THE KEYS OF THE DIGITAL WORLD

PUBLISHED BY **TERATEC**
ON THE OCCASION OF ITS
20th ANNIVERSARY

THE KEYS OF DIGITAL WORLD

PUBLISHED BY **TERATEC**
ON THE OCCASION OF ITS
20th ANNIVERSARY

Teratec

CONTENTS

INTRODUCTION

PERSPECTIVES

CONTENTS

TECHNOLOGIES

CONTENTS

USES

CONTENTS

SOCIETAL PERSPECTIVES

20 YEARS AFTER

Didier Besnard
Président, Teratec

Teratec is 20 years old. Thanks to the involvement of all its members, this Association has significantly contributed to France's sovereignty in high-performance simulation, both in terms of technology provision and industrial applications. Teratec is at the heart of a whole ecosystem gathering technology providers, HPC technology users, and research institutions. This has only been possible through strong coordination between its industrial and academic members: science takes place over time, while applications and innovation happen quickly. Teratec helps bring them together.

A hub for exchange and generation of initiatives and projects at the national and European level, Teratec is unique in its kind, capitalizing on science and innovation embodied by its members. Teratec's ecosystem, that of high-performance simulation and big data, including quantum computing and AI, is rapidly evolving. This is driven by enhanced parallelism between scientific progress and the emergence of industrial and societal applications: the ease for researchers to launch start-ups as well as national and European initiatives to support innovation are key drivers. In addition, the scientific dynamics of the various branches of high-performance simulation also leads to more powerful tools emerging. These are both challenges and opportunities, in science and technology on the one hand, and in terms of contributions towards companies within the ecosystem on the other.

IN A RAPIDLY EVOLVING ECOSYSTEM, TERATEC SEES ITS MISSION STRENGTHENED

Along these observations, Teratec can and must continue to be part of these scientific dynamics at work:

- Many of the industrial applications contemplated by Teratec members are critical applications. The accuracy and reliability of calculations must therefore be ensured. While the hybridization of physical simulation and AI offers opportunities to accelerate both modeling and numerical simulations, it must be possible to ensure compliance with physics laws while measuring the uncertainties of calculations. Furthermore, the controlled use of AI to process massive amounts of data obtained in all scientific and/or industrial fields complements the fundamental triptych of Simulation – modeling, numerical simulation, and experimental validation. The applications are limitless: data modeling, simulators and synthetic data, robotics... This is a strong area of focus for Teratec.

- Quantum technologies are also evolving at an ever-increasing pace. The Teratec initiative on quantum computing (TQCI), launched a few years ago, fits naturally into the current strong momentum. The expansion of participants, the integration of more start-ups involved in this field, and the establishment of the Maisons du Quantique by the SGPI are all actions that should be amplified.

- AI is a powerful driver of productivity and assistance, for users being less specialists of digital technology. It will enable easier access to high-performance simulation if simple tools can be developed, based on specialized AI. This is also the case for researchers who are not specialists using AI technologies in their own field of applications. There is therefore a challenge to overcome both in terms of tools and associated training.

OPERATIONALLY, TERATEC IS GOING TO AMPLIFY ITS CONTRIBUTION TO THE SIMULATION ECOSYSTEM

What are the current challenges?

• Improving bridges between science and innovation within the Association by strengthening the links between industrial and academic members.

Actions are being undertaken in this area with current members, but expansion to include new academic members remains to be done.

Working with start-ups is also in line with this objective. Many of them result from research currently carried out in research organizations. Support for start-ups must be strengthened, particularly those developing new technologies for intensive computing, quantum computing, or those that may specialize in AI.

• Teratec footprint must also be expanded by stronger links with science and technology players and institutional bodies at the European level. The European scale is indeed a guarantee of greater scope and effectiveness. From a technical point of view, promising results have already been achieved with European projects.

We must capitalize on results already gained by disseminating and reproducing the tools and/or training actions already developed, using them for the benefit of members of the Association.

• Through its members, Teratec is at the heart of the high-performance simulation ecosystem. However, it still needs to fit in more effectively into regional and national research and innovation initiatives. In that regard, the Association will have to launch or contribute to new sovereignty initiatives by its members.

• Finally, with such issues strongly arguing for the development of training and support axis, partnerships with diligent scientific and technical actors should be settled. For example, this involves promoting access to high-performance simulation technologies through simple, user-friendly AI-based tools within industrial sectors, including SMEs and SMIs. It also pertains to facilitating the use of AI-based development tools for non-specialists.

TERATEC COMBINED TO FUTURE

Through its real scientific and technical activities, strengthening interactions between members to address their needs, the expansion of its regional, national, and European footprint along with its quality training activities, Teratec will reinforce its unique nature and meet the demands of a rapidly evolving ecosystem.

Twenty years on, Teratec remains in tune with the developments underway in its field. The prospects are exciting. Challenges remain. The effort must go on. ■

TERATEC CELEBRATES 20 YEARS

Christian Saguez
Co-Founder and Honorary President, Teratec

Since Teratec inception in 2005, high-performance computing (HPC) has become a major strategic element. In the early 2000s, France no longer had control over HPC hardware or software technologies. The CEA's Simulation program revealed such a vulnerability, jeopardizing our sovereignty, scientific expertise, and economic development. Digital technology was to become the key to designing, innovating, producing sustainably, and exploiting innovations. Few leaders noticeably politicians, were aware of this and a fierce battle was to overcome to convince them, while a well-known consulting firm denied its importance.

Upon CEA initiative, a small group of industrialists created Teratec, as non-profit association under the French law of 1901, launching a dynamics that put France back among the major digital nations and coming to establish a unique model in Europe.

"It is not because things are difficult that we do not dare, it is because we do not dare that things are difficult," said Seneca. So, we dared, despite the many obstacles and opposition, and today Teratec is celebrating its 20th anniversary.

IGNITING MOMENTUM BY BRINGING TECHNOLOGY PROVIDERS, USERS, AND RESEARCH CENTERS TOGETHER

From the outset, key principles were established:

- Uniting technology providers, industrial users, and research centers in large collaborative projects.
- Developing co-design philosophy and bringing together hardware, software and architecture within multidisciplinary teams.
- Expanding uses in all economic sectors and disseminating them to companies of all sizes, including SMEs.

Use of HPC derived from mathematical modeling relying on PDEs (partial differential equations) and large algebraic-differential systems mainly concerned large manufacturing industries such as energy, aeronautics, and automotive mobilizing large numerical codes of finite elements and parallel architectures.

Today, computers with GPUs and other accelerators offer power increase. Computing platforms (cloud, edge) and current new business models such as SaaS (Software as a Service) have been developed. Data centers have multiplied, and new challenges have emerged: energy consumption, digital accuracy, security, and multidisciplinary algorithms. Over a decade, the whole sector has undergone profound changes.

In this context, Teratec is now a major player in Europe and must continue with actions and ability to bring people together, with strength in proposing new ideas and forward-looking approach in a sector undergoing such rapid transformation.

EXTREMELY FAST EVOLUTIONS AROUND HPC/AI/QUANTUM TECHNOLOGIES

Numerous technological and societal developments are going to stand out of a wide digital and parallel world. Among these, the following points are worth highlighting:

- The growing role of data, thanks to massive availability. This allows tools such as neural networks and Deep Learning to combine with conventional mathematical models (derived from physics), paving the way for a comprehensive approach to solving complex problems. This has led to widespread adoption of the concept of digital twins. The processing of data mixed with computing power is also fueling the rise of generative AI, which uses need to be mastered.

- Continued expansion of digital uses to all industrial and professional sectors as for everyday life. This pertains to crucial issues such as mobility, sustainable cities, agriculture and food, health and biology (new drugs, therapies and personalized medicine), but also climate, environment and biodiversity, not to mention areas such as the movie industry, video games, culture, commerce, fashion, textiles, or social networks.

- The arrival of new technologies, potentially disruptive, will undoubtedly bring new capabilities for analysis and discovery, such as storage on molecules and DNA, quantum computing, neuromorphic computers, etc. These technologies combined with current technologies will generate new developments at all levels: hardware, software, and algorithms. A necessary hybridization of all hardware, software, and algorithmic approaches is expected which requires continuous research efforts over the long term.

ONE FUNDAMENTAL CHALLENGE FOR EDUCATION AND MATHEMATICS

Our society must be able to master all these changes, understanding them, and put them to positive use while controlling some inevitable abuse. Initial and continuing education is another major challenge: ensuring that enough trained technicians and engineers can develop, create value and occupy jobs, but also enable everyone to use technology wisely and avoid abuses often associated with such radical innovations. All innovation brings about opportunities and risks as well but rejecting it would be very damaging to our society in the long term.

In this context, the major role of mathematics must be emphasized. Faced with the continuing decline in student standards, it is urgent to remedy practical measures and raise this fundamental discipline, conducting a broad promotional campaign to make it more attractive with the many exciting careers it offers, while better understanding its uses. This is essential to meet present and future challenges. As Einstein said, “One who has ceased to marvel has ceased to live”. So, let us continue to marvel at discoveries, go on support research, and remain confident in the future.

A NECESSITY AND AN OPPORTUNITY FOR FRANCE AND EUROPE

Digital technologies are and will remain a fundamental element for innovation, competitiveness, and value creation in about all sectors of the economy. They will be part of every professional activity and our daily life. France has many strengths, yet it must act quickly to maintain its position of excellence facing a growing global competition, investing in training specialists through a renewed and targeted scientific and mathematical education system. This must be part of a continuous process without giving in to fads, keeping up a strong European dimension.

In this context, Teratec must continue to play a leading role as a center of excellence, analysis, and foresight, supporting the regional, national, and European initiatives that are essential to ensuring sovereignty and competitiveness, while nurturing their promotion among all stakeholders. Let us seize this opportunity and remain confident in a future full of innovation and discovery, where digital technology will bring about many new developments, still unexpected as of today. ■

PERSPECTIVES

MASTERING DIGITAL TECHNOLOGIES

Hervé Mouren
Managing Director, Teratec

The rapid development of digital technologies, from high-performance computing and simulation to artificial intelligence and tomorrow, quantum computing upheaval is profoundly changing the way we work and design products and services. As early as 2005, major aerospace and energy companies recognized the profound transformation this would bring about in design and simulation.

Teratec now has eighty members and has developed into a diverse and highly complementary ecosystem, bringing together users, suppliers, teachers and researchers. Its objective is to ensure its members master digital technologies and to facilitate their dissemination and deployment. These two aspects, mastery and dissemination are key now, for digital technology makes it possible to design and produce products or services that were previously unimaginable, hence becoming one essential factors in industrial competitiveness. Its scope is now expanding to most fields with all sectors now involved and, gradually, all sizes of businesses.

For 20 years, Teratec has been actively supporting and contributing to the widespread use of high-performance computing and simulation. During this period, the computing power of supercomputers has increased a millionfold: the transition from teraflops to exaflops will enable us to tackle issues in 2025 that were unimaginable in 2005. In addition, the rapid development of artificial intelligence and quantum computing is opening a new and exceptional phase of developments, most of which are completely new. This is what we must prepare for.

To help raise awareness among an increasingly wider audience, Teratec organizes the annual Teratec Forum, as a major event in France for exchanges between researchers and manufacturers, suppliers and users.

Teratec is also very actively involved in the new European dynamics created around these strategic topics. We are one of the stakeholders in the European EuroHPC programme, which has a budget of €7 billion over six years, including the order for two exascale machines, the first of which, installed in Germany, is being built by French companies, and the second planned for next year in Bruyères-le-Châtel, in the Essonne region. As part of this European programme, Teratec is also involved in numerous projects and specifically operates a French center of expertise and competences for digital technologies with its partners, covering all fields of HPC, AI and quantum computing, providing support and training to all users in industry and research.

Today, digital technology impacts all industrial sectors and has been established as an essential component of our economic development. It has become one of the main differentiators in terms of competitiveness and efficiency for companies in all fields and at all levels, nationally, Europe wide and globally.

France and Europe have not yet fully grasped the strategic importance of developing scientific skills and mastering critical technologies. Other regions of the world have already made major advances in the last ten years. We have means to do the same, with an encouraging starting level: scientific excellence and high-quality training. This potential remains too little known; for example, some of the most widely used machine learning tools in the world were developed in our laboratories.

SKILLS AND TRAINING WILL MAKE THE DIFFERENCE

Digital technology is evolving very quickly, making training a strategic issue. Organizations that invest today in mastering these technologies will enjoy a competitive advantage. Those who master them best will be tomorrow's leaders. In initial training, it is necessary not only to develop dedicated courses, but also to prepare all future engineers for their best use. At the same time, our current research and engineering teams must integrate these tools without waiting for all specialists to come up. Continuing education is therefore essential. As digital technology is a race against time, we must accelerate the upskilling of present engineers, particularly in R&D. This is a major challenge and a key priority for Teratec.

Another challenge deals with the very difficult issue of data protection and, more broadly, cybersecurity. The protection of personal data, whether administrative, financial or health-related, is a major issue of central concerns for businesses, governments and all public and private organizations. We are living in an increasingly digitalized world witnessing a continuous rise in cyber-attacks and cybercrime. More than ever, it is essential to be able to guarantee the confidentiality and integrity of information shared online. This is the major risk of the coming period, through which we must be constantly vigilant.

There is another very important energy issue arising from the colossal investments planned for generative artificial intelligence. This headlong rush risks having serious consequences that are difficult to measure today.

Over the last few decades, digitalization has enabled the rise of email, teleworking and e-commerce, but also a wider dissemination of culture through digital platforms, with music and video being two vivid examples. Today, more sectors and industries capitalize on this know-how and these skills: agri-food, health, biology, transport, finance, insurance, culture, the environment and climate, and the entire field of human and social sciences. Digitalization will enable spectacular advances in the field of health. Digital technologies are opening new possibilities, particularly for treating conditions that were previously difficult to handle as well as for managing risk of epidemics. Major advances in healthcare lie ahead, as a wonderful challenge.

Teratec and its members are constantly monitoring developments, noticeably the rapid advancement of artificial intelligence but also quantum computing. First industrial applications are set to emerge soon, many of which being revolutionary as they are based on rather different approaches. In 2018, seven years ago, we launched the Teratec Quantum Computing Initiative (TQCI) to prepare for the arrival of quantum computing. This initiative has enabled us to carry out several projects in which many countries are now involved. This is a very good illustration of the principle of co-creation (and co-design) between suppliers and users, which is now particularly effective in a field such as ours.

France managed to maintain a leading position in research and training in all these areas. In the current international and geopolitical context, it is essential to preserve this position and even strengthen it by focusing still on training and giving our engineers and researchers the opportunity to integrate ongoing technological developments as early as possible. This is one of Teratec's objectives, both today and for the years to come. ■

AI FOR SCIENCE: FASCINATING SCIENTIFIC PERSPECTIVES, YET KEEPING VIGILANT ON OTHER TOPICS

Antoine Petit
CEO, CNRS

Artificial intelligence (AI) is transforming the world, particularly science with the exciting promise of accelerating scientific discovery, but also with risks that must be fully understood.

AI did not begin with ChatGPT. Nearly 70 years ago in 1956, the Dartmouth Summer Research Project on Artificial Intelligence was held, considered as the founding event in the field of artificial intelligence.

For a long time, scientific communities have been working in applied mathematics, computer science and data science, on reasoning and machine learning, on automatic signal processing in automation, robotics and human-machine interaction. And for almost as long, work has been done to use artificial intelligence (without necessarily using the term) in conjunction with other disciplines. Data processing in particle physics (particularly at CERN) and the development of bioinformatics are well-known examples.

The 2024 Nobel Prizes in Physics and Chemistry highlighted the contribution of AI to scientific discovery as well.

The CNRS has always aspired to help position French research at the highest international level in this field, by recruiting cutting-edge scientists, supporting research teams, structuring scientific communities, providing computing resources and strong cooperation with the economic framework and major international players.

This particular interest is reflected in the CNRS's 2019-2023 objectives and resources contract plan, in which artificial intelligence is one of six major social challenges (alongside climate change, energy transition, health and environment, the territories of the future and educational inequalities) for which the CNRS has set its ambition of contributing in a particularly energetic and proactive manner.

In 2021, the CNRS created the “AI for Science, Science for AI” (AISSAI) center to foster dialogue between all scientific disciplines around AI and encourage interdisciplinary projects. The AISSAI center inspired the declaration by the six leading European research organizations on AI.

This strategy is also being rolled out internationally. The CNRS has offices and research laboratories on every continent, including highly visible AI partnerships (notably in Canada, Japan and Singapore) with other ties currently under development (like in Australia and Morocco).

And in CNRS's latest contract of objectives, means and performance for 2024-2028, ‘Generative AI for science’ is highlighted alongside human brain, life in the universe, materials of the future, unlimited instrumentation and societies in transition.

One priority set by CNRS is to support this overall dynamics, to provide the research and innovation ecosystem with a sovereign HPC/AI convergent supercomputer, in partnership with the GENCI civil society and the ongoing support of the Ministry for Research (MESR) joined by the General Secretariat for Investment (SGPI), as part of the national strategy for artificial intelligence (SNIA).

The Jean Zay supercomputer has continued to evolve to keep pace with user demand. The fourth extension, installed in 2024, enables to perform 126 million billion operations per second!

Jean Zay's success is largely due to the conditions provided to users: one fast and simplified access for 'small' projects relying on a network of engineers to ensure support and training for user teams.

This model needs to be further strengthened and reachable in Europe as part of the AI Factory France project led by GENCI. In addition, further partnerships with players in the European cloud industry should enable the emergence of a decentralized, interoperable and sovereign AI ecosystem capable of competing on a global scale.

Beyond computing and support activities, CNRS is developing an integrated strategy to accelerate scientific discovery through AI, leveraging on one of its essential components with talent development. In 2021, the CNRS created specific positions to recruit scientists from other disciplines and highly familiar with data science and AI. The AISSAI center is therefore launching calls for postdoctoral researchers to promote such development of hybrid skills. Finally, the CNRS is supporting the SNIA's Choose France – Rising Talents program, being open to all disciplines.

An undeniable sign of success of this long-term policy comes as, more than one in five start-ups emerging each year from joint research units at the CNRS and its partners having a significant artificial intelligence component in their products or processes, almost all of which being linked to one or more other disciplines. The more promising prospects for innovation in France and Europe mostly set at intersections X+AI (where X can be medicines, materials, satellites, connected clothing...).

However, as in society at large, the deployment of AI in science cannot happen without risk. Research teams are working to analyze such risks, to control them and develop alternative approaches.

The energy cost and carbon footprint of AI represent a major handicap. In the short term, the aim is to optimize algorithms along with low-carbon infrastructure while developing foundations for frugal AI (favoring models with fewer parameters in particular), implementing policies to mitigate excessive uses. In the longer term, research into new computing components or paradigms (quantum, DNA-based, or even biomorphic) could open alternative avenues.

Another priority is to develop safer AI. The development of AI for science requires the ability to verify and explain results produced or assisted by AI. However, biases in scientific data can spread and amplify, the interpretation of results being made difficult by the increasing complexity of models, AI systems remain vulnerable to erroneous data or various attacks.

The use of AI in scientific practice also raises questions of scientific integrity in relation to the rules governing publications and peer review of scientific articles, the lack of data traceability and non-reproducibility of results.

The development of AI for science ultimately raises societal issues: unequal access for scientific communities due to the cost and concentration of computing resources and data among few players, higher dependence on proprietary tools, and potential risk of massive errors or manipulation of scientific results.

All of this requires greater transparency, explainability and governance of scientific AI. And it calls for the mobilization of the scientific community at European level, at the very least.

The story does not end with ChatGPT, nor does it begin with it. Research teams are fully aware of both the potential and limitations of generative AI, working to develop more reliable, frugal and powerful models to enable the future use of agentive AI capable of autonomous decision-making and action taking, followed by "physical" AI that can interact with the physical world and humans.

It is up to us to ensure that AI becomes a common good that benefits everyone, and is not monopolized by a few, whether states, large multinationals or individuals. This is a matter of governance for global democracy. ■

FROM CONTROL SYSTEMS AND COMMAND TO GENERATIVE AI: 40 YEARS OF DIGITAL TRANSFORMATION IN INDUSTRY

Patrice Caine
Chief Executive Officer of the Thales Group
President of the National Association for Research and Technology (ANRT)

Digital technology in industry already has a long history. To put it simply, we can say that it began in the 1980s with, amongst other examples, the transformation of control systems and the creation of programmable logic controllers, as well as the advent of CAD/CAM software. At that time, the objective was to increase productivity, normalise quality and improve the repeatability of industrial processes.

In the 2000s, widespread adoption of networked computing and ERP systems led digital technology to become a cross-functional management tool. Functions that had previously been organized in silos were connected: production, logistics, maintenance, finance and HR. The factory became an integrated, data-driven system capable of better adapting to demand, optimizing inventory, and making schedules more reliable.

The 2010s marked a major turning point with the emergence of what was sometimes tagged as 'Industry 4.0'. The Industrial Internet of Things, robotics, digital twins, augmented reality, cloud and big data found their way into engineering centres. Equipment communicates, learns and adapts. The approach became resolutely systemic, combining the physical and the virtual, simulation and real data, man and machine.

Today, we are entering a new phase of the digital revolution. Generative artificial intelligence, autonomous systems, high performance computing (and in the future quantum computers) will enable unprecedented system complexity, depth of modelling and innovation speed.

DIGITAL IS NOW OMNIPRESENT AT THALES

At Thales, we have been fortunate to be in a very favourable position to take advantage of these technologies over the decades. Beyond being a tool for improving our production efficiency, digital technology is indeed at the heart of our solutions and achievements in R&D. Our expertise in cybersecurity also enabled us to integrate innovations into our production units while guaranteeing a very high level of protection and reliability.

In our space activities, for example, satellite manufacturing now relies on extremely detailed digital twins that replicate thermal behavior, mechanical constraints, electromagnetic interactions, etc. These models, fed by large amounts of test data, have multiple advantages: they enable us to reduce physical iterations, quickly reconfigure production lines and anticipate potential defaults.

In Cannes, the SAPHIR robot, developed for the automatic installation of inserts on satellite panels, combines artificial vision, high-precision mechatronics and big data analysis. This combination has made it possible to increase production speed while improving quality and traceability. In Hasselt, Belgium, we

© Thales

inaugurated a photovoltaic panel factory in 2019 that was designed from the ground up according to Industry 4.0 principles. It integrates real-time data flows and digital interfaces between sites, directly connected to our centre of expertise in Cannes. The result is greater agility and better resource management.

In the defence sector, the development of complex systems now relies on end-to-end digital chains. System engineering, behavioural modelling, operational simulation, virtual testing, cybersecurity: everything is integrated upstream.

Digital technology has also facilitated collaboration between Thales teams and our ecosystem of SMEs, mid-sized companies and partner laboratories, thanks to secure collaborative environments.

AN ONGOING TRANSFORMATION

This full digitisation of the industrial chain has transformed jobs and skills sets. It is also changing the position of the factory in the global ecosystem, which is set to become one node in a vast distributed industrial system, capable of interacting dynamically with their stakeholders. Reaching this stage of development requires massive investment in digital infrastructure, cybersecurity, and training.

French industry is resolutely moving in this direction, across all sectors. our country still needs to speed up progress, by helping small and medium businesses to adopt digital technologies and set up shared digital platforms at a sector level.

The importance of this issue cannot be overstated, as digital technology is now much more than just one tool among many. It is truly the backbone of 21st-century industry. It has already transformed the way we design, produce and cooperate. Tomorrow, it will be central to developing an industry that is more resilient, sovereign, and sustainable. ■

TECHNOLOGIES

INTRODUCTION: TECHNOLOGIES: PROFOUND CHANGES

Jean-François Prevéraud
Journalist

Digital technologies being increasingly present in our professional and personal lives are constantly evolving, both in terms of components and physical implementations, as well as for system architectures and associated software. They no longer follow Moore's Second Law.

PERFORMANCE

As developing new chip technologies is becoming prohibitively expensive, foundries are to join forces to share the staggering costs of new factories where engineers must rethink processors. The days are gone for large monolithic chips. Long live chiplets combining elements of different types and technologies on the same substratum to achieve the expected performance at a lower cost and guarantee high-density interconnection, ensuring the extremely high bandwidth essential for HPC.

CPUs are complemented within a hybrid system architecture by a multitude of accelerators. And beyond GPUs, they are supported by Neural Processing Units (NPUs), Data Processing Units (DPUs), Infrastructure Processing Units (IPUs)..., which relieve them of multiple tasks. We are also attending the emergence of new physics, such as quantum, photonic, spintronic and neuromorphic which will change the game once mastered. Conventional computing will then give way to hybrid computing, mixing technologies according to needs which will have an impact on implemented architectures. The containerization of applications is one typical example.

FRUGALITY

But we must not forget that digital technology is one of the biggest consumers of energy, accounting for 11% of electricity consumption in France according to Ademe, which means engineers need to develop high-performance chips that should be much more frugal. This is even more necessary given that generative AI is causing demand for power to skyrocket.

Technological developments in components and system architecture also have an impact on data storage. There is no point in having fast calculations if you must wait for the data. The overall performance of a system is determined by its least performing link. This is why manufacturers develop their substratum and access technologies accordingly, especially as cloud storage is rising. Yet, we are already talking about Edge Storage.

SECURITY

Finally, we must not understate the security of such systems. In a world of crisis where alliances are being overturned, wars are breaking out, and companies are spying on each other. So, we must be increasingly vigilant against cyber threats coming from anywhere. Data and data intelligence are now the real added value of organizations, from businesses to governments. They must be protected.

We asked leading international experts to outline these current and future developments in the following pages.

EVOLUTIONS OF COMPONENTS FOR HIGH-PERFORMANCE COMPUTING: OUTLOOK FOR THE NEXT 5-10 YEARS

Marc Duranton
Senior Fellow, CEA

HPC systems have seen an unprecedented increase in performance over the past 20 years: comparing the BlueGene/L in 2005 with its 280 TFlops and El Capitan in 2024 (1,742 Pflops), representing a 6,222-fold increase in computing power with a moderate increase in power consumption, from 1.43 MW to 29.6 MW, resulting in an overall improvement in energy efficiency by a factor of 300.

However, we come to a crossroad: the rise of emerging workloads – particularly artificial intelligence – and slowdown in traditional CMOS scaling are redefining basic components. Over the next ten years, HPC systems will increasingly rely on specialized and heterogeneous architectures to maximize performance within strict power and cost constraints: systems based on chiplets and dedicated accelerators, devices using new technologies such as photonics, spintronic, or new technologies and innovative communication systems will need to address the challenges of power consumption, thermal limits, and reliability.

CHIPLET-BASED ARCHITECTURES AND ADVANCED PACKAGING

Over the next few years, processors and circuits for HPC will increasingly adopt chiplet-based designs rather than monolithic chips.

EXPONENTIAL INCREASE OF PERFORMANCE IN 20 YEARS

BlueGene/L - 2005
280 TFlops/s (280 x 10^{12})
1.43 MW

x 6222 ▶ **x 21**

El Capitan - 2024
1742 PFlops/s (1742 x 10^{15})
29.6 MW

Energy efficiency x 300 in 20 years

Figure 1 • Increase in energy efficiency of HPC machines over 20 years

Chiplets are small silicon modules, each performing a specific function, connected to a silicon substrat that enables high-density interconnection between them: the interposer. This assembly is then integrated into a package, as for a monolithic circuit. This assembly approach makes it possible to combine components manufactured in different technological nodes and specialized for distinct tasks into a single module – a 'circuit'. For example, a high-performance CPU core chiplet engraved in 5 nm can be coupled with analogue or input/output chiplets produced on older, less expensive nodes. The driving force behind this evolution is both economic and practical: manufacturing a single large chip is becoming prohibitively expensive and characterized by low yields, whereas several smaller chips improve yield capacity and reduce the cost per component. It also allows a wide range of accelerators and memory to be integrated side by side, maintaining diversity and specialization without the cost of completely redesigning a monolithic chip. Advanced packaging innovations are supporting this chiplet revolution. Silicon interposers with TSV-Through Silicon Via (2.5D integration) and direct die stacking (3D) provide the dense interconnectivity needed to connect chiplets with extremely high bandwidth. For example, AMD's MI350 chips merge CPU chiplets, GPU chiplets and 8 HBM (High Bandwidth Memory) memories in a 3D configuration within a single package for a total of over 150 billion transistors. By integrating the CPU and GPU on an interposer shared with memory, the MI325 achieves very high bandwidth, ideal for AI and HPC workloads.

