HiPEAC info 77

JANUARY 2026

HiPEAC 2026 Kraków

Digital sovereignty: what, why, and how?

How Michaela Blott tailors custom compute for efficient AI

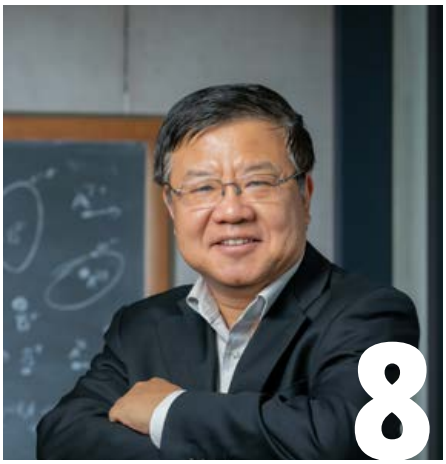
Frank Karlitschek on providing an alternative to US cloud platforms

Navigating the impact of AI on hardware, with Deming Chen

Contents

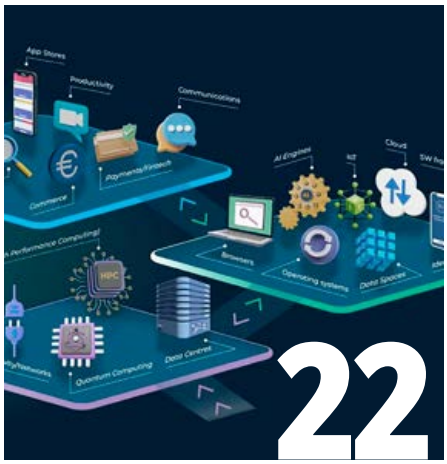
Michaela Blott on the efficiency of tailored compute

Frank Karlitschek on alternative collaboration platforms



How AI is transforming hardware, with Deming Chen

HiPEAC Technology Transfer Awards 2025



Building digital sovereignty and capacity in Europe

Photo credit: vectorfusionart | stock.adobe.com

EU technology initiatives for a sovereign Europe

Spanning the compute continuum from edge to cloud, HiPEAC (High Performance, Edge And Cloud computing) is a network of over 2,000 world-class computing systems researchers, industry representatives and students. First established in 2004, the project is now in its seventh edition. HiPEAC7 focuses on networking and roadmapping activities: bringing the computing community together in Europe, exchanging ideas, building thriving European value chains and exploring the long-term vision for computing systems.

hipeac.net

hipeac.net/linkedin

@hipeac / @hipeacjobs

hipeac.net/tv

Funded by the European Union

The HiPEAC project has received funding from the European Union's Horizon Europe research and innovation funding programme under grant agreement number 101069836. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Cover image: Oleksandr Bochkala | stock.adobe.com – Image generated with artificial intelligence

Design: www.magelaan.be

Editor: Madeleine Gray

The central theme of this magazine is digital sovereignty. Digital sovereignty is often used interchangeably with strategic digital autonomy, but they are quite distinct. Digital sovereignty as defined in the recently signed *European Digital Sovereignty Declaration*, is a matter of authority. It is the power to govern the digital space to European values. It allows the European Union (EU) to enforce frameworks like the General Data Protection Regulation (GDPR), the Digital Markets Act, the AI Act, the Data Act, etc. While Europe excels at regulation, there is mounting geopolitical pressure to deregulate – effectively asking the EU to cede a degree of its digital sovereignty.

Conversely, strategic digital autonomy is a matter of capacity. While sovereignty provides the legal authority, autonomy provides the 'muscle' – the ability to act independently by reducing strategic dependencies. This involves, for example, building the 'EuroStack', a robust foundation ranging from sovereign cloud infrastructure and high-performance computing to open-source ecosystems and semiconductors, ensuring that European laws are backed by the technical capability to implement them.

This distinction is crucial for our community: if Europe depends entirely on foreign proprietary hardware or black-box algorithms, its legal sovereignty is undermined by technical dependence. While the HiPEAC community may have limited influence over legal frameworks, we play a vital role in strengthening Europe's strategic digital autonomy by staying on top of the digital evolution.

Building the EuroStack is a vital step toward strategic autonomy, but it is not a destination. The US and China already have their own stack; building ours will at best allow us to catch up, but it will not allow us to lead. To compete in 2035 and beyond, we must look past today's requirements. Since the future is inherently unpredictable, our best path forward is an antifragile strategy: cultivating a generation of innovators and an entrepreneurial ecosystem that thrives on volatility. By transforming Europe into a 'startup continent' by 2040, we ensure that we don't just survive the coming geopolitical storms but grow stronger because of them.

My wish for Europe in 2026 is that it takes steps to massively unleash the creativity and entrepreneurship of its citizens. By facilitating this talent, we do more than just secure our autonomy and sovereignty – we will build a more resilient, prosperous society and consolidate Europe's position as the best continent on Earth to live, work, and run a business.

Koen De Bosschere, HiPEAC coordinator

A well-known figure in the computer-architecture community, HiPEAC 2026 keynote speaker Michaela Blott is a senior fellow at AMD with over 25 years of leading-edge computer architecture and advanced field-programmable gate array (FPGA) and board design. We caught up with Michaela in advance of the HiPEAC conference to find out about why she's passionate about computer architecture, her top trends to watch, and how AMD is focused on ever-more efficient technology for artificial intelligence (AI).

'Pairing novel algorithms with custom hardware architectures could enable step-change improvements in model capability, energy efficiency and performance'

What got you into computer architecture, and what keeps you passionate about the field?

Since university, I've always enjoyed gaining a deep understanding of everything, from the application all the way down to the transistor level. That, combined with a passion for building the best possible solution to a given problem, naturally led me to designing custom compute architectures early on, three decades ago. At the time, this was a niche focus, but custom hardware architectures are increasingly becoming mainstream.

The recent surge in demand for computing power and memory, particularly due to AI workloads, combined with technological constraints and customer challenges around monetizing high-cost inference, has accelerated this shift. Custom hardware architectures are exceptionally well suited to address these evolving industry requirements.

"The recent surge in demand for computing power and memory has accelerated the shift to custom hardware architectures"

Which of your professional achievements are you most proud of, and why?

That would be FINN, a project focused on creating highly customized inference accelerators on FPGAs which we started almost 10 years ago. It can autogenerate a dataflow architecture and was originally developed for fully binarized neural networks, delivering significant performance improvements over conventional AI accelerators.

We've since expanded it to support a range of custom data types, both integer and floating point, and we are currently combining this with the former Microsoft Brainwave team to explore optimized transformer accelerators.

What key trends in computer architecture (and computing systems more generally) are you keeping an eye on?

One key trend is the growing importance of densely integrated 3D-stacked memory. As compute architectures become increasingly memory-constrained, traditional dynamic random-

"FPGAs offer the flexibility to customize architectures for specific applications, thereby enhancing efficiency and performance"

access memory (DRAM) – including high-bandwidth memory (HBM) – is starting to show its limits in meeting the evolving capacity and bandwidth demands of AI workloads.

A second area is algorithm-level innovation, such as the resurgence of recurrent neural networks (RNNs), which may offer more scalable approaches to AI. This is an active and promising area of research, and one that compute architects should track closely.

Finally, rack-level optimization is becoming critical. As the industry moves beyond chip-level performance, the focus must shift to full-system efficiency to unlock the next wave of AI scalability.

Graphics processing units (GPUs) are currently hogging the spotlight in AI development. What role can other processors, such as FPGAs, play?

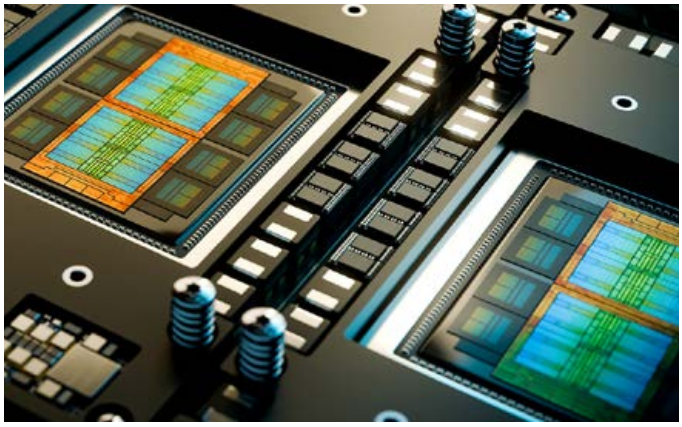
FPGAs offer the flexibility to customize architectures for specific applications, thereby enhancing efficiency and performance. By implementing only what is needed for a specific use case, we can avoid unnecessary passes through memory with dataflow, fuse more operators and in general process the data while it's in motion.

However, it must be said, not every workload is a good fit for FPGAs – and, likewise, not every workload is ideal for GPUs.

Why are hardware diversity and model optimization important in AI? Can't we just keep on scaling up with GPUs and large language models (LLMs)?

While large language models have driven remarkable progress, growing questions are emerging around their ability to generalize and understand fundamental concepts, among other things, limited by the availability of high-quality training data, and the increasing demands for compute and memory which have outpaced standard technology scaling.

This is where continued innovation will be key. Developing novel algorithms with better scaling properties – such as RNNs



The AMD Instinct MI350 series

– and pairing them with custom hardware architectures could enable step-change improvements in both model capability but also in energy efficiency and performance.

This may also represent a strategic opportunity, particularly for regions like Europe, to explore differentiated paths and help shape the global direction of AI development.

Are you concerned about the environmental impact of rolling out AI? If so, what actions are your team and AMD more generally taking to address this?

At AMD, energy efficiency has long been a guiding core design principle aligned to our roadmap and product strategy. For more than a decade, we've set public, time-bound goals to dramatically increase the energy efficiency of our products and have consistently met and exceeded those targets.

We recently surpassed our '30x25' energy efficiency goal, achieving a 31.7x improvement in energy efficiency for AI and high-performance computing (HPC) processors from 2020 to 2023. Building on that success, AMD has set a new, ambitious goal: to deliver a 20x improvement in energy efficiency for AI systems at the rack level by 2030. This reflects the industry's shift from chip-level to system-level optimization and underscores our commitment to responsible, high-performance innovation.

HiPEAC 2026 keynote speaker Frank Karlitschek is the co-founder and chief executive of Nextcloud, which provides a fully open-source, on-premises content collaboration platform. Since founding the company in 2016, Frank has become a vocal advocate for European digital sovereignty, as well as for open-source and privacy-preserving technological solutions. We caught up with Frank in advance of the HiPEAC conference in Kraków to talk about meaningful digital sovereignty, what makes Nextcloud special, and what he's learned in building the business.

'With Nextcloud, we have now an alternative to the big collaboration platforms from the United States'

What does 'digital sovereignty' mean, in your view? Why is it digital sovereignty important, both for individuals and for Europe?

In my opinion, sovereignty means the absence of strong dependencies on third parties. The more sovereignty we gain, the more it allows us to change our software or cloud provider whenever we think it is necessary. This makes us less open to blackmail. In the end it is a question of freedom.

What do you think needs to happen for Europe to achieve meaningful digital sovereignty? What role should companies, policy makers, etc. play in this?

We all should act strategically and improve our digital sovereignty. Governments need to act as anchor clients for digital open-source products coming from Europe. We see a lot of political initiatives which aim for more sovereignty, but we know that these take time. It is the local private sector that has to promote our economic independence in a bottom-up approach.

"Only open-source software prevents dependencies on individual providers and allows independent security audits"

Frank talking at the Nextcloud Enterprise Days Paris in November 2025

> *"The community is the engine room of our product, and generates huge value. If you put effort into your community, it is a major differentiator"*

What role does open source play in sovereignty? Do you think open source is important for technological development in Europe? If so, why?

Only open-source software prevents dependencies on individual providers and allows independent security audits. Everyone can take the source code and build its own solution if he or she wants to. If your cloud or software provider increases the prices, you can run it on your own servers. If you are not happy with what the provider does, you can take it and fork the project. You will always have the full choice of options.

In addition, you can rely on a community which has a constant eye on the code, improves it and spots vulnerabilities very quickly.

What inspired you to start Nextcloud? What makes Nextcloud unique, and why would you encourage people to use it?

I have been working with open source for over 25 years now and always wanted to create something that gives people back the control over their data. With Nextcloud, we have now an alternative to the big collaboration platforms from the United States. The main differentiator is that Nextcloud is 100% open source. For anyone who thinks data privacy is important, Nextcloud is the tool they should use.

What have you learned during the process of building the company?

There are a lot of important learnings, but I will pick three of them. First, I think it is important to have an ethical purpose for the company and not only increase shareholder value. This keeps you and your employees motivated even during rough seas and helps to build up a sustainable business. Second, in the open-source world the business model is an important choice to make. For us, the only feasible model is to sell enterprise subscriptions, as this does not compromise open source. The third learning is to value the community. It is the engine room of our product, and generates huge value. If you put effort into your community, it is a major differentiator.

The Nextcloud Community Conference in September 2025

What are the characteristics of your ideal cloud services, from the user perspective?

The core idea of the internet is decentralization. I think the perfect cloud service would need to work this way, too. You should be able to host software on your own server if you want to but collaborate with others through federation. Nextcloud, for example, is hosted on over 500,000 servers across the internet.

The second core aspect is open source. If you know the source code, you have the full control, and that is what counts.

What changes would you like to see in the technology sector in Europe 10 years from now?

We should start to value our own technology industry and stop saying that Europe can't do it. There are plenty of great tech companies which deliver alternatives to US products. If Europe had faith in its own solutions, this would be a huge push for our innovation ecosystem. Politicians can make a massive difference by actively advocating for open source and governments who implement it.

