Report on Future Technologies for Advanced Computation

Project Number: 101092562
Project Acronym: ICOS
Project Title: International Cooperation On Semiconductors
Responsible: CEA-Leti
Submission date: 11th October 2024

icos-semiconductors.eu

This publication is part of the work executed in Workpackage 3 (Technology scanning and foresight) of the ICOS project.

Responsible partner: CEA-Leti
Lead Author: Olivier Faynot (CEA-Leti)

Contributing Authors: VTT, Institut Polytechnique de Grenoble, SiNANO Institute, DECISION Etudes & Conseil, Tyndall National Institute, ST Microelectronics, Delft University of Technology, Infineon, VDI/VDE, imec, University of Ghent, IU.NET, Warsaw University of Technology (WUT) CEZAMAT, IMT-Bucharest, Fraunhofer IISB, Robert Bosch GMBH, NXP Semiconductors Netherlands BV, AENEAS

ICOS has received funding (2023 – 2025) from the European Union's Horizon Europe research and innovation programme under GA No 101092562.

Contents

Executive Summary

The days of CMOS technology with one-size-fits-all devices and architectures, no longer suffice to address the diverse needs of advanced computation systems – from artificial intelligence (AI) and machine learning (ML) requirements in graphic processing units (GPUs) and augmented and virtual reality (AR/VR) devices, to systems for autonomous vehicles and edge AI for the Internet of Things (IoT). Each of these emerging applications requires specialized solutions tailored to its specific performance, power, and latency requirements. As a result, there is a growing push towards more customized hardware architectures, including application-specific integrated circuits (ASICs), field-programmable gate arrays (FPGAs), and neuromorphic chips, to meet the unique challenges posed by the future of computing.

While Power-Performance-Area-Cost (PPAC) remains key for CMOS and DRAM, other key drivers for research and development in advanced computing have emerged to address the disparity between processor speed and memory performance ("memory wall"), the improved performance without a proportional increase in power consumption ("power wall") and sustainable manufacturing. In this report, several ingredients have been identified to maintain exponential growth in performance, although the rate of CMOS scaling has slowed since around 2008.

Scaling will continue to improve thanks to the current roadmaps for FinFETs, nanosheets, forksheets, complementary FET and Fully-Depleted Silicon-On-Insulator (FDSOI) architectures. Further advances will be driven by the introduction of new materials and devices for both logic and memory. Connectivity improvements will enhance performance and upcoming technologies like 6G wireless might even be used for improvement at a finer scale. In addition, new compute architectures will be essential to meet the growing diversity of applications. These will likely include SoC architectures enabled by CMOS 2.0, and specialized brain-inspired computing and quantum accelerators for specific tasks, designed to overcome the limitations of traditional computing systems.

International cooperation is essential for accelerating technological innovation and strengthening semiconductor value chains and is in line with the objectives of the EU Chips Act. In this context, the report does not only lay out the above technologies but also identifies various topics for EU to be active.

1 Overview

1.1 Purpose

The purpose of this deliverable is to present the findings from the analysis of existing international roadmaps (e.g., IRDS, IPSR-I, ECS-SRIA, NEREID) and their connections to other relevant CSAs (e.g., those focused on graphene and quantum computing), as well as insights from previous studies, knowledge from the consortium and partner network, desk research, and brainstorming sessions. This analysis focuses on future technologies in advanced computation that hold potential for exploitation through collaboration. Additional objectives include creating a comprehensive map of international research on future technologies and comparing the European status with that of other leading countries in areas identified as critical for Europe.

2 Future Technologies for Advanced Computation

2.1 Trends

With the world's digitalization, the amount of generated data is seen as exponentially growing in the coming years (Fig.1), not only through human activity, but also through smart factories, transportation, smart grids, and other critical infrastructures. The forecast of these data is already believed to be obsolete, with the huge wave of Artificial Intelligence that will dramatically impacts the generation of data.

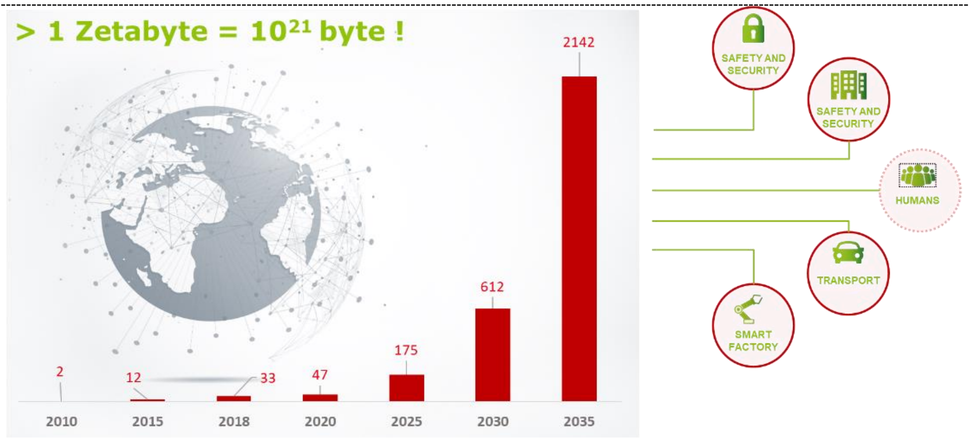
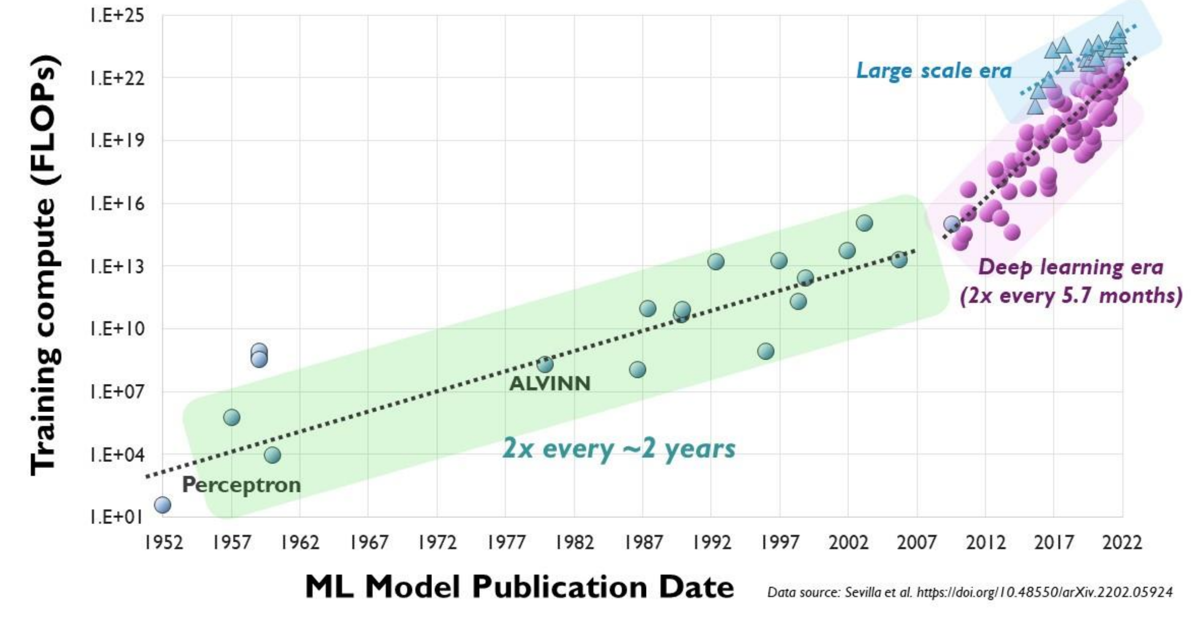


Fig.1: Evolution of the global data generation (actual and forecast)

Unsurprisingly, today's computational needs are largely dictated by AI and machine learning (ML) advancements, as demonstrated in Fig. 2. These technologies have fuelled an insatiable demand for processing power, with compute needs evolving significantly over the years.

To better understand this evolution, let's examine compute needs through the lens of machine learning. From the early days of ML in the 1950s, compute needs were modest and progressed steadily, largely following Moore's Law — doubling every two years as transistor density improved. However, since around 2010, the landscape has transformed dramatically. The advent of deep neural networks and more efficient training algorithms has caused compute requirements to skyrocket. In contrast to the previous decades, compute needs are now doubling every six months. This dramatic increase can be attributed to both the complexity and the scale of modern ML models.

In recent years, the emergence of extremely large models—some containing over a trillion parameters—has further driven up the demand for computational power by orders of magnitude beyond that required for conventional deep learning models. These massive models, designed to tackle increasingly complex tasks, highlight the unprecedented scale at which computational infrastructure must operate to support the future of AI. This trend emphasizes the need for advancements not only in AI algorithms but also in the hardware and infrastructure required to meet the growing demands of next-generation applications.



