Overcoming the data deluge challenges with greener electronics

J.-R. Lequepeys, M. Duranton, S. Bonnetier, S. Catrou, R. Fournel, T Ernst, L. Herault, D. Louis, A. Jerraya, A. Valentian, et al.

Overcoming the Data Deluge Challenges with Greener Electronics

J.-R. Léquepeys, M. Duranton*, S. Bonnetier, S. Catrou, R. Fournel, T Ernst, L. Hérault, D. Louis, A. Jerraya, A. Valentian**, F. Perruchot, T. Signamarcheix, E. Vianello, C. Reita
CEA-Leti, Univ. Grenoble Alpes, F-38000 Grenoble, France
*CEA-List, Université Paris-Saclay, F-91120, Palaiseau, France
**CEA-List, Univ. Grenoble Alpes, F-38000 Grenoble, France
jean-rene.lequepeys@cea.fr

ESSCIRC 2021 - IEEE 47th European Solid State Circuits Conference (ESSCIRC) | 978-1-6654-3751-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ESSCIRC53450.2021.9567836

***Abstract* - The complete ecosystem of electronic device manufacturers, from microelectronics, software and hardware designers to developers, producers, and integrators, is facing an immense new environmental challenge: to cope with the data deluge and to reduce the energy consumption of digital technologies. The purpose of this paper is to propose scientific and technical directions to reach global data and power sobriety while preserving computational efficiency. We present nine opportunities to lower the power consumption of computing units. A growth factor of 100 to 1000 in energy efficiency is achievable in the next 10 years if we take full advantage of all the potential improvements, working simultaneously at all five levels of the technological ecosystem (process steps, circuits, architecture, software and algorithms). We will indeed need to exploit all the possible technological advances to achieve this goal, including resistive memories, 3D stacking and new computing paradigms such as in-memory-computing, neuromorphic computing and quantum computing. Additionally, in order to maximize efficiency and performance, the research and development communities must work closer together and embrace a real culture of co-design to optimize applications and algorithms, algorithms and architectures, architectures and technologies jointly. We also propose to perform data processing operations as closely as possible to the source, in order to curtail the energy consumption that comes with data transport. Finally, we believe it is essential to accept the constraints for sustainable electronics now, and to change our mindsets quite radically by carrying out product life-cycle assessments in the very early phases of any new research.**

Keywords—3D, non-volatile memory, neuromorphic, FDSOI, GAA FET, quantum, co-design, architecture.

I. Introduction

The Fourth industrial Revolution, as described by the World Economic Forum, is relying on new digital technologies like Artificial Intelligence (AI), the Internet of Things and 5G/6G networks. Three interdependent trends have made this revolution possible. The first trend is the natural consequence of the exponential increase in processing and storage capabilities of electronic components at a reasonable cost and at constant overall power consumption. As a result, a growing part of the population is gaining access to smartphones and internet access. This democratization has led to the second trend, an exponential increase in the volume of data generated by humanity, stemming from both personal and economic activities. Increasing digital computing performance and the vast amounts of data generated, coupled with new classes of algorithms (Artificial Intelligence in general, Machine Learning in particular) to process the data, are transforming the way information can be used, and enabling the third trend: the development of new services, products and capabilities. While these technological advances offer great advantages, there is an urgent need to take into account and address the increase in worldwide energy and rare material consumption required to produce and operate digital technologies, and to orient worldwide R&D activities more and more towards sustainable electronics.

Confronted with this challenge, Europe has decided to respond with an ambitious action plan: on March 19, 2021, the 24 Member States plus Norway and Iceland, signed a declaration to accelerate the use of green digital technologies. The goal is to invest in and deploy green digital technologies to achieve climate neutrality and accelerate the environmental and digital transitions in priority sectors in Europe. Several actions will be taken at a national level including the development of energy efficient artificial intelligence solutions and low power hardware technologies, and the promotion of eco-designed products. Indeed, Europe is calling for a concerted effort to boost its capabilities in these key technologies, since they are enablers for other technological developments and provide a competitive edge to the European industry.

Computing in the era of the data deluge

According to IDC [1], humans and computers generated more than 64 zettabytes of data in 2020, and more than 2000 Zetabytes will be produced in 2035 (Fig 1). This escalation results from the fact that data is increasingly generated by machines (in 2018, only 44% of the data was produced by humans). It is expected that by 2022, 90% of all the data will be generated by machines (machine-to-machine communications) and by hundreds of billions of connected objects around the world. We are also witnessing a surge in multimedia data, such as image, video, speech and music, as opposed to "classic" digital data (text and numbers), and an increase in the resolution of videos and pictures (Fig 1).

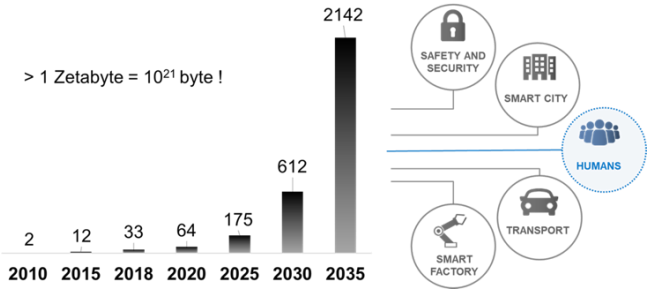

Fig 1: Data generation (actual and forecasted)[1]

To efficiently transform all of this raw data into useful information and services, computing will play an increasingly important role in the three stages of the data processing chain:

- Data transmission to processing units: Raw data is transported via wired networks (copper or optical) and

4G and 5G wireless networks to remote procesing units. Information can also be processed locally (at the edge) to decrease the amount of data transferred. According to IBS [2], in 2020 data traffic represented ~5% of the total generated data, the remaining 95% being either processed at the edge or lost;

- Data storage and processing: In 2025, it is expected that half of the data will be created and stored in enterprise servers or in the cloud and that more than 50% of the data will be real-time. This figure will probably reach close to 90% by 2050 (Source: IDC-Seagate data age) meaning that a large amount of data will be processed locally [3]. IDC also estimated that dark data [4] (data that are never used, unstructured or not indexed) would rise to 93% by 2020.
- Data analysis and exploitation: Massive usage of artificial intelligence algorithms will be needed to reduce the proportion of dark data.

