H2020-FETHPC-3-2017 - Exascale HPC ecosystem development

EXDCI-2

European eXtreme Data and Computing Initiative - 2

Grant Agreement Number: 800957

D2.2
Report on trends and potential synergies between electronics, photonics and HPC

Final

Version: 1.0
Author(s): JF Lavignon, TS-JFL; Marc Duranton, CEA
Date: 18/02/2020

Project and Deliverable Information Sheet

EXDCI Project	**Project Ref. №: FETHPC-800957**	
	Project Title: European eXtreme Data and Computing Initiative - 2	
	Project Web Site: http://www.exdci.eu	
	Deliverable ID: **D2.2**	
	Deliverable Nature: Report	
	Dissemination Level: PU *	**Contractual Date of Delivery:** 29 / 02 / 2020
		Actual Date of Delivery: 18 / 02 / 2020
	EC Project Officer: Evangelia Markidou	

* - The dissemination level are indicated as follows: **PU** – Public, **CO** – Confidential, only for members of the consortium (including the Commission Services) **CL** – Classified, as referred to in Commission Decision 2991/844/EC.

Document Control Sheet

Document	**Title: Report on trends and potential synergies between electronics, photonics and HPC**	
	ID: D2.2	
	Version: <1.0 >	**Status:** ***Final***
	Available at: http://www.exdci.eu	
	Software Tool: Microsoft Word 2013	
	File(s): EXDCI-2- D2.2.V1.0.docx	
Authorship	**Written by:**	JF Lavignon, TS-JFL Marc Duranton, CEA
	Contributors:	
	Reviewed by:	Bernd Mohr, FZJ John Clifford, PRACE
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	5/02/2020	Draft	For internal review
0.2	12/02/2020	Draft	Reviewer version
1.0	18/02/2020	Final version	Integration of reviewers' comments

Document Keywords

Keywords:	PRACE, , Research Infrastructure, post-exascale, photonics, electronics

Table of Contents

List of Figures

References and Applicable Documents

List all external web sites

[1] http://www.exdci.eu
[2] http://www.prace-project.eu
[3] http://www.etp4hpc.eu

List of Acronyms and Abbreviations

Below is a list of acronyms used within the EXDCI-2 project. The acronym specific to this report are explained inside the text where they are used.

AISBL	Association Internationale Sans But Lucratif (International Non-for-Profit Association)
BDEC	Big Data and Extreme-scale Computing
BDV	Big Data Value
CoE	Centres of Excellence for Computing Applications
cPPP	contractual Public-Private Partnership
CSA	Coordination and Support Action
D	Deliverable
DG	Directorate General
DoW	Description of Work
EC	European Commission
ECMWF	European Centre for Medium-range Weather Forecasts
EESI	European Exascale Software Initiative
ENES	European Network for Earth System modelling
EPOS	European Plate Observing System
EsD	Extreme scale Demonstrators
EU	European Union
FET	Future and Emerging Technologies
FP7	Framework Programme 7
GDP	Growth Domestic Product
H2020	Horizon 2020 – The EC Research and Innovation Programme in Europe
HPC	High Performance Computing
IDC	International Data Corporation
IESP	International Exascale Software Project
INVG	Istituto Nazionale di Geofisica e Vulcanologia (National Institute of Geophysics and Volcanology)
ISV	Independent Software Vendor
IT	Information Technology
KPI	Key-Performance Indicator
M	Month

OS	Operating System
PM	Person Month
Q	Quarter
R&D	Research and Development
R&I	Research and Innovation
RFP	Request for Proposal
ROI	Return On Investment
SHAPE	SME HPC Adoption Programme in Europe
SHS	Social and Historical Sciences
SME	Small and Medium Enterprise
SRA	Strategic Research Agenda
SWOT	Strengths, Weaknesses, Opportunities and Trends
TRL	Technology Readiness Level
US	United States
WG	Working Group
WP	Work Package

Executive Summary

The current technologies used for HPC systems will not be able to sustain the performance increase requested by the HPC/HPDA/AI[1] application communities. Since the end of the Dennard scaling and with the approaching end of the Moore's law, the standard CMOS[2] technology has to be complemented by other approaches if we want to continue to deliver more performance.

The three main HPC ecosystems beside Europe, namely US, China and Japan have undertaken significant research initiatives to work on these new approaches. The investigations include topics such as new materials that are more efficient than CMOS, new architecture, photonics, analog system or quantum technologies. The level of investment is high in these three countries but we do not believe that they have already a competitive edge over Europe.

Europe has its own research programmes for new upstream technologies mainly through ECSEL and Horizon2020 calls. The resulting projects are not always linked with the design of future HPC systems.

To foster this interaction, EXDCI-2 organized a workshop gathering European experts from photonics, electronics and HPC. It has confirmed the existence, in Europe, of research ideas with high potential for HPC (and for high performance edge). The main conclusion is that research projects involving upstream technology providers and HPC teams could deliver potential new solutions for HPC systems. Some science fiction success stories have been written by workshop participants that illustrate the benefits for science, industry and society of such an investment.

The new approaches are a combination of the three (or at least of two of them) degrees of freedom that can be played with to deliver more performance in an energy efficient way:

- Switching from computing centric execution used by processors and GPU (akin to Von Neumann architecture) to a data centric paradigm to reduce the overhead introduced by the data movement;
- Changing what is called an operation by playing with operand precision or introducing of multi-bits or analog coding or other ways of encoding information (e.g. Quantum),
- Introducing new materials that will deliver more efficient ways (in terms of timing and/or energy) to store, switch and/or process information.

This gives a very broad set of options but only a few will emerge due to economic constraints, critical mass issues, industrialization aspects, legacy and usability problems. The most promising options are presented in this report.

To conclude, we propose four recommendations to make Europe a significant potential player in the field of new technologies for future HPC systems:

- R1 Establish a continuous dialogue between photonics, electronics and HPC communities under the supervision of Photonics21, AENEAS and ETP4HPC.
- R2 Undertake small actions to specify research objectives, benchmarks and test data sets at the interface of two research communities.
- R3 Work on European specifications for the integration of heterogeneous chips.

[1] High Performance Computing/ High Performance Data Analytics/ Artificial Intelligence
[2] Complementary Metal Oxide Semi-conductor

- R4 Launch a research programme to develop new ideas coming from upstream technologies to provide new solutions for upcoming HPC systems.

1 Introduction

The objective of Task 2.2 of the EXDCI-2 project is to develop synergies between the HPC technology ecosystem and other European research activities especially in the field of electronics and photonics. HPC is an important driver for pushing the technology limits of electronics and photonics and can benefit from new technologies developed in these fields. This report presents the findings of this EXDCI-2 activity.

Chapter 2 discusses the limitations that the current technologies used by HPC systems face. They are the reasons why new research paths have to be studied if we want to sustain the performance increase. Most of these research potential options are developed in the photonics and electronics ecosystems.

In Chapter 3 we present the situation in the three main HPC ecosystem beside Europe: US, China and Japan. The different initiatives aiming to develop the technologies that will be needed for future HPC systems are explained. This analysis is made with the objective of gathering information on the paths taken by these ecosystems to prepare the post exascale era.

Chapter 4 is dedicated to a survey of the situation of Europe. After a brief description on how the photonics and electronics ecosystem are organized, some of the Horizon2020 programmes that can contribute to developing technologies for the post exascale time frame are analysed. Again the objective is to see what the potential relevant paths for future HPC systems are.

Among the activities of this task, a two days' workshop was organized in November 2019 to gather experts from electronics, photonics and HPC. Chapter 5 presents this workshop which aimed to assess potential new technologies for HPC systems but also to discuss how to develop a more integrated and efficient technology value chain in Europe. The main findings and recommendations issued by this interaction are also presented in this chapter.

Throughout this task, information about the potential technologies for future HPC systems has been gathered. Chapter 6 present a synthesis of this information and the challenges that are attached to their development. Despites the efforts of the authors, it has been difficult to be very precise about the expected maturity date of the technologies as, most of the time, research roadblocks are still to be removed.

The last chapter presents some recommendations for Europe to strengthen its position in technologies relevant for future HPC system. These recommendations could be implemented through EuroHPC actions and/or initiatives of the future Horizon Europe programme.

In the annexes, some additional information on the November 2019 workshop and the science fiction success stories that can take place if Europe is successful in developing key technologies for post-exascale HPC systems, can be found.

2 Why looking for synergies with photonics and electronics?

The end of Dennard scaling and excessive cost of Moore's law:

Moore's law[3] was for a long time accompanied by Dennard's law, i.e. doubling the operating frequency of the processors with each generation was related to voltage reduction resulting in a constant energy density. The IC architects therefore experienced a few tens of years of "happy scaling" where performance was automatically improved with an increase of the frequency and number of transistors without impacting power consumption (see Figure 1). In the 2000s, the frequency stabilized and multi-core architectures were developed in accordance with Moore's law: as you can still increase the number of transistors per mm^2, but they cannot go faster, the logical idea is to duplicate the computing resources. However, the supply voltage now remains almost constant across generations of technology, which consequently increases the energy density in active mode. In addition, the thinness of the transistors drastically increases the leakage current which become predominant. Today, we have reached the dissipation limits of silicon. Therefore, the main challenge is now to reduce power consumption in integrated circuits, which will also stabilize the TCO (total Cost of Ownership) of "computing" and HPC centers by reducing the electricity bill.

Parameter (scale factor = a)	Classic Scaling	Current Scaling
Dimensions	1/a	1/a
Voltage	1/a	1
Current	1/a	1/a
Capacitance	1/a	>1/a
Power/Circuit	$1/a^2$	1/a
Power Density	1	a
Delay/Circuit	1/a	~1

Source: Krisztián Flautner "From niche to mainstream: can critical systems make the transition?"

Figure 1: "classical scaling" shows the parameters when Dennard's scaling was still active, when the geometrical size of the technology (the technology "node"[4]) was reduced by factor "a". "Current scaling" shows the evolution of the parameters on small technology nodes, where Dennard's scaling is not anymore valid.

Figure 2 shows the consequences on the performance of microprocessors: from about 2005 the frequency of processors stop increasing, and as the number of transistors can still increase (Moore's law) the number of cores per processor increased, while the power dissipation reaches the limit of what can be dissipated at affordable cost per chip. To further increase the global system performance, the number of processors drastically increased in HPC and datacenters, with a correlated increase of power consumption. It is believed that 30 to 50 MW is the practical limit for the consumption of a HPC or datacenter, and it is the main practical limitation for exascale machines: reaching an energy efficiency that allow having exascale capabilities in this

[3] These are called « laws » but they are not physics law, only observation and forecasts.

[4] The **technology node** (also **process node**, **process technology** or simply **node**) refers to a specific semiconductor manufacturing process and its design rules. From https://en.wikichip.org/wiki/technology_node

power and dissipation budget. It is also why there is a concern about the power consumption of ICT (see Figure 3).

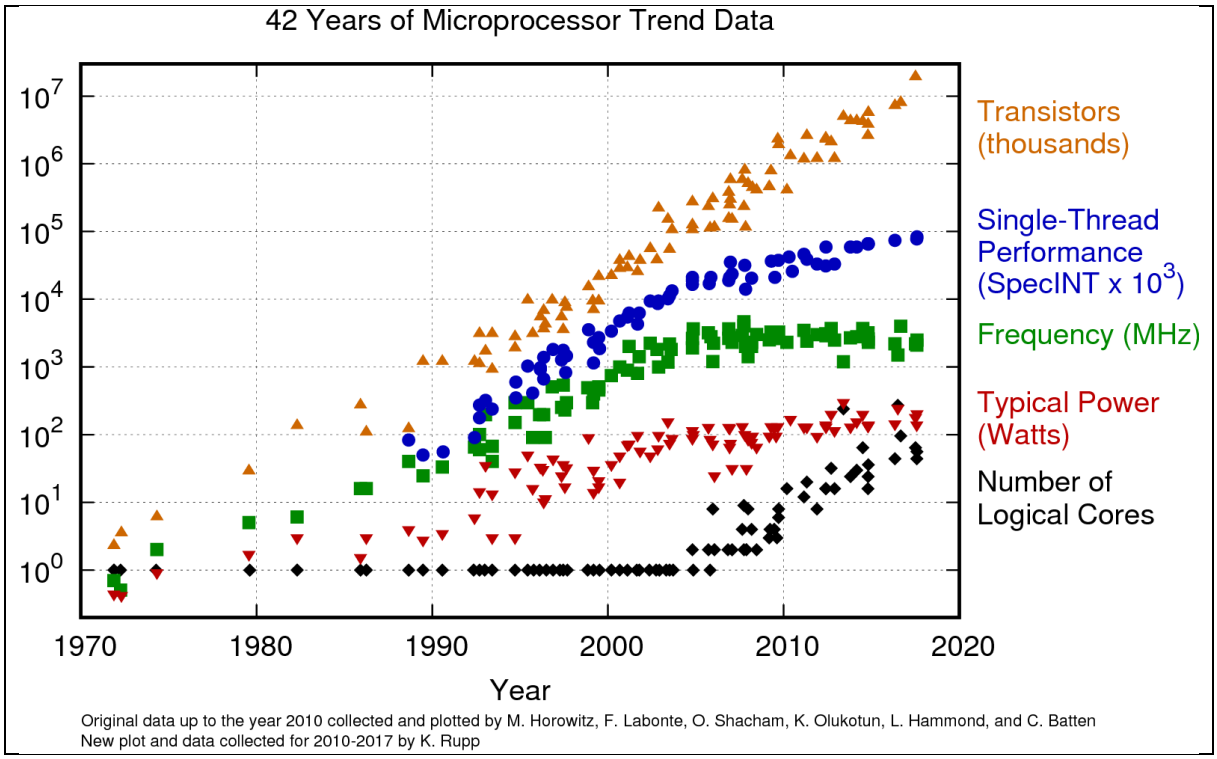


Figure 2: Evolution of processors over time

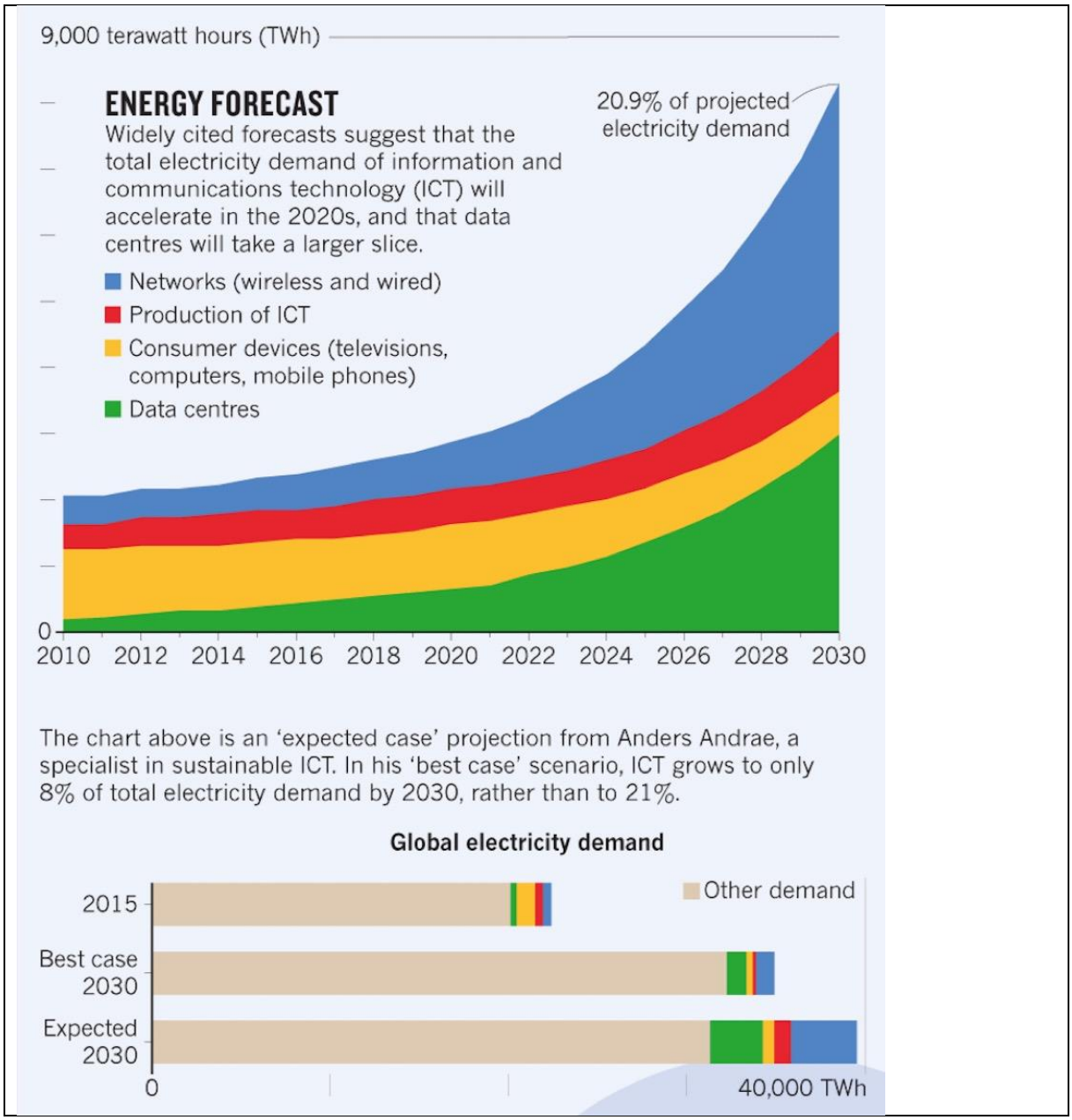


Figure 3: energy consumption of ICT (from Nature, September 12, 2018)

If we analyze the source of energy dissipation in more details, we see that transfer of data are the main source of loss and heat generation: Figure 4 shows that getting data from external DRAM takes 800 times more energy than making operations on those 64 bit data extracted from the DRAM. This is the drive of the approach of chiplets and interposers, where different "chips" are connected on a silicon interposer with a much smaller footprint than with a PCB (Printed Circuit Board). It also drives emerging architectures where computing and storage are more intertwined, like in "near memory computing", "in memory computing" or "processing in memory" architectures. Figure 5 shows this potential evolution over time, where the end of Dennard's scaling drove a rise in many-core architectures, the quest for better efficiency introduced heterogeneous architectures with a plurality of co-processors or accelerators, then a possible rise of processing in memory-based systems.