Data source: Sevilla et al. https://doi.org/10.48550/arXiv.2202.05924

Fig.2: Compute needs (number of floating points operations per second) as function of the introduction year of machine learning (ML) models

AI and ML have been the primary drivers of GPU advancements, as the parallel processing capabilities of GPUs are uniquely suited to handle the immense computational demands of training and running complex machine learning models. Beyond the use of GPUs for AI training and inference, a variety of emerging applications are driving unique workloads that demand different hardware capabilities and technology specifications. Each of these applications places specific requirements on compute infrastructure that go beyond traditional GPU-focused solutions.

For instance, GPUs used for training ML models must support high-throughput parallel compute. They require not only the ability to handle massive data processing tasks but also the capacity for high-bandwidth data movement, both within the system's DRAM and between other GPUs in multi-GPU configurations.

In contrast, augmented and virtual reality (AR/VR) solutions place a premium on low power consumption and compact form factors. These systems must also support high data bandwidths with minimal latency to seamlessly display high-definition video, ensuring a smooth and immersive user experience. The balance between power efficiency and performance is critical for mobile and wearable AR/VR devices, where energy constraints and form factor limitations are key design challenges.

Similarly, autonomous vehicles present their own set of demands. These systems rely on real-time data aggregation from multiple sensors, such as cameras, LiDAR, and radar. The compute architecture must be capable of processing and interpreting this sensor data quickly enough to enable real-time decision-making and responsive action. This requires low-latency, high-reliability compute systems optimized for real-time inference in unpredictable environments.

The days of CMOS technology with one-size-fits-all devices and architectures may no longer suffice to address these diverse needs. Each of these emerging applications requires specialized solutions tailored to its specific performance, power, and latency requirements. As a result, there is a growing push towards more customized hardware architectures, including application-specific integrated circuits (ASICs), field-programmable gate arrays (FPGAs), and neuromorphic chips, to meet the unique challenges posed by the future of computing.

2.2 Challenges for the semi-conductor industry in terms of data deluge

Needless to say, the semi-conductor industry is facing a number of challenges. We can identify four critical areas that need to be addressed to support the continued growth and development of advanced computing technologies:

1. **CMOS and DRAM PPAC (Power-Performance-Area-Cost):** As technology scales, traditional CMOS and DRAM technologies are hitting limits in terms of their ability to improve power efficiency, performance, area (size), and cost simultaneously. These components are essential for many computing systems, but new innovations are required to enhance their PPAC metrics to keep up with growing computational demands. Fig. 3 (left figure) shows that while traditionally the CPU single-thread performance has grown significantly during the early years, with an impressive rate of 50% improvement per year up to around 2008, today this rate has dramatically slowed down to a mere 5% annual improvement.

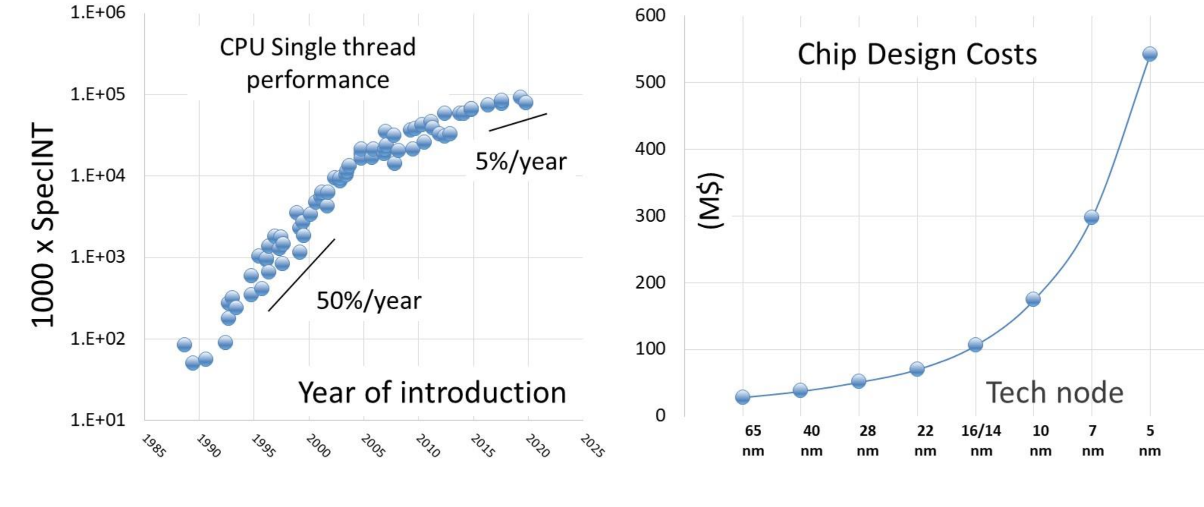


Fig.3: (left) CPU performance as function of the introduction year; data taken from M. Horowitz, F. Labonte, O. Shachan, K. Olukotun, L. Hammond, C. Batten. Additional data compiled by K. Rupp; (right) Design costs as function of CMOS technology node. Source: "AI Chips and why they matter", S. Khan and A. Mann, 2020

Next to that, the complexity of advanced CMOS technology nodes has led to increasing costs not only in developing and manufacturing the technology but also in designing in these nodes. As the semiconductor industry moved towards smaller nodes (7nm, 5nm, etc.), the cost of designing chips has skyrocketed, as shown in the graph on the right. This raises the question: While we design these ICs for AI/ML applications, how far can AI help in addressing the design cost challenges?

AI-driven tools have the potential to optimize various stages of chip design, such as automating parts of the design process, improving verification and testing, and enhancing predictive modeling for performance and thermal management. Additionally, AI could help in balancing the trade-offs between power, performance, and area (PPA), potentially allowing designers to create more efficient chips at lower costs. However, it's still an open question whether AI can substantially mitigate the rising costs and complexity.

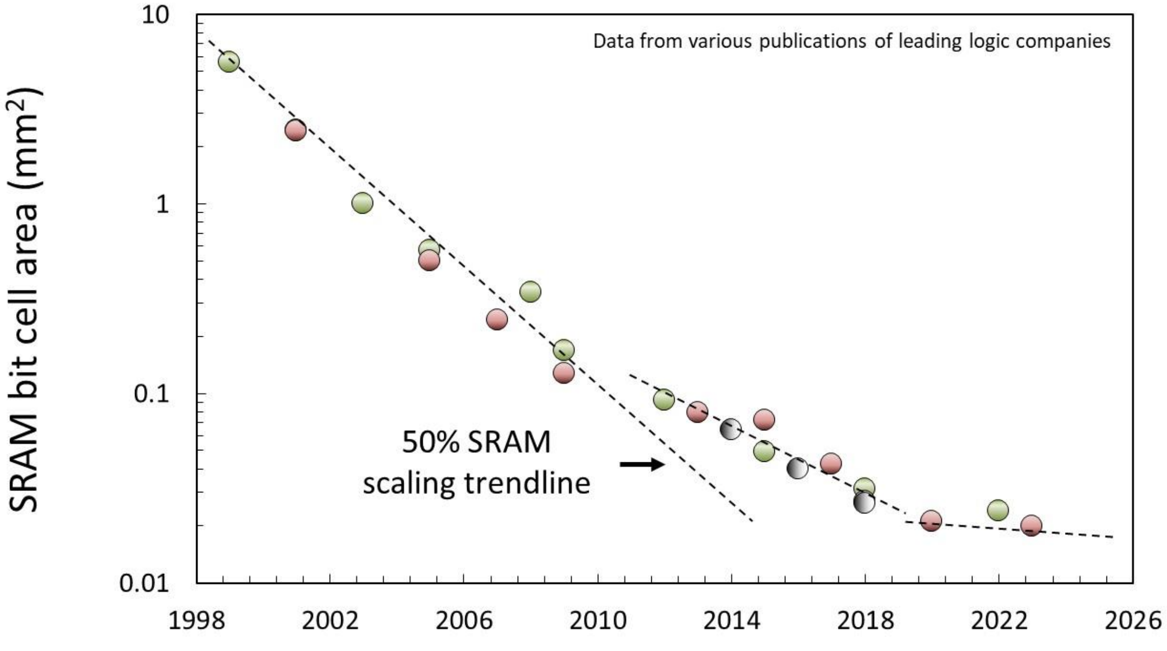


Fig.4: SRAM size as function of introduction year.

Historically, SRAM, a fundamental building block in logic technology, served as the benchmark for scaling. Each new technology node used to achieve a 0.7x linear scaling, translating to approximately 50% area reduction for SRAMs, which helped improve both performance and power efficiency (Fig. 4).

However, around 2008, this consistent progression began to falter as gate-pitch scaling, a critical factor in transistor density improvements, started to slow down. In recent years, the slowdown has become even more pronounced, significantly affecting the ability to sustain previous rates of miniaturization and efficiency gains. As a result, alternative technologies and architectural innovations are increasingly being explored to maintain progress in the post-Moore’s law era.

2. **Memory Wall:** The "memory wall" refers to the growing disparity between processor speed and memory performance. As compute power increases, the ability to access memory quickly enough becomes a bottleneck, slowing down overall system performance. Overcoming this requires innovations in memory architectures and technologies to ensure fast and efficient data transfer between memory and processors.

