis.
- **Technologies beyond the scope of CLERECO**. This section identifies those technologies that do not meet the CLERECO objectives and, thus, are not foreseen to be much influential for the next generation of computing systems.
- **Conclusions**. Finally, we draw the main conclusions of this deliverable and the impact on the CLERECO project.

# 1. Introduction

System reliability has become an important design aspect for computer systems due to the aggressive technology miniaturization, which introduces a large set of different sources of failure for hardware components. Unreliable hardware components affect computing systems at several levels. Raw errors are strongly related to the technology used to build the hardware blocks composing the system and are caused by effects such as physical fabrication defects, device aging/degradation (e.g., NBTI, HCI, etc.), environmental stress (e.g., radiations), etc.

Any system nowadays is built on top of a silicon substrate. Understanding the technology foundations of this silicon substrate and their evolution into new forms of transistors through new materials is fundamental to the construction of future computing platforms. Therefore, the role of the underlying technology in the overall system reliability is carefully considered in CLERECO.

The portion of the system reliability stack considered in WP2 is shown in Figure 2. D2.1 is the first step in this direction and analyzes the technologies that are currently available and those that are foreseen and it selects the most promising ones to be used in the upcoming years.

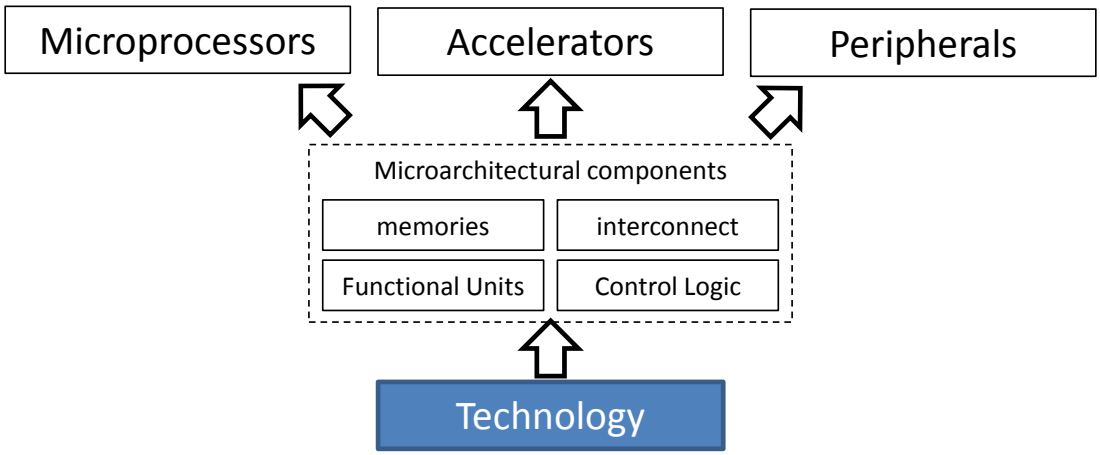



**Figure 2: The portion of the system reliability stack considered in this deliverable**

# 2. Technology requirements of the computing continuum

## 2.1. High Performance Computing

HPC computer systems have experienced a tremendous speed-up in performance during the last decades. If current trends in performance continue, exascale systems (systems 1000x

faster than today's) will be a reality by 2020. However, without breakthroughs in technology, this is highly unlikely to happen. Current systems have recently hit a power wall and have reached their limits. Not only high-performance computing platforms are affected but also mobile and desktop systems, as we will see in the next sections.

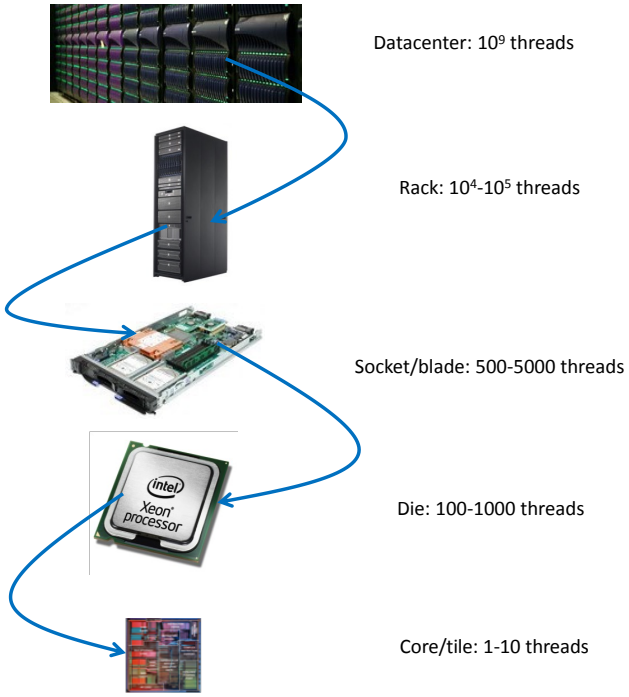



**Figure 3: Exascale HPC breakdown**

Figure 3 depicts the expected exascale system and its expected amount of components. To enable exascale performance, technologists and architects face two key challenges: power and memory performance. Power has to be cut down dramatically (around 400x) while memory bandwidth needs to grow around 50x as well. This power reduction is not achievable just with technology scaling: lower voltage levels will be required. However, lower voltage levels will result in a tremendous increase in reliability problems with underlying technologies. The issue of reliability poses a tremendous threat to the possibility of exascale systems. While reliability, in general, has gathered some attention in recent years for current technology, the new materials foreseen and the low voltage resiliency aspects in the exascale territory are still unexplored.

Memory systems account for 75% of the power budget for the CPUs and main memory [***1***]. Moreover, their design determines the memory bandwidth. Focusing research on memory systems has the potential to yield key advances regarding both the power and memory performance challenges. These memory systems are essential to exascale computing. Innovation in this area will determine computing performance increases over the next decade.

HPC systems are used in a variety of scenarios. Table 1 summarizes the availability (i.e., what percentage of time the machine is on) for different computer segments. It can be clearly seen that HPC systems (in this case banking, medical and defense) can be just 31.5 seconds "off" every year. This is a tremendous challenge when bearing in mind that, nowadays, the mean time between failures (MTBF) is measured in days. Should the resiliency of the components re-

main untouched, considering that we will have on the order of a million more components, this leads to an MTBF of a couple of minutes. Considering that the simplest checkpoint in such a system takes 30 minutes, it yields the whole system useless (i.e., there is no time to backup before another failure occurs).

**Table 1: Availability for different computer segments [2]**

| 9's | Availability | Downtime/Year | Examples |
|---|---|---|---|
| 1 | 90.0 % | 36 days, 12 hours | Personal Computers |
| 2 | 99.0 % | 87 hours, 36 min | Entry Level Business |
| 3 | 99.9 % | 8 hours, 45.6 min | ISPs, Mainstream Business |
| 4 | 99.99 % | 52 min, 33.6 sec | Data centers |
| 5 | 99.999 % | 5 min, 15.4 sec | Banking, Medical |
| 6 | 99.9999 % | 31.5 seconds | Military, Defense |

Consequently, HPC systems will make use of the newest technology available to meet the specification requirements.

## 2.2. General Purpose and Mobile Computing

General purpose and mobile systems (personal computers) have different market characteristics than the HPC systems. While availability is of prime concern, it does not get to the level of an HPC system (as Table 1shows). On the other side, these two markets are very competitive and thus cost is a big concern. This means that these market segments will not make use of the latest technology but just close to the latest.

In particular, on the mobile computing segment, battery life is also of prime concern. Thus, we expect to see specific "low power" versions of any technology to feed this market. These versions will be an evolution (and refinement) of pre-existing ones. Therefore, we expect some delay between the HPC segment adopting a technology and its low power counterpart.

## 2.3. Industrial and Safety/Mission Critical Embedded Systems

The majority of devices used in the industrial market (as well as the safety/mission critical systems) are based on embedded systems consisting of Programmable Module Controllers (PLC), input-output devices, actuators, and sensors. The embedded market is conservative with respect to acceptance of emerging technologies. One of the reasons is the long life cycle of industrial products, often in the range between 20 and 30 years. Beside the necessity of provid-

ing reliability, the system must be available – ideally 24/7. This requires a robust, reliable and conservative design. The "Proven in use" argument is much more accepted than theoretical estimations or testing in the industrial sector[5]. Often electronic components have to be "pre"-purchased in high quantities, because their production will expire before the industrial product life cycle has ended.

Next, we list some important future trends recognized for the embedded industry.

- Due to growing demands for increased performance together with fan-less design requirements, low-power multicore systems capable of executing both safe and non-safe functionality are gaining attention. For such systems, proper separation between safe and non-safe operation modes is required (e.g., by using MMUs, PMUS, core separation, cache separations, etc.)
- Increased hardware complexity (e.g., system-on-chip, SoC, designs) will eventually require employment of "safety certified chips" for safety critical applications, with support for safety certified system software like hypervisors, RTOS (microkernels), drivers, etc. By using safety certified tools for software development, the overall productivity is supposed to increase.
- With time, there has been a significant increase in the demand for system availability. In case of a fault or system failure, power-off is no longer an acceptable solution in the industry like it has been for avionics in ages. In such situations, system must be able to operate in a degraded mode. In robotics this is already a key design requirement, whereby powering off makes the robot loose its current path and can lead to severe damages. By maintaining power, the recommended design allows the robot to keep motors energized forcing it to follow its original path and in addition, it stops faster.
- Connected everywhere all the time (i.e., for the emerging concept of internet-of-things, IoT) will require highly secure solutions with authentication and openness. For such designs, safety and security will merge to assure availability and at the same time maintaining integrity.
- Increasing software complexity with smart/intelligent devices having multiple features will integrate into single devices (e.g., by employing M2M, control info between devices, camera, etc.). Adapted ecosystems (debugger, RTOS, compilers, IDE, programming methodologies, code generators, etc.) will increase productivity.
- Wireless connectivity tends to fall behind in the industrial sector by some years (compared to the application of wired technologies), especially for the safety domain. However, at the low end (i.e., sensors), it will be a key player with focus on low cost, low power, and short-range operation.
- Improving Lifecycle culture with reuse of legacy code (i.e., proven in use). New development will be done in a much more modular way focusing on reusability.