HiPEAC 2026 keynote speaker Deming Chen is the Abel Bliss Professor in the Grainger College of Engineering at the University of Illinois Urbana-Champaign. His research interests include machine learning and AI, system-level design methodologies, hybrid cloud systems, security and confidential computing, and reconfigurable and heterogeneous computing, and he has won multiple recognitions for his work. In this in-depth interview, he reflects upon major shifts in computer architecture and hardware design over the last 25 years, considers the impact of machine learning on hardware design, and discusses the demands artificial intelligence (AI) workloads are placing on hardware.

'While AI-assisted hardware design is far from a solved problem, it is well positioned to fundamentally reshape how hardware systems are conceived, built, verified, and deployed'

What are some of the most profound changes in the fields of computer architecture and hardware design that you've witnessed so far in your career?

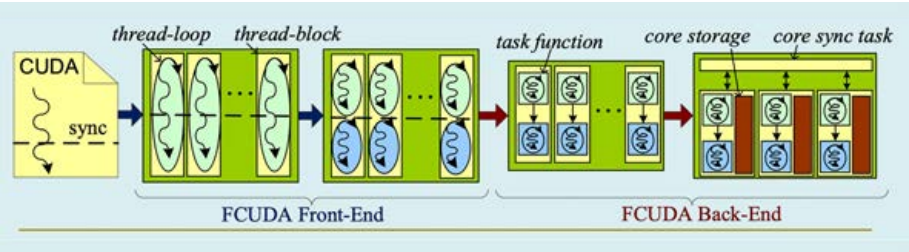
The most profound shift I have witnessed over the past 25 years is the transition from general-purpose computing centred on the central processing unit (CPU) to heterogeneous, accelerator-rich systems. Early in my career, architectural innovation was largely driven by instruction-level parallelism and frequency scaling. As power and memory walls became dominant constraints, the field pivoted toward specialization, reconfigurability, and energy-efficient architectures.

Another major change is the rise of hardware-software co-design as a necessity rather than an option. Compilers, runtimes, and architectures are now deeply intertwined, a trend reflected in my work on FCUDA, DNNBuilder, SkyNet, and ScaleHLS. The emergence of AI has further accelerated this shift, fundamentally altering architectural priorities toward memory-centric design, data movement minimization, and domain-specific accelerators.

Equally transformative has been the rise of open-source ecosystems and deep industry-academia collaboration. Tools and ideas now propagate far more rapidly, allowing research prototypes to transition into production systems, as exemplified by the adoption of our recent work, Medusa, in NVIDIA TensorRT-LLM. At the same time, AI itself is emerging as a powerful tool for architecture and system design, reshaping how future hardware will be conceived, optimized, and validated.

Which of your research results have had most impact?

Several of my research contributions stand out for their sustained impact across academia and industry. FCUDA was the first compiler framework to enable efficient translation of CUDA programs to field-programmable gate arrays (FPGAs). FCUDA created a new programming environment where both graphics processing units (GPUs) and FPGAs could be programmed using the same language in a heterogeneous compute system. This vision and effort preceded similar commercial OpenCL-to-FPGA flows offered by Intel and Xilinx by five to six years. It also significantly lowered the barrier for GPU programmers to leverage reconfigurable hardware. This work helped catalyse a broader movement toward productivity-driven accelerator design.



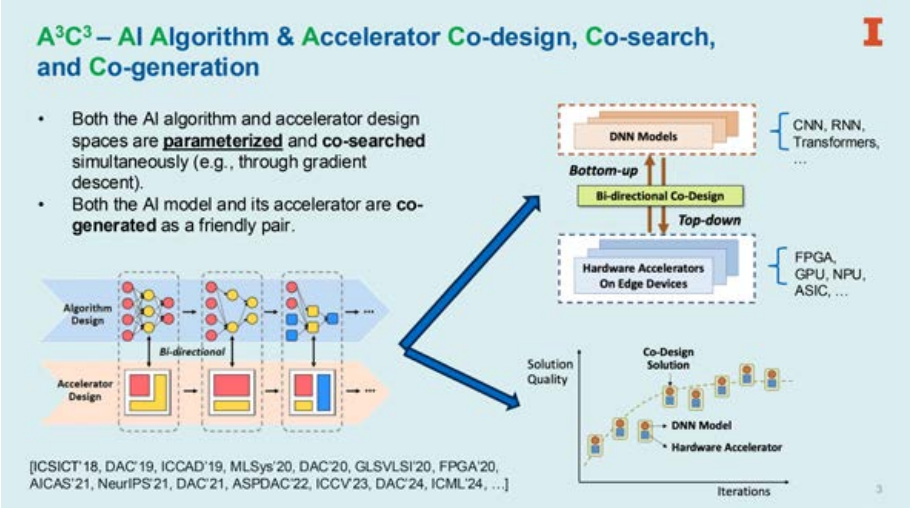
FCUDA was the first compiler framework to enable efficient translation of CUDA programs to FPGAs

DNNBuilder and ScaleHLS represent another impact area. DNNBuilder enabled automated mapping of deep neural networks onto FPGAs, while ScaleHLS became the first compiler to map PyTorch models directly to customized FPGA accelerators. With thousands of downloads worldwide, ScaleHLS has influenced both research workflows and industrial prototyping.

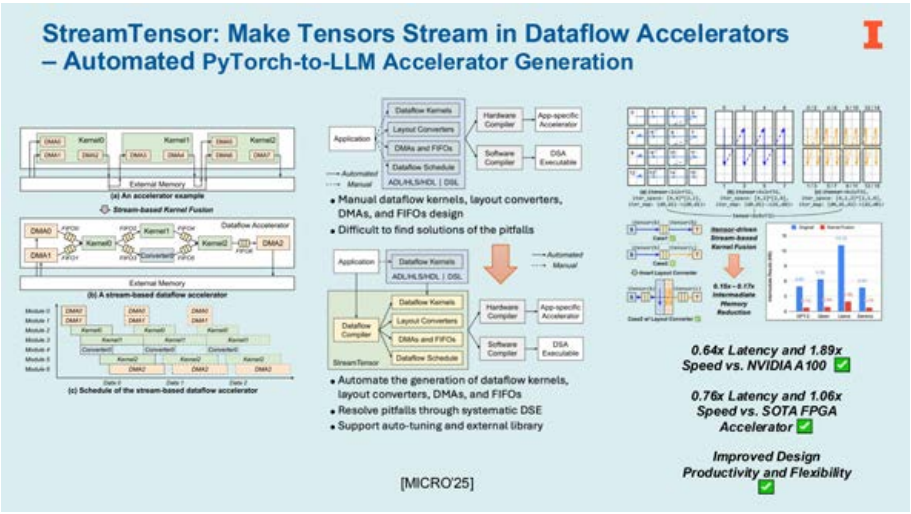
As mentioned above, Medusa's integration into NVIDIA TensorRT-LLM, delivering up to 3.6× inference speedup, exemplifies successful translation from research to production. Also, my A3C3 methodology

(AI Algorithm and Accelerator Co-design, Co-search, and Co-generation) has had lasting influence by framing AI and hardware as co-evolving artefacts rather than separate optimization targets. Following the A3C3 design methodology, we developed the SkyNet model, which won double championships in 2019 in the DAC System Design Contest for both GPU and FPGA tracks for low-power image object detection, outperforming 100+ competitors worldwide. Later, SkyNet has been used by many other contestants and several companies.

Finally, the startup company I co-founded, Inspirit IoT, Inc., recently launched a product called StreamTensor, a compiler framework that tackles the data-movement problem in AI hardware by transforming PyTorch models into optimized dataflow implementations. It includes a novel iterative tensor (itensor) type that systematically encodes stream information, describing how data moves, while the compiler identifies optimal tensor tiling, kernel fusion, and hardware-resource allocation. Based on FPGA evaluations on large language models (LLMs), StreamTensor achieves up to 0.64x lower latency and up to 1.99x higher energy efficiency compared to GPUs.



The A3C3 promotes hardware-software co-design for holistic optimization



StreamTensor, a commercially available product from Inspirit IoT, compiles PyTorch LLM models into stream-oriented dataflow designs for FPGAs

What impact has the application of machine learning to the design of hardware had? What are some of the promises and perils of AI in hardware design and computing systems more generally?

This is a very timely question. Machine learning (ML) has begun to meaningfully influence hardware design by introducing data-driven techniques into design space exploration, hardware code generation, and system- and tool-flow optimization. In my own work, ML-based methods have been applied to accelerator selection, high-level synthesis optimization, and scheduling decisions, enabling systems to adapt more effectively to workload characteristics and design constraints. Although these approaches are still at an early stage, they already demonstrate strong potential to augment human expertise and significantly improve design productivity.

At the same time, substantial challenges remain. High-quality industrial design data suitable for training ML models is often scarce or proprietary, which limits model robustness and generalization. Hardware code generated by AI today can be incomplete or error-prone, raising serious concerns about correctness, verification, and trustworthiness. In addition, AI-driven design tools can make internal design decisions less transparent, complicating debugging, validation, and security assurance.

Recognizing both the promise and the challenges, in 2024 I co-founded the first IEEE International Workshop on LLM-Aided Design (LAD) together with Dr Ruchir Puri of IBM, which evolved into the inaugural IEEE International Conference on LLM-Aided Design in 2025. This venue focuses on leveraging LLMs to assist the design of circuits, software, and computing systems with improved quality, productivity, robustness, and cost efficiency. The conference has attracted hundreds of researchers and practitioners, many from industry, and has emerged as a highly impactful and promising forum. Looking ahead, while AI-assisted hardware design is far from a solved problem, it is well positioned

to fundamentally reshape how hardware systems are conceived, built, verified, and deployed, provided these challenges are addressed through careful co-design, rigorous verification, and responsible deployment.

What are the most important requirements hardware must respond to in the AI era? How has this affected your approach to hardware design?

In my view, hardware in the AI era must address three dominant requirements: energy efficiency, memory efficiency, and adaptability. Modern AI workloads are increasingly memory bound, driven by massive parameter counts and the cost of data movement rather than raw computation. This shift has elevated the importance of memory-centric designs, near-memory computing, dataflow-based architectures, and high-bandwidth interconnects. As noted above, our recent work, StreamTensor, the first compiler to map PyTorch LLM models directly onto FPGAs, demonstrates substantial latency and energy efficiency gains by generating custom dataflow-based LLM accelerators that significantly reduce data movement (see StreamTensor figure on p.9). This approach has already gained notable traction in both industry and the research community.

Adaptability is equally critical. AI models evolve over a timescale of months, while hardware platforms evolve over years. This mismatch has reinforced my focus on reconfigurable and programmable architectures and compiler-driven specialization. My work on FPGA-based accelerators, heterogeneous systems, and dynamic reconfiguration reflects the need to maintain flexibility without sacrificing efficiency, allowing hardware to keep pace with rapidly evolving AI algorithms.

Finally, energy efficiency and scalability now dominate system-level design in the AI era. Hardware must operate efficiently from edge devices to large-scale data centres under strict power and thermal constraints. These realities have pushed my research toward cross-layer co-design, where algorithms, compilers, hardware, and systems are optimized jointly to maximize intelligence per joule rather than peak throughput. In this context, I recently planned and co-organized the National Science Foundation workshop 'AI+HW 2035: Shaping the Next Decade' and am contributing to a vision paper outlining a 10-year roadmap toward achieving a 1000× improvement in AI training and inference efficiency through deep integration across abstraction layers involving both AI model and hardware (HW) innovations.

What role do you think open source will play in hardware development in the future?

Open source will play a foundational role in the future of hardware development, much as it has in software. As hardware systems become increasingly complex and heterogeneous, no single organization can innovate effectively in isolation. Open-source tools, benchmarks, and infrastructure enable shared progress, reproducibility, and rapid iteration across the community.

My own work reflects this philosophy. Our open-source frameworks such as FCUDA, DNNBuilder, ScaleHLS, SkyNet, CSRNet, and Medusa have achieved broad adoption, fostered community-driven improvement, and delivered real-world impact. These platforms also serve as powerful educational resources, helping train the next generation of researchers and engineers in modern hardware-software co-design practices.

Looking ahead, open source will be essential for addressing AI-era challenges. For both government and industry, open ecosystems reduce risk, avoid vendor lock-in, and accelerate workforce development. Through my leadership roles in the IBM–Illinois Discovery Accelerator Institute and the AMD Center of Excellence, I actively promote and support researchers in building and sustaining open-source ecosystems that deliver lasting impact to industry and society.

The NSF AI+HW 2035 workshop featured a keynote by Yann LeCun and leading researchers from academia and industry, including HiPEAC associate members Kunle Olukotun and Subhasish Mitra

Witamy w Krakowie!
Welcome to Kraków!

HiPEAC 2026 General Chair Tomasz Kryjak (AGH University of Krakow) gives us a glimpse into this year's conference's beautiful host city.

Why is Kraków a good location for the HiPEAC conference?

Kraków, the former capital and second-largest city of Poland, is home to leading Polish universities and research institutes, including the Jagiellonian University, AGH University of Krakow, Cracow University of Technology, and Kraków University of Economics. Together, they conduct internationally recognized research in computer architecture, embedded systems, artificial intelligence (AI), and software engineering – areas that closely align with HiPEAC's core focus.

Kraków also hosts ACK Cyfronet AGH, Poland's leading high-performance computing centre, which plays a key role in European research infrastructures, supports large-scale simulations and data-intensive research, and currently runs the Gaia AI Factory project. In addition, the city has a vibrant ecosystem of research and development (R+D) centres and technology startups, making it a natural meeting point for academia and industry.