Despite their promise, chiplet architectures face several challenges. One of the most pressing stakes is the lack of industry standards for interfaces, power distribution and physical specifications. The industry is therefore striving to converge towards open standards; the Universal Chiplet Interconnect Express (UCIe) is one example, targeting data rates of 32 Gbps per pin between chiplets. Known Good Die (KGD) testing is crucial to prevent a defective chiplet from causing the scraping of an entire expensive package. Embedded sensors (temperature, current, error detection) and redundancy mechanisms are also being explored to detect and manage failures in situ.

AI ACCELERATORS AND FIELD-SPECIFIC ACCELERATORS

The rise of machine learning is driving HPC systems to integrate new specialized accelerators. For the past decade, GPUs have prevailed thanks to their massive parallelism. From now on, Neural Processing Units (NPUs) – processors optimized for neural networks – are gaining ground for low-latency inference. Modern CPUs also incorporate AI instructions or small on-chip accelerators. This trend is reinforced by the reduction in numerical precision: where simulations used to compute in double precision (FP64), AI models often accommodate 16-bit, 8-bit, or even 4-bit formats with minimal impact on the final result, resulting in major gains in performance and energy efficiency (FP8 arithmetics can offer ca. 8x more performance than FP64).

Other specific accelerators are emerging. Reconfigurable hardware such as FPGAs can accelerate specific algorithms or serve as a test bed for new architectures. DPUs (Data Processing Units) or IPUs (Infrastructure Processing Units) relieve the CPU of network, storage and security tasks. We also find the emergence of accelerators based on the open RISC-V ISA, illustrating the search for advanced customization with fewer constraints. As the gains from CMOS technology diminish, the next decade will see a profusion of accelerators, each targeting a class of workloads (AI, graphs, cryptography, data movement) within heterogeneous platforms.

HETEROGENEOUS COMPUTING AND MEMORY HIERARCHIES

With the proliferation of accelerators, HPC system architecture becomes one heterogeneous model within which various processors must cooperate. Standards such as Compute Express Link (CXL) will enable CPUs, GPUs, NPUs and memory devices to access a common coherent memory space. Memory technology is diversifying, but the gap between computing and storage capacities is widening over time: HBM (High Bandwidth Memory) provides enormous throughput but remains expensive and limited in capacity. Alternatives such as DDR5/ LPDDR5X and MCR-DIMM (Multiplexer Combined Ranks DIMM) pledge for a better cost / power/

SCALING OF PEAK HARDWARE FLOPS, AND MEMORY/INTERCONNECT BANDWIDTH

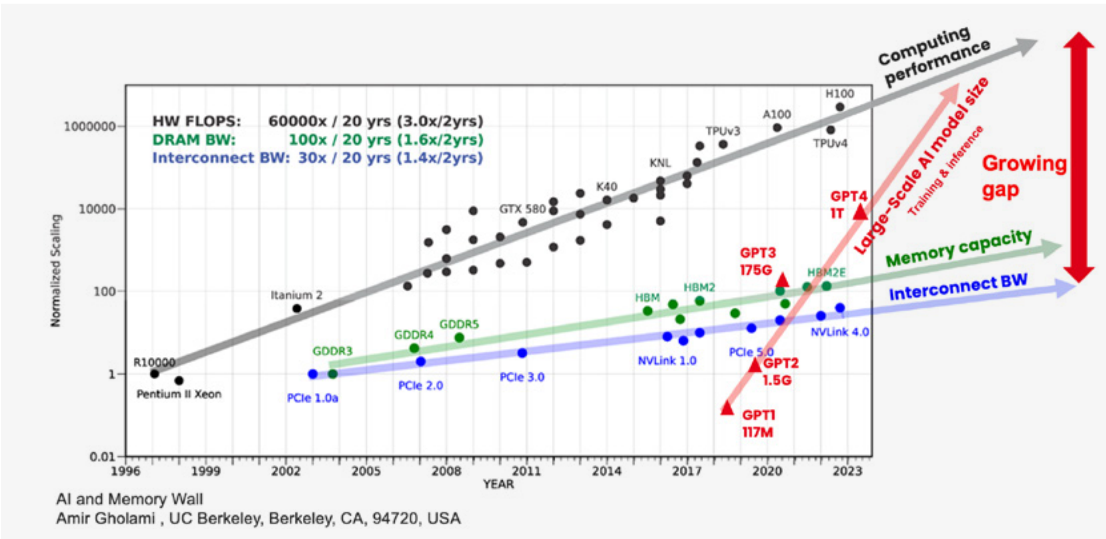


Figure 2 • Increasing gap between AI requirements, computing performance and memory

bandwidth compromise. Hierarchical memory architecture will combine HBM or SRAM on the chip, a large DDR pool for capacity, and potentially non-volatile or remote memory via CXL. Processing-in-memory (PIM) concepts which are still limited today could eventually reduce data traffic by executing certain operations within memory chips themselves.

Specialized interconnects (NVLink, Infinity Fabric, etc.) aim to link clusters of GPUs to form 'super-nodes' that share memory, as required by running models involving hundreds of billions of parameters on thousands of GPUs. The future will undoubtedly see disaggregated architectures implemented where resources (memory, accelerators) are dynamically allocated over a networked framework, supported by optical links and protocols such as CXL. The goal remains to minimize the latency and energy cost of data movement. A successful heterogeneous HPC architecture will take the presence of multiple specialized chips as transparent as possible to programmers, achieving performance and efficiency as if it were a single unified machine.

EMERGING TECHNOLOGIES: NEUROMORPHIC, PHOTONIC, AND SPINTRONIC PROCESSORS

Photonics is gaining ground in communication and even computing. Co-packaged optical interconnects (CPOs) could replace copper links, delivering immense throughput and lower energy per bit transmitted. Photonic accelerators capable of optical matrix multiplication are also emerging. Main challenges include integration, the thermal impact of on-chip lasers and reliability, but the potential gains in bandwidth and energy efficiency fuel intense R&D efforts. Photonic interposers are also emerging.

Spintronics taps electron spin for memory and logic. Non-volatile MRAM memories are their closest commercial example. They could be used as fast persistent cache or checkpoint storage in HPC. In the longer term, spintronic based accelerators (Ising machines) could solve optimization problems directly in hardware. Advances in materials and devices (reduction in variability, scaling) will determine the speed of spintronics integration.

This paves the way for accelerators using physical phenomena to perform calculations. One example points neuromorphic processors, which operate asynchronously and are event-driven, with extreme energy efficiency for certain algorithms. However, challenges related to scaling remain to be resolved: these devices often do not support time folding like conventional digital architectures, because we need as many physical devices as the size of the problem to be solved. It is generally not possible to reuse a device in another part of the computation as storing its state remains very ineficient. These devices often have variability that also limits scaling. In the next 5–10 years, these chips will remain specialized coprocessors ultra-parallel, energy-efficient systems.

ENERGY EFFICIENCY, POWER AND THERMIC MANAGEMENT

Energy consumption is the ultimate constraint: an exascale machine must remain within an envelope of around 20–30 MW. Specialized accelerators, precision reduction and 26 DVFS (dynamic voltage and frequency scaling) optimization aim to maximize flops per watt. Further increases in efficiency will require disruptive approaches: near-threshold voltage operation, new materials, cryogenic computing, 3D architecture and increased specialization.

SYSTEM ARCHITECTURE, RELIABILITY, AND SUSTAINABILITY

The convergence of these trends will transform the architecture of supercomputers, making them less monolithic and more composed of varied resources. But this complexity increases the risk of failures. Studies show that breakdowns are yet frequent and will increase with the number of components. MTTF (mean time to failure) could be 14 minutes for a job of 131,072 GPUs[1]. Future systems will therefore incorporate multi-level fault tolerance mechanisms, from on-chip sensors to predictive maintenance algorithms based on telemetry. The new technologies needed to continue increasing performance and efficiency are still poorly characterized on a large scale and will require careful integration phases to ensure reliability constraints are met. ■

1. https://arxiv.org/abs/2410.21680

QUANTUM COMPUTING: TOWARDS A NEW ERA FOR INTENSIVE COMPUTING

Pascale Senellart
Director of Research, CNRS
Center for Nanoscience and Nanotechnology – Université Paris Saclay, CNRS
Cofounder and Scientific Advisor, Quandela

The past decade, and most likely the next to come, represents a pivotal period for the introduction of quantum computing into the high-performance computing landscape. The rapid evolution of this field is remarkable. Twenty years ago, so-called 'second-generation' quantum technologies were still a distant prospect, confined to academic laboratories. Apart from a few pioneering initiatives such as D-Wave, the field truly reached a turning point in 2016 with the first quantum computing machines brought online by IBM and the involvement of other major companies in this technological race. The ecosystem then expanded rapidly, with the creation of most start-ups specializing in quantum computing between 2015 and 2020.

Today, the first quantum computers started being integrated into high-performance computing (HPC) infrastructures all over the world. Through the EuroHPC program, Europe is also investing along this path. The promises associated with this technology are considerable: beyond simple acceleration, it paves the way for calculations that are inaccessible even to the most powerful supercomputers, thanks to the introduction of new computational primitives based on the fundamental properties of quantum physics. Such prospects have led many countries to consider quantum computing as a strategic and sovereign technology. However, considerable challenges remain, whether in terms of integration into current HPC uses, increasing hardware performance or developing relevant applications.

THE FOUNDATIONS AND PROMISES OF THE SECOND QUANTUM REVOLUTION

Current supercomputer architectures already take advantage of quantum mechanics concepts, notably through the invention of the transistor and laser in the mid-20th century. This 'first quantum revolution' shaped the contemporary digital age. The impact of this historic milestone allows us to gauge the potential of the 'second quantum revolution,' this time focused on exploiting more subtle phenomena such as quantum entanglement and superposition.

Transition from the classical bit to the quantum bit (qubit), capable of simultaneously representing states 0 and 1, paves the way for new algorithmic paradigms. Certain theoretical algorithms, such as Shor's algorithm for factoring large prime numbers, exemplify the disruptive power of this approach. Others, more recent, promise exponential or polynomial accelerations in such fields as quantum chemistry, the design of new materials, finance or logistics.

However, these prospects face a major constraint: quantum coherence. This coherence, which is essential for maintaining superposition and entanglement, is extremely fragile and tends to disappear under the effect of environmental disturbances. For several decades, the scientific community has been working to isolate and control these systems, and the progress made over the last ten years has enabled the implementation of the first algorithms on quantum devices. Nevertheless, optimal, large-scale control remains a long-term goal.

TECHNOLOGICAL DIVERSITY AND THE CHALLENGES OF SCALE-UP

Like the early days of classical computing, quantum computers today rely on a variety of hardware technologies. Qubits can be implemented using superconducting circuits, photons, silicon-based components, or even single atoms trapped in a vacuum.

Each approach has specific advantages but also limitations. Photonic or atomic systems offer natural coherence preservation. However, to scale up to several million qubits, architecture based on superconducting or semiconductor materials appear better suited to industrialization, although their quantum stability is more difficult to maintain. Ultimately, hybrid architecture combining different approaches is likely to be the most promising route to achieving universal performance.

FROM EXPERIMENTAL DEVELOPMENT TO FIRST USE CASES

The initial development of quantum computing was driven by physicists then, gradually by specialists in quantum algorithms. Interactions with the HPC and applied computing communities, although late in coming, are now growing rapidly. This convergence is essential to accelerate the identification of relevant use cases.

The first experiments are based on devices that are still limited and imperfect. However, the online accessibility of quantum processors and their integration into several computing centres in Europe, the United States and Asia have led to a proliferation of exploratory studies. Players such as IBM, Quandela, Quantinuum and Pasqal are testing applications in a variety of fields, including industrial process optimization, chemical simulation and machine learning. The link between artificial intelligence and quantum computing is also a growing area of research. Recent initiatives, such as the Perceval Quest and the Airbus-BMW Quantum Challenge, exemplify such a convergence, where quantum computing opens new possibilities for AI, while AI itself contributes to the optimization of quantum architectures. Beyond the question of qubits and specific hardware characteristics of each platform, the fundamental contribution of quantum computing lies in the introduction of new computational primitives. Whereas conventional architectures have only evolved in scale over several decades – massive parallelism, graphic or vector accelerators – quantum processors are based on completely different mechanisms, derived from superposition and entanglement. These unconventional primitives do not substitute existing intensive computing resources: they complement each other. The major challenge is to integrate these novel building blocks into high-performance computing chains, leveraging supercomputers for most tasks while offloading certain steps to quantum processors. There are two application prospects: on the one hand, in the field of quantum machine learning, where the expressiveness of models can be enriched by quantum circuits; on the other hand, in specialized algorithms exploiting these primitives (combinatorial search, simulation of complex quantum systems, optimization). In this context, the aim is not just to compare the performance of quantum and classical machines, but to design new hybrid computing flows capable of combining the robustness of HPC infrastructures with the emerging capabilities of quantum processors.

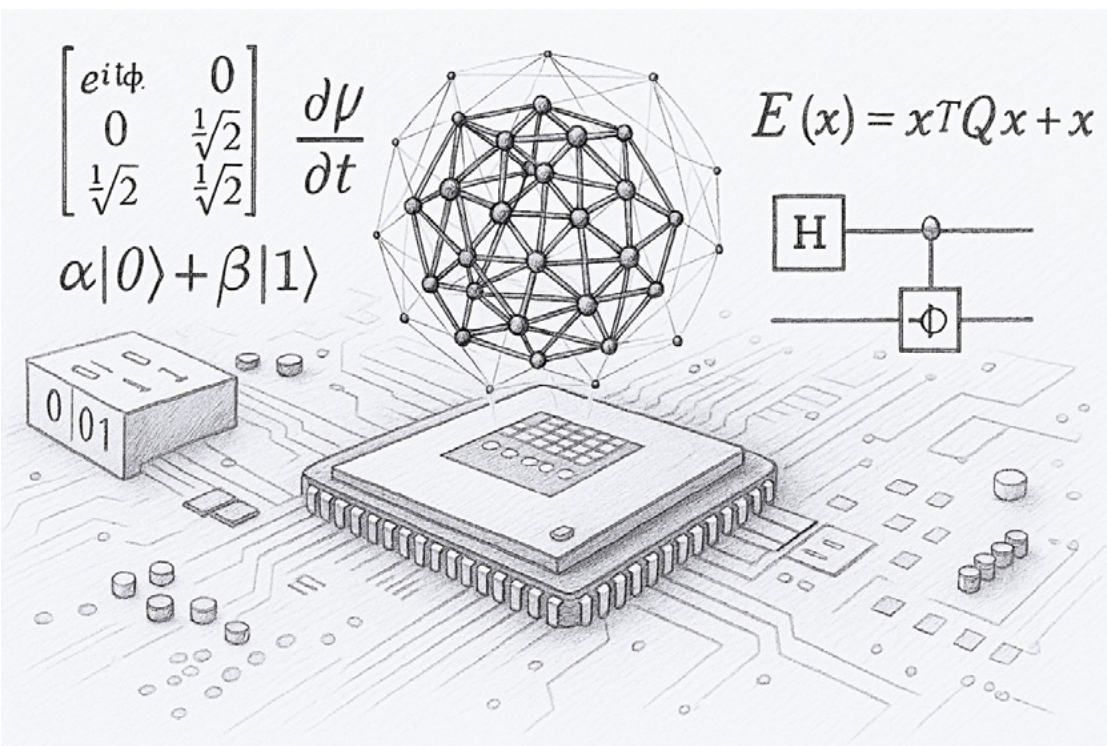
In the short term, the goal is to achieve what is known as ‘quantum utility’, characterized by measurable gains in computing time, energy consumption or precision, through the joint integration of classical and quantum processors. This transitional stage will promote maturation of technologies and guide future developments towards universal quantum computing with error correction.

1. https://www.scaleway.com/en/blog/1st-perceval-quest-a-journey-into-quantum-machine-learning/
2. https://www.airbus.com/en/innovation/digital-transformation/quantum-technologies/airbus-and-bmw-quantum-computing-challenge

CHALLENGES, RIGOR, AND PROSPECTS

The success of this transformation depends on the development of a robust ecosystem. This involves developing new programming paradigms, making quantum development tools accessible, and promoting the expansion of the user community.

Two pitfalls must be avoided. On the one hand, excessive skepticism, which underestimates the progress made and the medium-term prospects; on the other hand, excessive and sometimes misleading communication, which contributes to maintaining unrealistic expectations. Between these two extremes, the path to follow must be marked by scientific rigor and technological ambition. Quantum computing is indeed a unique scientific adventure, mobilizing multidisciplinary skills and generating enthusiasm comparable to that of the great technological revolutions of the previous century. In ten years, it will be possible to measure how far we have come and to assess the position this technology will have taken in the high-performance computing ecosystem. If future is not limited to 'faster computing' but rather to the integration of new, unconventional primitives into HPC infrastructures, its impact will be profound. Everything indicates that quantum computing will establish as one of the major breakthroughs of the 21st century. ■



AI-generated illustration

THE ENERGY CHALLENGES FACING QUANTUM COMPUTERS

Olivier Ezratty
Cofounder, Quantum Energy Initiative

One of the motivations behind the creation of the Quantum Energy Initiative (QEI) in 2022 was to bring together academic researchers and industry vendors interested in the issue of energy consumption in quantum technologies, particularly quantum computers. The aim was to study it as early as possible in the design of these systems and avoid the unpleasant surprises that abound in artificial intelligence with pharaonic projects for computing centers consuming GWs of electrical power.

The approach consists in creating a scientific basis for understanding, evaluating and optimizing this consumption, starting with the fundamental physics of qubits and then integrating all other quantum computers components, including qubit control electronics, cryogenics and all software stacks. One of the key actions of QEI is to lead a working group at IEEE to standardize the evaluation and measurement of quantum computer power and energy consumption.

Since 2022, the quantum computing ecosystem has evolved with the proliferation of demonstrations of logical qubits, particularly with superconducting, cold atom and trapped ion qubits, and the publication of detailed roadmaps by industry vendors now working to create default-tolerant computers. The error correction required to create these machines is costly. It increases the hardware infrastructure by one to several orders of magnitude compared to that of current noisy computers known as NISQ (Noisy Intermediate Scale Quantum). As a result, it may create various challenges related to the integration of future large scale quantum computers in existing data-centers, particularly, those that are using supercomputers. How far will conventional computing infrastructures need to adapt to the integration of these newcomers?

This raises several other questions. The first concerns how these machines will be used. Will they be dedicated to highly specialized calculations in fundamental and applied research, such as chemical or new material simulations, or will they be extended to production applications for daily use by many companies, for example to solve optimization or machine learning problems? What will be the scientific and economic added value for users, whether they come from academia or industry?

ESTIMATING THE POWER CONSUMPTION OF QUANTUM COMPUTERS

Initial estimates from suppliers indicate that the electrical power required to power quantum computers supporting a few thousand logical qubits could range from 1 to 100 depending on the technology, with a similar difference in terms of machine cost and carbon footprint. That is, between tens to hundreds of kW to around 100 MW. In some cases, these machines will therefore be more expensive in terms of power than the largest current supercomputers, which range from 1 MW to 40 MW at peak power. This will correspond to situations where these machines will be able to solve problems that are inaccessible to conventional supercomputers.

Their energy acceptability will then have to be defined. It will depend on their return on investment, current and future practices in high-performance computing investment, and capacities and constraints of electricity grids.

All this goes with other operational constraints related, among other things, to cryogenics which is necessary for almost all types of qubits, temperature and vibration stability, and the m aintenance of certain components.

INTEGRATING CONVENTIONAL COSTS INTO QUANTUM CALCULATION

A final energy constraint that is rarely assessed is the cost of standard computation surrounding quantum processors. These correspond to various key functions for which estimates are still scarce.

The first is the classical cost of real-time error correction. With many physical qubits supporting the corrected logical qubits required for default-tolerant computing, this correction relies on numerous readout and syndrome detection cycles. This could be a costly operation in terms of control electronics, signal processing and, in some cases, classical computing.

The second is the compilation and optimization of quantum circuits. With thousands to millions of physical qubits, this compilation could involve significant computing time and energy costs, especially when the operation is frequently repeated as the problem data evolves. It will also need to incorporate the quantum circuit partitioning techniques required to distribute quantum computing across multiple quantum processors connected by quantum interconnect technologies.

The third challenge is the cost of classical pre- and post-processing of quantum algorithms. In chemical simulations, these processes are substantial. Particularly, it is necessary to prepare the circuit describing the quantum state of the quantum system to be simulated, which must be as close as possible to the fundamental ground state of that system.

These three classical tasks have in common that they sometimes must solve ‘NP-complete’ problems, meaning that their cost can increase exponentially as the power of quantum computers increases. However, various optimization techniques are being developed for these various classical computing tasks. We will therefore need to evaluate not only the energy consumption of quantum processors and their immediate environment, but also that of all standard computing tools necessary for their operation.

BALANCING COMPUTATION TIME, POWER, AND ENERGY

It will also be necessary to find the right balance between computing time, cost and energy consumption for quantum processors. Indeed, some estimates of computing time for complex simulations are in weeks and montgs. To speed up these computing times, it may be possible to distribute the execution of (identical) quantum circuits across several quantum computers operating in parallel, provided that their cost is acceptable.

All this tends to confirm the need to adopt a systemic and interdisciplinary approach integrating quantum physics, enabling technologies, quantum algorithms, error correction, standard computing as well as economic considerations. ■

EVOLUTIONS OF SUPERCOMPUTER ARCHITECTURES

Jean-Philippe Nominé
Fellow, CEA DIF

Patrick Carribault
Fellow, Research Director, CEA DIF

Jean-Christophe Weill
Fellow, CEA DIF

High-performance computing relied on mainframe computers in the 1960s and 1970s, then on vector computers until the late 1990s – in a market limited to a small number of large organisations typically purchasing Cray or NEC machines. These highly specialised machines had the capacity to pipeline data processing sequences (organized into 'vectors') and high memory bandwidth. This sector gradually dried out with the rise of massively parallel machines in the late 1990s. These cluster-type architectures benefited from the integration of widely available processors mass market COTS components that were much more affordable than vector processors and could be used in large amounts to increase performance. This model has prevailed ever since with an overall machine organisation comprised of clusters of rather complex and homogeneous nodes, sharing core local memory in a more or less uniform manner. The nodes are interconnected through a network with appropriate performance and are generally interacting with a mass storage system through that same network.

CLUSTER ARCHITECTURE MODEL

For more than 20 years, the cluster architecture model dominated with various refinements, variations and developments:

- The emergence of multi-core processors to continue offering gains in computing power without causing increase in energy consumption to skyrocket ;
- For the same reason, clock speeds have plateaued (1 to 3 GHz);
- Addition of vector extensions to CPUs;
- Addition of GPUs as coprocessors, until they became the main fuel of computing power; indeed, the ratio of computing power to electrical power consumption favors GPUs, although they are more difficult to program than CPUs for most applications (even though unfeasible for some, at least not at a reasonable porting cost).

IMPACT OF ARCHITECTURE ON PROGRAMMING

Parallelism is required at all levels in these approaches:

- intra-node or even intra-processor multi-core, expressed in OpenMP or by explicit threads, for example;
- inter-nodes via interconnection network, based on a message communication system or with distributed task programming (MPI-type layer); the computer's global memory is the aggregation of all nodes' local memories;
- inputs and outputs parallel proxy processes able to gather partial data flows on nearby servers or shared files;
- the storage system generally uses a set of dedicated servers and performs operations in parallel (example: LUSTRE, as a parallel file system).

Supercomputeur EXA-HE, CEA/DAM – 180 Pflop/s
BullSequana XH3000, with Grace Hopper Superchip 72C 3GHz processors, NVIDIA GH200 Superchip
BullSequana eXascale Interconnect (Quad-Rail BXI v2)
Cooling with lukewarm water

The main difficulties in the effective use of such architecture are:

- The necessary coexistence and entanglement of different levels of parallelism (from instruction level parallelism at the computing core level, the possible use of local vector extension units for each processor, intra-processor or intra-node threads, the distribution of coordinated processes or tasks between nodes);
- The limitations and relatively slower growth in performance of processor/memory bandwidths; in unfavorable cases, a calculation instruction may require several dozen or hundreds of cycles to bring data into the register or proximity cache (it has been observed that most applications are generally memory-bound, i.e. limited in performance by this effect, never saturating the processor with calculations);
- In general, data movements at all levels are a major source of performance degradation, between simulation processes, between simulations and input or output files.

Indeed, HPC is a constant (we could say: Sisyphean) struggle to adapt and optimize software on architectures that are evolving faster than the hardware. In general, optimum performance is not achieved until the hardware is changed. We also know that, in many cases, greater efforts to improve software efficiency would yield significant gains (less energy-intensive, faster calculations; more calculations possible on available resources and energy). But human resources are unfortunately often the scarcest resource which rarely allows us to get closer to an optimum in this area. This heterogeneity has therefore become dominant, without any major standard for productive programming and portable performance emerging. MPI+X (X=OpenACC, OpenMP, for example) still reigns supreme in HPC. Methods and tools for better abstracting and separating the expression of parallelism from the expression of the model to be solved (platforms such as Kokkos, DSLs, etc.) and others for source-to-source translation and code optimization have emerged, again without any dominant approach.

EVOLUTION OF USE CASES: THE EMERGENCE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence (deep learning, with its inference and training requirements) has joined the fight for several years now. With its appetite for high computing and processing power, it has channeled developments in the field, not always to the benefit of HPC which had provided a foundation of parallelism techniques:

- The reduced precision requirements of AI are diverting processor manufacturers away from the double-precision floating-point calculations required for 'high-fidelity' HPC computing. What's more, as GPU prices are skyrocketing and facing the uncertainty of this timeframe, HPC is suffering from limited investments particularly in the public sector, Should we expect manufacturers to provide 64-bit emulation on 4-, 8- and 16-bit arithmetic units? Still unsure. Should we adapt applications as much as possible in terms of precision and reduce double-precision floating point to the bare minimum (mixing different precisions)? This is a huge problem, perhaps worse than what we experienced when parallelizing older vector applications. Are there any general lower-level methods for effective software emulation with higher precision? Not yet (see Ozaki, which only does matrix products so far – below reference*).
- Node architectures tend towards multiple GPUs per CPU, raising the question of efficient local memory sharing. The once-promising path of APUs (accelerated processing units) with memory integration between CPUs and GPUs is neither becoming mainstream nor certain.
- Network requirements are also being driven by AI, for the efficient injection of massive training data into GPUs. Here, we are not this far removed from HPC, as the cause may be common to motivate virtuous uses of interconnection photonics in HPC and AI.
- AI does not have the same data management needs as HPC which is generally more structured and hierarchical. But for exascale computing, HPC is increasingly looking to object storage to overcome the limitations of current parallel file systems, such as POSIX.

Despite tensions and differences, HPC and AI architectures will remain linked. HPC will have to evolve to stay in the race, serving high-fidelity modelling methods and simulations that will remain indispensable for a long time to come. ■

*Ozaki : https://arxiv.org/html/2504.08009v1

HYBRID ARCHITECTURES, MYTHS AND REALITIES

Arnaud Bertrand
Senior Vice President 3DS Outscale R&D

Let's start by dispelling a common misconception: there is no such thing as a 'one-size-fits-all' architecture created by a so-called Sauron – divine spirit of data centers – that can do everything perfectly. First and foremost, computing architecture is designed to serve a specific use case. It is designed with one idea in mind: to achieve the best possible compromise for that use case between:

- Cost;
- Performance;
- Sustainability, encompassing environmental impact (including electricity), operations (availability) and the ability to evolve over time.