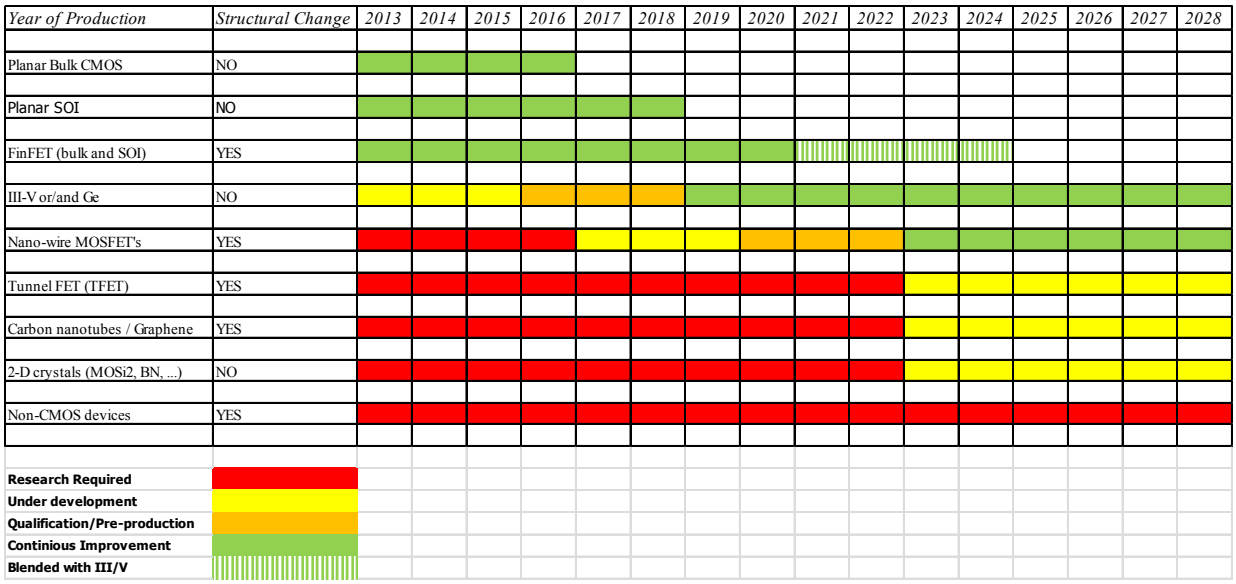
[5] *This will change in IEC 61508 3rd edition which is currently being prepared. The IEC 61508 will follow the ISO 26262 for which the statistical estimations based on detailed computation are accepted and widely used.*

# 3. State of the Art for Manufacturing Technologies

Scaling solid-state devices has the peculiar property of improving cost, performance, and power, which has historically given any company with the latest technology a large competitive advantage in the market. As a result, the microelectronics industry has driven transistor feature size scaling from 10um to ~30 nm during the past 40 years. During most of this time, scaling has simply consisted in reducing the feature size. However, during certain periods, there were major changes as with the industry move from Si bipolar to p-channel metal-oxide semiconductor (MOS), then to n-channel MOS, and finally to complementary MOS (CMOS) planar transistors in the 1980s, which has remained the dominating technology for the past two decades. The big challenge going forward is that the continuous downscaling of the planar CMOS transistor sizes seems to end as the transistor size quickly approaches tens of nanometers. How the industry evolves after this limit is reached is unclear.

## 3.1. Core Logic

A major portion of semiconductor device production is devoted to digital logic. In this section, both high performance logic and low-power logic -which is typically for mobile applications-, are included and detailed technology requirements and solutions are considered. Key considerations are speed, power, density requirements, and reliability. One key theme is continued scaling of the MOSFETs for leading-edge logic technology in order to maintain historical trends of improved device performance. This scaling is driving the industry toward a number of major technological innovations, including material/process changes (e.g., higher-K gate dielectrics and strain enhancement), and the emerging new structures in the near future (such as gate-all-around (nanowire) and alternate high-mobility channel materials). These innovations are expected to be introduced at a rapid pace, and hence understanding, modeling, implementation, and development of these technologies in a timely manner is expected to be a major issue for the industry. Figure 4 shows the predicted timeline for these technologies. Using the ITRS predictions as a baseline, we extended their data to have a wider view of available technologies. For each of them, we depicted their lifetime according to its development stage and year.



| Year of Production | Structural Change | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | 2025 | 2026 | 2027 | 2028 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Planar Bulk CMOS | NO | | | | | | | | | | | | | | | | |
| Planar SOI | NO | | | | | | | | | | | | | | | | |
| FinFET (bulk and SOI) | YES | | | | | | | | | | | | | | | | |
| III-V or/and Ge | NO | | | | | | | | | | | | | | | | |
| Nano-wire MOSFET's | YES | | | | | | | | | | | | | | | | |
| Tunnel FET (TFET) | YES | | | | | | | | | | | | | | | | |
| Carbon nanotubes / Graphene | YES | | | | | | | | | | | | | | | | |
| 2-D crystals (MOSi2, BN, ...) | NO | | | | | | | | | | | | | | | | |
| Non-CMOS devices | YES | | | | | | | | | | | | | | | | |

| | |
|---|---|
| **Research Required** | |
| **Under development** | |
| **Qualification/Pre-production** | |
| **Continious Improvement** | |
| **Blended with III/V** | |

**Figure 4: Technology Roadmap (ITRS predictions + CLERECO additions)**

Through these technologies, silicon engineers will try to keep up with Moore's law. The predicted design points per-year are shown in Figure 5. Note that the industry node (first row) is not anymore consistent with the manufactured channel length, as it was in the past. For conciseness, Figure 5 just reports high-performance design characteristics. In this table, to move to low power characteristics we just need to increase slightly the effective channel length in order to reduce leakage power.

| Year of Production | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
|---|---|---|---|---|---|---|---|---|
| Logic Industry "Node Range" Labeling (nm) [based on 0.71x reduction per "Node Range" ("Node" = ~2x Mx) | "16/14" | | "11/10" | | "8/7" | | "6/5" | |
| $L_{ch}$: Effective Channel Length (nm) | 16,0 | 14,4 | 13,4 | 12,2 | 11,1 | 10,2 | 9,3 | 8,5 |
| $V_{dd}$: Power Supply Voltage (V) | | | | | | | | |
| Bulk/SOI/MG | 0,86 | 0,85 | 0,83 | 0,81 | 0,80 | 0,78 | 0,77 | 0,75 |
| EOT: Equivalent Oxide Thickness | | | | | | | | |
| Bulk/SOI/MG (nm) | 0,80 | 0,77 | 0,73 | 0,70 | 0,67 | 0,64 | 0,61 | 0,59 |
| Dielectric constant (K) of gate dielectrics | 12,5 | 13,0 | 13,5 | 14,0 | 14,5 | 15,0 | 15,5 | 16,0 |
| Physical gate oxide thickness (nm) | 2,56 | 2,57 | 2,53 | 2,51 | 2,49 | 2,46 | 2,42 | 2,42 |
| Channel Doping ($10^{18}$/cm $^{3)}$ | | | | | | | | |
| Bulk | 6,0 | 7,0 | 7,7 | 8,4 | 9,0 | | | |
| SOI/MG | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 |

| | |
|---|---|
| Manufacturable solutions exist, and are being optimized | |
| Manufacturable solutions are known | |

**Figure 5: ITRS predictions on technology**

### 3.1.1. Planar CMOS and Bulk-Si

Planar CMOS on Bulk-Si is the usual name to identify planar MOSFETs built in bulk silicon CMOS processes. As any other transistor, they can be thought as an ideal switch. When it is active the two terminals are connected and thus current flows through the switch, when it is inactive, the two terminals are not connected. This communication in a switch translates in the de-

vice to an active channel through which charge carriers, electrons or holes, flow from the source to the drain. The conductivity of the channel is a function of the potential applied across the gate and the source terminals.

The transistor terminals are: Source (S), through which the carriers enter the channel; Drain (D), through which the carriers leave the channel; Gate (G), the terminal that modulates the channel conductivity (i.e., the switching terminal – by applying voltage to G, one can control the current); and Body (B), the terminal that can –only slightly- modulate the channel conductivity. Figure 6 shows a schematic of the planar CMOS transistor.

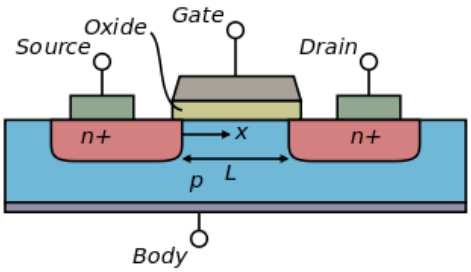



**Figure 6: Planar CMOS transistor**

Channel control only on one-side of the channel (i.e., gate) is ineffective. To compensate for such limitation, manufacturers dope the channel with materials to improve the conductivity and responsiveness. Nevertheless, inserting dopants is not deterministic. Thus, variations in the fabrication process (some even caused by the randomness of matter) end up with an unwanted (and uncontrolled) distribution of transistors with different electrical properties.

On top of that, transistor scaling limits arise also from practical limits related to leakage current at small gate lengths (L). The problem at small gate lengths is that the drain voltage reduces the barrier height at the source, thereby causing a low source-to-channel barrier height even with the gate voltage off, which leads to undesirable, large off-state leakage. This phenomenon is referred to as drain-induced barrier lowering and/or degraded short channel effect (SCE). For evidence that CMOS planar transistors are approaching their minimum practical size, one only needs to look at the off-state leakage trends for the industry. CMOS was initially promised as a technology that dissipated negligible power in the standby state. In present day high-performance logic technologies designed for microprocessors, the leakage power of CMOS transistors is approximately 20-30 W (out of a total power budget of 100 W).

This magnitude of leakage is already at the practical limit since it increases packaging cost (because of cooling) and, even more importantly, energy cost (both in terms of utility bills and the infrastructure to get energy into corporate server computer rooms). To prevent further increases in leakage, the rate of gate length scaling has already slowed in the recent 90 nm and 65 nm technology nodes; thus, rendering Moore's law an illusion. There is no hard limit on the minimum size of a planar CMOS device, but practical considerations on leakage limit the physical gate length to ~20 nm [***3***].

### 3.1.2. FinFETs

FinFETs emerged in high-end microprocessors in the last years [***4***] as improvements in the controlling manufacturing processes allowing a better manufacturing of 3D structures, which first appeared during the 1990's [***5***]. It can be seen as a step towards idealized gate control (i.e., gate all around). Figure 7 shows the basic structure of a FinFET. In contrast to the planar CMOS, the channel is surrounded by the gate on three sides and not on just one side. This configuration allows for better channel control and, thus, better "on-off" behavior (i.e., higher currents for when on, and lower currents –leakage- when off).

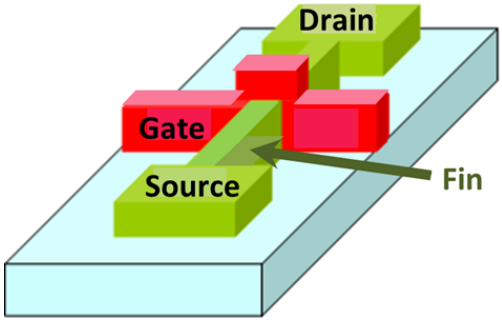



**Figure 7: FinFET structure**

Multi-gate transistors are already in large scale manufacturing for Intel [**6**]. The other semiconductor companies have shown plans to soon produce them as well. FinFETs reduce the need of a doped channel and, thus, they eliminate the variations caused by random dopant implantation. But this does not come for free. Other issues arise though. The channel surface roughness may introduce problems in carrier transport and reliability. Plus, lithography deviations may have a wider impact in such small devices. In this scenario, any minimal change impacts negatively the current density (hence, speed). It is estimated that in current technologies, series resistance degrades the saturation current by 1/3 from that of ideal case. This proportion will likely become harder to maintain or worst with scaling.

