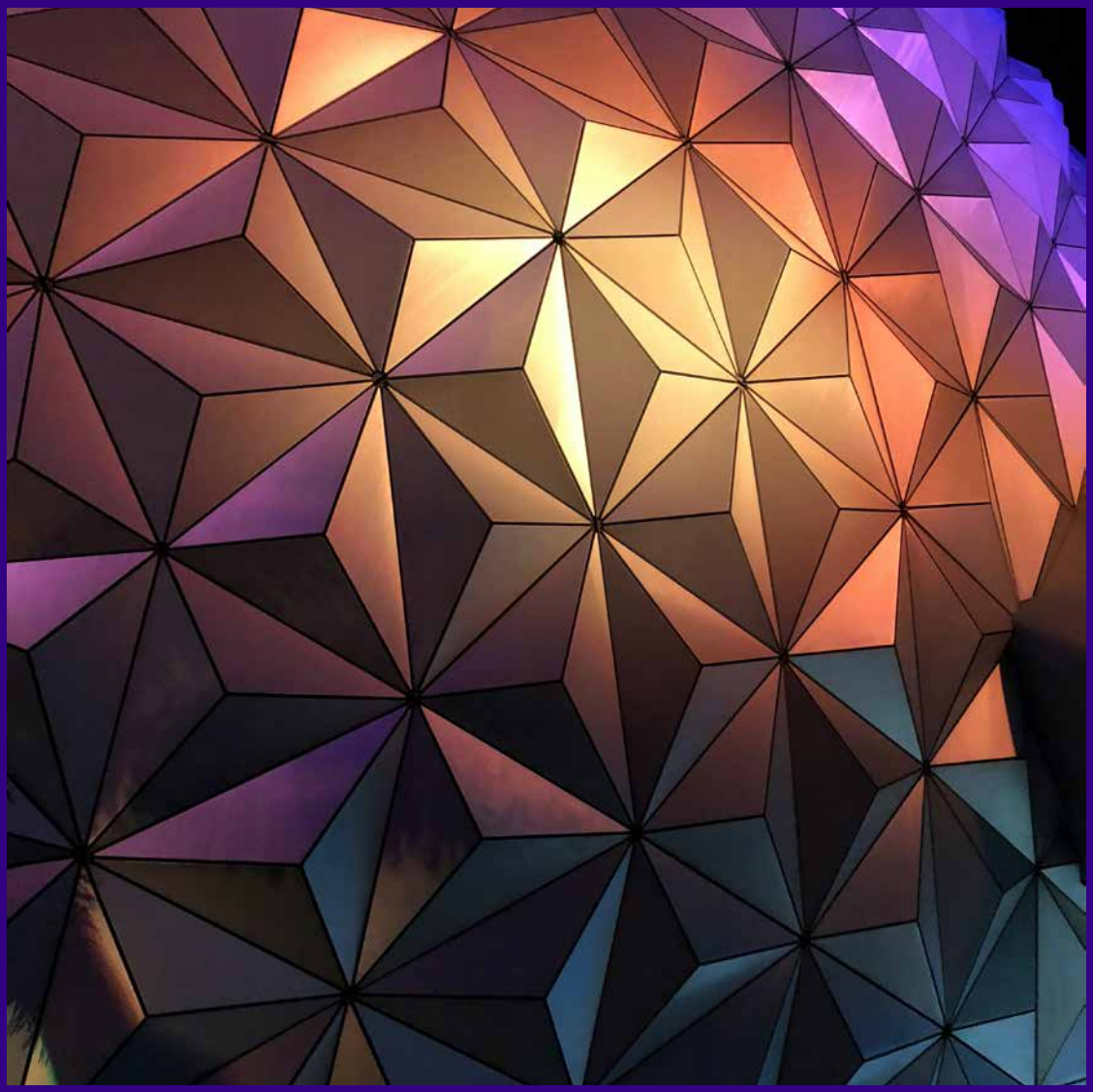
Inside magazine



February 2024 — Issue 06

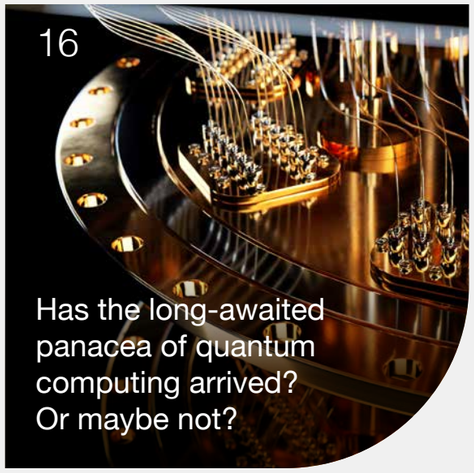
In this issue:

- **A hardware leash to keep AI under control**
- **Advancing Europe's semiconductor ecosystem**
- **Has the long-awaited panacea of quantum computing arrived? Or maybe not?**

Inside
Industry Association

In this issue

February 2024 — Issue 06



Dear reader,

With the Chips JU well and truly up and running, this magazine issue makes a brief retrospective of the launch event in Brussels whose message under the banner of ‘Chips For Europe’ was loud and clear to the representatives ranging from industry to public authorities: chips are vital to European key applications and strategic autonomy. EU Commissioner for the Internal Market, Thierry Breton, emphasised the need to consolidate and strengthen European competitiveness, remarking the necessity to attract private investments to Europe, increase international cooperation and massively invest in R&D&I. But the intervention of the CEO of prominent European industries revealed a mutual understanding of the ambitions and opportunities of the Chips Act, which can be realised only through inclusive cooperation along the ECS value chain and concretely bringing semiconductor technologies into real systems and applications.

For instance, few sectors are as committed as mobility in terms of innovation, collaboration between key players in the entire value chain and R&I ecosystem, and talent attraction. Innovation is understood as a change of attitude, a willingness to integrate knowledge in an accelerated way. Collaboration is considered an enabler for strategic initiatives like the software-defined vehicle (SDV). The talent is the ability to create, attract and develop the capabilities of people, of today’s and tomorrow’s professionals. The European Automotive Intelligence Centre (AIC), an INSIDE valuable member, comprises 32 organisations from nine countries and champions the need to respond ambitiously and responsibly to these challenges, especially with regard to ongoing technology and market transformations.

During the Chips Launch Event the ECS-SRIA 2024 was also unveiled. It will provide the basis for the calls of the Chips JU and support the European Chips Act with its long-term research directions. This year the ECS-SRIA increased attention on quantum technologies, seen as the next frontier in computing technology, which are attracting a lot of attention and investment from governments, industries and venture capitalists worldwide. However, despite substantial progress and investments over the past decade, ‘practical’ quantum computers capable of solving real-world problems remain elusive due to the inherent challenges in building reliable hardware. The question this article poses is whether the long-awaited panacea of quantum computing is around the corner or not.

On the AI theme, which seems to break all the hypes record week after week, we consider the potential risks associated with AI uncontrolled development and deployment. Many are now considering the integration of constraints into vital AI HW components to tackle these risks limiting the power of AI algorithms. Leveraging the symbiosis between HW and SW to contain the potential threats in AI-based systems is a new approach that includes the idea of embedding regulations that govern and control the training and deployment of advanced algorithms directly into the chips on which they run. It could offer a robust solution to prevent AI misuse by rogue nations, irresponsible companies, criminal organisations, hackers, individuals, and even autonomous systems.

AI is really top-of-mind, from developers to consumers, from technology to ethics, and along the entire ECS value chain. Consider, for example, the semiconductor industry which has also enthusiastically embraced AI to enhance productivity and address complex manufacturing challenges. The adoption of AI-based solutions in this domain poses unique challenges due to the industry’s intricate nature and specialised requirements, and here we explore critical considerations for the AI successful adoption and also present a practical case study on anomaly detection to illustrate its real-world implementation.

We also look at the possibility to extend our brain with AI, a domain characterised by a race to develop brain implants, a solution more invasive than what we presented in the magazine Issue 4. Neuralink, Elon Musk’s brain-implant company, implanted its brain-computer interface (BCI) device in the first volunteer, with promising results. While the milestone may not signify the merging of humans with AI, it marks a significant advancement for BCI, a vibrant market with several competitors offering very heterogeneous solutions.

As always we introduce to our community new members, starting from Sorbonne University and the Centre National de la Recherche Scientifique (CNRS), represented by Professor Andrea Pinna, a new member of the INSIDE Scientific Council. Andrea describes the activities, objectives and the areas of interest of his research team, along with the parallels with INSIDE and the underlying motivations for Sorbonne University and CNRS to become a member.

An interview with Sinetiq CEO Karl-Johan Gramner reveals “the link between the theoretical and the real.” The CEO compares their role to a physical architect who draws up the plan for a house and makes sure that everything is in order at the construction site but leaves the actual fabrication of the walls to another party. This approach has given them deep practical experience with third-party products in component-based systems and a knack for collaborative innovation.

And finally, in the context of the EUCEI initiative, INSIDE presents the 2024 RIAs Challenge, recently launched to highlight Research and Innovation Actions (RIAs) which has achieved relevant results and exploitation opportunities within the “Edge to Cloud Continuum”. In conjunction with the ECS Brokerage Event 2024, this initiative allows the selected projects to build a bridge with the industry, encourage the exploitation of their significant results and propose follow-ups.

Once again, I hope that the diverse and fascinating series of articles in this magazine will inspire you and perhaps prompt you to pitch your own story and/or views to us. Our INSIDE community is dynamic and growing quickly, and we encourage and value active participation.

Paolo Azzoni
Secretary General

Technology Frontiers

Technology Frontiers

A hardware leash to keep AI under control

Paolo Azzoni

Amidst the rapid evolution and diffusion of artificial intelligence, there is a growing concern about the potential risks associated with its uncontrolled development and deployment, and many are starting to consider strategies to mitigate these risks, integrating constraints into vital AI hardware components like GPUs to limit the power of AI algorithms. This approach goes hand in hand with other measures already established by some governments (like the US), such as export control, which avoids access to dual-use AI-based technologies but also significantly harms the business of companies in the AI domain (Nvidia lost billions due to export restrictions).

Indeed, regardless of their sophistication and complexity, AI algorithms are ultimately bound by the functionalities and capabilities of the hardware they run on. Researchers are exploring options to leverage the symbiosis between hardware and software to contain the potential threats associated with AI-based systems. This approach includes the idea of embedding regulations that govern and control the training and deployment of advanced algorithms directly into the chips on which they run.

This new strategy is currently a hot topic in theoretical debates about dangerously powerful AI and could offer a robust solution to prevent AI misuse by rogue nations, irresponsible companies, criminal organisations, hackers, individuals, and even autonomous systems. Unlike conventional laws or treaties (see e.g. the AI Act), this strategy might prove more complex and challenging to evade, relying on limitations introduced by design at the hardware level. As a confirmation, a recent report by the

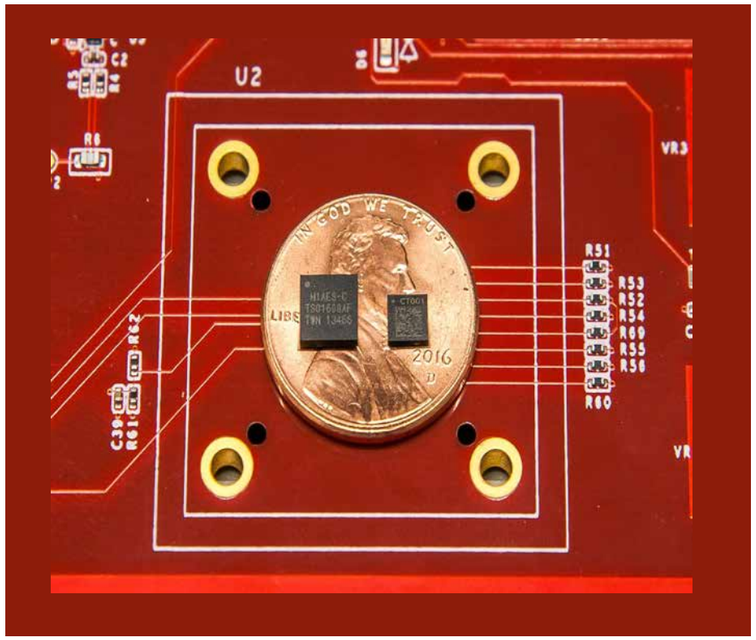


Figure 1 - Google's Titan server chip (left) and first-generation Titan M security chip (right)

Investigating microelectronic controls embedded in AI chips represents a potential EU security priority

Center for New American Security[1] highlights how carefully restricted silicon could strengthen various AI controls.