THE ULTIMATE GENERAL-PURPOSE ARCHITECTURE: CLOUD COMPUTING

Cloud computing architecture was designed from the outset to provide an acceptable compromise across all three of these areas. They combine:

- Industrialization of deployments (homogeneous hardware choices, backbone network, 1-2 types of servers, shared storage);
- Optimal resource allocation through an orchestrator;
- Resource sharing through abstraction, most often using virtual machines (VMs).

The orchestrator maintains a constantly updated mapping of resource allocation. It allows a quantity of resources to a dedicated abstraction without mastering the actual use that will be made of this resource (difference between the number of cores allocated and the actual processor usage rate).

The orchestrator intends to optimize the cost of ownership for the cloud provider. However, without control over the client application, it is not possible to optimize the actual use of resources. This has two consequences:

- Data centers heat up to run servers that are less than 50% busy;
- The purchase cost for the customer is not optimal – let's also dispel the myth that 'the cloud is cheaper'. It is more convenient,

quicker to implement and simpler, but not cheaper than optimized architecture for a specific use case.

THE VIRTUAL MACHINE, THE ULTIMATE ABSTRACTION OF THE CLOUD, ONCE AGAIN A STORY OF COMPROMISE

The abstraction used until now was the VM. It has the advantage of providing excellent memory partitioning and fine-grained resource reservation. But it also has limitations: it does not support 'borderline' uses (memory or computation) properly, and the (mille-feuille) multi-layer applications (hypervisor, emulator, VM, VM OS, and middleware) which prevents an optimal use of resources and complicates diagnosis.

The main advantage being for this type of architecture to meet computing needs adequately in 'average' cases, yet poorly in 'extreme' cases. For these cases, specialized architecture had to be designed:

- For memory, through large coherent memory systems, otherwise known as 'scale-up' or SMP servers;
- For computing, for which optimized servers have been designed, e.g. to handle FP64.

The first gave rise to numerous generations of mainframe servers, associated with a specific operating system (System z, GCOS) or a specific application (e.g. SAP HANA for databases). This type of architecture prioritizes transaction consistency over computing performance.

The second type is used for high-performance computing (HPC). It first led to the design of ultra-dense CPU-based machines and then saw the arrival of more specific accelerators (manycore, reprogrammable FPGA components, GPU graphics processors).

THE CRUCIAL IMPORTANCE OF NETWORKING IN ARCHITECTURES

This architecture also required the development of new low-latency networks capable of managing congestion.

Today, the network is everywhere. It allows centralized storage and guarantees access to data at any point in the infrastructure. It ensures communication between GPUs and relieves central processors of multiple memory loads. It also ensures that data is transported securely, segregating usages from users.

It is the true backbone of modern architecture. Without a suitable network, it is not possible to deliver use cases effectively.

In 2025, cloud computing seeks to unify all uses. At the same time, supercomputing which has been 'haute couture' until now is seeking to become industrialized. As in geopolitics, boundaries are blurring, and hybrid architecture is now essential.

Until now, none has been able to orchestrate resources and distribute computing tasks efficiently at the same time.

Containerization is changing this and provides an abstraction capable of gathering different use cases. Containers are portable. They can be orchestrated as well. Task execution can be planned. And they can also have large memory capacities.

Modern hybrid architecture combines container abstraction with different types of computing islets:

- General purpose;
- Intensive computing;
- GPUs;
- FGPA, QPUs, SMP machines...

All this is achieved by using the network to ensure high-quality exchange between storage and computing, enabling direct access to components (direct GPU, or even disaggregated components) and performing part of the data processing overtime.

Some islets are responsible for ensuring data persistence and therefore its protection, while others must be able to serve as transient computing devices without persistence, yet with recovery capabilities. These principles pave the way for Cloud to extend to Edge.

AI IN AMBUSH, ACTING AS A CATALYST FOR CHANGE

The advent of AI – with mixed use cases involving supercomputing (model running), cloud computing (model inference) and SMP (databases) – is accelerating the adoption of these hybrid architectures.

The Sauron of data centers may not be that far, even if it is not a matter of a single ring but about collaborative rings contributing to shape the future of our IT architectures. ■

NVIDIA: DARING TO PIONEER

Jensen Huang
CEO, NVidia

NWe met Jensen Huang, CEO of NVidia, to share his vision of new digital technologies, about their uses, future developments and their impact on society. We also discussed how NVidia will strive to reinvent itself and respond to these changes..

AI: A FOUNDING DECADE AND A FORTHCOMING REVOLUTION

Over the past ten years, we witnessed the birth of modern AI. It evolved from perception AI to generative AI, then to reasoning AI, and today we come to the cusp of physical AI. Each step expanded what computers could do, transforming industries from healthcare to manufacturing as well as research, thus creating new applications the world had never seen before.

Looking ahead to the next ten years, we embrace a new industrial revolution. AI will be everywhere: in every business, every society, and even in our daily lives. Just as electricity became the key infrastructure of the previous industrial revolution, 'AI factories' will be the infrastructure of the next decade. AI factories will power the world's industry supplying global industries with 'intelligence', as a new essential commodity.

And the most exciting thing is, every one of us will be supported, augmented and enhanced by AI in the future. This is the greatest equalizer in history. For the first time, a new technology offers the possibility of elevating all individuals, equally and simultaneously everywhere in the world. While technology was difficult to use in the past, AI is now changing the game. Everyone can benefit.

HPC AND AI: DRIVERS OF UNIVERSAL TRANSFORMATION

High-performance computing (HPC) and artificial intelligence are revolutionary technologies indeed. They do not just make existing processes faster; they create entirely new application industries. Agent-based AI, reasoning AI, robotic AI... These are just some of the technologies set to transform all sectors, from healthcare to education, mobility to manufacturing. Every industry will be revolutionized.

Today, some people are concerned: if we automate too much, won't productivity simply translate into fewer jobs? History shows us the opposite. Over the sixty past years of the computer era, when productivity has increased, employment has also increased. Why? Because when we come up with more ideas for a better future, productivity is what allows us to realize them.

AI is not about protecting a few individuals or countries; it is for everyone. It will create new jobs, new businesses and new industries. It will advance society at large. For the first time, our industry developed as useful as universal a technology meddling every field, from the arts to the life sciences, from customer service to the physical sciences, from media to manufacturing. Combined, AI and HPC represent not just a new wave of innovation, but the beginning of a whole new industrial revolution.

THE FUTURE WILL BE QUANTUM... AND HYBRID

The whole industry recognizes that the right path is quantum-classical hybrid computing. Far from the image of an isolated quantum computer bedded in a corner of a laboratory, the magic happens when you connect it to a GPU supercomputer, the GPU system doing all the control, the error correction and operations orchestration.

Progress made in error correction is groundbreaking which opens decisive prospects. And I can tell you that all the high-performance computing centers I have encountered are turning to the quantum-classical hybrid approach without exception.

The future will therefore be quantum with a major impact but always maintaining symbiosis with classical computing supported by GPU supercomputers.

NVIDIA: DARING TO PIONEER

Jensen Huang
CEO, NVidia

TOWARDS PHYSICAL AI AND GENERALIZED ROBOTICS

Everything that moves will one day become autonomous. We are entering the era of physical AI, an AI that understands the laws of physics, gravity, inertia, cause and effect, the permanence of objects, and clear common sense applied to the real world.

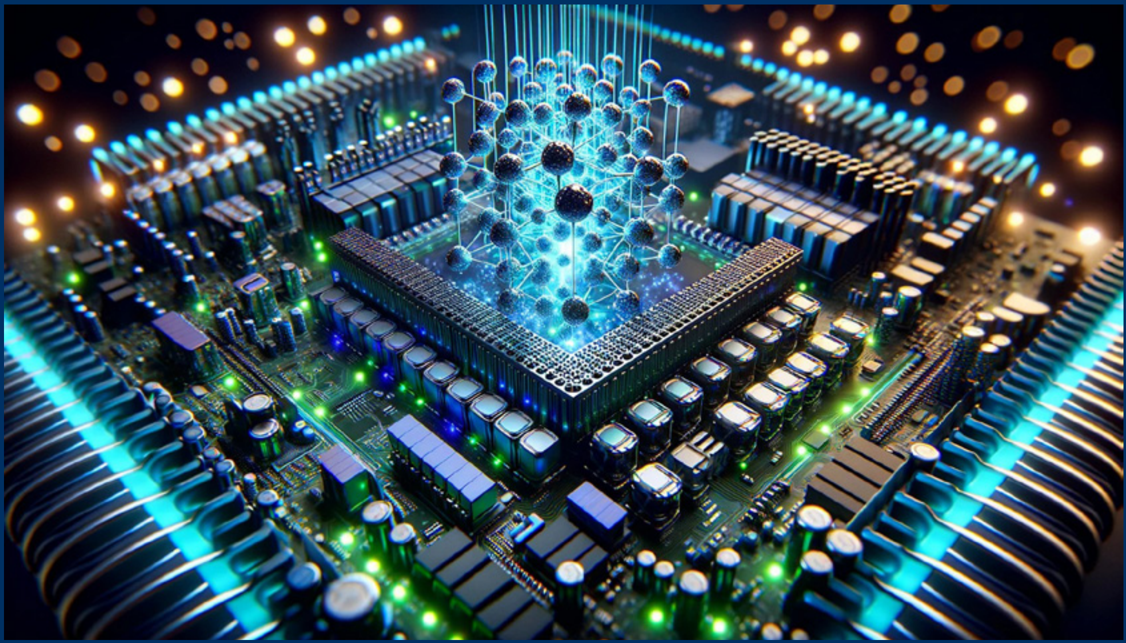
This means that all manufacturing companies will need two factories. For example, a car manufacturer will have one factory to physically produce its cars and an AI factory to create the intelligence to power them. This idea of two factories, one for the physical product and the other for its digital intelligence, is the future of all industries.

In the context of global labor shortages, with a shortfall of 30 to 40 million skilled workers, robotics will take over. It will enable industries to develop for economies to grow and societies to prosper. That is why I believe the 'ChatGPT moment for Robotics is imminent, just around the corner. Technology is in place – simulation, training, testing, real-world deployment – and we will see revolutionary advances in general robotics that will surprise us all very soon.

NVIDIA AND THE ART OF DARING TO PIONNEER

NVidia has always stood out for its willingness to take on challenges that no one else dares to tackle. This is how we approach every breakthrough, with no fear yet with curiosity and conviction.

So, asking how Nvidia will reinvent itself for the next era – whether it's physical AI, robotics or hybrid quantum-classical computing – the answer is simple: exactly the way we always have, remaining true to our founding spirit – daring to be first, making the right choices with significant contribution enough to give meaning to its teams and advance the world. ■



NVIDIA Quantum Cloud © NVidia

THE CHALLENGES FOR STORAGE: EVOLVING IN A FRAGMENTED LANDSCAPE AND INNOVATING IN THE AGE OF AI

Jean-Thomas Acquaviva
Team Leader Research Group, DDN Storage

The digitization of the world continues, driven by an insatiable thirst for Artificial Intelligence and the widespread use of data acquisition devices: from IoT to smartphones, connected cars and large scientific instruments. This continuous growth in data is reshaping the landscape of storage technologies, challenging established paradigms and standards. The coming decade promises a revolution in the way we store, access and process data. A fundamental transition is underway, from a human-centered, file-based management approach to a more fragmented continuum focused on machine-to-machine communications.

Historically, storage systems were human-centric, organized around file systems and the POSIX standard. However, those days are over. Object storage has already captured a significant share of the market, and we anticipate growth in machine-to-machine flows with a proliferation of specific data formats. This transition from a standardized to a fragmented world requires a reassessment of storage strategies. Following the saying 'software is eating the world', it is highly foreseeable that software will eat storage and occupy a pre-eminent position in all future data solutions.

CHANGING TECHNICAL AND MARKET DYNAMICS

Storage substratum technologies are constantly evolving yet not disappearing; they are stacking up. While flash drives are replacing hard drives in terms of speed and density, hard drives remain a viable solution for cold data. Furthermore, there is a growing market for Zero-energy storage, today using magnetic tapes, tomorrow using hard drives or even DNA. The industrialization of persistent memory devices temporarily halted by the failed Optane experiment will resume soon. All these technologies will remain invisible to end users, as cloud storage has become the dominant force, presenting only an abstraction to applications.

EMERGING TRENDS AND DATA CONTINUUM

Among the major trends expected over the next ten years is integrated edge storage. Another upcoming change is the gradual blurring boundary between storage systems and network technologies thus creating a 'data continuum.' A crucial concept for this continuum is data gravity, whereby data attracts data and related services thereto. Effective management of data gravity requires innovative storage solutions capable of managing and transferring data in increasingly distributed environments. Data gravity also drives storage disaggregation through computational storage, i.e. the integration of computing capabilities within storage devices.

The full implementation of the continuum requires efficient and robust data logistics. However, this is not solely a technical issue; it requires organizational changes that represent a significant obstacle to its development.

SECURITY AND SAFETY IN THE POST-EXASCALE ERA

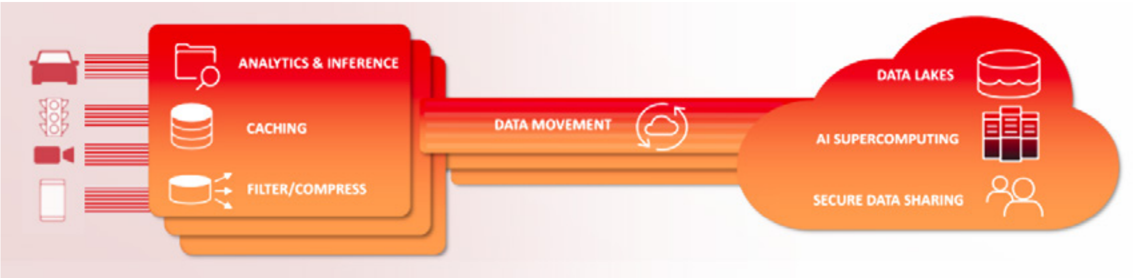
As we move towards an increasingly interconnected world, cybersecurity reveals paramount. Protecting data from unauthorized access and cyber threats requires effective cybersecurity strategies, using encryption, rigorous access controls and advanced intrusion detection mechanisms. This 'cybersecurity for an open world' is an essential component of the data continuum.

POST-EXASCALE AND DATA CONTINUUM

The symbiotic relationship between AI and storage is also deepening: 'Storage for AI, and AI for storage.' AI-based monitoring and analytics tools will become crucial for managing and optimizing complex storage systems. These tools will enable real-time data analysis, predictive maintenance, and automated problem resolution significantly improving reliability.

Our vision for the post-exascale era of storage solutions is one of a fragmented landscape characterized by the absence of universal standards, where flash memory will play a predominant role. We anticipate a fluid data continuum, where data gravity urges proximity to computing, with advanced, native integrated cybersecurity. This continuum implies a certain loss of uniqueness for the high-performance computing sector, with a need for integration into a broader and more diverse technological ecosystem.

By promoting collaboration, adopting open benchmarks and breaking down barriers between AI start-ups and the HPC community, Europe can stimulate innovation. With their rich scientific and industrial ecosystems, France and Europe have significant resources at their disposal: data generation, human expertise and first-rate infrastructure. The capability of our institutions to act in the long run, as shown by Teratec celebrating 20 years of activity, puts us, for once, in a leading position in this new race. ■



THE FUTURE OF MASSIVE DATA STORAGE FOR HPC AND AI

Gary Grider
Deputy Division Leader MS B260 HPC, Los Alamos National Laboratory

There have been some recent changes in HPC storage but there is little one would classify as breakthrough. Bandwidth and capacity advances in flash storage are putting pressure on magnetic storage markets in the performance area. New long term storage media types like ceramics and DNA have emerged. Of course, object storage is a big business as well and is looking more like file storage in many ways, humans seem to need hierarchy.

THE GROWING IMPORTANCE OF DATA

Mostly, the change came from the market swing toward data centric computing. With the advent of AI, it's not just classical HPC sites that need scalable storage technologies, everyone engaging in AI needs scalable storage solution. This has led to more commercial scalable storage systems with proprietary interfaces. AI training use cases have made compute node local storage become popular. In many ways this data centric economic trend has hurt open community solutions.

The biggest surprise is the emergence of Parallel NFS (PNFS) which was started in the 2002-time frame to attempt a common parallel file system client for all operating systems. Slow community and industrial progress and is now within striking distance of being near best in class for many scalable workloads like many writers to many files and one file, scalable metadata, read storms, distance, etc. The PNFS community is even tackling client erasure with a new approach that should make client unaligned erasure writing as efficient as server-side erasure, ending needing fail-over server pairs. There are four companies selling or working actively on PNFS. If the PNFS community is successful, a state-of-the-art scalable storage system including clients, metadata servers, and data servers could be available via the open community and distributed with your operating system.

TOWARDS PROCESSING CLOSER TO DATA STORAGE

Another important new capability is pushing processing to near the storage media. Indexing, vector embeddings, and other data centric processing is more efficiently done near the data storage. Many solutions are leveraging the rich Apache Analytics ecosystem which includes common formats for at rest, in memory, and in-flight data formats as well as standards for query pushdown techniques for reductions near storage.

Given the uncertainty the next decade holds regarding the use of information science to change our world, workload uncertainty, and the new world order of cloud and AI factories being the volume drivers in the computing, networking, and storage industries, two directions seem obvious. The broader HPC/AI community needs alternatives to just using cloud/factory resources and commercial proprietary tools. Having open source/standards-based community solutions and leveraging large open ecosystems seems obvious. Since the future relies on data, both existing holdings and new data, rich and security aware tools for adding and managing value of the data is also clear. ■

THE DNA, MEMORY OF THE FUTURE

Marc Antonini
Research Director, CNRS

Humanity's memory depends on our ability to manage ever-increasing volumes of data. By 2040, several thousand zettabytes (10^{21} bytes) will need to be stored for periods ranging from a few years to several centuries. Such a mass of information represents an unprecedented challenge: current storage technologies are already proving insufficient. Their widespread deployment in data centers would require considerable energy resources and raise serious issues related to carbon footprint, maintenance and rapid obsolescence of equipment.

In this context, one of the most promising alternatives is molecular storage, particularly synthetic DNA. This molecule has unique advantages: extremely high storage density, stability over very long periods of time, and relatively simple storage conditions. It offers a credible path for designing sustainable, energy-efficient solutions capable of meeting major data preservation needs of tomorrow.

The MoleculArXiv exploratory PEPR is fully in line with this approach. Its objective is clear: over a five-year period, develop the research and infrastructure needed to increase the current speed of the DNA writing/reading cycle by a factor of 100, while reducing its cost by a factor of 100. Achieving this goal would represent a decisive step forward to the industrialization of molecular storage.

This program, being part of France 2030, has been entrusted to the CNRS for scientific management, with the ANR acting as operator. With a budget of €20 million over seven years, it brings a multidisciplinary scientific community together with expertise in chemistry, biology, computer science, microfluidics and nanotechnologies. Four targeted projects were launched at the end of 2022, laying the foundations for this new technology. In 2025, three new projects selected through a call for proposals further enriched this dynamic and broadened the spectrum of research being conducted.

While DNA is the focus of MoleculArXiv due to its information density and proven stability, the program is also exploring alternatives, including synthetic polymers. These could offer additional advantages, such as increased writing speed or better scalability for certain specific applications. This dual approach focused on DNA while opening the door to other polymers, makes it possible to address a wide range of scientific and technological challenges.

By bringing teams together from a variety of disciplines, MoleculArXiv aims to accelerate the emergence of a disruptive technology that will permanently transform the way we store information. This program aims to build French leadership and position national academic research at the highest international level in a field that is strategic for the future of the digital society.

VIRTUALISATION AND ORCHESTRATION

Guillaume Colin de Verdière
Director of Cross-functional Digital Simulation Skills Programme, CEA

Gilles Wiber
Deputy Head of Department, CEA

Jean-Philippe Nominé
Fellow, CEA

Over the past ten years, pressure from users and cloud technologies has contributed to a change in how supercomputers are used. Machines are no longer seen as monolithic but as more flexible sets of resources. As a result, virtualization and orchestration software are increasingly used on supercomputers.

VIRTUALISATION AND CONTAINERIZATION

Virtualisation (and its variant known as containerisation) makes applications independent of hardware resources and/or their system environment. Virtual machines replicate an entire physical machine. Containers, which are lighter, embed an application and its software dependences; once built, they can then be deployed on various machines. These approaches offer numerous advantages for users but also for system administrators.

For an administrator, these approaches make it possible to test operating system upgrades or reserve all or part of the machine resource for experimentation without disrupting production. They also make it possible to contain user tasks in a secure environment, only by exposing the minimum resources required or even granting increased rights to resources locally. In the case of specific hardware (particularly proprietary graphics or network cards), it will be necessary to make specific adjustments with ad hoc drivers.

From a user standpoint, placing oneself within a container (such as Docker or Kubernetes) allows to run an application in a fully controlled environment, regardless of the versions of the hosted native system of the host machine. Users will have guaranteed data protection and the ability to move their code from one machine to another depending on their needs or on the resources allocated, with performance still warranted.

These methods are deployed on the CEA/DAM's EXA1 supercomputer, using the open source PCOCC* (Private Cloud On a Compute Cluster) software. PCOCC can manage and instantiate either containers or virtual machines with the same user interfaces. The introduction of containers also makes it possible to isolate computational tasks from one another, thereby increasing the security of the computing center (isolation can be achieved down to the network level). This work will be pursued for implementation in the future French EuroHPC exaflop computer (Alice Recoque) to be installed at CEA/TGCC.

ORCHESTRATION AND WORKFLOWS

Furthermore, increasing simulation-based studies and big data analyses require diversified processing phases, often linking different calculation codes with pre- and post-processing steps which may alternate. Going well beyond the rigid resource allocations and scheduling of traditional HPC, **orchestration** of these different processes leads to complex *workflows*;

* https://github.com/cea-hpc/pcocc

numerous tools and APIs are being developed for this purpose, but no major standard has emerged yet in this area. *Workflow* managers can feature *low-code, no-code* graphical interfaces describing all the desired processes, as for a digital twin for instance.

Virtualisation and orchestration clearly contribute to the democratisation of HPC usage as it will no longer need a specialist in supercomputers and their environment to be productive, once a suitable abstraction of the computing center and its resources will have been defined and intrumented.

Communities of AI researchers and neural network users who do not necessarily come from the traditional HPC world, as well as industrial are eagerly awaiting these methods. To meet their needs, the use of virtualisation and orchestration software is becoming a necessity to provide them with flexibility, efficiency and security. ■

In this allegory, the walls of a wide kitchen show the circuits of a computer. The shelf contains mini-kitchens (virtual machines with their utensils represent virtualization tools, above low-level resources) and meal boxes (ready-made meals represent applications bedded in containers). At the center of the room, an orchestration operator coordinates the sequence of activities – different modules – so that overall tasks are performed in a consistent manner. (Image generated by Gemini AI).

HPC AND EXASCALE: POWER PUT TO TEST SOFTWARE COMPLEXITY

Luc Giraud
Research Director, Inria Center at University of Bordeaux, Airbus CR & T, CERFACS

Brice Goglin
Research Director, Inria Center at University of Bordeaux, LaBRI

The arrival of the first exascale machines in 2022 is the culmination of a long technological race that began in the early 2000s. From the outset, designers of these machines understood that the main obstacle would be energy. This dictated radical choices: graphics processing units (GPUs) replaced traditional processors (CPUs) for their superior performance; the number of computing units exploded; reducing data movement became an obsession, as transporting data between two nodes can now cost more energy than recalculation in the moment.

EVER MORE COMPLEX PLATFORMS FACING STRATEGIC CHALLENGES

The transition to exascale computing has been accompanied by a profound transformation in hardware architectures. Modern computers are based on hybrid combinations of CPUs and GPUs, sometimes using heterogeneous memories...

Software stacks are often based on NVIDIA's CUDA environment. As alternatives are not mature enough yet, there is a potential risk of technological dependence on a few players about to weaken the strategic autonomy of industry and governments.

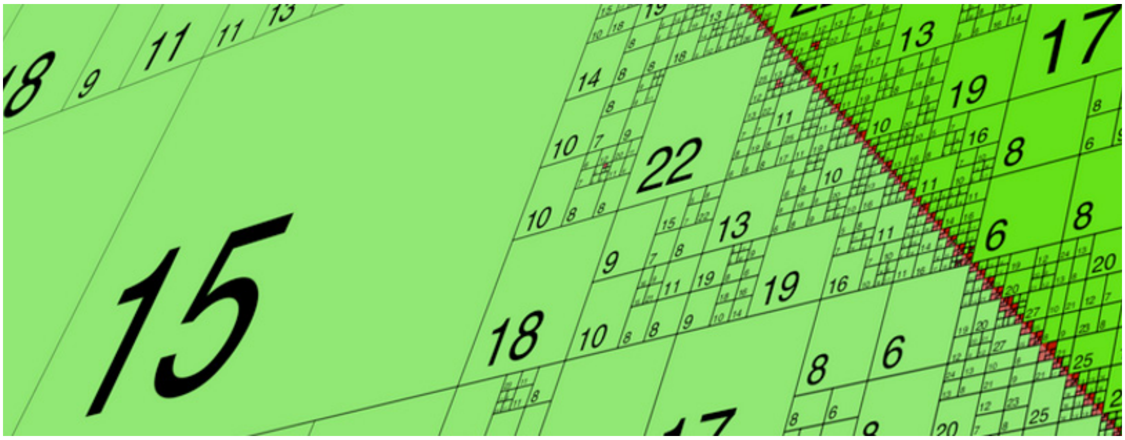
The ecosystem is struggling to keep up: software components are numerous, heterogeneous, and often specific to certain hardware. This undermines the portability of applications. This risk is partially offset by the extension of machine lifespans derived from environmental and financial reasons, but in return requires the stability of all software components – compilers, libraries, runtime environments, applications – to be guaranteed over extended periods, despite their asynchronous evolution.

Software standards are still a long way off. The venerable MPI protocol remains the essential common foundation for communication between nodes. However, no consensus has emerged yet for detailed management of tasks or execution graphs. Initiatives such as OpenMP are progressing, while models based on task graph programming attract growing interest. Although still complex to implement, they foresee a possible future standardization that could facilitate independence from manufacturers' proprietary solutions.

SCIENTIFIC APPLICATIONS IN THE MIDST OF CHANGE

Applications operating these supercomputers are becoming increasingly complex. Today's numerical codes incorporate multi-physical, multiscale, and sometimes temporal phenomena making four-dimensional simulation a necessity. In such fields as quantum mechanics and astrophysics, models even explore much larger state spaces.

Another major transformation: the emergence of uncertainty. Large simulations increasingly integrate uncertainty quantification or data assimilation approaches aiming to cross-reference numerical



Zooming in on a hierarchical matrix - University of Bayreuth, Germany

calculations with experimental observation data. A single deterministic exascale calculation is generally of little interest, as massive realization of varied sets of calculations becomes the norm.

This sophistication comes up against another often intrinsic difficulty: the irregularity of the data, hence of algorithms intended to handle it. The computing load is no longer distributed evenly among the processors; some nodes wait while others get bogged down in complex tasks. Dynamic load balancing, work diversion from one node to another, or even intelligent scheduling of task graphs resulting from a high-level expression of modern algorithm implementation are becoming essential elements of efficient computing.

ARTIFICIAL INTELLIGENCE: A FORETOLD REVOLUTION IN HPC

The emergence of artificial intelligence (AI) in the world of scientific computing is undoubtedly the most unexpected transformation of recent years. Hardware accelerators designed for AI require new digital formats (FP16, BFLOAT16) which are much more energy efficient than the classic double precision (FP64) favored by engineers and physicists. Ultimately, it is possible that these formats will become dominant, if we except calculations where extreme precision remains essential.