3. **Power Wall:** The power wall represents the challenge of improving performance without a proportional increase in power consumption. As compute workloads grow, the energy required to power these systems becomes a limiting factor. Addressing this issue involves optimizing power efficiency, especially for applications like AI and machine learning that demand high computational throughput.

4. **Sustainable Manufacturing:** The environmental impact of manufacturing is an increasing concern. The IC manufacturing processes have a significant environmental impact, and as

demand for integrated circuits continues to rise, the need for sustainable solutions grows [1]. The information technology sector is estimated to contribute up to 5% of total global emissions in terms of equivalent CO_2 [2]. This includes the impact of all devices, from datacenters to mobile phones, as well as their use, manufacturing, and the energy consumed by networks. Although only a portion of this impact can be attributed to semiconductor manufacturing, it remains a resource-intensive process, consuming large amounts of energy, water, chemicals, and raw materials through a complex supply chain.

While the Life Cycle Assessment (LCA) is a well-established method for evaluating the environmental impact of products, conducting an accurate LCA for integrated circuits is still challenging due to the scarcity of current information. Much of the data used today is outdated, leading to errors and variability in the LCA of electronic products, which complicates efforts to assess their environmental impact accurately.

Closing this data gap requires collaboration across the industry. Stakeholders must work together to improve data collection, develop new measurement techniques, and establish common standards and frameworks for reporting environmental impact. Such efforts will promote transparency, accountability, and innovation, while ensuring that the industry contributes to addressing pressing environmental challenges. One of these platforms is imec.netzero [3]. This is a modeling platform which provides a detailed, bottom-up view of the environmental impact associated with various semiconductor manufacturing processes. By simulating the energy, water, mineral consumption, and greenhouse gas emissions across different technology nodes, imec.netzero helps the semiconductor industry identify and reduce high-impact processes, contributing to its goal of achieving net-zero emissions by 2040.

In the remainder we will address some of the challenges mentioned in previous paragraphs in more detail, as well as their potential solutions.

2.3 The ingredients for continued compute system scaling

Despite the challenges mentioned before, future compute systems are expected to maintain exponential growth in performance, although the rate of CMOS scaling has slowed since around 2008. Despite this, advances will be driven by the introduction of new devices and materials for both logic and memory. Connectivity improvements, particularly through 3D integration and photonics, are already enhancing performance, and upcoming technologies like 6G wireless are poised to play a crucial role in improving e.g. datacenter connectivity but might even be used at a finer grained scale.

Additionally, new compute architectures will be essential to meet the growing diversity of applications. These will likely include SoC architectures enabled by CMOS 2.0 and specialized quantum accelerators for specific tasks, designed to overcome the limitations of traditional computing systems.

2.3.1 CMOS scaling but not as we (used to) know it

Transistors will remain vital to general-purpose computing, despite the emergence of new paradigms like e.g. quantum computing. While quantum technology shows great promise for specific complex problems, it is unlikely to replace traditional computing entirely. The efficiencies gained through decades of optimizing CMOS technology will continue to be leveraged.

Over the last decades, the industry has defied expectations through breakthroughs in EUV lithography, new materials, and device architectures like FinFETs and high-k dielectrics. These innovations have extended Moore’s law. Especially EUV has been crucial to extend the CMOS scaling roadmap. By the 2nm technology node, it is expected that the fourth generation of EUV patterning will be employed. However, pitches below 28nm will require multiple-patterned EUV steps, which add significantly to the complexity, cost and ecological footprint of the process.

To address this challenge, high numerical aperture (NA) extreme ultraviolet (EUV) systems with a 0.55 NA are projected to become available around 2026-2027. This next-generation EUV technology will allow some multi-patterned pitches to be reduced to single-exposure steps, simplifying the process and reducing costs while advancing scaling capabilities.

Next to that, Fully-Depleted Silicon-On-Insulator (FDSOI) technologies can be seen as complementary to the FinFET-based track. It is a semiconductor technology that enables efficient performance and energy savings, making it highly suitable for low-power applications. It is particularly valuable in mobile devices, automotive applications, IoT (Internet of Things) devices, and RF communications due to its ability to reduce power leakage and operate efficiently at low voltages. Additionally, FDSOI is known for providing excellent control over transistor behavior through back-biasing, which allows fine-tuning of performance and power trade-offs.

The FDSOI roadmap includes continued development of smaller nodes, with current technologies around 22nm and 12nm, and future plans targeting 10nm and beyond. This progression ensures that the technology will remain competitive in power-sensitive domains, offering a cost-effective alternative to FinFET in specific applications.
Going back to the FinFET-based CMOS scaling roadmap (Fig. 5), as feature sizes continue to shrink, the industry faces increasingly difficult physical and technical challenges. To overcome these, the transition from FinFET to nanosheets is underway for the 3/2nm technology node, with forksheet as the next step. This architecture allows for area scaling by reducing n-to-p spacing, thereby shrinking standard cell heights.

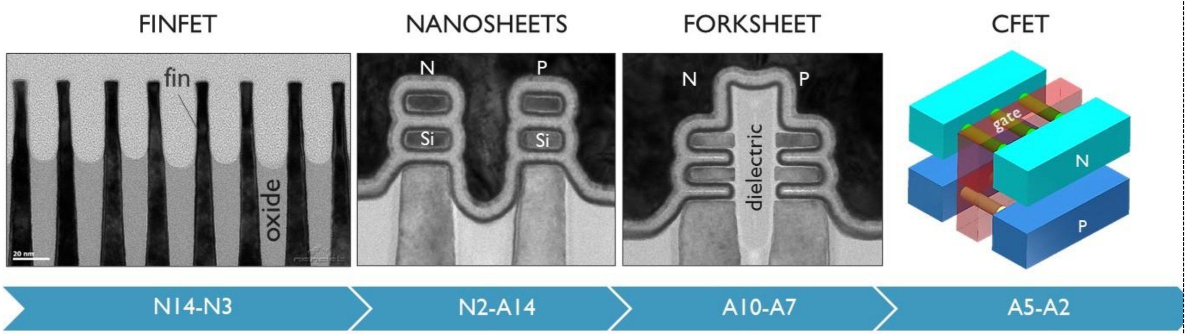


Fig.5: Scaling roadmap logic transistor. Nomenclature: Nxx=xx nm node, Axx=xx Å node [4]

Looking ahead, the semiconductor industry is exploring the complementary field-effect transistor (CFET) for post-1nm scaling. By vertically stacking n- and pMOS devices, CFET promises further optimization of channel width and drive current, offering another path for area scaling and potentially reducing cell heights to four tracks or less.

While CFET offers significant potential, it is pushing the limits of materials and process tools. Many of the required technologies are still in early R&D stages, making large-scale manufacturing of CFET devices a long-term prospect, likely more than a decade away.

2.3.2 New materials and device concepts

While Si-based transistors have dominated electronics for decades, the future likely involves a variety of new switch types built from different materials and governed by alternative physical principles. Among these, 2D materials are emerging as one of the most promising areas of research, thanks to their exceptional properties such as high mobility, flexibility, and tunability.

2D materials like tungsten disulphide (WS2) and hafnium disulphide (HfS2), part of the transition metal dichalcogenides (TMD) family, offer unique advantages in CMOS technology. Their atomic-thin layers enable conduction channels that can be scaled below 10nm, significantly reducing short channel effects while maintaining performance. Unlike the previous excitement surrounding III-V materials, which ultimately lost their edge over silicon at scaled dimensions, 2D materials maintain their high mobility and lower dielectric constants at nanometer-scale dimensions, making them better suited for continued scaling beyond silicon’s limits.

However, several challenges remain before 2D materials can be integrated into commercial transistors. Issues such as material growth quality, contact resistance, doping, and gate dielectric formation need to be resolved. Despite these obstacles, the use of 2D materials in nanosheet or CFET architectures appears highly likely, as their ultra-thin structures are ideal for stacking to improve drive current and reduce footprint.

There are additional opportunities to integrate 2D materials earlier in the development process. For example, they could be implemented as power switches in the backend or on the wafer's backside, where performance requirements are less stringent. In the context of CMOS 2.0, 2D materials can also serve as a low-capacitance, low-drive logic layer for driving short interconnects, offering a practical solution before full-scale integration into more performance-critical applications.

Another technology on the roadmap is the Dirac cold source FET. This concept leverages 2D materials to create a more energy-efficient transistor by reducing thermal excitations, offering significant power-saving advantages. While this technology is still in development, its potential to achieve near-ideal subthreshold swing makes it highly attractive for future low-power devices.