Certain chips already feature trusted components to protect sensitive data or prevent misuse. For instance, iPhones store biometric data in a trusted area called Secure Enclave, a dedicated hardware component separate from the main processor and provided with its own isolated memory and processing resources. The biometric data is encrypted and stored in the Secure Enclave, preventing unauthorised access or tampering and restricting access only to authorised processes. Similarly, Google adopts the Titan M chips[2] family, a RISC-V CPU with its own memory and cryptographic accelerator which is used in Pixel smartphones as a secure element that stores sensitive information, performs cryptographic operations, and verifies the integrity of the device's software during the OS boot. Moreover, to protect cloud infrastructure and services, Google adopts custom security-oriented chips in its data centres, including cryptographic accelerators, secure storage modules, and hardware-based encryption engines.

This approach could be adopted to leverage similar features in GPUs or incorporate new functionalities into future chips to limit AI access to computing power without proper authorisation. And the authorisation could be bound to licenses issued by government or international regulators and periodically renewed, enabling the restriction of AI training access by withholding new licenses. And this is not just theoretical: all chips from Nvidia that are adopted to train AI and are crucial for the creation of the most powerful AI models already contain a secure cryptographic module.[3] In November 2023, a research team at the Future of Life Institute,[4] a nonprofit dedicated to protecting humanity from existential threats, adopted, in cooperation with security startup Mithril Security,[5] the security module of an Intel CPU to develop a demo in which a cryptographic scheme can prevent unauthorised use of an AI model, set a computing threshold to limit processing load and remotely disable the use of the model. Complementing this approach, CNAS proposes that government or international regulators define protocols to deploy models only after they meet specific safety evaluation criteria.

While some view hard-coding restrictions into computer hardware as extreme, there is precedence in establishing infrastructure to monitor or control significant technology and enforce international treaties. An excellent example is represented by the adoption of this approach in modern seismometers: seismometers play a crucial role in monitoring underground nuclear tests, thus supporting treaties on underground weapon testing.

Although the approach to including hardware limitations in AI chips is not purely theoretical, the implementation of this approach faces political obstacles and technical challenges but also represents a very promising and profitable research area for the stakeholders of the ECS community. Developing hardware controls and limitations for AI would require new hardware features in future AI chips but also a new generation of cryptographic software schemes. Additionally, in a so critical domain, this approach requires a strengthened ECS value chain to ensure EU strategic autonomy and avoid being bypassed by other regions of the world with advanced chipmaking capabilities. This area of research and engineering is currently uncertain but will certainly require significant investments because, considering the current geopolitical situation, investigating microelectronic controls embedded in AI chips represents a potential EU security priority.

1 https://www.cnas.org/
2 https://safety.google/pixel/
3 https://www.nvidia.com/en-us/data-center/solutions/confidential-computing/
4 https://futureoflife.org/
5 https://www.mithrilsecurity.io/

Inside Magazine 6

Advancing Europe's semiconductor ecosystem

Highlights from the Chips Joint Undertaking launch event

Paolo Azzoni

Josh Grindrod

On November 30th and December 1st, a landmark event took place in Brussels: the launch of the Chips Joint Undertaking (JU)[1], which will drive investment into cross-border interdisciplinary research, development and infrastructure to bolster semiconductor technologies across Europe. This also serves as the main implementer of the Chips Act, aiming to double the EU's share of the global semiconductor market by 2030. As a committed member of the Chips JU, the three Industry Associations (AENEAS, EPOSS and INSIDE) were at the heart of the event organisation, which represents the start of an exciting new chapter for Europe.

A widely recognised need

Under the banner of 'Chips For Europe', the event indicated unprecedented interest in the topic of semiconductor manufacturing: over 800 individuals participated in the launch. Once an invisible force in our lives, chips have been thrust into the spotlight by the supply chain disruptions and global shortages that came hand in hand with the COVID-19 pandemic. For the general public, this crisis effectively served as an advertisement for chips, and the importance of semiconductors is now more widely recognised in terms of both their advanced applications and the need for strategic autonomy within Europe.

This message has been felt loud and clear, with the launch event attracting representatives from diverse backgrounds spanning large industry, SMEs, academia, research organisations and public authorities. The presence of EU Commissioner for the Internal Market, Thierry Breton, was another notable highlight as this marked the first ever attendance of a commissioner at such an event – again underscoring the importance of the semiconductor industry to Europe's future. In his introductory speech, Commissioner Breton emphasised the need to bridge the gap from lab to fab and outlined the three pillars of the Chips Act: the attraction of private investment to Europe, increased international cooperation and massive R&D&I investment. The latter will see €11 billion go into research up to 2030, making the Chips Act the largest investment of this kind in Europe.

Regarding the other two pillars, Commissioner Breton highlighted the creation of a European cloud-based design platform and four new pilot lines with a focus on extending Moore's law, scaling down FD-SOI technology, creating next-generation wide-bandgap materials and integrating heterogeneous technologies with advanced packaging. The platform will provide easy access to IP libraries, electronic design automation tools and support services, aiming to lower the barrier to entry for start-ups, SMEs and academics by enabling them to refine their designs at a lower cost. Access to both the platform and the pilot lines will be facilitated by a network of at least one Competence Centre per participating state, spreading the benefits across the continent.

Open to opportunity

Through such efforts, the Chips JU has already made a strong statement of intent on the balance between fundamental research and practical applications, the latter of which was exemplified by the extraordinary turnout of CEOs at the event. Nikolai Setzer, CEO of Continental AG, and Jaime Martorell, Special Commissioner for Microelectronics and Semiconductors in Spain, delivered keynote speeches on the topic "EU strategic autonomy and economic security". They were followed by a panel discussion on this topic featuring Thomas Skordas, Deputy Director-General of DG CNECT, European Commission, Pierre Barnabé, CEO of Soitec; Roger Dassen, CFO of ASML; Luc Van den hove, President and CEO of imec; Frédérique Le Grévès, President of STMicroelectronics France and EVP of Europe & France Public Affairs; Cinzia Silvestri, CEO of Bi/ond; and Joost van Kuijk, CEO/CMO of Adimec.

The following session was focused on "Maintaining and boosting European technology leadership", and was introduced by Jochen Hanebeck, CEO of Infineon Technologies AG, and Jo Brouns, Flemish Minister for Economy, Innovation, Work, Social Economy, and Agriculture, followed by a panel discussion included prominent figures like Signe Ratso, Deputy Director-General of DG RTD, European Commission; Sigrid Johannisse, European Semiconductor Board Member from the Netherlands; Stefan Finkbeiner, CEO of Bosch Sensortec GmbH; Maurice Geraets, Executive Director of NXP Semiconductors Netherlands B.V.; Sébastien Dauvé, CEO of CEA-Leti; and Eva Maydell, Member of the European Parliament.
The participation of these prominent figures in a variety of panel discussions revealed a mutual understanding that the ambitions of the Chips Act can only be realised through cooperation at all levels, from up-and-comers to industry giants. To this end, a Chips Fund will facilitate access to debt financing and equity, particularly for start-ups, scale-ups and SMEs. A positive start has already been made, with more than €100 billion of public and private investment so far generated by the Chips Act.

Talent was another hot topic at the event: how can we attract and retain the necessary workforce to carve out an approximately 20% share of the global semiconductor market by 2030? With sustainability a focal element of the Chips JU and Chips Act, several CEOs noted a strong alignment with the values of the younger generation and the need to promote careers in technology as the defining means to achieve green principles. In addition, plenty of panel members reminded us that – despite the name – the European Chips Act is not all about Europe. Strategic autonomy does not mean isolation; a degree of interdependence with other parts of the world is inevitable and we must remain open to opportunities with like-minded partners.

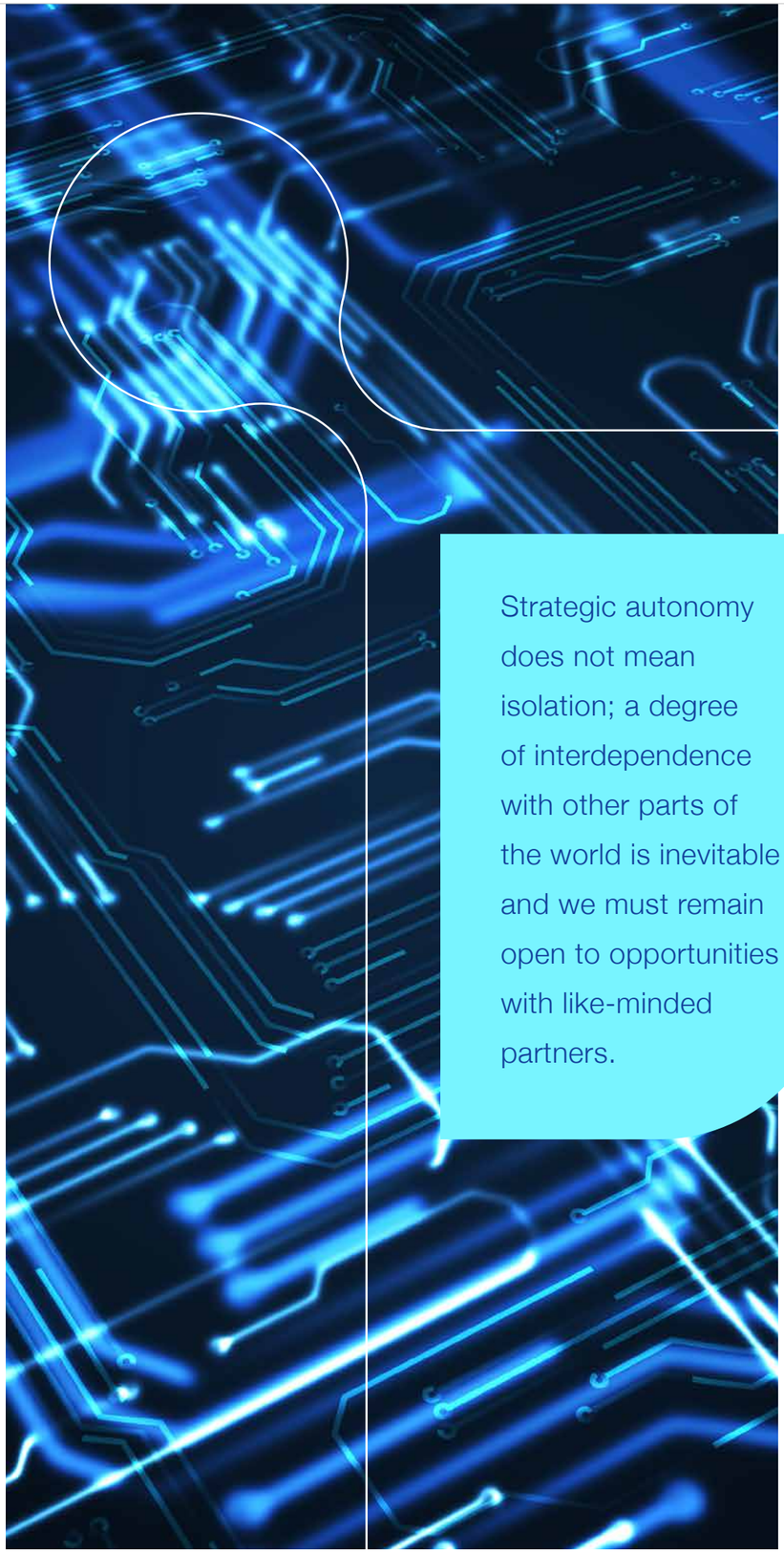
Propelling Europe forward

For INSIDE, the launch event was a source of pride as the 2024 version of the Electronic Components and Systems Strategic Research and Innovation Agenda (ECS-SRIA) was also unveiled. Created by the three Industry Associations, coordinating a community of over 300 experts, this will provide the basis for the calls of the Chips JU and support the European Chips Act with its long-term research directions.

Crucially, Chips For Europe also gave ample room for informal meetings and discussions, not just at the dinner and social event in the Art & History Museum of Brussels, but in an exhibition space showcasing the results of 30 significant projects in the area of the Electronic Components and Systems. Simply put, the collaboration required to make the Chips Act a success is impossible without a diverse and interdisciplinary community. As the event so comprehensively demonstrated, this spans the entire Electronic Components and Systems value chain, from research and development to manufacturing and applications. With the impressive turnout and clear dedication of both the INSIDE community and the wider world of intelligent digital systems, we encourage all of our members to immerse themselves in this ecosystem to propel European Electronic Components and Systems research and industry forward.