### 3.1.3. Silicon-On-Insulator

With Silicon-On-Insulator (SOI) wafers, transistors are formed in thin layers of silicon that are isolated from the main body of the wafer by a layer of electrical insulator, usually silicon dioxide. Isolating the active transistor from the rest of the silicon substrate reduces the electrical current leakage that would otherwise degrade the performance of the transistor. Since the area of electrically active silicon is limited to the immediate region around the transistor, switching speeds are increased and sensitivity to "soft errors" is reduced [***7***] [***8***].

When compared to planar (bulk) CMOS, SOI MOSFETs provide low drain currents, lower source junction capacitances and, consequently, lower leakage currents. Plus, due to their structure and smaller soft error sensitivity, they are able to operate in harsh environments. In this sense, some types of SOI devices, using radiation-resistant buried insulators, will increase the reliability and functionality of communication satellites and other orbiting and deep-space systems. SOI devices also extend the operating range of silicon devices to high temperature environments such as built in diagnostics and controls for automotive and other combustion engines.

Types of SOI-CMOS transistors are characterized by the thickness of the Si-SOI layer. For partially-depleted SOI-CMOS, the device Si layer is thicker than the depletion layer under the channel, in the range of 100 to 200 nm. As CMOS gates are scaled down, CMOS devices will be formed in thin Si layers, which are fully-depleted in the channel region between the source and the drain junctions. For fully-depleted CMOS, the Si device layer is of the order of 50 nm and shrinking towards 10 nm, also known as the "nano-SOI" regime. Fully-depleted CMOS devices will take advantage of the ability of advanced SOI fabrication processes to provide wafers capable of forming dual-gate transistors, with control gates both above and below the thin channel. Figure 8 shows the predicted evolution of SOI devices together with their expected physical design.

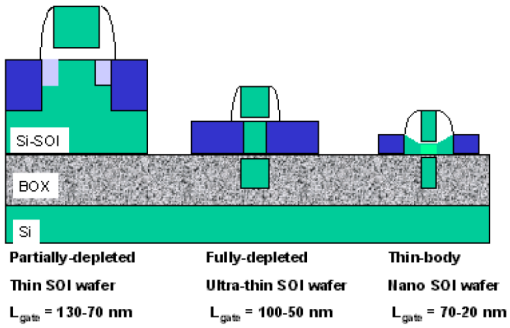



**Figure 8: Different types of SOI devices**

The implementation of fully depleted SOI and ultra-thin body SOI will be challenging. Since such devices will typically have lightly doped channels, the threshold voltage will not be controlled by the channel doping. The problems associated with high channel doping and stochastic dopant variation in planar bulk MOSFETs will be alleviated, but numerous new challenges are expected. Among the most critical issues, there is the control of the thickness and its manufacturing variability for these ultra-thin bodies, and to establish a cost-effective method for reliably setting the threshold voltage.

### *3.1.4. III-V HEMT Technologies*

It has been well recognized that new device engineering is indispensable in overcoming difficulties of advanced CMOS and realizing high performance circuits under 10 nm. In this scenario, the channel materials with high mobility and, more essentially, low effective mass, are preferable under the quasi-ballistic transport expected in ultra-short channel regime. From this viewpoint, strong attention was recently paid to Ge and III-V semiconductor channels. Because of extremely high electron mobility and low electron effective mass of Ge and III-V semiconductors such as GaAs, InP, InGaAs and InAs and extremely high hole mobility and low hole effective mass of Ge, Ge and III-V materials are suitable for high performance CMOS applications. The ITRS 2010 is predicting that the timeline for introducing Ge and InGaAs channels is set to the next few years (see Figure 4).

Transistors using these materials must be fabricated on Si substrates in order to utilize Si CMOS platform, meaning the necessity of the co-integration of III-V/Ge on Si, which is often called heterogeneous integration. Also, those channels must be ultrathin body structures such as ultrathin films, fin structures or nano-wire structures, because of their better control of short

channel effects. The gate stacks composed of high-K gate insulators and metal gates are regarded as mandatory for scaled CMOS.

In order to realize this CMOS structure, there are still many technological issues to be solved for realizing Ge/III-V MOSFETs on Si substrates: (1) high quality Ge/III-V film formation on Si substrates, (2) gate insulator formation with superior MOS/MIS interface quality, (3) low resistivity source/drain (S/D) formation, and (4) total CMOS integration.

The reason for the requirement of the high-mobility materials to be grown on Si substrate is not only for the established processing steps, but also for the expectation that Si components will be included in the same chips. Examples of these Si based components are embedded DRAM and nonvolatile memories, active analog devices including power devices, analog passives, and large circuit CMOS blocks that do not require high performance but better yield. Integrating these different materials with different process requirements is a huge challenge. Take as an example to integrate Si CMOS with III-V/Ge CMOS. There would be likely three kinds of high-K dielectrics required. Different kinds of metal gates are also required to provide different work functions to yield the necessary threshold voltages. And all processes have to be compatible within a single chip.

### *3.1.5. Usage for each Computing Segment*

High-performance computers drive the technology innovation. They are the early adopters of new technologies. While FinFETs are already available now at 14nm, Intel's leading role in technology shows the production of smaller technologies further ahead. Thus, 22nm FinFETs are now the most common in the HPC market and in the next 2 years it is expected to shift to 14nm (as manufacturing yield rises and prices drop).

On the other side, the mobile segment stays slightly behind as it is a very cost-sensitive segment. As seen in the projections of Samsung (Figure 9), it is now involved in the transition to manufacture FinFETs, which they expect to sell starting on 2015.

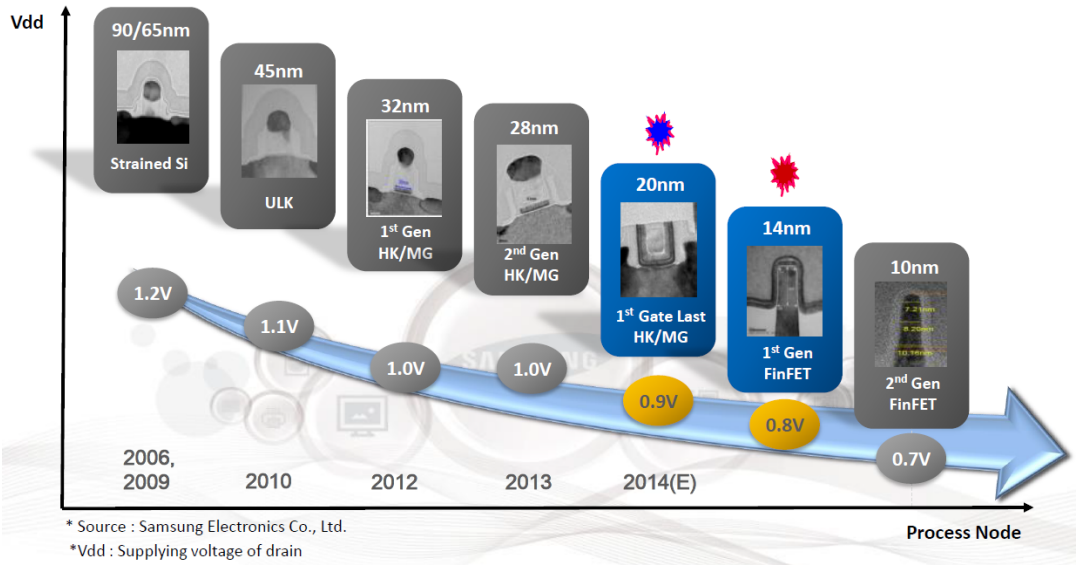



**Figure 9: Samsung technology roadmap**

In the embedded space, due to low volume markets, Non-Recurrent Engineering (NRE) costs of domain-specific solutions are high. Use of Commercial-Off-The-Shelf (COTS) processors is intended for applications requiring high levels of performance. High-performance processors are thus similar to the ones used in other embedded applications. Due to this reason, the core logic in commonly used industrial platforms (in-use or employed in the near future) is planar CMOS, developed using process technology ranging from 130nm – 28nm (130nm technology-based platforms exist in most of the currently employed systems, while 45nm and 28nm correspond to the emerging/upcoming platforms/designs). Table 2 shows the current and near-future technologies adopted in the avionics domain.

**Table 2: Technologies used in the avionics domain**

| Current technologies | Roadmap |
|---|---|
| COTS processors:<br>- Planar CMOS technologies on bulk or SOI substrates.<br>- Examples of technologies used : Global-Foundries C45SOI, TSMC 28HPM<br>In-house design:<br>- Altera HardCopy structured ASIC technology (TSMC 40nm / 28nm)<br>- Flash-based FPGA | - Technologies beyond 28nm and Fin-FET/FD-SOI technologies at the condition to increase confidence on their long term reliability.<br>- SRAM-based FPGA |

A special consideration needs to be made for the space domain. As shown in Figure 10, there is a lag of around ten years in the space domain between the throughput of mainstream processor and the throughput of the space processor.

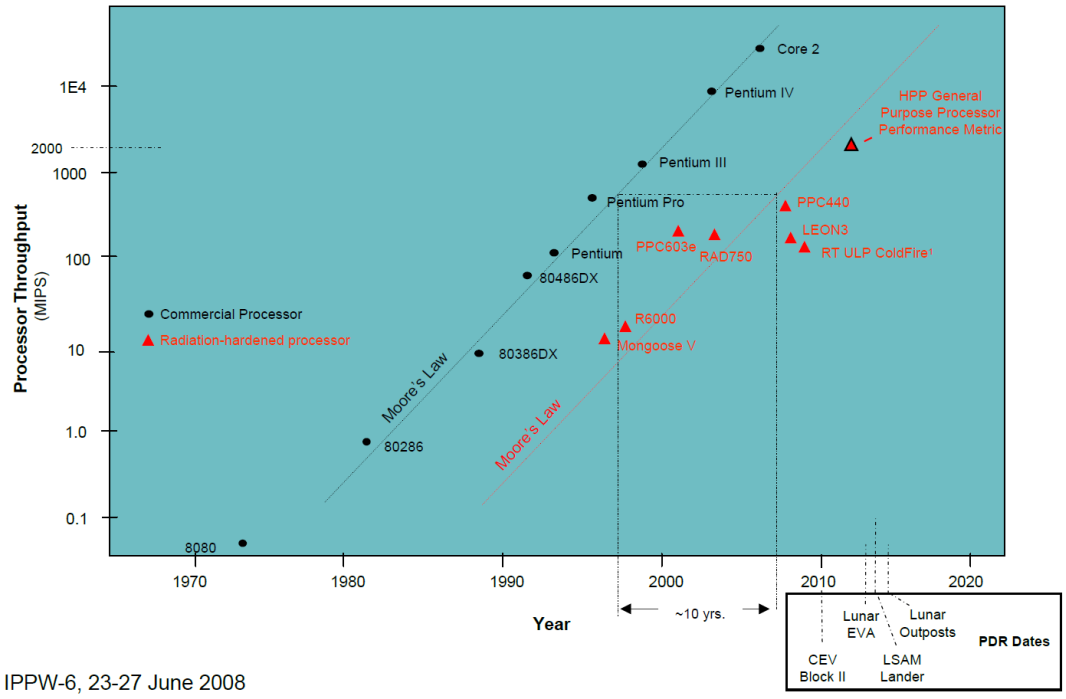



**Figure 10: Lag between performances of commercial processors and space processors [9]**

In this especially harsh environment, specific rad-hard technologies are used for the design of components operating in a spatial environment. Table 3 lists the different technologies available for space applications.