Tell us about the local computing ecosystem.

Kraków hosts a large number of high-tech R+D centres operated by major international companies, including Google, IBM, Cisco, Ericsson, Nokia, Aptiv, ABB, Hitachi, Honeywell, Woodward, Genetec, Arteris and Comarch. This industrial base is complemented by local innovation drivers such as the Kraków Technology Park, startup accelerators, and specialized ecosystems focused on 5G, the internet of things (IoT), and smart technologies, which help translate academic research into real-world applications and foster cross-sector collaboration. In addition, Kraków hosts a growing number of startups operating across the broader computing ecosystem. Ultimately, however, the city's most valuable asset is its talented pool of young people, who, after completing their education at Kraków's universities, actively contribute to and sustain the local and international R+D ecosystem.

What initiatives at your university would you highlight?

A significant proportion of research initiatives focus on artificial intelligence. At the heart of this is ACK Cyfronet AGH with the Gaia AI Factory, which provides the high-performance computing infrastructure essential for AI-driven research across disciplines. At AGH University of Krakow, numerous research groups explore different aspects of AI, including the Center of Excellence in Artificial Intelligence, while at the Jagiellonian University, many AI-focused groups are coordinated by the recently established Jagiellonian Centre for Artificial Intelligence, fostering both foundational research and innovative applications.

In the embedded systems domain, my Embedded Vision Systems Group began with embedded vision implemented on field-programmable gate arrays (FPGAs), and now we are expanding to new sensor modalities, including event cameras and LiDARs, while also exploring embedded control for AI-driven robotics. I strongly believe that embedded computing is a cornerstone for embodied AI – from drones and self-driving cars to walking robots – and this area is poised for rapid growth across AGH and other Kraków universities.

Any recommendations for things to do in Kraków?

Very close to the conference venue, on the opposite bank of the Vistula River, lies Wawel Hill with its Royal Castle, the former seat of Polish kings and a symbol of the nation's history. From there, a pleasant walk leads to the Main Market Square (Rynek Główny), one of the largest medieval squares in Europe, home to the iconic St Mary's Basilica and the historic Cloth Hall. The nearby Kazimierz district, with its cafés, restaurants, and vibrant atmosphere, is an excellent place to continue conversations after conference sessions.

The UNESCO-listed Wieliczka Salt Mine, about 45 minutes from Kraków, offers an extraordinary underground world of tunnels, chambers, and sculptures carved entirely from salt. Nature lovers may also consider a trip to Zakopane in the Tatra Mountains. Finally, no visit to Kraków is complete without sampling traditional Polish cuisine – especially żurek soup, pierogi (dumplings), and bigos.

Winners of the 2025 HiPEAC Technology Transfer Awards

Since 2012, the HiPEAC Technology Transfer Awards have been celebrating examples of boundary-pushing research from the HiPEAC community reaching the market. For the 14th edition, five transfers were selected for an award, three of which resulted in new European technology companies. The winners will be recognized in an awards ceremony at the HiPEAC conference in January, and first-prize winners receive a cash prize of € 1,000.

ACE3: Democratizing large-scale AI through groundbreaking software acceleration

Xiaoyang Sun, University of Leeds, UK

ACE3 is an advanced software platform that removes memory bottlenecks in hardware accelerators, enabling training of large AI models without massive supercomputing resources.

Large neural networks often hit hardware limits, as even powerful accelerators are constrained by memory bandwidth and capacity. ACE3 addresses this with a software-based acceleration technique that optimizes data handling and load balancing across heterogeneous compute units, unlocking more performance from existing hardware and enabling much faster training of complex AI models without new specialized accelerators.

After being commercialized through a spinoff company, ACE3 AI LTD, this technology is now available through a user-friendly platform that delivers these optimizations to industry and research teams. In partnership with industries, the ACE3 platform has shown significant reductions in training time and computing costs on real-world workloads. ACE3 AI LTD's broader vision is to democratize advanced AI by easing hardware constraints and enabling more organizations to use state-of-the-art models without massive infrastructure investments.

The ACE3 team members are Dr. Xiaoyang Sun, Professor Jie Xu, and Professor Zheng Wang.

ACE3 website: ace3ai.com
Xiaoyang Sun's website: xshaun.github.io
Jie Xu's webpage on the University of Leeds website: eps.leeds.ac.uk/computing/staff/331/professor-jie-xu
Zheng Wang's website: zwang4.github.io

Belfort: Hardware acceleration for computing on encrypted data

Ingrid Verbauwhede, KU Leuven / Belfort, Belgium

Belfort is a new KU Leuven spinoff. It has launched the world's first available product that accelerates encrypted compute in hardware, making it possible to process encrypted data without ever decrypting it. It allows servers to compute directly on encrypted inputs, so sensitive information is never exposed, even during execution. Historically, this approach has been too slow and costly for real-world use. With Belfort's custom hardware architecture, encrypted compute is now viable in real time for the first practical workloads. From fraud detection to genomic analysis, privacy-preserving applications are moving from theory to reality.

Belfort is a spinoff from KU Leuven's COSIC (Computer Security and Industrial Cryptography) lab, a global leader in secure computation and cryptographic hardware. The research underpinning Belfort's core technology was developed over several years, and matured through several grants and research projects, including two grants from the European Research Council.

Belfort website: belfortlabs.com
COSIC lab: esat.kuleuven.be/cosic

The Belfort founding team: Laurens De Poorter (COO), Furkan Turan (Head of Engineering), Michiel Van Beirendonck (CEO) and Ingrid Verbauwhede (Head Scientist). Photo by Fred Paulussen (Fredography)

Accelerating early software development: A high-performance parallel hardware-simulation framework

Nils Bosbach, RWTH Aachen University, Germany

The Institute for Communication Technologies and Embedded Systems (ICE) at RWTH Aachen University has transferred a parallelization and co-simulation technology for SystemC-based virtual prototypes to MachineWare GmbH. The package combines a generic intra-process parallelization strategy for virtual central processing unit (CPU) models with a multi-process co-simulation orchestration layer. Integrated into MachineWare's ecosystem, the technology delivers faster virtual-platform execution, enabling broader CI/regression coverage, earlier bug discovery, and more efficient pre-silicon software development for heterogeneous system-on-chips (SoCs).

MachineWare website: machineware.de
ICE RWTH Aachen website: ice.rwth-aachen.de

Novel intermediate representation for efficient recursive query processing

Amir Shaikhha, University of Edinburgh, UK

Recursive query processing underpins a broad range of applications – from databases and knowledge-graph management to declarative networking, artificial intelligence, and machine learning. This knowledge transfer involves a novel intermediate representation (IR) for recursive queries, called TempoDL, together with its associated compilation and optimization techniques, introduced in 'Optimizing Nested Recursive Queries' (SIGMOD 2024). This transfer enables the direct incorporation of TempoDL and its associated compilation pipeline, developed in collaboration between the University of Edinburgh and RelationalAI, into an industrial relational knowledge-graph platform, allowing customers to express rich, recursive business logic with production-grade performance, scalability, and semantic guarantees.

University of Edinburgh Informatics website:
informatics.ed.ac.uk
RelationalAI website: relational.ai

Ximplic: Simplifying computing-in-memory

Simranjeet Singh and Farhad Merchant, University of Groningen, The Netherlands

The embedded system-on-chip (SoC) market is rapidly expanding, driven by the need for compact, energy-efficient, and high-performance computing solutions across sectors such as edge AI, the internet of things (IoT), automotive, and industrial automation. Valued at USD 169.5 billion in 2023 and projected to reach USD 295.5 billion by 2030 (CAGR ≈ 8.5%), this market's growth is increasingly constrained by the traditional von Neumann bottleneck, where excessive data movement between memory and processor dominates energy consumption and latency.

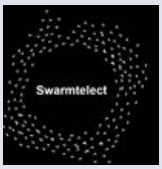
Computing-in-memory (CiM) directly addresses this limitation by integrating computation within memory arrays, thereby minimizing data transfer, enhancing parallelism, and enabling orders-of-magnitude improvements in energy efficiency. From a market-demand perspective, embedding CiM functionality into SoCs offers a disruptive advantage for real-time AI inference, sensor-edge processing, and low-power autonomous systems. As major semiconductor companies and startups alike invest in CiM-enabled architectures, the technology is poised to redefine the performance-efficiency balance of embedded SoCs and capture a significant portion of this rapidly growing market.

However, a major challenge in deploying CiM architectures lies in the extended development cycle associated with their complex integration into SoCs. To address this critical bottleneck, Farhad Merchant, and Simranjeet Singh, researchers at the University of Groningen, have founded a spinoff company, Ximplic, dedicated to accelerating the design of computing-in-memory architectures through advanced electronic design automation (EDA) tools and intellectual property blocks (IPs).

Ximplic's EDA platform and IPs tackle these challenges with a device-agnostic ecosystem tailored for CiM workflows. It supports any underlying device technology while unifying modelling, simulation, logic synthesis, and hardware prototyping within a single intelligent environment. Designers can seamlessly move from concept to silicon, independent of the memory technology, enabling rapid exploration and implementation. By simplifying complexity across the entire hardware design flow, Ximplic empowers faster innovation from device physics to system-level realization, paving the way for the next generation of energy-efficient, memory-driven computing.

Ximplic website: ximplic.com
Groningen Cognitive Systems and Materials Center (CogniGron) website:
rug.nl/research/fse/cognitive-systems-and-materials

Lakeside Lab researchers announce Swarmtelect spinoff



Researchers at Austria's Lakeside Labs have announced a future spinoff, Swarmtelect.

Swarmtelect offers intelligent workload allocation and orchestration solutions that use decentralized swarm intelligence to optimize complex systems like production lines, edge-fog-cloud computing environments, and energy networks. It provides autonomous decision-making, predictive resource allocation, and self-healing capabilities that the researchers say keep systems running smoothly even during disruptions, while integrating seamlessly with existing infrastructure.

'What really sets Swarmtelect apart is its focus on decentralized collective intelligence. Instead of relying on a central controller, systems can self-organize and adapt in real time. On top of that, we offer explainable resource management and turn advanced research into solutions that actually work in real industrial environments,' says Marija Gojkovic, one of Swarmtelect's founders.

'Swarmtelect is the result of a full year of sharpening an idea we truly believe in: making decentralized swarm intelligence practical, reliable, and ready for real-world infrastructures. What started as research has matured into a solution that can autonomously orchestrate resources where complexity is the norm. I'm excited to finally take this step from theory to impact,' added co-founder Melanie Schranz, a member of the DISCOVER-US network.

The founders of Swarmtelect – Melanie Schranz, Marija Gojkovic, Péter Forgács and Khalil Al-Rahman Youssefi – are open to partnerships, pilots, and collaborations.

swarmtelect.com
linkedin.com/company/swarmtelect

Anyway Systems spun off from EPFL's Distributed Computing Lab

Anyway Systems, recently spun off from EPFL's Distributed Computing Lab (DCL), is commercializing software that eliminates the need for data to be sent to third-party cloud services when artificial intelligence (AI) is used to complete a task. Developed by EPFL researchers Gauthier Voron, Geovani Rizk and Rachid Guerraoui, the software allows users to download open-source AI models and use them locally.

According to the launch announcement by EPFL, Anyway Systems combines distributed machines on a local network into an on-premise cluster. It uses robust self-stabilization techniques to optimize the usage of underlying local hardware. The researchers say that this means that huge AI models can be deployed on Anyway Systems in a few minutes, requiring no more than four machines with one commodity graphics processing unit (GPU) each.

'For years people have believed that it's not possible to have large language models (LLMs) and AI tools without huge resources, and that data privacy, sovereignty and sustainability were just victims of this, but this is not the case and smarter, frugal approaches are possible,' said Rachid Guerraoui, head of the DCL.

'Anyway Systems represents an interesting technology that optimizes resource usage while ensuring data security and sovereignty,' said HiPEAC member David Atienza, associate vice-president of research centres and technology platforms at EPFL. 'Its sustainable approach aligns perfectly with the needs of EPFL's advanced computing platforms and will play a pivotal role in shaping the trajectory of future AI development at EPFL to consume fewer resources with the deployment of LLMs such as Apertus.'

Anyway Systems was recently chosen as one of six inaugural grantees of the Startup Launchpad AI Track – powered by UBS. The software is currently being tested in companies and administrations across Switzerland.

Announcement on EPFL website:
go.epfl.ch/d8e896
anyway.dev
linkedin.com/company/swarmtelect

Cloudberry announces Europe-focused semiconductor VC fund

Venture capital (VC) firm Cloudberry, which is based in Helsinki and London, has launched a fund of € 30 million, which will be dedicated to semiconductors, photonics, and advanced materials startups. The fund will invest at pre-seed and seed stages, focusing on technologies linked to compute, connectivity, sensing, and power. Cloudberry is backed by Finnish state-owned investment company Tesi and partners including Global Foundries and Radiant Opto-Electronics.

According to the press release announcing the launch, the team plans to invest in up to 20 companies across Europe, ranging from early research and development (R+D) spinouts to teams commercializing innovations within the fund's core focus areas.