Being able to trust information systems will be another important factor in their deployment. Systems need to be resistant to attacks and must protect corporate and personal data by complying with the new European GDPR regulations. Cryptography will undoubtedly be more widely deployed in the future and will also require additional computing capacity.

II. THE RISE IN ENERGY CONSUMPTION DUE TO THE SLOW DOWN IN SCALING BENEFITS

The power requirements of Information and Communication Technologies (ICTs) (Fig 2) have been relatively constrained so far, thanks to technological breakthroughs that have compensated the end of Dennard's scaling.

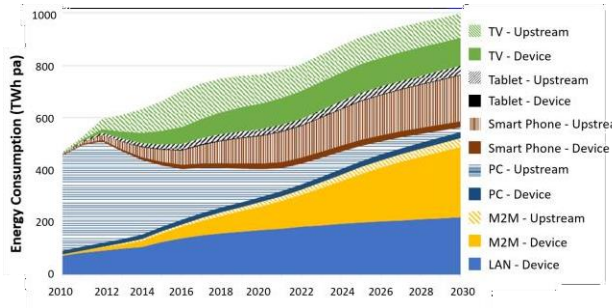

Fig 2: Energy forecast for ICT. (Source: IEA–4E TCP -EDNA– June 2019 –Total Energy Model for Connected Devices)

Until 2023, the power consumption of datacenters should remain stable at around 200 TWh but, with the slowing down of scaling benefits, we could see an exponential growth in their energy consumption as early as 2024, unless new technological innovations come into play (Fig 3) [5].

This exponential increase in energy consumption would not be sustainable. Part of the population is already reacting to it and the result down the road could be a rejection of some digital technologies, as it is already happening in some countries (e.g., France's situation with strong opposition to 5G systems). Conversely, a proper and well thought out deployment of these technologies would bring new useful services to society. It is therefore necessary for the electronics community at large (technology and electronic circuit architects, and circuit, software and electronic systems providers) to come up with breakthrough solutions to curtail the rise in the power consumption of digital systems. In this article, we analyze new digital technology opportunities and propose computing solutions that will facilitate the move towards sustainable and responsible electronics.

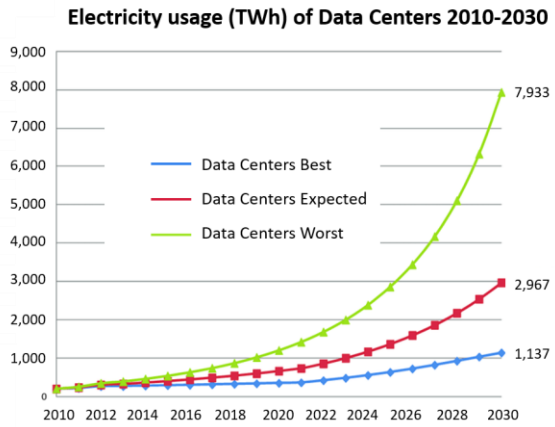

Fig 3: Data center electricity needs and the effects of ending Moore's law [5].

A. The limits of energy efficiency in computing systems

Both the energy efficiency and performance of computing systems have increased steadily in the past decades. In central processing units (CPUs), computations per kilowatt-hour have doubled every 1.5 years or so [6]. The performance of supercomputers went from 2 Gflops in 1985 (Cray2) to 415 Pflops in 2020 (Fugaku), i.e., a 10^8-fold increase in 33 years. Today we could replace the 1985 Cray2 computer with two Nvidia Jetson Nano compact computers and achieve the same computing performance while lowering the power consumption from 200kW to 20W. Although this increase in energy efficiency is quite impressive, it only amounts to a factor of 10^{-4}. The improvement in the number of computations performed per kWh is called "Koomey's law", and it states that "the number of computations per joule spent double every 18 months or so"[7]. In terms of energy performance, mechanical engines are currently only a few orders of magnitude away from a thermal engine's thermodynamical efficiency limit, whereas computing systems are quite far from the ultimate computational limit.

Figure 4 illustrates the performance evolution of supercomputers since the year 2000: a 10-fold improvement every four years in both performance (FLOPS) and energy consumed per operation (J/FLOP). The data suggests that further improvements in energy efficiency will soon be restrained by the physical limits of charge carriers in CMOS devices. In the future, significant power consumption reductions will rely less on improved fabrication processes (scaling, 3D integration and packaging) than on new computing architectures and paradigms, some of which are already being deployed (e.g., computing accelerators). ICT frugality will also be a key factor in reducing the world's digital energy footprint, and possibly the only way to avoid a Jevons paradox scenario. The Jevons paradox is that efficiency enables growth. The more efficient the technology, the more it gets used. New digital technologies that can do more with fewer resources allow the economy as

a whole to produce more, and the rate of consumption of that resource rises due to increased demand [8].