Jean-Luc di Paolo-Galloni
Valeo and Chips JU PMB

Beyond the celebration of the event and the networking part, as chair of the Private Members Board of the Chips JU and as President of Inside Industry Association, I cared to insist during the launch days on the importance of the extension of the undertaking's perimeter. If I compare it with the past ECSEL, it calls for more inclusivity and interdependence throughout the value chain of the electronics ecosystem, more shared challenges, interoperability and efficiency of the funding schemes to sustain all types of players, a reduction in any unneeded complexity, and an uplifting of the strategic level of the impacts, from the lab to fab and from the hardware to the software. Let's think and act European with our dedicated allies, all together to scale up and speed up the transfers to the industries for a better digital world.



Strategic autonomy does not mean isolation; a degree of interdependence with other parts of the world is inevitable and we must remain open to opportunities with like-minded partners.

Stefan Finkbeiner
CEO Bosch Sensortec and Chairman of EPOSS

The Chips JU is a great opportunity from my perspective as CEO of Bosch Sensortec as well as Chairman of EPoSS. For us the Chips JU represents more than only advanced silicon chips. It will cover hardware technology diversity, advanced nodes, and their applications. Yes, we need advanced silicon made in Europe as well as advanced integration technologies, and hardware and software co-design, but if we do not bring these technologies into systems and applications, we can not enable our automotive, medical or automation industry, all industries which are strong in Europe. We will not be successful in the end without embedded software and embedded AI going up to real AI. If you look at future applications like video processing for autonomous driving for example, it will be the software that makes the systems successful at the end. Hardware has to be standardised to be scalable and software, which is able to run on different hardware platforms, will be needed. All these topics need to be addressed simultaneously by the Chips JU.

But how do you really enable this? We need to be successful by even more efficient developments. In the pre-development and the pre-competitive phase especially, we have to leverage the synergies along the value chain by cooperation. A single company cannot cover all necessary developments along the entire value chain. Silicon chips are very expensive, as are certain software platforms (e.g. a car OS and SW frameworks), for which we need standards that all of the ecosystem's partners can benefit from, optimising efficiency.

To reach our goals, we need a wider community and must attract new experts to fill the lack of skills Europe is currently experiencing. If we want to grow and drive hardware development, software development and the whole ecosystem, we need to use a different language and include arguments that people, specifically young people, understand and that are associated to e.g. green ECS. We need to address not only 'latest transistor technologies' but topics like climate change, renewable energy, air quality sustainability, etcetera, where our contribution from the ECS community is huge.

It's therefore also very important that all activities that we drive in the EU have a long-term scope and are sustainable and address our values. Otherwise, we will not earn the benefits out of what is being seeded with the set-up of the Chips JU right now. Speed is crucial! To avoid being late, developments cannot be sequential! We already have to think about how we can leverage the new Chips JU pilot lines for pre-development of systems, in view of quick qualification and industrialisation. We need to define how we can go from ideas to applications as fast as possible. If you consider, for example, how fast the Chinese automotive industry is moving forward, we need to gain speed in Europe. In the Chips JU we must strengthen together the key technologies in Europe and develop new applications and markets.

[1] https://www.chipsjulaunchevent.eu/

Follow-up

Follow-up

The race to connect the brain to the digital world

Paolo Azzoni

In magazine issue 4 we described a technology to detect the brain waves which represents a non-invasive solution to create a direct connection between the brain and the digital world, a domain that is experiencing a race currently characterised by more invasive solutions based on the implantation of very small sensors directly in the brain.

In September, Neuralink[1], Elon Musk's brain-implant company, announced the start of recruiting volunteers for a clinical trial to test its brain-computer interface (BCI) device. The BCI collects electrical signals from neurons and translates them into commands to control external devices, initially aiming to assist paralyzed individuals in controlling a cursor or keyboard with their thoughts.

Neuralink recently raised $43 million in venture capital, despite having faced several controversies over its treatment of research animals and regulatory violations, which represent a problematic obstacle for scientific research in this domain. Government funding, particularly from agencies like DARPA and the NIH, has also played a significant role in supporting the research on BCI technologies.

In the race to connect the brain to the digital world, other companies are also making strides such as Synchron[2] which demonstrated the long-term safety of its implant in patients, while startups like Precision Neuroscience and Motif Neurotech have introduced novel brain implant devices. These efforts represent a culmination of decades of academic research in this domain, now translating into concrete applications and commercial opportunities.

Innovations in BCI technology focus on creating wireless, less invasive systems with improved neural data capture capabilities. For example, Synchron has developed a stent-like brain implant that can transmit neural signals from inside a blood vessel in the brain without requiring open brain

Figure 1 - Neuralink device

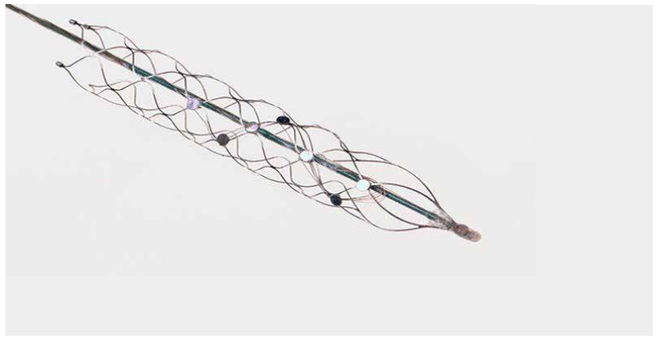


Figure 2 - Synchron's Stentrode® Minimally-Invasive Neural Interface

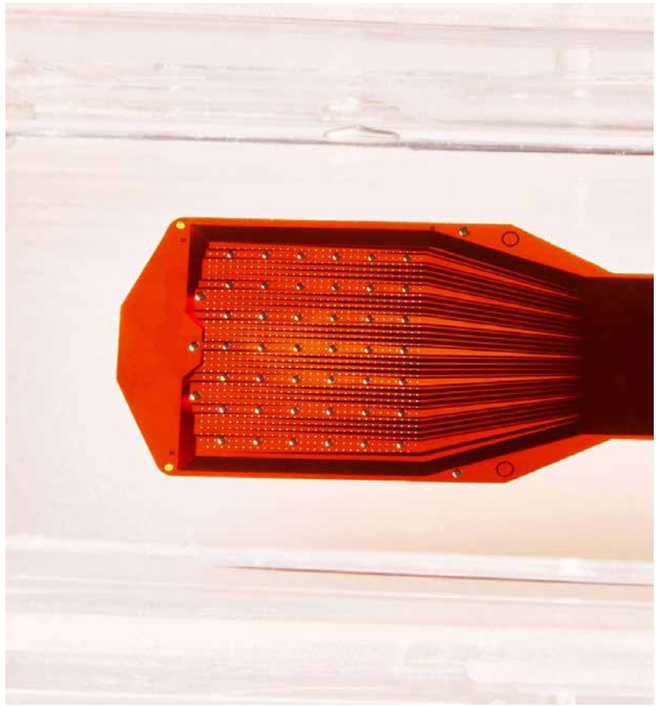


Figure 3 – Precision Neuro Layer 7 Cortical Interface

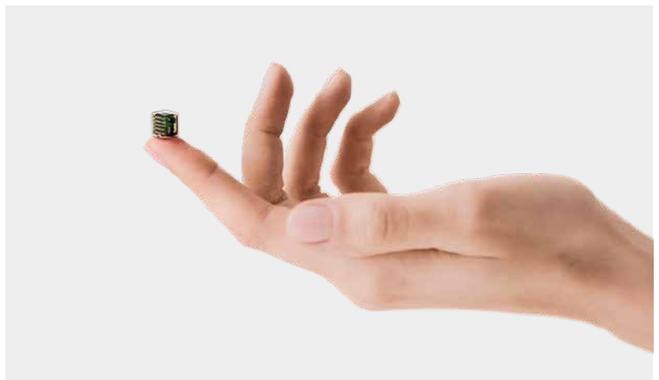


Figure 4 – Motif Neurotech device.

surgery, and have already implanted BCIs in patients, demonstrating safety and functionality.

Other companies like Precision Neuroscience[3] and Motif Neurotech[4] are developing thin film arrays and electrical stimulation devices to treat neurological disorders such as depression and dementia. Forest Neurotech aims to use ultrasound technology in neural implants for therapeutic stimulation. Neuralink's solution seems to increase wireless capabilities and features over 1,000 electrodes distributed across 64 threads, significantly surpassing the capabilities of previous BCIs.

These advancements represent a diverse range of approaches to brain-computer interface technology, from electrical stimulation to sound wave-based devices. While challenges remain, such as ensuring safety and effectiveness, the progress in this field holds promise for revolutionizing medical treatments and improving the lives of individuals with neurological conditions.

In this context, on the 29 of January, Neuralink announced has successfully implanted its BCI device, called Telepathy, into the first human patient. Neuralink's main goal is to achieve symbiosis with artificial intelligence, but for now, the focus is on exploiting the technology for medical purposes. The first application for example consists in enabling paralyzed individuals to control a cursor or keyboard using their brains. In this case the trial targets participants with quadriplegia due to spinal cord injury or ALS, aged at least 22, and expects to span six years.

The first implant seems to be successful and with "promising neuron spike detection", Elon Musk said: the device, a coin-sized device records and transmits brain signals wirelessly to an app for decoding, has been implanted into the brain's movement control region. But it is too early to assess whether the patient can effectively utilize the implant.

While the milestone may not signify the merging of humans with AI - Neuralink's main objective - it marks a significant advancement for their BCI device and keeps the race to connect the human brain to the digital world extremely vibrant.

1 https://neuralink.com/
2 https://synchron.com/
3 https://precisionneuro.io/
4 https://www.motifneuro.tech/

Technology Frontiers

Technology Frontiers

Has the long-awaited panacea of quantum computing arrived? Or maybe not?

Paolo Azzoni

The pursuit of quantum computing, seen as the next frontier in computing technology, is attracting a lot of attention and investment from governments, industries, and venture capitalists worldwide. Quantum computing aims to revolutionise computation by harnessing quantum properties such as superposition and entanglement to perform calculations at speeds unattainable by classical computers. A quantum computer is based on devices called qubits, which are not limited to manipulating 0s and 1s but can utilise a quantum mechanics phenomenon which allows a third state that is a superposition of both 1 and 0 at the same time. However, despite substantial progress and investments over the past decade, 'practical' quantum computers capable of solving real-world problems remain elusive due to the inherent challenges in building reliable hardware. And with such a low TRL, the hypothesis of a quantum pilot line is premature and represents a very risky investment.

In addition, the concrete possibility of generating economic benefits remains uncertain and, according to a recent study published in Nature by Chander Velu and Fathiro H. R. Putra,[1] the introduction of quantum computers could potentially slow down economic growth. Despite quantum computers offer exciting possibilities for various application domains, Prof. Velu clarified in a recent interview that "to overcome the specific hurdles in adopting quantum computers, the first crucial step is to demonstrate their practical value in tackling real-world industrial or societal challenges. This means showcasing their capabilities and effectiveness in solving complex problems that are currently difficult or infeasible for classical computers to handle."[3] He believes that, like the wide diffusion of digital computers in the '70s and '80s, the introduction of quantum computers will generate enormous economic gains, but it could be characterised by an initial period of stagnation for productivity growth, exactly as happened with the wide diffusion of digital computers. He estimates a potential economic loss in terms of global GDP of USD 13,000 per capita.

> The road to 'practical' quantum computing remains scattered with challenges and requires huge investments.