**Table 3: Technologies for space applications**

| Available technologies | Technologies in development |
|---|---|
| Atmel 0.18um CMOS, 0.15um SOI<br>DARE 180nm / 90nm (manufactured by UMC)<br>Honeywell 0.28um/0.15um SOI process<br>Aeroflex UT130nHBD – UT90nHBD<br>Microsemi (CMOS 0.15um 0.13um)<br>Xilinx Virtex FPGAs(CMOS 90nm, 65nm) | ST 65nm (RH-CMOS65LP)<br>Atmel 0.15um SOI, mixed signal (ATC77) |

Wrapping up, the technology/segment distribution is as follows:

**Table 4: Summary of technologies foreseen per segment**

| Core Logic / segment | | High-Performance | General Purpose / Mobile | Industrial and Safety/Mission Critical |
|---|---|---|---|---|
| **Structure** | Planar CMOS | YES (diminishing) | YES (diminishing) | YES |
| | FinFET (3D) | YES (arising) | YES (arising) | NO |
| **Substrate** | Bulk | YES (dominant) | YES | YES |
| | SOI | YES | YES | YES |
| | III/V | Experimental | Experimental | Sensors |

## 3.2. Embedded Memories

Memory is the core of computation – be it human or machine, we cannot process anything unless we have a place to store data, and that's why memory has always been one of the core components in computer design. When talking about memory, most assume that we are referring to RAM but there is more than that.

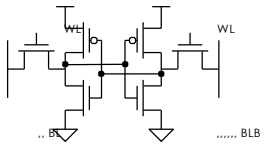
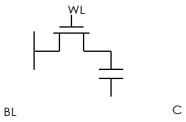
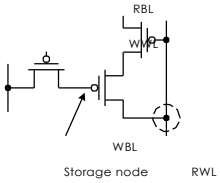
Memory is classified into two major categories, Static RAM (SRAM), and Dynamic RAM (DRAM). Static RAM uses a special arrangement of transistors (a ring of two inverters plus access transistors) to make a memory cell. One memory cell can store 1-bit of data. Most modern SRAM cells are made of six CMOS transistors, and are the fastest type of memory available.

In contrast, Dynamic RAM lines up one transistor with a capacitor to create an ultra-compact memory cell. On the flip side, the capacitor needs to be refreshed after a specific period to keep the charge in the capacitor, which introduces latency in memory access, as not all cells may be available at any time.

While DRAM has an obvious size advantage over SRAM, its speed cannot even get close to those offered by static memory cells (because they don't need to be refreshed and are always available for access). That's why faster memory is always made out of SRAM cells – like Registers in the CPU and Caches used in numerous devices. But thanks to much higher space requirements, SRAM is expensive and cannot be used as primary memory of a system.

DRAM on the other hand is quite dense, and therefore is employed in most places that do not require instantaneous access but large capacities – like the main memory in a computer. Table 5 describes and compares the most widely used SRAM and DRAM designs.

**Table 5: Comparison of various memory technologies for on-die caches [*10*]**

| | (A) SRAM | eDRAM | |
|---|---|---|---|
| | | (B) 1T1C | (C) Gain cell |
| **Cell schematic** |  |  |  |
| **Process** | CMOS | CMOS + Cap | CMOS |
| **Cell size[6] ($F^2$)** | 120 - 200 | 20 - 50 | 60 - 100 |
| **Data storage** | Latch | Capacitor | MOS gate |
| **Read time** | Short | Short | Short |
| **Write time** | Short | Short | Short |
| **Read energy** | Low | Low | Low |
| **Write energy** | Low | Low | Low |
| **Leakage** | High | Low | Low |
| **Endurance** | $10^{16}$ | $10^{16}$ | $10^{16}$ |
| **Retention time** | - | < 100 us * | < 100 us * |
| **Features** | (+) Fast (-) Large area (-) Leakage | (+) Low leakage (+) Small area (-) Extra process (-) Destructive read (-) Refresh | (+) Low leakage (+) Decoupled read/write (-) Refresh |

### *3.2.1. SRAM*

The first traces to SRAM date back to 1964, when 64-bit Metal Oxide Semiconductor (MOS) Static RAM was developed at Fairchild Semiconductor. However, the breakthrough came when Intel developed its first 256-bit static RAM (SRAM), the 1101 chip in 1969 and formally launched it in 1971.

[6] *Cell size is mesured as a function of F. F is called half-pitch. The pitch is the minimum distance between the first level metal lines on the integrated circuit.*

The SRAM cell consists of a bi-stable flip-flop connected to the internal circuitry by two access transistors (Table 5A, the ones connected to WL). When the cell is not addressed (WL=0), the two access transistors are closed and the data is kept to a stable state, latched within the flip-flop. The cell needs the power supply to keep the information. The data in an SRAM cell is volatile (i.e., the data is lost when the power is removed). However, the data does not "leak away" like in a DRAM, so the SRAM does not require a refresh cycle.

The *de facto* SRAM 6T design is shows in Table 5A. It is built as any combinational logic in the circuit, thus it does not require any more process steps. It is fast but costly in terms of area and leakage power. As technology shrinks, several other designs have been proposed that increase the cell robustness (i.e. susceptibility to noise, couplings, etc.) and they reduce leakage power. The most known designs are the 8T cell and the 10T cell. Obviously, adding more transistors has a cost –at least- in area and power. Nevertheless, the performance and robustness boost they bring meets the design constrains of very selective environments.

SRAM memories are fundamental components in any computing system nowadays. Almost all on-chip memories in all processing chips are SRAMs (e.g. caches, register files, buffers, tables, etc.).

### *3.2.2. eDRAM*

In 1996, Mitsubishi took a standard 16-Mbit DRAM and wedged a RISC CPU into the middle. The M32R/D cost more than separate processor and DRAM, and it did not catch on. Almost 20 years later, embedded DRAM's (eDRAM) allow system designers to use high bandwidth, high performance and high-density memory near the processing core (for high-performance chips) or within the System On Chip (SOC). Logic based embedded DRAM technologies are now present in some ST microelectronics devices, as well as, in some high-performance devices from IBM and Intel cores. Embedded DRAM is therefore a powerful tool if it can be made cost effective. Logic or stand-alone DRAM technologies have been used to realize eDRAM's. Logic based technologies offer the advantage of high performance, and compatibility with existing cores, which is essential. The main challenge of logic based eDRAM is to find the right compromise between added process cost and memory density.

As far as applications are concerned, one of the main benefits of logic based eDRAM is higher performance. Some examples of high speed and high-resolution applications are graphics and networking, using wide bus width. Another benefit is cost reduction as a system can be designed with fewer components. This applies to digital consumer applications like printers, cell phone and camcorders, as well as most embedded systems. In that case, cost analysis of added complexity versus memory density is key.

Process choices for logic-based eDRAMs are driven by 2 factors: compatibility with the logic transistor, and cost. For compatibility, the logic process has to remain unchanged so that standard cell libraries and IPs can be directly usable. To maintain the compatibility with the logic process, low thermal budget recipes are mandatory to build the capacitor: low temperature nitrides or colder solutions using MIM (Metal Interdielectric Metal) with high-K dielectrics. For low cost requirements, CUB (Capacitor Under Bitline) is the preferred choice to minimize the number of added masks, using the first interconnect level as bitline. Retention time is not considered as a critical factor as long as design solutions allow the implementation of hidden refresh and ECC (Error Correction Code) to improve the DRAM robustness.

For trade-offs between process cost and DRAM density, the stacked capacitor lies between 2 other architecture choices: planar cells and deep trench cells. The planar cell allows a very easy integration with only one added mask, but cell sizes remain 2 to 3 times larger than stacked or trench cells, restricting the use to small DRAM capacity. On the other side the trench cell is very competitive in terms of size, allowing high memory density, but with added process complexity, close to a stand-alone DRAM process.

### *3.2.3. Usage for each Computing Segment*

While all systems will rely on SRAM memories for their on-chip memories (i.e. caches), in the HPC segment we foresee an adoption of eDRAM as a solution to include larger caches to maximize the on-chip memory capacity. Currently, both IBM and Intel HPC high-end processors include eDRAM.

In many industrial embedded devices the CPU chip holds both built-in ROM (Flash, 512K++) and RAM (128K++) adequate to run required software – often with a tiny RTOS with communication and a small application. The variety in this segment of chips has increased enormous with all kind of performance and behavior. Very often the same CPU core is present with huge variations in peripheral composition tailor made for specific application areas. Depending on the application scenario, it typically ranges from the low cost end with low power, slow CPU, little ROM/RAM and few peripherals (referred to as low-end devices) to costly full-scale high performance solutions (known as high-end devices).

Wrapping up, the technology/segment distribution is as follows:

**Table 6: Summary of on-chip memory technologies foreseen per segment**

| On-chip Memory usage / segment | High-Performance | General Purpose / Mobile | Industrial and Safety/ Mission Critical |
|---|---|---|---|
| **SRAM** | YES | YES | YES |
| **eDRAM** | YES | NO | NO |

## 3.3. Main Memory and Storage

CMOS logic and memory together form the predominant majority of semiconductor device production. The types of memory considered in this document are DRAM and non-volatile memory (NVM). The emphasis is on both commodity and embedded memory chips are expected to follow the same trends just with some time lag. For both DRAM and NVM, detailed technology requirements and potential solutions are considered.

For DRAM, the main goal is to continue to scale the foot-print of the 1T-1C cell, to the practical limit of $4F^2$. The main challenges in manufacturing are the creation of vertical transistor structures, the introduction of dielectrics to improve the capacitance density, and meanwhile keep the leakage low.

The NVM discussion in this section is limited to devices that can be written and read many times; hence read-only memory (ROM) and one-time-programmable (OTP) memory are not included although many such memories are important both for standalone and embedded

applications (as they usually hold the boot and setup data). The current mainstream NVM is Flash memory. NAND and NOR flash memories are used for quite different applications –data storage for NAND and code storage for NOR flash. There are serious issues with scaling for both NOR and NAND flash memories that are dealt with at some length in Section 3.3.2. Other non-charge-storage types of NVM are also considered in Section 3.3.3, including ferroelectric RAM (FeRAM), magnetic RAM (MRAM), and phase-change RAM (PCRAM). These emerging memories promise to continue NVM scaling beyond Flash memories. However, because NAND Flash and to some extent NOR Flash are still dominating the applications emerging memories have been used in specialty applications and have not yet fulfilled their original promise to become dominating mainstream high-density NVM.