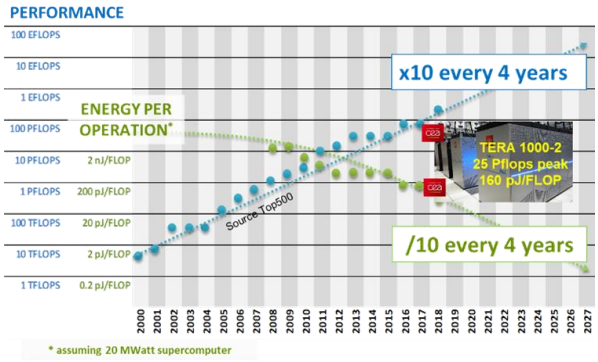

Fig 4: Evolution of the performance of supercomputers.

B. *Co-design as a way to improve computing efficiency*

Effective performance (and correlated energy efficiency) is a combination of multiple factors: technology, system architecture, software and the right choice of algorithms. In the early years of microelectronics, design and optimization in these three domains were not tackled simultaneously. The slow-down of Moore's law [9] now calls for a new paradigm centered on co-designing technological processes, architectures and algorithms.

After the year 2000 and the end of Dennard's scaling law [10], the complexification of integrated circuit structures limited the increase in processor frequency whereas the transistor density kept following Moore's law. To keep up with the need for higher performance, despite a much slower increase in frequency, multiple core architectures were introduced, giving birth to "multi"/"many" core architectures ("multi" = many identical cores, "many" = many specialized cores). This, in turn, required a new generation of high "quality" algorithms for parallelization. Today, the co-design of the whole semiconductor chain (process, architecture, algorithm and software) is the key factor to improve overall computing performance and efficiency. However, such improvements are currently limited by the fact that not all algorithms lend themselves well to parallelization, constraining overall efficiency (« Amdahl's law » [11]). Figure 5 below illustrates the evolution of processor performance over the past 30 years.

Novel technologies and co-design approaches can open up new ways to improve performance and curtail energy consumption. Many-core/multi-chip architectures, for example, take advantage of 3D fabrication and assembly techniques. Passive or active silicon interposers act as a substrate on which the multiple optimized chips/cores are assembled in a complex but optimized architecture. This approach requires both Wafer-to-Wafer (W2W) and Die-to-Wafer (D2W) bonding having the right alignment accuracy, specific surface preparation, and dedicated material and process developments. An example of one of CEA-Leti's many-core architectures is the "INTACT" demonstrator shown in Figure 6, which includes 96 cores and 6 chiplets [12][13].

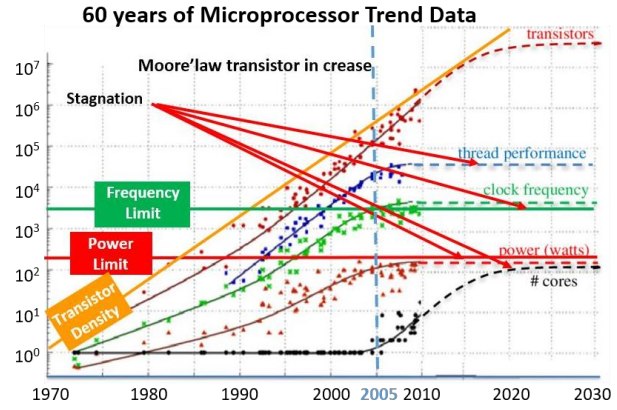

Fig 5: Evolution of processor performance adapted from [14].

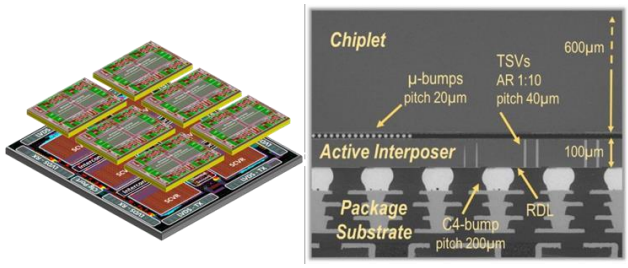

Fig 6: CEA-Leti's INTACT demonstrator, from concept to 3D cross-section.

***First research track*: develop architectures and demonstrators based on active silicon interposers, chiplets and a 3D toolbox. Co-designing semiconductor processes and novel 3D electronic architectures can help improve computing performance significantly.**

C. *A strong need for more specialized accelerators*

The move towards parallelization has promoted the emergence of dedicated accelerators. Recently, energy and performance constraints led to the introduction of new architectures dedicated to specific purposes like, for example, Graphics Processor Units (GPUs) for image processing, as illustrated in Figure 7, and Neural Network accelerators (NPUs) for learning and inference in machine learning applications. Dedicated full computing architectures based on specific hardware accelerators can be up to two orders of magnitude more efficient for the same calculation than general-purpose programmable solutions (Fig 7). Ideally, a dedicated hardware solution for a given problem would be the most energy efficient solution, but economic considerations still impose a certain degree of versatility in integrated circuits so that they may be used in several applications. A good tradeoff between cost and adaptability has been obtained with the emergence of solutions for a relevant class of problems like neuromorphic reconfigurable architectures, minimization of functions, etc.

***Second research track*: develop new electronic architectures using dedicated but versatile accelerators that exhibit optimized power consumption in a wide range of applications.**

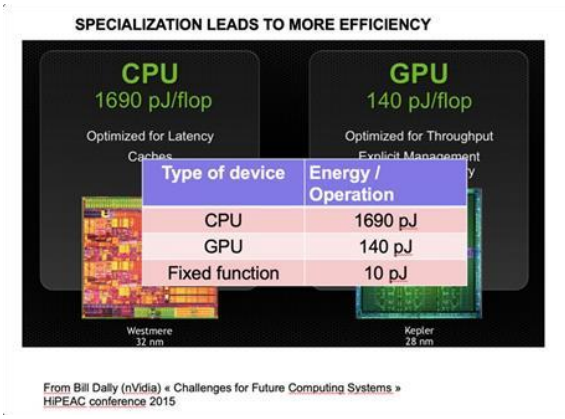

Fig 7: Specialization in computing strongly increases energy efficiency.