This paradox is generated by the necessity of investing in expensive equipment and by the learning curve required to master the new technologies, as well as by the necessity of changing processes and business models. He highlights three major challenges in the adoption of quantum computing: (i) high integration costs and low short-term return of investment; (ii) difficulty in translating quantum concepts for business managers and engineers; (iii) the cryptographic threat of quantum computers. In any case, on a

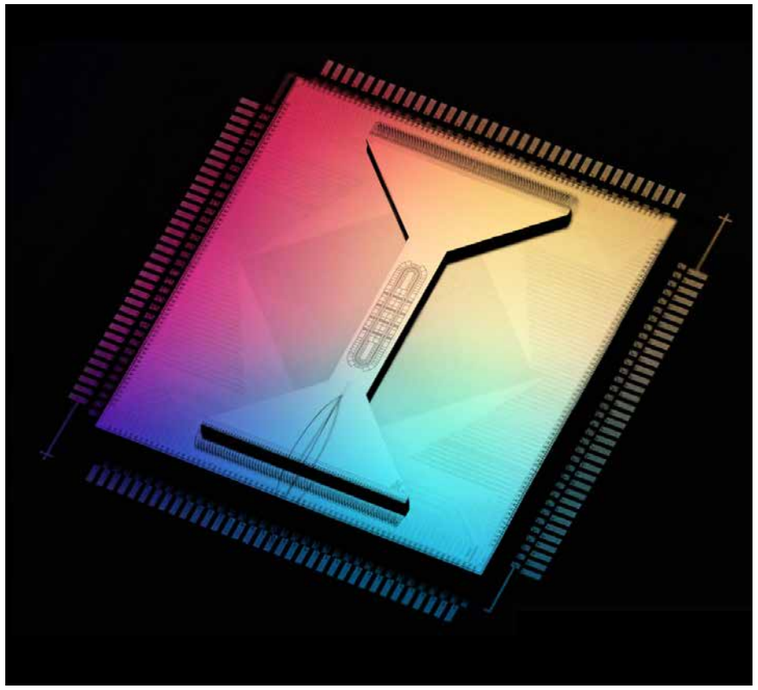


Figure 1 – Quantinuum System Model H2

timeframe of 10-15 years, it is difficult to say whether the technological immaturity or the lack of skills will be the real bottleneck. Prof. Velu's and Putra's article is not pessimism, but a wakeup call to avoid any potential trap in the future introduction of quantum computers.

Moreover, it is essential to notice, quite reasonably, that quantum computers will not replace standard computers. At a theoretical level, a 'universal' or 'general purpose' quantum computer could compute anything, replacing every category of today's computers, from laptops to supercomputers. But despite being theoretically possible, its concrete feasibility is extraordinarily difficult. Building quantum computers that solve commercially valuable problems requires an immense technological effort, as well as enormous investments. In our future daily life, even if quantum processors with very high fault tolerance will be available, it is difficult to imagine that we will use them for office applications, web browsing, multimedia, programming, etc., just as today we are not using high-performance computers (HPC) in computing centres for these applications. Most probably, quantum computers will be devoted for a long time to non-R&D production workloads, scientific research (which, for the resolution of certain problems,

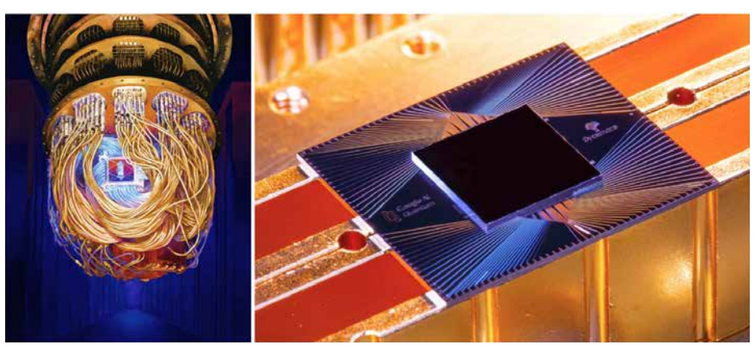


Figure 2 – Google's Sycamore quantum processor.

requires millions of qubit!), and special applications. Someone even argued that we should not have called quantum computers 'computers' because of their nature and specificities.

Leaving these speculations, recent breakthrough experiments conducted by researchers from Google and the startup Quantinuum bring us to the current state of the art, having reignited hopes for 'practical' quantum computing. Both teams independently reported advancements in developing a crucial component called a topological qubit,[3] which promises to address some of the reliability issues affecting current quantum hardware designs (e.g. reducing computational errors and enabling more complex algorithms). "This could well be a transistor moment for the quantum computing industry," said Quantinuum founder Ilyas Khan.[4]

These topological qubits are envisioned to enable more robust information storage and manipulation in quantum computers, potentially unlocking a wide range of applications from drug discovery to financial modelling and artificial intelligence…but there is some disagreement on the definition of a topological qubit, and the interpretation of these experimental results is the subject of an intense debate among researchers. While both Google and Quantinuum demonstrated key mechanisms necessary for topological qubits, their interpretations of the research results differ:

- Using a computer based on Sycamore processor,[5] Google's collaboration with Cornell University[6] showcased experiments indicating the creation of non-Abelian anyons, fundamental to topological qubits. However, Google researchers dispute the achievement of a topological qubit (they don't call it a 'topological qubit'), highlighting the need for error correction and practical utility.
- Quantinuum[7] collaboration with Harvard University and Caltech claims success in creating a topological qubit using its quantum computer, despite experiencing challenges similar to Google and remarking again that both experiments are based on materials and designs that are too fragile for a practical use.

This disagreement highlights the complexities and challenges inherent in the field of quantum computing, where theoretical concepts often clash with practical implementation. Despite the uncertainties surrounding the interpretation of these results, the pursuit of topological qubits represents a significant milestone in advancing our understanding of quantum mechanics and its potential applications. Moreover, these experiments represent an excellent example of collaborative efforts between academia and industry in pushing the boundaries of quantum computing research.

While the road to 'practical' quantum computing remains scattered with challenges, these breakthrough experiments offer valuable insights and could pave the way for further advancements in the field. As researchers continue to fight with the technical complexities and theoretical nuances of quantum mechanics, the quest for practical quantum computing is still trying to overcome what is called the NISQ (Noisy Intermediate Scale Quantum) era, an era where quantum computing is bound to and limited by error correction technologies which represent an obstacle for its wide diffusion and commercial exploitation.
In this phase, research and industry still depend on fundamental science, but Google and Quantinuum experiments represent promising milestones towards higher levels of technology readiness.

1 https://doi.org/10.1038/d41586-023-02317-x
2 http://www.eng.cam.ac.uk/news/how-introduce-quantum-computing-without-slowing-economic-growth
3 https://en.wikipedia.org/wiki/Topological_quantum_computer
4 https://www.quantinuum.com/news/for-the-first-time-ever-quantinuums-new-h2-quantum-computer-has-created-non-abelian-topological-quantum-matter-and-braided-its-anyons
5 https://quantumai.google/hardware
6 https://doi.org/10.1038/s41586-023-05954-4
7 https://doi.org/10.48550/ARXIV.2305.03766

Sorbonne/ CNRS and INSIDE sitting on the same page

Andrea Pinna
Professeur des Universités at Sorbonne Université

Chris Horgan

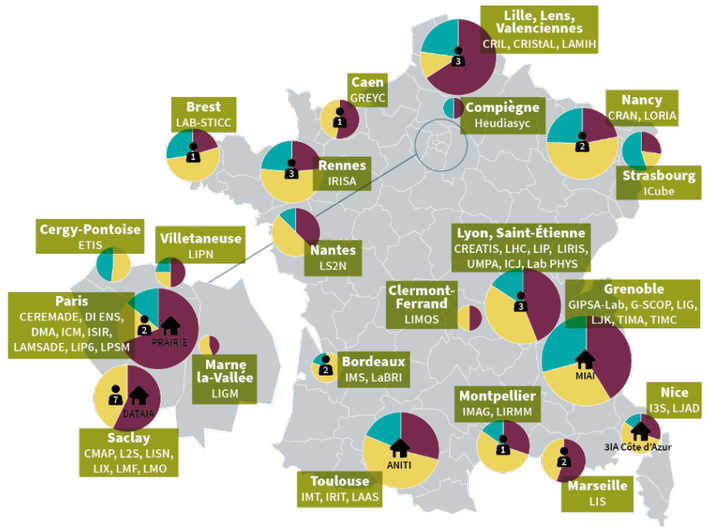
Inside Magazine 6

Sorbonne University (SU) and the Centre National de la Recherche Scientifique (CNRS) recently became a new INSIDE member and Andrea Pinna a new member of the INSIDE Scientific Council. Andrea Pinna has a PhD in electronics systems from University Pierre and Marie Curie, Paris. He is currently Associate Professor with the French Laboratory of Computer Science, Sorbonne University in the French capital. From 2006 to 2011, he worked in private industry in semiconductor and information technology. His research activities are based on e-health embedded system for smart medical device design and on embedded system co-design. "The best way to successfully valorise and handle a technological transfer from SU/CNRS to the industrials partners at European level is to become member of INSIDE Industry Association", he says.

Sorbonne University (SU) is structured by three faculties – Arts and Humanities, Health, and Science Engineering – and comprises 55000 students, with more than 7000 researchers in 135 research units. With a high-level scientific network and a global reputation, Sorbonne University has set itself the objective of developing its impact and attractiveness at the European and international level. At the heart of this strategy is active participation in European Union programs and initiatives as well as European networks of universities and research stakeholders, thereby strengthening the mobility of students and staff in the construction of the 4EU+ alliance. SU is involved also in the European Innovation Communities (KIC) Health, Digital and Climate.

Frontier research and groundbreaking innovations

CNRS is a public research organisation under the administrative supervision of the French Ministry of Higher Education, Research and Innovation and is committed to conducting



frontier research for the advancement of science in order to advance knowledge through cutting-edge research for the overall benefit of the society and support world-class basic research in all disciplines as well as promote interdisciplinarity, in particular with regard to major societal issues. CNRS works hand in hand with industrial and economic players on groundbreaking innovations and is a key driver in promoting the international visibility of French research, in particular within the framework of large-scale European programmes and infrastructures.

One of the largest publicly-funded research organisations in the EU, with a workforce of nearly 32000-strong, its 1137 units are spread throughout mainland France, French overseas and abroad. 97% of the units are jointly operated with other academic partners (universities, research operators) or industrial companies which share resources (HR, finance), premises as well as governance. CNRS develops productive relationships with industry and helps laboratories enhance their research and technology transfer to the business world (1000 research contracts each year, 140 common structures with companies comprising 1800 people, 6500 patents, 1200 active licenses, 80-100 annual startups, 8000 new jobs). There are ten scientific institutes coordinating and implementing the scientific policy: Biology, Chemistry, Ecology & the Environment, Humanities & Social Sciences, Engineering & Systems, Mathematics, Nuclear & Particle Physics, Physics, Information Sciences, Earth & the Universe. Noticeably, the CNRS was the first beneficiary of the Horizon 2020 programme, so far keeping this position in Horizon Europe as well.

Computer science

LIP6 is a Joint Research Institute (UMR 7606) of SU and CNRS. With more than 500 members, 200 of them permanent, LIP6 is one of the largest computer science institutes in France. The 18 research teams cover a wide area of Computer Science: from electronics to artificial intelligence. Collaborations at LIP6 are as much about fundamental (modelling and resolution of fundamental challenges) as applied research (implementation and validation in real conditions). The activities of the Institute unfold around four transversal axes:

- Artificial intelligence and data science
- Architecture, systems, and network
- Safety, security and reliability
- Theory and mathematics of computing

The industrial collaborations of the institute include startups, SMEs and big companies. “Our expertise is well established in various areas: IoT, communications, security, systems reliability, e-health, banking, energy, transports and space,” Andre explains. “Those collaborations have led to the creation of several start-ups and 84 software products and patents.”

Information science

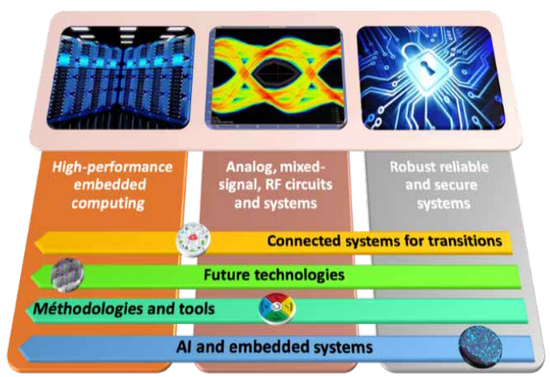
The CNRS Institute for information Science addresses societal challenges like Biology, Health, Wellbeing, Environment, Sustainable Development, Transport, Energy and Digital Humanities through five main topics Data science, knowledge management & AI, Software science, Cyber-security, Image processing and Robotics. AI is one of the main topics and the scientific priorities lie on the foundations of AI; trustworthy and responsible AI, privacy, reliability, fairness, bias; intelligent robotics and interaction with humans; new (green) computing paradigms; shaping new application fields; strengthening interdisciplinary collaborations. The interdisciplinary actions and collaborations on these topics are empowered by the National Research Working group GDR, one of which is very close to INSIDE on AI embedded architectures and high-performance computing.