'We see this as the start of a long-term effort to build Europe's semiconductor ecosystem,' said Veera Pietikäinen, founding partner at Cloudberry. 'Europe has world-class talent and deep technology capabilities in semiconductors and photonics, but has lacked investors who truly understand how to support and scale them. We are now building a specialist platform that helps these companies grow and strengthen Europe's technological sovereignty in the process.'

cloudberry.vc

Embedded IoT cybersecurity scaleup Exein raises € 170m in 2025

Exein, a company providing embedded internet-of-things (IoT) cybersecurity solutions, recently announced a successful €100 million funding round, bringing the total raised by the company in 2025 to €170 million. Founded in 2018 and headquartered in Rome, Exein also has offices in Germany, Taiwan, and the United States.

Exein's solutions embed advanced security directly into device software, leveraging edge artificial intelligence for real-time threat detection and response. Announcing the funding round, Exein stated that the new investment will support their mergers and acquisitions (M&A) strategy as the company continues scaling across new markets and launch their next-generation technology in 2026.

The round was led by Blue Cloud Ventures, joined by HV Capital, Intrepid Growth Partners, Geodesic Capital and the investment bank J.P. Morgan.

exein.io
linkedin.com/company/exein

NobodyWho raises € 2 million in pre-seed funding for on-device inference engine

Copenhagen-based startup NobodyWho recently announced that they have raised € 2 million in pre-seed funding to bring smaller artificial intelligence (AI) models directly onto users' devices. Founded by entrepreneur and artist Cecilie Waagner Falkenstrøm, along with co-founder and chief technology officer Asbjørn Olling, NobodyWho's open-source inference engine enables small language models (SLMs) to run locally on laptops and mobile phones, meaning that data does not have to leave the device.

Announcing the funding round on LinkedIn, Cecilie Waagner Falkenstrøm said: 'We began challenging the "bigger-is-better" logic of cloud AI and pushing toward another future – one where AI runs on-device, works offline, respects privacy and data sovereignty, and is far more cost-efficient and climate-aligned.'

The funding round was backed by Nordic early-stage investors PSV Tech and The Footprint Firm, with participation from Norrsken Evolve.

nobodywho.ooo

Wodan AI closes € 2 million pre-seed round for encrypted AI technology

Wodan AI, a Spanish–Belgian startup, recently announced a successful €2 million pre-seed funding round. According to the company website, Wodan AI enables secure, privacy-preserving artificial intelligence (AI) by converting standard models into encrypted runtimes using homomorphic encryption (HE). The platform combines advanced cryptography with containerized orchestration to keep data private, including during computation.

The funding round was led by Spanish venture capital firms JME Ventures, Swanlaab, and Adara Ventures, with additional backing from Belgium-based ScaleFund, the company announced this week.

According to a report in Tech Funding News, Wodan AI is relocating its global headquarters to Madrid and consolidating its research and development (R+D) operations in Spain. 'Spain has the technical talent required to develop critical sovereign AI technologies. With our headquarters, R+D centre and team growth in Madrid, we aim to consolidate a European hub of reference in private, secure AI designed for the most demanding sectors,' the report quotes Manuel Pérez Yllan, chief technology officer and co-founder of Wodan AI, as saying.

wodan.ai
techfundingnews.com/wodan-ai-2m-encrypted-ai-europe/

S2C, MachineWare and Andes launch RISC-V co-emulation solution

S2C, MachineWare, and Andes Technology recently announced a collaborative co-emulation solution designed to address the increasing complexity of RISC-V-based chip design. The solution integrates MachineWare's SIM-V virtual platform, S2C's Genesis Architect and Prodigy FPGA prototyping systems, and Andes' high-performance AX46MPV RISC-V CPU core, providing a unified environment for hardware and software co-verification.

Announcing the solution, MachineWare noted that as RISC-V designs move toward high-performance, multicore, and highly customized architectures, pre-silicon software development and system validation have become more challenging. The announcement states that this solution supports a 'shift-left' verification approach, allowing hardware and software teams to work in parallel, resulting in reduced development time and lower project risk.

Based in Aachen, MachineWare was launched with its high-speed functional RISC-V simulator, SIM-V, in 2022. Its founders form part of the HiPEAC network.

machineware.de
s2cinc.com
andestech.com

Launch of Sainet: Siena Artificial Intelligence Network

Launched at the end of 2025, Sainet is a not-for-profit initiative creating a community around artificial intelligence in Siena, Italy. Co-founded by Pietro Pianigiani, currently working on hardware at the Barcelona Zettascale Lab founded by Barcelona Supercomputing Center, Sainet is linked to the University of Siena and SAILab, the local artificial intelligence lab. Sainet aims to promote knowledge-sharing, startup acceleration and community building, with a particular focus on education and technology transfer.

sainet.it

New paper enables predictable reconfiguration in safety-critical systems

Giacomo Valente, University of L'Aquila

Researchers from the University of L'Aquila and Collins Aerospace have developed a new approach that enables embedded systems to meet strict size, weight, and power (SWaP) constraints while maintaining real-time high performance. Their work, titled 'Leveraging Traffic Injection and Quality-of-Service to Control the Reconfiguration Delay', focuses on systems based on runtime reconfigurable hardware and addresses one of their most persistent challenges: the variability of reconfiguration delay.

Many advanced systems use dynamic reconfiguration to swap hardware modules on the fly. This allows multiple accelerators to share the same field-programmable gate array (FPGA), helping to meet SWaP constraints. However, the time needed to load a new module, the reconfiguration delay, is notoriously unpredictable because it depends on interference from other components accessing shared resources such as memory or communication buses. This unpredictability significantly complicates the adoption of dynamic reconfiguration in safety-critical contexts, including aerospace systems. Existing solutions to bound this delay typically require additional hardware, intrusive architectural changes, or impose performance penalties on other tasks.

The new approach instead takes advantage of what modern systems-on-chip already offer. By accurately modelling the interference generated by concurrent tasks and exploiting built-in quality-of-service (QoS) mechanisms, the approach provides designers with a safe and tight upper bound on reconfiguration delay, without adding hardware or consuming extra system resources.

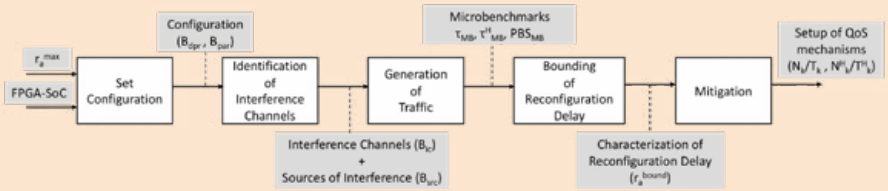
The researchers validated their solution on a Xilinx Zynq UltraScale+ platform through two extensive experiments. In uncontrolled conditions, reconfiguration delay was predicted to increase up to 6.3× compared to isolation. With the proposed QoS configuration, automatically generated by their tool, the delay remained strictly bounded to 1.8×, while competing tasks experienced only a 14% slowdown. This represents a substantial improvement over existing approaches.

By making runtime reconfiguration predictable using standard architectural features, this work strengthens timing assurance in safety-critical systems across domains such as aerospace, robotics, and autonomous vehicles.

FURTHER INFORMATION:

G. Valente, V. Muttillo, F. Federici, L. Pomante, T. Di Mascio, 'Leveraging Traffic Injection and Quality-of-Service to Control the Reconfiguration Delay', *Journal of Systems Architecture*, Volume 168, 2025, DOI: 10.1016/j.sysarc.2025.103570

doi.org/10.1016/j.sysarc.2025.103570



The five steps of the proposed approach

Jeronimo Castrillón awarded ERC Consolidator

HiPEAC member Jeronimo Castrillón Mazo, Chair of Compiler Construction at the Faculty of Computer Science at TUD Dresden University of Technology (TUD), has been awarded a European Research Council (ERC) Consolidator Grant. His project, COMpilers for ExTreme Heterogeneity (COMETH), which will run for five years and receive €2 million in funding, aims to develop a new generation of tools that will simplify the programming of increasingly complex computer systems.

As new technologies emerge, such as in-/near-memory computing or quantum computing, computers are set to combine many fundamentally different types of hardware. This makes programming considerably more difficult. The COMETH project is building models capable of understanding different computing paradigms and knowing how they can work together to solve a task, allowing complex programming code to be automatically translated into the detailed instructions required by computers.

bit.ly/TU_Dresden_COMETH

COMETH

New architecture developments discussed at REACH 2025 in London

Sarah Sennett, Institute of Engineering and Technology

'An architectural framework is for life, not just for Christmas!' So said Arm's Executive Vice President and Chief Architect, Richard Grisenthwaite, as he delivered the opening keynote at REACH 2025 – the Institution of Engineering and Technology's computer architecture conference, now in its second year.

Richard's talk looked at how computing, and Arm's technology, will develop in view of the rise of artificial intelligence (AI) and machine learning. He predicted that the future will include very specific AI, used to solve very specific problems, and that finely targeted systems will be more efficient than large data centres. AI has expedited the range of systems that can be developed; however, innovation will be determined by the trust in the systems that they run on.

AI was a common theme over the two days of the conference. Michaela Blott (AMD) – also the HiPEAC 2026 opening keynote speaker – noted that the evolution of AI is accelerating, and highlighted some of the challenges presented. Agility – and accelerated accelerator development – will be key to resolving the challenges, she said. In his talk, Jangwoo Kim (MangoBoost), explored how modern datacentre and server architectures are being reshaped by the increased use of AI. Smart network scaling will be the key, he said.

The topic of AI was continued in a panel discussion titled 'The Chip Architecture for Future AI – Does it Exist?', which included contributions from Sophie Wilson CBE (Broadcom), Karu Sankaralingam, (NVIDIA), Partha Maji (Microsoft), Atiq Bajwa, (Ampere Computing), Carlo Luschi, (Graphcore), and Wayne Luk (Imperial College London).

Panel session at REACH 2025

Sustainability in computer design was another important topic discussed. HiPEAC member Lieven Eeckhout (Ghent University) pointed out that we need to reduce device carbon emissions by 15.5% per device, per year, to meet the Paris Agreement for carbon reduction. However, the transition to green energy is not moving fast enough.

How will computer architecture develop over the next 25 years? Boris Grot (University of Edinburgh), a HiPEAC member and member of the REACH organizing committee, posed the question of what the sector will look like in 2050 in an interactive panel session. The panel's thoughts included:

- More customized processors for defined tasks – Richard Grisenthwaite (Arm)
- Lots of opportunity to rethink memory as a system that can process – Onur Mutlu (ETH Zürich)
- Generative AI is going to have a huge impact, including on the labour market – Karu Sankaralingam, (NVIDIA)
- Memory architectures will become more distributed. Lots more static scheduling, with associated power benefits – Chris Wilkerson (Intel)
- Data centres will grow exponentially to meet computing needs, providing 24/7 information and suggestions streams, personalized to the user's data – Adi Yoaz (Huawei)

Closing the conference, Hidetsugu Irie (University of Tokyo) gave the first talk in Europe on the STRAIGHT architecture . This approach retains the classical sequential programming model, while reducing per-instruction renaming overheads and opening new scaling paths.

In addition to a lively lecture room, REACH 2025 featured a poster session of the latest research in computer architecture and related technologies. The best poster, as voted for by the delegates, was awarded to HiPEAC member José Cano and his student Jude Haris (University of Glasgow), for their poster 'Accelerating AI at the Edge: The Power of Efficient HW/SW Co-Design'.

REACH 2026 will return to London in November 2026.
The call for posters will open in early 2026.
More information is available from spkl.io/6005AcYi5

Christian Pilato appointed chair of ACM SIGDA

HiPEAC member Christian Pilato (Politecnico di Milano) has been appointed Chair of ACM SIGDA (Special Interest Group on Design Automation). In a post announcing the appointment, Christian said: 'The design automation community is evolving rapidly, driven by AI-powered design, heterogeneous computing, and the growth of open-source hardware and EDA ecosystems. Now more than ever, collaboration and inclusiveness are key to advancing our field.

'As Chair, my focus will be on broadening SIGDA's global membership, strengthening engagement through active member contributions, bridging academia, industry, and open-source communities, and ensuring our events remain sustainable and accessible. I look forward to collaborating with colleagues and partners worldwide.'

For further information on how to get involved, visit the ACM SIGDA webpage:
acm.org/special-interest-groups/sigs/sigda

Koen De Bosschere admitted to Royal Flemish Academy of Belgium

HiPEAC coordinator Koen De Bosschere was formally admitted to the Royal Flemish Academy of Belgium on 20 December 2025 for his scientific and societal impact in the field of computer science. His inaugural speech in the Class of Technical Sciences was entitled 'From classroom to boardroom: the entrepreneurial ecosystem in Ghent'.

Koen De Bosschere (second from left) admitted to the Royal Flemish Academy of Belgium

ACM Gordon Bell Prize for Climate Modelling awarded to 26-person team

Multiple European institutions represented, including HiPEAC member Torsten Hoefler's team at ETH Zürich.

The 2025 ACM Gordon Bell Prize for Climate Modelling was presented to a 26-member team for developing a full Earth simulation at 1 km resolution during the Supercomputing conference in St. Louis, Missouri, United States, in November 2025. The award honours innovative contributions to parallel computing toward solving the global climate crisis.

Announcing the award, ACM noted that, while digital simulations of the Earth are among the most effective tools to understand climate change, simulating how human activity influences the climate is an extraordinarily difficult challenge. This is due to the huge number of variables that need to be taken into consideration – such as the cycles of water, energy, and carbon, how those factors relate to each other, and how diverse physical, biological, and chemical processes interact over space and time.

The 26-member team was the first to develop a full Earth simulation at a resolution of 1 km, using compute power provided by the Alps and JUPITER supercomputers, based in Switzerland and Germany, respectively.