In general, technical requirements for DRAMs become more difficult with scaling (see Figure 11 with the ITRS projections). In the past couple of years, DRAM was introduced with many new technologies (e.g. 193 nm argon fluoride (ArF) immersion high-NA lithography with double patterning technology, improved cell FET technology including fin type transistor [***11***] [***12***], buried word line/cell FET technology [***13***] and so on). Due to new technologies, DRAM will continue to scale with 2-3 year cycle and 20 nm HP (minimum feature size) DRAM will be available by 2017.

| *Year of Production* | *2013* | *2014* | *2015* | *2016* | *2017* | *2018* | *2019* | *2020* |
|---|---|---|---|---|---|---|---|---|
| Logic Industry "Node Range" Labeling (nm) [based on 0.71x reduction per "Node Range" | "16/14" | | "11/10" | | "8/7" | | "6/5" | |
| Half Pitch -F- (Contacted line) (nm) | 28 | 26 | 24 | 22 | 20 | 18 | 17 | 15 |
| DRAM cell size ($\mu m^2$) | 0,00470 | 0,00406 | 0,00346 | 0,00194 | 0,00160 | 0,00130 | 0,00116 | 0,00090 |
| DRAM storage node cell capacitor dielectric: equivalent oxide thickness EOT (nm) | 0,55 | 0,5 | 0,4 | 0,3 | 0,3 | 0,3 | 0,3 | 0,3 |
| DRAM cell FET structure | RCAT+Fin | RCAT+Fin | RCAT+Fin | VCT | VCT | VCT | VCT | VCT |
| DRAM Cell Transistor Gate material (Burried/Planer/Vertical+Gate material) | Buried/TiN | Buried/TiN | Buried/TiN | Vertical/TiN | Vertical/TiN | Vertical/TiN | Vertical/TiN | Vertical/TiN |
| Minimum DRAM retention time (ms) | 64 | 64 | 64 | 64 | 64 | 64 | 64 | 64 |
| DRAM soft error rate (fits) | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |
| Cell Size Factor: a (DRAM size/$F^2$) | 6 | 6 | 6 | 4 | 4 | 4 | 4 | 4 |
| ***Gb/1chip target*** | 4G | 8G | 8G | 8G | 8G | 16G | 16G | 16G |
| | | | | | | | | |
| | | | | | | | | Manufacturable solutions exist, and are being optimized |
| | | | | | | | | Manufacturable solutions are know n |
| | | | | | | | | Manufacturable solutions are NOT know n |

**Figure 11: DRAM technology outlook (excerpt from ITRS)**

## *3.3.1. DRAM*

Robert Dennard invented DRAM at the IBM Thomas J. Watson Research Center in 1966/1967. Next year, the first known DRAM chip ever developed was a 256-bit device created by Lee Boysel at Fairchild Semiconductor. Later, Boysel founded Four Phase Systems in 1969 and developed 1,024-bit and 2,048-bit DRAMs. Intel released the 1103, the industry's first mass-produced DRAM device, in 1970.

Aside from the energy consumption and the high performance (and power) dependence on ambient temperature, there are still plenty of technical challenges as well as the issue of increased process steps to sustain the cost scaling. Fundamentally, there exist several significant process flow issues from a production standpoint, such as process steps of capacitor formation, or high aspect ratio contact etches requiring photoresists with hard mask pattern transferring layer that can stand up for a prolonged etch time. Furthermore, continuous improvements in lithography/hard mask and etch will be needed. In addition, lower wordline/bitline resistance is necessary for getting the same or better performance.

Although 3-D type cell FETs like saddle-fin FETs are introduced and have revolutionized the one transistor-one capacitor (1T-1C) cell (see Figure 12), it is getting more difficult to design due to the need to maintain a low level of both subthreshold leakage and junction leakage current to meet the retention time requirements. To optimize these operation windows in future devices, fully depleted type FET device (like a surrounded gate) will be needed to reduce the BL capacitance to get the sense margin. Another challenge is a highly reliable gate insulator. A highly boosted gate voltage is required to drive higher drain current with the relatively high threshold voltage adopted for the cell FET to suppress the subthreshold leakage current. The scaling of the DRAM cell FET dielectric, maximum word-line (WL) level, and the electric field in the cell FET dielectric are critical points for gate insulator reliability concern. To keep the electric field to a sustainable level in the dielectric with scaling, novel manufacturing process requirements for DRAMs such as front-end isolation, recess-FET formation, conformal oxidation process, gate filling process, and damageless recess process are all needed for future high-density DRAMs.

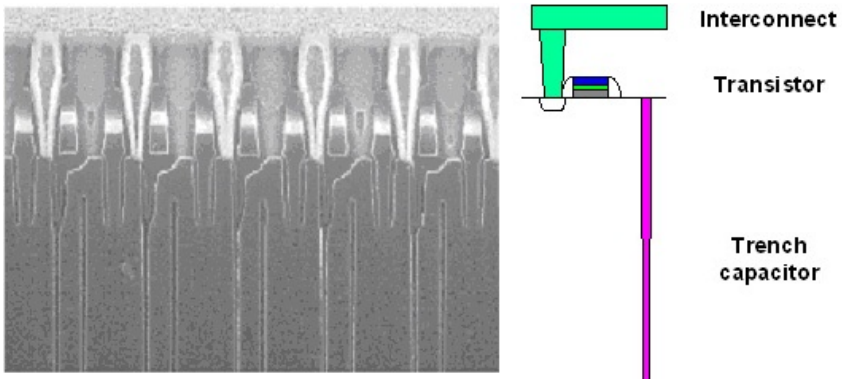



**Figure 12: DRAM cell schematic (right), implementation by Quimoda (left)**

**Potential Solutions**

Since the DRAM storage capacitor gets physically smaller with scaling, the EOT (Electrical Oxide Thickness) must scale down sharply to maintain adequate storage capacitance. To scale the EOT, dielectric materials having high relative dielectric constant (K) will be needed. Therefore MIM (metal-insulator-metal) capacitors have been adopted using high-K (ZrO2/Al2O/ZrO2) as the capacitor of 40-30's nm half-pitch DRAM. And this material evolution and improvement are continued until 20 nm and ultra-high-K (such as perovskite) materials will be needed as of 2016 [***14***]. Also, the physical thickness of the high-K insulator should be scaled down to fit the minimum feature size. Due to that, capacitor 3-D structure will be changed from cylinder to pillar shape.

On the other hand, with the scaling of peripheral CMOS devices, a low-temperature process flow is required for process steps after formation of these devices. This is a challenge for DRAM cell processes, which are typically constructed after the CMOS devices are formed, and therefore are limited to low-temperature processing.

The other big topic is the migration to $4F^2$ cell size. As the half-pitch scaling becomes very difficult, it is impossible to sustain the cost trend. The most promising way to keep the cost trend and increasing the total bit output by generation is by changing the scaling of cell size factor (a) (where a = [DRAM cell size]/[DRAM half pitch]). Currently, $6F^2$ is most commonly used cell

size factor. To migrate to $4F^2$ cell is very challenging. For example, vertical cell transistor must be needed but still a couple of challenges are remaining.

All in all, maintaining sufficient storage capacitance and adequate cell transistor performance are required to keep the retention time characteristic in the future. However, continued scaling of DRAM devices and the demand of bigger product sizes (i.e. >16 Gb) have made the underlying requirements more difficult.

The main near-future solutions are listed in Figure 11 and in the paragraphs above, but new future technologies will necessary be beyond the "8/7" node. The most promising alternatives are described in Section 3.3.3.

### *3.3.2. Flash memory*

Toshiba's Fujio Masuoka invented Flash memory in the early 1980s [***15***] [***16***]. Masuoka discussed and detailed flash (NOR and NAND) for the first time.

Flash memories are based on simple one transistor (1T) cells (see Figure 13), where a transistor serves both as the access (or cell selection) device and the storage node. Several non-conventional non-volatile memories that are not based on charge storage (Ferroelectric or FeRAM, Magnetic or MRAM, phase-change or PCRAM, and resistive or ReRAM) are often called "emerging" memories. They are described in the next section. These memory elements (the storage node) usually have a two-terminal structure (e.g. resistor or capacitor) thus do not serve as the cell selection device. The memory cell must include a separate access device in the form of 1T-1C, 1T-1R, or 1D-1R.

Floating gate Flash devices achieve non-volatility by storing and sensing the charge stored "in" (on the surface of) a floating gate. The NAND array consists of bit line strings of now 64 devices or more with a string selection device at each end. This architecture requires no direct bit line contact to the cell, thus allows the smallest cell size. During programming or reading, the unselected cells in the selected bit line string must be turned on and serve as "pass" devices, thus the data stored in each device cannot be accessed randomly. Data input/output are structured in "page" mode where a page (on the Word line) is of several KB (8KB – 16KB today) in size. Both programming and erasing are by Fowler-Nordheim tunneling of electrons into and out of the floating gate through the tunneling oxide. The low Fowler-Nordheim tunneling current allows the simultaneous programming of many bits (page), thus gives high programming throughput, suitable for handling large amount of data. Since devices in the same bit line string serve as pass transistors their leakage current does not seriously affect programming or reading operation (up to a limit), and without the need for hot electrons junctions can be shallow. Thus the scaling of NAND flash is not limited by device punch through and junction breakdown as in NOR flash.

Gate coupling and floating gate to floating gate cross talk are difficult challenges when scaling below 20nm. Both can be alleviated by adopting high-K IPD and using a planar structure. Successful implementation of this new innovation in the 20nm and 16nm nodes recently gives hope to scale 2D NAND using a planar cell structure into the ~ 10nm regime. Although high-K also helps to reduce the program/erase voltage, the voltage reduction does not catch up with the rate of 1/2 pitch scaling, thus WL-WL electric field continues to increase and breakdown becomes a serious scaling limitation. Low-K dielectric is already not effective and air gaps between word lines are now adopted to improve the breakdown tolerance. Further scal-

ing, however, still faces this very difficult challenge as the electric field increases at each new node.

Since the tunnel oxide scales very slowly, or not at all, the fringing field of the scaled device becomes less controlled (by the control gate) thus both degrades the device performance (larger subthreshold swing) and also increases the cell-to-cell interference. The number of storage electrons decreases linearly with the area of the device, in principle, and thus eventually will be too low and will cause unacceptable retention time distribution and severe random (telegraph) noise.

(Planar) NAND Flash has now already scaled to 16nm node and further scaling to near 10nm seems possible. Beyond that, WL-WL breakdown, neighboring cell interference and statistical fluctuation of number of storage electrons must be overcome to further scale.

**Charge Trapping NAND Flash**

Currently all NAND products are fabricated with floating gate devices. The difficult challenges of maintaining or increasing the gate coupling and reducing the neighboring cell cross talk may be reduced by using charge trapping devices, but since rapid progress in planar HK/MG device has already alleviated these issues it is unlikely that 2D charge trapping devices will be adopted. Most 3D NAND devices, however, use charge trapping devices thus their principle and operation are described.

Charge trapping devices have only one single gate that controls the MOS device channel directly and thus there is no coupling issue, and the cross talk between thin nitride storage layers is either insignificant or at least much reduced. Nitride trapping devices may be implemented in a number of variations of a basic SONOS (*Silicon-Oxide-Nitride-Oxide-Silicon*) type device –see Figure 13.