Beyond hardware, AI also promises to accelerate computation itself: by building scale models capable of approaching costly simulations, optimizing numerical solvers, or by dynamically predicting the best distribution of tasks. However, such uses remain precocious for the moment, even though promising in the field of large-scale industrial simulations where mathematical explainability may still prevail over AI approaches.

BEYOND POWER: THE SILENT CHALLENGES OF EXASCALE COMPUTING

Exascale computing isn't just a simple upgrade of existing computers; it's a complete shift in the way scientific computing is designed, programmed, and used. This silent revolution requires researchers, engineers, and decision-makers to anticipate technological choices that will impact industrial and scientific capabilities for decades to come. ■

HPC AND ARTIFICIAL INTELLIGENCE: INNOVATION AND CONVERGENCE

Gabriel Antoniu
Research Director, Scientific Lead of the KerData Team, Inria

Bruno Raffin
Research Director, Scientific Lead of the DataMove Team, Inria

Olivier Beaumont
Research Director, Scientific Lead of the Topal Team, Inria

A Artificial intelligence (AI), and particularly deep learning, has transformed the field of high-performance computing (HPC). Once distinct, these two disciplines are now converging to address increasingly complex scientific and industrial challenges. This convergence is highlighted in strategic agendas for European research and is currently the subject of rich discussions within international working groups such as InPEx. HPC, traditionally based on the numerical resolution of complex equations such as partial differential equations, is taking advantage of advances in AI to accelerate calculations, optimize models, and explore new computational paradigms. Conversely, AI relies on HPC infrastructures to train increasingly large models, requiring large-scale computing power and data management. This article discusses recent developments at the intersection of these fields, highlighting the methodological approaches, algorithmic innovations, and infrastructure developments that are redefining the landscape of scientific computing and its applications.

DEEP LEARNING AND DIGITAL SIMULATION

Neural networks (NNs), at the heart of AI, transform the numerical simulation that underpins high-performance computing (HPC). Unlike traditional methods which approach the solution of complex equations, such as partial differential equations (PDEs), through discretization and iterative calculations, NNs learn from data to construct functions.

NNs are combined with classical PDE solvers. The solver calls the NN which implements a computational kernel built by learning from data. This call can be executed locally or off boarded to a GPU. The resulting gain in precision allows the solver to work at a lower resolution level, reducing computation time.

NNs are trained to completely replace solvers by learning from the data they produce. The resulting model can be seen as a highly efficient (lossy compatible) data compressor, coupled with interpolation capabilities. Compact and fast, it can be used to interactively explore simulation data or to be coupled with data sensors as a digital twin. It can also be used for parameter space exploration at a reduced cost, inaccessible with the original solver, which is reused to validate promising solutions being identified by the NN.

The architecture used, such as convolutional networks or transformers, often comes from classical AI, but specific models such as PINN (Physics-Informed Neural Networks) or FNO (Fourier Neural Operators) are also developed.

Finally, advanced techniques such as automatic differentiation integrated into neural network (NN) frameworks have enabled the development of hybrid architectures that closely combine NNs and solvers. In these architectures, the solver is incorporated into the NN and contributes to calculating the output. During training, the loss function, a central concept of NNs, must be deviated for backpropagation, a key step in learning. This also involves deriving the solver. Thanks to NN frameworks, these derivatives are calculated automatically via automatic differentiation.

OPTIMIZING OF HPC INFRASTRUCTURES FOR LEARNING

High-performance computing infrastructures are being profoundly modified by the training of large-scale artificial intelligence models. Memory, network, storage, and computational accuracy are subject to specific constraints. Architectures have undergone significant changes: graphics processing units (GPUs) have become central elements, and reduced-precision computations (FP16, bfloat16) have become widespread, allowing for a better balance between performance and accuracy. At the same time, learning algorithms integrated techniques such as dematerialization and compression to reduce memory consumption.

IMPACT ON NETWORKS AND STORAGE

Collaborative communication plays a crucial role in distributed training which often relies on multiple forms of parallelism (data, model, or computation). This places significant pressure on interconnection networks and requires specialized networks and mechanisms to overlap computation and communication efficiently. On the storage side, large models typically access large volumes of data sequentially, while the preparation phases remain demanding.

ACCEPTABILITY OF APPROXIMATIONS

Unlike numerical simulation which produces data dynamically and requires high numerical precision, high reproducibility, and regular intercrossing communications, machine learning can tolerate approximations, losses of numerical precision, truncated or compressed collective communications, and allow bolder optimizations with a greater variety of algorithmic solutions and computation-memory tradeoffs.

Ultimately, we can probably expect more hybrid architectures combining computing and data processing at the edge. Innovative approaches such as in-memory computing or neuromorphic chips could also emerge, alongside specialized systems like those from Cerebras or Graphcore. In any case, learning is already a central driver of evolution for all intensive computing.

AI AND HPC: A DECISIVE CONVERGENCE

The convergence of high-performance computing and artificial intelligence marks a decisive step in the evolution of computational sciences. Advances in deep learning, combined with the power of HPC infrastructures open unprecedented perspectives for numerical simulation, computational optimization, and the exploration of complex parameter spaces. Neural networks, whether used to analyze data, augment traditional solvers, or replace them with surrogate models, are redefining methodological and software approaches. At the same time, the adaptation of HPC architectures to the requirements of AI, with the rise of graphics processors, reduced-precision computing and techniques such as dematerialization, is pushing the boundaries of computational efficiency. In the future, the increasing integration of hybrid paradigms and the emergence of innovative technologies such as in-memory computing, should strengthen this synergy, making it possible to meet ever more ambitious scientific and industrial challenges. Such dynamics establishes AI as a central driver of innovation for HPC. ■

PROGRAM AGENCY FOR DIGITAL TECHNOLOGIES, ALGORITHMS, SOFTWARE AND USES

Sophie Proust
Executive Director of the Digital Programmes Agency, Inria

Created at the beginning of 2024 by decision of the President of the French Republic and entrusted to Inria, the Digital Programs Agency - Algorithms, software and uses aims to develop and manage research and innovation programs responding to the major scientific and technological challenges in the digital field, in particular ensuring good coordination between the various players in research and the economic framework.

Its main missions include the construction of a concerted prospective vision, the emergence of new strategic programs, establishment of high-impact strategic partnerships and supporting the ecosystem, the transfer of scientific discoveries to industrial solutions, training and improving the skills of researchers and engineers throughout their careers, and obviously coordination with European and international actions on digital technology.

The Digital Program Agency thus plays a key role in structuring and coordinating technological advances in France. As a national pillar for digital innovation, it stands out for its ability to bring together programs related to strategic areas, in support of national acceleration strategies such as high-performance computing (HPC), cybersecurity, network and cloud dual system, artificial intelligence (AI), quantum technology, virtual worlds, as well as digital health, digital serving the environment, and digital and learning.

AN OPEN AND CROSS-FUNCTIONAL MODEL

One of the Agency's main strengths lies precisely in its ability to break down barriers between disciplines, fostering rich interactions between sectors that historically often operated in silos. Uniting highly complementary programs such as high-performance computing, AI, quantum computing, and cloud, it offers a platform where synergies materialize around innovative projects.

Breaking down barriers allow us to design and develop cross-functional technological objects that combine the strengths of different fields. For example, the AI Factory France project led by GENCI gained support from the Agency, enabling a close collaboration between experts in AI, high-performance computing, and cloud.

BETTER ANTICIPATING MAJOR SCIENTIFIC DEVELOPMENTS

At a time when Europe acquires exascale supercomputers, the community must already prepare for the post-exascale era where it will be necessary to deal with specialized computing architectures that consume less energy, and where it will be critical to support uses and application combining data capture and processing, digital simulation and artificial intelligence. These forthcoming uses call for a true convergence of methods and tools implemented in such fields previously compartmentalized. This is the next challenge of tomorrow's high-performance computing!

AI CHANGES DESIGN MODE

Herbert Taucher
VP for Industrial Research & Pre-development for Integrated Circuits and Electronics in Foundational Technologies, Siemens

We asked a major industrialist also operating as publisher of design, manufacturing and operating software, about how new digital technologies impact their industrial activities and how they integrate those into their software solutions to increase automation.

Siemens is a major international industrial group, employing more than 300,000 people, including 50,000 in R&D, and offering more than 10,000 products in its catalog. Many of these are increasingly intelligent hardware, thanks to the software driving them. But a wide range of software for design (mechanical, electronic, software), manufacturing, and the operation of industrial activities is also part of our solutions.

Having worked as a design engineer for many years, I am now heading the group's electronics business, from fundamental research to design software (EDA). We operate in conjunction with the group's other activities to produce multi-technology systems. I also advise the group's board of directors on electronics acting as the interface on these matters for the European Commission, local governments, and industry associations.

AI, and even more so Generative AI will change the way Siemens products are designed as well as the associated design software in all areas. This is especially true as we increasingly use digital twins from design to operation, and we must address new parameters such as cyber threats and sustainable development. Our customers also expect products that can evolve throughout their lifecycle. We must therefore integrate specifications, functional structure, verification, code, 3D, digital simulations... into digital design twins. Then, we offer these solutions as scale models to be executed in runtime to simplify and optimize the operation of our finished products.

SOFTWARE-SPECIFIED HARDWARE

To meet customers' sustainability expectations, we are increasingly developing «software-defined hardware» ie., hardware designed so that core functionality relies on software control, making it scalable over time to offer more services to users.

This requires a better understanding of how use cases are evolving to define degrees of freedom that may interest our customers in the future. These are documented guesses, but the richer your data sources and models, the better you anticipate potential new features.

AI ACCELERATES DESIGN

Given the scale of issues to be addressed, designers must rely on data and a modular set of integrated models, helping them not to find local optima but one global optimum.

AI and HPC should help us achieve this. AI has already been introduced into several design software programs, but with very different levels of maturity. This ranges from innovative ideas to technology demonstrators, often not tested yet on a large scale. The question remains whether it will be adapted to the complexity and scale of problems, in terms of capacity and performance.

Learning mechanisms already makes it possible to characterize product's parameters in a multidimensional manner to create a scale simulation model just as effective, yet less computationally intensive and faster to execute. We have thus reduced by three orders of magnitude the number of Spice simulations required to optimize an electronic product. And the same goes for the characterization of standard libraries or analog simulation. These are the very first applications to benefit from the contribution of machine learning technologies and AI.

We develop design tools using Generative AI, knowledge graphs, graph neural networks (GNN), large language models (LLM)..., for the entire Siemens PLM portfolio. Such technological update of expert systems is made possible by the explosion of computing power

AI CHANGES DESIGN MODE

Herbert Taucher
VP for Industrial Research & Pre-development for Integrated Circuits and Electronics in Foundational Technologies, Siemens

allowing models not only made of a few million but of a few billion parameters.

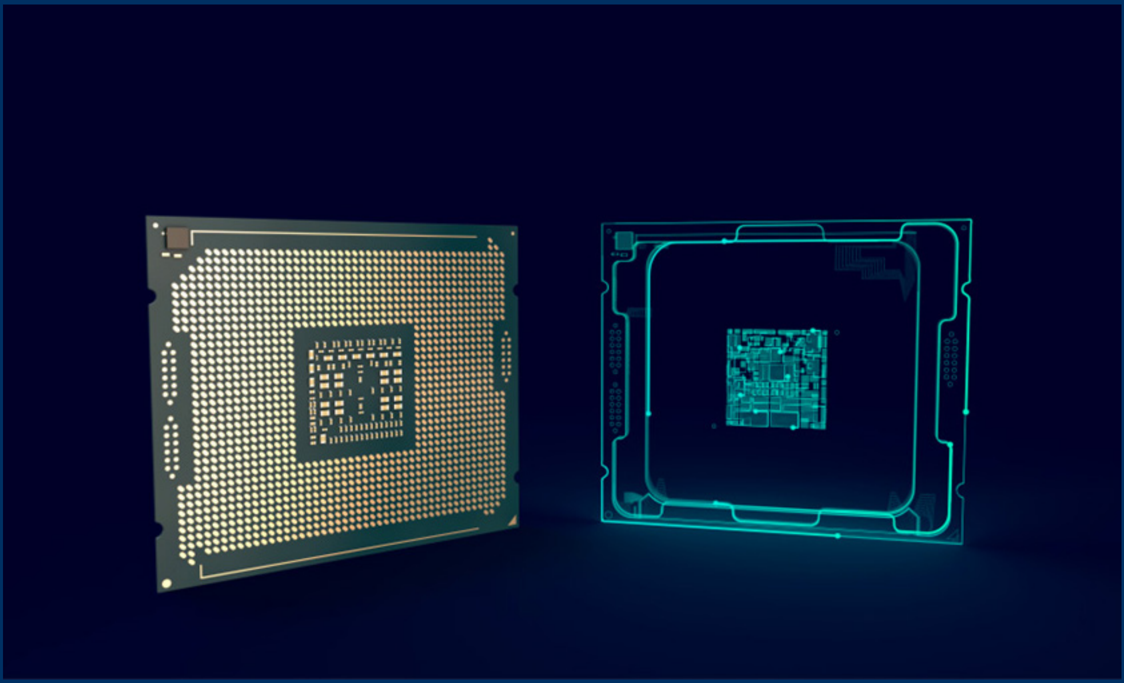
This kind of automation will also help us address future withdrawals of design, manufacturing, and operations staff, as well as the shortage of experts affecting our industries. This is a priority for Siemens as a company, but also for our customers.

That's why we introduced co-pilots into our software, not just for design but also for product manufacturing and operation. This increases user expertise and productivity.

QUANTUM MUST FIND ITS PLACE

Quantum computing seems interesting for certain uses (logistics optimization...) by boosting productivity and quality of results, but it comes in competition with other technologies (neuromorphic...) that are generally less expensive. This remains a subject of research in quick progress, particularly around the classical/quantum mix. There is still a long way to go, both in terms of hardware and software rewriting as in understanding proper use, but we will experience a comparable evolution like Generative AI and the arrival of LLMs.

Ultimately, all these technologies should allow us to greatly automate design processes, in all areas, starting with around twenty lines of specifications and a few key standard values. A designer will focus more on usage scenarios of their product as well as their evolution over time, yet retain decision-making power. ■



Digital Twin Semiconductor image source Siemens

CALCULATION, A CENTRAL CONCERN FOR THE ASIC PROGRAMME AGENCY

Jean-Philippe Bourgoin
Director of ASIC Programme Agency for Digital Components, Systems and Infrastructures, CEA

Announced by the President of the French Republic in December 2023, the program agencies, including the ASIC agency entrusted to the management of the CEA, were set up at the beginning of 2024 with mission in agency's domain of building a strategic vision, proposing and implementing national research programs, more generally, monitoring all existing programs and mobilizing the national academic community, liaising with industry.

Computing, whether at the component, system, or digital infrastructure level, is a subject of particular concern to the ASIC Agency. It plays a decisive role in our information and communication societies, and its mastery is more strategic than ever. Indeed, geopolitical tensions, as well as the profound scientific and technological transformations underway, notably the deployment of artificial intelligence (AI) and the race for quantum, are profoundly challenging established alliances, development dynamics, and value chains.

For example, while Europe and France remain very well positioned in embedded AI, they are totally dependent on high-power AI components: this is why the ASIC Agency has proposed two programs to the State. The first on component design – in connection with the European design platform being set up as part of the chips act – aims to develop the design of components based on 3D microelectronic integration, and tools using generative AI. The second, jointly with the Agency led by Inria, aims to develop ultra-efficient AI components through hardware-software co-design, taking advantage of breakthroughs such as computing in memory, neuromorphic approaches, photonic interconnections and 3D integration, to get back in the race within an 8-year horizon.

The Agency is also implementing a reflection on developments in AI and quantum computing infrastructures from edge to cloud, and actions to be implemented so that France and Europe retain sovereign control.

From analyses carried out by the Agency and beyond scientific, technological and industrial issues to be addressed, it is already clear that expanding the national pool of academic and industrial skills is an essential component.

In this context and over 20 years, Teratec has initiated a remarkable path to follow in terms of awareness, acculturation, structuring and facilitating access to computing in all forms. ■

USES

INTRODUCTION:
USES: FROM INDUSTRY TO HUMAN BEINGS

Jean-François Prevéraud
Journalist

Initially, digital simulation technologies were contained to industrial uses (space, aeronautics, automotive, energy…), for calculation of structures or fluid flows. These sectors were at the origin of major design software and calculation codes (Ansys, Catia, Fluent, Nastran…). These are the tools for CAD (Computer-Aided Design), CAE (Computer-Aided Engineering), Fluid Mechanics, Thermal Simulation…

For a long time, these specific tools to Engineering Sciences were limited to product validation ending the design phase, because low computing power imposed prohibitive calculation runs. This led to long and costly iterations in the event of design changes.

FROM VALIDATION TO PREDICTION

Gradually, digital technologies have been investing in other fields such as nuclear and defense, which have also developed their own solvers. Then these digital technologies finally left the design offices focusing on production and management: CAM (Computer-Aided Manufacturing); CAPM (Computer-Aided Production Management); ERP (Enterprise Resource Planning)… Research and the world of finance have also begun to show an interest.

With computing power becoming more widespread and off-the-shelf software offerings expanding, all sectors of activity have finally benefited from suitable digital simulation tools. This has also enabled a shift in usage, from validation to prediction. Today, we simulate before designing, manufacturing a product, or launching a service. And the arrival of HPC and Generative AI reinforces this approach.

UPON GENERAL PUBLIC

While major industrial sectors are still the major users, a multitude of other sectors have also adopted digital simulation tools. Life science specialists use them to explore the living nature (basic research on diseases and epidemics, drug development, digital clinical trials, human digital twins, personalized medicine, neuroscience…).

Similarly, Earth science specialists make extensive use of digital technologies to study the earth, water and air, to find new resources, predict and anticipate natural phenomena, better manage the environment, predict climate change, facilitate sustainable development…

And we now also find digital technologies in materials science, culture, heritage, land management… with more consumer applications thanks to 3D and augmented or virtual reality technologies.

Current and future developments are presented to you by the best international experts in the following pages. ■

DIGITAL BIOLOGY TO EXPLORE AND PREDICT THE LIVING WORLD

Marjorie Domergue
Project Manager, DIGIT-BIO Metaprogram, INRAE

Hervé Monod
Director, DIGIT-BIO Metaprogram, INRAE

Carole Caranta
Deputy Executive Director for Science and Innovation and Co-Director of DIGIT-BIO, INRAE

In recent years, we have witnessed a real explosion of data in biology, thanks to new data acquisition technologies and the rise of open science.

Research in the life sciences has been profoundly impacted by this influx of data and by advances in the tools developed for their analysis (modelling, artificial intelligence, high-performance computing). Digital biology is now making it possible for us to develop more integrative approaches to biological systems and explore them in all their complexity through the multilevel integration of different spatial scales.

Against this background, the mission of the DIGIT-BIO metaprogram[1], launched by INRAE in 2021, has been to support interdisciplinary research at the interface between the formal and life sciences, seeking to improve our understanding of the behavior of biological systems, predict their evolution in changing environments, and facilitate their management. The projects described below illustrate how digital and data science now permeate research in the life sciences. Future developments will make even greater use of methodologies based on AI and generative AI, also drawing on digital twins[2] as a novel framework to improve our ability to act on systems by mobilizing their digital representations.

USE OF AI FOR GENETIC SELECTION IN PIGS

The OBAMA project combines artificial intelligence and genomics to improve pig breeding by predicting the effects of genetic mutations on key traits such as growth, behavior, and meat quality. Genomics has revolutionized animal genetic selection. By sequencing the DNA of large numbers of animals, thousands of genetic variants associated with traits of interest, such as fertility or behavior, can be identified. However, most of these variants are located in non-coding genomic regions, making it difficult to understand their true effect. To meet this challenge, two researchers from INRAE's Occitanie Toulouse center – Julie Demars, a geneticist, and Raphaël Mourad, an expert in Deep learning – created the interdisciplinary OBAMA project, with a two-fold objective:

- Develop new approaches to deep learning capable of accurately identifying causal variants of interest for traits, through the orthological augmentation of data from the genome sequencing of several species.
- Experimentally validate the prediction of the phenotypical effects of variants obtained using these models on a given trait of interest.

The project draws its data from a pioneering pig genetics program (PorcQTL) to crossbreed Large White with Meishan

1. https://digitbio.hub.inrae.fr/
2. https://digitbio.hub.inrae.fr/thematiques/vers-des-jumeaux-numeriques

Pigs. The data collected includes classic traits, such as weight and meat composition, alongside behavioral traits, such as docility and maternal attention.

Genome-wide association studies (GWAS) make it possible to identify regions of the genome that are statistically linked to these traits. Each region can contain thousands of variants of which only a few play a causal role. This is where deep learning steps in, making it possible to predict the impacts of mutations, not directly on the phenotype, but on certain molecular phenomena such as gene transcription and chromatin activity... The traditional supervised learning approach is constrained by the limited amount of functional data available from pigs. To overcome this limitation, OBAMA proposes a data augmentation strategy based on orthology, inspired by image processing techniques: the large number of mammalian genomes that have already been sequenced but not yet annotated make it possible to considerably enrich training sets to develop more robust, accurate models. Ultimately, the project could lead to a novel strategy for the identification of mutations responsible for complex traits in several livestock species.

INTEGRATIVE EPIGENETICS TO PREDICT THE ADAPTIVE CAPABILITIES OF BIO-AGGRESSORS

In the current context of reduced fertilizer use and the effects of climate change, agriculture must now cope with the development of numerous pathogens. Understanding and predicting the adaptive capacity of these crop pests and diseases to their environment is critical to effective control, and is addressed by the EPIPREDICT project.

Epigenetic variations are reversible and heritable modifications of gene expression that do not alter the DNA sequence. In response to environmental constraints over short periods such as episodes of extreme heat, these modifications are a remarkable adaptive mechanism allowing organisms to produce new phenotypes, promoting survival and

development. The study of this epigenetic code relies on high-throughput sequencing techniques that generate massive quantities of data which current analytical methods struggle to exploit fully.

In the EPIPREDICT project, two INRAE genomics researchers, Nadia Ponts (MycSA) and Gaël Le Trionnaire (IGEPP) have teamed up with an expert in statistics, David Causeur (IRMAR), to remove this obstacle. The ambition is to develop innovative statistical and mathematical methods to identify the epigenetic elements responsible for variations in gene expression using two study models: the pea aphid and the phytopathogenic fungus Fusarium graminearum. To understand the relationship between the epigenome and the transcriptome, the project has tested and compared various methodologies, demonstrating the superiority of functional regression mixture models. These perform as well as deep learning methods in predicting expression, and provide relevant biological information on the different co-existing association patterns between chromatin opening and gene expression within the genome. This project will enable the construction of new predictive models, a more detailed understanding of epigenetic mechanisms, and contribute to the design of resilient and sustainable agroecosystems. ■

HEALTH & BIOLOGY: PRECISION MEDICINE AND ARTIFICIAL INTELLIGENCE

Claude P. Bertrand
Executive Vice-President Research & Development, Chief Scientific Officer, Servier

Patients don't have time to wait! Yet, R&D process in healthcare is long and costly because of potential risk and high regulation.

Today, it takes at least 10 years to bring a new drug to market, considering the full cycle in Research & Development (R&D). The required investment amounts to approximately €2 billion with potential attrition rate of up to 90%!

Every month and every year saved is essential for patients waiting for treatment.

Artificial Intelligence (AI) and new technology related to data play a vital role in pharmaceutical R&D. We believe that combining scientific innovation with power of technology will give way for revolutionary new treatments, particularly for patients suffering from rare diseases. We estimate that AI has potential to improve the probability of success in drug development by 10 to 15% and potentially reduce time to market by two to four years.

IN THE BEGINNING WAS PRECISION MEDICINE

Precision medicine is an approach that considers the genetic, environmental and lifestyle variations of groups of individuals to design appropriate prevention, diagnostic, and treatment strategies. Unlike traditional medicine which applies standardized treatments to all patients, precision medicine seeks to identify the most effective interventions based on the specific characteristics of each patient group. Ultimately, we could even imagine a shift toward personalized medicine where treatment would correspond to one patient.

In practical terms, precision medicine has already enabled significant advances at several stages of the drug life cycle:

- Early diagnosis using genetic tests to detect diseases at an early stage, known as «companion diagnostics.»
- Patient stratification: The ability to isolate subgroups of patients based on their genetic and clinical characteristics, easing more accurate diagnoses.
- Targeted therapies: By specifically targeting genetic mutations or biological abnormalities in a patient, the likelihood of a future drug's effectiveness is significantly increased.

APPLICATIONS AND BENEFITS INCREASED TENFOLD WHEN USING DATA AND AI

These technologies act as an unprecedented accelerator for analyzing the components of precision medicine, hence profoundly transforming the way to proceed and, more importantly, how we succeed in R&D.

Two essential use cases:

Healthcare databases: Using and sharing large datasets from electronic medical records, clinical studies, and research to identify trends and correlations.

AI algorithms: Applying machine learning algorithms to analyze data and provide personalized recommendations.

Thanks to technological advances and increasingly powerful data analysis, precision medicine promises to transform the way diseases are diagnosed, treated, and prevented.

AI is being integrated into many aspects of precision medicine, from drug discovery to treatment personalization and post-marketing pharmacovigilance.

AI is essential to precision medicine because it enables the processing and analysis of complex and massive data, identification of biomarkers and therapeutic targets, personalization of treatments, prediction of treatment responses, acceleration of research and development, improvement of diagnostic accuracy, and facilitation of preventive medicine. Thanks to AI, precision medicine can deliver more tailored, effective, and safe care, transforming the way diseases are diagnosed, treated, and prevented.

AI is transforming healthcare R&D, hence getting to faster processes, more precise, and more personalized, which can lead to revolutionary medical discoveries and more effective and safe treatments.

AI AND DATA FOR PATIENTS SERVICE

- Personalization of treatments.
- Use of data to deliver personalized therapies tailored to individual patient needs.
- Analysis of genetic and clinical data to work out specific treatment plans to targeted patient populations.
- Monitoring and Prevention.
- Implementation of patient monitoring systems based on real-time data to anticipate health problems.
- Development of predictive algorithms to early detection of diseases or relapses.
- Patient-Centered Care System: Use of data to improve communication and coordination between various healthcare professionals involved in the patient's care pathway. ■

KEY FIGURES

AI investments in healthcare:

- In 2011, global investments in AI for healthcare were relatively modest, with less than $100 million granted.
- In 2021, these investments exploded, reaching approximately $6.6 billion, according to reports from CB Insights and other venture capital sources. A significant portion of these investments are dedicated to oncology applications.

Number of scientific publications:

- The number of scientific publications on the use of AI in oncology has grown exponentially. There were approximately 200 publications on this topic in 2011.
- By 2021, this number had exceeded 2,500 publications per year, according to databases such as PubMed and Google Scholar.

Clinical trials using AI:

- In 2011, there were very few clinical trials using AI in oncology, with less than 10 registered trials.
- By 2021, the number of clinical trials using AI diagnosis, treatment and management of cancer had grown to over 300, according to ClinicalTrials.gov

CONTRIBUTIONS OF HPC AND AI TO DIGITAL HEALTH: ISSUES AND CHALLENGES

Yves Vandenbrouck
Research Director at Maison de la Simulation, CEA

According to the French Digital Health Agency (ANS), 30% of global data currently look at health, and this volume doubles every 73 days. The UK Biobank Initiative, with its genomic and imaging data from more than 500,000 individuals, represents 30 petabytes. In 2024, European investments in digital health increased by 17%, for a global total of $24.5 billion nearly half of which made in the United States. The exploitation of this data by AI or digital twins finds clear prospects: discovery of molecules, identification of biomarkers, virtual clinical trials, diagnosis, personalized medicine, prevention, and optimization of patient care. The main challenges relate to the management of massive and heterogeneous data, the reduction of research timing, and the personalization of treatments, all requiring advanced analytical capabilities.