However, there is another way to decrease the energy consumption of moving data: using photons instead of electrons.

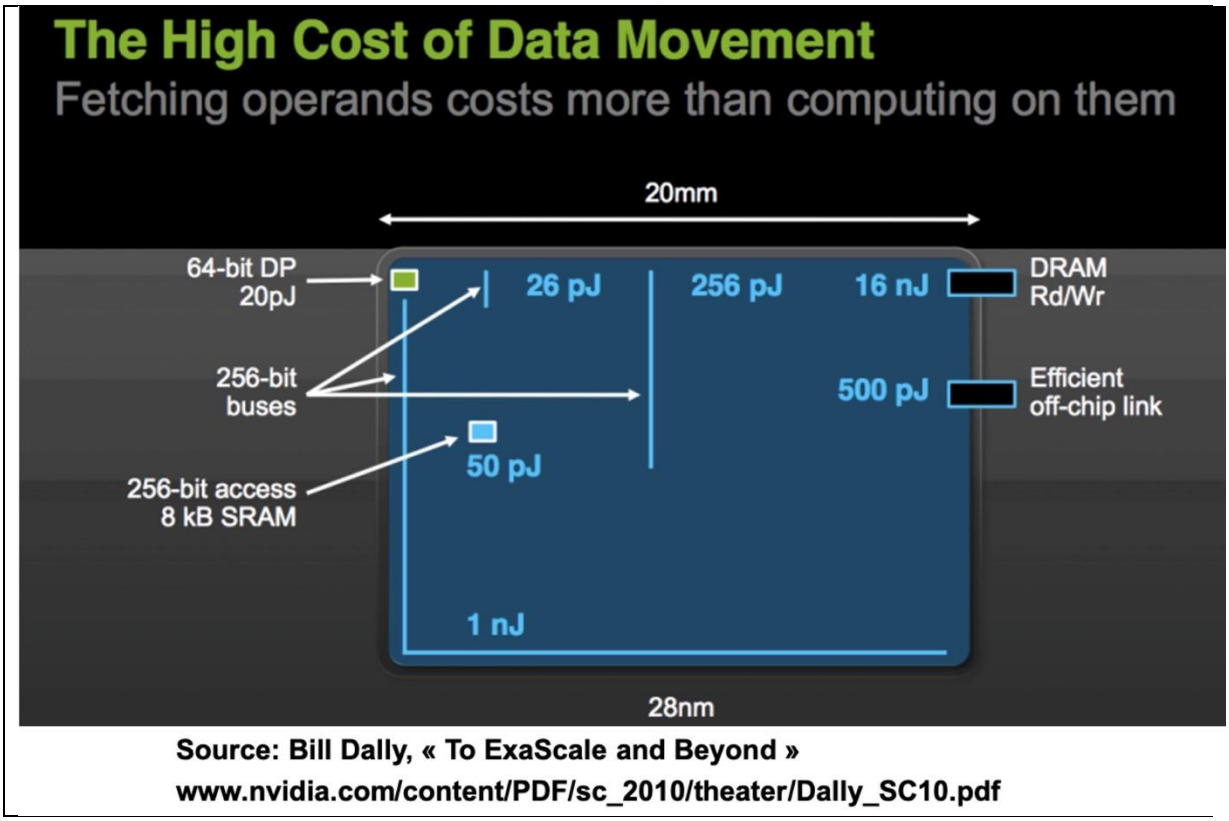


Figure 4: cost of moving data

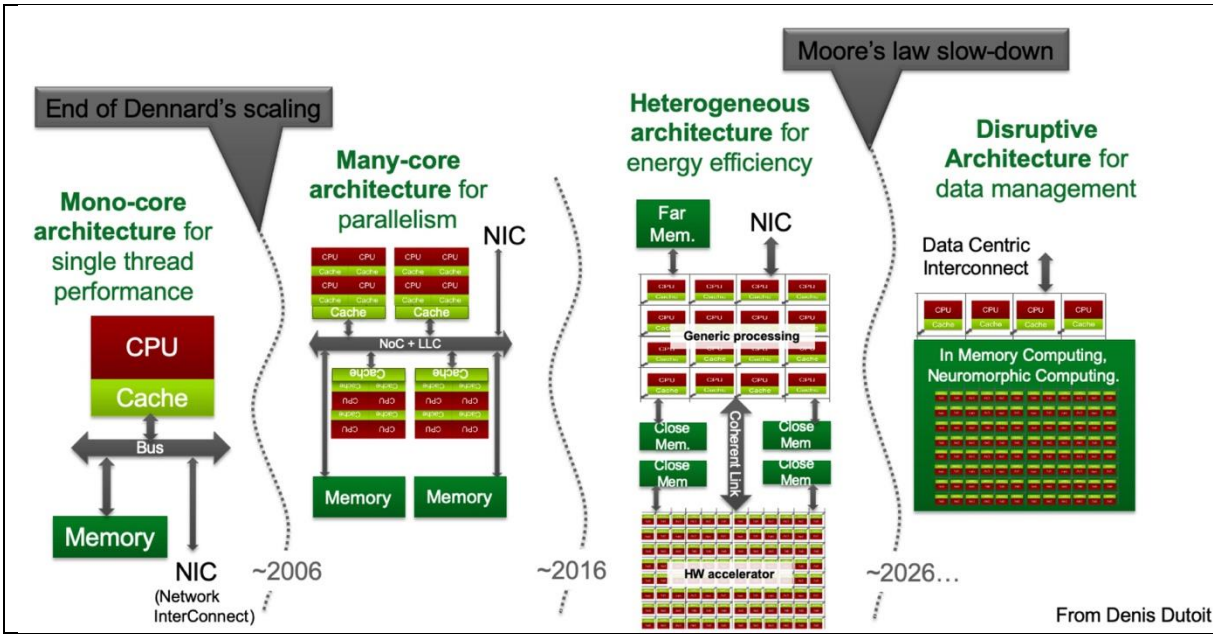


Figure 5: Evolution of computing systems over time, driven by more and more efficiency (picture from Denis Dutoit, CEA).

Electrons require little energy to create a signal and receive it, but Ohm's law means that the transmission of electrons results in the dissipation of a lot of energy. On the contrary, photons are relatively expensive to create (lasers) or to receive (sensors), but once created, they can travel on long distance with minimal attenuation. A factor, expressed in Mbit per second per km and per watt, once reached, means that optics are more efficient than electronics at transmitting information. As the throughput of current systems is always increasing, the

distance where optics are efficient decreases, perhaps down to interconnecting chiplets on an interposer, as shown in Figure 6 and Figure 7.

Without even considering computing with photons, we can see that there is a strong rationale to look for synergies between photonics and electronics for very high performance systems that will process a very large amount of data.

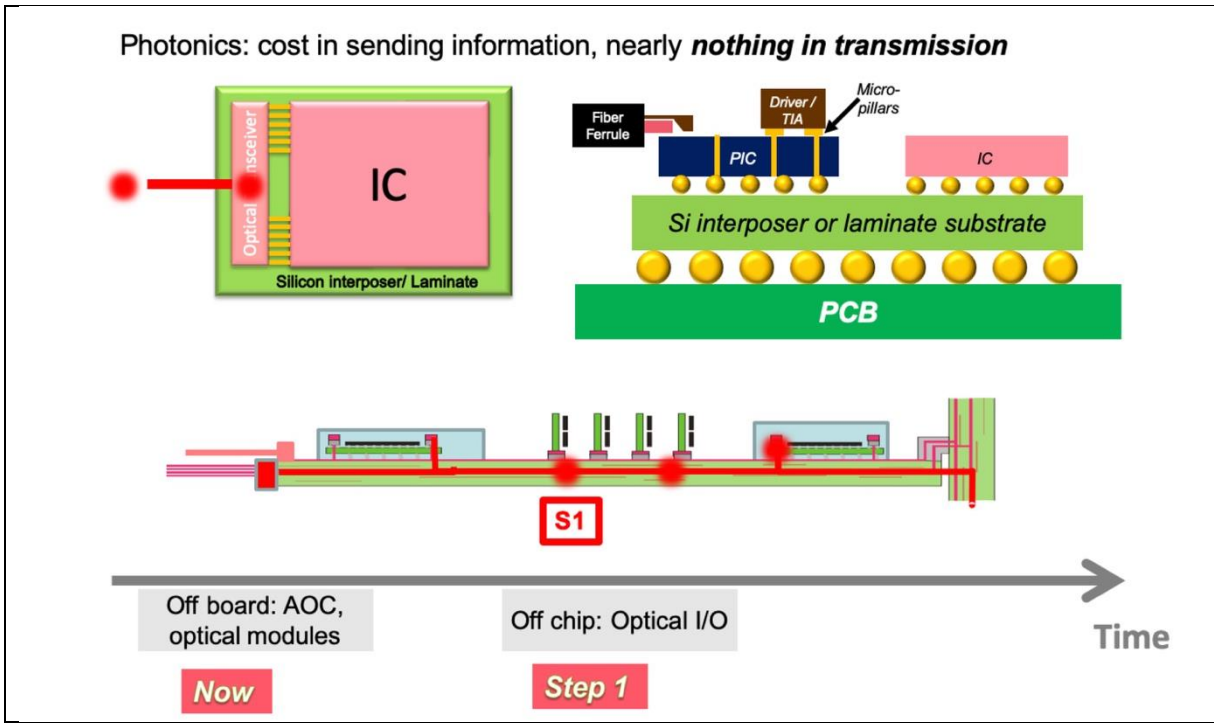


Figure 6: optical interconnect is efficient down to board, and perhaps to chip, where a serdes (electrical interconnect) is replaced by a Photonic Interconnect Circuit (PIC).

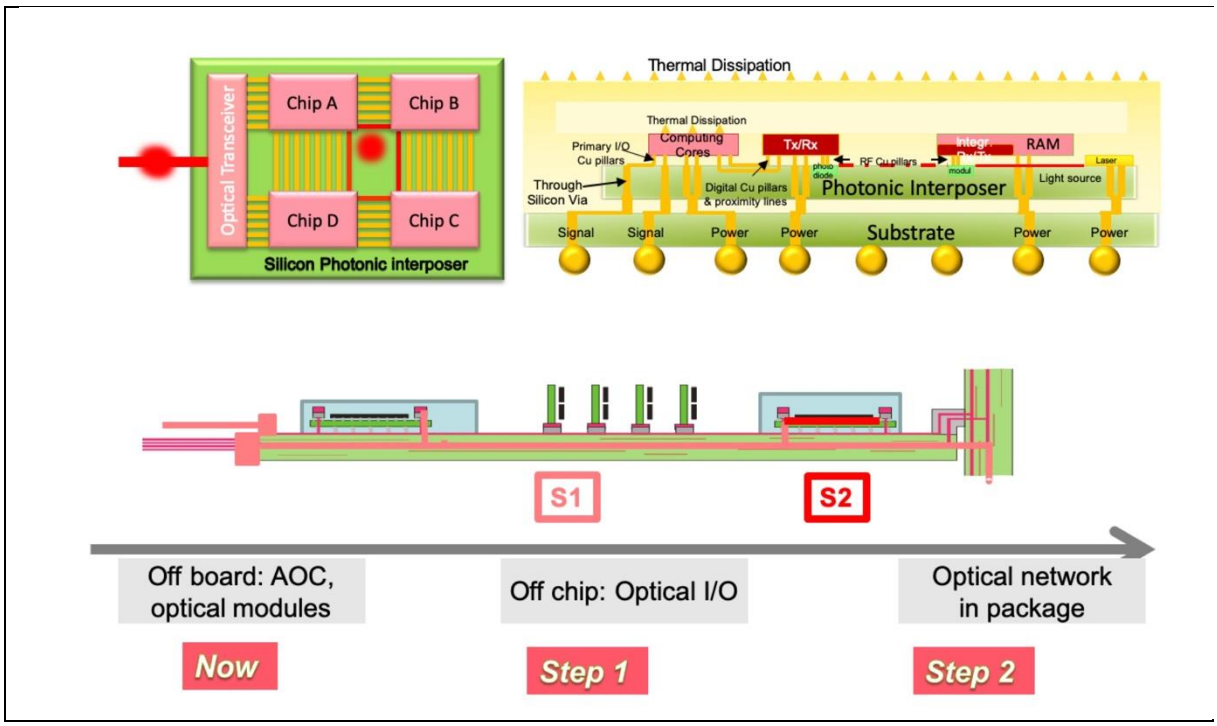


Figure 7: or efficient even at the chiplet level, with a photonic interposer.

In summary, in the post-exascale time frame, HPC system performance will not increase any further using the current technologies. There is a need for new solutions that can come from new heterogeneous architectures and new research paths being explored mainly by the photonics and electronics communities.

3 International landscape

Before discussing the European situation, it is interesting to look at what is going on in the main international HPC ecosystems. For this analysis, we have focused on the US, China and Japan which are the countries with the most advanced programmes for post exascale HPC technologies.

3.1 USA

The United States, as usual in Information Technology, has been the first ecosystem to acknowledge that a disruption has to happen to sustain the growth in performance of HPC systems. Several reports have been issued for high level advisory committees. One of them "Future High Performance Computing Capabilities" was issued as a "Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC)[5]" back in 2017. These reports were the starting point of actions to launch ambitious research programmes. DARPA (Defense Advanced Research Projects Agency) is very active in this field and has launched several programmes relevant for HPC that are having a broad impact way beyond the defense domain:

- ERI Electronics Resurgence Initiative
- ACCESS Accelerated Computation for Efficient Scientific Simulation
- PIPES Photonics in the Package for Extreme Scalability

A short description of these programmes can be found below.

3.1.1 *ERI*

Announced in 2017, ERI is a five-year, upwards of $1.5 billion investment to jumpstart innovation in the electronics industry. To address the impending engineering and economic challenges confronting those striving to push microelectronics technology forward, DARPA is nurturing research in circuit design tools, advanced new materials, and systems architectures through a mix of new and emerging programmes. The first phase of ERI is organized around 6 programmes:

- Materials & Integration
  - Monolithic Integration of an SoC in Three Dimensions (3DSoC)
  - Framework for Novel Compute (FRANC)
- Architecture:
  - Software Defined Hardware (SDH)
  - Domain-Specific System on Chip (DSSoC)
- Design
  - Intelligent Design of Electronic Assets (IDEA)
  - Posh Open Source Hardware (POSH)

The Three Dimensional Monolithic System-on-a-Chip (3DSoC) programme seeks to develop the monolithic 3D technology required to build logic, memory, and input/output (I/O) on a single die using a legacy lithography node while improving performance by more than 50X when compared with leading edge technology nodes. To achieve its goals, 3DSoC seeks to develop fabrication technology as well as the design flows required to take advantage of the technology's capabilities.