III. THE POWER OF BETTER ALGORITHMS

Algorithms can have a very strong impact on performance: there can be a difference of many orders of magnitude between a "naive" algorithm and an optimized algorithm. A good illustration is Ewin Tang's algorithm. This 18-year-old teenager proved that ordinary computers can solve an important computational problem (the "recommendations" problem, crucial for the success of services like Netflix) with performances potentially comparable to those of a quantum computer. In 2016, computer scientists Iordanis Kerenidis and Anupam Prakash published a quantum algorithm that solved the recommendations problem exponentially faster than any known classical algorithm, but they did not prove whether a fast classical algorithm could do the same [15]. Like Kerenidis and Prakash's quantum algorithm, Tang's classical algorithm, published in 2018 [16], works in polylogarithmic time - meaning that the computation time is proportional to the logarithm of features such as the number of users and products in the dataset - and is also exponentially faster than any previously known classical algorithm. Thus, we can see that finding the right algorithm for a given application allows phenomenal speed-up factors that can increase exponentially with the size of the problem.

***Third research track*: investigate new algorithms and innovative approaches that could lead to implementations that are far more efficient in solving application problems.**

IV. THE INCREASING DEMAND FOR ORES AND RARE EARTHS

The data deluge will have a big impact on the demand for data storage technologies and, as a result, on the demand for the materials required to make them. Recent estimates made by researchers from the Catholic University of Leuven indicate that the storage of 10% of the expected 2025 global datasphere will require up to 8 kilotons of neodymium, which is close to 12 times the current yearly European demand for this material. The development of new generation memory technologies will have to take into account such critical raw material challenges.

Electronic products are also having a negative ecological impact in other ways. For example, the exploitation of highly polluting mines that consume a lot of water and the recycling and burial of obsolete electronic products, often done in third world countries, are contributing to destroy ecosystems and lower living conditions of local populations. If we want to make a systematic change in the way we design and fabricate electronic components, we must carry out life cycle assessments of future products during the earliest stages of their design and development processes. We must also be ready to find alternatives to the rare, polluting and difficult to recycle materials that are presently used in those products. This huge endeavor will last for well over a decade.

***Fourth research track*: develop sustainable electronics by accepting and deploying the huge R&D efforts it will require and by promoting the new mindset that must accompany this change.**

V. THE STABILITY OF THE SUPPLY CHAIN - GEOPOLITICAL CONSIDERATIONS

The semiconductor fabrication process requires a wide range of materials (Fig 8). Only 15% of these materials comes from recycling. Figure 9 illustrates the type and quantity of materials used for different kinds of microelectronic memories. Currently, many of those materials are extracted in only a few countries: Australia produces 50% of the world's lithium, Chile 25% of the world's lithium and 25% of the world's copper, South Africa 70% of the world's platinum, and China supplies 95 % of most of the rare earths.

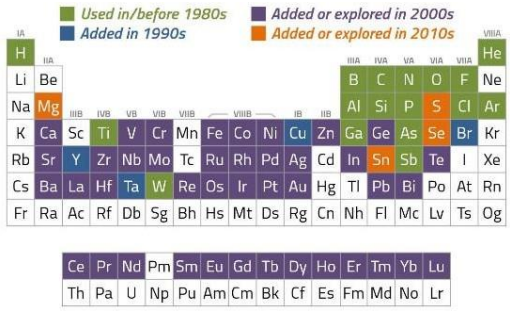

Fig 8: Increase in materials used for semiconductor fabrication. (Courtesy of Lam Research)

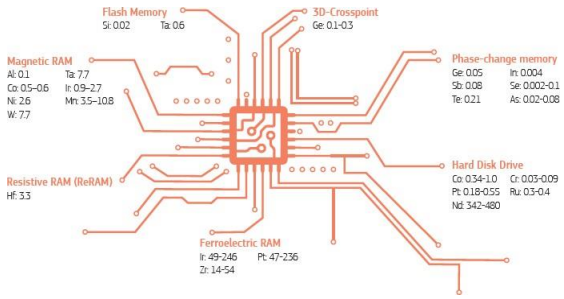

Fig 9: Estimated material intensity factors for different memory technologies. Amounts are in tons per Zettabyte. (Source: Critical Raw Materials for Strategic Technologies and Sectors in the European Union - September 2020)

Consequently, there is a high potential for supply chain disruption and an added risk associated with the world's evolving political and economic strategies. Nevertheless, there are new and important deposits of rare earth elements in other parts of the world. In Greenland and Japan, for

example, significant deposits were recently discovered in far eastern territorial waters. At any rate, technologies that consume less rare materials are urgently needed. Ultra-thin film wafer bonding, 2D material growth, die to wafer bonding and local epitaxial growth developed at CEA-Leti are powerful levers to limit critical material usage in microelectronics. 3D technologies and advanced packaging can also help reduce the need for noble metals.