The primary objective of the GDR SOC2 research network is to pioneer novel approaches for the design and validation of embedded systems, particularly for Internet of Things, edge computing and embedded artificial intelligence. We prioritise the study of hardware architectures, considering their interactions with both software elements covering applications, operating systems and reconfigurability, and the surrounding environment, encompassing analog, radio frequency and high-speed communication fields as well as sensor and actuator technologies.

“The overarching challenges we address revolve around minimising energy consumption to enhance the autonomy of embedded systems, managing the carbon footprint of exascale computing, ensuring robust security and integrity of electronic systems, and optimising the cost-effectiveness of the design and validation processes. Our aim is to tailor integrated systems to meet the diverse requirements of various application sectors such as smart cities, e-health, security and safety of people and assets, industry 4.0 etc.”

With a focus on complex integrated systems integrating hundreds of billions of elementary devices on a single silicon chip (SoC) and using new computing paradigms, new technologies or integrative approaches (such as 3D integration for example), these systems necessitate a broad spectrum of multidisciplinary skills spanning electronics, micro-nanoelectronics (digital, analog, RF), embedded computing, real-time operating systems and physics for emerging technologies.

The SOC2 research network adopts a holistic and cross-cutting approach to tackle these challenges, fostering collaboration among diverse communities and laboratories. Structured into seven threads - three thematic and four transversal - the group aims to address high-performance embedded computing, analog and RF circuits and systems, robust reliable and secure systems, connected systems for transitions, future technologies, methodologies and tools, and artificial intelligence in embedded systems. Bringing together 67 laboratories from across France, predominantly affiliated with CNRS and INRIA, the SOC2 research network benefits from the collective expertise of over 800 permanent staff, ensuring a robust and collaborative effort towards advancing system-on-a-chip and embedded systems research and innovation.



Removing scientific obstacles

Sorbonne Université and CNRS participate in the national Priority Research Programme and Equipment (PEPR) on AI. This programme aims to remove the scientific obstacles linked to the technological pillars of the national strategy, on embedded and trusted AI, by integrating doctoral training and going from fundamental research to proofs of concept. With a 73 million euro endowment, it will contribute to French excellence in research in this area and the rapid transfer of its results. The research supported by this PEPR focuses on the following main topics:

- Embedded AI (nanoelectronic components and architectures, software layers and component/software interfaces) and frugal AI (in data / computing power / energy efficiency)
- Decentralized AI (complementarities and alternatives between decentralised architectures and the cloud for AI) and trusted AI (robustness, generalisation, absence of bias, interpretability/ explainability, cyber risks, native data protection technologies)
- Mathematical foundations of AI to strengthen interfaces between mathematics and other disciplines

European impact vision

“On a more personal note, as a bit of a newcomer in the INSIDE Community my experience with the Scientific Council, for instance, has been excellent. I am very much aware of the interaction between the industrial part and the Scientific Council, the ways in which the latter can help industry not only translate the science but also empower industry, to provide the capabilities to valorise it. So, the scientific experience comes from both the academic and industrial parts. In this configuration, communication between the partners is key to prioritising what is important for industry and for the European Community so that the right direction can be given in terms of all the effort that needs to be made, with efficiency. I think that the INSIDE Steering Board also has this kind of responsibility from the perspective of technical documentation and strategy. I would like to make a stronger link between the French side and the Joint Undertaking. It’s important to collaborate, it empowers us, so my role also in the Scientific Council is to try to make a discreet and diplomatic interface to enable better synchronisation and, ultimately, collaborate rather than compete. A fruitful relationship. At European level, I want to exchange with European colleagues in order to have a really European impact vision and not a local, devoted focused vision.”

Projects

Adaptive architectures for embedded artificial INtelliGence – AdaptING

Lead / co-lead: *Alberto Bosio (ECL-INL, Lyon), Ivan Miro Pandes (CEA LIST)*

Partners: *CEA, CNRS, INRIA, École Centrale de Lyon, Sorbonne Université, Université de Rennes, Nantes Universités, Université Bretagne Sud*

Labs/institutes: *CEA LIST, INL, LIP6, IETR, IRISA, Lab-STICC*

The increasing need to distribute AI applications from the cloud to edge devices is becoming a pressing concern for addressing data privacy, bandwidth limitations, power consumption reduction and low latency requirements, especially for real-time, mission- and safety-critical applications. Consequently, there is an ongoing effort to design custom and embedded AI hardware architectures (AI-HW) that can support energy-intensive data movement, speed of computation, and large memory resources that AI requires to achieve their full potential. However, the current AI-HW architectures are mainly based on GPU, TPU, or specialised designs, which are devoted to improving the energy/performance efficiency for a specific class of AI applications, such as Convolutional Neural Networks. Thus, they are not designed to provide the high flexibility and massive parallelism needed to support a wide range of AI algorithms, including dynamic networks, Recurrent NNs, Transformers, etc. To address these limitations, the AdaptING project proposes a new architectural paradigm called adaptive architecture, which aims to make HW adaptable to any given AI application and its constraints in terms of accuracy, energy, latency, and reliability. The adaptive architecture is designed to provide flexibility, efficiency, sustainability, and reliability for embedded AI. This approach goes beyond the current state-of-the-art HW architectures and targets the next generation of AI by investigating and designing flexible, efficient, sustainable and reliable embedded AI on adaptive architectures.

Four scientific challenges are addressed in this project:

Learning-on-Chip: to address on-chip learning capabilities, we will focus on sustainable continual and incremental learning approaches where the embedded database size and the training energy is reduced and compatible with edge devices. Moreover, the type of operators and the data movement are different between inference and training. Thus, mutualising the same HW for inference and training is not trivial. The embedded learning approaches will be studied in order to identify the key operators allowing area/energy efficient switching between training mode to inference mode using the same hardware resources (e.g. floating point vs 4-bit operators).

Flexibility: we intend to follow a top-down approach: starting from class of AI applications down to its HW implementation. We leverage on the outcomes from "Algorithmic foundations of frugality", to include *novel* classes of *operations*. On the other hand, novel computational kernels, for example memories able to execute operations thanks to the use of *emerging technologies*, will also be integrated in the architecture with special focus on its electrical/accessibility constraints. Finally, the *interconnection* of hardware components has to be effective and guarantee high HW resources utilisation to enable time-domain multiplexing (i.e., scheduling different kernels on the same hardware component) to target large and dynamic AI algorithms.

Energy efficiency: this objective completes the previous ones through a bottom-up approach. First of all, one of the keys to achieve efficiency is to optimise the *memory access* operations. For this reason, we must reinvent the memory hierarchy and deploy computation resources to each level of the memory to reduce as much as possible the need for *data movements*. Secondly, we will support not only different *data precisions* but also different *data types* (at fine grain (i.e., HW elements can work with different data types and precisions). Here the goal is not to identify the best data representation for a given model, but to provide an architecture capable of supporting the desired type/precision pairs across all hardware resources. Power and frequency domains will be identified on the architecture at different granularity levels to improve the overall energy efficiency.

Reliability: the main goal is to make reliable AI by understanding how hardware errors (due to variability, ageing, external perturbations) can impact AI decisions and how to mitigate those impacts in an efficient way. The adopted methodology will start by the analysis of the possible failure mechanisms affecting the hardware. From the knowledge of failure mechanisms, we will derive the hardware errors (i.e., the logical representation of a failure mechanism) and we will analyse their impact on the AI results (impact analysis) in terms of accuracy degradation and how to determine if such degradation is critical (e.g., higher than a threshold). The complexity stems from the fact that the degradation metric strongly depends on the AI application (i.e., classifiers, object detector, segmentation). Thanks to the impact analysis, the last step of the methodology will be devoted to the design of low-cost *health monitors* to efficiently detect HW errors thus ensuring the correctness of the hardware (AI Hardware aware). Moreover, the knowledge of the occurred errors and their impact on the AI execution will be available to assist qualification methods, namely for explainability.

HOLlistic approaches to GReener model Architectures for Inference and Learning – HOLIGRAIL

Lead / co-lead: *Olivier Sentieys (IRISA, Rennes), Olivier Bichler (CEA LIAE)*

Partners: *CEA, CNRS, INRIA, Université de Rennes, Université Paris Saclay, Université Grenoble Alpes, Grenoble-INP, INSA Lyon*

Labs/institutes: *CEA LIST, CEA LIAE, IRISA, TIMA*

Accelerators of artificial intelligence algorithms currently consume much more power than they should, in particular in the learning phase. The many aspects of this question are too often considered in isolation. Based on the complementary expertise of the partners, and thanks to the integration into the rich community build by the PEPR on foundation of frugal AI, we will instead systematically look at a holistic, global comprehension of all these issues in established and upcoming AI algorithms. We will therefore combine more compact and efficient number representations, hardware-aware training algorithms that enhance structured sparsity, coding compactness and tensor transformations, with their adaptation to efficient hardware mechanisms and compiler optimisations. Our ambition is to provide breakthroughs in efficiency when running inference and training algorithms on specialised hardware. The results are intended to be integrated into development solutions for embedded systems, in particular within the DeepGreen national platform for the deployment of deep learning in embedded systems.

Six research challenges will be addressed in this project:

Extreme quantization: quantization is a key enabler to efficient implementations, both on existing off-the-shelf components (embedded GPU or MCU) where it enables low-precision data-level parallelism, and on specialized hardware architectures. We aim to explore very low quantization levels (below 4-bits) on complex network topologies, using integer as well as non-standard number representation formats such as logarithmic or custom floating-point. Quantization of cyclic computing graphs (e.g. recurrent networks like LSTM or GRU), which remains challenging, will also be addressed. This is complementary to work planned in DeepGreen, which focuses on the implementation and improvement of acyclic, integer quantization technics available in the state-of-the-art.

Structured sparsity (locality, regularity) for both weights (static) and activations (dynamic). Sparsity is intrinsically present in deep learning models, partially due to the use of linear rectifier activations and regularization in the loss function. However, it is naturally unstructured and thus difficult to exploit by data-parallel hardware. The main drawback of current, generally a posteriori, state-of-the-art compression or event representation remains the potentially high overhead associated with the indirection implied by these approaches. Our aim is to propose methods that create structured sparsity right from the initial design and training of models, enabling more direct and efficient parallel implementations, in terms of both latency and power consumption.

Maximum entropy coding: the activation values in deep neural networks (DNN) typically follow an exponential decay distribution: larger values are much scarcer than smaller ones, with a significant number of values being zero (high sparsity). The same is true when considering the activation rate in spiking neural networks. This means that the Shannon entropy in a neural network is much lower than the actual number of bits passed from one layer to another, even when the network is heavily uniformly quantized. This observation is still poorly formalized in the state-of-the-art and only indirectly leveraged in the latest developments in spike-based models using backpropagation though time and surrogate gradient technics. We aim to address fundamental questions regarding the weights and activations distributions in deep neural networks, and in particular for convolutional and attentional models.

Tensor methods for hardware-aware network architecture approximation. Tensor methods have been successfully considered as generic and efficient tools for DNN speed-up and compression. For example, most NN layers, i.e., the trainable weights and non-linear activation functions, can be decomposed into products of low-dimensional tensors, leading to an impressive compression rate. We aim to exploit recent algorithmic breakthroughs in the field of tensor methods that open the possibility to improve significantly the performance of compressing neural networks with less loss of accuracy than the current state-of-the-art. We will investigate how to consider in these methods the inherent sparsity in deep learning models and study the combination of tensor decompositions with other approaches such as quantization and distillation.

Energy-efficient training: we seek to use fewer bits during training, relying on number representations and bit-width that can be dynamically (i.e., at runtime) adapted during epochs and among layers. We will address this problem at the arithmetic and algorithmic levels and explore new mixed numerical precision arithmetic operators or kernel structures that are more efficient, both in terms of speed and energy. We seek to demonstrate that it is possible to train modern networks on 8 bits or less with an accuracy close to the reference. The key point here is to adapt the number representation with the distribution of the values during the training epochs.

Compiler and architectural support: going from a model that has been optimized with previous methods to a high-performance (in terms of latency and power) hardware implementation is a non-solved task. Indeed, current runtime and compiler infrastructures either rely on highly optimised operators from DNN specific libraries or pattern specific (for tensors) compiler optimisation strategies. The first approach (libraries) only supports a limited number of operators, failing to fully exploit compiler optimization such as fusion, specialisation, etc. The second approach (DSL - domain specific language - compilation) is competitive only on simple mainstream architectures (CPU, GPU). Revisiting pattern-specific optimisation strategies, becomes even more important in the current context (sparsity, extreme quantization), as the combinatorial optimisation search space is increased. Doing so involves the development of both analytical and statistical performance models. A particularly interesting and important challenge is the development of "high-level" performance models, that is, the ability to evaluate how friendly to optimisation is a partially/non optimised DNN. This is a required step for hardware-aware quantization or pruning, but also design-space exploration.