COMBINING AI, HPC, AND BIG DATA TO TRANSFORM BIOMEDICAL RESEARCH AND THE MEDICINE OF THE FUTURE

Advances in Deep Learning, combined with high-performance computing and big data, have revolutionized 3D protein modeling. Of the 200 million known proteins, only 20% have an identified three-dimensional structure. For years, in-silico approaches aimed to predict these structures, understand pathologies, and design new treatments. A definite breakthrough comes from the work of David Baker (with “de novo” protein design) and the DeepMind team (Demis Hassabis, John Jumper) who were awarded the 2024 Nobel Prize in Chemistry. Their results produced with AlphaFold have increased prediction accuracy from 40 to over 90%, thanks to an AI based on attention mechanisms trained on over 2 billion sequences, deployed on a 128 TPUs v3 device optimized for neural networks (see Figure 1).

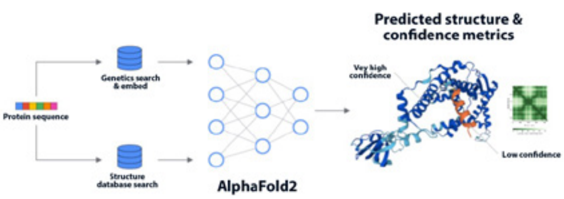


Figure 1. Simplified diagram of the AlphaFold2 inference flow (source ©EMBL-EBI)

Despite persistent challenges (such as antigen-antibody recognition), the success of AI in healthcare relies on the combination of multidisciplinary skills and powerful computing infrastructures. These advances open new perspectives, particularly for discovering treatments. In 2025, generative AI made it possible to identify a target for pulmonary fibrosis and create a promising inhibitor molecule, successfully tested in phase II by in-silico Medicine – a world first. Given the costs (€50 to €100 million) and low success rate (10 to 20%) to develop a candidate, these results exemplify the potential of AI. Start-ups like IKTOS and InstaDeep are leveraging cloud and

AI to explore immense chemical spaces, automate synthesis, and plan robotization equivalent to the productivity of 30 chemists. Furthermore, foundation models (FM) emerging in biology can exploit vast volumes of unlabeled data. A national CEA project aims to develop one FM to analyze the human cortex folding's, based on brain images of more than 100,000 individuals (UK Biobank) to identify early markers of psychiatric disorders. However, despite these promises, the level of maturity of AI remains inconstant depending on the field and requires ethical supervision from the design stage («ethics by design»).

MATURITY LEVELS OF USE CASES AND AI IN HEALTHCARE

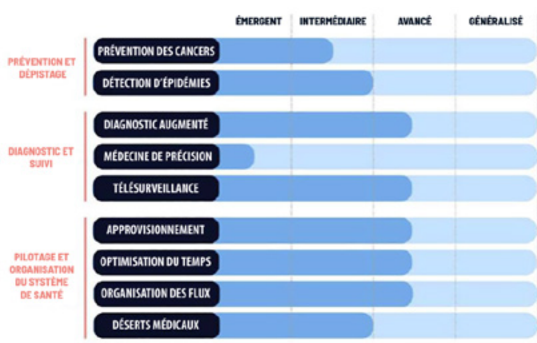


Figure 2. Degree of maturity of AI use cases in healthcare (© HealthCare Data Institute).

STRUCTURING THE DIGITAL HEALTHCARE ECOSYSTEM IN FRANCE: PROGRESS MADE AND OBSTACLES TO OVERCOME

Launched in January 2021 as part of France 2030, the digital health acceleration strategy aims to structuring a globally competitive French ecosystem. It is built around training (€80 million), research via the «Digital Health» PEPR (€40 million), and the assessment of digital medical devices (€20 million). A key challenge lies in the deployment of powerful and secure digital infrastructures complying with regulatory requirements (GDPR, CNIL, SecNumCloud). A €75 million call for projects was launched in 2022 to strengthen hospital data warehouses, with 16 winning institutions aiming to create a national network for secure health data sharing. In a rapidly expanding market, GAFAM, such as AWS (UK Biobank) and Microsoft Azure (SNDS) are setting up, sparking debates on data sovereignty. The national health data platform (or HDH) will have to offer certified sovereign hosting by 2026. This context is encouraging French and European stakeholders to adapt their offerings and guarantee data security and access. Interoperability remains a priority for effectively reusing data often scattered, heterogeneous, and poorly documented. Removing barriers to access (bureaucratic burden, silos) and promoting a culture of open sharing between producers and users is crucial. The Marchand-Arvier report (2024) calls for strengthened interministerial management and an overhaul of HDH governance. The publication of the European regulation on the European Health Data Space (EEDS), enforcing a one-stop shop (ORAD) in each country by 2026, should accelerate coordination and streamlined access to health data.

BUILDING THE DIGITAL HEALTH ECOSYSTEM

The growth of digital health in France relies on the ability to master and combine HPC, big data, and AI in a highly competitive international context. This profound change, as much cultural as technological and organizational requires investments in training (data science, cybersecurity…) for sovereign and secure infrastructures upon reliable, auditable, and interpretable AI software architectures. In addition to national initiatives, programs such as NumPex (Exascale/AI) or European projects (EuroHPC, EBrains, EEDS), co-led by France, offer opportunities to pool skills and resources. The "HPC/AI for Digital Health" workshop run at the Teratec 2025 Forum clearly illustrated the need for cooperation between manufacturers, researchers, clinicians, and data producers. While synergies still need to be strengthened, talents and projects live. It is therefore essential for all stakeholders – public and private – to work together and build a collaborative ecosystem that support biomedical research to improve patient journeys. ■

NEUROSPIN: MESHING PHYSICS AND DIGITAL TECHNOLOGY TO SERVE NEUROSCIENCE

Jean-François Mangin
Director of Baobab Unit (UMR) at Neurospin, CEA

The world of physics has long understood the need for very large instruments to answer certain fundamental questions. The world of life sciences has not yet experienced such a phase transition. Neurospin was founded by the CEA some twenty years ago to attempt to initiate such a process, with an exceptional MRI called Iseult at the heart of its project. Its 11.7 T magnet was designed by physicists who usually work in the field of high-energy physics. The first lesson of this project is that physics timing is not that of biology: to the surprise of the instigators of this project, it took almost twenty years to produce the first in vivo images. This milestone marked a new step, passing of the baton to the physicists of Neurospin and their imaging expertise. The image in Fig. 1 depicts the highly wrinkled morphology of the cerebellum, confirming the unparalleled potential of Iseult. Without such an exceptional resolution of this image (300 isotropic microns), it is impossible to visualize the almost fractal aspect of this anatomy. Efforts to improve spatial resolution are now focused on functional imaging. The human cortex is organized into layers and in columns less than a millimeter thick. Zooming in to a resolution that allows us

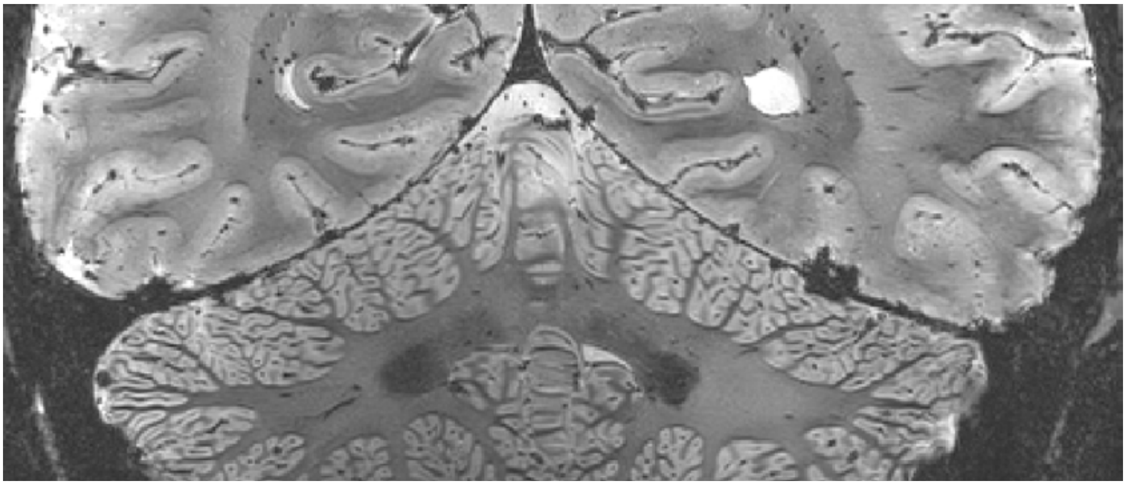


Fig. 1: Human cerebellum imaged in vivo at 11.7T (300 microns isotropic) - © Courtesy, N. Boulant & A. Vignaud

to distinguish the signals from these elementary components is very promising.

FROM PHYSICS TO ALGORITHMS

Other professions are now being called upon to move on the adventure leading to first discoveries in neuroscience: data analysis experts who will develop algorithms dedicated to the specificities of images from Iseult, and researchers in cognitive or clinical neuroscience who will design the first experiments. Since the MRI technique is completely digital, digital technology will now play a leading role alongside physics and at all levels of the analysis chain.

MRI is a machine that can read the «Fourier transform» of the object being isolated in its tunnel. This reading in Cartesian mode can take a considerable amount of time. The MIND research group (CEA-INRIA) is using AI to invent new trajectories that optimize the number of reading points, by combining «compress sensing» vision and modeling of the constraints induced by the imager's electronics. This strategy is going to minimize acquisition times.

Numerical modeling is also essential when combining so many raw images to reveal a complex structure. The Ginkgo team (CEA) has thus set a world record for mapping brain connections. By accumulating 5,000 hours of post-mortem acquisitions on an 11.7T magnet designed for rodents and after dividing the human brain into cubes, they achieved an isotropic resolution of 200 microns. The calculated wiring image results from a massive optimization process carried out at the TGCC, to interpret the perturbations of random water movements induced by axon membranes. The underlying model is inspired by spin glasses in statistical physics.

IMAGING ANALYSIS BY AI

Ultimately, digital technology is clearly taking off when it comes to the analysis of very large databases. Most recent developments in open science now allow the analysis of more than 100,000 MRI brain scans, as in the Champollion project part of the risk-stake research program Audace from the CEA, aiming to bring out an AI capable of deciphering the hidden language behind patterns of one's cortex folding's (Fig. 2). ■

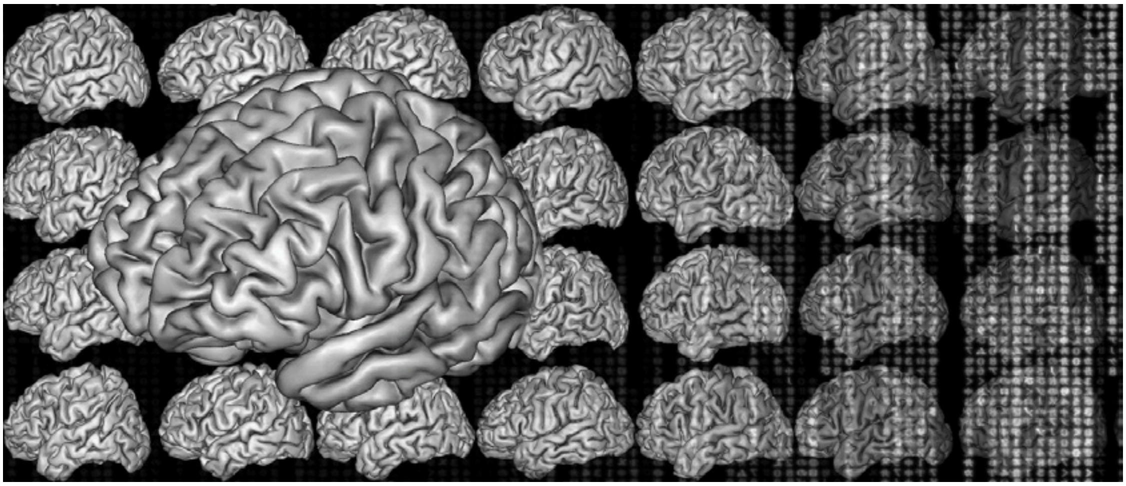


Fig. 2: A fraction of the Rosetta Stone of 100,000 brains used by the Champollion AI to decipher the "ideograms of the cortex"

THE VIRTUAL HUMAN TWIN: A CRUCIAL LEAP FORWARD FOR SCIENCE

Steven Levine,
Director Virtual Human Modeling, Dassault Systèmes

Ignited about fifteen years ago, the digital modeling of human organs accelerated during the pandemic, resulting in a true Human Virtual Twin (HVT), today usable both in research and clinical practice.

Starting with modeling of the heart as Living Heart Project, this work has been extended to other organs (brain, liver, lungs, kidneys, pancreas...). With its technical collaborative exchange platform developed by Dassault Systèmes, used by thousands of experts to share their research and through public funding, we rely on multi-organ and multi-scale models supported by parametric data to reflect human variability; from a very high precision level of a few million, we have gone to hundreds of millions of degrees of freedom; these are now usable in clinical practice as they have been validated by regulatory authorities.

FDA-APPROVED MODELS WITH EXCITING FEATURES

The accuracy of the models depends on the organ and the context of use, but two key criteria guide their validation: regulatory acceptance and clinical use. A 5-year study conducted with the FDA validated the use of the heart virtual twin as a surrogate population for in-silico clinical trials. Several medical devices have thus obtained approval. The clinical use of virtual twins represents the last step. It has expanded, for example, to guide complex surgical procedures in pediatrics. Experience shows that surgeons using virtual twins obtain better results than those without. The first virtual twins focused on a few well-comprehended phenomena to validate the technology. Today, their scope is only limited by the availability of data. By merging mechanistic physiology with clinical data, connected objects, genomic or multiomic data, a virtual twin can predict outcomes across the entire spectrum of human activities.

THE IMPACT OF DIGITAL TECHNOLOGIES

Continuous technological advances (Simulation, AI, HPC, Quantum, etc.) are the driving force that distinguishes HVT from animal experiments or clinical tests to predict human response. Hence, heart models adapt instantly to any cardiac morphology. Similarly, the Living Liver Model combines pharmacokinetic modeling with that of fluids and structures, thus covering several orders of magnitude in time and space. The computing power available in the cloud allows companies and hospitals to be more responsive by running in-silico clinical trials in parallel with thousands of virtual patients. Finally, AI quickly creates models specific to one pathology or patient, with less risk of error than humans. Similarly, initial work carried out with quantum engines shows they are particularly suited to this type of application.

HARDWARE-IN-THE-LOOP IS COMING

These HVTs are also used to test equipment currently under development. Eventually, specialized "Heart-on-Chip" ASIC circuits could be considered, updating electromechanical fields in real time by analyzing ECG, pressure, or strain data from sensors, to control new smart medical devices.

While the potential of HVTs seems infinite, it is limited still by fragmentary knowledge of the human body's functions. The quality and accuracy of measuring patients' physiological data to feed the models will also be crucial. Finally, moving from observation-based to predictive medicine requires a profound paradigm shift that medical training will have to anticipate. This is an ongoing change that must be accelerated. ■

DIGITIZING NATIONAL HERITAGE: FROM STONE TO DIGITAL TWIN

Florian Moreno
Value Creation and R&D Director, AGP

For three decades now, Art Graphique & Patrimoine (AGP) has been exploring a simple yet powerful idea: making digital technology an ally in service to the memory of places.

Since our founding, we have seen digital technology transform our professions. Initially, it was about helping to understand, document, and draw what exists – to restore and protect it, sometimes simply to ensure it is not forgotten. Today, we enhance sites and make them accessible through immersive experiences and interactive digital duplicates. And behind all this: precise measure surveys; extensive modeling; terabytes of data to manage. There is an entire ecosystem in motion, and AGP is playing its part with rigor and pragmatism.

A JOURNEY THAT BEGINS ON THE CONSTRUCTION SITE

When Gaël Hamon founded AGP in the 1990s, he came from a background in stone trade, in the literal sense. As a stonemason, he knew monuments, the construction techniques, and constraints from the ground. But also, he foresaw the dawn of something new: information technology which would, sooner or later, revolutionize the way we view, measure, and think about heritage.

From the outset, the goal was clear: to produce rigorous, reliable, and usable surveys for heritage architects and specialized firms. Digital technology was a tool, far from a gadget. It had to be learned with the same rigor as a compass or a ruler.

At that time, we worked mainly on building heritage: facades, vaults, sculpted portals. But around 2005, we broadened our scope. Museum works of art (statues, paintings, decorative objects) became a field of activity. Here too, the need to document, preserve, and promote was strong. And tools required have followed.

ABOUT TOOLS, A SILENT REVOLUTION

For years, the technical priority was due to precision. Each new scanner promised more detail and density. And so, we moved forward. Until, about ten years ago, we reached a ceiling: beyond a certain level of detail, the human eye no longer perceives the difference, and neither do for uses. Since then, speed, flexibility, and compactness have guided innovation. Whereas it took two operators and four scanners to survey a church, a single person

Drone survey of the 5th century baptistery of Saint Léonce Cathedral in Fréjus and the resulting digital model. Doc AGP

Drone survey of the 5th century baptistery of Saint Léonce Cathedral in Fréjus and the resulting digital model. Doc AGP

equipped with a state-of-the-art scanner, a machine can do better, faster, and without compromising on quality. Registrations become semi-automatic, interfaces more intuitive and, above all, it's possible to combine approaches: terrestrial laser-gram metrics; ground-based or aerial photogrammetry, with the use of drones. Each monument is captured from all angles, with millimeter precision and flexible intervention.

But all this has a very concrete consequence: the more powerful the tools, the larger files become.

THE SNOWBALL EFFECT OF DATA

Between 1994 and 2004, all our projects fit in with a single hard drive. Today, it can take several terabytes to store an entire project. This isn't just a technological anecdote. It's a daily challenge.

So, the faster and more comprehensive the capture, the larger the files. This requires robust infrastructure, long-term archiving, reliable servers, maintenance, and backups. As a digital heritage, in short, that must be preserved with as much care as stones it represents.

But it's not just about storage. It also requires processing. Cleaning the data, assembling and optimizing it to make it usable - be it a professional software or a mediation application. And that requires ingenuity.

For the past ten years, we have been integrating the logic of PBR (Physically Based Rendering). We are no longer content to simply apply a texture. We are working with «Normal Maps» which allow us to simulate the relief of a surface without resorting to complex geometry feigning reliefs, roughness effects, shine... All this allows us to produce stunning, but above all lightweight renderings. It is a subtle compromise between perceived quality and file size. An invisible but essential task.

WHAT CALCULATION IS CHANGING (AND WILL STILL DO)

With such volumes, manual processing is no longer sustainable. High-performance computing (HPC) has become a discreet but essential partner for us. Point cloud registration, mesh generation, Baking of texture to reduce model size: all of this is now processed in parallel on servers equipped with powerful graphics cards. And without this, some projects would remain stuck at a file processing stage.

Digital twin used in the ongoing rehabilitation of the Salon de Diane at the Versailles Estate. AGP Doc

Plots cloud from metrics of the Arc de Triomphe in Orange. Doc AGP

This is just the beginning. Artificial Intelligence is beginning to take over tedious tasks: cropping hundreds of images; classifying gigabytes of data; identifying shapes; proposing alignments... It doesn't stand in human expertise, but it relieves teams, speeds up workflows, and opens new perspectives.

Beyond image recognition, it will help, as in medical fields, to diagnose the pathology of a building or a work of art. In mediation, it will enable a virtual mediator to respond accurately and in a personalized manner to visitors.

We are entering the era of the digital twin 2.0: no longer simply visualizing a monument, but tracking and monitoring it, while simulating scenarios. A 3D model is becoming an analysis tool and mediation platform. This was noticeably the case with the digital twin used for Notre-Dame de Paris after the fire. Coupled with physical sensors, it can even react, signal deterioration, and anticipate restoration needs.

WHAT ABOUT NOW?

Today, AGP has joined the Memorist cluster, alongside five entities with a shared goal: to preserve, digitize, enrich, and transmit heritage. Such linking was accompanied by a change in management team, but the DNA remains the same: rigor; innovation; respect for material; and attention to detail.

Tomorrow, it's about integrating digital technology into the long-term perspective of heritage. Records will no longer be simple snapshots, but regular, documented, and updated flows. Each monument or work will be able to exist in a living, reachable digital space.

Heritage will be more widely accessible. A visitor will explore a sculpture in augmented reality from home; a restorer will virtually test an intervention; a curator will follow the evolution of an object through cross-referenced measurement surveys and dynamic visualizations.

This type of approach, already common in medicine, is emerging in the heritage world. Even if everything remains to be invented, the potential is immense: preventive conservation; training; transmission; opening to the public at large.

And behind all this, AGP will remain in its own right: making connections. Between hand and machine. Between material and model. Between the past we receive and the future we prepare for. ■

ROLE AND CONTRIBUTIONS OF HPC, AI AND QUANTUM TECHNOLOGIES TO ADDRESS CLIMATE AND ENVIRONMENTAL CHALLENGES

Philippe Drobinski
Research Director, CNRS – Professor at Ecole Polytechnique
Laboratoire de Météorologie Dynamique – Institut Pierre Simon Laplace (LMD-IPSL)
Director of Energy4Climate (E4C) interdisciplinary center

The role and contribution of HPC, AI, and quantum technologies in addressing climate and environmental challenges.

Faced with climate emergency, cutting-edge technologies such as high-performance computing (HPC), artificial intelligence (AI), and quantum computing are proving essential tools for understanding, modeling, and mitigating effects of climate change. These technologies enable us to process massive volumes of data, simulate complex scenarios, and develop innovative solutions for a more sustainable planet.

HIGH-PERFORMANCE COMPUTING (HPC): THE PILAR OF CLIMATE MODELING

HPC is at the heart of modern climate modeling. Supercomputers can simulate atmospheric, oceanic, and terrestrial phenomena with unprecedented accuracy. For example, general circulation models (GCM) emulate HPC to forecast long-term climate trends and assess the potential impacts of greenhouse gas emissions [1]. Institutions such as the National Center for Atmospheric Research (NCAR) in the United States, the European Center for Medium-range Weather Forecast (ECMWF) in Europe, Météo France, and the Institut Pierre Simon Laplace (IPSL) in France also use supercomputers to improve the accuracy of weather and climate forecasts. These simulations help decision-makers develop more effective adaptation and impact mitigation policies.

ARTIFICIAL INTELLIGENCE (AI): A CATALYST FOR SUSTAINABLE SOLUTIONS

AI is transforming the way we address environmental challenges. It is being used to analyze large climate datasets, optimize energy

systems, and monitor biodiversity. For instance, AI algorithms can predict extreme weather events, helping communities prepare and respond more effectively. Initiatives like Aardvark Weather, developed by the University of Cambridge, are using AI to provide accurate weather forecasts in regions where resources in supercomputing are limited [2]. In addition, AI is being used to monitor deforestation, track methane emissions as the Kayrros company does, and optimize precision agriculture, hence contributing to the reduction of greenhouse gas emissions.

QUANTUM COMPUTING: A PROMISE FOR THE FUTURE

While quantum computing is only just beginning, it gives considerable potential for solving complex climate-related problems. Quantum computers could simulate chemical reactions to develop new carbon capture materials or optimize energy grids for better integration of renewable energy [3]. Companies like PsiQuantum work on quantum architectures capable of simulating such complex chemical processes, giving way for advances in carbon capture and storage. However, technical challenges remain, particularly related to qubit stability and the energy consumption of cooling systems required to operate quantum computers.

TOWARDS RESPONSIBLE AND SUSTAINABLE USE

While these technologies lend promising solutions, it is crucial to consider the inner environmental impact. Data centers powering AI and HPC consume significant amounts of energy and water, contributing to greenhouse gas emissions as well [4]. It is therefore essential to adopt sustainable practices, such as using renewable energy, optimizing energy efficiency, and recycling electronic equipment. Initiatives are being taken for technologies to become more environmentally friendly. For example, some companies are developing data centers powered by renewable energy sources, while others are working on more efficient AI algorithms, thus reducing their carbon footprint. Specific research on frugal digital is also being conducted at research centers such as the Energy4Climate (E4C) center in France.

CONCLUSION

HPC, AI, and quantum technologies play a crucial role in the fight against effects of climate change. They enable us to better understand our planet's complex systems, predict future impacts, and develop innovative solutions for a sustainable future. However, their deployment must be accompanied by a reflection on their own environmental impact to ensure that these tools truly serve the cause of climate. ■

References:

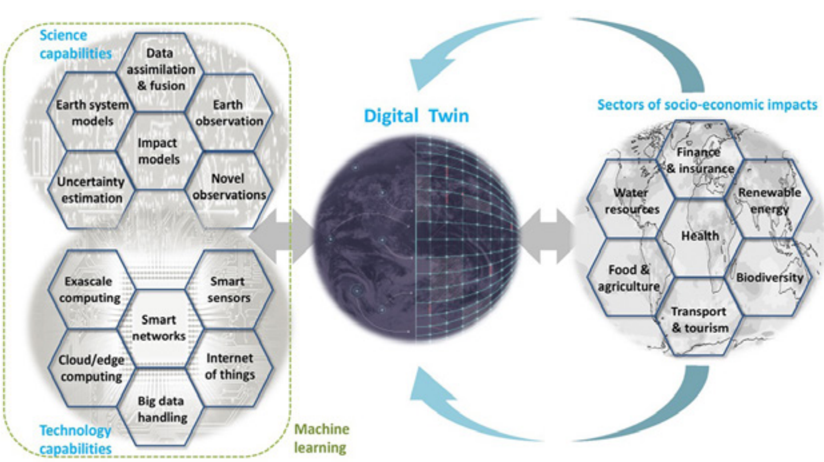
1. Kulkarni, P., Manoharan, S., & Gaddi, A. (2024). Advancing Climate Modelling through High-Performance Computing: Towards More Accurate and Efficient Simulations. EAI Endorsed Transactions on Energy Web, 11(4). https://doi.org/10.4108/ew.7049publications.eai. eu

2. Allen, A., Markou, S., Tebbutt, W. et al. End-to-end data-driven weather prediction. Nature 641, 1172–1179 (2025). https://doi.org /10.1038/s41586-025-08897-0

3. Giani, A., and Z. Goff-Eldredge (2022), How Quantum Computing Can Tackle Climate and Energy Challenges. Eos, 103, https://doi. org/10.1029/2022EO220500

4. MIT News. (2025). Explained: Generative AI's environmental impact. Retrieved from https://news.mit.edu/2025/explained-gene rative-ai-environmental-impact-0117MIT News

5. Bauer, P., Quintino, T. & Wedi, N. (2022) From the Scalability Pro gramme to Destination Earth. ECMWF Newsletter, 171. https:// www.ecmwf.int/en/newsletter/171/earth-system-science/ scalability-programme-destination-earth



Credit : ECMWF [5]

Digital twins of the Earth system enable the overall socio-economic sectors to access optimal data quality and create tools to extract information targeted to their applications. A digital twin is based on the convergence of cutting-edge scientific and technological skills, based on simulations and observations, leveraging the full range of digital technologies. It relies on machine learning methodologies for all components.

DIGITAL TECHNOLOGY AT THE HEART OF GEOSCIENCES FOR A CHANGING WORLD

Michaël Chelle
Chief Digital Officer, Bureau des Recherches Géologiques et Minières (BRGM)

Romain Chassagne
Deputy Program Director in charge of Digital Twins, Bureau des Recherches Géologiques et Minières (BRGM)

Simon Lopez
Deputy Program Director in charge of Computing and digital simulation, Bureau des Recherches Géologiques et Minières (BRGM)

The subsurface has become essential to human activities in a context of climatic, environmental, geopolitical and digital upheavals. It provides water, energy, metals, storage solutions (CO2, heat, waste) but also generates risks on the surface (earthquakes, ground movements, submersion...). However, this natural environment is complex and costly to explore, given the different spatial-temporal scales (from micrometric minerals to thousands of square kilometers of a lithospheric plate; from the few seconds of an earthquake to the millions years of geological eras), the heterogeneity of its structures, the non-linearity of physic-chemical processes occurring and the extreme conditions (temperature, pressure...) that can prevail thereto. To manage the subsurface in such context, public and private decision-makers need a broad, up-to-date 3D and multi parameter mapping. This ambitious objective requires the transformation of geosciences through digital technology. This is why BRGM has made it a scientific challenge aiming to develop quantitative and predictive geosciences, using data and knowledge to simulate physical and chemical processes of the subsurface, while quantifying uncertainties. This will better support public decisions by quickly producing reliable and traceable data, as for the evacuation of an area threatened by marine submersion. This requires applying digital technology through different prisms: data, digital simulation, the use of artificial intelligence (AI), and their integration into digital twins of the subsurface.