VI. The need to process data closer to the source

We can see in Figure 10 that there is a strong tendency to process more and more data at the source (or edge) or very close to it. While only 20% of all the data generated worldwide was processed locally in 2015, it is projected that 80% will be processed locally in 2030 [17]. However, the data deluge will impose more computing systems at both ends, cloud and edge. The main reason for the redistribution of data processing between cloud and edge is the need for more real-time, autonomous, local and private/secure applications in production processes and in services in general. Edge computing can provide solutions to those needs and, at the same time, reduce the overall energy required to process the data. The major benefits of computing at the edge are ensuring:

- Data protection of data: Maintaining personal or corporate data locally is the best way to ensure their integrity and keep control at the owner's level. This is particularly critical for healthcare applications and industrial proprietary data. Moreover, the protection of personal data is a legal obligation under the European GDPR, which lays out specific requirements for businesses and organizations established in Europe or serving European clients/users, and regulates how businesses can collect, use, and store personal data;
- Very low latency: The transmission time between a sensor, or any data generating system, and a data processing server is limited by the speed of data transmission (which can be, at best, the speed of light). This translates into a 10ms round-trip latency for a server located 1500km away, not taking into account the time required for processing. Such latency is unacceptable for applications such as autonomous vehicles, remote robot control, and production control in a factory;
- Operational reliability: Many critical applications, such as self-driving cars, cannot depend on the quality and availability of today's communication networks or channels. A car's behavior in a critical situation should not have to depend on the status of a 4G/5G connection or of a line-of-sight photonics connection;
- Optimal use of transmission channels and storage: More than 90% of the data sent to the cloud is used at best once, but still requires energy for transmission and storage. By pre-processing the raw data at the source, it would be possible to reduce overall data bandwidth and storage requirements;
- Autonomy of decision. The use of local data and local processing would allow a much higher degree of customization of the response, e.g., medical devices that provide continuous personalized treatment adapted to the history of the patient and not depending only on generic cloud-based applications.

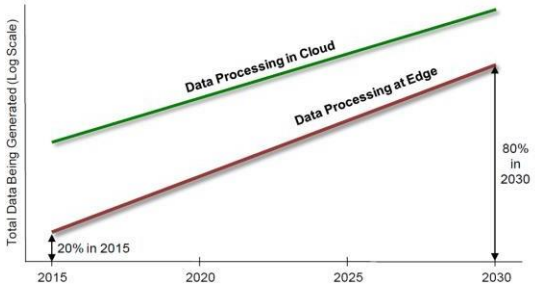

Fig 10: Evolution of data processing at the Edge. (Source: IBS 2020-09)

Fifth research track: **develop edge-computing solutions to reduce the overall energy requirements and improve service quality.**

VII. The evolution of AI

Most traditional Artificial Intelligence methods are based on convolutional neural networks (CNN) [18], although other types of networks are emerging. Convolutional neural networks are creating an important computational burden. In fact, the ever-increasing storage capacity and processing resources required by CNN are fueling the demand for higher performance computing capabilities. The complexity of these networks results in the need for very large data sets, especially during the training phase, and regular access of stored weights during the inference phase. All these operations contribute to increase the required energy per operation. Indeed, transferring and storing one gigabyte of data through the internet uses between 3.1 kWh and 7 kWh, instead of 0.000005 kWh when done locally [19]. And, as mentioned earlier, latency is also affected: an operation at the edge lasts µs whereas cloud operations take closer to 10 seconds.

The edge may be viewed as a hierarchy that includes essentially three segments (Fig 11), each segment requiring four kinds of AI-specific ASICs having different power consumptions and different latency and response times. The three segments are:

- Edge AI - which is on-premises (as opposed to cloud) computing and may be embedded. Typical examples are set-top boxes, smart speakers and autonomous vehicles. Edge AI operates with a power budget in the range of 10W and a latency in the range of 10 milliseconds (ms);
- Portable device edge (embedded) AI – which is applicable to medical devices, smartphones and other mobile devices and operates with a power budget in the range of 1W to 10W and a latency close to one ms;
- Deep Edge AI (or AI for the Internet-of-Things) – used for intelligent sensors and the IoT. Deep Edge AI must respect strong cost and power constraints - in the range of several hundreds of mW to 1 µW- and a latency of less than one second.

A fourth segment, the so-called Near Edge segment, also exists. It interfaces with the cloud and can function as a mini datacenter. It is composed of systems operating with an intermediate latency and power consumption ranging

anywhere from 1 to hundreds of kW, depending on the required performance. Examples are shopfloor servers for automated production and local datacenters connected to 5G base stations.

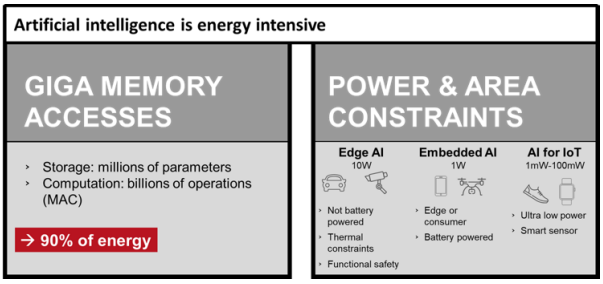

Fig 11: Artificial Intelligence is energy intensive.

In order to optimize power consumption, dedicated ASIC solutions have become a necessity, and the market for AI accelerators is expected to grow, as shown below (Fig 12).

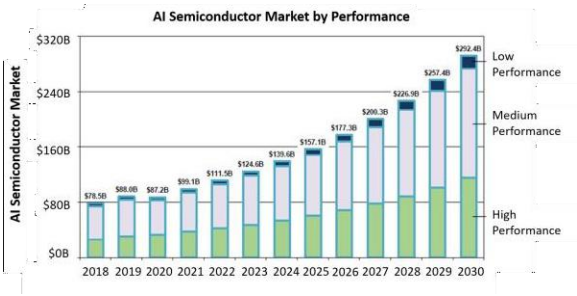

Fig 12: AI accelerator market perspectives. (Source: IBS 2020)

We also need to limit the amount of data that moves from a memory unit to a computing unit, since reading data from memory and storing it back into the data bank represents ~90% of the power consumed by a chip (Fig 13). To evolve beyond computing architectures based on the Von Neumann model, we are moving computation closer to the memory (Near Memory Processing) and even into the memory (In Memory Computing). These concepts can be implemented with SRAM memories or non-volatile memories and the calculations can be performed in analog or digital form, the gains varying from 30 to 1000, depending on the implementation scenario.