[5] https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/201712/ASCAC-Future-HPC-report.pdf?la=en&hash=2FEB999A02D5D4C30EAC01A0C090AAFCC49996E9

The Foundations Required for Novel Compute (FRANC) programme aims to develop innovative approaches to advance compute technologies beyond the Von Neumann topology. Leveraging recent advances in materials, devices, and integration technology, the programme seeks to develop novel memory-centric compute topologies that break the traditional separation of processors and memory components to realize dramatic advances in compute efficiency and throughput of the workload, especially for applications constrained by size, weight, and power (SWaP). Innovative compute architectures and new, fast non-volatile storage and memory-centric computing devices will be explored under FRANC to enable low latency compute near or inside the data storage elements. Such approaches are particularly suited for applications relevant to artificial intelligence (AI) where in-memory computation provides unique advantages over traditional Von Neumann computation.

The goal of the Software Define Hardware (SDH) programme is to build runtime-reconfigurable hardware and software that enables near ASIC[6] performance without sacrificing programmability for data-intensive algorithms. Under the programme, data-intensive algorithms are defined as machine learning and data science algorithms that process large volumes of data and are characterized by their usage of intense linear algebra, graph search operations, and their associated data-transformation operators. The SDH programme aims to create hardware/software systems that allow data-intensive algorithms to run at near ASIC efficiency without the cost, development time, or single application limitations associated with ASICs.

Domain-Specific System on Chip (DSSoC) intends to demonstrate that the tradeoff between flexibility and efficiency is not fundamental. The programme plans to develop a method for determining the right amount and type of specialization while making a system as programmable and flexible as possible. DSSoC wants to de-couple the programmer from the underlying hardware with enough abstraction but still be able to utilize the hardware optimally through intelligent scheduling. DSSoC specifically targets embedded systems where the domain of applications sits at the edge and near the sensor. Workloads consist of small chunks of data but often with a large number of algorithms required in the processing, meaning that high compute power and low latency at low power are required.

The Intelligent Design of Electronic Assets (IDEA) programme seeks to develop a general purpose hardware compiler for no-human-in-the-loop translation of source code or schematic to physical layout (GDSII) for SoCs, System-In-Packages (SIPs), and Printed Circuit Boards (PCBs) in less than 24 hours. The programme aims to leverage advances in applied machine learning, optimization algorithms, and expert systems to create a compiler that could allow users with no prior design expertise to complete physical design at the most advanced technology nodes.

The Posh Open Source Hardware (POSH) programme seeks to enable mathematically provable secure electronics and create an open source hardware IP ecosystem, along with accompanying validation tools. Under the programme, researchers will work to develop methodologies, standards, and simulation as well as emulation technologies for the verification and mathematical inspection of analog and digital IP to provide proof of functionality and security. The program also aims to develop and release a number of silicon-proven analog and digital IP blocks on an open source platform to serve as a foundation for rapid design of complex secure SoCs at leading edge technology nodes.

The initial programmes are expected to be followed by additional initiatives such as PIPES (see section 3.1.3).

[6] Application Specific Integrated Circuit

3.1.2 *ACCESS*

In comparison to ERI, ACCESS is a much smaller programme but it is interesting because it illustrates well the objective of finding disruptive computing solutions. The budget of ACCESS is small (in the range of $1M for each project) and it has the rather focused objective of developing technologies for the acceleration of scientific simulations of physical systems characterized by coupled partial differential equations (PDEs).

The Accelerated Computation for Efficient Scientific Simulation (ACCESS) programme seeks innovative ideas for computational architectures that will achieve the equivalent of petaflops performance in a benchtop form-factor and be capable of what traditional architectures would define as “strong” scaling for predictive scientific simulations of interest.

The design and development of the prototypes are envisioned to leverage advances in optics, MEMS, additive manufacturing, and other emerging technologies to develop new non-traditional analog and digital computational means and to overcome some of the current known limitations of these means, such as precision and stability. Of particular interest are hybrid analog/digital architectures that replace numerical methods and memory-intensive computational parallelization with nonlinear and/or intrinsically parallel physical processes to perform computations.

Unfortunately, we have not been able to find the list of the projects funded by this programme.

3.1.3 *PIPES*

After the ERI first phase and the launch of 6 research areas (see section above), it appears that other domains have to be investigated to complement this first effort. Photonics was one of these and the ERI phase II includes a new programme PIPES for this field.

The Photonics in the Package for Extreme Scalability (PIPES) programme, seeks to enable future system scalability by developing high-bandwidth optical signalling technologies for digital microelectronics. Working across three technical areas, PIPES aims to develop and embed integrated optical transceiver capabilities into cutting-edge MCMs and create advanced optical packaging and switching technologies to address the data movement demands of highly parallel systems.

The first technical area of the PIPES programme is focused on the development of high-performance optical input/output (I/O) technologies packaged with advanced integrated circuits (ICs), including field programmable gate arrays (FPGAs), graphics processing units (GPUs), and application-specific integrated circuits (ASICs). Beyond technology development, the programme seeks to facilitate a domestic ecosystem to support wider deployment of resulting technologies and broaden their impact.

The second technical area investigates novel component technologies and advanced link concepts for disruptive approaches to highly scalable, in-package optical I/O for unprecedented throughput. The objective is to answer the need for enormous improvements in bandwidth density and energy consumption to accommodate future microelectronics I/O.

The third technical area of the PIPES programme will focus on the creation of low-loss optical packaging approaches to enable high channel density and port counts, as well as reconfigurable, low-power optical switching technologies. This aims to enable the development of massively interconnected networks with hundreds to thousands of nodes that are expected due to the advance in the 2 previous areas.

The total budget is foreseen in the range of $65M for the three areas. The selected projects have started during the second half of 2019.

3.1.4 *Global view*

The US effort is not limited to the above described programmes. Nevertheless, they give an impression of how the US ecosystem acts and how the priorities are defined.

Besides these activities, it is worth mentioning:

- The US AI initiative launched in February 2019;
- The National Quantum Initiative Act issued in December 2018. This initiative has announced a budget of $1.2B to develop US leadership in quantum.

In conclusion, the US is undertaking a large research effort to maintain a dominant position in computing and to prepare the technologies that will replace the current CMOS based chips.

3.2 China

Assessing the efforts of China on the future of HPC technologies, is difficult as access to information related to the actual research projects is limited. This is why we have focused our analysis on the views presented in a journal[7] by members of the Chinese HPC ecosystem "Special Issue on Post-exascale Supercomputing" issued in November 2018.

The main challenges identified by the Chinese HPC research community are:

1. **Energy efficiency bottlenecks**: The US Department of Energy's exascale research programme sets a goal of 1 exaflops at 20–40 MW, or 25–50 gigaflops/W, probably around the year 2022. The US DARPA's JUMP programme sets a more ambitious long-term goal of 3 peta operations per second per watt, or 3 peta operations per joule (POPJ), possibly by around 2035. Here an operation is not necessarily a 64-bit IEEE floating-point operation. Cambricon-1A, which was developed at the Institute of Computing Technology of the Chinese Academy of Sciences in 2015 and targets machine learning on small terminals such as smartphones, reached over 1 tera operations per joule (TOPJ)
2. **Order-of-magnitude better devices**: Emerging technologies, such as 3D-stacking, fully optical communication, magnetic semiconductors, and memristors, are challenging mature technologies used in today's supercomputers which are based on CMOS
3. **Novel systems architectures**: Systems architecture has played a critical role in the history of modern supercomputing. Architectural innovations, from vector supercomputers, SMP, ccNUMA, and MPP, to clusters, have enabled the exponential growth of performance and scalability
4. **Effective co-design of software and hardware**: Currently, there is a wide gap between the peak performance and the sustained performance that real applications can achieve, especially with new applications with sparsity and irregularity, such as data analytics and complex multi-modal applications
5. **Ecosystem for diverse applications**: the existing ecosystem has a tradition of scientific and engineering computing, which is not enough for the new diverse applications that converge numeric simulation, big data, and artificial intelligence. China proposes to build up a new supercomputing ecosystem for application development, which supports the mixed or converged workloads of arithmetic-intensive, data-intensive, and intelligent applications.

[7] https://link.springer.com/journal/11714/19/10

In order to tackle these challenges, the Chinese researchers work on the different HPC system components.

At computing node level, the solution that is proposed in based on the following choices:

- Processor: many-core architecture with each processing core supporting scalar double precision floating-point processing instead of vector processing
- Co-processor: acceleration for specific applications, such as traditional scientific computation and engineering applications, and emerging applications including data analytics and deep learning
- Sharing of high bandwidth memory
- Inter-processor link

For the network, the expected technologies will serve to interconnect large nodes that are a heterogeneous mix of central processing units (CPUs), accelerators, co-processors, and field programmable gate arrays (FPGAs)/application specific integrated circuits (ASICs). The CPUs could also be a mix of strong and weak cores. The nodes will also have a large amount of memory of different technologies, such as non-volatile random-access memory and three dimensional (3D) stacked memory. The network technologies will:

- Be heterogeneous with wired or wireless interconnects. The on-chip interconnect technology would have matured to incorporate wireless interconnection among the components within a node. Similarly, photonic technologies would have matured to be used within a node or a rack. This can provide a large number of concurrent communications among different components (CPUs, accelerators, and memories) without contention. As each node will be dense, the nodes will need to be connected to the overall network speed of terabits per second with multiple adapters/ports. This will facilitate a good balance between inter- and intra-node transfers;
- Both wireless and photonic technologies will allow the establishment of high-dimensional topologies within intra-core, intra-node, and intra-rack levels. These technologies will also facilitate one-to-many, many-to-one, and many-to-many communications in a flexible manner with a good performance
- Allow capabilities to be increasingly offloaded to the network leading to an era of ‘in-network computing’

The analysis of the Chinese community is not limited to hardware technologies. In terms of software, they acknowledge that progress is also needed here if we want to have efficient post-exascale systems. The directions that are highlighted are:

- programming effort shifting from computation to data,
- precision optimization,
- programming hardware instead of software.

In summary, the Chinese ecosystem is already working on technologies for post-exascale systems both at the hardware and software levels.

3.3 Japan

To analyse the situation in Japan is more difficult because most of the documents are in Japanese and we do not have found a specific publication like for China which summarizes the vision of the Japanese HPC ecosystem on post-exascale technologies. However, an impression of the current Japanese efforts in technology can be given by looking at their current initiatives in the domains of supercomputing, Artificial Intelligence and quantum technologies.

3.3.1 *Supercomputing*

Japan has a long history of development of supercomputers. The Earth Simulator was dethroned in November 2004 as the top supercomputer in the world, but the Fujitsu's K computer, based on 68544 SPARC64 VIIIfx CPUs, each with eight cores, for a total of 548,352 core, processors developed in house, was at the top of the TOP500 in 2011 with 10 petaflops. It was developed for RIKEN (an Institute for Physical and Chemical Research). It should be noted that the K Computer did not use graphics processors or other accelerators. The K Computer was also one of the most energy-efficient systems, and while dethroned on the TOP500, it stayed or a long time at a good position on the Green500. It was also a very "equilibrated" machine, with a good ratio compute/storage/communication, making it quite efficient on the HPCG benchmark where it was dethroned only in 2018.

Since 2016, the fastest supercomputer in Japan has been Oakforest-PACS hosted by the JCAHPC (Joint Center for Advanced HPC), a joint center between the University of Tsukuba and the University of Tokyo. This supercomputer used an Intel Xeon Phi (Knights Landing) and Intel Omni Path Architecture interconnection network with 8208 nodes for 25 petaflops.

The new generation of exascale range machine developed in Japan for the RIKEN by Fujitsu is the Fugaku machine which should be fully operational in 2021. Its computing capabilities will be 100 times those of the K computer, but it will follow the K philosophy about being a "balanced" machine, aiming to have the best level of practicability in the world, thanks to a "codesign" approach. Unlike the US machines (Summit and Sierra) which rely on GPU to reach top performance, the Fugaku machine does not use discrete accelerator chips. Instead, the ARM v8.2-A cores, custom designed by Fujitsu, use long vector SVE (Scalable Vector Extension) extensions with a SIMD length of 512 bits developed in collaboration between ARM and Fujitsu. The chip has 48 cores (+ 2 or 4 for OS) reaching more than 2.7 teraflops per 48 cores at 2.0 GHz with boost to 2.2 GHz. The 7nm FinFET chip uses low power logic design, allowing to reach 15 GF/W @ dgemm. It should also be mentioned that the software environment is ported to the ARM processor instruction set.

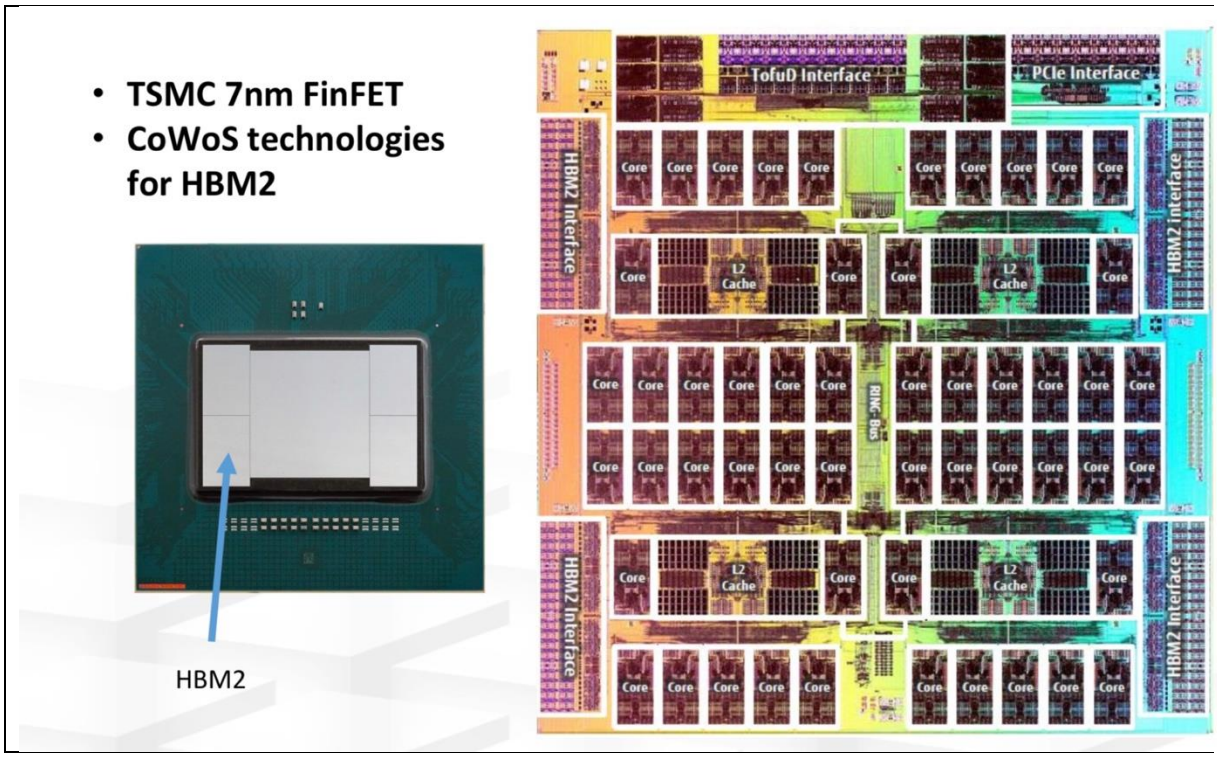


Figure 8: the Fujitsu A64FX chip, core of the Fugaku computer

The machine will have two types of nodes: Compute Node and Compute & I/O Node connected by Fujitsu TofuD, a 6D mesh/torus Interconnect.

In term of energy efficiency, the prototype of Fugaku was at the top of the Green500 at the end of 2019 with 16.9 gigaflops/watt.

3.3.2 *Artificial Intelligence*

Japan is offering supercomputing capacities to their researchers (and companies) working on Artificial Intelligence. In 2015, The Ministry of Economy and Industry (METI) started the AIRC (Artificial Intelligence Research Center), hosted by the AIST (Advanced Institute for Science and Technologies). The 2016 budget was 19.5 billion yen. They installed the "AI Bridging Cloud Infrastructure" (ABCI) which reached about 20 petaflops in 2019. "AI Bridging Cloud Infrastructure (ABCI) is the world's first large-scale Open AI Computing Infrastructure, constructed and operated by National Institute of Advanced Industrial Science and Technology (AIST), Japan. It delivers 19.9 petaflops of HPL performance and the world's fastest training time of 1.17 minutes in ResNet-50 training on ImageNet datasets as of July 2019. ABCI consists of 1,088 compute nodes each of which is equipped with two Intel Xeon Gold Scalable Processors, four NVIDIA Tesla V100 GPUs, two InfiniBand EDR HCAs and an NVMe SSD. ABCI offers a sophisticated high performance AI development environment realized by CUDA, Linux containers, on-demand parallel filesystem, MPI, including MVAPICH, etc." (from https://insidehpc.com/2019/09/the-abci-supercomputer-worlds-first-open-ai-computing-infrastructure/).