Follow us on Social Media!

Let's get connected for our latest news, events and updates!

/inside-association

/Inside_IA

Inside
Industry Association

Emergences: Near-physics emerging models for embedded AI

Lead / co-lead:	*Marina Reyboz (CEA LIST), Gilles Sassatelli (CNRS LIRMM)*
Partner:	*CEA, CNRS, Université Côte d'Azur, Université Aix-Marseille, Université de Bordeaux, Université de Lille, Université Paris-Saclay, Université Grenoble Alpes, INSA Rennes*
Labs/institutes:	*EA LIST, LIRMM, CEA LETI, LEAT, CRISTAL, IMS, Spintec, INL, C2N, UMPHY, IM2NP, IETR, LPNC, INT*

Contemporary machine learning (ML) has incurred profound changes in the scientific, societal and economic landscapes alike. After a decade of sustained progress AI as a discipline is still making regular breakthroughs on many fronts, at the expense of an ever-increasing amount of consumption of compute resources. Modern language models feature hundreds of billion parameters and training energy consumption alone probably falls in the GWh range, with a logical forecast worsening the already prohibitive carbon footprint of AI.

The *Emergences* project aims at advancing the state-of-the art on near-physics emerging models by collaboratively exploring various computation models leveraging physical devices properties. Effort will focus on three distinct fronts: Event-based models, Physics-inspired models (from physical systems dynamics) and innovative near-physics ML solutions (exploiting device properties). The investigations will be focused on embedded systems for Edge AI that call for increased energy efficiency for inference and learning, which could be incremental. They will apply to several application domains ranging for instance from the monitoring of the environment to health. Other important tasks such as common tools, performance metrics definition and model scalability analysis and will be conducted through as a collaborative transverse initiative.

It is commonly accepted that 1+ order-of-magnitude gains can be achieved through deeply rethinking the inner nature of the underlying AI computational models and architectures for leveraging model properties in target implementation technologies. Conversely, it is interesting to exploit specific technology properties for easing the implementation of specific functions. We outline three promising research directions in which state-of-the-art research shows promise, as follows: spiking neural networks / event-based models, physics-inspired models and near-physics design for machine-learning. In each of these three directions, we intend to leverage specific features of these technologies to enable significant gains over past approaches in terms of energy efficiency. Also, we will study embedded training which has so far been considered to be out of reach, thanks to close collaboration with neuroscience/cognitive sciences partners and recent findings in these fields. We will pursue these three directions with an ambitious interdisciplinary approach enabled by the broad spectrum of expertise gathered in the consortium, ranging from device physicists to computer science and architecture experts.

The *Emergences* project will thereby comprise a horizontal track aimed at consolidating datasets and benchmarking/analysis activities as well as finding scalability-enabling techniques through researching specific design patterns that will make it possible to meet performance, functionality and energy-efficiency levels in selected application areas proposed by consortium partners.

AI in semiconductor manufacturing

Case studies and guidelines for greater efficiency - AIMS 5.0

Gian Antonio Susto
Associate Professor at University of Padova

Inside Magazine 6

Over the past decade, companies spanning diverse industries have increasingly turned to artificial intelligence (AI) tools to optimise operational efficiency, innovate new functionalities and services, and advance automation. This trend has seen significant acceleration following the emergence of generative AI tools, which have not only captured the public's imagination but also propelled the adoption of AI solutions across various sectors. The semiconductor industry, renowned for its stringent quality standards and relentless pursuit of efficiency, has also enthusiastically embraced AI to enhance productivity and address complex manufacturing challenges. However, the adoption of AI-based solutions in semiconductor manufacturing poses unique challenges due to the industry's intricate nature and specialised requirements. In this article, we delve into the specific challenges associated with AI in semiconductor manufacturing, exploring critical considerations for successful adoption while also presenting a practical case study on anomaly detection to illustrate its real-world implementation.

AI in semiconductor manufacturing: challenges to address for effective adoption

The constantly expanding availability of data, coupled with the rapid advancement in computational power and the evolution of artificial intelligence (AI) algorithms, has ushered in a new era of transformation in the manufacturing industry, particularly in the semiconductor sector. This industry, known for its data-intensive nature and stringent quality standards, has witnessed a significant diffusion of AI-based technologies in recent years. With the global chip shortage of 2020 highlighting the critical importance of efficiency and productivity, the adoption of AI has become more imperative than ever.

Semiconductor manufacturing demands precision and reliability, making it a fertile ground for AI applications aimed at optimising operations, reducing waste, and enhancing throughput. Technologies such as predictive maintenance, fault/anomaly detection, and soft sensing/virtual metrology hold promise in revolutionising quality control and operational efficiency in semiconductor fabrication facilities. However, realising the full potential of AI in this domain comes with its own set of challenges[1].

Firstly, there are limitations on the availability and usability of data. While semiconductor manufacturing generates vast amounts of data, much of it may lack the necessary labelling or structure required for effective AI training. This scarcity of tagged or sufficiently labelled data poses a significant obstacle to developing robust AI models.

Secondly, the 'human-in-the-loop factor' introduces complexities related to trust, acceptance, and actionability of AI-driven insights. Despite the potential benefits, there may be scepticism or reluctance among human operators or decision-makers to fully embrace AI solutions. Bridging this trust gap and translating AI recommendations into actionable strategies aligned with business objectives is crucial for successful implementation.

Moreover, the lack of AI-ready architectures presents interoperability challenges. Integrating AI into existing manufacturing systems may prove challenging due to compatibility issues and the absence of

Semiconductor manufacturing demands precision and reliability, making it a fertile ground for AI applications aimed at optimising operations, reducing waste, and enhancing throughput.

dedicated tools for monitoring and refining AI solutions in the semiconductor manufacturing context.

Furthermore, the absence of off-the-shelf solutions compounds the complexity. Semiconductor manufacturing processes often require specialised modelling approaches that may not align with standard AI techniques. Consequently, developing tailored AI solutions capable of addressing the unique challenges of semiconductor fabrication becomes paramount.

Scalability also emerges as a critical concern. Effective AI solutions must be scalable to accommodate growing data volumes, changing equipment configurations, and evolving technological landscapes. Seamless integration with existing workflows and adaptability to new equipment recipes and technologies are essential for long-term viability and relevance.

Addressing these multifaceted challenges demands a collaborative effort across academia, industry, and technology providers.

A successful case study: explainable anomaly detection

In the framework of the AIMS5.0 (Artificial Intelligence in Manufacturing leading to Sustainability and Industry5.0) project[2], spearheaded by Infineon Technologies AG and supported through the HORIZON-KDT-JU-2022-1-IA-Topic-1 call, the University of Padova (UniPD) and Statwolf, a leading provider of end-to-end AI-powered data analytics solutions, are collaborating with 51 other partners across 12 countries. Together, they are developing AI-based solutions to tackle the aforementioned challenges and promote the adoption of AI technologies with a positive impact on sustainability. These solutions aim to reduce scrap wafer and resource usage while enhancing product quality.

In particular, one of the first solutions developed in this project by the team of the University of Padova and Statwolf is an innovative anomaly detection system that addressed several of the challenges listed above. Anomaly detection, also known as outlier detection, is an approach used to identify patterns or data points that deviate significantly from the majority of the data. Anomalies are data points, events, or observations that are rare, unexpected, or do not conform to the expected behaviour of a system: in the context of semiconductor manufacturing, anomalies may be associated with, for example, defective wafers or strange equipment behaviour. Multivariate anomaly detection can complement classic univariate control charts by capturing a broader spectrum of anomalies, including those that may be missed by univariate approaches. Furthermore, it is adept at handling complex data distributions, such as non-Gaussian or multimodal data, commonly encountered in semiconductor manufacturing environments. One distinguishing feature of multivariate anomaly detection is its ability to provide a unique quantitative indicator known as the 'anomaly score'. This score serves as a comprehensive summary of the equipment or process status, offering valuable insights into the severity and nature of detected anomalies. By leveraging this metric, operators and process engineers can swiftly prioritise and address anomalous events, thereby mitigating potential risks and optimising production efficiency.

The solution developed by the University of Padova and Statwolf transcend conventional anomaly detection methodologies, incorporating cutting-edge algorithms and a seamlessly integrated architecture. This innovative approach addresses several challenges outlined previously:

1. **Tagged Data Requirement**
   Unlike traditional approaches reliant on labelled data, the developed solution circumvents the need for abundant tagged data. This is particularly advantageous in semiconductor manufacturing, where faults are relatively rare in production processes.

2. **Multivariate Analysis**
   By accounting for the multivariate nature of semiconductor manufacturing equipment, equipped with numerous sensors to monitor and regulate underlying processes, the solution offers a more comprehensive understanding of system behaviour.

3. **Integration with Existing Systems**
   The solution boasts full integration with fault detection and classification systems, manufacturing execution systems (MES) data, and internal IT infrastructure. This seamless integration enhances operational efficiency and facilitates holistic data analysis.

4. **Root Cause Analysis**
   Leveraging innovative eXplainable AI (XAI) methodologies, the solution provides detailed root cause analysis information. This goes beyond mere anomaly detection by identifying feature importance, guiding process engineers in troubleshooting and decision-making[3,4].

5. **What-If Scenarios**
   In the realm of XAI, the solution introduces 'what-if' scenarios, empowering non-experts in AI to combine black box outcomes with subject matter expertise. This facilitates practical decision-making and enhances collaboration between stakeholders.

6. **Continuous Improvement**
   Through the incorporation of new data and human feedback, the solution evolves over time, adapting to changing manufacturing dynamics and improving its predictive capabilities.

7. **MLOps Monitoring**
   To ensure the reliability and effectiveness of AI outcomes, the solution incorporates robust machine learning operations (MLOps) systems. These systems facilitate continuous monitoring, maintenance, and optimisation of AI models, ensuring consistent performance in real-world settings.

By addressing these key aspects, UniPD and Statwolf collaboration offers a sophisticated anomaly detection solution tailored to the unique challenges of semiconductor manufacturing. This not only enhances operational efficiency and quality but also fosters innovation and agility within the industry.

Guidelines for effective AI adoption in semiconductor manufacturing

While the anomaly detection solution described above may appear complex for deployment, with the right combination of expertise and key factors, it becomes achievable for many companies. We strongly advocate for companies to:

- Harness the Active Ecosystem of AI: Although relatively few researchers are specifically focusing on the intersection of AI and semiconductor manufacturing, the AI landscape is continually evolving with novel approaches and open-source software. Engaging AI experts with vertical knowledge can be invaluable. For instance, UniPD has developed 'CeRULEo: Comprehensive Utilities for Remaining Useful Life Estimation Methods', an open-source library facilitating the rapid evaluation of cutting-edge approaches in predictive maintenance[5]. Additionally, it's advisable to collaborate with AI experts already immersed in the semiconductor domain to discern and filter reliable approaches suitable for semiconductor settings.

- Involve Users Throughout Development: User involvement at every stage of development is paramount. Adopting XAI approaches enhances trust in the developed solutions and promotes their adoption. We must advocate for a perspective where process engineers, both with and without AI expertise, collaborate synergistically. While improperly designed AI solutions prove futile, effective ones can significantly enhance company operations and empower operators and decision-makers. The involvement of subject matter experts is crucial for gaining a clear understanding of their needs. Ultimately, the path to effective AI adoption lies in empowering users rather than displacing them.

- Ensure Availability of Suitable Digital Architecture: Having a suitable digital architecture for data handling, integration, and monitoring of AI solutions is essential. Rather than reinventing the wheel, leveraging existing AI architectures is advisable. For example, at Statwolf, the software development is not built from scratch. Instead, the Statwolf team has crafted an end-to-end data and AI platform that enables the development of tailored solutions with the performance of off-the-shelf ones. This platform facilitates the deployment of a wide array of algorithmic approaches, ranging from simple statistics to complex neural networks, ensuring versatility and effectiveness.

The time for embracing AI adoption in semiconductor manufacturing is now, if properly done!