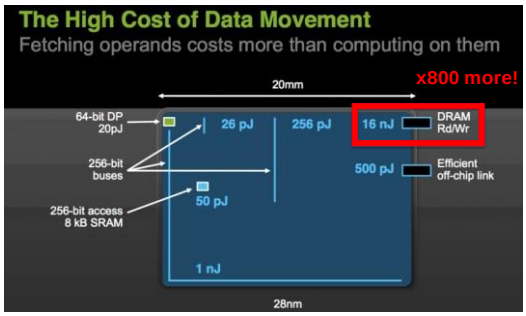

Fig 13: Data movement is costly in terms of power consumption. (Source: Bill Dally, "To ExaScale and Beyond", 2010)

***Sixth research track*: implement dedicated ASIC + AI solutions to move less data and lower power consumption, and develop in-memory computing and near memory processing solutions with the adequate software and EDA toolbox.**

VIII. Bio-inspired Artificial Intelligence solutions

Bio-inspired neuro-computing appears to be a very promising approach to lower energy consumption (Fig 14). The human brain is well suited for many complex tasks, such as image recognition, which it performs with very high energy efficiency. However, it is not able to make relatively simple arithmetic operations like multiplying large numbers. A bee's brain has very low computing power but it deploys clever survival strategies that consume very little energy. Several neuromorphic research chips exist today, like the SpiNNaker (Human Brain Project), IBM TrueNorth, Neurogrid (Stanford) and Intel's Loihi family chips, and startups are on their way to commercializing chips based on Spiking Neural Networks (Brainchip, Innatera, Synsense, GrAIMatterLabs or WestwellLabs).

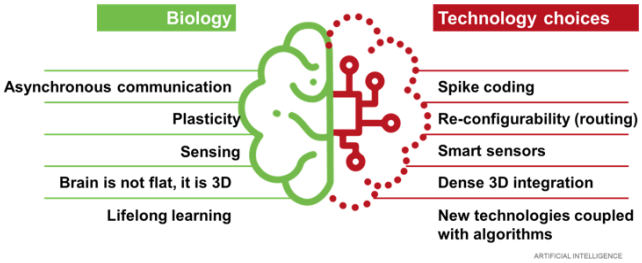

Fig 14: Bio inspiration to fill the gap with natural intelligence.

At CEA-Leti, we are developing neuromorphic solutions with neurons and synapses, integrating very compact and low cost resistive RAM (OxRAM) in the back-end of line (Fig 15).

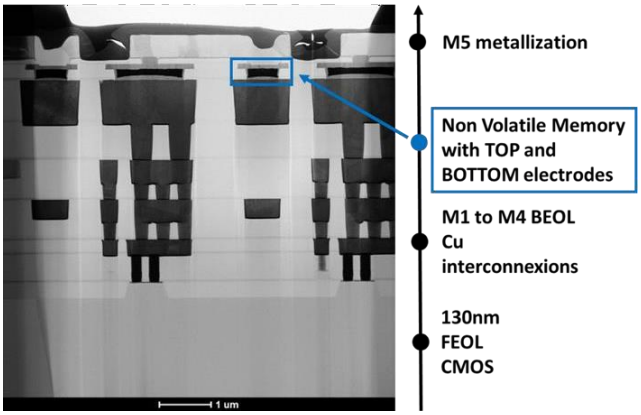

Fig 15: Non-volatile memory in the back-end of line.

The first generation of this type of solution, the SPIRIT neuromorphic circuit (Fig 16), was designed to recognize hand-written characters with an energy efficiency of 3.6pJ per synaptic event [20]. The second version of the circuit, manufactured using 28nm FDSOI technology, is a real scale-up with 130,000 neurons and 75 million synapses using OxRAM non-volatile memories. Its power consumption is expected to be less than 1pJ per synaptic event and should facilitate the processing of LiDAR signals.

Replacing SRAM, Embedded Flash and stand-alone memories with non-volatile resistive memories comes with many benefits, including an increase in memory density, lower power consumption, improved latency and the possibility of deploying in-memory-computing (IMC) building blocks, including neuromorphic architectures. By assembling resistive memories in a 'crossbar' arrangement, IMC can be naturally implemented by simply relying on

Kirchhoff's current law. Additionally, the ability of resistive memories to change their resistive state makes them promising candidates to emulate synaptic plasticity and enable on-chip learning. However, there is no guarantee that resistive memories will function in state-of-the-art neural network topologies due to the memories' multiple non-ideal properties, such as device variability. Only through strong interaction between hardware and software developments can we expect to overcome this difficulty. For example, it has been demonstrated that non-ideal traits of resistive memories can provide compact ways to implement stochastic synapses in Bayesian Neural Networks. Intrinsic device variability is exploited to implement Markov Chain Monte Carlo (MCMC) sampling, thus enabling low power in situ learning (few µJoules) [21].

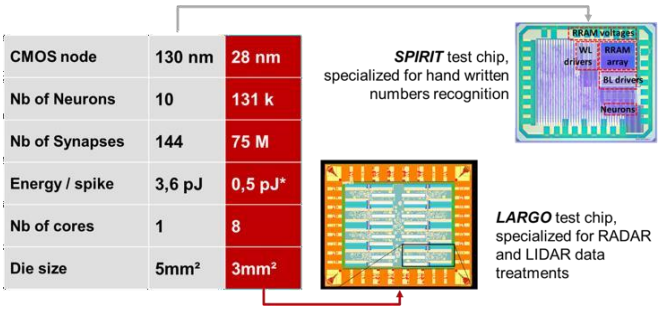

Fig 16: SPIRIT chip with embedded OxRAM (IEDM 2019) and LARGO chip.