The team comprised researchers from the following organizations: ETH Zürich (including HiPEAC member Torsten Hoefler), Deutsches Klimarechenzentrum, NVIDIA, Max Planck Institute for Meteorology, University of Hamburg, Forschungszentrum Jülich and the Swiss National Supercomputing Centre.

awards.acm.org/bell-climate

Michael O'Boyle, Michele Magno and Thomas Ernst named IEEE Fellows



The January 2026 cohort of IEEE Fellows included two HiPEAC members. The IEEE Fellow grade is conferred by the IEEE Board of Directors to individuals with outstanding qualifications and significant contributions to engineering, science, and technology.

HiPEAC founding member Michael O'Boyle (University of Edinburgh) was selected for his contributions to machine learning–based compilation and parallelization. Meanwhile, HiPEAC member Michele Magno (ETH Zürich) was recognized for his contributions to low-power wake-up radios and energy harvesting for smart sensors.

Thomas Ernst (CEA Leti), a contributor to several editions of the HiPEAC Vision, was also named IEEE Fellow for his contributions to 3D-stacked gate-all-around nanosheet and nanowire transistors.

Further information is available in the announcement by the IEEE Computer Society:
computer.org/press-room/2026-class-fellows

Videos of lectures by Michael O'Boyle and Michele Magno are available in the ACACES 2023 playlist:
bit.ly/ACACES23_videos

A HiPEAC Vision webinar featuring Thomas Ernst is available in the HiPEAC webinars playlist:
bit.ly/HiPEAC_webinar_videos

Ewa Deelman receives IEEE Sidney Fernbach Memorial Award

In November, DISCOVER-US member Ewa Deelman (University of Southern California) was named recipient of the IEEE Sidney Fernbach Memorial Award, given for outstanding contributions in the application of high-performance computers using innovative approaches. The award was made in recognition of her work in workflow-based parallel and distributed computing. Ewa leads the design and development of the Pegasus Workflow Management System, a widely adopted platform that enables large-scale, reproducible, and open computational science across domains. Her research spans workflow optimization, resource provisioning, data management, provenance capture, and the use of cloud and hybrid platforms for scientific discovery.

Timothy Jones inducted into MICRO Hall of Fame

In October, HiPEAC member Timothy Jones was inducted into the ACM/IEEE MICRO Hall of Fame, which recognizes authors who have had eight or more papers published at the IEEE/ACM International Symposium on Microarchitecture (MICRO).

The MICRO Hall of Fame includes numerous HiPEAC members; additions in recent years include Lieven Eeckhout (Ghent University) and Stefanos Kaxiras (Uppsala University).

MICRO Test of Time Award for Gabriel Loh and co-authors

In October, HiPEAC associate member Gabriel Loh and co-authors were awarded the MICRO Test of Time Award for their 2006 paper 'Die Stacking (3D) Microarchitecture'. The award recognizes the most influential papers published in prior sessions of the IEEE/ACM International Symposium on Microarchitecture (MICRO). The paper was selected 'for pioneering the microarchitectural evaluation of 3D die stacking, demonstrating its potential for substantial performance/power gains and shaping future 3D die-stacking research'.

Ingrid Verbauwhede awarded FWO Excellence Prize

In a formal ceremony in November, HiPEAC member Ingrid Verbauwhede (KU Leuven / Belfort) was awarded the 2025 FWO Excellence Prize, one of the highest scientific honours in Belgium. The prestigious award, which is awarded every five years and has a prize of €100,000, was made in recognition of Ingrid's pioneering work in hardware and semiconductor security. Ingrid is the recipient of multiple international honours, including two ERC Advanced Grants, the IEEE Don Pederson Award (2023), and the EDAA Achievement Award (2024).

bit.ly/COSIC_Ingrid_Verbauwhede_FWO

Isak Karabegović wins Lifetime Achievement Award Bosnia and Herzegovina

In December, HiPEAC member Isak Karabegović received the Federal Lifetime Achievement Award in Science for 2025, awarded by the Government of the Federation of Bosnia and Herzegovina, for his longstanding and outstanding contribution to the development of science, higher education and international cooperation.

Accepting the award, Isak thanked his collaborators and added: 'I would like to send a message to young people not to be afraid of knowledge, to take knowledge into their own hands, to be innovative and creative because our future depends on their research and work.'

fmon.gov.ba/Obavjest/Pregled/1597

Dates for your diary

HiPEAC webinars
Check the HiPEAC website to keep up to date on forthcoming dates
hipeac.net/webinars

DATE 2026: Design Automation and Test in Europe Conference
20-22 April 2026, Verona, Italy
date-conference.com

EuroSys 2026: European Conference on Computer Systems
27-30 April 2026, Edinburgh, UK
2026.eurosys.org

ISC High Performance
22-26 May 2026, Hamburg, Germany
isc-hpc.com

Photo credit: ItalyDrones | stock.adobe.com

ACACES 2026: Advanced Computing Architecture and Compilation for High-performance Embedded Systems Summer School
12-18 July 2026, Fiuggi
HiPEAC summer school
Applications open March 2026
hipeac.net/acaces/2026

ACM Summer School on HPC Computer Architectures for AI and Dedicated Applications
20-24 July 2026, Barcelona
Deadline for applications: 15 February
europe.acm.org/seasonal-schools/hpc/2026

NorCAS 2026: IEEE Nordic Circuits and Systems Conference
27-28 October 2026, Tampere, Finland
events.tuni.fi/norca
Special session proposals deadline: 15 June 2026
Paper submission deadline: 15 August 2026
Further information: Jari Nurmi, Tampere University
norcas@tuni.fi

Europe's technological dependencies – and their impact on the ability of the European Union (EU) to enforce its own laws – are under scrutiny. Launched in September 2024, EuroStack is an industrial-strategy project and not-for-profit foundation led by the European tech industry. HiPEAC spoke to EuroStack Chair Cristina Caffarra and Board Member Stéfane Fermigier to find out what constitutes 'digital sovereignty', why Europe should take urgent action to rebuild its technology sector, and what EuroStack is doing to address this issue.

'By not controlling our critical digital infrastructure, we have ceded control over our economic future and our ability to uphold our own laws and values'

While Europe's position in the information technology sector has been waning for decades, recent events have resulted in an ever-louder chorus of voices concerned about the continent's ability to control its digital destiny.

There are several pressing reasons behind this concern. First, a weak technology sector in Europe has critical consequences on the economy. 'For years, Europe has outsourced its digital foundations, becoming what many call a "digital colony". This leads to a recurrent, massive outflow of resources: a recent study estimates that European businesses' reliance on US-based cloud and software services benefits the US economy to the tune of €264 billion annually – a figure comparable to our total energy-import bill, or which represents three million jobs in Europe,' says Stéfane Fermigier, founder and chief executive of the open-source software vendor Abilian and a co-founder of the EuroStack Initiative Foundation.

Cristina Caffarra speaking at the European Digital Sovereignty Summit, November 2025

The implications are grave, according to the economist and EuroStack Chair Cristina Caffarra, who agrees with the conclusions of Mario Draghi's 2024 report, *The future of European competitiveness*. 'We are an exporting economy which is being clobbered by an active trade war with China, and loss of competitiveness with both China and the United States (US). While productivity growth was similar 20 years ago between the US and Europe, there has been a major decoupling as the US pulled ahead, driven by its tech sector. Digital is – and artificial intelligence (AI) will be – an all-purpose technology. If you are behind with adoption and diffusion of these technologies, you lose out in every sector of the economy. So to me this is not remotely an anti-US campaign: US suppliers have filled a vacuum we have left for them to fill, kudos. But if we don't focus on building our own capabilities in this space, we will fall further and further behind.'

Second, recent geopolitical events have thrown Europe's vulnerability, through its reliance on imported technologies, into sharp relief. 'Global instability means that Europe can no longer take for granted the reliability of its traditional partners. Our dependencies are no longer simply a commercial issue, but a critical strategic vulnerability,' says Stéfane. There is also growing awareness of the legal and security risks that reliance on extraterritorial technology entails: 'European companies' data and intellectual property are exposed to extraterritorial laws like the US CLOUD Act and FISA, which can compel disclosure to a foreign government without regard for European privacy and business secrets standards. This is a direct threat to our competitiveness and security,' adds Stéfane.

Defining 'digital sovereignty'

For Stéfane, digital sovereignty is 'the ability of individuals, organizations, and governments to independently control, manage, and use their digital infrastructure, data, and

technologies, free from undue external systems in a self-determined manner'. In addition to the governance of critical infrastructure, this also entails 'the ability to make autonomous decisions about the use and management of applications and technologies, safeguarding privacy, data integrity, and digital independence'.

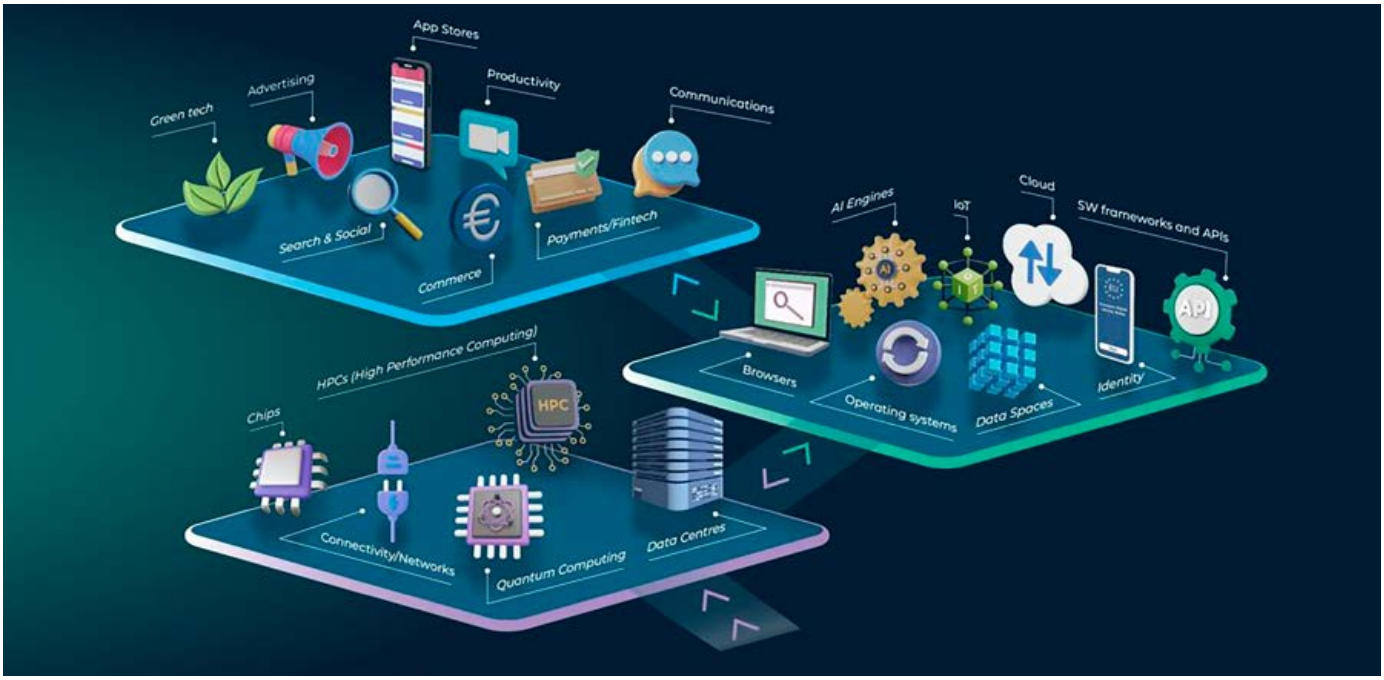
However, as Koen De Bosschere notes in his introduction to this magazine edition, the power to govern the digital space is dependent to a large extent on the technical capacity to implement European laws – i.e. reducing dependency on technologies from other regions. For Cristina, attempting to tame technology companies through regulation is a failed approach; the emphasis must shift to building up the technology sector in Europe. 'We are now a digital colony, because while we were busy regulating Google Search, Apple's App Store, Meta's social networks, companies were building data centres, laying cables, appropriating the data – which effectively means that the digital infrastructure we rely on today is not ours.'

Cristina fiercely rejects the negative attitude exhibited by some which suggests that Europe has missed the boat on technological development – noting that Europe has 'incredible capabilities, incredible assets, some of the best engineers', with dynamic ecosystems – even for chips – in places like Dresden, Eindhoven, and Grenoble – but laments that Europeans are 'complacent beyond words'.

Sovereign European Provider: The EuroStack framework

EuroStack, which has evolved from a broad, informal coalition into a non-profit association, aims to provide a pragmatic response to this situation. To avoid vague definitions and what Stéfane refers to as 'sovereignty washing', where 'non-European providers use marketing and superficial corporate structures to create an illusion of compliance', EuroStack has translated the concept of 'digital sovereignty' into a concrete, legally robust framework built on five pillars, as set out below:

1. **Jurisdiction and governance – accountability to EU law**: the ultimate parent entity of the provider must be headquartered and legally incorporated in the European Territory (EU, EEA, EFTA), and must be free from decisive non-EU control.
2. **Technical sovereignty – freedom from technological lock-in**: an antidote to the scenario in which Europeans' data is in Europe but trapped within proprietary foreign technology. 'We mandate that services be built predominantly on open-source software and implement open standards, guaranteeing data portability, interoperability, and even operational reversibility,' explains Stéfane.
3. **Operational sovereignty – control over the entire service-delivery chain**: 'It's not enough for data to reside in the EU; it must also be managed from the EU. We require that the entire physical infrastructure – and, critically, the operational control plane – are located and operated from within Europe,' says Stéfane. 'All personnel with privileged access must be European residents, employed by a European entity, and working from within the EU.' ▶



EuroStack graphic showing the digital ecosystem

4. **Data sovereignty – verifiable protection of all data**: exclusive EU residency for all data, including metadata and backups, with additional credits for verifiable technical measures that make it impossible for any unauthorized party – including the provider itself – to access unencrypted data.
5. **Economic sovereignty – a net contributor to the European economy**: 'This ensures that public money is used to build our own industrial base. We require that the majority of a provider's research and development (R+D) expenditure and personnel for the core technology are located in Europe, and that their business model is fair and transparent, free from punitive tactics like exorbitant data egress fees,' Stéfane says.