1 Susto, G. A., Diebold, A., Kyek, A., Lee, C. Y., & Patel, N. S. (2022). Guest Editorial Process-Level Machine Learning Applications in Semiconductor Manufacturing. IEEE Transactions on Semicon-ducto

2 AIMS5.0 Project Website https://aims50.eu/

3 Dandolo, D., Masiero, C., Carletti, M., Dalle Pezze, D., & Susto, G. A. (2023). AcME—Accel-erated model-agnostic explanations: Fast whiten-ing of the machine-learning black box. Expert Systems with Applications, 214, 119115.

4 Carletti, M., Terzi, M., & Susto, G. A. (2023). Interpretable anomaly detection with diffi: Depth-based feature importance of isolation forest. Engineering Applications of Artificial Intelligence, 119, 105730.

5 Lorenti, L., & Susto, G. A. (2023). CeRULEo: Comprehensive utilitiEs for Remaining Useful Life Estimation methOds. Journal of Open Source Software, 8(88), 5294

Biography

Gian Antonio Susto is an associate professor at the University of Padova, Italy. His research interests are focused on artificial intelligence and control applications in the context of semiconductor manufacturing, with a focus on technologies such as predictive maintenance, fault and anomaly detection, virtual metrology and dynamic sampling. He is an associate editor for the journal IEEE Transactions on Semiconductor Manufacturing in the area of process modelling. He has been involved in several EU-funded projects in collaboration with the major European semiconductor manufacturing companies. He is one of the co-founders of Statwolf, a company developing end-to-end data analytics solutions for semiconductor manufacturing and Industry 5.0.

SME Focus

SME Focus

Flexibility and full reuse

Sinetiq’s solutions for service-oriented architectures

Karl-Johan Gramner
CEO at Sinetiq

Chris Horgan

Despite spinning off from its parent company just under three years ago, Sinetiq is no newcomer to the INSIDE community: as part of BnearIT, this commercial SME was a key player in the original Arrowhead project of 2013 and remains actively involved in developments around the Arrowhead Framework. Alongside this, they apply more than 20 years of experience in system integration, service-oriented architectures and component-based systems to help national and international customers improve their profitability and competitiveness. As CEO says, “we are the link between the theoretical and the real.”

The start and the end

“We would like to be mentors to our customers on how to accomplish a service-oriented architecture. Once they’ve figured out how to build it, we can stand at the receiving end and help them out on the components or the different systems to be integrated,” begins Karl-Johan. “When you create a system of systems, I would say that in 100% of cases, at least one of the components will not fulfil the specification; either the API or the service itself will not have been created the way it was specified and something will crash. You then have to figure out, ‘okay, which of these 50 components is not behaving the way it should’. So, we tell you the way to structure your systems or components at the start and how to make the system of systems work in the end. Most IT consultancy companies help out with the big chunk of implementation in the middle, but we are not like most companies; we help you to integrate.”

To this end, Sinetiq offers a wide range of knowledge and services, including product packaging, guidelines and training, architecture and design principles, interface maintenance and compliance, and testing, verification and validation. Standard solutions based on a service-oriented architecture are included and refined within their self-developed products, which enable end-users to obtain new data and functionalities efficiently and continuously. For instance, their consultancy workshop has been packaged into an integration agility assessment that allows companies to discover whether they can easily swap components in a plug-and-play manner. Such offerings enable controlled flexibility, full reuse of investments and a reduction in time to market through quick access to the necessary data for decisions.

A bridge to practice

To illustrate this further, Karl-Johan compares their role to a physical architect who draws up the plan for a house and makes sure that everything is in order at the construction site but leaves the actual fabrication of the walls to another party. This approach has given them deep practical experience with third-party products in component-based systems and a knack for collaborative innovation – both of which made them an ideal partner to join the first Arrowhead project 11 years ago. Orchestrated by Luleå University of Technology, this sought to address challenges in cooperative automation and legacy systems. The resulting Arrowhead Framework abstracts elements of Internet of Things (IoT) to services that enable interoperability between almost all IoT elements. This ultimately facilitates scalable, secure and flexible information sharing for the design and implementation of automation systems in domains as diverse as production, mobility and energy.

“If you look at Arrowhead now, we were a big part of how the architecture is set up and structured,” Karl-Johan explains. “By the time the first Arrowhead project started, we had been doing real implementations for ten years. So, we were asked to participate as we had that practical experience, whereas many of the other participants had a lot of academic or theoretical knowledge. We could say, ‘that’s fine in theory, but you have to do like this in practice.’ That’s where we contribute.”

Learning and sharing

Following the release of the Arrowhead Framework in 2016, several other EU projects have furthered its development, including Productive4.0 and Arrowhead Tools. In 2020, Arrowhead joined the Eclipse Foundation due to its proven governance framework and processes for collaboration on open-source software, becoming Eclipse Arrowhead. Collectively, Arrowhead projects have brought together more than 100 participating organisations and, for Sinetiq, their continued involvement represents a chance to learn from some of the greatest minds in Europe.

“The first reason we are involved in Arrowhead is to stay on top of cutting-edge technology. This is our chance to work with universities, professors, PhDs and whoever else really wants to innovate in this field,” Karl-Johan continues. “In the Arrowhead projects, they have verticals and horizontals. The horizontals are usually the technical work packages, like microservices or translation work packages in which researchers and innovators explain and define how things are. Then you have the verticals, which are the use cases in which different companies try out the technology that the technical work packages define. At Sinetiq, we will contribute to the technical work package on microservices, for example, but then we will also support the use cases by making use of the technical advancements made. In these projects, we are a bridge between the deep tech and the practical.”

Of course, Karl-Johan is also grateful to the strong ecosystems in Europe that make such collaboration possible. “I understand that INSIDE is one of the facilitators, making sure that these EU projects come to play. It’s very, very important there are such organisations. Without them, there would be no projects. It shows that the industry has an interest in this kind of innovation and it’s hard otherwise for industry to tell the EU that this is important. That’s where INSIDE can be a strong voice, rather than one of us going there ourselves trying to say why they should start a project. I also recently learned of the opportunities at INSIDE, like the Brokerage, where you can listen to which new projects are about to be started and present your company. It’s another great way for European companies can stay competitive.”

Making it work

And while this collective mentality helps make the entire European ECS domain stronger, Karl-Johan is also keen to emphasise how it makes individual companies more streamlined by avoiding doubled work. A core tenant of Sinetiq is never recreating what can already be found; participation in ecosystems for collaboration and knowledge sharing allow them to live up to this principle. “Our products are components in a service-oriented architecture, which means that you should develop some products yourself if they are your unique selling points as a commercial company. We will never build anything that already exists in open source or is available at another provider. We don’t want to reinvent the wheel, so we always have the following steps in mind. First, look at open source. Second, look at other commercial providers. Third, build it yourself. Fourth…okay, we’ll build it for you. Or we might already have it, of course.”

“When I look into the future, we want to be both a provider of components or products and an expert consultant. That’s a difficult trip but it’s not impossible,” concludes Karl-Johan. “However, we only want to have products or components within a service-oriented architecture, and that’s the final reason for us being part of Eclipse Arrowhead[1]: if our components support Arrowhead or are compatible with it, it makes this easier for us and brings us a big benefit. So, we are really eager for Eclipse Arrowhead to make it as a well-used open-source community and I hope that this is also the focus for all parties in the Arrowhead projects. We already have a common ground to start from, so all of us can make this work together.”

[1] https://arrowhead.eu/eclipse-arrowhead-2/

Inside Magazine 6

Innovation as a mindset, talent as the future

Mikel Lorente

Few sectors are as committed to innovation and talent as the automotive sector, now called mobility. Innovation is understood as a change of attitude, a willingness to integrate knowledge in an accelerated way. Talent is the ability to create, attract and develop the capabilities of people, of today's and tomorrow's professionals.

This requires a certain climate that facilitates both objectives in an organised and natural way. There is nothing better than the example of the environment itself to set the guidelines for behavior in these important matters.

But this does not come out of nowhere. For an ecosystem to come about, there must be instruments that unite and strengthen the different elements scattered within a physical and virtual environment.

This is the case with AIC-Automotive Intelligence Center, a young project less than 15 years old, which has become the ideal place to co-develop world-class capabilities.

This is because AIC is a European centre of value generation for the automotive sector based on a concept of open innovation where companies strengthen their competitiveness through cooperation.

This distinctive concept is market-oriented, integrating activities of a different nature in different areas: people, developing the best professionals; processes, leading advanced technologies; and products, driving the requirements of future vehicles.

In this sense, it is equipped with modern facilities of more than 60 thousand square meters, where the resident companies can deploy their activities and take advantage of the center's equipment, as well as the knowledge that is generated among the different team members that make up the centre.

Within its activities, AIC meets all the strategic needs of the automotive sector: competitive intelligence, analyzing different future scenarios of the automotive sector; training, preparing automotive professionals in different qualifications; research, researching priority areas for the sector

Within its activities, AIC meets all the strategic needs of the automotive sector.

and for companies; industrial development, developing new industrial initiatives from their initial stages; and new business, supporting all those initiatives that add value to the sector by attracting projects in those areas that are fundamental to the future of the industry.

To date, 32 organisations from nine countries have joined the project. It also has a wide network of international collaborators with other agencies. In short, a place to innovate and develop talent.

To better understand its activities, we have the founder and CEO of the centre, Inés Anitua.

Inside Magazine 6

A rising number of partnerships show a growing openness to join forces.

Would you explain your approach to innovation?
AIC has developed its idea of innovation in a very particular way: an innovation that is not only technological, but much more comprehensive. To this end, AIC has developed capabilities in different areas that it understands to be key to accelerating innovation, such as competitive intelligence, applied research, training and industrial development, all integrated into a single pack/one-stop-shop that allows it to know precisely where it is heading and at what pace.

How is this innovation structured?
Our objective is first to understand what is happening in the different areas of knowledge in order to be able to analyse in real time where industry is heading and how it affects companies. We have a very practical view of things because there is a bottom line behind any organisation that needs to be taken care of. It's not innovation for innovation's sake. And controlling the timing is very important.

What are the most important knowledge streams?
Participating in organisations such as INSIDE, ERTRAC, 2ZERO and other international partnerships gives us a very accurate overview of what is happening. In addition, companies with very different characteristics also enrich our reality and support our strategic decisions on where to move forward. That is why, in recent years, we have been opening our scope from the more traditional aspects to electrification, electronics and software approaches...

What do you see as the main challenges?
Interestingly, talent is one of the lines that we look after the most because we know that it is people who make the difference. In this sense, we want new professionals to emerge with the qualities that we consider fundamental, such as ambition, love of risk, ability to adapt to change, multicultural mindset, etc.

And from a technological standpoint?
We are facing extremely complex times in which regulatory impulses, changes in consumer mentality, brutal competition from other non-European countries and the geostrategic chaos in which we find ourselves make the challenges all the more formidable.

In a context of vehicles becoming increasingly electric, automated, connected and service-oriented, ongoing technology and market transformations, especially towards the software-defined vehicle (SDV), are leveraging the need for collaboration between key actors representing the whole value chain and R&I ecosystem (OEMs, suppliers, tech companies, chip manufacturers, cities and mobility service providers, universities and research and technology organisations, etc.).

The rapid increase in software complexity requires more standardisation across players, a need for an end-to-end software platform (including an operating system and middleware layer), building on the SW layer abstracting the underlaying HW, fostering the smartisation of products at a system and subsystem level, etc. A rising number of partnerships show a growing openness to join forces. Strengthening European collaboration becomes crucial to accelerating this recent trend.

To this challenge, we can add further topics such as solid batteries and hydrogen, without overlooking advanced manufacturing, etc.

Finally, how do you see European industry in the coming years?
Europe has to be very smart about its commitment as a sector. It has to be ambitious from the point of view of sustainability but, at the same time, it has to think from the point of view of global competition. I think we have had a few years of some disarray, of a lack of common sense in making some decisions, of having let ourselves be carried away by voluntarism... We have to go back to our basic principles to be able to respond with ambition and responsibility to today's challenges.

Inés Anitua
CEO AIC-Automotive Intelligence Center

RIAs Challenge Unveils Winners

Inside Magazine 6

Paolo Azzoni

Tiziana Fazio

In the context of the European initiative EUCEI[1], INSIDE Industry Association recently launched the 2024 RIAs Challenge[2]. INSIDE Industry Association is a partner of EUCEI, an initiative composed of two CSAs, Open Continuum and UNLOCK-CEI, whose intention is to build a European continuum among the industries involved in edge, IoT and cloud technologies and applications. The 2024 RIAs Challenge has been organised to highlight Research and Innovation Action projects (RIAs) within the "Edge to Cloud Continuum" which has achieved relevant results and encouraging exploitation opportunities.