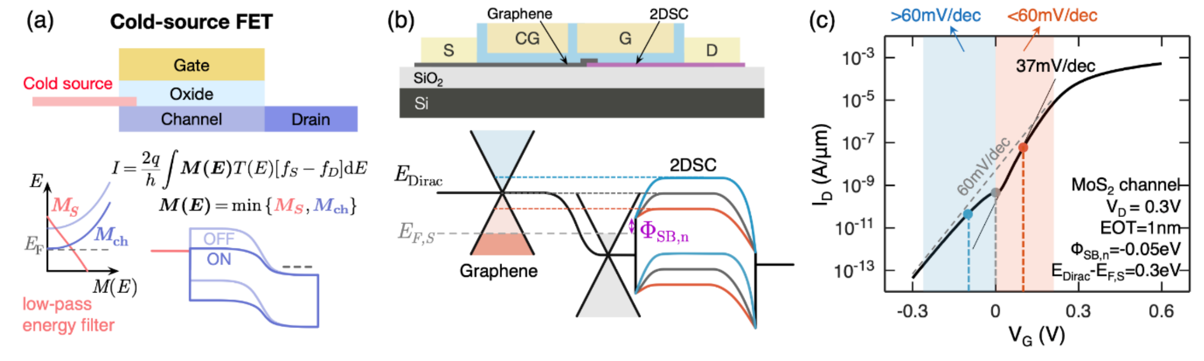


Fig.6: Operation principle of the Dirac cold source FET. Taken from [5].

In addition to 2D materials, carbon nanotubes (CNTs) are also gaining renewed attention. CNTs possess excellent electrical properties, including high mobility and current-carrying capacity, making them ideal for high-performance transistors. However, challenges like the difficulty in growing and aligning uniform nanotubes have delayed their widespread adoption although significant progress has been made over the last decade.

In general, the pros of 2D materials and CNTs include their ability to operate efficiently at ultra-small scales, their high mobility, and their potential for low-power operation. On the downside, material growth, integration, and contact formation remain significant technical challenges that need to be addressed before these materials can replace or complement silicon in mainstream electronics.

2.3.3 CMOS 2.0 – The next generation of scaling?

CMOS 2.0 represents a significant evolution in semiconductor technology, aiming to tackle the power, performance, and scaling limitations of traditional silicon-based CMOS as it reaches its physical boundaries. This new approach combines heterogeneous technologies, 3D transistor stacking, and advanced materials to enhance chip performance, efficiency, and scalability. The evolution from homogeneous platform (CMOS 1.0) to heterogeneous platform (CMOS 2.0) is shown in Fig. 7.

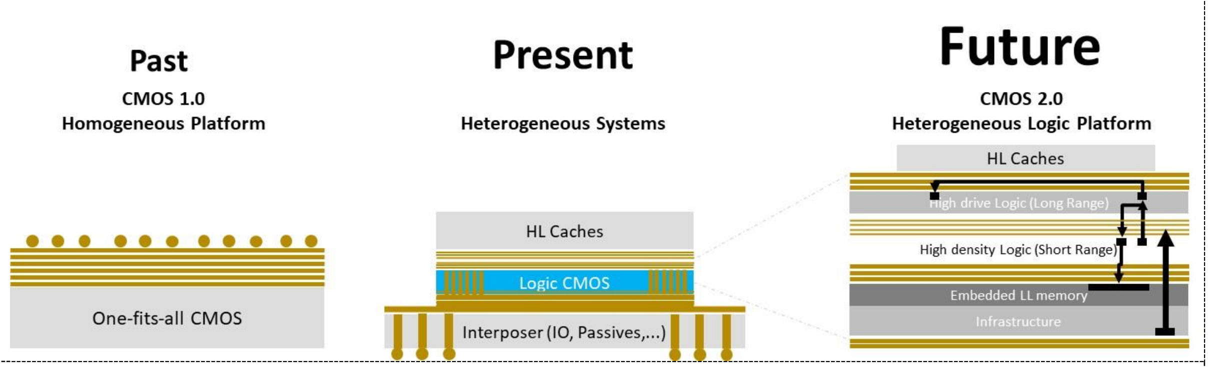


Fig.7: Evolution from a homogeneous platform where CMOS was a one-device-fits-all technology to a finer-grained heterogeneous platform.

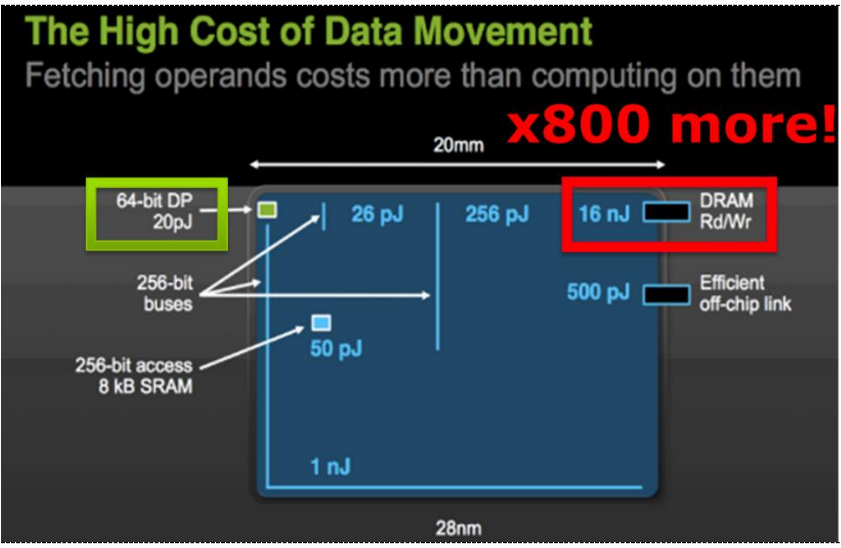
One of the key innovations is heterogeneous integration, where various components like logic and memory are brought together on a single chip, eliminating inefficiencies and reducing power consumption. Additionally, 3D stacking allows for the vertical layering of transistors, increasing component density without expanding the chip's physical size, which opens new possibilities for more powerful yet compact devices. Also, we will see more and more the use of the wafer and chip backside, going from implementing passive structures like backside power delivery connections to active devices.

CMOS 2.0 also incorporates the use of new, often dissimilar materials as compared to Si, such as e.g. GaN for power management. Another exciting development is neuromorphic computing, which mimics the structure of the human brain, enabling more energy-efficient processing for applications like artificial intelligence and advanced computing tasks. These innovations make CMOS 2.0 crucial for addressing the needs of next-generation technologies such as AI, IoT, and high-performance computing.

In general, CMOS 2.0 offers a platform where advanced compute and advanced functionalization can coexist and complement each other. However, with a zoo of different devices, features and materials, a System-Technology-Co-Optimization (STCO) framework is required to optimize the performance of these complex systems [6].

2.4 Addressing the memory wall

Among all the power consumption tasks in an Integrated Circuit (IC), data movement and storage are today known as the major ones, and consume much higher energy than that required to process the data, as illustrated in Fig.8 (800 times more energy for operand fetching). By moving away for the Von Neumann architecture, we can expect to reduce by 90% the energy consumption. Lots of efforts need to be done in that field.



The High Cost of Data Movement

Fetching operands costs more than computing on them

x800 more!

Fig.8: Part of the energy consumed in an IC, to compute or to store the data. Source: Bill Dally, “To ExaScale and Beyond”, 2010

By comparing the performances (Computation Performance in GOPS versus the Computation efficiency in GOPS/W) between existing ICs and bio Systems like Honey Bee brain or Human brain, we clearly see that the ICs are well under the performance of the bio systems (Fig.9).

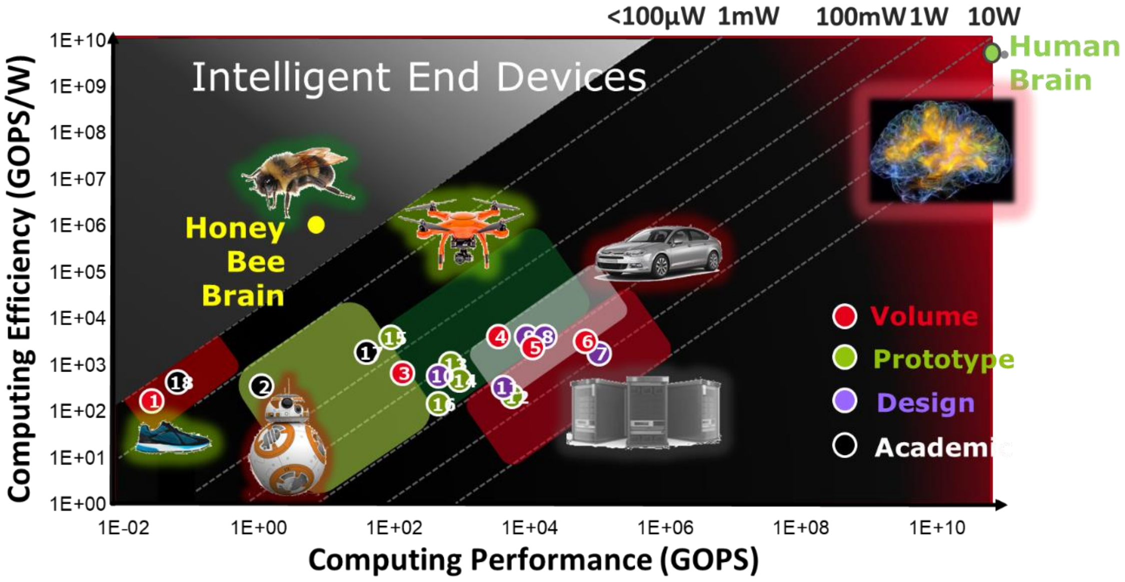


Fig.9: Comparison of the computing efficiency and the computing performance of existing Integrated circuits and biology systems.