***Seventh research track*: use non-volatile memories to facilitate the implementation of neuromorphic chips that can deploy on-chip learning algorithms and a smart in-memory computing approach.**

IX. More Moore Technologies Continue to Bring Competitive Advantages

Scaling linked to Moore's law continues. The 5nm FinFET technology has been commercialized and Gate All Around (GAA) nanowire and nanosheet FETs are currently being developed for nodes 5/3/2 (Fig 17) [22].

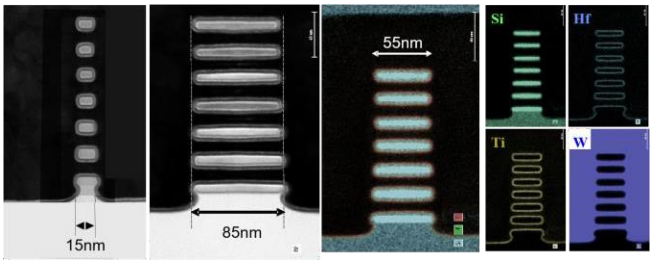

Fig 17: Stacked nanowire and nanosheet technology developed at CEA-Leti.

FDSOI is still the best-in-class solution for ultra-low power thanks to its dynamically controllable threshold voltage which can adapt static and dynamic dissipation depending on the instant task requirement. The technology can be scaled down to at least a 10 nm channel length, by thinning down the Tbox to improve electrostatic control [23][24]. Nevertheless, some boosters are required (Fig 18) in order to obtain, at the same time, a sufficiently big drive current. All the individual boosters (Fig 18) have been validated. They are deployable and have been included in industrial roadmaps.

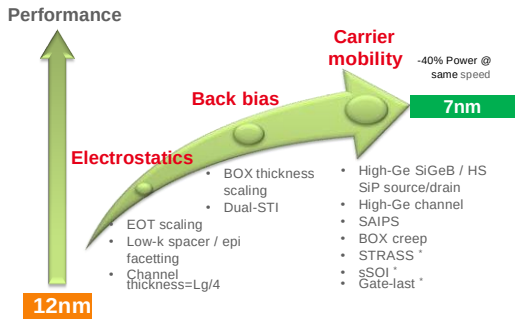

Fig 18: FDSOI boosters and roadmap towards a 7nm node.

At the 28nm node and below, conventional embedded Flash memories are facing strong challenges in terms of cost, speed and reliability. For advanced technology nodes, Back End Of Line resistive memories are more appealing. They are currently implemented as a 1T1R structure, i.e., with one MOS transistor (1T) used for accessing one resistor (1R) [25]. The access transistor limits the memory cell footprint. Promising results have recently demonstrated an increase in the memory density by stacking multiple 1T1R thanks to monolithic 3D technology [26], or replacing the MOS transistor by a stacked nanowire transistor [27], or by a backend selector [28].

***Eighth research track:* explore silicon-based devices and 3D structures to pursue equivalent scaling, down to the ultimate physical limits.**

X. Quantum Computing Offers New Opportunities

The promise of Quantum Computing, near instant computation made possible by replacing the deterministic and serial nature of classical computing by a probabilistic and simultaneous operating mode, will provide access to unchartered territories. This new computing power, based on superposition, entanglement and interference of qubits, should be well suited to solve complex algorithms such as those required in transport and logistics for traffic optimization and in healthcare, via molecular simulations, for new drug discoveries. Quantum computing will also impact strategic domains like energy, materials, finance and defense. These potentialities come with the need for broad research actions in all the domains we mentioned previously (devices, low temperature circuits [29][30], architectures, software and algorithms) because the quantum computing paradigm is intrinsically different from the digital one, and has its specific challenges.

Among these challenges, there is the choice of the technological option for the physical system that will implement the qubits required for quantum computing, and it is too early in the process to declare a "winner". Indeed, different approaches might be needed for different applications, but there is good reason to think that the silicon-based approach, in which semiconductor devices are used to create arrays of electrostatic-potential wells to isolate spin elements, is a leading candidate to meet critical criteria [31][32]. While the quality of the qubits is still an obstacle, the controllability, repeatability, scalability and manufacturability of very large arrays by well-proven processes and materials coming from the semiconductor industry is unparalleled by any other system.

Many of the basic challenges of quantum computing are intertwined. The need for cryogenic operating temperatures, for example, affects all approaches to quantum technology. All approaches will also require electronics for qubit control, read-out, and interfacing with classical computing systems. This highlights the broad importance of exploring and developing low-temperature CMOS technologies, especially advanced technologies such as FD-SOI, which has the unique property of being able to dynamically control the threshold voltage (Vth) of devices at different temperatures. Finally, the overall system design for this type of novel low-temperature system will require careful and clever partitioning of different functions and different elements (such as analog and digital hardware) at different temperature stages. For this reason, CEA-Leti's quantum computing development teams include experienced analog and RF designers working closely with hardware architects and low-level-software engineers to develop a clear pathway to a fully operational integrated system.

Ninth research track: **explore a silicon-based approach for quantum computing including low-temperature CMOS technology for qubit control and readout, packaging based on a Silicon interposer that can host qubits and electronics, and suitable error-correction codes.**

XI. Conclusion

There are many opportunities to lower the power consumption of computing units. By working simultaneously at five levels of the technology (process steps, circuit, architecture, software and algorithms), we should be able to improve power efficiency by a factor of 100 to 1000 in the next 10 years. However, to achieve this goal, we must exploit all the technological breakthroughs such as resistive memories, 3D integration, new computing paradigms (in-memory-computing, neuromorphic and quantum), process data as close as possible to the data source, and adopt a co-design approach throughout the whole microelectronics community. In parallel and immediately, we must take into account the constraints for sustainable electronics and change our mindsets quite radically by carrying out product life-cycle assessments in the earliest stages of all new technological developments.