3.3.3 *Quantum Computing*

Japan is also active in quantum computing, but mainly in "analog" quantum computing[8] (quantum annealing). There are a few start-ups on quantum computing, and also a flagship project financed by MEXT and working on transmon[9] based qubits (for the "universal" quantum computer). Most applications are based on combinatorial optimization for industry, transportation or traffic. Software stacks and simulations are also being developed, also for the "universal quantum" approach, e.g. by the MDR company (https://mdrft.com/).

On the hardware side, NEC is developing a superconducting flux qubit for quantum annealer, somewhat a competitor of the D-Wave machine. They claim that their forthcoming machine will be more efficient than D-Wave because they will support a more generic interconnect, event better than the new "Pegasus" interconnect of D-Wave. Optical based computing has a long history in Japan, and NTT is developing a quantum annealer based on laser and coherent ising principle. The Cabinet Office ImPACT "Quantum Neural Network" involves organizations such as NTT, NII (National Institute of Informatics). Fujitsu, Hitachi and Toshiba have digital solutions for the hardware implementation of simulated annealing. Fujitsu's Digital Annealer[10] is an ASIC which can be used as coprocessor. The first generation (1024 "bits" with 16 bits inter-bit coupling accuracy) is accessible by the cloud, while the second generation will support 8192 "bits" with 64 bits for coupling accuracy. By combining chips by software, the 2019 servers can support up to 1M "bits". Hitachi has developed a CMOS Annealing machine[11]. Toshiba[12] is developing hardware to support their quantum-inspired algorithm "Simulated Bifurcation Algorithm".

[8] See section 6.7 for the definition of "analog" and "universal" quantum computers

[9] A **transmon** is a type of superconducting charge qubit that was designed to have reduced sensitivity to charge noise. The transmon was developed by Robert J. Schoelkopf, Michel Devoret, Steven M. Girvin and their colleagues at Yale University in 2007 (from Wikipedia)

[10] https://www.fujitsu.com/global/services/business-services/digital-annealer/index.html

[11] https://www.hitachi.com/rd/portal/contents/story/cmos_annealing2/index.html

[12] https://www.toshiba-sol.co.jp/en/pro/sbm/sbm.htm

4 European effort

In this research domain for future computing solutions Europe has been active and various organisations or programmes work to develop relevant technologies. This section presents a quick overview of the main stakeholders or initiatives.

4.1 ECSEL-Aeneas

The ECSEL (Electronic Components and Systems for European Leadership) JU (Joint Undertaking) is undertaking a huge research effort in the digital technologies for Europe with more than 3.4 B€ of funding committed to 64 projects since 2014. Even if HPC was not a priority at the beginning of ECSEL, this has changed as it can be seen in the last multi-annual Strategic Plan[13] where the HPC field is well connected to the goals of the Chapter "Computing and storage".

We have been in contact with ECSEL to discuss how to develop the synergy between the electronic community at the heart of ECSEL and the HPC ecosystem. It is obvious that in the technology process field, there is currently not a good fit between the directions taken by the European industry players and the trend in HPC to use the more advanced technology nodes to reduce the number of chips to interconnect and build an HPC system. Nevertheless, there are other technical domains developed in Europe within ECSEL that are important for HPC like energy management, cooling, integration of heterogeneous chips or new architectures.

There is an agreement to foster the interaction between HPC and ECSEL. The main driver for the interaction is the participation of experts from both domains to the activities of ECSEL and of ETP4HPC. This is expected to continue with the forthcoming JU on Key Digital Technologies.

One of the main interests of ECSEL is the strong industrial focus of the programme. Thanks to the strong participation of the AENEAS (Association for European NanoElectronics ActivitieS) members in the research projects of ECSEL, there has been a very good exploitation track of the ECSEL project results. We have also been in contact with AENEAS in order to have a good connection with this ecosystem that can help to bring to market new electronic technologies relevant for future HPC system. Again, there are already links with the participation within AENEAS of experts with HPC background and focus. To foster the interaction, AENEAS was invited to and attended the workshop organized by EXDCI-2 to discuss future computing technologies (see Chapter 5).

4.2 Photonics21

The European Technology Platform Photonics21[14] has been our contact to link with the photonics industries and relevant R&D stakeholders. Photonics21 has more than 2500 members. This shows that there are various applications fields such as ICT, lighting, industrial manufacturing, life science, safety as well as in education and training. So computing is only part of the scope of this community, which is more fragmented as the different value chains are sometimes not connected. Within Photonics21 we have been mainly in contact with their Working Groups (WG) 1 and 7 which deal respectively with "information and communication"

[13] https://www.ecsel.eu/sites/default/files/2020-01/ECSEL%20GB%202019.134%20-%20MASP%202020%20and%20Annex.pdf

[14] https://www.photonics21.org/index.php

and "research, education and training". In this photonics community, you also find the main stakeholders of the silicon photonics domain which is one of the most relevant research domains for HPC.

In order to foster the relationship between the two ecosystems, we presented the HPC cPPP and EuroHPC JU during the Photonics21 annual event. The Photonics21 association was also invited to and attended the workshop organized by EXDCI-2 to discuss future computing technologies (see Chapter 5).

4.3 ICT and FET calls

The European Commission has been active in developing future technologies for computing with different calls coming from either the FET (Future and Emerging Technologies) or ICT (Information and Communication Technologies) programmes. Some of the calls were the implementation of the Photonics cPPP (driven by the EC with Photonics21) and some other calls were complementary to the ECSEL actions with a lower TRLs (Technology Readiness Levels) target.

The main calls (for which we already know the funded projects) and the most interesting projects (for which we can already see of the results) are presented in the following paragraphs.

4.3.1 *Photonics calls*

One of the first photonics calls of Horizon2020 was ICT-27-2017. This call had a specific topic on data centre communication. The challenge was to address low-cost, energy-efficient photonic devices supporting radically new system and network architectures driven by the emergence of exa-scale cloud datacentres. The projects had to focus on optical inter- and intra-data centre transmission, switching and interconnects facilitating Tb/s interface speeds and Pb/s network throughput.

This call has funded important projects like:

- WIPE - Wafer scale Integration of Photonics and Electronics
- L3MATRIX - Large Scale Silicon Photonics Matrix for Low Power and Low Cost Data Centers
- ICT-STREAMS - Silicon Photonics Transceiver and Routing technologies for High-End Multi-Socket Server Blades with Tb/s Throughput interconnect interfaces
- Teraboard - High density scalable optically interconnected Tb/s Board

As an illustration of the relevance of these projects for HPC, the ICT-STREAMS has worked on a processor-processor optical interconnect (see Fig 9).

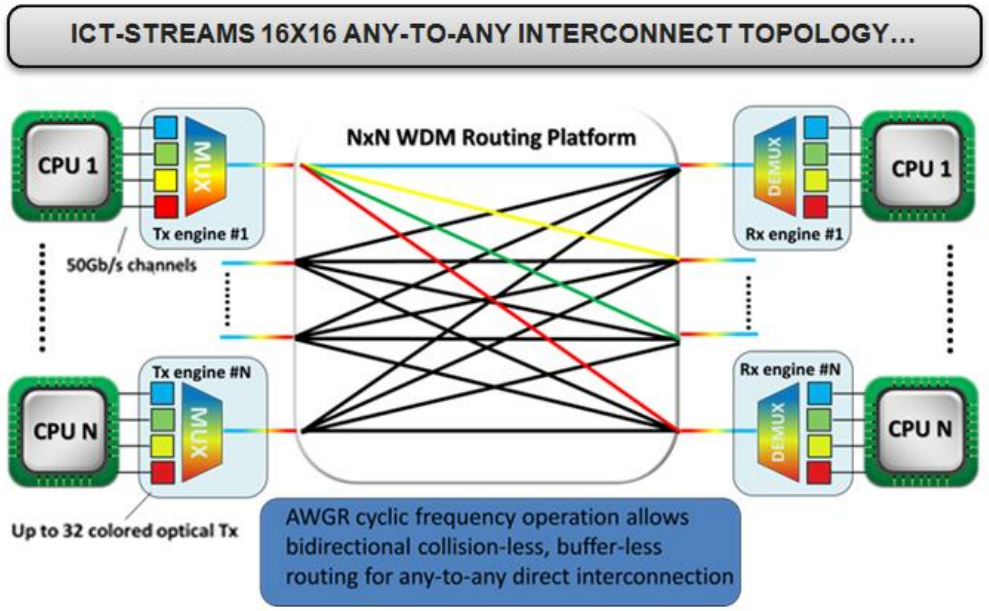


Figure 9: ICT-STREAMS interconnect topology

The FETOPEN-01-2016-2017 call has also selected research projects aiming to develop new computing options. Among the projects are:

- PHASE-CHANGE SWITCH: Phase-Change Materials and Switches for Enabling Beyond-CMOS Energy Efficient Applications
- CHIRON - Spin Wave Computing for Ultimately-Scaled Hybrid Low-Power Electronics
- COPAC - Coherent Optical Parallel Computing
- SiLAS – SiliconLaser

The ICT-30-2017 KET Photonics has a focus on the component level. One of its objectives was to achieve major advances in chip integration technology, enabling a cost-effective volume manufacturing of PICs with significantly enhanced performances (e.g. integration complexity, footprint, energy efficiency, speed…) or new functions. So, some of the selected projects can have interesting contributions to the emergence of new silicon photonics devices:

- plaCMOS - Wafer-scale, CMOS integration of photonics, plasmonics and electronics devices for mass manufacturing 200Gb/s non-return-to-zero (NRZ) transceivers towards low-cost Terabit connectivity in Data Centers
- MOICANA - Monolithic cointegration of QD-based InP on SiN as a versatile platform for the demonstration of high performance and low cost PIC transmitters
- 3PEAT - 3D Photonic integration platform based on multilayer PolyBoard and TriPleX technology for optical switching and remote sensing and ranging applications
- PASSION - Photonic technologies for progrAmmable transmission and switching modular systems based on Scalable Spectrum/space aggregation for future agIle high capacity metrO Networks
- PICTURE - High Performance and High Yield Heterogeneous III-V/Si Photonic Integrated Circuits using a Thin and Uniform Bonding Layer.

The ICT-03-2018-2019 - Photonics Manufacturing Pilot Lines for Photonic Components and Devices was also dedicated to the development of new approaches for photonics components. However, the selected projects are not targeting components that are relevant for computing.

The ICT-05-2019 - Application Driven Photonics Components call partially focused on computing. One of the topics was photonics System on Chip/ System in Package for optical interconnect applications. The actions are supposed to address advanced techniques for the intimate combination of photonic integrated circuit technology with other enabling circuits, devices and mother boards to realise major advances in the capability, performance and complexity of photonic system-on-chip and system-in-package components targeting photonic interconnect applications in the network, datacentre and consumer communication space. For this topic, the following projects have just started:

- POETICS CoPackaging of Terabit direct-detection and coherent Optical Engines and switching circuits in mulTI-Chip moduleS for Datacenter networks and the 5G optical fronthaul
- TWILIGHT Towards the neW era of 1.6 Tb/s System-In-Package transceivers for datacenter appLIcations exploiting wafer-scale co-inteGration of InP membranes and InP-HBT elecTronics
- NEBULA Neuro-augmented 112Gbaud CMOS plasmonic transceiver platform for Intra- and Inter-DCI applications.

4.3.2 *Electronics calls*

The ICT-4-2015 has been a first attempt in Horizon2020 to target new technologies for low power computing. The call's target was highly performing low-power low-cost micro-servers, using cutting-edge technologies like, for example, optical interconnects, 3D integrated system on chip, and innovative power management, which can be deployed across the full spectrum of home, embedded, and business applications.

Among the funded projects are:

- SAFEPOWER - Safe and secure mixed-criticality systems with low power requirements
- TULIPP - Towards Ubiquitous Low-power Image Processing Platforms
- ARGO - WCET-Aware Parallelization of Model-Based Applications for Heterogeneous Parallel Systems
- OPERA - lOw Power heterogeneous architecture for nExt generation of smaRt infrastructure and platforms in industrial and societal Applications
- dReDBox - Disaggregated Recursive Datacentre-in-a-Box
- HERCULES - High-Performance Real-time Architectures for Low-Power Embedded Systems
- VINEYARD - Versatile Integrated Accelerator-based Heterogeneous Data Centres
- LPGPU2 - Low-Power Parallel Computing on GPUs 2
- M2DC - Modular Microserver DataCentre
- UniServer - A Universal Micro-Server Ecosystem by Exceeding the Energy and Performance Scaling Boundaries
- TANGO - Transparent heterogeneous hardware Architecture deployment for eNergy Gain in Operation
- PHANTOM - Cross-Layer and Multi-Objective Programming Approach for Next Generation Heterogeneous Parallel Computing Systems

In this list the HERCULES, VINEYARD and LPGPU2 were projects with interesting contributions for HPC.

The ICT-25-2015 was also among the first Horizon2020 calls to address innovative nano-electronics solutions relevant for high performance computing. Some of the objectives were:

- Integration of functionalities in a system-on-chip (SoC) or system-in-package (SiP) by using nanostructures and/or nanodevices.
- New computing paradigms like quantum computing and neuromorphic computing with a focus on their future integration with Si technologies.

This call has funded the following projects that are very relevant for the future of computing:

- NeuRAM3 - NEUral computing aRchitectures in Advanced Monolithic 3D-VLSI nano-technologies
- CONNECT - CarbON Nanotube compositE InterconneCTs
- EUROPRACTICE 2016 - EUROPRACTICE training, CAD and prototyping services for European universities and research institute
- INSIGHT - Integration of III-V Nanowire Semiconductors for next Generation High Performance CMOS SOC Technologies
- NanoStreeM - NANOmaterials: STRategies for Safety Assessments in advanced Integrated Circuits Manufacturing
- METRO4-3D - Metrology for future 3D-technologies
- IONS4SET - Ion-irradiation-induced Si Nanodot Self-Assembly for Hybrid SET-CMOS Technology
- NEREID - NanoElectronics Roadmap for Europe: Identification and Dissemination
- SUPERAID7 - Stability Under Process Variability for Advanced Interconnects and Devices Beyond 7 nm node
- PETMEM - Piezoelectronic Transduction Memory Device
- STREAMS - Smart Technologies for eneRgy Efficient Active cooling in advanced Microelectronic Systems
- GREAT - heteroGeneous integRated magnetic tEchnology using multifunctional standardized sTack (MSS)
- MOS-QUITO - MOS-based Quantum Information TechnOlogy
- REMINDER - Revolutionary embedded memory for internet of things devices and energy reduction
- Nanonets2Sense - Nanonets2Sense
- PHRESCO - PHotonic REServoir Computing

For example, the NeuRAM3 project has been an important project for the European research on neuromorphic architectures.

The ICT-31-2017 call was focused on new topics still at the research stage (TRLs 2-3). The objective was the development of new approaches to scale functional performance of information processing and storage substantially beyond the state-of-the-art technologies with a focus on ultra-low power and high performance. This call has funded 3 projects that are interesting in the context of new computing architectures:

- MNEMOSENE - Computation-in-memory architecture based on resistive devices
- Fun-COMP - Functionally scaled computing technology: From novel devices to non-von Neumann architectures and algorithms for a connected intelligent world
- 3eFERRO - Energy Efficient Embedded Non-volatile Memory Logic based on Ferroelectric Hf(Zr)O2.

The more conventional approach of the ICT-05-2017 - Customised and Low Energy Computing call needs also to be mentioned. This call has funded the Montblanc 2020 project which is a precursor of the European Technology Initiative (EPI) (funded under a Framework Partnership Agreement). EPI is the major European effort to develop technologies for HPC processor and accelerator.