2.4.1 How to mimic the Human Brain?

Try to mimic those bio systems is thus the key to some big breakthrough in terms of Power Efficiency. The use of Back-End-Of-Line (BEOL) Memories is appearing to be a key enabler as they allow to mimic neuron networks, as illustrated in Fig.10. This examples is shown for PCRAM but can be extrapolated to any other types of Memories (MRAM, OxRAM, FeRAM,…). Key performances of the existing Non-Volatile Memories are summarized in Fig.11, with a comparison of the Flash Technology available widely in the Industry.

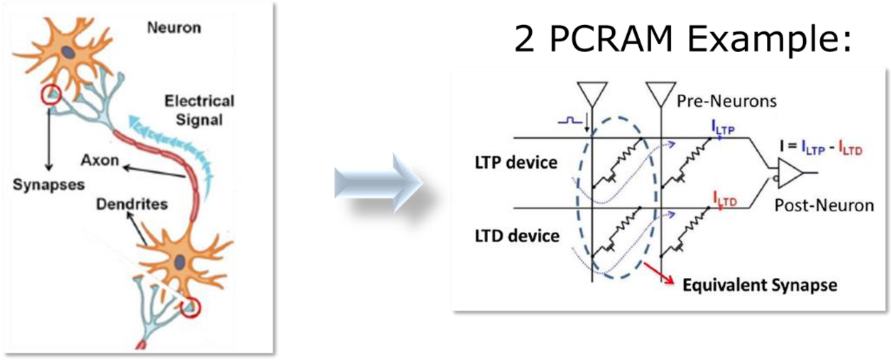
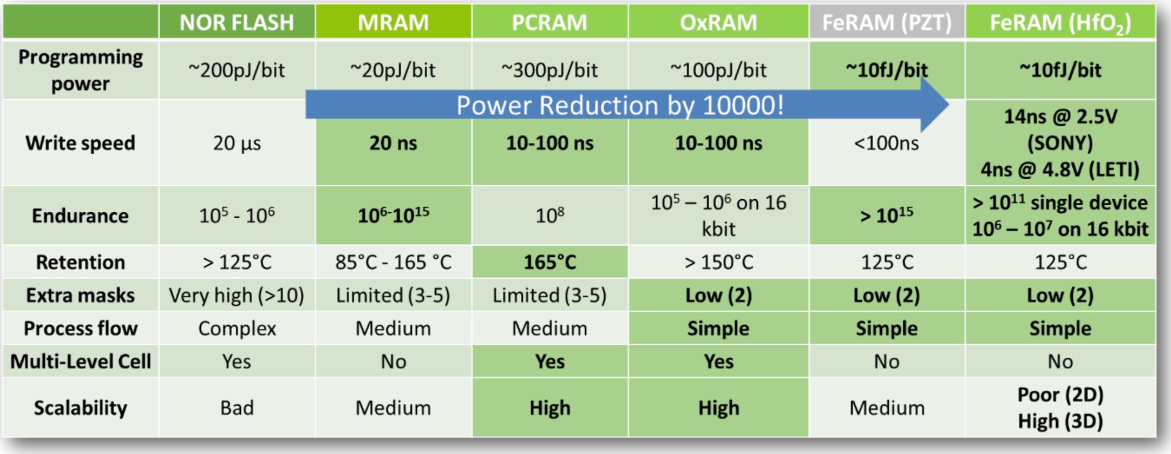


Fig.10: Neuron network simulation using PCRAM network.



	NOR FLASH	MRAM	PCRAM	OxRAM	FeRAM (PZT)	FeRAM (HfO_2)
Programming power	~200pJ/bit	~20pJ/bit	~300pJ/bit	~100pJ/bit	**~10fJ/bit**	**~10fJ/bit**
Write speed	20 µs	**20 ns**	**10-100 ns**	**10-100 ns**	<100ns	**14ns @ 2.5V (SONY) 4ns @ 4.8V (LETI)**
Endurance	10^5 - 10^6	**10^6-10^{15}**	10^8	10^5 – 10^6 on 16 kbit	**> 10^{15}**	**> 10^{11} single device 10^6 – 10^7 on 16 kbit**
Retention	> 125°C	85°C - 165 °C	**165°C**	> 150°C	125°C	125°C
Extra masks	Very high (>10)	Limited (3-5)	Limited (3-5)	**Low (2)**	**Low (2)**	**Low (2)**
Process flow	Complex	Medium	Medium	**Simple**	**Simple**	**Simple**
Multi-Level Cell	Yes	No	**Yes**	**Yes**	No	No
Scalability	Bad	Medium	**High**	**High**	Medium	**Poor (2D) High (3D)**

Power Reduction by 10000!

Fig.11: Comparison of the performance of the existing BEOL Memories.

Human brain exhibits typical characteristics, like asynchronous communication, plasticity, sensing, 3D structure and learning during our entire life. It is essential to be able to reproduce those characteristics in the Integrated Circuit. This can be done, as shown in Fig.12, by using Spiking coding, re-configurability, smart sensors, dense 3D integration and new algorithms allowed by the use of new technologies.

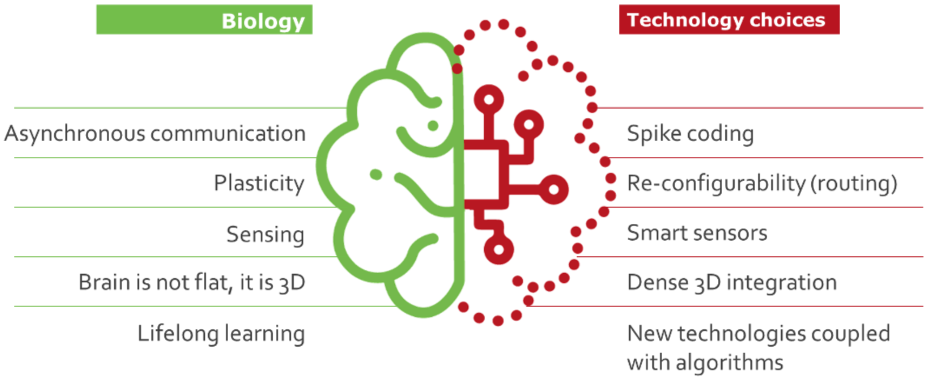

Biology
Technology choices
Asynchronous communication
Plasticity
Sensing
Brain is not flat, it is 3D
Lifelong learning
Spike coding
Re-configurability (routing)
Smart sensors
Dense 3D integration
New technologies coupled with algorithms

Fig.12: Characteristics of Human Brain with the corresponding feature for the mimic Integrated Circuit

By utilizing available on-chip memories, including SRAM and embedded Non-Volatile BEOL memories, it is possible to shift from the traditional Von Neumann architecture towards Near-Memory Computing. Logic and arithmetic operations can now be redefined using these memory resources, effectively bringing memory closer to the logic.

Resistive memories allow for further advancements by enabling In-Memory Computing, where memory units are directly involved in computation. This eliminates the energy consumption associated with data movement.

Unlike CMOS-based memories such as static or dynamic random access memories, which store one bit per unit cell, resistive memories can be programmed to intermediate states between their lowest and highest resistance values. This feature allows for the compact storage of synaptic weights in neural networks. Additionally, using fundamental laws of electric circuits, arrays of memristors can implement the core operation of deep learning—Multiply and Accumulate (MAC). In this case, the multiply operation is governed by Ohm’s law, while accumulation is performed according to Kirchhoff’s current law (Fig.13).

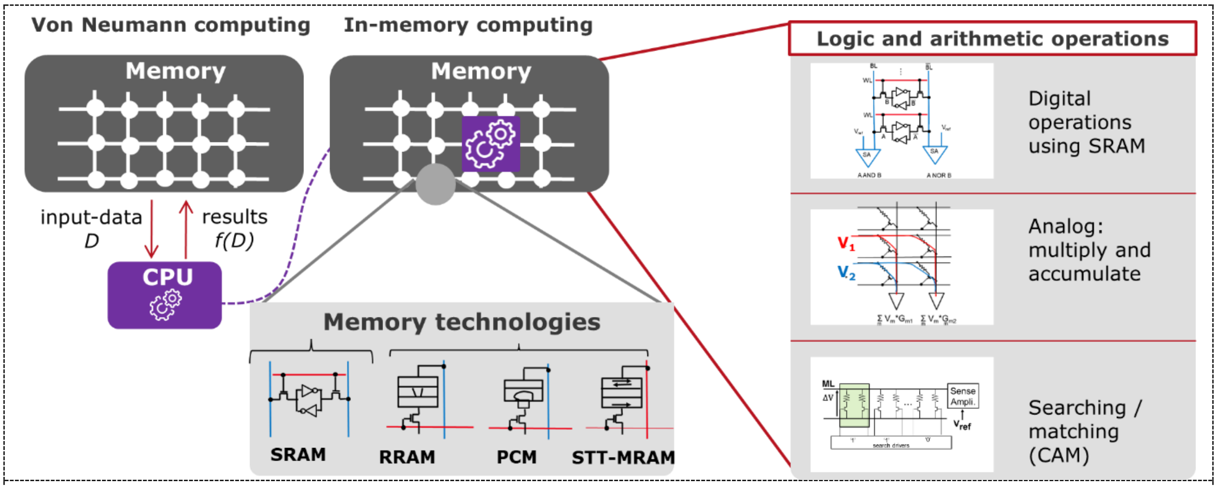


Fig.13: Evolution from Von Neumann architectures towards Near Memory computing, by implementing logic and arithmetic operations with Memory technologies

This concept has been successfully demonstrated on-chip, achieving high energy efficiency and performance in the tera operations per second range, with flexibility to support diverse models and accuracy comparable to software implementations. However, practical realization faces challenges due to memory variability, imperfections in analog CMOS circuits, and voltage drop effects. These challenges can be addressed through specialized programming schemes [7], circuit optimizations, or by combining memory arrays with emerging computing paradigms (see Fig. 14).