From ideas to action: EuroStack's priorities for digital sovereignty

With all this in mind, how should European actors – in both the public and private sectors – move forward to reverse the trend towards digital subjection?

EuroStack has defined three priorities for action, as follows:

1. **'Buy European': Drive demand through strategic procurement**
   'We must amend EU public-procurement directives to mandate that a significant and growing share of public IT spending is directed towards sovereign European solutions,' says Stéfane. 'This isn't protectionism; it's a strategic investment in our own industrial base. It creates the vital customer revenue that our companies need to scale and compete.' Cristina agrees unequivocally: 'Every other region in the world says, "buy from yourself". The default in Europe is "buy from everywhere". It might have made sense originally because we don't want individual states in Europe to just buy from themselves – but now it is being weaponized against us by the hyperscalers,' she says. 'The outflow of public procurement to the US is around €80 billion a year, so if we reserved 20–30% to European suppliers via strategic public procurement rules, the "shot in the arm" would be sizeable.'

   In addition to specifying European providers, EuroStack also wants procurement processes to be simplified to avoid disadvantaging small and medium enterprises (SMEs).

2. **'Sell European': Organize and strengthen European supply**
   The success of the US hyperscalers is not accidental: they offer user-friendly, effective, vertical solutions. Cristina emphasizes that customers are understandably partial to integrated solutions (convenience, performance and reliability) so European suppliers will realistically need to develop comparable bundles, by collaborating in various ways. 'What is the European alternative to the stack that goes from chips, to compute, to software, to AI that I can use instead of Amazon or Microsoft?' she asks. 'The truth is that we don't have one. We have lots of bits and pieces, but they need to be put together. This cannot be the responsibility of the customer; it needs to be the responsibility of the supplier. Improve your product. Join forces with someone. Create a portfolio. This is what is needed, and we are seeing that it is beginning to happen. EuroStack suppliers are clubbing together to meet demand from industrial customers.'

   Of course, following this advice is more challenging for the patchwork of European SMEs than it might be for a single large enterprise. A central plank of EuroStack's strategy is to make it easier for buyers to find trustworthy European solutions. 'We are creating a vetted, dynamic catalogue of European digital capabilities, performing strategic gap analyses, and aggregating offers through open standards. The goal is to break down siloes and allow our innovative SMEs to compete collectively,' explains Stéfane.

Cristina appearing on Bloomberg Daybreak Europe

Bloomberg Tech London debate, 'Is Europe too late to compete in the chips war?'

'But a catalogue is not enough. We are also facilitating contacts between demand (customers) and suppliers who can create work together to create collaborative integrated solutions to respond to customers' needs,' adds Cristina. 'EuroStack is aiming to play an active role in creating these solutions and architectures – "matching" demand and supply. The success of European tech depends entirely on customers moving their purchases away from current patterns and starting to "buy European" for real. The public sector is not enough: we need to get customers shifting.'

3. **'Fund European': Mobilize strategic capital**

'We must be smarter about how we use public and private funds,' says Stéfane. 'An immediate action should be to stop the subsidization of non-European technologies and redirect funds to where they can have most impact. We also proposed a dedicated "EuroStack Fund of Funds" to de-risk adoption, support foundational open-source projects, and help our tech companies scale up, with an investment logic that accounts for a "sovereignty dividend".'

In terms of R+D, Stéfane advocates for a strategic approach: 'Key R+D priorities must include genuinely sovereign cloud and edge infrastructure, trustworthy artificial intelligence, next-generation cybersecurity, and continued investment in foundational open-source components that underpin our entire infrastructure,' he says.

Stéfane Fermigier speaking at Open Source Experience, December 2025
Photo credit: William Jezequel

The role of open source

According to Stéfane, the role of open source in European digital sovereignty is transformative. 'It is simultaneously the technical key to unlocking technological autonomy, the foundation upon which we can build a collaborative and innovative ecosystem, and the central instrument in a pragmatic policy strategy to reclaim our digital future. It is, quite simply, the European way to build digital sovereignty.'

To understand the role of open source, Stéfane refers to the definition laid out in a recent report by the European Alliance for Industrial Data, Edge and Cloud [see 'Further reading', overleaf]: 'Digital sovereignty = data sovereignty + technological autonomy. While data sovereignty – controlling where our data is stored – is important, it is insufficient. Without the ability to control the underlying software and hardware that processes this data, we risk creating "sovereign prisons" where our data is in Europe but we are completely locked into foreign technology,' adds Stéfane.

Open source offers a way out of vendor lock-in, says Stéfane, as it provides full control over the code, 'giving us the ability to maintain our systems, innovate independently, and switch providers'. The ability to audit the source code also allows for collective, independent verification, ensuring there are no hidden backdoors or vulnerabilities. It also allows Europe to build on its strengths by enabling companies, research institutions and public bodies to pool resources and collaborate on cutting-edge technologies that no single entity could build alone, he adds.

However, open source must be approached strategically if Europe is to fully benefit from it, says Stéfane, citing five critical gaps currently preventing Europe from fully leveraging open source. Standards and interoperability are required to escape the fragmented market where dominant players create proprietary ecosystems disguised as open. Financial viability is important given the lack of long-term funding for European solutions, as is market visibility, which is also currently lacking. The shortage of expertise in sovereign open-source technologies needs to be addressed, as does the governance and influence of open-source projects: 'While Europe contributes a vast amount of code globally, we lack proportional influence over the strategic direction of key projects, which are often governed by foundations outside the EU.'

Next steps – and how HiPEAC can help

In line with EuroStack's pragmatic outlook, the next steps for the EuroStack Initiative Foundation will be building operational capacity to deliver its vision, Stéfane says. 'Over the coming months, we will be implementing our governance model, defining a formal membership program to bring hundreds of aligned companies into the fold, and expanding our working groups focused on concrete outcomes in integration, procurement, and communication. Our focus is on tangible outcomes: interoperable technology, reliable procurement frameworks, and stronger ties between European innovators,' he adds.

HiPEAC can play a critical role in contributing to the full-stack endeavour – from silicon to software – required for technological sovereignty, according to Stéfane. 'EuroStack's role is to create the market demand and industrial strategy for sovereign technologies. The HiPEAC community's role is to deliver the foundational research and development that makes these technologies possible. We see this as a necessary partnership to build a competitive European digital sector.'

In particular, Stéfane highlights three main areas where HiPEAC could contribute:

1. **Ensure a viable, sovereign hardware-software stack**
   'Initiatives like the European Processor Initiative are fundamental,' says Stéfane. 'What's needed from our point of view is to ensure the output of such projects is not only technically excellent but also commercially viable and designed for integration. We need a collaborative approach to co-design, ensuring that Europe develops a coherent and optimized stack where sovereign software can run efficiently and securely on the next generation of European processors.'

2. **Inform strategic and technical roadmaps**
   'We need the expertise within the HiPEAC community and the analysis from resources like the HiPEAC Vision to help us identify critical technological dependencies and strategic opportunities. This will directly inform our gap analyses and help guide the "Fund European" pillar of EuroStack to make targeted, high-impact investments,' says Stéfane.

3. **Develop sovereignty-oriented talent**
   'We need the HiPEAC community's help to shape programmes that produce engineers, architects and researchers with a full-stack, sovereignty-oriented, and open-source / open-hardware mindset,' says Stéfane. 'This means cultivating talent that understands not just deep technical challenges, but also the requirements for building secure, open, transparent and interoperable systems.'

Cristina speaking at IAPP Data Protection

A new horizon for European technology

Although sobering, the reckoning with Europe's technological dependency is finally provoking action. 'Our leaders are starting to wake up to the fact that by not controlling our critical digital infrastructure: we have ceded control over our economic future and our ability to uphold our own laws and values,' says Stéfane. 'Concrete action is needed now to reclaim our autonomy, and we believe it has to come from a multifaceted approach.'

The end goal for EuroStack is not to answer all of Europe's needs with homegrown technology, according to Cristina. 'EuroStack is not about autarchy protectionism: again, we are not anti-American and the goal is not to replace American suppliers,' she clarifies. 'But at the moment we've shrunk to 10-20% of the market in several segments – can we reach 30-40% of our own market for digital technologies?'

FURTHER READING:

Deploying the EuroStack: What's needed now
EuroStack, 2025
eurostack.eu/the-white-paper

Open source way to EU digital sovereignty & competitiveness
European Alliance for Industrial Data, Edge and Cloud, 2025
bit.ly/EAIDEC_OpenSource_2025

HiPEAC partner Stanisław Krzyżanowski (CloudFerro) explains how the European Union (EU) is addressing cloud sovereignty.

Encoding cloud sovereignty

How Europe is getting serious about homegrown cloud

In the early 2010s, Europe and the rest of the world experienced a cloud explosion. Microsoft, Google, IBM and others joined Amazon and launched cloud service platforms. Simultaneously, NASA and Rackspace initiated OpenStack, which remains the most popular open-source cloud software to this day.

The key promise of cloud was speed, efficiency and scale. The idea was simple: let engineers create value instead of maintaining servers. Cloud became a utility – the same as internet access, water or electricity. The provider of the utility was not a concern, especially since the providers were primarily based in the United States. This attitude reflected strategic relations and the zeitgeist.

The dominance of hyperscalers – an industry term for Amazon Web Services, Microsoft, and Google that underscores their immense scale – in the European cloud market stems from a combination of factors. Their early entry, deep access to capital, operational agility, and readiness to embrace risk provided a significant competitive edge. Over the past decade, these three companies swiftly captured market share, becoming virtually synonymous with the concept of 'public cloud'.

However, their influence extends far beyond cloud infrastructure alone. As hyperscalers expanded, they leveraged their scale and technological capabilities to integrate vertically and horizontally across the digital value chain. For example, Google evolved into the world's largest digital advertising exchange, while other hyperscalers diversified into platforms, productivity tools, and critical digital services. This has allowed big tech companies to entrench themselves not just as cloud providers, but as indispensable providers of essential services across multiple facets of the economy, amplifying their presence and deepening their penetration throughout society.

National and European representatives at the 2025 European Digital Sovereignty Summit ©BMDS/Woithe

This combination of the all-encompassing role of Big Tech and shifting geopolitical dynamics sparked debate and concern. Europe's pivot began when cloud stopped being 'someone else's IT' and became the control layer for essential services. Legal uncertainty around cross-border access, the everyday costs of lock-in, and geopolitical tensions reframed the question. It became less about compliance paperwork and more about agency: who can reach the systems, who can change them, and how quickly Europe can exit.

That is why Brussels is encoding sovereignty – ultimately into regulations, first into procurement. In October 2025 the European Commission launched a €180 million 'sovereign cloud' competition under the Cloud III Dynamic Purchasing System and attached a Cloud Sovereignty Framework to it. The framework breaks sovereignty into eight objectives – from strategic anchoring and legal jurisdiction to operational control, supplychain transparency, technological openness, security, environmental considerations and EUlaw compliance – and converts those dimensions into a quantified 'sovereignty score' used in award evaluation.

At the European Digital Sovereignty Summit in Berlin on 18 November 2025, the same logic moved from tender documents to political narrative. Cloud and artificial intelligence (AI) were framed as competitiveness and resilience issues, and leaders argued openly for reducing dependencies through European infrastructure choices.

Here is a bigger story. Once sovereignty becomes a repeatable method – score, certify, audit, and preserve the right to switch – it spreads beyond cloud. Space shows the trajectory. The EU's Space Strategy for Security and Defence explicitly targets technological sovereignty by reducing strategic dependencies, and IRIS² is presented as secure connectivity supporting Europe's autonomy and digital sovereignty. Even access to orbit is now treated in similar terms, with the EU emphasizing European launchers such as Ariane 6 and Vega C for Galileo and Copernicus infrastructure.

bit.ly/Cloud_Sovereignty_Framework

Alberto P. Martí is vice president of open-source innovation at OpenNebula Systems and the chair of the Industry Facilitation Group of the Important Project of Common European Interest on Next Generation Cloud Infrastructure and Services (IPCEI CIS). We asked him how he defines digital sovereignty, how Europe should approach this issue, and how the IPCEI-CIS seeks to build out the technology ecosystem in Europe.

'Digital sovereignty has profound implications for the way in which our democracies operate'

What are the pitfalls of relying on technological infrastructure from other parts of the world?

Many European organizations need to use technological infrastructure that spans different continents. The cloud-edge continuum is also a geographical continuum that has to be able to offer infrastructure resources on demand wherever they are required. However, we need to offer industrial users a secure way to use resources from different cloud and edge providers without having to rely on proprietary platforms or fall into critical dependencies or vendor lock-in situations – like the ones we are currently witnessing with hyperscalers. That's where strategic autonomy in the digital sector comes in.

How do you define 'digital sovereignty'?