ENHANCING THE VALUE OF DATA

For decades, BRGM has been archiving subsurface data, like a national geological survey in other countries. One of the current challenges is to make this heterogeneous yet lacking heritage more accessible and usable. AI holds promise for this purpose, particularly for automatically structuring and interpreting such data. Geophysical measurements, on the other hand, require mathematical and numerical approaches to solve reverse problems and figure deep geological entities. These calculations, which are very CPU-intensive, rely on parallelized codes making the most of new architecture.

NUMERICAL SIMULATION OF UNDERGROUND PROCESSES

The quantitative approach to many subsurface issues is based on multi-physical modeling of transfers and flows. These models can be very complex, depending on the degree of coupling data between physical phenomena and heterogeneous and discontinuous nature of geological geometries, therefore being computationally intensive (Fig. 1). In addition, they are used to produce maps and spatialized data, which requires multiplying simulations as well as quantifying uncertainties. Parallel computing has expanded their use, but required simulation time remains a barrier, especially in crisis situations. To overcome this, these models must be simplified (reduced order modelling), by lowering their complexity and/or developing less computationally intensive models (like with meta-modelling).

AI TACKLES BIG MODELS

Although limited by a scarcity of subsurface data, uses of AI are rapidly developing in geoscience. Machine learning is an excellent tool for simulating subsurface phenomena, despite limited knowledge of the mechanisms, and offers rapid execution in direct mode. Such pace is also leveraged to develop metamodels from simulations of «large» multi-physics models. However, the training process involves long simulations and therefore requires optimal sampling of hyperspace factoring, while making sure to limit uncertainty. Generative AI, which is very GPU-intensive, is being explored via LLMs and PRAGs to exploit unstructured data (reports, old maps, mining archives), but also to create synthetic yet realistic data via GANs compensating for the difficulties of direct observation.

DIGITAL TWINS

Digital twins are also emerging to comprehend the subsurface, whether for industries exploiting the subsurface (geothermal energy) or to support public policies (integrated water management, field-risk assessment). As true catalysts for decision-making, they rely on the continuous assimilation of monitoring data streams into sophisticated simulations. They foster a dynamic and predictive understanding of the Earth system, while raising the need for high-performance scientific computing.

This digital transformation of geosciences, essential to address societal challenges related to the subsurface, leads to growing needs for computing power, with CPUs and parallel systems as well as GPUs. Being aware of the environmental and geopolitical challenges facing mineral resources, BRGM nevertheless aims to ensure a rational use of these resources in computing. ■

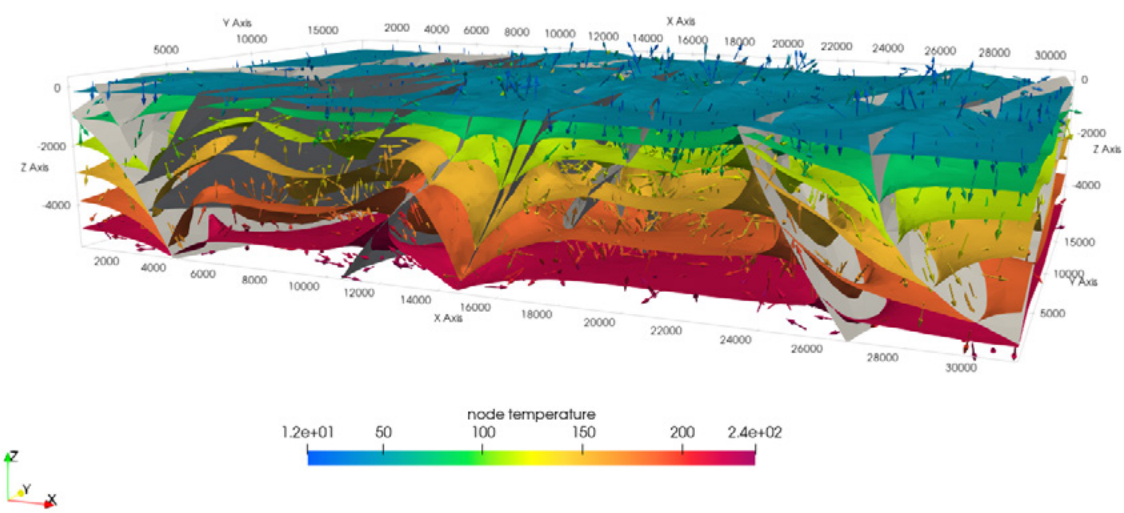


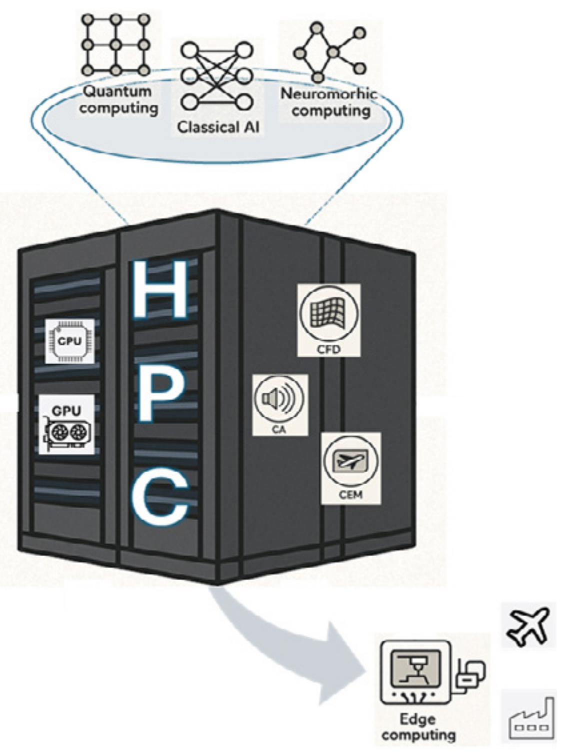
Figure 1: Temperature range in a geothermal system of the Rhine Rift, simulated by a hydrothermal dynamics model, based on the ComPASS code (Armandine Les Landes A. et al, 2025, https://doi.org/10.1016/j.cageo.2024.105752)

EXTENSION AND HYBRIDIZATION OF HPC IN AERONAUTICS

Thierry Chevalier
VP Chief Engineer Manufacturing & Process | Engineering R&D, Capgemini
Senior adviser complex system of systems modelling & simulation applied to climate transition

FOR THE LAST 10 YEARS

HPC architecture has changed significantly with the emergence of specialized processors (GPU, TPU, parallel ARM, manycore Kalray, etc.), enabling the containment of simulation energy costs. This evolution required a redesign of codes, often involving a shift from implicit to explicit approaches, or toward more computationally demanding methods such as Large Eddy Simulation (LES) and Detached Eddy Simulation (DES), which have become feasible thanks to Exascale computing systems.



The disciplines traditionally reliant on High-Performance Computing (HPC)—such as aerodynamics, acoustics, electromagnetism, and structural mechanics—have leveraged these advances to reduce, or even replace, physical testing, and in some cases contribute directly or indirectly to certification. Integrated multiphysics optimization has emerged, enabling coupled analysis of complex phenomena (e.g. multidisciplinary coupled analyses of aerodynamics/structure/acoustics/thermal aspects for engine installations including pylon and nacelle).

Quantum computing, still marginal, is being explored through hybrid HPC/quantum approaches, particularly for fault detection. Artificial Intelligence (AI) is increasingly integrated into numerical algorithms: rapid initialization of solution fields by AI to reduce convergence time, and robust, high-performance

AI-based turbulence models applicable to highly diverse configurations in Computational Fluid Dynamics (CFD).

Advances in storage (e.g. RAM, disks, distributed file systems) and hardware have made it possible to broaden the range of cases addressed: finer meshes, increased data volumes, and more complex Machine Learning (ML) models. This improvement in Input/Output (I/O) performance, combined with the integration of AI and ML with HPC systems, has revolutionized data processing and analysis. It has fostered the development of digital twins and predictive maintenance. The surge in the number of airlines that have switched to flight-hour-based maintenance is highly indicative of this success, which ensures "Zero Aircraft On Ground." Trajectory optimization has also delivered gains comparable to a fleet modernization (e.g. EasyJet).

Finally, the widespread adoption of HPC type of techniques at the edge has revolutionized operations (e.g. flexible onboard image processing controlled from the ground) and manufacturing (e.g. 3D printing with initial optimization of parameters/shapes to be printed and parameterization learning models, subsequently embedded in edge systems near the printer to control parameters in real time).

IN THE NEXT DECADE

Neuromorphic computing (e.g. Tesla Dojo) improves energy efficiency and enables learning from reduced datasets thanks to its native ability to process temporal sequences. This approach is expected to benefit preventive maintenance by providing more efficient and scalable models (requiring smaller training datasets for recent technologies). However, it is overshadowed by the success of Large Language Models (LLMs), which are transforming the generation and use of technical documentation, support services, and even the analysis of simulation results.

The same algorithm at the core of LLMs - Google's "Transformer" - also opens perspectives beyond language, such as uncertainty and margin management (critical for complex products like aircraft) and data flow organization (e.g. on an avionics bus). Numerical simulation is increasingly improving the structure of the neural networks on which it relies (topological or categorical optimization of the network during its creation to generate "sparse" networks), making it possible to achieve equivalent performance with models reduced by several orders of magnitude, suitable for use in edge or embedded systems.

Quantum-classical integration is progressing, notably through logical qubits, with applications in quantum chemistry for the study of new materials or alternative fuels - a key challenge for post-kerosene aviation. The increase in computing power paves the way for near real-time simulations and optimizations, reducing exploration and design cycles—crucial at a time when the traditional "tube-and-wing - kerosene" architecture is being reconsidered.

The improvement of energy efficiency of HPC is expected to continue offsetting the increase in its usage, in a context where the boundaries between simulation and AI are becoming blurred. Hybrid approaches are becoming the norm (e.g., workflows partially driven by AI, combinations of simulation models and ML models). Notably, AI capabilities to analyze code, suggest technology porting, and even assist or automate the configuration and maintenance of simulation chains can greatly facilitate the evolution and upkeep of the extensive existing corpus of numerical simulation in aeronautics, at a time when hardware architectures are evolving rapidly.

Finally, embedded HPC is becoming widespread, both in aircraft and in factories. Combined with optimized AI models, it enables onboard simulation, analysis, and real-time adaptation close to physical systems. This opens the way to more resilient and adaptive systems, as well as continuous performance optimization during flight (e.g. adjustments via onboard Light Detection and Ranging (LIDAR)), with enhanced pilot assistance. ■

DIGITAL TRANSFORMATION AND NUMERICAL SIMULATION: THE ACCELARATOR FOR VEHICLE PROJECTS

William Becamel
Expert Leader, Numerical Simulation
Global Engineering Renault Group

Olivier Colmard
Digital Transformation Director, CAE & PLM
Global Engineering Renault Group

The automotive sector faces numerous challenges. Amongst regulatory issues increased international competitive pressures, and the energy and technological transition (vehicle electrification, driver assistance, and connected systems), Renault Group's response is agile and rapid engineering. Coming up with this situation, vehicle development now takes place over two years.

To achieve such level of performance, digital technology is central. The story did not begin today, taking place in the 2000s with significant investments in these technologies. Constantly evolving computing infrastructure made it possible to create increasingly specific and reliable models (Fig. 1). An example with passive safety is particularly illustrative. Progress in this area with such levels of correlation make it possible to achieve the expected performance targets from the first crash.

SIMULATION IN VEHICLE DEVELOPMENT: WHY AND HOW

The generalization of digital simulation has become inevitable in vehicle development; it is one substantial condition among major milestones

ACCURACY IS THE KEY

INCREASE OF THE SIZE OF THE MODELS

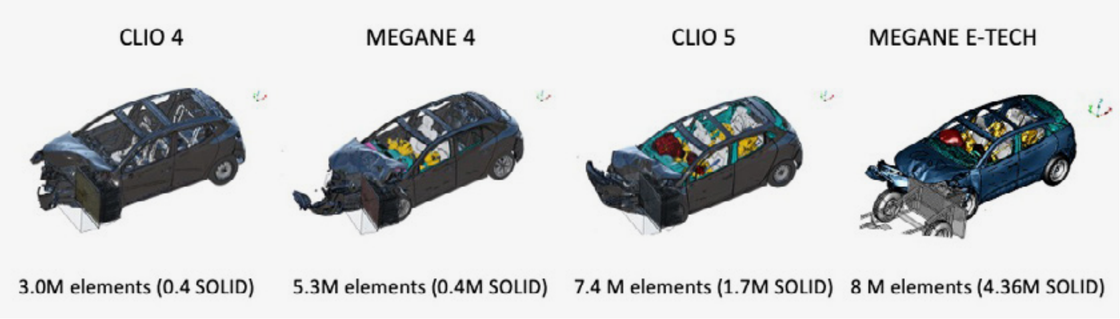


Fig. 1: Increase in model sizes between different vehicle generations. Gradually, restraint systems (airbags / belts) and dummies were, for example, integrated. This evolution was achieved without impact on time spent sharing data with engineers, thanks to developments in HPC infrastructure and optimization of calculation codes.

to overcome product design and industrialization. Company's organization has evolved to meet this need, notably with the professionalization of occupations to manage CAD in configuration, to create meshing and modeling, to launch and analyze calculations. In just 3 weeks, we are now able to produce the report of all structuring operations in a vehicle (aerodynamics, thermal comfort, forbearance, endurance, acoustics, crash...).

If requirements are not met, daily iterative loops are carried out by engineers to resolve any difficult issues and provide solutions to the product. This digital twin of the vehicle allows for rapid design, initiation of physical prototype phases with a high level of confidence and acceleration of final validations. The results are worth the investment here, considering that validation costs were reduced by 53% between 2019 and 2022 (fig.2).

ACCELERATING TO STAY IN THE RACE: LEVERAGES FOR DIGITAL SIMULATION

However, we can't simply stop there. Competition remains strong, and disruption is necessary. What are the prospects?

1. Continue to improve our computational coverage (i.e., number of validation activities we can simulate) and maturity of such simulations (i.e., their accuracy).

2. Increase our simulation and modeling speeds. Contribution of GPUs in this area is much valuable depending on the physical domain.

3. Combine virtual and physical resources to anticipate vehicle development before starting the

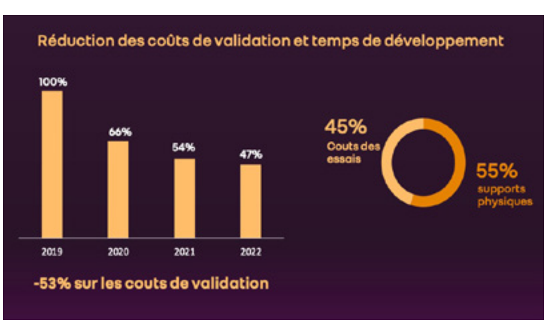
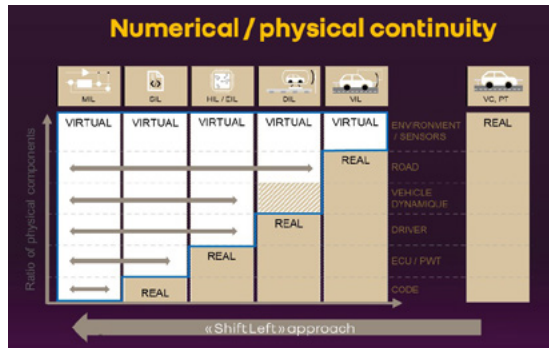


Fig.2: Evolution of Renault Group validation costs between 2019 and 2022.



industrial phases. We are talking here about the continuum of XiL tools (of the Model in the Loop – MIL, Software in the Loop – SIl, Hardware in the Loop HiL...)

4. Finally, the last element offering major performance opportunities is Artificial Intelligence. We should talk rather about multiple artificial intelligences, with the many algorithms, neural networks and trained tools to accelerate long-lasting tasks in product development. Some of these tasks are performed upstream of digital simulation to bring about concepts, others come in substitution or in complement of simulation to expedite or improve the result. Thanks to this technology, design loops could be accelerated.

AI can also help with many other tasks: searching for information (standards, similar parts, quality data...), generating lines of code for embedded systems, defining and performing validation tests...

All learning and inference duties for these AIs still demand significant computing resources. Maintaining our competitiveness requires high-level computing infrastructures, agile enough to allocate massive resources matching those needs. Effective computing infrastructures are thus a prerequisite for the competitiveness of our industry.

Overall, digital simulation and more generally digital transformation of engineering are central to accelerate vehicle development, enabling us to design innovative and high-performance products at the best cost for our customers. ■

HPC AND AI TO SERVE NAVAL SUPERIORITY

Cédric Leblond
Project Manager for Scientific Computing, Naval Group

Naval Group designs, builds, integrates, maintains in service, modernizes, dismantles and deconstructs submarines and surface ships for the French Navy and 50 allied navies. Such positioning is unique as the company is involved in the entire life cycle of armed vessels, from preliminary studies to design, through to deconstruction, including the physical and functional integration of systems.

The company is heir to nearly four hundred years of innovation, contributing to the technological superiority of its customers and improving its competitiveness. This innovation is currently reflected in six major areas: Smart Naval Force, Invulnerable Ship, Smart Ship, Blue Ship, Smart Availability, and Smart Industry.

Digital simulation, in the broadest sense, is now a major cross-functional tool within the company, present throughout the entire ship lifecycle. It is also a true driver of innovation in each of the six areas.

EVOLUTION OF HPC USE

For over 30 years, high-performance computing (HPC) has been central to the detailed design phase of ships, particularly for multi-scale and multi-physics simulations. Its use now extends to the entire development cycle. Starting with the bid phase, it provides for the verification of expected performance where to compare several variants. In the upstream design phase, it facilitates the rapid integration of innovations and contributes to achieving the «right in first time» standard, a key factor in optimizing costs, deadlines, and quality across the entire program.

The needs for detailed design have intensified: detailed simulation is now used systematically to verify that requirements are met at each iteration, according to the principles of Verification and Validation. Furthermore, exchanges between design engineers and computer scientists occur in near real-time, as made possible by a widespread use of HPC, guaranteeing responsiveness, robustness, and reliability of the design.

HPC is also used in production. Welding simulation helps ensure industrial practices and optimizes assembly sequences. Simulation of additive manufacturing processes provides for the definition of the right parameters, reducing testing, costs, and lead times. Fluid-structure analyses are conducted to verify the seismic resistance of ships. In the maintenance of operational functions, platforms are also modernized to meet new threats and capacity requirements.

The integration of these innovations under tight time constraints positions HPC as an essential decision-making tool, ensuring both technical performance and operational availability.

SHORT AND MEDIUM-TERM ISSUES

The challenges of HPC in naval defence are multiple. Growing regional tensions require accelerating production rates, with shorter design phases, while maintaining an equivalent level of reliability and robustness. This involves increasing computing power and ensuring greater flexibility through outsourcing, while matching strict requirements for cybersecurity, technological sovereignty, and compliance with defence standards. Processing sensitive data requires controlled, resilient, and sovereign infrastructures without compromising scalability or operational responsiveness.

Evolving threats and the prospect of high-intensity conflicts are redefining the capability requirements of combat vessels. Platforms must be delivered ready for engagement, interoperable, adaptable, under cost control conditions. One solution involves the development of ship-wide digital twins, made possible by HPC and machine learning. These twins enable virtual simulation, optimization, and validation of the ship's architecture and systems before being built, reducing technical risks and commissioning times. In operations, they provide navies with real-time analysis, diagnosis, and performance optimization capabilities, thereby strengthening the tactical advantage and resilience of naval forces.

The integration of AI into engineering processes and its convergence with HPC, also become essential. AI is transforming every step: automatic generation of experimental designs, assistance with data entry, orchestration of calculations, acceleration of solvers, analysis of results, and data valorization. This evolution requires acculturation of HPC experts to AI, and the recruitment of profiles with dual skills. To the extent that AI is mobilizing a growing share of budgets, redefining technical standards (GPU supremacy, return to single precision), and attracting the most sought-after talents, a marginalization of the traditional HPC sector is plausible. The controlled integration of AI and its fusion with HPC are therefore a key issue for preserving know-how, attractiveness, and competitiveness. More generally, the joint mastery of HPC and AI is an essential lever for ensuring technological sovereignty and naval operational superiority.

Illustration of the digital ship, a true digital clone of an existing vessel. Copyright: Naval Group

NUCLEAR ENERGY: THE ESSENTIAL LINK TO DIGITAL TECHNOLOGY AND HIGH-PERFORMANCE COMPUTING

Patrick Blanc-Tranchant
Deputy Director for Energy Programs, CEA

NEW DYNAMICS AT WORK IN THE NUCLEAR INDUSTRY...

Faced with climate emergency, nuclear power establishes as an essential component of the imperative energy transition of societies, with its capacity to produce massive amounts of decarbonized, controllable electricity and open perspectives beyond electricity, for decarbonizing industrial heat, producing low-carbon hydrogen and even synthetic fuels.

In France, the recognition of its strategic role in achieving carbon neutrality, since 2020, has driven a new political and industrial dynamics aimed at extending operation of the historic park, the deployment of power reactors intended to take over in due time, but also the development of highly innovative reactors benefiting from a new approach to construction, being mass produced in part directly in the factory (SMR type reactors, for Small Modular Reactors).

... REQUIRING STRONGER SUPPORT FROM DIGITAL TECHNOLOGY AND ADVANCED SIMULATION

Modeling, then simulation have accompanied the development of nuclear power since its emergence, but the new dynamics it is experiencing will lead to unprecedented mobilization of these tools and techniques, and of digital technology more broadly.

Data mining and the development of digital twins which rely on techniques and approaches within the field of Artificial Intelligence (AI) will make it possible to optimize predictive maintenance, improve reactor control, anticipate unforeseen events, and more. In short, digital technology will (continue to) profoundly transform industrial practices.

Many new fields of application for modeling / simulation will also arise:

- The extension of existing reactors will require in-depth studies in addition to those conducted to date, to ensure materials and fuel behavior or the mechanical strength of components, for instance.
- New reactors will also induce extremely significant needs for their design, sizing, or to support the necessary safety tests and demonstrations before they come to service. The new fuels used will also create a very high demand for research.
- Finally, the integration of nuclear power into a more diversified energy mix than past solutions will also lead to new demands: the need for more maneuverable reactors, which will have repercussions, particularly for their fuel, for them to contribute to the production of heat or hydrogen, for example, also requiring a great deal of new research.

All these major developments will reinforce the need for high-fidelity simulation of reactors, as close as possible to the underlying physics.

IN PRACTICAL TERMS: HIGH-FIDELITY SIMULATION...

This need for advanced, more predictive simulation will require strengthening so-called «multi-scale» and «multi-physics» approaches: better understanding and modeling elementary physical phenomena, combining them through «up-scaling» approaches, and better integrating physical disciplines, still most often treated separately today for using models specific to each field.

The expected progress must cover all key physical disciplines involved in reactor operation: neutronics combined with nuclear physics, radiation protection, fluid mechanics, structural thermos-mechanics, materials science (simulation of fuel and other plant components), chemistry (particularly corrosion simulation), as well as their coupling, given de facto interdependence of physical phenomena to be modeled.

This is complemented by the desire to make progress in evaluating uncertainties associated with simulations aiming of optimization, by reducing the margins that these uncertainties require us to take today. This will be achieved by new means and approaches which also fall within the field of AI (use of neural networks, meta-models...).

The corresponding work is already being deployed. In the various physical disciplines, more detailed simulations are being implemented progressively: more systematic use of Monte Carlo techniques for neutronic transport theory, detailed CFD-type simulations in thermo-hydraulics for example, far from the historically necessary approximations. And coupling between these different disciplines is also progressing rapidly. But the whole proceedings will have to be considerably amplified to move from calculations that are still relatively limited in size to the possibility of processing the scale of complete reactors.

... AND THE NEED FOR HPC COMPUTING

Continuing to make progress along this way, first to perform reference calculations that should support industrial calculations and, in due time, contribute directly to these industrial calculations which will rely heavily on accessible computing power.

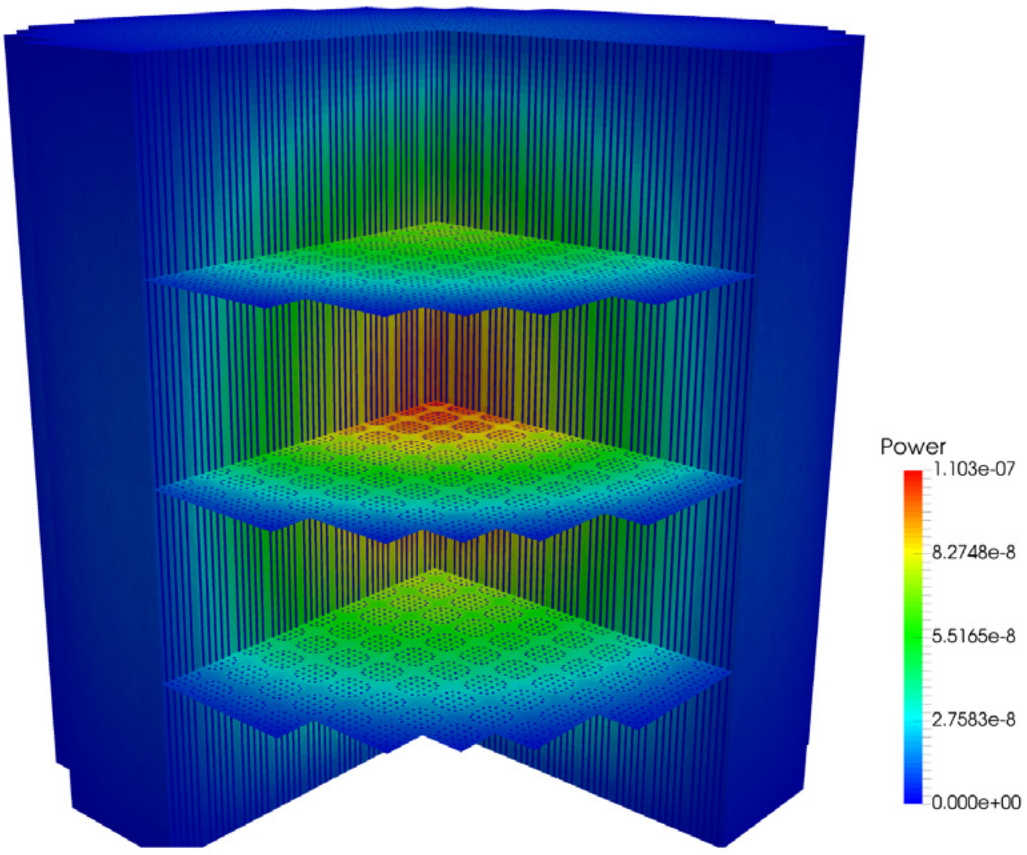
To this end, we assessed the needs for simulating an accidental transient in a PWR reactor (current reactor) with a modeling level disrupting from what is conventionally done today for this type of study, and by closely coupling the three major disciplines of neutronics, thermo-hydraulics and materials/fuel physics, hereto disrupting from current studies which do not deal yet with these three different physics simultaneously.

The assessment showed that to achieve the desired level of precision which would establish a major step forward, around 500 million hours of computing (CPU) would be required; compared to the most precise calculations carried out today that require only a few million hours of computing. Two different orders of magnitude, then.