Binarized neural networks represent one such paradigm, where both synaptic weights and neuronal activations are limited to binary values (+1 and −1). This reduces the need for multi-level programming and, in turn, minimizes variability. These networks are well-suited for various embedded memory technologies, such as RRAM, PCM, and MRAM. Recently, a binarized neural network was demonstrated based on filamentary resistive memories and powered by a miniature wide bandgap solar cell optimized for edge applications [8].

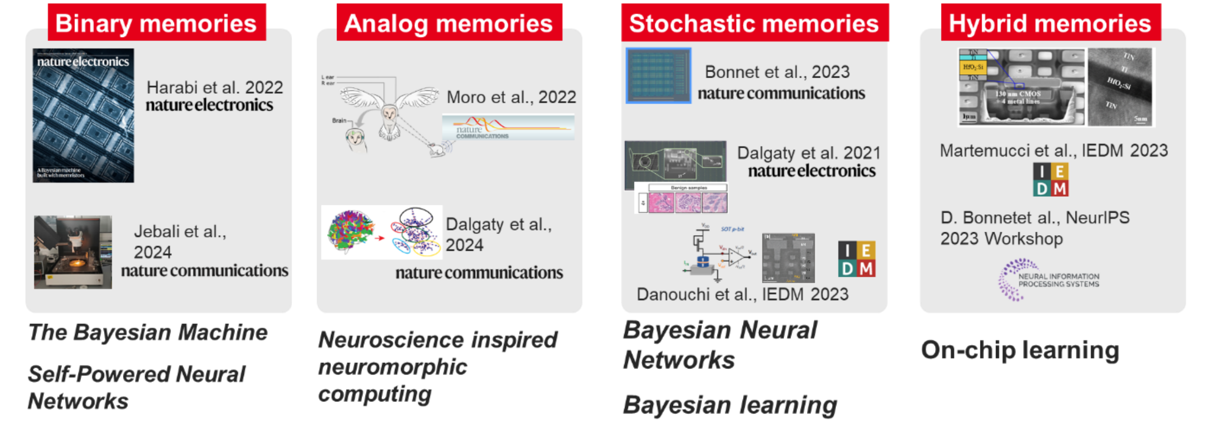


Fig.14: Different approaches pushed for In-Memory-Computing Artificial Intelligence

Another approach incorporates computational principles inspired by the brain into neuromorphic circuits and architectures. These circuits are naturally tolerant to variability, mirroring biological systems' ability to make accurate decisions despite operating on unreliable substrates with imprecise neurons and synapses. Unlike traditional artificial neural networks

(ANNs), biological neurons display complex internal temporal dynamics and interact through sparse, event-driven signals (spikes). The brain also processes information hierarchically, at different temporal scales—from milliseconds at the synapse level to seconds at dendrites and even minutes to hours within the neural network. Replicating these multi-timescale processes within compact CMOS technology remains a significant challenge. A promising solution is to leverage novel nanodevices, such as resistive memories, which have demonstrated the capability to implement varying time constants across synapses [9], neurons [10], dendritic arbors [11], and the broader neural network [12].

A third approach embraces the inherent randomness of memory devices, using their properties to compute efficient Bayesian algorithms. Bayesian neural networks offer a major advantage in sensory processing tasks, as they handle limited data effectively and estimate uncertainty with precision. In these networks, synaptic weights are not fixed values but are instead modeled as probability distributions. The natural variability in filamentary resistive memories and phase-change memories can represent these distributions as multi-level random variables [13], [14].

A fourth approach involves Ising networks, which utilize binary stochastic spintronic devices to generate stochastic bit streams, known as probabilistic bits or p-bits. While Magnetic Tunnel Junctions (the fundamental components of MRAM) may not offer the same degree of synaptic granularity as resistive memories, they excel in autonomous binary sequential sampling. This makes them particularly well-suited for physically implementing interacting binary stochastic neurons in Ising machines [15].

Despite significant progress in low-power AI using in-memory computing devices, current implementations are still limited to pre-programmed inference hardware. The next challenge is to develop adaptive learning capabilities that allow systems to handle real-world dynamics more effectively. Continual learning, which enables the accumulation of knowledge without catastrophic forgetting, is especially valuable for edge devices that frequently interact with environmental data. However, implementing edge AI remains difficult with current memory technologies, particularly when balancing the conflicting requirements for training and inference under strict energy constraints.

Previous approaches for edge models small enough to fit on-chip have used separate SRAM and RRAM macros for training and inference, respectively. This design results in area overhead and delays due to data transfer. To address this, the concept of "hybrid memory" was introduce which integrates two distinct memory technologies at the device level with fine granularity to enable on-chip learning.

The monolithic integration of two on-chip memory technologies—oxide semiconductor gain cells and Resistive RAM— was demonstrated into a compact joint memory cell on a Si CMOS platform [16]. Additionally, a unified memory stack was proposed based on silicon-doped hafnium oxide and a titanium scavenging layer. This stack functions as both a memristor and a ferroelectric capacitor and is integrated into the back-end-of-line (BEOL) of a standard CMOS process. This technology offers an efficient and cost-effective solution for AI devices with learning capabilities, requiring no additional masks and only a few extra process steps [17-18].

Finally, our vision is that hardware innovation will continue to meet the ever-increasing demand for computing power through 3D integration technologies. This technology enables the vertical stacking of logic, memory, and sensing components using unconventional fabrication processes. It can also be employed to increase on-chip storage capacity, allowing massive neural network weights to be fully hosted on-chip. Additionally, 3D integration reduces latency and power consumption due to shorter interconnects, improves bandwidth between stacked layers by enabling multiple connections, and facilitates the integration of heterogeneous layers, each optimized for a specific function [19].

2.4.2 3D and Chiplet approach to enable heterogeneous integration for Power Efficiency

It is now well established that the combination of different kinds of integrated circuits by 2.5 and 3D technologies is a way to combine the best of different technologies and the miniaturization to provide higher performance, flexibility and modularity with drastic energy efficiency benefits.

Instead of using a large monolithic die, that can suffer from a lower yield or increased parasitics due to a larger distance between CPU and Memories, it is well understood that moving to a chiplet concept is key. The chiplet approach is a disagregation of a large IC into smaller ICs with similar functionalities at smaller scale, approach (see Fig.15) that can significantly improve the overall performance.

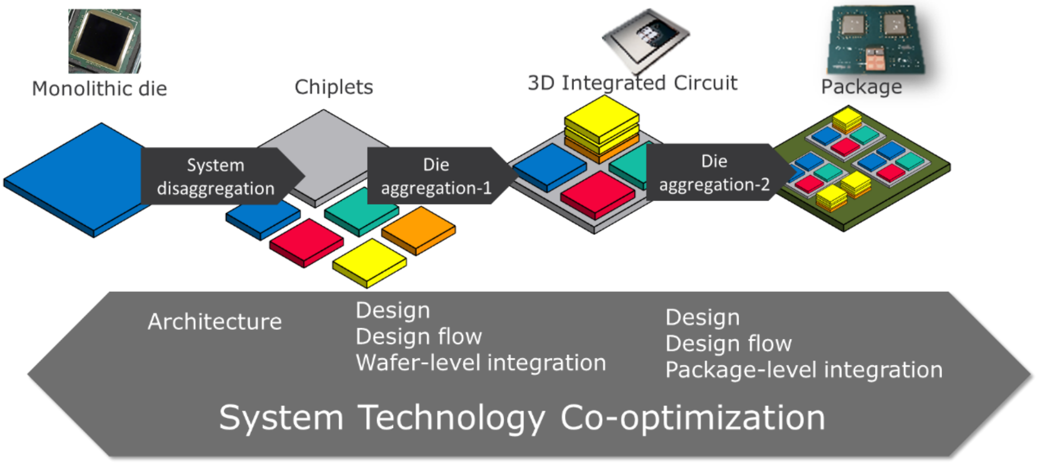


Fig.15: Disaggregation of a monolithic chip into several chiplets, re-aggregated together using 3D integration.