As we are discovering lately, digital sovereignty involves many different aspects in our daily lives, and it is a concept that has profound implications for the way in which our democracies and institutions operate. Technology, in that sense, is just a means – a means that can be used to exert influence and coercion on others, or a means to make sure one can make decisions in an autonomous manner, with no fear.

The main challenge that we have in Europe right now when it comes to digital sovereignty is the need to, once and for all, tackle the dependency complex that we suffer collectively. For decades, Europeans have assimilated a mentality that undermines our own capacities and makes us assume that the best technologies and digital services come from across the Atlantic. We have turned into passive and docile consumers of products developed outside Europe, which has led to many people in Europe displaying now a self-defeatist attitude about our chances to stand up when other countries start exploiting the EU's strategic dependencies in order to impose their political will on us.

What should Europe prioritize in its steps towards digital sovereignty?

Europe is in a unique position to lead a global resistance movement against Big Tech. We should prioritize the establishment of strategic alliances with other democratic partners around the world, a multi-national community of countries willing to stand together against the many pressures from the US and China to impose their respective technological stacks and commercial terms. One could see this collective fight for digital sovereignty as a sort of modern-day, global decolonization movement, a process in which every country should be able to contribute their best resources to create alternative infrastructures, products and economic models for the digital world. Europe, in that sense, has a lot to learn from countries such as India and Brazil, but it also has a lot to offer, although it might require a new approach to our traditional economic relationship with partners in the Global South.

Alberto moderating a panel with EU Member State representatives at the NexusForum2025 Summit

What role can open-source technologies play in enabling digital sovereignty?

Open-source is the only way to implement an innovation model that is inclusive and flexible enough to accommodate active collaborations from many different partners, including among competitors. Unfortunately, open-source has also suffered a process of gradual colonization by Big Tech, and we now face a situation where European companies willing to work together on

the technologies that they need have to rely on open-source entities and development infrastructures controlled by non-EU corporations. GitHub, which was purchased by Microsoft some years ago, is a perfect example of that sad reality. The first priority for Europe right now should be to regain control over the open-source 'means of production', and make sure that European open-source technology providers receive the support they deserve for making sure that EU alternatives can keep competing in the market.

How does the IPCEI-CIS promote sovereign technological development in Europe?

The IPCEI-CIS offers a unique opportunity to promote technological sovereignty in Europe because it is the first time that the European Union has devoted €3 billion for companies from across the continent to develop open-source software together. Cloud and edge computing are critical enabling technologies for artificial intelligence (AI), simply because we cannot expect to have sovereign AI in Europe if most of the underlying infrastructure is controlled by non-EU corporations.

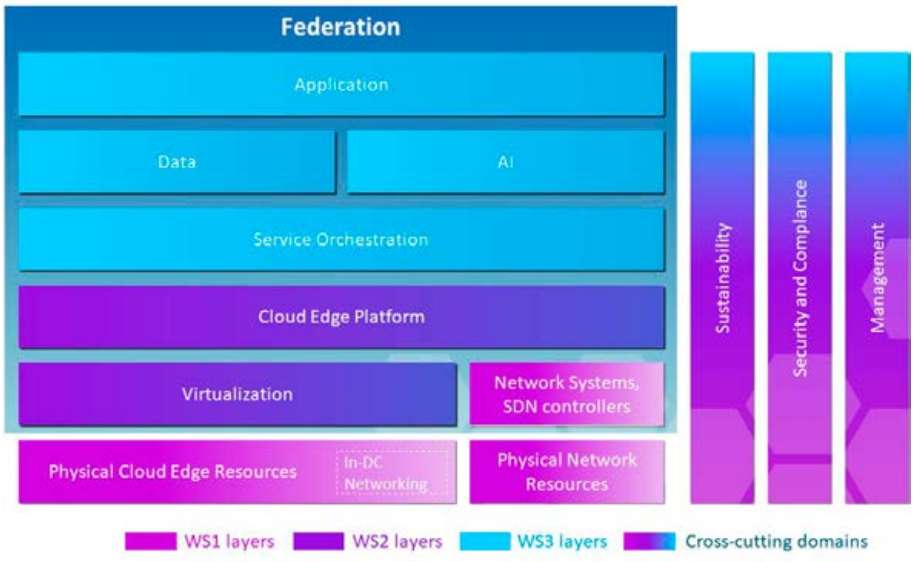
The IPCEI Cloud, and the upcoming IPCEIs on AI and Edge Computing, will consolidate an ecosystem of European technology vendors and cloud providers that, finally, will be offering an alternative stack of open-source solutions that are interoperable and adaptable to different sectors. One example is the Fact8ra.AI integration pilot, which is creating a European stack for federating AI factories across the EU, aggregating graphics processing unit (GPU) servers from high-performance computing (HPC) facilities and across the cloud-telco continuum.

How can the HiPEAC community get involved in this effort?

We need all hands on deck, industry and academia. Through projects like NexusForum.EU, we are trying to bridge the gap between EU-funded research projects that are exploring new technological solutions and those European industrial partners that are willing to incorporate them into new, innovative products. The HiPEAC community is a key actor in this process, and their potential to contribute to the adoption and improvement of sovereign technological solutions in Europe is huge. We will work together to make sure that there is a gradual alignment between projects at different technology readiness levels (TRLs) and help these collaborations to bring real value to the European open-source technologies that industry is adopting.

What would you say to people who feel that pursuing digital sovereignty is not worth the effort, as it is impossible to 'catch up' with other world regions?

I don't think it is a question of 'catching up' with other regions; it is about reclaiming the right to decide what kind of digital world we want. Right now, these decisions are made for us by a handful of megalomaniac billionaires living thousands of kilometres away, with zero empathy for the millions of people who use their platforms and products. The struggle for digital sovereignty is not only about Europe having its own data centres; it is about bringing democracy into an area of our economy that has become vital for many counties and private companies, but also for individual citizens. To those people who feel overwhelmed by the challenge I would ask: yes, it is a hard task, but what's the alternative if we don't even try?



Layers and domains of the IPCEI-CIS Reference Architecture

FURTHER READING:

IPCEI-CIS: Important Project of Common European Interest on Next Generation Cloud Infrastructure and Services
bit.ly/IPCEI_CIS

8ra initiative
8ra.com

Fact8ra Federated AI Factory
fact8ra.ai

NexusForum
nexusforum.eu

Advancing Europe's digital sovereignty

The role of EPI, EUPEX, EUPILOT, and DARE

Since launching the European High Performance Computing Joint Undertaking (EuroHPC JU) in 2018, Europe has made significant strides towards European technologies for high-performance computing (HPC) systems. In this article, Katarina Vukušić (University of Zagreb) and Romana Konjevod (Barcelona Supercomputing Center) show how key projects financed by the EuroHPC JU have laid the technological foundations for a thriving European HPC ecosystem, and what the future holds for HPC technology development in Europe.

Europe's push for digital sovereignty has become a defining strategic priority. According to the conclusions of the European Council from 18 April 2024, 'the European Council will ensure an integrated approach across all policy areas to increase productivity and sustainable and inclusive growth throughout Europe, build a robust, innovative and resilient economy and advance Europe's unique social and economic model that will boost the Union's competitive transition towards digital sovereignty in a self-determined and open manner'. This ambition goes beyond political statements; it reflects a fundamental need for Europe to control the technologies, data, and infrastructures that underpin modern society.

In a rapidly changing geopolitical environment, Europe is increasingly aware of the need to strengthen control over key digital technologies. Building the capacity to shape its own hardware, software, and innovation pathways has become an important strategic priority. This means strengthening collaboration across countries, creating integrated technology ecosystems, and investing in long-term, forward-looking design rather than patchwork solutions.

The global race toward exascale computing started years ago, led by the United States and China. Recognizing the strategic stakes, Europe established the EuroHPC Joint Undertaking to accelerate the deployment of world-class supercomputers built with European technologies and intellectual property, systems that will ultimately empower European research and development (R+D) and industrial development.

Europe's strategic response

To help build these kinds of capacities, the European Commission (EC) started a special framework partnership agreement in 2017, out of which the **European Processor Initiative (EPI)** was born. The EC's aim was to support the creation of a world-class European high-performance computing (HPC) and big data ecosystem built on two exascale computing machines, which would rank among the top three supercomputers in the world.

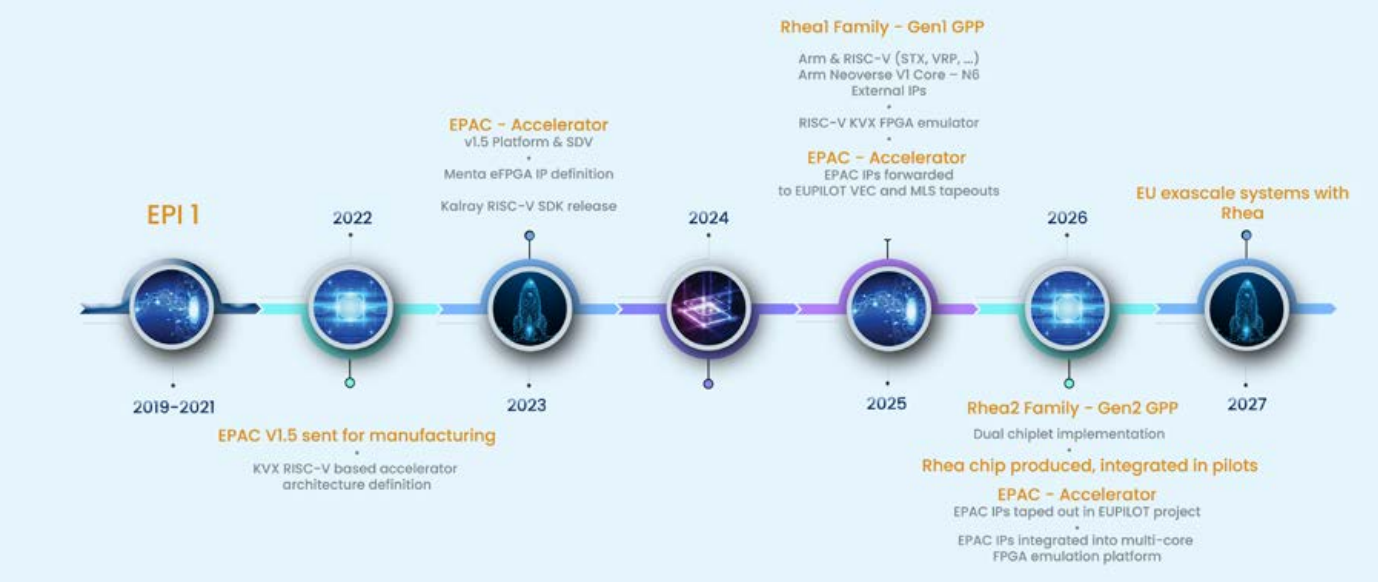
In 2018, the project started its initial stage with the goal of designing and implementing a roadmap for a new family of low-power European processors for extreme scale computing, high-performance big data, and a range of emerging applications. It continued in a second stage started in 2022 which was complemented by two pilot projects, EUPEX and EUPILOT, funded by the newly established EuroHPC Joint Undertaking.

As EPI had two main pillars – an Arm processor family developed by a company born out of EPI (SiPearl), and a RISC-V-based European processor (EPAC) – the two pilot projects were established to help carry out the basis for the future of digital sovereignty along these two pillars.

EPI and the pilot projects – EUPEX and EUPILOT

EUPEX's key objectives are to: co-design a modular exascale-pilot system; build and deploy a pilot hardware-and-software platform integrating European technology; demonstrate the readiness and the scalability of the pilot technology in general and the project's signature 'Modular Supercomputing Architecture' (MSA) in particular, for exascale computing; and, perhaps most notably, prepare applications and European users to efficiently exploit future exascale machines.

EUPILOT was created to deliver the first all-European open-source and open-standards-based software-and-hardware integrated systems by creating accelerators designed and implemented in Europe. It has reached a major milestone with the successful tape-out of its VEC and MLS accelerator chips, one targeted at HPC workloads and the other towards artificial intelligence (AI) workloads.



European Processor Initiative timeline

As EPI draws to an end, SiPearl's first processor, Rhea1, is currently being manufactured at TSMC, and it will be available for sampling in 2026. It will equip JUPITER, the first of the two exascale systems planned in Europe. Rhea1 has 80 arm® Neoverse™ cores, two Scalable Vector Extension (SVE) units of 256 bits per core, built-in high-bandwidth memory (HBM) with four stacks of HBM, four DDR5 interfaces supporting two DIMMs per channel (2DPC) and 104 lanes of PCIe Gen5 interface.

The future is DARE

To take over the baton in the open-hardware space, which is crucial for the aspects of sovereignty mentioned above, we reach the latest development in EuroHPC JU's portfolio: **DARE (Digital Autonomy with RISC-V in Europe)**, which is a groundbreaking initiative driving Europe's independence in high-performance computing and AI.

DARE sets out to lay the technological foundations for European digital autonomy in HPC and AI. By combining open RISC-V architectures, chiplet technologies, and a co-designed software ecosystem, DARE aims to deliver working prototypes, shape the European Union's HPC roadmap, and boost Europe's ability to build and sustain its own supercomputing value chain.

DARE marks the next step in Europe's long-term HPC roadmap, ensuring that the technologies developed across EPI, EUPEX, and EUPILOT evolve into a sustainable, fully European supercomputing value chain.

Conclusion

Thanks to strategic R+D investment, Europe has significantly strengthened its capacity to build next-generation advanced computing systems. Over the next few years, the European HPC community will build on this foundation to advance the technologies underpinning the European HPC value chain.