The ICT-06-2019 call Unconventional Nanoelectronics was targeting more future approaches to computing components. Its focus is to demonstrate new concepts at the transistor or circuit level which bring the potential of highly improved performance for generic or specific applications. This can be based on materials, computing unit architecture (transistor or beyond) as well as at circuit level. Still the focus is on devices and components, as well as related processing technologies. The funded projects under this call are just starting. The list is:

- NeurONN – Two-Dimensional Oscillatory Neural Networks for Energy Efficient Neuromorphic Computing
- SEQUENCE – Cryogenic 3D Nanoelectronics
- BeFerroSynaptic – BEOL technology platform based on ferroelectric synaptic devices for advanced neuromorphic processors
- ZeroAMP – Nanomechanical Switch-Based Logic and Non-Volatile Memory for Robust Ultra-Low Power Circuits
- NEoteRIC – Neuromorphic Reconfigurable Integrated photonic Circuits as artificial image processor
- MUNFAB – Modeling Unconventional Nanoscaled Device FABrication
- PlasmoniAC – Energy and Size-efficient Ultra-fast Plasmonic Circuits for Neuromorphic Computing Architectures
- MeM-Scales – Memory technologies with multi-scale time constants for neuromorphic architectures

Besides these existing projects, Horizon2020 still has calls that are very relevant for the development of future computing technologies. The FETPROACT-09-2020 will address neuromorphic computing technologies. The objective is to target new computational substrates and engines, based on new materials and engineering principles for efficient and low-power neuromorphic computing; together with new theories, architectures and algorithms for neuromorphic computation (classification, control,…), learning (including unsupervised, incremental, single-shot and/or event-based) and adaptation/plasticity for and in such new neuromorphic hardware.

4.3.3 *Summary of European position*

Europe has a strong scientific background and is on par with other countries concerning emerging technologies. It has developed a research ecosystem relevant for proposing new options for computing/storage/networking within the Horizon2020 programme. However, this research network is less deeply connected to the industry than in the US for example.

The level of funding for this domain is also not at the same level as the US effort. This could be a risk as larger investments could be needed to bring to the market some emerging technologies. The lack of risk taking (for European venture capitalists, and even for companies) compared to some other part of the world might prevent the emergence of future key players, as we don’t know yet which technology will be really key in the future.

5 Workshop with electronics and photonics ecosystems

To foster the relationship between the photonics, electronics and HPC ecosystems, we organized first a small workshop with the 3 main RTOs (Research and Technology Organizations), namely CEA/Leti, IMEC and Fraunhofer in November 2018 in Brussels. The objectives were to introduce the HPC challenges to the photonics and electronics communities and to analyse what are the main research areas in these fields that are relevant to tackle these issues. It was also a way to gather a core team of people that have been involved in the follow-up actions of this task.

Amongst the follow-up action was the organization of a larger workshop in November 2019 again in Brussels. This workshop is described in the next sections.

5.1 Introduction

The preparation of this workshop was done by a core team with experts coming from photonics, electronics and HPC. We decided to focus the workshop on technologies relevant for high performance chips that can deliver computing power either at the edge or in HPC systems. This choice was made because there was a consensus that the same technologies can applied for both domains and that it was interesting to target a broader market.

For the participation we decided to keep the attendee number under 50 in order to have more interactive sessions and to foster the discussions between experts. As we had more than 20 speakers the ratio of participants without a presentation was around 50%. This part of the invitees was mainly experts from industry. We also had representatives from the European Commission and people representing Aeneas and Photonics21. The other CSA (Coordination and Support Action) on HPC, Eurolab4HPC, was also invited to coordinate the vision of the two CSAs. The industry representatives were from:

- Atos/Bull
- Global Foundries
- IBM
- Infineon
- Lighton (a French SME developing optical analog computing solutions)
- NXP
- ST
- VTT

It was important for us to have such participation from industry, as one of the objectives is not only to organize a new research network able to work on new technologies but also to develop a connection with European industry that could bring to market these innovations.

A list of the participants is given in Annex 9.2.

5.2 Agenda

The agenda of the workshop started with a session to present today's challenges in HPC. This session covered both the state-of-the-arts and current limitations of HPC systems and the evolution of HPC applications. The last topic is important to really address the requirements of tomorrow's HPC usages and to be able to propose disruptive approaches interesting for HPC applications but not in line with current HPC systems.

The second session was dedicated to the progress of traditional electronics. As there are still some advances to expect from the CMOS technology, it was important to get a view on how far this approach can go and what are the things that may change at the very end of Moore's law.

The third session covered some of the European research in the field of new architectures to support AI. As HPC usage evolves with the support of compute intensive AI applications and the integration of AI technics in conventional HPC applications, this topic is of importance. There are also a lot of different approaches that can be either in competition or complementary.

The fourth session focused on new technologies that can replace CMOS and the "in memory computing" architecture. The two were presented together, as most of the time, IMC architectures are based on new materials to store data and propose operations on those data.

The fifth session was dedicated to silicon photonics and optical analog computing. The idea was to see how using photons instead of electrons could be beneficial in future HPC systems.

The workshop was ended by a session on the integration of heterogeneous technologies that could be a real enabler for integrating new technologies while continuing to benefit from the best of CMOS technology.

One could be surprised that there was not a session on quantum computing. This was a thoughtful decision based on the following reasons:

- The focus of this task is to investigate technologies that could reach the market in the 2025-2035 time frame;
- The quantum universal model (based on entangled qubits in superposition states and on gates to drive the interdependencies of the qubits) has important roadblocks to solve before it can be expected to get a real system implementing this model;
- The other quantum computing approaches are closer to analog solutions (some being similar to simulated annealing technics). It has not been totally proven that they are more powerful than classical computing or other analog devices.

After each session, a debate was organized to compare the points of view and to discuss:

- What could be a time line for the emergence of the different technologies and which are the roadblocks and the milestones
- How to develop a value chain for these technologies in Europe.

The main technical findings of the workshop are integrated into the technical vision that has been developed through-out this task and which is presented in Chapter 6.

At the end of the workshop, considerable time was dedicated to a more global discussion about how Europe can progress toward a more integrated technology value chain and regain its position in the IT landscape. The recommendations coming from this discussion are presented in Section 5.4.

5.3 Science Fiction Success Stories

During this workshop, we also discussed with the participants the possibility to write Science Fiction Success Stories (SFSSs) about the technologies that were discussed. A science fiction success story is a story that has not happened yet but could be envisioned or dreamed of.

Such a story explains the result that can be achieved in the future with as many details as possible about the quantitative benefits for HPC systems and the timing and the team that is needed to obtain the result.

This is a good way to have a vision of the importance of the benefits, the potential steps from research to market and the approximate time frame for delivering the promises of the technology.

We had some volunteers to go through this exercise and to present the future in a entertaining way. The SFSSs generated after the workshop are included in Annex 9.3 and cover the following topics:

- Photonics interconnect
- Integration of heterogeneous chips for healthcare
- New silicon technologies for automotive system
- Security
- Integration of cryogenic systems
- Near memory computing.

5.4 Discussion and recommendations

It is worth mentioning that during this workshop, besides the SFSS exercise that was a success, we also tried to develop links between the different layers of the research value chain. The underlying idea is that usually the research is done in a rich and complex ecosystem with upstream teams that produce results and downstream teams that can use the results. Some exchanges of information between the people of these connected fields can help progress move faster.

So, it was proposed to the participants as a follow-up action to describe:

- Some elements that will help guide their research, to assess it potential for downstream users, to compare it to the state-of-the-art and other technological solutions. This can be:
  - Data sets
  - Benchmarks
  - Small application kernels
  - Communication patterns
- Symmetrically the challenges that they would like upstream teams to solve. If they are able to define what they are interested in getting, it can help other research teams to focus on their concerns.

This proposal was not a success as we did not get any such description. Despite, this poor result we still think that this idea involving the producer/consumer in the research value chain has to be developed to increase the efficiency of the research done in Europe.

During the discussion at the end of the workshop we got some recommendations from the participants. The main ideas are listed below:

- As we need some disruptions to progress toward future computing solutions, the research calls of the European Commission should be very open and not limited to very narrow subjects as in the current ICT calls.
- The interaction between the research teams and the companies providing photonic components has to be fostered.
- An open standard for chiplet would help people to invest in the approach. This standardization would help to minimize risks as several options may exist. Europe should be more present in initiatives like Open Compute which has started to work in the domain of hardware interface.

- Certain export legislation is a risk for people working on advanced computing solutions. The definitions of optical computing or neuromorphic computing are not detailed and so solutions in these domains could fall under these export control laws.
- The different ecosystems in touch with the EC (photonics, electronics, edge, AI, Big Data, Security, HPC) should work more closely together and think about a common call. The Horizon2020 approach with budgets marked for each ecosystem has to be surpassed for a more integrated way to work together.
- The definition of common goals that will align the work of several communities could foster some important achievements at the European level. Whether these are goals or integrations, design, prototypes, or specifications, they can help to produce concrete results for the benefit of everyone.
- It would be nice for non-partners to be able to follow projects. Some process to organize such "observer status" could be put in place.
- We have still to work on defining a common language to make sure that the different communities have a good understanding of each and other.

6 Main technical findings

This section presents the main technical findings about how the electronics and photonics technologies will impact the future HPC system. This information is a synthesis of the discussions undertaken with the electronics and photonics ecosystems and of the analysis of other technical documents.

6.1 Introduction

Today's HPC system architecture is dominated by the standard CPU+GPU[15] solution. This architecture has been effective at offering the performance increase requested by HPC users while challenging them to exploit the massive parallelism and heterogeneity offered by this solution.

We foresee little changes in the 2020-2023 time frame with the first exascale systems based on this approach. After sustaining the growth in the number of operations per watt, new solutions will have to be found as Moore's law will fade and Dennard's scaling gone. Progress can be made in three axis:

- New architectures
- New data representation schemes
- New technologies (compared to CMOS[16])

Most of the new approaches are a combination of the three (or at least of two of them) but it is important to understand that we have these three freedom degrees that can be played with:

- Switching from computing centric execution used by processors and GPU (akin to Von Neumann architecture) to the data centric paradigm to reduce the overhead introduced by the data movement;
- Changing what is called an operation by playing with operand precision or introducing multi-bits or analog coding or other ways of encoding information (e.g. Quantum),
- Introducing new materials that will deliver more efficient way (in terms of timing and/or energy) to store, switch and/or process information.

This gives a very broad set of options but only a few will emerge due to economic constraints, critical mass issues, industrialization aspects, legacy and usability problems. The following sections present some of the most promising paths.

6.2 Enhancements of current CMOS technologies

6.2.1 *CMOS scaling*

Even if we are close to the limit of CMOS scaling, there is still room for improvement in this domain. The leading foundries (TSMC, Intel, Samsung) are investigating for at least 2 more technology nodes compared to their current technologies. This could provide a way of putting roughly about 4 times more transistors in the same surface of silicon compared to today. However, this scaling comes with the cost of very expensive equipment (e.g. Extreme ultraviolet lithography - *EUV* or *EUVL*), and the power density of those technologies is still not known, perhaps limiting the number of devices active at the same moment on the die. It should

[15] Computing Processor Unit + Graphical Processor Unit
[16] Complementary Metal Oxide Semi-conductor

also be noted that even if the technology nodes are labelled with the same name (e.g. 7nm), all these nodes might not be equivalent:

Nominal node		28nm	22nm	20nm	18nm	16nm	14nm	12nm	10nm	7nm	5nm (ITRS)
Intel	Lg		24				20		16	~12nm	10
	Fin Pitch		60 FinFET				FinFET 42		FinFET 34	FinFET	12
	CPP		90				70		54		32
	M1		80				52		36		16
	SRAM		HD 0.092µm2				0.0588µm2		0.0312µm2	0.027µm2	
	Year Publication		VLSI 2012				IEDM 2014		IEDM 2017/ISSCC2018	IEDM 2016	
	Risk Prod		2011				2014		1Q18		
Samsung	Lg	32		25	25		30		~20	~16	
	Fin Pitch	BULK		BULK	FDSOI		48 FinFET		Single Fin 42	Dual thin EUV 27	
	CPP	114		86	86		78		68	54/57	
	M1	90		64	64		64		51	36	
	SRAM	0.152µm2		0.084µm2			0.064/0.08µm2		0.04µm2	HD 6T SRAM 0.026µm2	
	Year Publication	ICSIST 2011		VLSI 2012			JSSC 2014		ISSCC/VLSI 2017	VLSI 2017/ISSCC2017-2018	
	Risk Prod	2011		2013			4Q-2015		1Q2017	2H-18	
TSMC	Lg	30	30	30		33		*25*	~20	~16	
	Fin Pitch	BULK	BULK	BULK		FinFET 45		*FinFET 45*	FinFET	FinFET 4th	
	CPP	118	105	90		90/80		*90/80*	64	57	
	M1	90	80	64		64		*64*	42	40	
	SRAM	0.155µm2	0.155µm2			0.07µm2			0.03µm2	0.027µm2	
	Year Publication	VLSI 2012	VLSI 2012	VLSI 2014		IEDM 2013		*6Track*	VLSI 2016	IEDM 2016	
	Risk Prod	2011	2018	2013		4Q-2015		*3Q2016*	4Q2016	3Q-17	
GF	Lg		28				30				
	Fin Pitch		FDSOI				48 Fin FET				
	CPP		90				78				
	M1		78				67				
	SRAM		0.110µm2				0.110µm2				
	Year Publication		IEDM 2016				IEDM 2016			C.Reita, C.Fenouillet-Beranger - CEA-LETI - 2018	
	Risk Prod		2016				2H-2016				

Figure 10 : Nominal vs. actual node dimensions (Source : CEA Leti)

CMOS scaling is also related to the evolution of the structure of the transistor. After FDSOI[17] and FinFet[18], the structure of the transistor could be based on silicon nanowires.

In this domain, one of the challenges for Europe is that these technologies will only be developed by foreign players. As HPC system performance is highly dependent on density, the most advanced chips are mandatory and Europe need to have access to these technologies. This a strategic risk that needs to be monitored.

On the technology side, the challenge for the last CMOS technology nodes is to get a reduction of the energy by transistor while increasing the number of transistors. Even if we cannot expect as in the past a decrease of the energy by the same factor as the increase of transistors, progress in energy efficiency is mandatory to make the new CMOS technology nodes a success for HPC systems.

6.2.2 *2.5D/3D stacking*

2.5D/3D stacking provide a way of reducing the latency and energy and avoiding package bandwidth bottlenecks when we want several chips to communicate together. 2.5 D stacking is the concept of small dies (called chiplets) integrated on a common substrate (the interposer) that can be organic, passive silicon, active silicon or using photonic technologies. 3D stacking is the stacking of layers of integrated circuits on top of each other. It can be done either by wafer to wafer, chip to wafer stacking, or by monolithic 3D which allow a finer granularity (down to the level of transistors). HPC is already benefiting from this technology with the first HPC systems using high bandwidth memory (HBM or HMC[19]) and processor manufacturers (AMD, Fujitsu, Intel…) having already used 2.5D in the latest products. The boost in memory bandwidth is a great improvement for memory bound applications and a must for architectures with accelerators that require this kind of bandwidth to deliver their performance. 2.5D also allows for mixed chiplets with various technologies, and for example, with active interposers,

[17] Fully Depleted Silicon On Insulator
[18] fin field-effect transistor
[19] High Bandwidth Memory and Hybrid Memory Cube

having power conversions integrated "in the chip", providing globally a better energy efficiency.

It can also be a path for production of hybrid packages mixing chips of different architectures (see Section 6.3) or even chips manufactured with different technologies (see Section 6.5) Nevertheless, stacking "compute" chips with a higher heat dissipation than memory chips leads to thermal problems that today limit the number of chips that could be put in a package.

Europe is present in this field with excellent know-how in the three main RTOs[20] and some research installations to test new concepts. For an industrial development, as "High end" computing is more and more important for the automotive market (Advanced Driver Assistance Systems and self-driving), the European automotive industry might be a driver for having European actors in 2.5 D and the integration of complex systems on dies or interposers.

The European EPI (European Processor Initiative) plans to use2.5 D technology, in the same way as AMD, Intel and Fujitsu.

In this domain, one of the main technical challenges is the thermal dissipation of the stack. Innovations are needed to solve this problem before 3D stacking can scale and deliver all of its promises.

Another challenge is the set-up of European industrial options. Even if today the RTOs have research installations to design new concept chips, the industrial solutions are mainly coming from Asia. Having a European based option would be important from an independent and economic stand point because of the potential of this technology.

To foster the development of European chiplets that can be integrated via 2.5/3D stacking, it could also be strategic to work on integration standards. Having a commonly defined interface will provide an exploitation path for new initiatives in the area of accelerator design.

6.2.3 *Precision of operations*

The trend in the past has been to provide more and more precision for operations as HPC was focused on simulations where stability and convergence of algorithms depended on this feature. Today, for new applications coming mainly from neural networks, high precision is not mandatory (the "learning phase of Deep Neural Networks can be done in most case with 16-bits floating point operations) and switching to low precision can save space (i.e. transistors and data paths) and energy.

The availability of 16 and 8 bit operations in processors and accelerators is a trend that will allow to adapt precision to the needs of algorithms with an important saving in some cases.