This chiplet approach is clearly enabled by advanced 3D technological bricks (high density through silicon vias, wafer-to-wafer and die-to-wafer direct hybrid bonding, 3D sequential integration… that are illustrated on Fig.16.

This toolbox provides different techniques, not necessarily equivalent in terms of density of connections. Among those techniques, the Hybrid Bonding Technology (based on Cu/SiO2 structures) appears to be the most promising one as it allows an interconnection pitch in the range of 1-2µm for the most advanced research groups (illustrated on Figs. 17 & 18).

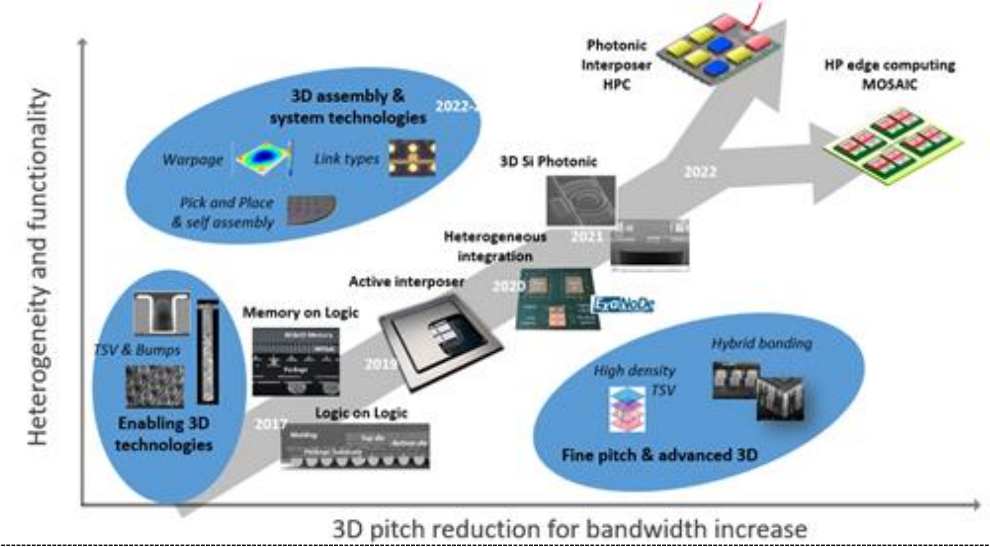
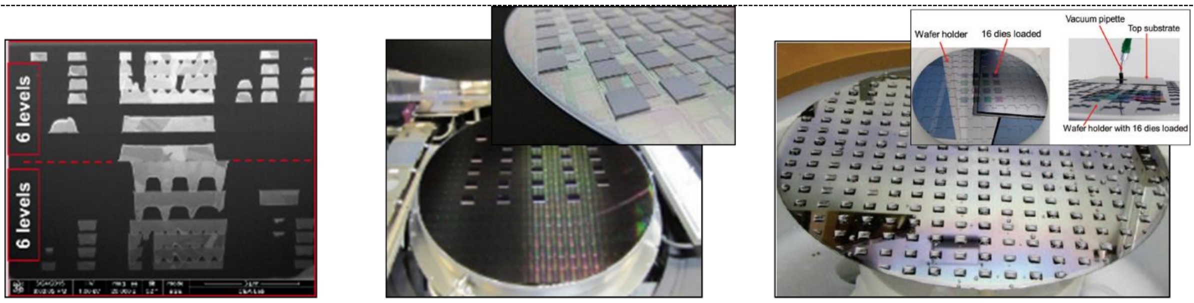


Fig.16: 3D integration technological toolbox used for chiplet approach



- Direct bonding of metal and dielectric
- Down to 1 micron pitch interconnects

- Wafer-to-wafer (W2W) or Die-to-wafer (D2W) technologies
- High heterogeneity allowed by D2W

- Collective D2W approaches
- Self-assembly for high precision & high throughput

Fig.17: Various options of Hybrid Bonding Technologies: Wafer to Wafer, Die to Wafer with or without the Self-Assembly technique.

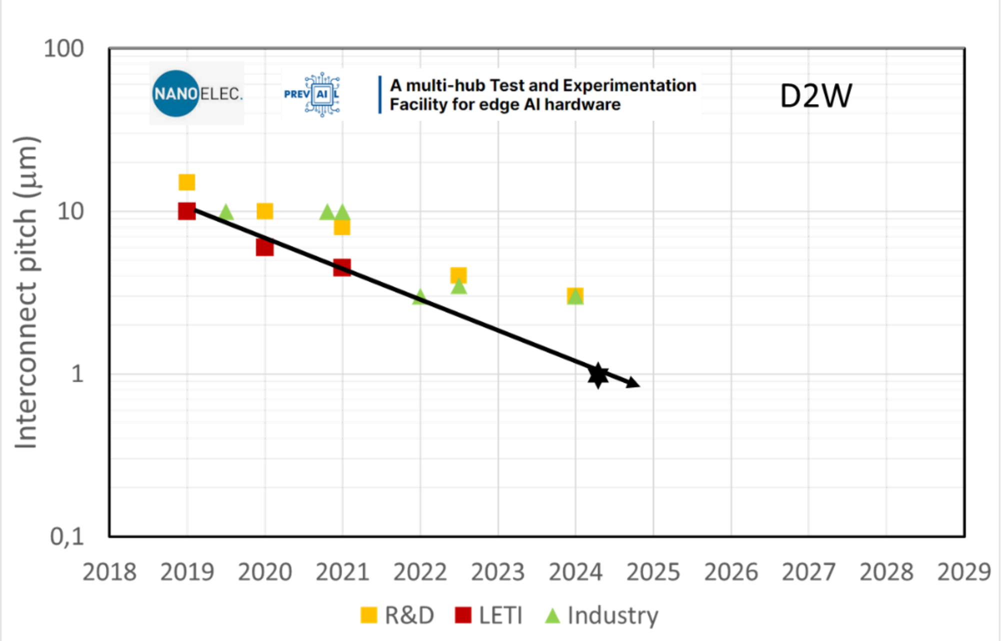


Fig.18: Alignment performance of Die to Wafer Hybrid bonding technology. Comparison of Industry and R&D data.

2.5 EU and non EU actors in the field of advanced Computing Technologies

It is very important to identify the fields where EU is active in terms of Research and Manufacturing. Preliminary radar plots have been elaborated for technologies on Advanced Computation during the ICOS project with the intent to highlight where EU is strong and where strong improvements are needed. Same plots have been done for US and Asia. Those plots are inserted in Fig. 19 to 21.

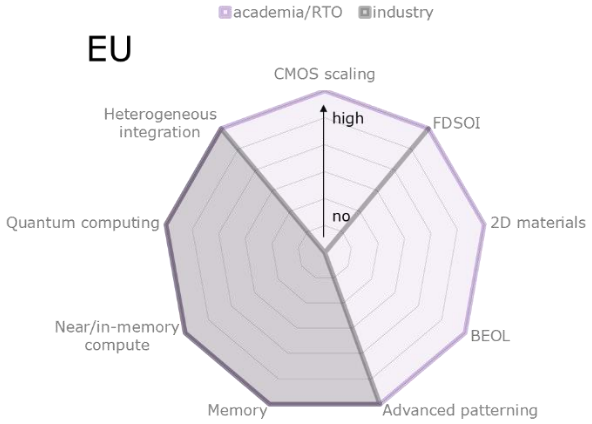


Fig.19: Radar plot highlighting the positioning of EU actors in the field of Advanced Computing.

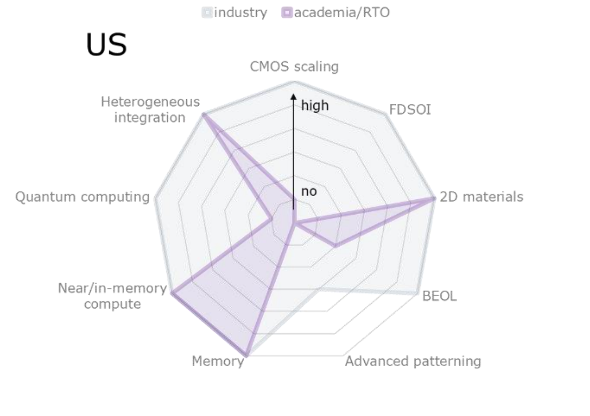


Fig.20: Radar plot highlighting the positioning of US actors in the field of Advanced Computing.

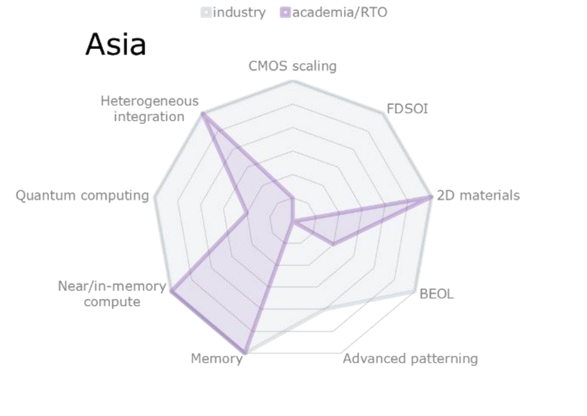


Fig.21: Radar plot highlighting the positioning of Asian actors in the field of Advanced Computing.