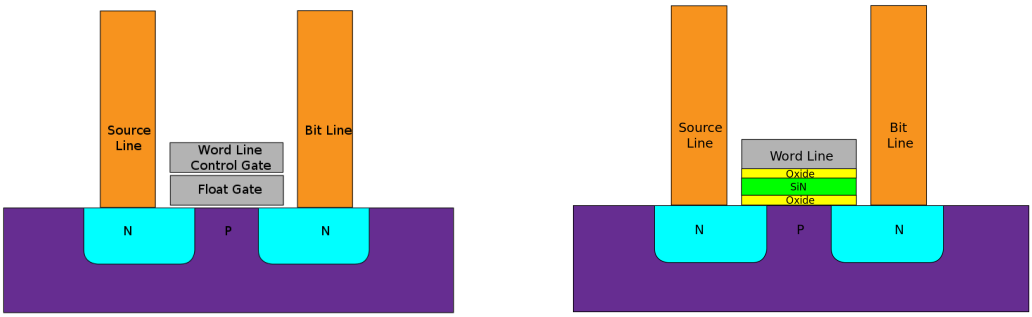



**Figure 13: Conventional Flash transistor (left) SONOS Flash transistor (right)**

Although charge trapping NAND can help alleviate coupling, cross talk issues and scaling below 20nm; it does not help the fundamental limitations such as word line breakdown and too few electrons. Therefore, in the flash roadmap [***14***] it occupies a transition role between planar and 3D NAND. When charge-trapping devices are used to build 3D NAND, the larger device size naturally solves the electron number and the word line breakdown issues.

### 3.3.3. Emerging technologies

Since the ultimate scaling limitation for charge storage devices is too few electrons, devices that provide memory states without electric charges are promising to scale further. Several non-charge-storage memories have been extensively studied and some commercialized, and each has its own merits and unique challenges. Some of these are uniquely suited for special applications and may follow a scaling path independent of Flash. Some may eventually replace flash memories. Logic states that do not depend on charge storage eventually also run into fundamental physics limits. For example, small storage volume may be vulnerable to random thermal noise, such as the case of super-paramagnetism limitation for MRAMs.

One disadvantage of this category of devices is that the storage element itself cannot also serve as the memory selection (access) device (transistor) because they are mostly two-terminal devices. Therefore, these devices use 1T-1C (FeRAM), 1T-1R (MRAM, PCRAM and ReRAM) or 1D-1R (PCRAM and ReRAM) structures. Figure 14 shows the basic structure of any Non-Volatile Memory (NVM). The access transistor is drawn at the bottom part of the figure, then, the extra capacitor, resistor or magnetic material is connected to the Drain (in this case, a Ferroelectric layer). In any of these designs, it is challenging to achieve small ($4F^2$) cell size without innovative access devices. In addition, because of the more complex cell structure that must include a separate access device, it is more difficult to design 3D arrays that can be fabricated using just a few additional masks like those proposed for 3D NAND.

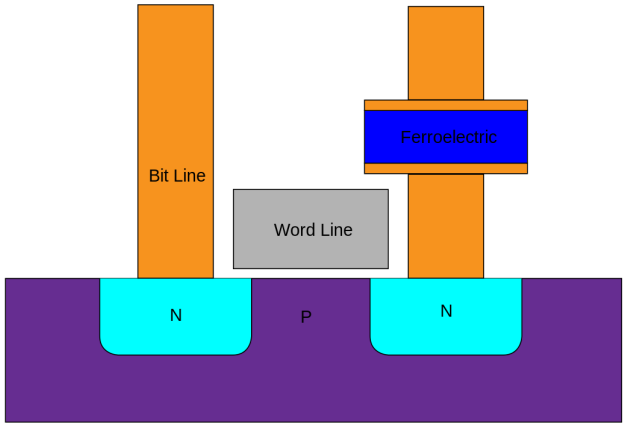



**Figure 14: Basic NVM structure**

**FeRAM**

FeRAM, or Ferroelectric RAM, still feels somewhat exotic today, but its history dates back to 1952, when MIT graduate student Dudley Allen Buck described the principle of FeRAM in his master's thesis [***17***]. It took more than 30 years for the idea to be picked up again. The technology idea was completed in 1991 at NASA's Jet Propulsion Laboratory.

FeRAM devices achieve non-volatility by switching and sensing the polarization state of a ferroelectric capacitor. These memories have a destructive read operation (i.e. the "value" stored in the cell is deleted when reading. Because of this, it is a challenge to find ferroelectric and electrode materials that provide both adequate change in polarization and the neces-

sary stability over extended operating cycles. The ferroelectric materials are foreign to the normal complement of CMOS fabrication materials, and can be degraded by conventional CMOS processing conditions. Thus, the ferroelectric materials, buffer materials, and process conditions are still being refined. So far, the most advanced FeRAM [***18***] is substantially less dense than NOR and NAND Flash, fabricated at least one technology generation behind NOR and NAND Flash. However, FeRAM is fast, low power, and low voltage and thus is suitable for RFID, smart card, ID card, and other embedded applications. In order to achieve density goals with further scaling, the basic geometry of the cell must be modified while maintaining the desired isolation.

**MRAM**

First steps in developing MRAM devices were made by Nobel laureates Peter Grünberg [***19***] and Albert Fret [***20***]. Later, IBM in 1989 discovered the "giant magneto-resistive" effect in thin-film structures.

MRAM devices employ a magnetic tunnel junction (MTJ) as the memory element. An MTJ cell consists of two ferromagnetic materials separated by a thin insulating layer that acts as a tunnel barrier. When the magnetic moment of one layer is switched to align with the other layer (or to oppose the direction of the other layer) the effective resistance for current flow through the MTJ changes. The magnitude of the tunneling current can be read to indicate whether a "1" or a "0" is stored. Field switching MRAM probably is the closest to an ideal "universal memory" since it is fast, non-volatile and can be cycled indefinitely, thus may be used as NVM as well as SRAM and DRAM. However, producing a magnetic field in an IC circuit is both difficult and inefficient. Plus, the big challenge is to reach the adequate magnetic intensity fields to accomplish switching in scaled cells, where electromigration limits the current density that can be used. Therefore, it is expected that field switch MTJ MRAM is unlikely to scale beyond 65nm node.

Recent advances in "spin-torque transfer (STT)" (also referred to as spin-transfer torque) approach, where a spin-polarized current transfers its angular momentum to the free magnetic layer and thus reverses its polarity without resorting to an external magnetic field, offered a new potential solution. During the spin transfer process, substantial current passes through the MTJ tunnel layer and this stress may reduce the writing endurance. Upon further scaling, the stability of the storage element is subject to thermal noise.

**PCM**

Phase-change memory (PCM) is still in its nascent stages today, more than 50 years after its invention. In his 1969 dissertation, Charles Sie of the Iowa State University explained that a phase change memory device would be "feasible" by integrating chalcogenide film with a diode array. However, some work had been done prior to that by Stanford Ovshinsky at Energy Conversion Devices, who believed that the properties of chalcogenide glasses could be used as a potential memory technology. Intel co-founder Gordon Moore also published a paper describing phase-change memory in 1970.

PCRAM devices use the resistivity difference between the amorphous and the crystalline states of chalcogenide glass (the most commonly used compound is Ge2Sb2Te5, or GST) to store the logic "1" and logic "0" levels. The device consists of a top electrode, the chalcogenide phase change layer, and a bottom electrode. The leakage path is cut off by an access transistor in series with the phase change element. The phase change write operation consists of: (1) RESET, for which the chalcogenide glass is momentarily melted by a short elec-

tric pulse and then quickly quenched into amorphous solid with high resistivity, and (2) SET, for which a lower amplitude but longer pulse (usually >100ns) anneals the amorphous phase into low resistance crystalline state. The 1T-1R (or 1D-1R) cell is larger or smaller than NOR Flash, depending on whether MOSFET or BJT is used, and the device may be programmed to any final state without erasing the previous state, thus it provides substantially faster programming throughput. The simple resistor structure and the low voltage operation also make PCRAM attractive for embedded NVM applications. The major challenges for PCRAM are the high current (fraction of mA) required to reset the phase change element, and the relatively long set time. Since the volume of phase change material decreases rapidly with each technology generation, there is a hope that both the aforementioned issues become less problematic with scaling. Interaction of phase change material with electrodes may pose long-term reliability issues and limit the cycling endurance and is a major challenge for DRAM-like applications. Because PCRAM does not need to operate in page mode (no need to erase) it is a true random access, bit alterable memory like DRAM.

The limited cycling endurance and the smaller bandwidth (due to high current required for writing) make PCRAM unsuitable to replace DRAM. However, it is otherwise similar to DRAM and its scalability may make it less expensive than DRAM in the future. Moreover, since PCRAM is nonvolatile it saves both the refreshing power, and more important, the dead time for refreshing which becomes increasingly a problem for DRAM. Therefore, a hybrid memory using small amount of DRAM and mostly PCRAM can be a low cost solution for high performance memory.

**ReRAM**

While the term for resistive RAM was recently coined by a group of researchers at Sharp and the University of Houston in 2002 [***21***], the discovery of switching resistance materials was first published during 1964 by Nielsen and Bahara in the University of Nebraska [***22***] and with other materials by Gibbons and Beadle at Stanford the same year [***23***]. It was not until 1967 that Simmons and Verderber at Standard Telecommunication Laboratories Ltd. In the UK, show the use as a memory device [***24***].

Nowadays, still many of these resistive memories are still in research stage. Resistive memories promise to scale below 10nm and the focused R&D efforts in many industrial labs make this technology widely considered a potential successor to NAND (including 3D NAND).

Resistive memories change the built-in resistor conductivity by atomic processes, thus are not limited by the number of storage electrons. In principle, it should eventually also be limited by the number of atoms that provide the electrical characteristics. In resistive switching memory cells (ReRAMs), ions behave on the nanometer scale in a similar manner to a battery. The cells have two electrodes, for example made of silver and platinum, at which the ions dissolve and then precipitate again. This changes the electrical resistance, which can be exploited for data storage. Still, there is not enough understanding of the atomic details to project when this will limit the scaling of ReRAM. In the device level, < 10nm ReRAM has been reported. In the array level, 20nm 1Gb 2-layer 3D ReRAM has been published. However, high-density ReRAM still must overcome several difficult challenges to be cost competitive to NAND.

### *3.3.4. Usage for each Computing Segment*

DRAM memories will keep dominating all the markets. The incursion of DDR4 memories will be first in the HPC segment and then into the mobile and embedded ones. Flash is widely used

nowadays across all the segments. Other technologies have yet to prove their endurance, performance and scalability before they become a real alternative.

With rapid progress of NAND Flash and the recent introduction of 3D NAND that promises to continue the equivalent scaling, the hope of STT-MRAM to replace NAND seems remote. However, its SRAM-like performance and much smaller footprint than the conventional 6T-SRAM have gained much interest in that application, especially in mobile devices which do not require high cycling endurance as in computation.

For industrial embedded devices, when external memories are used, it always consists of Flash for program storage (and eXecute-In-Place) and DRAM for both execution and data (DDR2). For small embedded devices without a Graphical User Interface (GUI), typically Flash size is in the range 1 to 16MB and DRAM in the range 1 to 8MB. DDR3 memories are expected to be replacing DDR2s in the near future in such systems.