On the reverse side, sometimes convergence of algorithms could benefit from high precision to reduce the number of iterations and to save time and energy. New coding schemes are possible and alternative to the IEEE 754 arithmetic, like UNUM[21]. Its implementation in processors could be more efficient then software emulation of this high precision scheme.

The trend, to have flexibility in the representation of data, is a challenge for software development. The choice of the representation could be made automatically with analysis of the pro and cons of the different options. This could lead to the automatic selection of libraries optimized for different data formats.

[20] Research and Technology Organizations
[21] Universal Number

This rethinking of data representations is also a precursor to a more disruptive approach with the use of analog options (see Section 6.6).

6.3 New architectures

Today standard processors and GPU accelerators are based on a Von Neumann architecture where a controlled execution applies operations onto data that are stored in registers (fed by caches, fed by memory). This architecture is very flexible but can be costly in terms of transistor, data paths and energy compared to what is just needed for an application. This implies a lot of moves and duplications of data, which is not efficient (bringing data from external memory is 3 orders of magnitude more energy demanding than a floating-point operation on those data). However, there is a research path what proposes architectures that will be more efficient for some classes of problems. Some of these new architectures can be implemented using standard CMOS technology or providing opportunities to introduce new technologies that will be more efficient than CMOS (see section 6.5).

Some concepts of new architectures are generic (see section on data flow or IMC[22] below) or target a specific class of algorithms (see section on neuromorphic, graph and simulated annealing below).

6.3.1 *Data flow*

In data flow architectures, data move between modules that perform the computation on the data. You do not have any program counter that controls the execution of the instructions as in a Von Neumann architecture. Deep Learning architecture (see section on neuromorphic architecture below) can be implemented as a specific dataflow architecture (the main operations are matrix-based). The investigation of dataflow architectures is linked to FPGA (Field Programmable Gate Array) as most of the ideas have not led to the tape out of specific circuits but have been tested and implemented with FPGA.

With the slowdown of standard processors performance increase, development of data flow architectures can provide an alternative to deliver this performance increase. The development of reconfigurable architectures (like the Intel CSA Configurable Spatial Accelerator) and progress toward flexible reconfigurable FPGA will be an asset for implementing data flow architectures.

6.3.2 *IMC/PIM (In Memory Computing; Processor In Memory)*

These architectures couple the storage with some computing capabilities. The idea is that bringing the computation to the memory will be cheaper in resources than moving data to the computing units. Most of the time this approach is mixed with a standard architecture to allow computation on several data.

The architecture is also related to the development of Non-Volatile Memory (see Section 6.4) and appealing as long as the cost of the in-memory computation is low.

[22] In Memory Computing

6.3.3 *Neuromorphic*

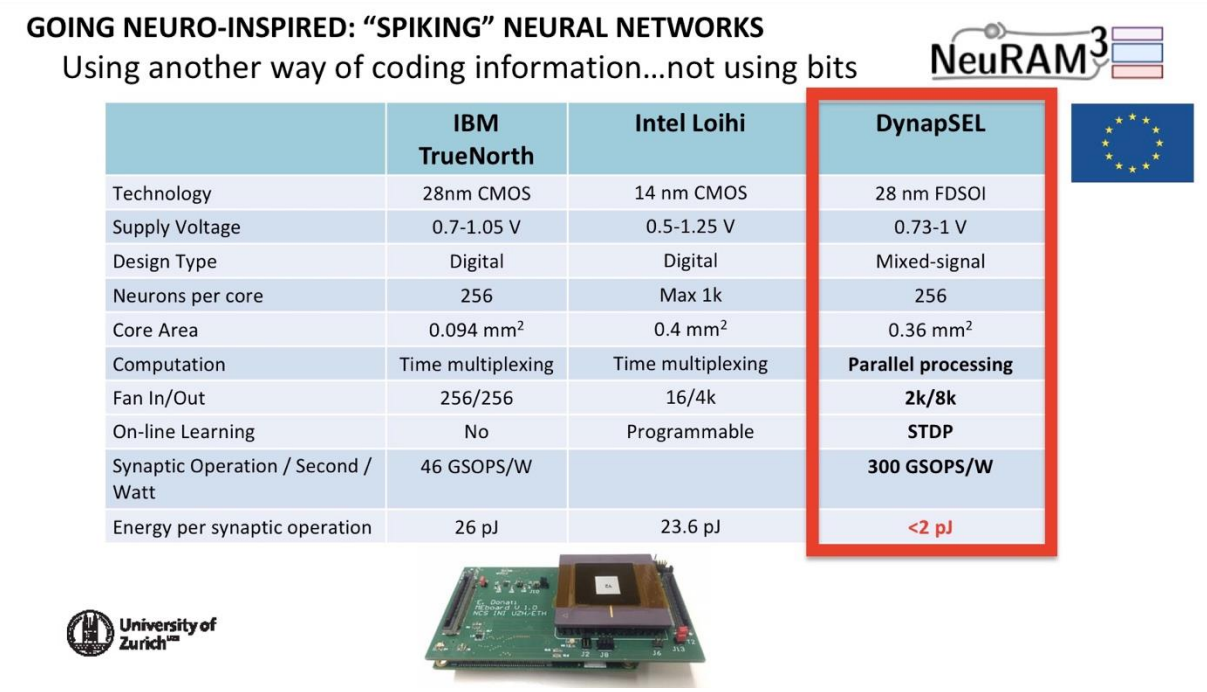
The development of AI, and especially applications using Deep Learning techniques, has led to a huge interest for neuromorphic architectures that are inspired by a theoretical model of a neuron. This architecture can be used for AI tasks but can also be viewed as a generic classification function or a function approximation.

As more and more applications (or a part of an application) are mapped to this paradigm, it is worth developing specific circuits that implement only the operations and data paths mandatory for this architecture. Several examples already exist such as the Google Tensor Processing Unit chip or Fujitsu Deep Learning Unit chip. These efforts have not exploited all the possible options and have not developed all the interesting features of the architecture, so research in this area is still valuable.

We can distinguish various kind of possibilities:

1. Using classical digital arithmetic, but designing more specialized architectures (examples: TPU and DLU)
2. Using another way of coding information, like "spikes" or their representation in AER coding (Address-Event Representation) (see fig 12).
3. Using « physics » to make computation (e.g. Ohms law for products and Kirchhoff law for summation; see Section 6.6 "Analog computing").

Of course, the approaches can be combined. Typically, most people call "neuromorphic" the approaches using option 2, because it is closer to the way the nervous system communicates. One important aspect is that this architecture is a good candidate to introduce an alternative to CMOS (see Section 6.5).



	IBM TrueNorth	Intel Loihi	DynapSEL
Technology	28nm CMOS	14 nm CMOS	28 nm FDSOI
Supply Voltage	0.7-1.05 V	0.5-1.25 V	0.73-1 V
Design Type	Digital	Digital	Mixed-signal
Neurons per core	256	Max 1k	256
Core Area	0.094 mm^2	0.4 mm^2	0.36 mm^2
Computation	Time multiplexing	Time multiplexing	**Parallel processing**
Fan In/Out	256/256	16/4k	**2k/8k**
On-line Learning	No	Programmable	**STDP**
Synaptic Operation / Second / Watt	46 GSOPS/W		**300 GSOPS/W**
Energy per synaptic operation	26 pJ	23.6 pJ	**<2 pJ**

Figure 11: NeuRAM3 approach

6.3.4 *Graph computing*

Graphs play an important role in data representation and in some AI or optimization problems. As standard processors have poor performance due the non-regular access to data, developing a specific architecture for this problem can be relevant.

To our knowledge there is no current initiative trying to implement this path (the Graphcore company has developed what they call a IPU (Intelligence Processor Unit) which is focused on AI applications and does not use a graph computing focused architecture). Nevertheless, this path could lead to an important acceleration for graph oriented problems.

6.3.5 *Simulated annealing*

Simulated annealing is a method for solving complex optimization problems. It can be implemented by software on classical Von Neumann processors but you can also design an ASIC that will significantly speed-up the computation by mapping directly the variables and their interactions and by providing a hardware based random generator.

This approach has been implemented by Fujitsu with its "Digital Annealing" processor. This project has developed a standard CMOS ASIC and a software stack to map the optimization problem to the circuit.

Other efforts use quantum devices (see section 6.7) to target the same class of problems (this approach requires cryogenic operation which is not the case with CMOS based developments).

6.4 Hybrid of CMOS and other technologies: NVMs, silicon photonics

6.4.1 *NVMs*

Different technologies are being developed to propose Non-Volatile Memory. Besides the existing NAND, resistive memory (memristor), phase change memory (PCM), metal oxide resistive random access memory (RRAM or ReRAM), conductive bridge random access memory (CBRAM) and Spin-transfer torque magnetic random access memory (STT-RAM) are interesting technologies. The developments in this domain have several impacts for HPC. The energy to retrieve data is decreased, the latency to read the data is reduced and the density can be increased (especially with solutions implementing multi-states storage for each cell).

NVMs also play a role in providing an easy implementation of the IMC/PIM architecture when compute elements can be associated as in Memristive Computing.

6.4.2 *Silicon photonics*

Silicon photonics can be used either to compute or to provide interconnect between computing elements.

Compute

The properties of light can be used to perform computation. For example, the interaction of light whose phase has been modulated according to inputs can produce operations over these inputs. This idea can be used to implement neuromorphic architecture where the main operation is a scalar product.

This approach is promising but several steps are still to be achieved: assessment of the value proposal in term of energy efficiency and industrialization path of the technology.

Another path is to use the massive parallelism of optics to perform complex operation (typically where the complexity is not a linear increase versus the size of the problem). An example is the system proposed by the start-up LightOn, integrated in an OVH cloud server (see Section 6.7 on analog computing).

Interconnect

Photonics is already used for long distance communication in HPC systems (electrons are easy to create and interface, they display attenuation with distance (Ohm's law), while photons are energy demanding for creation and interfacing but have low attenuation with distance). The technology is also appealing for rack level communication. But perhaps the most interesting aspect will be at the package level with the development of active interposer with embedded silicon photonics networks between chips or chiplets. The bandwidth and the energy efficiency can be increased compared to current CMOS solutions.

Again, for these applications, a silicon photonics industrialization path has to be developed. As seen in Section 4, European projects have proposed or are working on interesting technologies. A solution for transferring these results to market has to be found.

6.5 New solutions more efficient than CMOS

CMOS has been such an industrial success story that it has reduced the effort on alternative solutions to implement transistor or computing elements. With the end of CMOS progress more emphasis will be put on these other options even if it is still to be proven that they will be able to deliver more computing performance than CMOS.

6.5.1 *Superconducting*

With the use of superconducting material, the expectation, based on the zero resistivity of the interconnects, is that power consumption could be up to two orders of magnitude lower than that of classical CMOS based supercomputers.

Nevertheless, superconducting circuits have still to overcome several drawbacks like density, switching time, interfacing with external systems or noise reduction, become they can be seen as a potential solution for HPC. Most of the time the implementation uses Josephson junctions and so has the same disadvantages as analog computing.

6.5.2 *Magnetoelectric and spin-orbit MESO*

Researchers from Intel and the University of California, Berkeley have proposed a new category of logic and memory devices based on magnetoelectric and spin-orbit materials. These so-called "MESO" devices will be able to support five times the number of logic circuits in the same space as CMOS transistors. In these devices, logic and storage bits will be encoded by the spin state (up or down) of bismuth-iron-oxide, a multiferroic material. Compared to CMOS the switching energy is better (by a factor of 10 to 30), switching voltage is lower (by a factor of 5) and logic density is enhanced (by a factor of 5). In addition, its non-volatility enables ultralow standby power.

This path is promising even if the roadblocks for industrialization are still difficult to assess.

6.5.3 *Memristive devices*

Besides the uses of the resistive memory for NVM and analog neuromorphic architectures (see Sections 6.3 and 6.4), memristive devices can be interesting for implementing logic gates and computing. Even if the switching time may be slower than CMOS, they can provide a better energy efficiency. The integration of memory into logic allows to reprogram the logic, providing low power reconfigurable components and can reduce energy and area constraints in principle due to the possibility of computing and storing in the same device (computing in memory). Memristive devices can also be arranged in parallel networks to enable massively parallel computing.

Again for this technology, it is difficult to assess when it will be mature to propose a credible alternative for computing.

6.5.4 *Other materials*

Research has been done on new materials that could lead to new ways to compute including carbon nanotubes, graphene or diamond transistors. Nevertheless, at this stage of the research, it is too early to assess whether these options will propose a valuable solution for HPC systems.

6.6 Analog computing

Analog computing is when a physical (or chemical) process is used to perform a calculation. (An analog computer or analogue computer is a type of computer that uses the continuously changeable aspects of physical phenomena such as electrical, mechanical, or hydraulic quantities to model the problem being solved. – Wikipedia)

6.6.1 *Optical systems*

Optical systems can be used to compute some functions thank to the properties of light and optical devices like lenses. This approach is extremely energy efficient compared to traditional computers. This technology cannot suit every application but a number of algorithms as scalar products, convolution-like computations (e.g. FFT, derivatives and correlation pattern matching), are naturally compatible. Some demonstrations have been made by the EsCAPE project with the computation of spectral transforms by an optical system. The precision of the results can be a problem if the spectral transform is the input of a subsequent algorithm needing high resolution. Nevertheless, this method is well suited for correlation detection, sequence alignment testing or pattern matching applications.

Optical systems have also been used to implement reservoir computing. Reservoir computing and Liquid State Machines are models to solve classification problems and can be seen as “part” of neuromorphic architecture. Nevertheless, this approach is often coupled with research to implement this model with analog optical computing.

Optical computing is the more advanced field of analog computing with already two European start-up (Optalysys and LightOn) proposing products for accelerating recognition applications.

6.6.2 *Other options*

Other options are possible like using thermal or electrical systems to find solutions of some differential equation problems.

The mix of analog/digital computing inside a CMOS chip can be a way to provide more energy efficient solutions.

6.7 New computing paradigm: quantum computing

Quantum computing is a new paradigm where quantum properties are used to provide a system with computing capacity. Today research in this field can be split into 2 categories:

1. The "universal" quantum computers based on qubit and gates performing operation on these qubits. It uses two quantum properties, superposition (capacity to be at the same time in a superposition of 2 states) and entanglement (capacity to link the state of an element to the measure made on another element). From these properties, a mathematical model of a universal quantum computer has been developed. In this model a system of qubits can be put in a state that represents the superposition of all the values of the computed function (i.e. this system has in "parallel" computed the values of a function for all the 2^N inputs).
2. The quantum annealers, or quantum simulators, represented for example by the D-Wave machine, *use quantum physics to escape from local minima in optimization functions using quantum fluctuations.* This class of machines is limited to problems that can be modelized such as minimization of function, like the travelling salesman, flow optimization, molecular simulation etc. Other possibilities are to use known quantum processes (like (ultra)cold atoms) to modelize other quantum related phenomenon, like in chemistry.

Most efforts have targeted the first approach. Nevertheless, developing a physical system that behaves like the "universal" model is at the level of research and will need to solve hard problems such as the decoherence of the qubits, a reliable measurement system, error correction and the NxN interconnection between the qubits.

The EC flagship on quantum technologies addresses the topic with a good level of support compared to the risk/reward of this domain.

6.8 Transversal questions

6.8.1 *Integration within "classical" HPC systems*

Most of the technologies presented in the previous sections are more complementary to current HPC system technologies than a complete replacement solution. This leads to the question of the integration of these new options within the current framework. Some of them can be viewed as accelerators that will take care of some parts of the application while the rest will be computed by a classical system. Others such as NVM or IMC needs a reshaping of the data and computing hierarchy.

Integration of accelerators or innovative data architectures (see fig 13) leads to the question of how to choose either at compile time or at runtime between the several options for the execution

of a computation and storage of the data. The decision process will need the emergence of new compilation schemes or of new runtime software.