We can mention that EU is very strong in R&D in all the domains of Advanced Computing. This is mainly due to its strong RTOs (in 200 and 300mm) and its large portfolio of Universities in the different countries.

Regarding the manufacturing, we notice a weak part in the field of advanced CMOS technologies (with both FEOL and BEOL technologies), as well as in the field of advanced materials. For all the other topics (Memories, Heterogeneous integration, advanced patterning, Quantum computing), EU is quite well positioned.

US and Asia have similar profiles, for both R&D and Manufacturing. It is interesting to note that they both look stronger in manufacturing than in R&D. This might be a consequence of the strong positioning of their key players in the field of Semiconductors (Foundries, Fabless) that are doing a lot of R&D internally.

3 Conclusion

The "Report on Future Technologies for Advanced Computation" underscores the vital importance of international cooperation in the semiconductor sector, particularly in the context of emerging technologies that promise to drive significant advancements in various applications of computational systems. The latter include artificial intelligence and machine learning in graphic processing units, augmented and virtual reality devices, advanced driver-assistance systems and their advancement for autonomous vehicles, edge AI for the Internet of Things. Each of these emerging applications requires specialized solutions tailored to its specific performance, power, and latency requirements.

The underpinning diversification of ‘advanced compute’ needs implies that one-size-fits-all devices and architectures based on CMOS technology will no longer suffice. This calls for an interconnected global landscape where fostering collaboration among nations, research institutions, and industry leaders is essential to accelerate technological innovation and strengthen the semiconductor value chains. This alignment is not only crucial for meeting the objectives of the EU Chips Act but also for ensuring that Europe maintains its competitive edge in the global market.

Besides the optimisation of Power-Performance-Area-Cost (PPAC) for CMOS and DRAM, key research and development has emerged to address the disparity between processor speed and memory performance (“memory wall”), the improved performance without a proportional increase in power consumption (“power wall”) and sustainable manufacturing.

Transistors will remain vital to general-purpose computing, despite the emergence of new paradigms. For their continuous scaling several architectures will continue to develop according to the various roadmaps. These include FinFETs, nanosheets, forksheets, CFET and FDSOI. To this end, advancements will also be required in key enabling technologies such as high-NA EUV for small feature (~nm) patterning and heterogeneous integration for increasing component density without expanding the chip’s physical size. Further advances will be driven by the introduction of new materials (e.g., transition metal dichalcogenides, carbon nanotubes, oxide semiconductors and ferroics) and devices (e.g., steep subthreshold swing transistors and non-volatile memories such as PCRAM, MRAM, OxRAM, FeRAM).

New computational paradigms will be essential to address the power and memory walls as well as the growing diversity of applications. In the near future, BEOL memories will enable brain-inspired architectures, e.g., near-memory, in-memory and neuromorphic computing. In the longer term, quantum computing accelerators may be used to overcome the limitations of traditional computing systems in a large family of optimisation problems. For such architectures and further performance improvements (e.g., in connectivity) heterogeneous integration technologies will also be a key enabler.

To summarise, key Advanced Computation‘ topics for EU to be active are:

- Classical’ Logic Scaling Roadmap beyond FinFET technology that extends devices structures through sub nm nodes (e.g., GAA and CFET architectures)
- Exploration of ‘Fully Depleted SOI’ technology for Power Efficient Analog and RF applications
- Exploration of alternative channel materials (e.g., 2D materials)
- Extension of the scaling of BEOL technologies, through the use of Ru, Airgap or Graphene-based metallization, by reducing the associated RC network
- Added BEOL functionality through the introduction of new materials such as 2D, oxide semiconductors and ferroics
- Exploration of the use of BEOL Non-Volatile Memories (using for example resistive RAM such as FeRAM, MRAM, PCRAM) for Power Efficient Neuromorphic-based architectures applied e.g., to embedded systems
- Photonic chips for optical interconnects and quantum information processing
- Demonstration of the capability of the ‘Buried Power Rail delivery’ to decongest the interconnection density that is becoming the most limiting factor for the scaling at 2nm and below
- Enablement of the High-NA EUV lithography for the patterning of 2nm nodes and beyond
- Usage of 3D integration to desegregate the classical large area chips into chiplets that will be much more power efficient when reconstruct using 3D integration design flow and associated toolbox.

The EU is very strong in R&D in all the domains of advanced computing and the findings of this report illuminate semiconductor technologies for advanced computing systems that may benefit from international collaboration. By leveraging shared knowledge and resources, stakeholders can address existing challenges, drive innovation, and create a sustainable future in the semiconductor industry.

References

[1] M. G. Bardon. “DTCO including Sustainability: Power-Performance-Area-Cost-Environmental score (PPACE) Analysis for Logic Technologies”, IEDM, 2020.

[2] U. Gupta et al., “Chasing Carbon: The Elusive Environmental Footprint of Computing”, 2021 IEEE International Symp. on High Performance Computer Arch. (HPCA), 2021, pp. 854-867.

[3] https://netzero.imec-int.com/

[4] N. Collaert, "Advancements in IC Technologies: A look toward the future," in IEEE Solid-State Circuits Magazine, vol. 15, no. 3, pp. 80-86, 2023.

[5] https://www.mit.edu/~pengw/research/csfet/

[6] “Unlocking system scaling bottlenecks with STCO”, interview Julien Ryckaert, https://www.imec-int.com/en/articles/unlocking-system-scaling-bottlenecks-system-technology-co-optimization

[7] E. Esmanhotto *et al.*, “Experimental Demonstration of Multilevel Resistive Random Access Memory Programming for up to Two Months Stable Neural Networks Inference Accuracy,” *Adv. Intell. Syst.*, vol. 4, no. 11, p. 2200145, Nov. 2022, doi: 10.1002/aisy.202200145.

[8] F. Jebali *et al.*, “Powering AI at the edge: A robust, memristor-based binarized neural network with near-memory computing and miniaturized solar cell,” *Nat. Commun.*, vol. 15, no. 1, p. 741, Jan. 2024, doi: 10.1038/s41467-024-44766-6.

[9] F. Moro *et al.*, “Neuromorphic object localization using resistive memories and ultrasonic transducers,” *Nat. Commun.*, vol. 13, no. 1, p. 3506, Jun. 2022, doi: 10.1038/s41467-022-31157-y.

[10] M. Payvand *et al.*, “Self-organization of an inhomogeneous memristive hardware for sequence learning,” *Nat. Commun.*, vol. 13, no. 1, p. 5793, Oct. 2022, doi: 10.1038/s41467-022-33476-6.

[11] S. D’Agostino *et al.*, “DenRAM: neuromorphic dendritic architecture with RRAM for efficient temporal processing with delays,” *Nat. Commun.*, vol. 15, no. 1, p. 3446, Apr. 2024, doi: 10.1038/s41467-024-47764-w.

[12] T. Dalgaty *et al.*, “Mosaic: in-memory computing and routing for small-world spike-based neuromorphic systems,” *Nat. Commun.*, vol. 15, no. 1, p. 142, Jan. 2024, doi: 10.1038/s41467-023-44365-x.

[13] D. Bonnet *et al.*, “Bringing uncertainty quantification to the extreme-edge with memristor-based Bayesian neural networks,” *Nat. Commun.*, vol. 14, no. 1, p. 7530, Nov. 2023, doi: 10.1038/s41467-023-43317-9.

[14] T. Dalgaty, N. Castellani, C. Turck, K.-E. Harabi, D. Querlioz, and E. Vianello, “In situ learning using intrinsic memristor variability via Markov chain Monte Carlo sampling,” *Nat. Electron.*, vol. 4, no. 2, pp. 151–161, Jan. 2021, doi: 10.1038/s41928-020-00523-3.

[15] K. Danouchi *et al.*, “Designing networks of resistively-coupled stochastic Magnetic Tunnel Junctions for energy-based optimum search,” in *2023 International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA: IEEE, Dec. 2023.

[16] Shuhan Liu *et al.*, “Edge Continual Training and Inference with RRAM-Gain Cell Memory Integrated on Si CMOS,” *to be presented at 2024 International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA: IEEE, Dec. 2024.

[17] M. Martemucci *et al.*, “Hybrid FeRAM/RRAM synapse circuit for on-chip inference and learning at the edge,” in *2023 International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA: IEEE, Dec. 2023.

[18] M. Martemucci *et al.*, “Unified Ferroelectric/Memristive Memory for Neural Network Inference and Training,” under review at *Nat. Electron*.

[19] E. Vianello and M. Payvand, “Scaling neuromorphic systems with 3D technologies,” *Nat. Electron.*, vol. 7, pp. 419–421, June 2024, doi: https://doi.org/10.1038/s41928-024-01188-y.

ICOS
International Cooperation
On Semiconductors

icos-semiconductors.eu

@ICOS_horizon

ICOS - International Cooperation
On Semiconductors

ICOS has received funding (2023 – 2025) from the European Union's Horizon Europe research and innovation programme under GA No 101092562.