With its launch at the close of January 2024, this initiative gathered forward-thinking project proposals from numerous consortia and, following meticulous deliberation, we are pleased to unveil the distinguished winners of this initiative. These outstanding projects will be celebrated at the forthcoming ECS Brokerage Event 2024[3], on 20 and 21 February in Brussels, at the Edge to Cloud Continuum space. We decided to organise the 2024 RIAs Challenge in conjunction with the ECS Brokerage Event 2024 to allow the selected projects to showcase their added value, potential impact and innovation, and to build a bridge with the industry to encourage the exploitation of their significant results and the creation of new project proposa

In this article we present a summary of the four winners of the 2024 RIAs Challenge. In the next INSIDE Magazine issue we will present a second group of projects which applied to the Challenge and deserve attention for their promising results and potential impact.

NEPHELE

What is your project about?

NEPHELE is a RIA (Research and Innovation Action) project funded by the Horizon Europe programme under the topic "Future European platforms for the Edge: Meta Operating Systems". Its vision is to enable the efficient, reliable and secure end-to-end orchestration of hyper-distributed applications over programmable infrastructure that is spanning the computing continuum from Cloud-to-Edge to IoT, removing existing openness and interoperability barriers in the convergence of IoT technologies against cloud and edge computing orchestration platforms, and introducing automation and decentralised intelligence mechanisms powered by 5G and distributed AI technologies.

What areas of research does your project cover?

We cover the research areas of orchestration mechanisms for the computing continuum, virtualisation of IoT devices, the convergence of IoT technologies with edge and cloud computing technologies, and the interoperability of IoT technologies. Machine learning techniques are used for the development of the various orchestration mechanisms.

How would you describe your project's added value, impact, innovation and results?

The NEPHELE project introduces two core innovations, namely:

1. An IoT and edge computing software stack for leveraging the virtualisation of IoT devices at the edge part of the infrastructure and supporting openness and interoperability aspects in a device-independent way. Through this software stack, the management of a wide range of IoT devices and platforms can be realised in a unified way, avoiding the need for middleware platforms, while edge computing functionalities can be offered on demand to efficiently support the operations of IoT applications.
2. A synergetic meta-orchestration framework for managing the coordination between cloud and edge computing orchestration platforms, through high-level scheduling supervision and definition, based on the adoption of a "system of systems" approach. The NEPHELE outcomes are going to be demonstrated, validated and evaluated in a set of use cases across various vertical industries, including areas such as disaster management, logistics operations in ports, energy management in smart buildings and remote healthcare services. The NEPHELE outcomes are going to be made available as open-source and provided for adoption and extension to the research community.

Website	https://nephele-project.eu/
Contact person	Anastasios Zafeiropoulos
	tzafeir@cn.ntua.gr
LinkedIn	https://www.linkedin.com/company/nephele/

[1] https://eucloudedgeiot.eu/

[2] https://www.inside-association.eu/post/rias-challenge-2024

[3] https://www.inside-association.eu/post/save-the-date-to-join-the-electonic-components-and-systems-community-at-the-ecs-brokerage-event-2024

MYRTUS

Multi-layer 360° dYnamic orchestrion and interopeRable design environmenT for compute-continUum Systems

What is your project about?
MYRTUS aims to unlock the new living dimension of CPS, embracing the principles of the TransContinuum Initiative, integrating edge, fog and cloud computing platforms. MYRTUS leverages an AI-powered cognitive engine to orchestrate collaborative distributed and decentralised agents and components. Additionally, components must be augmented with interface contracts covering both functional and non-functional properties.

What areas of research does your project cover?
Computer sciences; Information science; Bioinformatics; Computer hardware and architecture; Design environment; Dynamic orchestration; Computing continuum; Interoperability; AI.

How would you describe your project's added value, impact, innovation and results?
MYRTUS contributes to creating new knowledge in the computing continuum domain, with methodologies and tools for node execution and processing portability over edge-fog-cloud, including dynamic and seamless orchestration. The goal is to become a reference in the computing continuum, offering solutions that overcome the problem related to vendor/platform lock-in, therefore promoting and facilitating the adoption of MYRTUS technologies among startups and SMEs, reducing their development time and cost. MYRTUS embraces the sustainable and responsible computing paradigm, promoting obsolescence avoidance (supported by MYRTUS principle of openness, interoperability, and portability) and resource saving and energy efficiency (supported by HW specialisation and optimisation techniques). Collaboration is a key driver of innovation and knowledge exchange that can lead to more efficient research outcomes and a better understanding of the broader research landscape. MYRTUS has a strategy to establish synergies with other projects and initiatives, including important associations (HIPEAC, INSIDE, Gaia-X, etc.), technology communities, the IPCEI initiatives and the projects that will be funded in the upcoming Cluster 3 calls.

Website https://www.linkedin.com/company/myrtus-eu/
Contact person Katiuscia Zedda
katiuscia.zedda@abinsula.com

FLUIDOS

Flexible, scaLable, secUre, and decentralIseD Operating System

What is your project about?
FLUIDOS (Flexible, scaLable, secUre, and decentralIseD Operating System) is an innovative initiative working towards a dynamic and trustable cloud to edge to IoT continuum. FLUIDOS develops a software framework with modular components that can adapt to different hardware/software platforms. It enables a single FLUIDOS node (either an individual device, an edge server or a cloud cluster) to support and expose a wide range of resources and services, which can be seamlessly shared and consumed by other nodes, either nearby or remote. The project's vision for the future is based on five essential aims: (1) fluidify edge and cloud computing through decentralised, autonomous resource integration; (2) shift computing gravity beyond data centres, fostering cross-provider community computing; (3) orchestrate services based on energy efficiency parameters using AI; (4) implement a Zero-Trust security approach for authenticated, authorised access to dispersed resources; and (5) cultivate a multi-stakeholder edge services market, promoting European digital autonomy.

What areas of research does your project cover?
On a wider level, FLUIDOS covers the following research areas: Open Source Computing Continuum; Cloud Computing; Edge Computing-Internet of Things. More specifically, FLUIDOS has three main use cases: 1. Intelligent Power Grid – Energy; 2. Smart Viticulture – Agriculture; 3. Robotics Logistics. FLUIDOS envisions extending these use cases to cover other research areas such as Software-Defined Vehicle (SDV) in the automotive industry.

How would you describe your project's added value, impact, innovation and results?
The IT landscape has evolved into a world of hyperconnectivity, where devices and information systems communicate and exchange data on numerous applications. FLUIDOS' added value derives from the ability to leverage the enormous, unused processing capacity at the edge, scattered across heterogeneous edge devices that struggle to integrate with each other and to form a coherent seamless computing continuum. FLUIDOS seeks to create impact through a user-centric, intelligent system that optimises design, functionality, security, privacy, cost-effectiveness, carbon footprint and resource sharing in various domains by: (1) developing a software platform with modular components that can adapt to different hardware/software platforms and enable a single node to support and expose a wide range of resources and services; (2) establishing a decentralised architecture where FLUIDOS instances interact horizontally and vertically, enabling seamless orchestration across devices, edge and cloud, supporting hyper-distributed applications and non-cloud services; (3) ensuring secure interactions within its infrastructure through a zero-trust approach through the employment of security mechanisms, hardware security anchors, and AI-based techniques for attack detection and mitigation while preserving user privacy; (4) developing an energy- and carbon-aware computing model to optimise energy consumption, carbon emissions, and costs through the utilisation of lightweight computing nodes, carbon-aware orchestration, and harmonised energy information protocols; (5) allowing different actors to choose between sharing, selling or renting resources and services, thereby supporting intermediate brokering systems, open protocols, and resource and service-sharing standards across administrative/business domains; and (6) fostering an open and collaborative ecosystem involving developers, users and stakeholders, whereby participation in industry-driven events, contributions to open-source communities and the establishment of a community of early adopters is encouraged. Among the flagship tangible results in FLUIDOS, we can cite: (a) the REsource Advertisement and Reservation protocol (REAR), that enables FLUIDOS nodes to advertise, negotiate, reserve and acquire resources (e.g., CPUs) and services (e.g., DB-as-a-service); (b) the meta-orchestrator for resource and service placement supporting different optimisation strategies; (c) the infrastructure-oriented, dynamic and transparent computing continuum based on the Liqo.io open-source project, including strong security characteristics; and (d) the policy-oriented orchestration framework to facilitate the seamless deployment and orchestration of resources and services.

Website https://www.fluidos.eu/
Contact person Alexia Zafeiropoulou

ICOS

Towards a functional continuum operating system

What is your project about?
The unstoppable proliferation of novel computing and sensing device technologies, and the ever-growing demand for data-intensive applications in the edge and cloud, are driving the next wave of transformation in computing systems architecture. The resulting paradigm shift in computing is centred around dynamic, intelligent and yet seamless interconnection of IoT, edge and cloud resources in one computing system, to form a continuum. A continuum, today also referred to as cloud continuum, IoT continuum, edge-to-cloud or fog-to-cloud, is expected to provide the means for data processing both in the edge and cloud, while inferring and persisting important information for post-mortem and offline analysis. We envision a holistic approach towards the solutioning of this technology trend in future systems, by architecting, designing and implementing the continuum as extensible, open, secure, adaptable, AI-powered as well as highly performant and technology agnostic, managed through a metaOS: IoT2Cloud Operating System (ICOS).

What areas of research does your project cover?
The ICOS project covers the set of challenges that emerge when addressing the continuum paradigm, proposing an approach embedding a well-defined set of functionalities, resulting in the definition of an IoT2Cloud Operating System. Indeed, ICOS is designing, developing and validating a meta operating system for the continuum, addressing the following challenges: i) the volatility and heterogeneity of devices, continuum infrastructure virtualisation and diverse network connectivity; ii) optimised and scalable service execution and performance, as well as consumption of resources, including power consumption; iii) guaranteed trust, security and privacy; and iv) reduction of integration costs and effective mitigation of cloud provider lock-in effects, in a data-driven system built upon the principles of openness, adaptability, data sharing and a future edge market scenario for services and data. Everything is demonstrated through four use cases in different domains: In-car advanced infotainment and multimedia management system; Agriculture operational robotic platform; Railway structural alert monitoring system; and Energy management and decision support system.

How would you describe your project's added value, impact, innovation and results?
The main ICOS result consists of three key layers: Meta-Kernel, Security and Intelligence layers, as well as two additional modules, ICOS Shell and Data Management. The Meta-Kernel Layer is responsible for providing the principal OS functionalities to the continuum. It closely integrates with the Security Layer responsible for guaranteeing security and trust provisioning, as well as with the Intelligence Layer that will enrich any action to be taken with innovative AI approaches. Moreover, ICOS also includes a Shell to interface users as well as Data Management to handle all data related issues. Thus, the main innovations and impact can be summarised as follows:

1. An open, intelligent, multi-platform, plug-and-play and technology-agnostic meta operating system encompassing a green orchestration strategy to suit the specific technical and business needs and requirements of a smart continuum scenario.
2. Novel, autonomous, intelligent and adaptable data and resource utilisation methods enhancing the abilities of app developers to take advantage of all continuum resources.
3. A trustworthy set of resources offering security, resilience, reliability and privacy functionalities in the continuum. Everything is included in an open innovation environment to facilitate the engagement of the scientific and engineering community.

Contact person Francesco D'Andria
francesco.dandria@eviden.com

Online version is available at Inside-association.eu

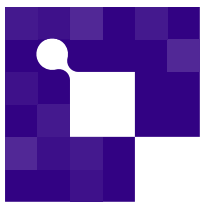
Publisher
INSIDE Industry Association
High Tech Campus 69-3
5656 AG Eindhoven, The Netherlands

Design and Creative lay-out
Studio Kraft – Veldhoven, the Netherlands

Acknowledgements
With thanks to INSIDE involved persons for any assistance and material provided in the production of this issue. With thanks to the interviewees, project participants, INSIDE Industry Association office, the INSIDE Industry Association Presidium and other INSIDE Industry Association-involved persons for any assistance and material provided in the production of this issue of the INSIDE Magazine.

Contributions
The INSIDE Industry Association office is interested in receiving news or events in the field of Intelligent Digital Systems. Please submit your information to info@Inside-association.eu

Inside
Industry Association



Inside

Industry Association

Inside-association.eu