DRAM and Flash technologies are generally used in mission-critical applications with some lag between their appearance in consumer electronics and their use in these applications. For mission-critical applications, the long-term reliability of 3D NAND technology is currently a source of concern. Due to their low maturity level, these technologies are generally not used in critical applications. Emerging memory technologies have also a low maturity level. However, their low sensitivity to SEU is a very interesting property for these applications.

Summing up, the technology/segment distribution is as follows:

**Table 7: Summary of main memory technologies foreseen per segment**

| Main Memory / segment | High-Performance | General Purpose / Mobile | Industrial and Safety/ Mission Critical |
|---|---|---|---|
| **DRAM** | YES | YES | YES |
| **FLASH** | YES (storage) | YES | YES |
| **Other…** | Test-chips | NO | NO |

## 3.4. Interconnects

Driven by continuing scaling of Moore's law, chip multi-processors and systems-on-a-chip are expected to grow the core count from dozens today to hundreds in the near future. Interconnect has become a primary bottleneck in integrated circuit design. As CMOS technology is scaled, it will become increasingly difficult for conventional copper interconnect to satisfy the design requirements of delay, power, bandwidth, and noise. Figure 15 shows the evolution of logic and interconnect delays making the distinction of local (intra-core) to global (inter-core) interconnects. It can be seen that in current technologies, the delay gap is between 2x and 3x. The introduction of copper interconnects alleviated the problem for some time (green line) nevertheless, nowadays it is not enough and new interconnect fabrics are being researched.

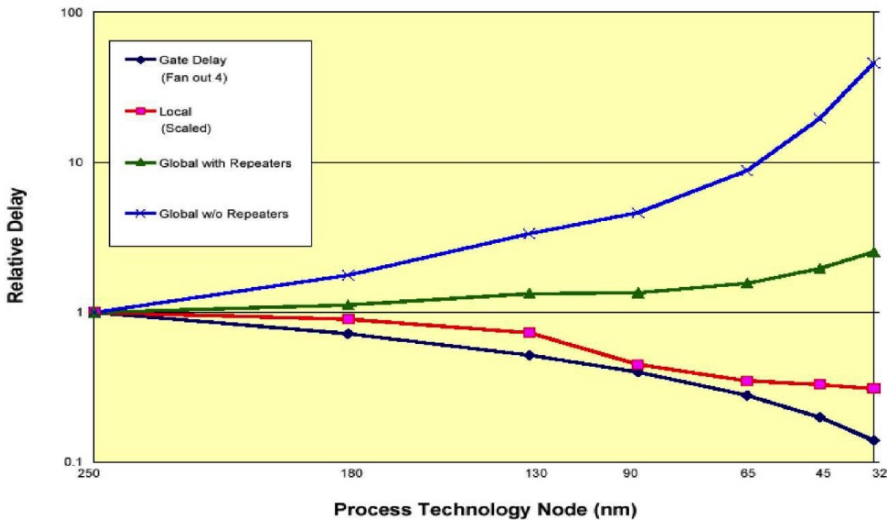



**Figure 15: Delay of logic (dark blue) and several on-chip network configurations (source:ITRS)**

On-chip optical interconnect has been considered as a potential substitute for electrical interconnect in the past two decades. Optical devices are widely used in the telecommunication area, and are commonly applied as board level interconnects. The concept of on-chip optical interconnect was first introduced by Goodman in 1984 [***25***]. Since electrical/optical and optical/electrical conversion is required, an optical interconnect is particularly attractive for global interconnects, such as data buses and clock distribution networks. Recently, several comparisons have been made between on-chip electrical and optical interconnects, the most complete is the work by Chen et al. [***26***]. They perform a more comprehensive comparison between optical and electrical interconnects at different technology nodes based on a practical prediction of optical device development. Still, this comparison is particularly challenging since optical interconnect is a young fast-developing technology, while electrical interconnect is relatively mature. Table 8 shows the predicted data for delay. It can be seen that 45nm becomes the trip point where optical networks achieve better overall transmission delay. Plus, Table 9 shows that at the same technology node optical networks are more efficient than electrical. Bearing in mind that these numbers are for global interconnects (i.e. intra-core); it would seem that should be present in any device in the newest technologies. Nevertheless, this is not the case. Combining optical transmitters and receivers on a silicon substrate without impacting the fabrication cost significantly has not been the case so far.

**Table 8: Delay (ps) distribution in a 1 cm optical data path as compared with the electrical data path**

| *Year* | *2004* | *2007* | *2010* | *2013* | *2016* |
|---|---|---|---|---|---|
| **Technology node** | 90 nm | 65 nm | 45 nm | 32 nm | 22 nm |
| **Optical transmitter** | 177.5 | 18.4 | 8.6 | 6.0 | 5.0 |
| **Optical receiver** | 0.4 | 0.3 | 0.2 | 0.3 | 0.3 |
| **Total optical** | 177.9 | 18.7 | 8.8 | 6.3 | 5.3 |
| **Electrical** | 7.5 | 12.7 | 15.8 | 22.8 | 31.2 |

**Table 9: Power consumption (mW) in an optical interconnect compared with the electrical one**

| *Year* | *2004* | *2007* | *2010* | *2013* | *2016* |
|---|---|---|---|---|---|
| **Technology node** | 90 nm | 65 nm | 45 nm | 32 nm | 22 nm |
| **Modulator driver** | 83.7 | 45.8 | 25.8 | 16.3 | 9.5 |
| **Modulator** | 114.0 | 52.1 | 30.4 | 20.0 | 14.3 |
| **Waveguide** | 46.7 | 46.7 | 46.7 | 46.7 | 46.7 |
| **Photo-detector** | 1.4 | 0.5 | 0.3 | 0.3 | 0.2 |

Despite the performance advantages, if we factor in cost, this technology is nowadays prohibitive. Still, there is much research and test-chips developed. The latest works in the area, point out that the implementation in III/V materials should be simpler and integration should become an easier matter. Given that III/V materials are expected to become a reality in the next few generations for logic processes, it seems that eventually, optical networks will stand a chance in the semiconductor market.

# 4. Technologies Beyond the Scope of CLERECO

CLERECO is targeting technologies until 2020. Still, there are a number of promising emerging technologies that will not be ready for mass manufacturing within this decade and require further research from the community. They are included in this section.

## 4.1. Post CMOS

### *4.1.1. Gate-All-Around, Nanowires*

Beyond the multi-gate (FinFET) structure, a natural progression would be the gate-all-around (GAA) nanowire structure [***27***]. This is the ultimate structure in terms of electrostatic control to scale to the shortest possible effective channel length. To accurately project the device performance, 3-D simulation is necessary and it demands much more effort. There is not enough information available on these devices yet, but there is the general agreement that they are the most evolutionary approach.

### *4.1.2. Tunnel FETs*

As scaling continues, the power density of the IC continues to go up with the transistor density, although the power per transistor goes down. An effective solution would be based on transistor actions that do not depend on the Boltzmann distribution which sets a lower limit of subthreshold slope of 60 mV of gate voltage per decade of channel current. One such conduction mechanism is tunneling. A class of transistor based on this effect is called tunneling FET (TFET) [***28***]. It is basically a p-n junction placed under an MOS gate. With a proper design of the heterojunction under the gate, ultra-low voltage operation is the goal.

### *4.1.3. Spin-logic*

Contrary to its storage counterpart, spin-logic (also referred as spintronics) will not be ready until 2020. It is because of non-existing manufacturing processes as well as requirements to redefine/redesign logic, since it is difficult/expensive to build computing structures used in CMOS (i.e. it re-quires multiple spin cells to perform an AND operation, but more efficient in building structures for comparison, etc.).

### *4.1.1. 3D integration*

As pointed out in section 3.4, the increasing number of cores and the increasing pressure on the interconnection network demands for novel architectures. The most promising solution to overcome the delay, power and bandwidth problem faced is the manufacturing of 3D chips. Figure 16 shows an example of such. In the figure, we can see how three different layers (or planes) of silicon are stacked together to provide a single solution. Initial efforts were made into manufacturing all the structure in a single wafer. This proved to be utterly expensive with high

yield-costs. Nowadays, the most promising solution is the stacking of different (i.e. independently manufactured) chips. In this scenario, inter-plane vias or Through-Silicon-Vias (TSVs) provide the necessary communication among the planes. This solution has several advantages: Reduced interconnect delay (i.e., shorter paths), the possible combination of disparate technologies (i.e., each plane can be implemented in a different technology) and a higher yield (i.e., stacking of good-chips, only). But also, there are drawbacks: increased crosstalk noise, inter-plane via density and, worst of all, thermal density.

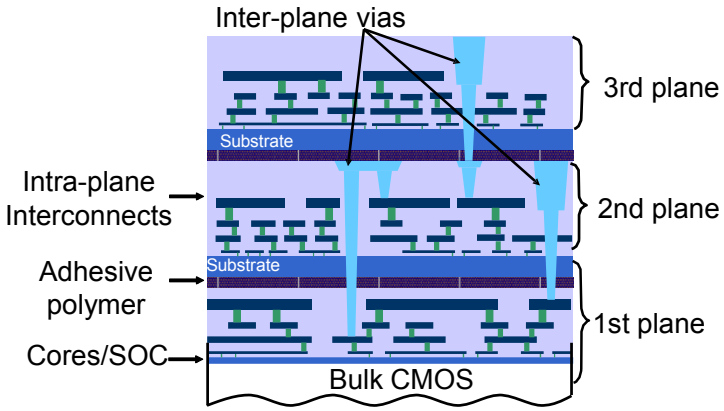



**Figure 16: 3D interconnected system**

The different planes in a 3D integrated chip do not necessarily have to be dedicated to the interconnect. For instance, they could also be used for stacking memory, as adopted by MICRON [**29**].

## 4.1. Beyond CMOS

Eventually later in the roadmap, more forward-looking solutions in the utilization of alternate channel materials to further enhance the transport will be adopted. It is anticipated the first solutions would be III-V (for n-channel) and Ge (for p-channel) combination, still based on MOSFET s the first product will be introduced in 2018. Other possibilities beyond these semiconductors are 2-D crystals. These include graphene, boron nitride (BN), dichalcogenides such as $MoS_2$, $WS_2$, $NbSe_2$, and complex oxides such as $Bi_2Sr_2CaCu_2Ox$.

Finally, beyond the outlook of this report, MOSFET scaling will likely become ineffective and/or very costly. Completely new, non-CMOS type of logic devices and maybe even new circuit architecture are potential solutions. Such solutions ideally can be integrated onto the Si-based platform to take advantage of the established processing infrastructure, as well as being able to include Si devices such as memories onto the same chip.

### 4.1.1. Carbon nanotubes

Carbon nanotube field effect transistors (CNFETs) are promising candidates as a potential extension to silicon transistors [***30***] [***31***]. With extraordinary electrical properties, such as quasi--ballistic transport and higher carrier mobility, CNFETs exhibit characteristics that surpass those of state-of-the-art and predicted Si-based MOSFETs.