AND MORE BROADLY?

To broaden the vision, let's emphasize that these enhanced contributions expected from simulation and high-performance computing for nuclear systems will also be needed in the future for other means of energy production or storage (fuel cells or batteries, for example) and, more broadly, to support the current transformations of our global energy system for sizing, management, and performance monitoring.

Tomorrow even more than today, simulation and high-performance computing will be "must have" for energy. ■



Visualization resulting from a detailed calculation of a power pad of a complete reactor core using the PATMOS software, a precursor to future Monte-Carlo simulations in reactor physics at CEA/DES carried out by the TRIPOLI5 code, currently under development (the so-called Hoogenboom benchmark, implemented within the framework of the OECD/NEA - https://inis.iaea.org/records/3jkzd j6f83). The corresponding calculation was a very large calculation on CEA HPC resources, in disruption from previous simulations for which the precision achieved here were inaccessible at the scale of a complete reactor core.

MAJOR CHALLENGES OF TRANSITION TO CARBON NEUTRALITY AND THE ROLE OF DIGITAL TRANSFORMATION

Alain Martin
Scientific Information System Manager, EDF R&D

Stéphane Tanguy
Systems and Information Technologies Director, EDF R&D

Background: Scientific computing and simulation are essential tools for ensuring the proper functioning of the electricity system today and for building its future.

For over 30 years, EDF has relied on scientific computing and simulation to guarantee the safety and performance of its facilities, to balance supply and demand with highly reliable consumption forecasts and meet the multiple challenges of the energy transition.

While EDF operates its own supercomputers (HPC) ranked among the 100 most powerful in the world, the Group's R&D has also developed its own portfolio of simulation codes that model the different physics and ensure the representativeness of our simulations. Some of these major codes are available as open source and used by a scientific community that extends far beyond France.

We have more recently introduced Artificial Intelligence algorithms into our codes to produce results more quickly and simply without significant loss of accuracy (for example, developing a metamodel to provide a heat map that helps prioritize welding repairs).

For several years now, we have been closely interested in quantum algorithms which will play a major role when large-scale quantum computers become available that will allow us to deal with problems where classical simulation only provides approximations.

SUPERCOMPUTER: KEY TOOL AND CORNERSTONE OF R&D FOR ENERGY MODELING

With computing power exceeding 11 petaflops (or 11 million billion operations per second), EDF's supercomputers are a key lever for R&D and are central for EDF's innovation efforts. They enable the company to model complex physical phenomena with greater precision while relying on a dedicated experimental base to validate and qualify its own calculation codes, hence ensuring the reliability and relevance of its simulations.

For example, these resources are used for:

- Simulating the physical phenomena that govern the performance and safety of our production resources (nuclear, thermal, hydraulic, or renewable) using field physics codes such as code_saturne, neptune_cfd, code_aster, and telemac;
- Forecasting consumption and optimizing production resources to ensure a constant supply-and-demand balance;
- Anticipating the impacts of climate change to ensure the continued operation of our current resources for decades to come.

AI ALREADY WIDELY DEPLOYED IN THE ELECTRICITY SYSTEM

AI has become a key strategic lever for EDF. Initially mastered by R&D, it has gradually been deployed across the Group, notably through the Data Analytics

Factories which support business lines in developing high-value use cases for the following needs:

With the rise of generative AI, EDF is exploring new opportunities while ensuring responsible and efficient integration of these technologies to guarantee their transparency, security, and low energy impact.

QUANTUM COMPUTING: A BREAKTHROUGH PROMISING TECHNOLOGY FOR EDF

Quantum computing represents the future of scientific computing and the promise of major scientific advances unattainable by classical computing. It also represents a complete breakthrough from an algorithmic perspective, making it essential to anticipate the porting of our simulation codes to this new technology. EDF is currently exploring several categories of promising use cases:

- Optimization: improving smart charging control for more efficient management of electrical networks;
- Materials: molecular-scale modeling to better understand the aging of materials under irradiation and optimize the lifespan of batteries and photovoltaic cells;
- Physics simulation using partial differential equations (PDEs): significant performance gains and access to simulation results unattainable with conventional approaches.

HPC CRONOS in EDF Data Center

In collaboration with academic and industrial partners, EDF is anticipating the integration of quantum computing into its simulation methods and infrastructure to remain at the forefront of innovation. This is a strategic challenge, both to strengthen its competitiveness and to meet demands of a rapidly changing energy sector.

AI AND QUANTUM COMPUTING ARE ESSENTIAL LEVERS FOR MEETING THE CHALLENGES OF OUR FUTURE ELECTRICITY SYSTEM

As we have seen, digital simulation has been an essential tool for ensuring the performance and safety of our electricity system yet, new challenges associated with the energy transition, climate change and the decarbonization of our economy will require more than additional PFlops. It is now a question of producing new, more complex models using previously unexploited data, which will only be possible with the contributions of artificial intelligence and quantum computing. ■

HYBRID AI FOR CONNECTED, SUSTAINABLE AND RESILIENT CITIES AND TERRITORIES

Francisco Chinesta
University Professor, Arts et Métiers Institute of Technology

Dominique Baillargeat
Scientific Executive Director
CNRS@ CREATE (Singapore)

Victor Champaney
Co-founder,
Duoverse

The main characteristic of a city or a territory, in terms of predictability, is its extreme complexity.

The existence of numerous interconnected and interacting systems, of large spatial and temporal dimensions, presenting different granularities (systems of systems) and subject to uncertainties and imprecisions that propagate within, makes modeling particularly complex. If we add to this the main protagonists, humans, whose behavior eludes deterministic models, the hope of achieving the desired level of diagnosis, prognosis, and support for decision-making is considerably limited.

MODELING FOR BETTER PREDICTION

Models from engineering sciences, designed to predict the behavior hence performance of a component or a multicomponent system, have experienced colossal progress, both in precision and efficiency, the latter being dependent on equally considerable advances in computing capacity, driven by the democratization of high-performance computing (HPC).

Discretization techniques have also played a key role, enabling the efficient solution of increasingly large models, both spatially and temporally. At the end of the last century, model reduction techniques made it possible to solve, often in real time, large-scale, multi-physics, multiscale, nonlinear, coupled, and complex problems. Today, speed and precision are no longer incompatible!

At the beginning of the 21st century, physical systems have increasingly been digitalized, aiming at fast and accurate prediction to anticipate and make optimal decisions, both during the design and operation of engineering systems.

In such a context, the digital twin emerged as a virtual replica of the observed reality. Sometimes used early in the design phase, it draws on assembled knowledge, historical data, and the contributions of decades of

Fig. 1: Augmented reality on model (in collaboration with Duoverse and IMMERSION)

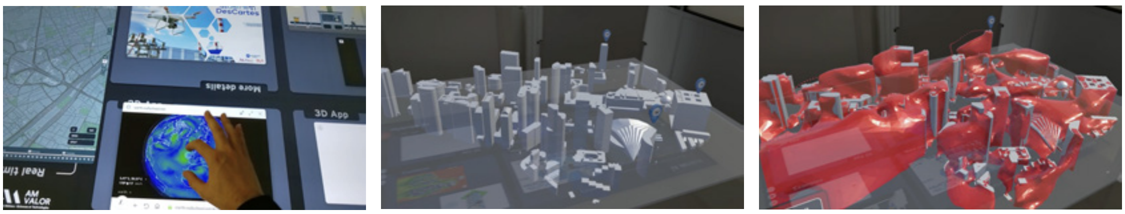


Fig. 2: Virtual reality on a collaborative engineering table (in collaboration with Duoverse and IMMERSION)

engineering science. Other times, it aims to optimize the operation of systems already manufactured and in use throughout their life cycle. This reinforces the two fundamental missions of an engineer: designing and operating. After four major revolutions – steam, electricity, electronics (automation), then data with AI – the fifth will integrate humans, who were traditionally treated as observers instead of actors within cities and territories. It is no longer a question of trying to automate city operations, such an attempt already taken and quickly abandoned facing failures, but of developing technologies with a low environmental and energy footprint, transparent to the user, and dedicated to the citizen: for their comfort, pleasure, well-being, safety, and efficiency.

FROM DIAGNOSIS TO ACTION: TOWARDS A HYBRID DIGITAL TWIN

Such technology will have three levels of action: (I) remote action triggered by a diagnosis and prognosis; (II) support for decision making for authorities based on real-time analysis of countless scenarios; and (III) information available to citizens allowing to adapt their choices, both in daily life and in emergency situations.

However, while diagnosis can be achieved through simple data analysis using AI techniques, prognosis requires projection into the future, a model-specific capability!

However, current modeling applied to complex systems very often reveals a notable gap between predictions and observations. Reality seems to contain elements that models fail to capture due to inaccuracies or uncertainties, that are sometimes epistemic in nature.

The «data-centered» approach is also not that viable, as it would require a massive amount of data to describe cities or territories with fine granularity. Imagine having to install millions of sensors in every building, every street, every infrastructure...

Both paradigms show their limitations, speaking for one hybrid modeling paradigm. Sensors should be retained, but in small number and placed in optimal positions. To achieve this, models derived from physics and knowledge are a valuable aid! Then, supplement the data with engineering science models that offer an estimate – admittedly not perfect, but not incoherent. Exploited by AI, the data will thus enrich these basic models to achieve a so-called hybrid model. A compelling combination of precision, speed, frugality (data conscious), and explainability (essential for certification)!

This twin, called hybrid, ensures faithful design, but also new management of systems breaking free from old modeling paradigms.

DESCARTES, THE AUGMENTED CITY LABORATORY

In the context of cities, the DesCartes project (https://descartes.cnrsatcreate.cnrs.fr/) led by the CNRS@CREATE subsidiary of the CNRS in Singapore, within the CREATE campus of excellence, and funded by the NRF (National Research Foundation) proposes a hybrid modeling methodology for several key themes: (I) the environment; (II) structural monitoring; (III) smart grids; (IV) mobility, particularly around drones; and (V) crisis management targeting cities, ports, and airports.

This methodology is designed to be adopted in other cities and territories, beyond the Singaporean context.

Thus, parametric modeling of wind in an urban environment, with metric resolution from wind data at altitude (direction and intensity) and from a few weather stations, makes it possible to estimate several key phenomena: the optimal trajectories of drones, temperature distribution with the identification of urban heat islands, air quality from real-time traffic data, or even the dispersion of pollutants from an harbor (e.g. ships) or accidental, hazardous emissions from industrial sites.

The impact of these phenomena can thus be estimated in real time with great precision to enable appropriate action.

Such hybrid modeling requires the convergence of several technologies: (i) parametric model reduction techniques applied to physical phenomena such as structural mechanics, flows, heat transfer, or electromagnetism (for example, for satellite or 5G coverage); (ii) machine learning methods used to enrich physics-based models; (iii) data assimilation techniques allowing models to be calibrated using real observations.

All of this relies on powerful high-performance computing (HPC) infrastructures, and more recently, on the capabilities of quantum computing (noticeably quantum annealing) for extracting reduced models, solving inverse problems (such as source identification or root-cause search), multidimensional assignment, and even coordinating collaborative agents. Results gained in these areas are already very promising to date. ■

Fig. 3 - Immersive virtual reality: Model (top), Temperature and Pollution (center and bottom) – produced in collaboration with Duoverse and IMMERSION

ACCELERATING THE DISCOVERY OF NEW MATERIALS: CONVERGENCE OF HPC AND AI

Jean-Yves P. Delannoy
Scientific Director Digital R&D, Arkema

The discovery and development of advanced materials have traditionally been a time-consuming and resource-intensive endeavor, reliant on trial-and-error experimentation. However, this paradigm is rapidly changing due to the convergence of High-Performance Computing (HPC) and Artificial Intelligence (AI). These technologies are transforming materials innovation and are vital to addressing global challenges in energy, sustainability, and manufacturing. Historically, it could take decades to go from initial concept to market deployment of material. Today, the integration of HPC and AI offers a path to accelerate this process significantly, reduce development costs, and explore areas of the chemical space previously inaccessible.

THE EVOLUTION OF MATERIAL DISCOVERY

Traditional materials discovery relied on manual, sequential, and labor-intensive workflows. Researchers synthesize and test multiple candidates individually, then iterate this process using their results. While fundamental, this method is limited by human intuition, resources, and the vast accessible space of materials combinations. Now, automation, parallel experimentation, and AI-driven modelling and simulation are reducing this complexity[1]. Despite remaining challenges – such as ensuring reproducibility – these innovations are transforming materials science. High-performance computing, multiscale modeling, and AI are central to this change. Advanced computational tools enable high-throughput screening of vast chemical spaces, improving predictive materials design.

HIGH-PERFORMANCE COMPUTING AND SIMULATION

HPC allows researchers to simulate material properties with increasing accuracy. Ab initio methods, based on quantum mechanics, give access to atomic interactions and materials properties before synthesis, reducing costs and increasing research efficiency. Molecular dynamics simulations add powerful insight into material behavior. These tools are particularly effective in the fields of energy storage and catalysis. With the introduction of clouds, access to supercomputing has been expanded and democratized, accelerating innovation. In polymer science, simulations now guide macromolecular design and reduce laboratory work. However, achieving a balance between precision and computational efficiency and speed remains a significant challenge. Coarse-grained models helps to study complex systems over longer timescales, while AI is pushing the boundaries of what is computationally feasible and interpretable in materials science[2].

ARTIFICIAL INTELLIGENCE IN MATERIALS DISCOVERY

AI is transforming materials discovery by generating new hypotheses, designing new materials, and accelerating decision-making[3]. With the rise of generative AI and large language models (LLMs), tools such as Microsoft's MatterGen and Google's GNoME (Graph Networks for Materials Exploration) can propose new molecular structures optimized for specific properties.

Conventional machine learning (ML) models also contribute by enabling rapid property prediction, as demonstrated in the development of sustainable polymers. ML models — particularly deep learning architectures — are highly effective at capturing complex relationships between structure and properties. These models significantly reduce the experimental search space, decrease the discovery time, and improve reproducibility.

Importantly, AI mitigates the traditional tradeoff between simulation accuracy and computational speed. This paradigm shift positions AI as a key enabler in addressing critical challenges in energy, sustainability, and advanced manufacturing.

HPC AND AI: ACCELERATORS OF NEW MATERIAL DISCOVERY

The convergence of High-Performance Computing and Artificial Intelligence marks a transformative era in the discovery of innovative materials. Together, they offer unparalleled capabilities to design and develop materials with tailored properties at an accelerated pace.

By harnessing the computing power of HPC and AI, researchers can now navigate vast chemical landscapes with previously inconceivable efficiency. As these technologies mature and become increasingly accessible, organizations that adopt them will position themselves at the forefront of next-generation materials innovation.

Rather than choosing between HPC and AI, the future of materials science lies in leveraging both strengths to accelerate the transition from concept to commercialization of innovative solutions. ■

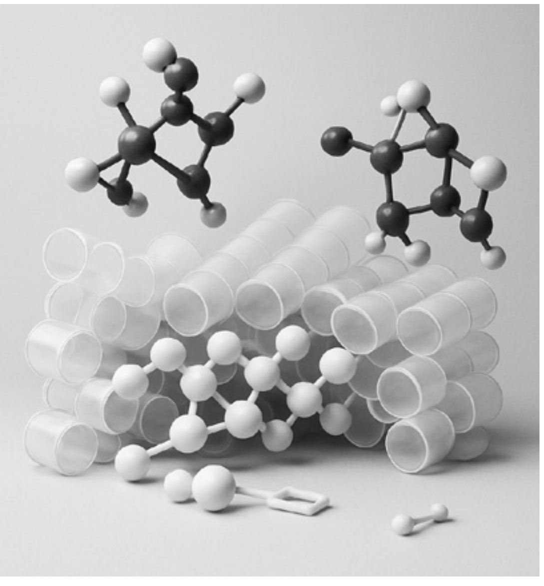


Fig. 1: Artistic Illustration of cellulose acetate plasticization - Generated using ChatGPT 4O

References

1. Edward O Pyzer-Knapp, Jed W Pitera, Peter WJ Staar, Seiji Takeda, Teodoro Laino, Daniel P Sanders, James Sexton, John R Smith, and Alessandro Curioni. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Computa tional Materials, 8(1):84, 2022.

2.. Thomas E Gartner III and Arthi Jayaraman. Modeling and simula tions of polymers: a roadmap. Macromolecules, 52(3):755–786, 2019.

3.. Yuanqi Du, Arian R Jamasb, Jeff Guo, Tianfan Fu, Charles Harris, Yingheng Wang, Chenru Duan, Pietro Liò, Philippe Schwaller, and Tom L Blundell. Machine learning-aided generative molecular design. Nature Machine Intelligence, 6(6):589–604, 2024.

AI, HIGH-PERFORMANCE AND QUANTUM COMPUTING TO SERVE TOMORROW'S FORCES

Emmanuel Chiva
General Delegate, Ministère des Armées

Michaël Krajecki
Head of Research Department, Ministry Agency for AI in Defence

20 YEARS ALREADY!

Teratec has become a key French and European player in digital simulation, big data, and AI! So yes, the keywords of the decade are AI, deep learning, LLM, digital twins... All these advances have burst into our daily lives and are the source of numerous technical, organizational, and ethical challenges. The foresight exercise is always difficult, and it is often easier to look in the rear mirror rather than anticipate the next bend down the road. The last decade has been very rich for Teratec; being spectators a few times, we were most often actors in these transformations. The Ministry of the Armed Forces has been strongly mobilized at your side: through the Defense Innovation Agency, it supports numerous projects led by researchers from our laboratories, by start-ups as well and even major manufacturers, all of whom contribute to Teratec's influence.

A dual technology – surviving several winters – AI has undergone its revolution thanks to the conjunction of three factors. First, information is widely available and natively produced in digital form, allowing the creation of gigantic data corpus. Yet above all, widely accessible computing power allows the implementation of learning mechanisms on which AI developed over the last decade. Finally, recent algorithmic advances, among which we can highlight deep learning or even adverse networks and attention mechanisms, have led to the development of new intelligent and often embedded applications. Within the Ministry of the Armed Forces, the Ministerial Agency for Defense AI (AMIAD) produces and develops AI technologies for the benefit of the forces. The General Directorate of Armaments (DGA) is leading the implementation of these AIs being central to weapons systems.

While AlexNet had already revolutionized the world of imaging in 2012, the advent of GPUs was confirmed from TOP500 to the next. In November 2013, the University of Reims – a member of Teratec – ranked its ROMEO in the TOP500 and in the GREEN500: for the first time, a French accelerated supercomputer was available for academic research. In 2017, while Stanford published the first edition of the AI Index Report, MP Cédric Villani, Fields Medalist in 2010, was tasked with leading a mission on the implementation of a French and European strategy in AI, still acting today as part of France 2030. In the summer of 2022, researchers from research organizations and the start-up world will benefit from the Jean Zay supercomputer to develop the GPT3 BLOOM model. By the end of 2022, ChatGPT had reached 1 million users in 5 days, and the momentum is growing. In France, Mistral – a company founded in April 2023 – offers so-called open-weight models.

The convergence of data, algorithms, and computing power has revolutionized our world in a dizzying way in just a few years!

This revolution is boosted by another, with quantum: sensors, cryptography, and computing. From an operational perspective, the DGA is focusing its efforts on quantum sensors (atomic clocks, cold atom gravimeters). However, regarding computing, the Ministry of the Armed Forces believes that civilian research will be able to accelerate and meet its needs. As early as 2018, ATOS offered its simulation solution, the QLM (Quantum Learning Machine) which made it possible to anticipate the arrival of the first quantum computing accelerators.

Since 2022, the HQI (Hybrid HPC Quantum Initiative) project has been offering a hybrid computing platform combining several quantum technologies with GENCI's Joliot Curie supercomputer at the TGCC (CEA). This project aims to foster the emergence of new use cases that can take advantage of quantum computers. This second revolution should certainly see new applications arising in the coming years. Teratec's academic and industrial stakeholders are at the heart of these revolutions and benefit from the support of France 2030 to promote the science developed in France, Europe, and throughout the world.

HOW THE MINISTRY OF ARMED FORCES OWNS THESE ISSUES?

AI technologies will play a leading role in the future operational superiority of the armed forces. However, the world of defense has specific characteristics that justify our continued investment in the development of AI for use in our armed forces and the transformation of the Ministry of the Armed Forces.

In addition to conventional considerations (ethics, employment framework, doctrine), the military environment presents specific oddities. Data scarcity: Data is scarcer in the defense sector, particularly sensitive, and often relies on sensors more specific to the military environment, such as radars or sonars. Beyond the sovereignty issue posed by public clouds, those imply hyper-connectivity systems quite unrealistic in Defense.

In 2025, the Ministry commissioned the most powerful classified supercomputer dedicated to AI in Europe. Located at the Mont-Valérien in Suresnes, the latter allows France to process confidential data sovereignly for the needs of the armed forces and defense companies. It is not connected to the internet, and its maintenance is carried out by French citizens entitled to maintain national defense secrecy.

Similarly, quantum applications for defense are many: inertial measurement systems, new encryption approaches, and computational applications. Hence, the PROQCIMA program, funded by France 2030 and led by the DGA, aims to develop in France two prototypes of fault-tolerant universal quantum computers, each equipped with 128 logical qubits, ready for industrial scale. Six French start-ups are now in the race to develop such a system like Pasqal, Alice and Bob or even Quandela...

WHAT ABOUT TOMORROW?

For the Ministry of the Armed Forces, strengthening its skills in the field of quantum and AI remains one utmost priority: from 2027, focusing on upstream research, the ministry will unite activities related to research in AI and quantum for defense in the Defense Research Institute together with the CIEDS (Interdisciplinary Center for Studies for Defense and Security) in a new totem building of the Polytechnic School.

Tomorrow, robots will come out from factories to interact in an open environment directly linked to the real world. The robot will become a new extension for AI in our world. In March 2025, the Minister of the Armed Forces announced the Pendragon project aiming to develop the robotic combat unit of tomorrow. Entrusted to AMIAD and the future Command to Combat, this experimental unit will be capable of conducting autonomous combat tactical actions to fulfill various missions, intended for rapid adoption by the forces and industrialization, starting in 2027. Keep in mind that AI continues to develop and needs exceptional computing resources to be ever more useful and precise. AI is algorithms and data. The Ministry of the Armed Forces will be at your side to bring French digital innovation to the highest international level. AID, AMIAD and DGA will mobilize to ensure that our Armed Forces have the best digital technologies to carry out their missions. From Jean Zay to the European Giga factories, Teratec players have their right in these new challenges! ■

Illustration of the PENDRAGON project – credits: Ministry of the Armed Forces

HIGH-PERFORMANCE COMPUTING TECHNOLOGIES FOR THE INSURANCE AND FINANCE SECTOR

Christophe Michel
Validation and Risk-Model Manager, Crédit Agricole CIB

THE FUNDAMENTAL LINK BETWEEN HPC AND FINANCIAL RISK MANAGEMENT

High-performance computing is a strategic pillar for the financial industry. The very nature of banking and insurance businesses relies on sophisticated multidimensional risk management: property damage risks, market risks affecting the valuation of balance sheet items, counterparty risks, operational risks, and liquidity risks.

The assessment and management of all these risks rely on mathematical and actuarial models of which implementation requires significant computing resources, both for data processing and for performing complex calculations.

The post-2008 crisis regulatory framework has significantly intensified these computing needs. The introduction of directives such as Solvency 2 for insurers and the Fundamental Review of the Trading Book for market activities imposed new risk indicators that are particularly demanding in terms of computing power.

A prime example relies to market activities. Market risk management is primarily focused on derivatives. Their valuation and hedging rely on intensive digital methods whose computational complexity increases with the sophistication of the products and the accuracy of the models. This sophistication meets two major competitive imperatives: on the one hand, precise adaptation to customer needs and, on the other hand, the performance of risk management models, which is crucial for the competitiveness of the business.

ARCHITECTURAL PRINCIPLES

Faced with these challenges, the architecture of computing systems must be organized around structuring principles. Separating functional and technical aspects and optimizing resources through orchestration capable of keeping computation close to the data to minimize latency and save potentially reusable data that is costly to compute.

To enable optimization at a global level, it is necessary to get an abstract of computational services in the form of tasks, data, and the interactions between the two. Finding its most suitable mapping is a fertile source of highly material optimizations.

Optimization strategies must adapt to different usage and contexts. Priority will be given to optimizing the duration of recurring processing operations with maximum use of available resources, or to unit response swiftness for real-time processing operations.

In any case, the limiting factors remain computing resources, memory, and their interactions. Orchestration must efficiently manage tens of thousands of computing cores distributed across different sites and cloud environments, while optimizing an overall cost function that incorporates the carbon footprint – a considerable technical challenge of its own.

THE EMERGENCE OF AI

Artificial intelligence is gradually transforming the appreciation of calculation in the financial sector. More recently, use cases with generative AI are spreading across the various service activities within the financial industry.

AI models themselves are resource intensive. This approach presents several distinctive characteristics, such as the use of standardized algorithms via external libraries, primarily in Python, and the use of specialized hardware accelerators (GPU/TPU). These characteristics reinforce the need to rely on architecture designed to adapt to the context of use.

Beyond its business applications, AI is a powerful technique for optimizing the overall distribution of computations, in hybrid environments including cloud. It can also be trained to learn the most computationally expensive business functions to drastically reduce their execution time.

Integrating AI into an information system raises new challenges. Controlling functions that evolve through their very own use or managing the often-unclear scope of validity of these functions gets tricky in practice. These considerations serve to guide the choice of first preferred use cases for AI. It is particularly suited to replacing heuristics for solving complex problems. However, vigilant caution is required when rigorous precision is imperative.

OUTLOOK: QUANTUM COMPUTING

Quantum computing appears to be full of promise. The first operational machines already make it possible to explore specific applications that exploit quantum properties to solve certain categories of issues with some potentially revolutionary efficiency.

The first operational use cases in the financial industry have emerged in the areas of combinatorial portfolio optimization and regulatory capital optimization. Many other avenues are currently being explored: for example, quantum algorithms applied to solving PDEs, to clustering, and AI.

This technology is already transforming more standard approaches with the exploitation of tensor networks and quantum-inspired methods. It could constitute the next computing paradigm for certain financial applications, complementing the arsenal of HPC and AI methods within the sector's technological ecosystem. ■

THE CLOUD: CORNERSTONE OF INNOVATION FOR TODAY AND TOMORROW!

Yaniv Fdida
Chief Product and Technology Officer, OVH Cloud

Industry experts all understand that digital technology follows cycles of innovation. These usually have the effect of redefining the ecosystem and landscape entirely. For example, at the end of the 1990s, the emergence of the internet led to the dotcom boom, with the birth of OVHcloud in 1999! Since then, there have been many other waves including smartphones, which put the computing power of computers in our pockets or on our wrists. The cloud also represents one of those cycles, allowing remote access to data, wherever the user is – and is accompanied by the general trend away from the physical and towards the digital. We see this trend in music, in video and in photos, for example. Even though it may be invisible to most people, the cloud is an essential part of our daily lives and consumption patterns.

More recently, the AI wave has been ever-present in the news. The rise of ChatGPT has plunged us into a new cycle whose full consequences have yet to be seen. One thing is certain: artificial intelligence is here to stay, and its impact is starting to have a long-term impact on certain fields already, from transportation to health and robotics. However, with the integration of AI into so many walks of life has come the need for increased vigilance. Data is now the black gold of AI. Data is used to train AI, and this development has meant a heightened need to use data responsibly and ethically. As an industry, we must ensure that this technology is on the right path and make sure that the uses of AI and how AI handles data are acceptable to all.

Looking ahead, we believe quantum computing will be the next disruptor for solving things that we consider unsolvable today. Although it is early days for quantum, Europe is already standing out, thanks to the innovative companies across the region, and the incredible diversity of approaches. Our founder, Octave Klaba, is fond of saying that the twentieth century was atomic, and the twenty-first will be quantum. With this in mind, we are working to use our knowledge and skills to connect the first quantum computers to the cloud and prepare the ecosystems of tomorrow, helping to bring their vision and computing power to as many people as possible.