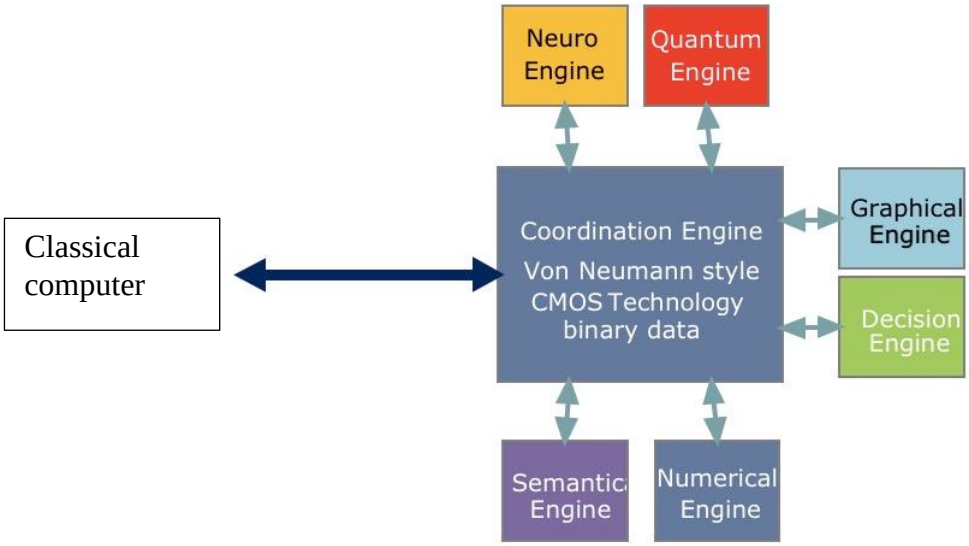


Figure 12: Potential future architecture of an HPC node with several accelerators

6.8.2 *Algorithmic and programming impact*

To take advantage of the new options, the development of applications needs to be reassessed and some new notions have to be explored:

- What is the minimum precision of the operation for either convergence or stability?
- How can the precision requirements and the tolerated errors be specified?
- Are there DSL[23] that can ease the exploitation of new architectures?
- How can we move from a monolithic application to a more modular one that could be mapped on different computing paradigms?

To integrate the advances the new technologies, the HPC community has to start to address these questions.

6.9 Summary

It is clear that continuing scaling the CMOS technology will not be the major factor of performance increase in the future, because we are reaching the limits of physics (and the cost to develop very advanced technology nodes becomes very high). This analysis shows that the research paths are multiple and diverse in continue increasing performance at affordable energy and cost. It is too ambitious to set priorities at this stage. A precise assessment of the roadblocks and a risk/reward analysis would require more interactions and significant work from the HPC, photonics and electronics research communities.

However, some of the technologies such as the 3D integration, dedicated and energy efficient accelerators such as neuromorphic architectures, silicon photonics interconnects and analog

[23] Domain Specific Language

systems have a maturity level which makes us believe that they will impact the HPC systems in a near future.

None of the technology paths will completely replace the current CMOS technology. The potential options are more complementary, where the CMOS-based system could continue to act as an orchestrator of a diversity of architectures and technologies. The integration of today's and tomorrow's computing/storage/networking paradigms will be one of the challenges, together with the development of efficient software stacks that will efficiently benefit from those emerging solutions while keeping the programming complexity to tractable levels.

7 Recommendations for a structured European effort

The interactions between the photonics, electronics and HPC communities have increased thanks to the efforts of the EXDCI-2 project. The stakeholders have started a dialogue on how to work together even if there are still some gaps in terms of language, expectations about maturity levels or working habits. These interactions have been productive and should be continue beyond the time frame of the EXDCI-2 project.

The way to implement a continuous interaction may be to have the domains ETPs (European Technology Platforms) or equivalents to take over the initiative started by EXDCI-2. ETP4HPC, Photonics21 and AENEAS can extend and amplify the dialogue to analyze the best way to work together and to assess the potential of different research paths. This dialog can be implemented for example via technical workshops such as the one organized by EXDCI-2 or other actions like webinars with mutual presentations.

R1 Establish a continuous dialog between photonics, electronics and HPC communities under the supervision of Photonics21, AENEAS and ETP4HPC.

In order to make this dialogue progress faster, it seems interesting to undertake small actions that will work at the interface of the research teams. Even if it has not been successful achieved by EXDCI-2, we believe that the development of a chain of producer/consumer research teams would be important to achieve progress. Small actions could be implemented involving producer consumer research teams to specify objectives, to propose common benchmarks, and to gather test data sets. These outputs could help to focus the upstream research and deliver a way to evaluate the progress and so the potential of different research paths.

These actions are not research projects per se but rather small efforts involving the two parts of a link in the research chain. Experts from the two sides can deliver the outputs in a short time frame (6 months seems enough) that will be used afterward for setting research project objectives and KPIs.

R2 Undertake small actions to specify research objectives, benchmarks and test data sets at the interface of two research communities.

After this study, we believe that in the next ten years the main path to increase HPC system performance will be to integrate at low level new architectures (based or not on new materials). We do not see the emergence of solutions that will radically replace the current technologies in this time frame. Rather the new options will be complementary to the current one. To be efficient in this integration, it is mandatory to do it at the lower level if we want to avoid costly data transfers and latencies. So, the integration of heterogeneous chips or analog technologies will be a must in the evolution of HPC.

Europe has interesting strengths in this field at the research level. If we want to have also a chance to be successful at the industry level, we need to establish a standard for integration of heterogeneous chips. This standard will allow the teams developing new accelerators to be sure that their work could be integrated in a complete solution. Perhaps the word standard is too strong as the maturity of the integration field is not yet achieved. Nevertheless, Europe needs to have experts to agree on common specifications (that could be reviewed and evolutive) that will enable the integration of heterogeneous chips. This effort could also be a way to push the European electronics industry to develop industrial solutions as the market could be more stable and broader with this initiative.

This recommendation is also one of the EXDCI-2's tasks on standards which has identified some strategic domains where a European action on standards is of the utmost importance.

R3 Work on European specifications for the integration of heterogeneous chips.

This EXDCI-2 initiative has showed that some European research ideas have good potential for future HPC systems. To unleash them, research projects coupling upstream technologies providers and the HPC community are mandatory. The results will not be for exascale systems even not for the first post exascale systems, but we need to start the research projects soon if we want Europe to be in the position to propose alternative solutions when the last CMOS technology processes will be reached.

The HPC systems and high performance edge devices can share some of the new technologies making this investment even more attractive for Europe. The science fiction success stories show that important benefits for science, industry and society depend on these new technologies.

The dialogue established within this task, also shows that the European stakeholders are motivated to work together and see some potential for new technologies. The end of Moore's law open the game and Europe can re-position itself in the IT market. If we want to construct this European technology value chain from basic technology up to HPC applications, we need to invest in a specific research programme.

EuroHPC has great ambitions to reposition Europe in HPC. To construct the technology value chain of the future HPC systems can be one of them.

R4 Launch a research program to develop new ideas coming from upstream technologies to provide new solutions for upcoming HPC systems.

With the implementation of these 4 recommendations, we believe Europe can regain leadership in HPC technologies.

8 Acknowledgments

We would like to thank the people who, through discussions with the authors, have provided valuable information and comments. Specifically, we thank Alun Foster, Bert de Colvenaer, Yves Gigase, Caroline Bedran, Patrick Cogez, Ursula Tober, Markus Wilkens, Sebastien Bigo, Jean-Luc Beylat, Joachim Pelka, Patrick Bressler, Rolf Aschenbrenner, Tolga Tekin, Anne Van den Bosch, Iuliana Radu, Peter Debacker, Yvain Thonnart, Pascal Vivet, Carlo Reita, Denis Dutoit, Said Derradji, Mathis Bode and Igor Carron.

We want also to express our congratulations to the authors of the Science Fiction Success Stories that have shared with us a bold scientific vision and have showed their creativity:

- Integration written by Maciej Wiatr (Global Foundries), Rolf Aschenbrenner (Fraunhofer IZM) and Michael Töpper (Fraunhofer IZM)
- New silicon technologies for automotive written by Rolf Aschenbrenner (Fraunhofer IZM) and Ferdinand Bell (NXP Semiconductors)
- Security written by Matthias Hiller (Fraunhofer AISEC), Maciej Wiatr (Global Foundries) and Ferdinand Bell (NXP Semiconductors)
- Integration of cryogenic systems written by Pekka Pursula (VTT) and Mikko Merimaa (VTT)
- Near memory computing written by Théo Ungerer (Augsburg University) and Denis Dutoit (CEA)

9 Annexes

The annexes are:

1. The November 2019 workshop presentation and agenda
2. November 2019 workshop participant list
3. The Science Fiction Success Stories
    a. Photonics interconnect written by JF Lavignon (Technology Strategy) and Marc Duranton (CEA)
    b. Integration written by Maciej Wiatr (Global Foundries), Rolf Aschenbrenner (Fraunhofer IZM) and Michael Töpper (Fraunhofer IZM)
    c. New silicon technologies for automotive written by Rolf Aschenbrenner (Fraunhofer IZM) and Ferdinand Bell (NXP Semiconductors)
    d. Security written by Matthias Hiller (Fraunhofer AISEC), Maciej Wiatr (Global Foundries) and Ferdinand Bell (NXP Semiconductors)
    e. Integration of cryogenic systems written by Pekka Pursula (VTT) and Mikko Merimaa (VTT)
    f. Near memory computing written by Théo Ungerer (Augsburg University) and Denis Dutoit (CEA)

9.1 Workshop documents

Presentation of workshop and agenda

Workshop: New technological paths for high performance chips targeting HPC and edge

Bedford Hotel, Brussels, November 5-6th 2019

HPC systems will reach the exaflops mark in the next years thanks to the latest CMOS technology nodes and the use of vector/graphic accelerators. After this achievement, it will be more and more difficult through this technology path to answer the demand for increasing performance in HPC systems. In parallel, we see a need for more and more powerful chips at the edge (autonomous systems and sensors with high processing capacities).

To fulfil these demands for extreme performance chips, new technological approaches have to be developed.

The objective of this workshop is to address this challenge and to discuss:

- What are the most relevant technologies and/or new architectures for future HPC/edge systems;
- How to accelerate the uptake of these technologies/architectures;
- How Europe can develop a value chain for these new approaches and get a strong position.

The workshop will gather European experts from the HPC, nanoelectronics and photonics ecosystems coming from research organizations and industry. It will be organized with short presentations and discussion time.

The aims at the end of the workshop are to:

- Set a list of promising technologies/architectures relevant for future HPC/edge systems and meaningful to develop in Europe;
- Be in position to write short "science fiction success story[24]" for some of these high potential technologies.

The outcomes of the workshop are expected to serve as arguments to get the adequate research calls in Horizon Europe and in EuroHPC JU programs.

Logistic information

The workshop is on invitation only. It is organized by the EXDCI-2 project[25]. It is free of charge[26] for the invitees.

The workshop is organized from Tuesday November 5th with a welcome from 12h30 and a start at 2pm until Wednesday November 6th at 4pm.

The workshop is located in Brussels at hotel Bedford Hotel[27] (Rue Du Midi 135, Brussels, 1000, Belgium); Each participant will have a room reserved in the hotel.

[24] See next page
[25] European Extreme Data and Computing Initiative a coordination and support action supported by the Horizon 2020 programme under the grant agreement n°800957
[26] The project will pay for accommodations and meals. The project could also provide travel support in specific cases (please ask if needed).
[27] http://www.bedfordhotelcongresscentre.com/

Recommendation to speakers

The objective of your presentation is not only to explain the technology/architecture that you foresee as useful for future HPC/edge systems. An analysis on the maturity and the way to develop this approach in Europe is also of interest for this workshop. So, we recommend that your presentation addresses the following topics:

- Technology/architecture presentation
- Why it is relevant for future HPC/edge systems
- What could be a time line for the emergence of this technology and which are the roadblocks and the milestones
- How to develop in Europe a value chain for this technology

Science Fiction success story (SFSS)

We call science fiction success story, a story that has not yet happened but could be envisioned or dreamed of.

Such a story explains the result that can be achieved in the future with as much as possible details about the quantitative benefits for HPC systems, the timing and the teaming that has been needed to obtain the result.

We expect at the end of the workshop to identify for the most interesting technologies a leader for different “SFSS” (and possibly co-contributor(s)) that will provide these documents.

The SFSS are expected to be short document (around 1 page, at most 2 pages).

D2Report on trends and potential synergies between electronics, photonics and HPC

Agenda

	start	end	duration	content	speaker	organisation
DAY 1 Tuesday November 5						
	12:30	14:00	01:30	Welcome lunch - registration		
	First session setting the scene					
	14:00	14:10	00:10	introduction of the workshop	JF Lavignon	EXDCI-2
	14:10	14:40	00:30	State of the art and challenge in HPC systems	Saïd Derradji	Atos
	14:40	15:00	00:20	HPC applications challenges	Mathis Bode	University of Aachen
	15:00	15:10	00:10	First finding of EXDCI-2	Marc Duranton JF Lavignon	EXDCI-2
	15:10	15:20	00:10	Questions		
	Second thematic session silicon process					
	15:20	15:40	00:20	Future computing devices	Francois Andrieu	CEA
	15:40	16:00	00:20	3D and packaging	Severine Cheramy	CEA
	16:00	16:10	00:10	Discussion on the thematic		
	16:10	16:30	00:20	Break		
	Third thematic session AI oriented and neuromorphic architecture					
	16:30	16:50	00:20	Hardware Accelerated Spiking Neural Networks for Low Power Signal Processing	Johannes Leugering	Fraunhofer
	16:50	17:10	00:20	Analog accelerators for neural networks	Bert Offrein	IBM Zurich
	17:10	17:30	00:20	Application of a photonic Si platform for neuromorphic applications	Benoit Charbonnier	CEA
	17:30	17:50	00:20	Embedded AI: From Standard Hardware to Neuromorphic Elements	Rainer Kokozinski	Fraunhofer IMS
	17:50	18:10	00:20	Neuromorphic research at IMEC	Peter Debacker	IMEC
	18:10	18:30	00:20	Discussion on the thematic		
	Forth session thematic new technologies and PIM first talk					
	18:30	18:50	00:20	Memristive technologies	Dietmar Fey	University of Erlangen
	18:50	18:50	00:00	End of the working day		
	18:50	19:30	00:40	Free time		
	19:30	21:00	01:30	Diner together		
DAY 2 Wednesday November 6						
	Forth session thematic new technologies and PIM continuation					
	08:45	09:05	00:20	PIM	Denis Dutoit	CEA
	09:05	09:25	00:20	Non volatile memories & PIM	Thomas Kämpfe	Fraunhofer FMD (IPMS)
	09:25	09:45	00:20	InP Photonic integrated circuits	Karl-Otto Velthaus	Fraunhofer FMD (HHI)
	09:45	10:10	00:25	Discussion on the session thematics		
	10:10	10:40	00:30	Break		
	Fith session thematic silicon photonics and analog computing					
	10:40	11:00	00:20	Photonics	Tolga Tekin	Fraunhofer
	11:00	11:20	00:20	Photonics interconnnect for interposer	Carlo Reita	CEA
	11:20	11:40	00:20	Photonics at IHP	Lars Zimmermann	FMD (IHP)
	11:40	12:00	00:20	Photonics at IMEC	Philippe Absil	IMEC
	12:00	12:20	00:20	Bringing Light to Artificial Intelligence	Igor Carron	LightON
	12:20	12:50	00:30	Discussion on the thematics		
	12:50	14:00	01:10	Lunch		
	Last session tranversal challenges, wrap up and next steps					
	14:00	14:20	00:20	System integration Technologies for High Performance Computing	Rolf Aschenbrenner	Fraunhofer FMD (IZM)
	14:20	16:00	01:40	Discussion: integration challenges, interaction of ecosystem, SFSC, next steps		
	16:00	16:00	00:00	End of the workshop		
	16:00	16:20	00:20	Tea-coffee break		

9.2 November 2019 workshop participant list

first name	last name	organisation	email
Philippe	Absil	IMEC	Philippe.Absil@imec.be
Francois	Andrieu	CEA	francois.andrieu@cea.fr
Rolf	Aschenbrenner	Fraunhofer	rolf.aschenbrenner@izm.fraunhofer.de
Ferdinand	Bell	NXP	ferdinand.bell@nxp.com
Mathis	Bode	Aachen University	m.bode@itv.rwth-aachen.de
Patrick	Bressler	Fraunhofer	Patrick.Bressler@mikroelektronik.fraunhofer.de
Igor	Carron	Light On	igor@lighton.ai
Marco	Ceccarelli	European Commission	Marco.CECCARELLI@ec.europa.eu
Benoit	Charbonnier	CEA	benoit.charbonnier@cea.fr
Severine	Cheramy	CEA	severine.cheramy@cea.fr
Marcello	Coppola	ST	marcello.coppola@st.com
Veronique	de Halleux	IMEC	Veronique.deHalleux@imec.be
Peter	Debacker	IMEC	Peter.Debacker@imec.be
Saïd	Derradji	Atos	said.derradji@atos.net
Marc	Duranton	CEA	marc.duranton@cea.fr
Denis	Dutoit	CEA	denis.dutoit@cea.fr
Dietmar	Fey	University of Erlanger	dietmar.fey@fau.de
Thomas	Geenen	ASML for ECSEL+Aeneas	thomas.geenen@asml.com
Emilie	Germetz	Neovia Innovation	emilie.germetz@neovia-innovation.eu
Matthias	Hiller	Fraunhofer	matthias.hiller@aisec.fraunhofer.de
Knut	Hufeld	Infineon	Knut.Hufeld@infineon.com
Thomas	Kämpfe	Fraunhofer	thomas.kaempfe@ipms.fraunhofer.de
Rainer	Kokozinski	Fraunhofer	Rainer.Kokozinski@ims.fraunhofer.de
Jean-Francois	Lavignon	TS-JFL	jf@ts-jfl.net
Johannes	Leugering	Fraunhofer	johannes.leugering@iis.fraunhofer.de
Mikko	Merimaa	VTT	Mikko.Merimaa@vtt.fi
Bert	Offrein	IBM Zurich	OFB@zurich.ibm.com
Carlo	Reita	CEA	carlo.reita@cea.fr
Werner	Steinhögl	European Commission	Werner.STEINHOEGL@ec.europa.eu
Jeorg	Stephan	Fraunhofer	joerg.stephan@mikroelektronik.fraunhofer.de
Tolga	Tekin	Fraunhofer	Tolga.Tekin@izm.fraunhofer.de
Theo	Ungerer	Eurolab4HPC	ungerer@informatik.uni-augsburg.de
Karl-Otto	Velthaus	Fraunhofer	karl-otto.velthaus@hhi.fraunhofer.de
Ursula	Vober	Photonics21	tober@vdi.de
Maciej	Wiatr	Global foundries	Maciej.Wiatr@globalfoundries.com
Lars	Zimmermann	IHP	lzimmermann@ihp-microelectronics.com

9.3 Science fiction success stories

First all photonics communication supercomputer

This week of November 2025, the new European supercomputer named “Light Speed” has started operation in the European HPC computing centre located at Babobruge. This HPC system should be ranked 1st in the next release of the Top500 that will be issued in 2 weeks.