In a CNFET the role of the channel is played by one or more CNTs for reliability reasons. A CNT is a graphene sheet rolled up to form a hollow cylinder. It can exhibit a metallic or a semiconducting behavior and it can present different diameters (at nanometer scale) depending on its chirality (i.e., angle of the atom arrangement along the tube). The first carbon nanotube transistor was fabricated in1998 at Delft University (The Netherlands). While there is still a lot of research to make this a feasible solution, its properties make them an ideal candidate – specially- to substitute electrical interconnects for inter-core communications.

### 4.1.1. Holographic and Molecular Storage technologies

Fancy storage technologies that are still in early stages of research are not expected to emerge until 2030 the earliest. Thus, they are fare beyond the scope of CLERECO.

Although holographic memory has been discussed since the 1960s, just recently, a team of researchers from the University of California, Riverside Bourns College of Engineering and Russian Academy of Science have demonstrated a new type of holographic memory device that could provide unprecedented data storage capacity and data processing capabilities in electronic devices [***32***]. The new type of memory device uses spin waves – a collective oscillation of spins in magnetic materials – instead of the optical beams. Spin waves are advantageous because spin wave devices are compatible with the conventional electronic devices and may operate at a much shorter wavelength than optical devices, allowing for smaller electronic devices to have greater storage capacity.

Molecular memory is a term for data storage technologies that use molecular species as the data storage element, rather than e.g. circuits, magnetics, inorganic materials or physical shapes. The molecular component can be described as a molecular switch, and may perform this function by any of several mechanisms, including charge storage, photochromism, or changes in capacitance. In a perfect molecular memory device, each individual molecule contains a bit of data, leading to massive data capacity. Several universities (e.g. MIT), research labs (e.g. NASA), and a number of companies (e.g. Hewlett Packard) have announced work on molecular memories, which some hope will supplant DRAM memory as the lowest cost technology for high-speed computer memory.

# 5. Conclusions

This deliverable has analyzed the different technologies that are currently available and those foreseen for future systems. It has been necessary to distinguish between the different computing segments, as their needs are different.

Such an extensive analysis was fundamental to identify those technologies that are more likely to be important in the development of future systems and therefore to focus the project activities and resources to these technologies. Therefore we guarantee concrete and exploitable results for all project partners. Figure 17 summarizes the findings.

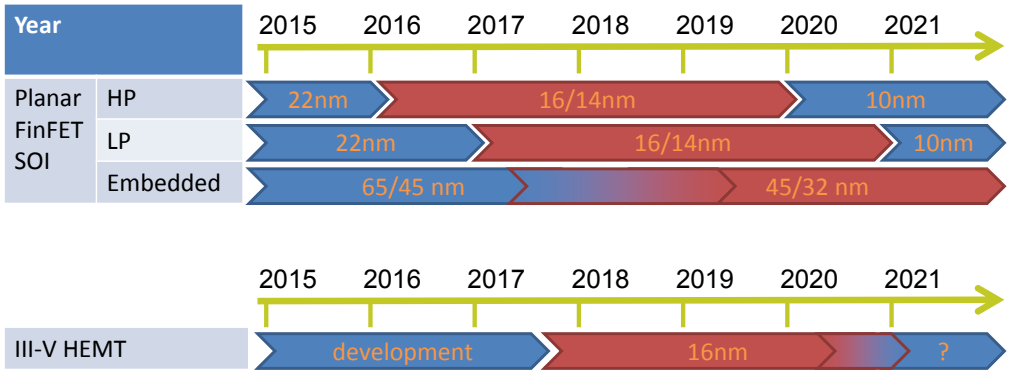



**Figure 17: Summary of technologies and their adoption in the future systems**

As CLERECO is focused on the near-term technologies, we will focus on all alternatives foreseen before 2020. This means: FinFET (22nm and 14nm), SOI (65nm and 45nm) and III-V HEMT (16nm). As mentioned in the introduction of this document, these technologies are going to be analyzed during the project in the framework of the WP2 activities in order to provide reliability data to be exploited as a knowledge base for the development of the reliability evaluation activities performed in WP3, WP4 and WP5; and for the development of the project demonstration activities of WP6.

# 6. Acronyms and Definitions

| | |
|---|---|
| CMOS | Complementary Metal-Oxide Semiconductor |
| CUB | Capacitor Under Bitline |
| DRAM | Dynamic Random Access Memory |
| ECC | Error Correction Code |
| eDRAM | Embedded Dynamic Access Memory |
| FET | Field Effect Transistor |
| HEMT | High Electron Mobility Transistor |
| IC | Integrated Circuit |
| MIM | Metal Interdielectric Metal |
| MOSFET | Metal-Oxide-Semiconductor Field-Effect Transistor |
| MTH | Magnetic Tunnel Junction |
| SRAM | Static Random Access Memory |
| STT | Spin-Torque Transfer |

# 7. Bibliography


[1] D. Guedes, W. Meira, R. Bianchini B. Diniz, "Limiting the power consumption of main memory," *ACM SIGARCH Computer Architecture News*, vol. 35, no. 2, pp. 290-301, 2007.

[2] Stephen and Engelmann, Christian Scott, "Advancing Reliability, Availability, and Serviceability for High Performance Computing," Oak Ridge National Laboratory, 2006.

[3] Scott E Thompson et al., "In Search of "Forever," Continued Transistor Scaling One New Material at a Time," *IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING*, vol. 18, no. 1, pp. 26-36, 2005.

[4] C Auth et al., "A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors," in *2012 Symposium on VLSI Technology (VLSIT)*, 2012.

[5] D Hisamoto, T Kaga, Y Kawamoto, and E Takeda, "A fully depleted lean-channel transistor (DELTA) - a novel vertical ultrathin SOI MOSFET," *IEEE Electron Device Letters*, vol. 11, pp. 36-39, 1990.

[6] M LaPedus, "Who's Winning The FinFET Foundry Race?," *Semiconductor Engineering*, 2014.

[7] E H Cannon, D D Reinhardt, M S Gordon, and P S Makowenskyj, "SRAM SER in 90, 130 and 180 nm bulk and SOI technologies," in *42nd Annual IEEE International Reliability Physics Symposium Proceedings*, 2004.

[8] S Kiamehr, T Osiecki, M Tahoori, and S Nassif, "Radiation-induced soft error analysis of SRAMs in SOI FinFET technology: A device to circuit approach," in *51st ACM/EDAC/IEEE Design Automation Conference (DAC)*, 2014.

[9] A. S. and Adams, J. H. and Darty, R. C. and Patrick, M. C. and Johnson, M. A. and Cressler, J. D. Keys, "Radiation Hardened Electronics for Space Environments (RHESE)," in *IPPW-6*, 23-27 June 2008.

[10] Mu-Tien and Rosenfeld, Paul and Lu, Shih-Lien and Jacob, Bruce Chang, "Technology comparison for Large Last-Level Caches (L3Cs): Low-leakage SRAM, low write-energy STT-RAM, and refresh-optimized eDRAM," in *Proc. 19th International Symposium on High Performance Computer Architecture (HPCA 2013)*, Shenzhen China, 2013.

[11] J Y Kim and et al, "S-RCAT (sphere-shaped-recess-channel-array transistor) technology for 70nm DRAM feature size and beyond," in *2005 Symposium on VLSI Technology. Digest of Technical Papers.* , 2005.

[12] Sung-Woong Chung and et al, "Highly Scalable Saddle-Fin (S-Fin) Transistor for Sub-50 nm DRAM Technology," in *2006 Symposium on VLSI Technology. Digest of Technical Papers*, 2006.

[13] T Schloesser and et al, "6F2 buried wordline DRAM cell for 40 nm and beyond," in *IEDM Technical Digest*, 2008.

[14] ITRS 2013 Executive Summary. [Online]. http://www.itrs.net/Links/2013ITRS/Home2013.htm

[15] M Asano, H Iwahashi, T Komuro, and F Masuoka, "A new flash E2PROM cell using triple polysilicon technology," in *International Electron Devices Meeting*, 1984.

[16] F Masuoka, M Momodomi, Y Iwata, and R Shirota, "New ultra high density EPROM and flash EEPROM with NAND structure cell," in *International Electron Devices Meeting*, 1987.

[17] D A Buck, "Ferroelectrics for Digital Information Storage and Switching," Cambridge, 1951.

[18] Y K Hong and et al, "130 nm technology, 0.25 μm2, 1T1C FRAM Cell for SoC (System-on-a-Chip)-friendly Applications," in *2007 Symposium on VLSI Technology*, 2007.

[19] P Grünberg, R Schreiber, Y Pang, M B Brodsky, and H Sowers, "Layered Magnetic Structures: Evidence for Antiferromagnetic Coupling of Fe Layers across Cr Interlayers," *Physics Review Letters*, vol. 57, no. 19, pp. 2442--2445, 1986.

[20] M N Baibich, J M Broto, A Fert, and F Nguyen Van, "Giant Magnetoresistance of (001)Fe/(001)Cr Magnetic Superlattices," *Physics Review Letters*, vol. 61, no. 21, pp. 2472--2475, 1988.

[21] W Zhuang et al., "Novel colossal magnetoresistive thin film nonvolatile resistance random access memory (RRAM)," in *IEEE International Electron Devices Meeting, IEDM '02.* , 2002.

[22] P H Nielsen and N M Bashara, "The reversible voltage-induced initial resistance in the negative resistance sandwich structure," *IEEE Transactions on Electron Devices*, vol. 11, no. 5, pp. 243, 244, 1964.

[23] J F Gibbons and W E Beadle, "Switching properties of thin Nio films," *Solid-State Electronics*, vol. 7, no. 11, pp. 785-790, 1964.

[24] J G Simmons and R R Verderber, "New thin-film resistive memory," *Radio and Electronic Engineer*, vol. 34, no. 2, pp. 81-89, 1967.

[25] J W Goodman and et al, "Optical Interconnects for VLSI Systems," *Proceedings of the IEEE*, vol. 72, no. 7, pp. 850-866, 1984.

[26] G Chen and et al, "Predictions of CMOS Compatible OnChip Optical Interconnect," in *International Workshop on system level interconnect prediction*, 2005.

[27] N Singh et al., "High-performance fully depleted silicon nanowire (diameter ≤ 5 nm) gate-all-around CMOS devices," *IEEE Electron Device Letters*, vol. 27, no. 5, pp. 383-386, 2006.

[28] U Avci and I Young, "Heterojunction TFET Scaling and Resonant-TFET for Steep Subthreshold Slope at Sub-9nm Gate-Length," in *IEDM*, 2013.

[29] Jon Fingas, "Hybrid Memory Cube receives its finished spec, promises up to 320GB per second," *Engadget*, Apr. 2013.

[30] S J Tans , A R-M Verschueren, and C Dekker, "Room-temperature transistor based on a single carbon nanotube," *Nature*, vol. 338, pp. 49-52, 1998.

[31] R Martel, T Schmidt, H R Shea, and T Hertel, "Single- and multi-wall carbon nanotube field-effect transistors," *Applied Physics Letters*, vol. 73, no. 17, pp. 2447-2448, 1998.

[32] F Gertz, A Kozhevnikov, Y Filimonov, and A Khitun, "Magnonic Holographic Memory," *Applied Physics Letters*, to be published 2014.