HOW IS OVHCLOUD TAMING ARTIFICIAL INTELLIGENCE?

Since 2017, OVHcloud has been weaving artificial intelligence into its solutions. We have been integrating the principles of machine learning into our DNA for applications including predictive maintenance and anti-fraud. This rigour means that we have an excellent understanding of how to democratize AI. By learning and applying the technology internally, we have now developed various simple and affordable services, enabling the acceleration of AI usage for our customers. Customers across the world can now benefit from support with model training to the deployment and provision of off-the-shelf open-source models for developers.

However, we always hold our ethical principles close. As we said, data is the black gold for AI and it is clear that there can be no compromise in how AI is deployed and used, especially when it comes to data sovereignty. It is with this clear goal in mind that we have developed a whole set of services to store, handle, ingest, and transform data in an ethical and trustworthy fashion.

It's also important to talk about AI's hardware requirements. GPUs are important, but that's not all: AI requires the co-operation of both CPU and GPU, as well as high-bandwidth network and storage capabilities to ensure low latency.

While artificial intelligence may rely on dedicated accelerators in the future, particularly for inference – which is something we're currently looking at – GPUs remain essential at present. This is as exciting

THE CLOUD: CORNERSTONE OF INNOVATION FOR TODAY AND TOMORROW!

Yaniv Fdida
Chief Product and Technology Officer, OVH Cloud

• • •

as it is challenging because GPUs have high energy consumption and as a result, generate significant heat. The environmental impact of GPUs is anything but trivial and datacentres have had to adapt to this to cater to AI's needs.

Since 2003, OVHcloud has cooled its servers with a closed-loop water circuit. This provides superior cooling and reduces the electricity and water consumption compared to the air conditioning systems widely used in traditional datacentres. Our teams in R&D have worked tirelessly to develop different generations of water-cooling systems and have perfected solutions to cool up to 8 GPUs per server.

CLOUD AND SOVEREIGNTY IN 2025

Finally, the current geopolitical instability today reinforces the relevance of OVHcloud's strategy and the Group's positioning on a key issue: the need for states, particularly in Europe, and for all organizations, to strengthen their digital sovereignty. Since it was founded, OVHcloud has been built on sovereign and open solutions. Europe's dependency on American-based solutions can have serious consequences, particularly on our economic activities, our privacy and, more generally, on our strategic autonomy. By securing its critical digital infrastructures, strengthening its capacity for innovation and choosing to promote an open and trusted cloud, Europe will be able to better face the current geopolitical arena. All decision-makers, public and private, as well as the investment community have understood this: it is now time to act. ■

SOCIETAL PERSPECTIVES

INTRODUCTION:
SOCIETAL ASPECTS: THE GOOD AND THE BAD

Jean-François Prevéraud
Journalist

If digital technologies have long been used in virtually all sectors of activity, their impact on society has been limited until recently. Admittedly, some jobs have had to evolve and adapt with necessary investment in training. But this contributed to greater comfort in the workplace.

On the other hand, the emergence and subsequent democratization of the internet, followed by generative artificial intelligence more recently, have changed the game with an overall impact on society.

A CHANGE IN LIFESTYLE

In the early 1990s, technological advances led to the arrival of consumer web applications, thanks to powerful online search engines (Mosaic, Netscape, Yahoo, Lycos, Altavista, Google, Internet Explorer...). Knowledge was no longer necessary; all you needed to know was how/where to look for it. And the integration of AI into search engines makes things much easier now, even anticipating requests.

The world of commerce was one of the first to be affected by the democratization of digital technologies. Platforms such as eBay, Amazon, Alibaba, Rakuten, and others have gradually taken over a large share of global commerce, driving many brick-and-mortar enterprises to bankruptcy. And today, services are undergoing the same evolution with online banks for instance, leading physical branches in retail banking to shut down.

As for the world of communication, it is undergoing similar revolutions. Mobile phones (GSM) have already virtually replaced landlines. Pushing the power of the Internet in our pocket entails everyone to access, create, and share information in real time, particularly via professional networks (e.g. LinkedIn...) or consumer social medias (Facebook, Instagram, Snapchat, X-Twitter, WhatsApp, YouTube, TikTok...) with derivative applications such as online games, music streaming or dating sites.

STAYING VIGILANT

Great ideas now realized, but some people are exploiting it to engage in targeted insidious marketing or worse, spreading fake news to promote harmful ideas threatening democracy. Social media has an undeniable impact as well on the most vulnerable including youngsters, as we can see from the devastating effects of cyber-bullying and cyber-addiction. So much so that governments now consider ever new regulations to be put in place.

Finally, let's not forget the “digital divide” that is widening between those who have access and use digital technologies daily and those who cannot based on their educational capacities and/or income.

Nevertheless, these digital technologies still have a positive impact on society at large. Just look at the effect of AI on education which is changing the way we teach and learn. Our experts outline these positive aspects in the following pages. ■

EDUCATION IN THE FACE OF THE DIGITAL WAVE: NEW KNOWLEDGE FOR NEW CHALLENGES

Jean-Christophe Jouhaud
Training & Education Manager, Cerfacs

The rapid rise of Artificial Intelligence (AI), High-Performance Computing (HPC), and quantum technologies is profoundly transforming skills requirements across all sectors. This shift is accompanied by a growing shortage of qualified profiles in both public and private sectors. This imbalance results from technological acceleration outpacing the adaptation of training programs, reinforced by intense global competition for expert talent.

To address this tension, several structural levers must be activated: the development of professional training pathways, the co-construction of curricula with companies, revaluation of scientific fields of study starting in secondary education, and the integration of hybrid AI/HPC/quantum modules into existing programs.

Europe has recognized the importance of this strategic issue by placing digital training at the heart of several major initiatives: the Digital Europe Program, EuroHPC JU, EuroQCI, EIT Digital, and the ELLIS network. The goal is clear: to build a European foundation of high-level digital skills.

France is actively contributing to this momentum, noticeably with two key projects:

- EuroCC2/CCFR (Teratec, Cerfacs, INRIA Academy, CRIANN, and ROMEO) which brings more than 30 national HPC competence centres in Europe together;
- EDIHs (European Digital Innovation Hubs) supporting SMEs and public actors in adopting digital technologies through training, experimentation, and strategic support.

These hubs take an essential role in continuing education, sectoral awareness (health, agriculture, industry), and the dissemination of digital skills across regions.

In terms of initial training, France is involved in the EUMaster4HPC program, a European master's degree of excellence run by several institutions (Sorbonne University, University of Strasbourg) and partners linked to the industrial world, such as Cerfacs.
This curriculum aims to train a new generation of specialists capable of combining scientific, technical, and applied expertise in HPC.

To meet the challenges of scale, specialization and agility, new teaching methods are essential. Distance learning plays a crucial role in this effort:

- MOOCs introduce a wide audience to acquire the fundamentals;
- SPOCs offer in-depth, supervised courses;
- Hybrid training combines digital content, practical workshops, and remote access to infrastructure such as HPC clusters or quantum platforms.

Delivering better, faster, and more widely accessible training is now essential for competitiveness and technological sovereignty. ■

UNPRECEDENTED STAKES IN A WORLD UNDERGOING PROFOUND CHANGE

Thierry Damerval
Special Envoy for science, technology and innovation
Ministry for Europe and foreign affairs
Ministry for higher education and research

2005-2025: WHAT UPHEAVALS IN THIS RAPIDLY CHANGING WORLD!

Let us remember that twenty years ago, discussions in Europe focused on the knowledge economy and the goal of devoting 3% of GDP to research and development. Nowadays, the defence economy and the share of GDP to be devoted also come to the agenda. The scientific world itself has undergone profound changes. Until the early 2000s, the scientific world had seen a certain stability in terms of the nation's ranking since the creation of the ISI (Institute for Scientific Information) in 1957. The United States was by far the leading scientific nation with four English-speaking countries (the United States, the United Kingdom, Canada, and Australia) accounting for more than 40% of global scientific output, hosting most students and postdoctoral researchers from around the world.

This global mapping of science has changed profoundly over the past twenty years, with new balances as new players arise. The United States' share of global publication output has fallen from 28% to 14%. China's share now stands at 20%, up from 3.9% in 2000, moving from 6th to 1st rank globally. India has risen from 13th to 3rd position. These are strictly quantitative indicators of course, subject to debate and criticism. A particular issue is the increase in the annual number of publications which has risen from 800,000 in 2000 to nearly 3,000,000 in 2023, a growth rate never seen before in the history of science, also pointing out poorly regulated practices. Such increase being much higher than the growth of researchers' population worldwide should be linked with the race to publish and raises many questions about the conduct and evaluation of research findings. There is also a global rise with issues related to quality, reproducibility, and fraud. The rate of article retractions due to proven errors or fraud has increased tenfold in 20 years.

In this evolving context, where the future is difficult to predict (as recent US presidential elections strikingly reminded us), two elements stand out that have marked the last two decades with their major challenges, both scientifically, technologically, economically, and diplomatically:

- The United Nations 2030 Agenda and the 17 Sustainable Development Goals adopted in 2015: after seminal reports (the Meadows report in 1972, the Brundtland report in 1987, and the first IPCC report in 1990), collective awareness and action have been slow to materialize.
- The extremely rapid rise of information and communication technologies and related sciences and technologies (high-performance computing, data management, artificial intelligence, quantum technologies).

The applications already in use and the potential of AI are considerable. They affect all areas of social and economic life, which is extremely rare in the history of science and technology. The health sector will undoubtedly be one of the major beneficiaries. They will also change education and affect all areas of research, monitoring activities, the design of research projects and even their evaluation, the production of synthesis, and scientific expertise. However, many questions arise at the international level regarding the regulation of uses, in a context where private actors prime over the public sphere and where the rapid pace of developments boosts uses even before the regulatory framework has been defined.

A TECHNOLOGY WITH SIGNIFICANT GEOPOLITICAL IMPLICATIONS

Vladimir Poutine declared in 2017 that “whoever dominates AI will dominate the world.” In 2020, the American think tank Brookings Institution echoed this insight, stating that “whoever is the leader in artificial intelligence in 2030 will dominate the world until 2100.” China has set for itself the goal of owning the entire AI value chain on its territory by 2030.

AI is primarily American, with major private players (GAFAM) and new players such as OpenAI playing a significant role, adding up with the volumes of data that facilitated deep learning and hardware infrastructure. The geostrategic stakes are obvious when one considers that the main chip manufacturer and supplier for the American company Nvidia is the Taiwanese company TSMC. China is obviously the other dominant power, with extremely rapid growth in the volume of data controlled, most of which is out of reach.

Alike Germany, Canada, Japan, and South Korea, France takes part in a group of intermediate powers, none of which can claim to play a dominant role on its own. For us, the challenges and opportunities are playing out within a European framework.

WHAT INTERNATIONAL REGULATION?

When it comes to regulating usage, it is interesting to draw a parallel with another major scientific revolution: the emergence of genetic engineering and its applications in the biomedical field. Faced with the potential applications of these new technologies, the global scientific community expressed its concerns at the Asilomar Conference in 1975. This mobilization led to the creation of UNESCO’s International Bioethics Committee and, in France, to the creation in 1983 of the National Consultative Ethics Committee for Life Sciences and Health. The 1980s saw numerous debates and discussions involving various stakeholders at the national, European, and international levels. A framework and limits were then established, such as the absence of intervention on germinal lines and the ban on cloning, a framework that has been respected internationally ever since.

An Asilomar conference did take place in 2017 on artificial intelligence, yet without the same impact as the one in 1975. At the European level, the AI Act clearly sets out a framework. Will it increase confidence in models developed within such boundaries and contribute to stability for developers and customers of AI systems? These are among major issues for Europe.

At the AI Action Summit, held in Paris in February 2025 and co-chaired by France and India, the declaration for trustworthy, sustainable, and inclusive AI was signed by 64 states, including India, China, Canada, Australia, the EU and its member states, and the African Union. However, it is significantly worth noticing that neither the United States nor the United Kingdom signed up.

AI and related technologies are changing the world and our daily lives, with undeniable benefits but also risks requiring control and monitoring that will be a major challenge to manage over the next 20 years! ■

FRAMING INNOVATION WITHOUT HINDERING: THE EUROPEANN APPROACH

Alexandru Mateescu
Université Paris1 Panthéon-Sorbonne

The rapid development of digital technologies – high-performance computing, simulation, artificial intelligence, and soon quantum computing – is transforming our societies. These tools bring about spectacular advances in health, education, industry, the environment, and other areas, while also raising major risks: misinformation, the digital divide, technological dependence, cybersecurity, and more. Faced with these challenges, the question is not to slow down innovation, but to regulate with judgement. Europe has chosen a unique, demanding, and open path where regulation becomes a lever for building trust, protecting citizens, and supporting competitiveness. This tension between protection and dynamism is central for the contemporary debate.

REGULATING WITHOUT HINDERING: A COMPLEX CHALLENGE

Regulating without hindering innovation is a delicate and ongoing task. As technologies advance, their uses become more diverse and complex. The fight against misinformation must not restrict freedom of expression. Making algorithms more transparent must not discourage research or compromise the confidentiality of trade secrets. Imposing strict cybersecurity requirements must not paralyze small, innovative organizations. Each decision involves a subtle balancing act between legitimate objectives that are often in conflict.

These tensions translate into regulation through a series of difficult choices. Rules must be clear enough to ensure the protection of citizens, but flexible enough to enable creativity and competitiveness. It is not only the content of texts that matters, but the entire process leading to enforcement: their development, implementation, and adoption by stakeholders in the field.

Such a balance cannot be achieved solely within legislative bodies; it takes shape where innovation and regulation meet. At least three types of ecosystems play a complementary role here:

- Business networks in France and Europe that bring companies together working in simulation, high-performance computing, and artificial intelligence. Alike with Teratec in France, these networks play a key role in promoting experimentation, technology dissemination, and training, while creating synergies between suppliers, users, and laboratories.
- On the academic side, numerous European institutions bring researchers from various disciplines together – law, computer science, philosophy, economics, arts – to explore the impact of digital technology, particularly AI, on society and contribute to the development of expertise with critical understanding of its uses.
- Finally, at the European level, DG Connect and the EuroHPC initiative pool high-performance computing resources among Member States, and

drive research and development policy in conjunction with Member States. These actions aim to ensure that Europe retains strategic sovereignty over its digital infrastructure, while supporting innovation and industrial competitiveness.

By linking all these levels of contribution – business, academia, European governance, and others – Europe is transforming regulation into a living process, where political vision, research, and industrial innovation reinforce each other as well as guide legislative solutions.

EUROPEAN SOLUTIONS

The European Union has gradually built a coherent regulatory framework to strike a fair balance: protecting citizens; supporting innovation; preserving collective control over digital technology.

The first step introduced the GDPR (General Data Protection Regulation) coming into force in 2018. Not only did it strengthen trust by giving control to every citizen over their personal data, but it also encouraged the development of sensitive digital services, such as telemedicine. While protecting privacy, Europe has thus laid a common foundation which has enabled the emergence of an integrated and dynamic digital market.

In 2022, Europe took a new step forward with the DSA (Digital Services Act) and the DMA (Digital Markets Act). The DSA requires platforms to actively manage systemic risks, while regulating these actions through transparency and rules for remedy. The complementary DMA reduces strict dependence on a few digital giants while opening markets to new players, hence stimulating competition and creativity.

In 2023, the Data Governance Act and the Data Act addressed the crucial issue of data circulation. They not only secure data flows but also promote shared value creation. This means that SMEs can not only access data produced by their connected machines, which was previously locked down by a single supplier, but also use it to optimize their production and create new services. By protecting sensitive data while expanding access, Europe is strengthening the strategic autonomy and competitiveness of its economic players.

In 2024, the AI Act complemented this package with one unprecedented framework for artificial intelligence, introducing a logic of proportionality:

- Critical uses, such as medical software or recruitment tools, are subject to strict transparency and robustness requirements;
- While everyday uses, such as cultural content recommendations, benefit from a more flexible framework that promotes adoption and innovation.

By doing so, Europe regulates high-risk practices with AI while enabling creativity and the rapid dissemination of less sensitive solutions.

These four pillars – privacy, online security, open markets, and AI oversight – show that well-designed regulation can protect without paralyzing and guide without constraining. By balancing individual protection, economic dynamism and collective sovereignty, Europe is charting a unique path, where regulation becomes a lever for innovation as much as a guarantor of freedoms.

BUILDING A RESPONSIBLE AND SUSTAINABLE DIGITAL FUTURE

By linking its efforts around these four pillars, Europe is making progress to show that it is possible to innovate without losing collective control of digital technology. This approach transforms regulation into a living process, connecting institutions, businesses, and researchers in shared dynamics. It is not only a model for a responsible and sustainable digital future, but also an invitation to move forward, nurturing the dialogue between innovation, trust, and sovereignty facing challenges ahead. ■

THE WORLD OF EDUCATION, DRIVING FORCE AND SUPPORT TO NEW DIGITAL TECHNOLOGIES

Guillaume Gellé
University Professor

To discuss training issues related to new digital technologies, we met with Guillaume Gellé, University Professor, former President of the University of Reims Champagne Ardenne, former President of France Universités, former member of the Genci Board of Directors, and former President of CINES. He shed light on the impact of AI on both students and faculty, as well as on teaching methods and training programs.

Alongside my academic career, I have held positions that gave me a broad overview of digital technology in the world of education. Recently, I submitted to the Minister of Higher Education, Research, and Innovation a report focusing on operators engaged in digital technology in France, which led to the announcement in early September 2025 of the merger of CINES, ABES, and AMUE *, to strengthen the consistency and quality of digital services for higher education and research institutions and their users. It also aims to closer adapt France's strategy to the European context.

EDUCATION ADAPTS TO AI

The impact of digital technologies on young people's education and the growing role of AI naturally requires changes in teaching methods, but it also means making young people better aware that this is a constantly changing world, full of pitfalls, which they need to master. This means teaching them from an early age not to believe everything they see on the internet. Critical thinking therefore remains a key part of their intellectual development.

There is now widespread awareness of these issues, and the digital education revolution is underway. This requires large-scale training programs for teachers with specialized trainers, who may not necessarily come from our own institutions.

As with all other disciplines, this effort to provide training in digital technology and AI must begin at an early age and continue throughout schooling. But digital technology is evolving very quickly, and this is a challenge that the educational community has taken up.

On the other hand, as in all other scientific disciplines, the fundamentals are constant. These are the ones we teach to the youngest children. We will also need to prepare our students to face some changes that we have not grasped yet, but which will impact society in the future.

TRAIN SPECIALISTS

Alongside this general training program, "digital technology and AI for all," we will also need to develop advanced training programs such as, "digital technology and AI for specialists." We have everything needed to achieve this: one among the best schools in mathematics; leading engineering schools; top-level university researchers; public funding, noticeably through France 2030. These high-level training programs for AI specialists are in place, focusing on cutting-edge technologies and technological breakthroughs we are tackling. It is also a race against time, no longer technological, but international also, with China and the United States moving at full speed, as well as international companies seeking to attract our best talent.

This training strategy must be settled on a few centres of excellence, but also on all our universities and engineering schools to ensure that knowledge is disseminated to all students throughout the country. Therefore, we must go on with the work carried out by Genci and programs such as MesoNet, spreading them from centres of excellence towards universities across the country. We must also develop the European strategy of university alliances and the development of specialized AI clusters. This is our only chance to avoid falling behind China, going very quickly.

TRAIN ALL ENGINEERS ON A MASSIVE SCALE

AI is becoming one of the fundamental areas of knowledge in many occupations, including engineering. All our generalist engineers must therefore be properly trained in AI, so that they understand how it works and can carry out small-scale developments, at the very least. This is a

THE WORLD OF EDUCATION, DRIVING FORCE AND SUPPORT TO NEW DIGITAL TECHNOLOGIES

Guillaume Gellé
University Professor

huge challenge because, even though we already succeed in training nearly 40,000 engineers per year in France, we would need to train up to 80,000 by 2030 to achieve objectives set for the country's reindustrialization.

I am not too concerned about our ability to train people, as we have a growing pool of teacher-researchers who specialize in AI with whom we will be able to mobilize more. This is also true in national research organizations (CNRS, INRIA, etc.), where many researchers work showing an extremely strong sense of public service. If our country needs trainers, they will respond to the major national challenges. The stake will be to find the computing resources they will need to meet the powering challenge, because there is no AI without energy.

COOPERATION MUST EXTEND TO AFRICA

We have strong ties, often based on special human relationships, with several African universities, whose professors we have trained up to the highest European standards. These universities contribute as fertile ground for motivated and brilliant young students who are very interested in AI. With their strong mathematical background, they make excellent developers and engineers.

This is even more important as AI can address the continent's major economic, social, and environmental challenges on the one hand, and on the other hand, because in times of weakened diplomatic relations, partnerships between universities prove to be an essential form of soft power.

MOTIVATED STUDENTS

Although students rushed to use ChatGPT to complete their assignments and unsupervised assessments more quickly, when it was released, they soon realized that generative AI was going to have an impact on society as a whole and that it was their interest to understand how it worked to use it to its full potential. The approach was similar among faculty members: to use it in professional practice; to adapt teaching and student assessment methods; to integrate this into teaching. Both students and teachers are now taking this step. One thing is certain: teaching and learning will never be the same again with AI. Let's take advantage to customize education with practical tools that support students where they really need it.

DEMANDING INDUSTRIALISTS

With support from regional, national, and European authorities, universities are opening their computing centres to businesses, often SMEs that lack the investment capacity or dedicated teams to develop the digital solutions they need. While pooling of infrastructure at the regional level is a first step, apprenticeship training is another asset, enabling engineering students to work on R&D projects meeting needs of local businesses.

The situation is different for large industrial companies which are equipped with infrastructure and teams capable of using these new technologies for their own needs. On the other hand, I believe it is essential for the industry to also contribute to training our teachers in the professions they will offer our graduate students.

NEXT CHALLENGE: QUANTUM TECHNOLOGY

Our country is very well positioned in the field of quantum technology research which has led to the emergence of some of the most promising start-ups in the world. The ecosystem is taking shape in France and Europe under the impetus of Genci and Euro-HPC. But our universities are not holding back and develop high-level training programs using quantum machine simulators. For example, the Roméo computing centre at the University of Reims acquired a simulator in 2021 and regularly organizes conferences and training courses in this field for students and companies, while coordinating the quantum component of MesoNet. ■.

* CINES (Centre Informatique National de l'Enseignement Supérieur) – ABES (Agence Bibliographique de l'Enseignement Supérieur) – AMUE (Agence de Mutualisation des Universités et Établissements)

BETWEEN INNOVATION AND RESPONSIBILITY, THE PARIS IA ACTION SUMMIT MARKED A HISTORIC MILESTONE

Anne Bouverot
France's Special envoy to the AI Action Summit

"NOTHING IN LIFE IS TO BE FEARED, IT IS ONLY TO BE UNDERSTOOD."

These are the words of the great scientist Marie Curie, reminding us of that progress stems from the exchange of ideas, cultures, and worldviews. It was the spirit of the AI Action Summit held in Paris in February 2025 at the Grand Palais, six months after the 2024 Olympic and Paralympic Games and fifteen months after the first AI summit in Bletchley Park in the United Kingdom.

A summit to introduce AI to the world, with Paris and Europe at its centre, hosted by France and co-chaired by French President Macron and Indian Prime Minister Modi.

The main objective of the summit was to reconcile science, standards, and solutions. The Grand Palais witnessed concrete examples, innovation, and investment announcements for trustworthy AI that serves the public interest. We wanted to go beyond the "science fiction" perspective of AI to reveal current tangible applications and attempt to reconcile the digital transition with the ecological transition, while ensuring that AI helps augment jobs rather than replacing them.

40,000 PARTICIPANTS

More than 800 contributors from over 100 countries, representing governments, international organizations, academia, industry, and civil society, worked with us to prepare and shape this summit. Nearly 40,000 people participated in more than 250 accredited events in France and abroad between July 2024 and February 2025, ahead of the largest global meeting on AI ever organized in Paris. The summit brought together 1,500 participants from around the world with 148 speakers from 43 countries, and a particular focus on gender equality as evidenced by the significant 48% rate of women among the speakers.

In addition to President Macron and Prime Minister Modi, the Summit notably brought together US Vice President JD Vance, European Commission President von der Leyen, Chinese Vice Premier Zhang Guoqing, the «godmother of AI» Fei Fei Li, Nobel Prize winners Demis Hassabis and Daron Acemoglu, Turing Prize winners Yann Le Cun and Yoshua Bengio, business leaders Sam Altman (OpenAI), Arthur Mensch (Mistral), Gundbert Scherf (Helsing), Choi Soo-Yeon (Naver), Sundar Pichai (Microsoft), Martin Kon (Cohere), Dario Amodei (Anthropic), and many others.

A ROADMAP

More than 100 concrete actions and commitments have been announced to promote trustworthy AI that is available to all and serves the public interest. We have called them the "Paris Actions for Artificial Intelligence," a roadmap for AI development based on sharing science, solutions, and common standards.

Among the major initiatives, we launched:

• Current AI: an international foundation for AI in the public interest, with an initial investment of €400 million over five years, supported by ten countries including France, India, Germany, and Kenya. This foundation will focus on data, openness, and accountability in AI models.

• The Coalition for Sustainable AI, initiated by France, the United Nations Environment Programme (UNEP), and the International Telecommunication Union (ITU), to assess the energy and environmental costs of AI and contribute to achieving the United Nations' sustainable development goals.

• The Pact for Trustworthy AI in the World of Work, signed by nearly 60 multinational companies, aims to promote social dialogue, occupational safety and health, productivity, and inclusiveness.

• The French National Institute for AI Assessment and Security (INESIA), to analyse systemic risks in national security and evaluate the performance and security of AI models, particularly in the fight against disinformation.

• Commitments to maintaining human control in AI-based weapons systems and laying the groundwork for a "global dialogue" and a "scientific panel" to strengthen international governance.

More than 60 countries signed a leaders declaration on AI, forging a common vision for trustworthy, sustainable, and inclusive AI.

330 BILLION EUROS

Unprecedented investments have been announced to position France and Europe at the avant-garde in AI:

• €109 billion in private investment in AI infrastructure in France;

• €200 billion in public and private investment in Europe through the InvestAI initiative and the EU AI Champions initiative, which brought 120 major European companies and start-ups together;

• €20 billion in investment by the European Commission in AI infrastructure (AI GigaFactories).

This Summit marked a turning point for future international meetings on AI: one of shared openness, action, and progress to develop AI in the public interest.

Diplomacy of AI is also scientific and cultural, and the multilateral system is the keystone of such edifice, which I intend to continue to support in future AI events such as the AI Impact Summit in Delhi next February. ■

LEGAL NOTICE

«The Keys of the Digital world» is published by Teratec
Campus Teratec – 2, rue de la Piquetterie
91680 BRUYÈRES-LE-CHATEL – France

infos@teratec.fr - www.teratec.eu

ADVERTISING AND EXTERNAL RELATIONS
Jean-Pascal Jégu - jean-pascal.jegu@teratec.fr

PUBLICATION DIRECTOR
Didier Besnard, President, Teratec

GRAPHIC DESIGN / PRINTING
Maindor Digital - o.rouquet@maindor.com

Legal deposit: October 2025

THE KEYS OF THE DIGITAL WORLD

The rapid development of digital technologies is profoundly changing the way we work and design products and services. In just a few years, digital technology – simulation and high-performance computing, and now AI and quantum computing – has established as an essential component of the whole economic development. It has become one of the main differentiators of competitiveness and efficiency for businesses and, at the same time, a major issue of sovereignty.

Teratec's goal is to ensure its members master these technologies and facilitate their dissemination and deployment. Training, at both initial and continued education levels, will be an essential component. Companies that invest most in these fields will benefit from a decisive competitive advantage.

And those mastering such technologies will become the leaders of tomorrow.

Hervé Mouren
Managing Director, Teratec