XII. References

[1] IDC's "Data Age 2025" whitepaper, 2018, see also "the Digitalization of the wsorld-data Age 2025" (video) Dave Reinsel, hosted by Seagate Technology

[2] "DoT (Data of Things) applications opportunities", IBS Newsletter, Sept 2020

[3] European's Commission 2020 Strategic foresight report

[4] Y. Liu et al., Deep Hash-based Relevance-aware Data Quality Assessment for Image Dark Data", ACM/IMS Trans. Data Sci., Vol. 2, no. 2, article 11, 2021

[5] M. Koot, F. Wijnhoven, "Usage impact on data center electricity needs: A system dynamic forecasting model", Applied Energy 291, 2021

[6] https://energy.mit.edu/news/energy-efficient-computing/

[7] J. G. Koomey, "Outperforming Moore's Law," IEEE Spectrum, vol. 47, no. 3, pp. 68–68, Mar. 2010

[8] R York, J. Mc Gee, "Understanding the Jevons paradox", Environmental Sociology, vol.2, 2016

[9] G. E. Moore, "Cramming more components onto integrated circuits," Electronics, vol. 38, no. 8, 196

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

[11] M. D. Hill; M. R. Marty "Amdahl's Law in the Multicore Era", IEEE Computer, vol.41, pp. 33-38, 2008

[12] P. Vivet et al., "2.3 A 220GOPS 96-Core Processor with 6 Chiplets 3D-Stacked on an Active Interposer Offering 0.6ns/mm Latency, 3Tb/s/mm2 Inter-Chiplet Interconnects and 156mW/mm2@ 82%-Peak-Efficiency DC-DC Converters," 2020 IEEE International Solid-State Circuits Conference, pp. 46-48,2020

[13] P. Vivet *et al*., "IntAct: A 96-Core Processor With Six Chiplets 3D-Stacked on an Active Interposer With Distributed Interconnects and Integrated Power Management", IEEE Journal of Solid-State Circuits, vol. 56, no. 1, pp. 79-97, Jan. 2021

[14] Original figure is available at: https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/

[15] I. Kerenidis and A. Prakash, "Quantum gradient descent for linear systems and least squares," Phys. Rev. A 101, 022316, Feb. 2020

[16] E. Tang, "A quantum-inspired classical algorithm for recommendation systems,"Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, pp. 217-228, 2019

[17] N. Wirth, "A plea for lean software," Computer, vol. 28, no. 2, pp. 64–68, Feb. 1995

[18] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time-series. In M. A. Arbib, editor, The Handbook of Brain Theory and Neural Networks. MIT Press, 1995

[19] V. Zhirnov, R. Cavin and L. Gammaitoni, "Minimum Energy of Computing, Fundamental Considerations", ICT - Energy - Concepts Towards Zero - Power Information and Communication Technology, 2014

[20] A. Valentian, F. Rummens, E. Vianello, T. Mesquida, C. Lecat-Mathieu de Boissac, O. Bichler, C. Reita, 'Fully Integrated Spiking Neural Network with Analog Neurons and RRAM Synapses', IEEE International Electron Devices Meeting, 2019

[21] T. Dalgaty, N. Castellani, C. Turck, K.-E. Harabi, D. Querlioz, E. Vianello, 'In situ learning using intrinsic memristor variability via Markov chain Monte Carlo sampling', Nature Electronics, vol. 4, pp. 151–161, 2021

[22] S. Barraud *et al.*, '7-Levels-Stacked Nanosheet GAA Transistors for High Performance Computing', IEEE Symposium on VLSI Technology, 2020

[23] V. Barral et al, Strained FDSOI CMOS technology scalability down to 2.5nm film thickness and 18nm gate length with a TiN/HfO2 gate stack, IEEE Internationnal Electron Device Meeting, 2007

[24] F. Andrieu et al, Design Technology Co-Optimization in advanced FDSOI CMOS around the Minimum Energy Point: body biasing and within-cell VT-mixing, IEEE Symposium on VLSI Technology, 2018

[25] L. Grenouillet, '16kbit 1T1R OxRAM arrays embedded in 28nm FDSOI technology demonstrating low BER, high endurance, and compatibility with core logic transistors', IEEE International Memory Workshop 2021

[26] E. Esmanhotto, et al., High-Density 3D Monolithically Integrated Multiple 1T1R Multi-Level-Cell for Neural Networks, IEEE International Electron Devices Meeting, 2020

[27] S. Barraud, et al. 3D RRAMs with Gate-All-Around Stacked Nanosheet Transistors for In-Memory-Computing, IEEE International Electron Devices Meeting, 2020

[28] D. Alfaro Robayo et al., Integration of OTS based back-end selector with HfO2 OxRAM for crossbar arrays, IEEE International Electron Devices Meeting, 2019

[29] L. Le Guevel et al, A 110mK 295µW 28nm FDSOI CMOS Quantum Integrated Circuit with a 2.8GHz Excitation and nA Current Sensing of an On-Chip Double Quantum Dot, IEEE Internationnal Solid-State Circuits Conference, 2020

[30] B. Cardoso Paz et al, Variability Evaluation of 28nm FD-SOI Technology at Cryogenic Temperatures down to 100mK for Quantum Computing, IEEE Symposium on VLSI Technology, 2020

[31] T. Meunier et al, Towards scalable quantum computing based on silicon spin, IEEE Symposium on VLSI Technology, 2019

[32] R. Maurand et al, A CMOS silicon spin qubit, Nature Communications volume 7, 13575, 2016