Beside given Europe the first spot for HPC performance, this supercomputer is also the first one to use photonics communication at all the levels of its architecture: at processor and System on Chip (SoC) level with photonic interposer technology, at board level for inter-processor communication, at rack level for inter-node communication and at data-centre level for inter-rack and storage communication. Thanks to advances in silicon photonics, the different levels use photons instead of electrons to communicate the data with increased bandwidth, reduced latency and energy saving. For example, at package (SoC) level, communication between processor and high-performance memory or between processor and accelerator can reach the bandwidth of 1TB/s with an energy saving of 60% compared to using electron.

This innovation has been prepared by research started in the Horizon2020 with projects like ICT-streams and Teraboard. After an additional research stage, a spin-off BrightSiCom has been created in 2021 and has industrialized the different technology bricks developed by the research projects. BrightSiCom will also soon propose its technologies for the chips that help you to control your car.

Thanks to this all photonics communication, “LightSpeed” will perform better on key applications with increase of performance from 30% for latency bound applications to more than 100% for memory bound applications. Compared to a supercomputer using electrons instead of photons for its communications, “LightSpeed” will also save more than 1 MW of energy. This represents more than 5M€ of saving during the operation of the supercomputer.

More importantly, “LightSpeed” will provide a new level of performance enabling progress for European citizens. For example, better previsions of weather hazards as flooding or high wind effects will help to manage climatic risks; more data assimilation will help to personalize clinical treatments.

The future of internet based human beings

A normal day in 2035:

It has been, again, a short night… Despite the fact that I slept only 4 hours I'm feeling extremely vital and relaxed. This little tiny chip under my skin is doing a really good job. Since I got the implant some months ago, I sleep extremely well and can concentrate much better at work and at home. Professional business and daily life have become so much easier these days.

Starting in 2020s European Initiatives for High-Performance Computing (HPC), Artificial Intelligence (AI) and trusted electronics enabled an incredible progress of solutions for medical applications, data computing and general business efficiency. All this in agreement with environmental aspects and humanity has led to the comfort I'm using today.

The chip under my skin consists of several specialized AI accelerators integrated together in one package stacked in third dimension with a high-performance, low-power compute and communication engine. Bio-sensors on the same chip feed the AI and computing architecture with all the necessary data to detect my activity, vital functions and body constitution. AI accelerators and the HPC engine are extremely computationally efficient. My body temperature and movement are fully sufficient to supply required energy to operate the chip without any battery – throughout my entire life. Calculation of raw data and the majority of the functions happens at the edge – under my skin. Only very complicated algorithms for extraordinary tasks are off-loaded to a data center.

The chip is connected to a powerful trusted cloud server, somewhere nearby, on a solid European soil. The European initiatives in the 2020s solved the major problem of that time, namely huge energy consumption of data centers. Already over 15 years ago it was crystal clear that we cannot withstand the exponential grow in amount of data to be processed with the conventional architecture of data centers, even if some experts were thinking Moore's law was going to continue forever. Projects initiated at that time made it obsolete to build data centers in Alaska, Island or in the Antarctic due to cooling problems. Instead trusted server farms have been established in standard environment in Europe using components from trusted sources and solar energy. I think I would have not decided to implant the chip if my data had to travel over the borders of continents to be processed in an unsecured environment.

The best thing though is that the chip can stimulate my body functions. It can cause me to fall asleep quickly and wakes me up at the right moment of my sleep phase to guarantee maximum possible rest and recovery. Based on the weather forecast and traffic situation from the cloud it happens earlier or later depending on my business calendar for the earliest task of the coming day. Embedded AI components learn analyzing my behavior to realize at the right moment when I'm ready to leave. My autonomous car, also connected to the cloud server and controlled by my chip, arrives in front of the house just at the moment I'm stepping out of the door.

This is why I'm so relaxed despite the "short night" and I'm looking forward to a busy but well organized working day as well as a nice evening with my family afterwards.

Future Car 2025

Entertainment, Meeting Area and Safety for all Road Users

Authors: Rolf Aschenbrenner (Fraunhofer IZM) and Ferdinand Bell (NXP Semiconductors) on a success shared with automotive application

Summer 2019. The "High-Tech" family is driving to the beach. The family's younger son is asking his older sister when they will finally have a car that drives on its own, so the family can play cards while on the road. His sister's smartphone battery is running low, so she turns her attention to him and tells him story about the Future Car of 2025.

"Listen up, little brother. Imagine a self-driving car, equipped with an array of high-performance sensors all around the car. By 2025, we'll be able to really talk to each other when we go travelling. Our parents can turn their seats around, and the car will drive on its own. We only have to tell the car where we want to go before we hit the road, and the car will look for, find, and take the best and safest route to our holiday destination.

But rotating seats so we can play cards or enjoy meals together in the car are not the only great invention you'll be able to discover in our Future Car. You'll also find displays with wireless gaming or movie streaming, even while the car is driving by itself. The window will be even more high-tech, because we´ll have Augmented Reality technology fitted into them. Look out and see the landscape, and you'll get some facts and figures about the history and the people living there. Or you can take a video call from our grandparents, with the caller's image shared on our car windows, just as if they were screens. It´s like a big TV studio. Grandma and grandpa will only be visible from us on the inside, but our entire family can see and speak to them.

And before you say it, Dad, you really don't need to worry about safety. There will be a lot of technologies in the future car that will make the roads safer for all people – us inside the car and the people outside."

The girl´s voice takes on a little robot sound, as she answers the most import question of her little brother:

But how can that be possible?

By bringing high-performance safe and secure computing to the Edge. Researchers and engineers will design systems with functional safety, vehicle safety, device reliability, and cyber security to enable smart vehicles. Radar, LIDAR, cameras, and V2X[28] sensors which will communicate with the environment and track the occupants' biometrics for mood and health ensuring a safe and secure trip. Miniaturization of all these electronic components will be based on hetero-integration, using panel-level embedding technology and Si-3D[29] integration. Sensor-based system solutions will be provided with transceivers (based on SiGe[30], CMOS or other advanced technology), microprocessors, power management, network interfaces, and software development kits for algorithm development, including machine learning and artificial intelligence. Combined with functional safety and cyber security, we can establish a fully autonomous safe and secure ecosystem for self-driving cars. Zero accidents, zero emissions, zero wasted time!

And the girl´s voice returns to its normal tone, and she whispers "I am dreaming about very miniaturized electronic inside our future car – smaller than a wisp of my hair, and with much more technology in it." At the end, she opens the window and blows a dandelion that she got as a gift from a researcher at their last stop. Her little brother is smiling and dreaming of 2025, when he will playing cards while driving and eating sandwiches with his family. All the while, their parents are dreaming

[28] Vehicle to everything
[29] Silicon technology using the 3 spatial dimensions
[30] Silicon-Germanium semi-conductor

about their future car as well and looking forward to how their work and meetings can be more efficient and leisure time more enjoyable in 2025.

Self-driving car in 2025 – better senses than the human driver. © NXP Semiconductors.

EuroHPCs remain online after global SYNAPSE attack

Earlier this month, on April 2nd 2029, a global attack corrupted more than 70% of the worldwide HPC capability. According to experts, the SYNAPSE attack exploited the user separation, escalated privileges and is currently running an unknown AI algorithm across the data of the various users. The Edge architecture and distributed HPC prevent a controlled shutdown of the architecture, so that experts are still searching for a solution. Again, the data centers based on the EuroHPC Architecture are still online, so that the investment in cybersecurity was proven to pay off more and more, especially over the turbulent last two years.

The EuroHPC architecture is rooted initially in the activities that were started in 2021 in the Horizon Europe Programme. Back then, first approaches for hardware trust anchors for HPC were developed and installed that allow to bind the executed operations to specific hardware. Also, secure enclaves for HPC were designed to follow the trend of multi-user HPC on the same hardware. The previous years have shown that this architecture is more resilient to attacks, compared to the traditional open architectures and especially reduced the scalability of attacks.

The seamless Edge integration technology gave new access to unused computation resources at the Edge for HPC and thus reduced the total energy consumption of HPC by 20%. It also enabled new business models to provide Edge-devices as a service, leading to some of today’s leading service providers such as GreenCompute. This led to the recent paradigm of tactile HPC, facilitating secure local and low-latency HPC computation.

First cryogenic co-processor in a supercomputer

Helsinki, Friday 11 December 2026

Early this month, a new ingredient has been added to the toolkit of high performance computing. A fibre connection has enabled a BullSequana XH2000 to communicate with a cryogenic co-processor operated in a special cryostat. The supercomputer is hosted in Kajaani, Finland, in the premises of CSC, the Finnish supercomputing centre. The unconventional optical link is the result of a recent European effort to develop cryogenic ultra-fast electro-optical interfaces with ultra-low power dissipation. VTT Technical Research Centre of Finland has lead the consortium including several European universities, RTOs and companies. The cryogenic operating environment implied at the same time challenges and unique opportunities, which led to novel integrated nanophotonic solutions with high speed and unprecedentedly low power consumption, in the order of 100 aJ[31]/bit and 500 aJ/bit in the down link and up link respectively. The tested cryogenic co-processor is based on single-flux-quantum (SFQ) technology, an ultra-fast superconducting technology that holds the promise to replace the power hungry GPU accelerators in supercomputers, with more than 1,000 times higher power efficiency. Similar photonic links can also be adapted to communicate efficiently with ultra-high speed cryogenic CMOS electronics, or with quantum co-processors that can be used as accelerators for specific tasks. Thanks to these novel cryogenic optical interfaces, it will therefore be possible to efficiently integrate cryogenic technologies in supercomputers, enabling novel architectures with unprecedented power efficiency and performance.

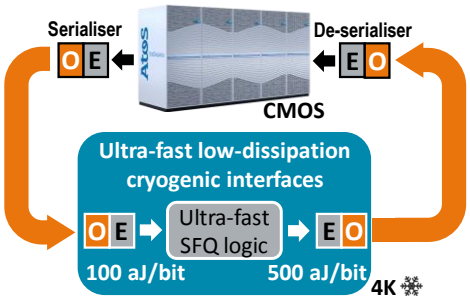


Figure 13 Schematic of an optical link between a supercomputer and a

The developed cryogenic datalink will enable tight integration between classical supercomputers and the 10,000-qubit quantum processor being developed as a successor of the European 100-qubit processor, launched in 2021 by the OpenSuperQ project under the European Quantum Flagship. First applications of these quantum-accelerated supercomputers are expected in molecular simulations for drug development, promising more effective and less expensive medicines in the future. Quantum co-processors are also being developed for solving big data analysis problems in healthcare using machine learning and artificial intelligence.

The developed technology also has major impact in other applications. In order to help companies and research labs leveraging the new technology in different fields, VTT has started offering multi-project-run services of superconducting devices, including also their combination with silicon photonics. The research has enabled companies to offer commercial cryogenic photonic integrated circuits as well as room-temperature silicon photonics solutions exploiting efficient modulators and detectors. The companies involved are also exploiting cryogenic electronics to offer a wider product portfolio based on single-photon detectors, including single-photon cameras.

The novel cryogenic interfaces and co-processors add a new all-European hardware to supercomputers, which also enables integrating European cryogenic processors in their architectures, both classical and quantum. This represent a major step in ensuring a leading position for Europe in supercomputing technologies.

[Text generated by VTT AI system]

[31] aJ atto joule or 10^{-18} joule

Near- and In-Memory-Based Accelerator for Supercomputers

In 2025 a new accelerator to the European supercomputer named “Light Speed” has started operation in the European HPC computing centre located at Babobruge. This HPC system should be ranked 1st in the next release of the Top500 that will be issued in 2 weeks.

Data movement has been identified already end of last decade as a main source of inefficiency in computing systems. High-Bandwidth Memory (HBM) was a first but still conventional step to more efficiency of CPU-centric computing, but in consequence the memory should be in the center of our notion of computing and lastly leading to processing in memory.

The new *Near-memory computing (Near-Memory Processing, NMP)* accelerator is characterized by processing in proximity of memory to minimize data transfer costs. Compute logic, e.g. small cores, is physically placed close to the memory chips in order to carry out processing steps, like e.g. stencil operations, or vector operations on bulk of data. Near-memory computing is a node-based hardware accelerator in combination with memory. The new near-memory computing accelerator replaces the memory controller of a die-stacked HBM to be able to perform logic operations on the row buffer. The new memory controller performs semantically richer operations than load and store, respectively cache line replacements. Processing by near-memory computing reduces energy costs by its vicinity to the main memory and goes along with a reduction of the amount of data to be transferred to the processor.

This new *NMP* accelerator is a first step, *In-memory computing (In-Memory Processing, IMP)* expected for 2030 will go a step further with memory cell used not only as storage cell but becoming an integral part of the processing step. This will further reduce the energy consumption and the area requirement in comparison to near-memory computing. This technology pushes the classical Von-Neumann architecture to a new computing model that requires significant changes in software and compiler. Therefore, *IMP* is considered as a more long-term solution.

Near- and future in-memory computing profit and will profit from several existing and upcoming semiconductor technologies:

- die stacking and 3D chip integration for integration at millimeter scale of logic with memory for Near-memory computing,
- monolithic 3D for integration at micrometer scale of transistors for processing into memory cell,
- memristors as memory technology do not need a refresh which is required by DRAM in HBM, and are able to perform operations (memristive computing),

- photonics for ultra-high bandwidth and low power communications required to scale-out In-memory compute nodes.

In 2030, the second generation of “Light Speed” *IMP* accelerator will benefit from monolithic 3D, memristors and photonic technologies to achieve extreme level of computing.

Applications for near- and in-memory computing accelerators could be data intensive applications often categorized as “Big Data” workloads. Such accelerators are especially fitting for data analytics, as they provide immense bandwidth to memory-resident data and dramatically reduce data movement, the main source of energy consumption. Analytic engines for business intelligence are increasingly memory resident to minimize query response time.

Near- and in-memory computing with memristors will influence the concept of Storage-class Memory, i.e., a non-volatile memory technology in between memory and storage, which may enable new data access modes and protocols that are neither “memory” nor “storage”. In-memory computing will also influence strongly edge computing approaches, in which new architectures have to be found that are characterized by processing data directly at sensors where the data is captured to reduce as described above the amount of data that has to be transferred to more-coarse grained cores for post-processing.

Assuming that near- and in-memory computing technologies will be mature, we need to change algorithms and data structures to fit the new design and thus allow memory-heavy “in-memory” computing algorithms to achieve significantly better performance. We may need to replace the notion of general purpose computing with clusters of specialized compute solution. Accelerators will be “application class” based, e.g. for deep learning (such as Google’s TPU and Fujitsu’s DLU), molecular dynamics, or other important domains.