FURTHER READING:

Results of the Special European Council, 17-18 April 2024
https://bit.ly/European_Council_April24

EuroHPC JU eurohpc-ju.europa.eu

European Processor Initiative
Framework partnership agreement number: 800928
Specific grant agreement number: 101036168 (EPI SGA2)
The JU receives support from Horizon 2020 and from Croatia, France, Germany, Greece, Italy, Netherlands, Portugal, Spain, Sweden, and Switzerland
european-processor-initiative.eu

EUPEX - Grant agreement number: 101033975
The JU receives support from Horizon 2020 and from France, Germany, Italy, Greece, United Kingdom, Czech Republic, Croatia
eupex.eu

EUPILOT - Grant agreement number: 101034126
The JU receives support from Horizon 2020 and Spain, Italy, Switzerland, Germany, France, Greece, Sweden, Croatia and Turkey eupilot.eu

DARE - Grant agreement number: 101202459
The JU receives support from Horizon Europe and Spain, Germany, Czechia, Italy, Netherlands, Belgium, Finland, Greece, Croatia, Portugal, Poland, Sweden, France and Austria dare-riscv.eu

SiPearl website sipearl.com

EPAC webpage
european-processor-initiative.eu/accelerator

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or of the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

In this article, Florian Wohlrab, head of the OpenHW Foundation, introduces us to a first-of-its-kind collection of EU-developed RISC-V components, giving industry and academia access to verified, industry-ready semiconductor intellectual property (IP) from leading European contributors.

Introducing the European Unified RISC-V IP Access Platform (UAP)

Technological sovereignty is a key strategic priority for organizations and governments around the world. Underpinning the digital products that drive modern economies, open-source semiconductor IP is increasingly seen as essential to technological sovereignty, enabling stronger collaboration, and more resilient innovation.

The new European Unified RISC-V IP Access Platform (UAP), created by the TRISTAN consortium and hosted by the OpenHW Foundation (an industry collaboration with the Eclipse Foundation), provides a single source of verified, industry-ready RISC-V components under various licences from TRISTAN, ISOLDE and other European research projects. The UAP is designed for engineers, researchers, startups, and established semiconductor teams who need reliable RISC-V IP with clear licensing and maturity information.

While development is ongoing, the UAP already includes multiple software and hardware components that any end user can adopt, marking an important step toward European technology sovereignty:

- Hypervisor support for CVA6 that follows the RISC-V hypervisor extension
- A Yocto-based Linux image for CVA6 processors
- An eXtension InterFace (CV-X-IF) to add custom accelerators into CVE2 RISC-V Cores
- Trace extensions for CVE4
- ELinOS embedded Linux for RISC-V, which includes the CODEO IDE for building industrial-grade embedded solutions

'The Unified RISC-V IP Access Platform is one of the most important initiatives to come from the TRISTAN project, ensuring that the contributions from consortium partners continue to have impact on the European stage long past the end of our funding. Critically, it enables us to build and nurture a community around European RISC-V that will drive ongoing innovation and collaboration that supports European technological sovereignty,' commented Rob Wullems, innovation strategist at NXP Semiconductors GmbH and the TRISTAN project lead.

From research to industrial adoption

The European Union sees RISC-V as a strategic lever for technological sovereignty and greater competition in the global semiconductor market, now worth roughly USD 700 billion. While the EU currently represents about 10% of this market, (see factsheet in 'Further reading', below), the 2023 European Chips Act aims to double Europe's share to 20% by 2030.

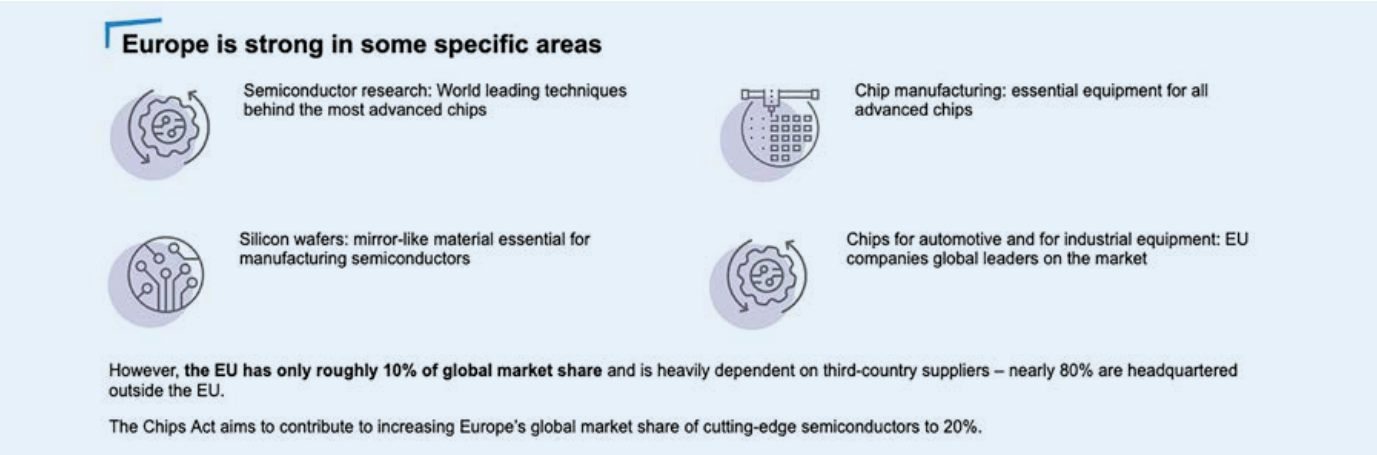
However, when it comes to semiconductors, achieving technological sovereignty is anything but straightforward. Indeed, the recently published paper 'Improving Chip Design Enablement for Universities in Europe' identifies multiple barriers to entry, including costs, legal constraints, and access restrictions. The paper also notes a key gap: too little semiconductor IP is developed and maintained in Europe.

European research projects such as TRISTAN and RIGOLETTO play a vital role in advancing Europe's semiconductor capabilities. But these are not without challenges. These projects have historically delivered strong technical results, yet many have struggled to achieve market adoption after the project ends. Alongside, short project timelines make it difficult to build communities that can maintain and evolve resulting IP.

The UAP addresses this by giving developers and organisations a single place to find verified, industry-ready RISC-V components developed in Europe

A shared catalogue to keep European IP thriving

The UAP acts as a static unified access page, pointing to repositories hosted on the OpenHW Foundation GitHub, automatically mirrored to a European-hosted GitLab instance and to other public forges, or maintained as private assets. As an evolving structure designed to better support integration across



Extract from the Chips Act factsheet

toolchains, accelerators, and infrastructure components, the UAP provides a clear view of each item's maturity, usability, licensing and integration workflow, including documentation, and status information.

The UAP is overseen by the Virtual Repository Task Group, which includes representatives from TRISTAN and ISOLDE. Other Chips JU and RISC-V-related EU projects such as Rebecca, RIGOLETTO, and Scale4Edge have begun to join. As more EU projects open-source their IP, they will be added as maintainers so that each project can curate its own catalogue and ensure continuity beyond the end of TRISTAN.

More features are planned for the UAP, including KPI-based adoption tracking, extended interoperability matrices, improved tooling, and long-term community development.

'The Unified RISC-V IP Access Platform is absolutely critical to supporting technological sovereignty in Europe, and the OpenHW Foundation and Eclipse Foundation are committed to developing it into a sustainable, interoperable, and community driven resource for the wider RISC-V ecosystem. Open-source collaboration is essential to ensuring a competitive playing field, and by working together, we will be able to go further, faster,' said Gaël Blondelle, chief membership officer at the Eclipse Foundation.

From EU research output to industry adoption

TRISTAN (Together for RISC-V Technology and Applications), which launched in 2023, aims to industrialize RISC-V cores – taking them from the lab to the real world, creating a sustainable open-source ecosystem to drive competitiveness and enable more agile innovation in Europe.

The TRISTAN consortium is a diverse group of 46 partners representing a wide range of stakeholders, from household-name organizations and small- and medium-sized enterprises, to research organizations, universities, and industry associations connected to RISC-V. Together, they catalyse expertise and resources from across Europe and beyond to drive innovation and collaboration in the field of RISC-V technology.

The TRISTAN project has resulted in a variety of RISC-V cores, from deeply embedded and microcontroller-performance cores, to verified application-ready cores. Multiple critical peripheral semiconductor IP has been developed, including for debug, trace, interrupt controllers, and hypervisor support. Alongside, software and simulators have been contributed by a variety of consortium partners.

Live now: how to access and contribute to the UAP

The UAP is now live, featuring both open source and proprietary RISC-V IP from TRISTAN, ISOLDE, Rebecca, and Scale4Edge. If you're building semiconductor products, we encourage you to explore the catalogue, leverage European RISC-V IP where it fits, and contribute back to strengthen the ecosystem.

Scan the QR code to browse the UAP Directory:

EXPLORE FURTHER:

European Unified RISC-V IP Access Platform (UAP)
github.com/openhwgroup/tristan-isolde-unified-access-page

OpenHW Foundation
openhwfoundation.org

Chips Act Fact Page, European Commission
digital-strategy.ec.europa.eu/en/factpages/chips-act

L. Krupp, I. O'Connor, L. Benini, C. Studer, J. Rodrigues and N.Wehn, 'Improving Chip Design Enablement for Universities in Europe', 2025
arxiv.org/pdf/2508.14907

As computing systems become increasingly dynamic and modular, featuring components from across the globe, traditional approaches to verifying trustworthiness are falling short. In this article, Alessandro Cilardo, a full professor at the University of Naples Federico II, sets out how the TrustWeave initiative helps engineers design, build, and operate complex infrastructures that remain verifiably trustworthy over time.

TrustWeave

Weaving verifiable trust into Europe's digital infrastructures

Europe's digital infrastructures are undergoing a profound transformation. High-performance computing (HPC), cloud, edge, and artificial intelligence (AI) systems are no longer monolithic platforms deployed once and trusted indefinitely. They are dynamic, composite ecosystems, assembled from hardware and software components sourced across global supply chains and continuously updated over time. This evolution creates a growing tension between the complexity of modern systems and the mechanisms traditionally used to establish trust in them.

Cybersecurity assurance has historically relied on static, product-centric approaches: point-in-time evaluations, documentation-heavy certification processes, and fixed assumptions about system boundaries. These mechanisms remain essential, particularly for regulated procurement and critical infrastructure, where rigorous standards such as Common Criteria are often required, but they increasingly struggle to capture the dynamic aspects of modern digital infrastructures. Questions such as whether a deployed system is still in a trustworthy state, whether its components were updated securely, or whether trust decisions can be automated at scale are difficult to answer with static assurance alone.

The TrustWeave initiative was born from this gap. It originated within the Italian National Centre on HPC, Big Data and Quantum Computing (ICSC), in the context of WP4, 'Trust, Security, and Privacy in HPC', which focuses on integrating hardware-rooted trust into large-scale computing systems. In that setting, TrustWeave emerged as a response to a very practical challenge: how to design, build, and operate complex infrastructures that remain verifiably trustworthy over time, across heterogeneous components and evolving supply chains, while remaining compatible with existing certification and regulatory frameworks.

TrustWeave is not a new certification scheme, protocol, or standard. It is a set of pragmatic guidelines aimed at helping all relevant actors prepare for a future in which static assurance and dynamic, cryptographic trust mechanisms coexist and reinforce each other. Its focus is on concrete design choices, architectural patterns, and operational practices that can already be adopted using mature technologies, and that will remain compatible with more ambitious assurance models as they emerge. Inevitably, these guidelines also speak to policy-makers and certification bodies, suggesting how existing frameworks could gradually evolve through pilots and incremental alignment rather than disruptive change.

At the heart of TrustWeave is a shift from product-centric to component-centric trust. Modern systems are composed of independently developed and managed components: processors and accelerators, firmware layers, hypervisors, orchestration software, AI frameworks, and complex build and update pipelines. Each of these elements may have a different vendor, lifecycle, and threat model. TrustWeave introduces the notion of 'trustable digital components', or TDCs, to capture this reality. A TDC is any hardware, software, or service component capable of establishing a cryptographic identity, binding that identity to its lifecycle state, and producing verifiable evidence about its configuration and provenance.

This perspective aligns naturally with recent advances in hardware-rooted trust, remote attestation, and software supply-chain security. Over the past decade, industry and standards bodies have developed powerful tools for runtime assurance, including attestation architectures and evidence formats, software bills of materials, build provenance, and transparency mechanisms. These approaches excel at providing up-to-date, machine-verifiable trust signals and scale well to large, distributed infrastructures. However, they typically evolve outside formal certification frameworks and lack the regulatory anchoring that certification provides.

"At the heart of TrustWeave is a shift from product-centric to component-centric trust"

Image credit: Nil | stock.adobe.com

TrustWeave's key insight is that these approaches should not be seen as competitors. Certification schemes such as the European Union Cybersecurity Certification Scheme on Common Criteria (EUCC) remain invaluable as trust anchors, providing deep assurance that a component or system was designed and evaluated correctly against a defined threat model. Dynamic assurance mechanisms, in turn, can answer a different question: whether a specific deployed instance is trustworthy right now. TrustWeave proposes to weave these layers together by combining certification artefacts, industry audit results, attestation evidence, and provenance metadata into a coherent, layered trust model.

This convergence is particularly relevant for Europe. Digital sovereignty increasingly depends not only on where technology comes from, but on the ability to verify and control trust across complex supply chains. Confidential computing, federated AI, and privacy-preserving data processing, all central to Europe's HPC and AI strategies, fundamentally rely on hardware-rooted trust and attestation to deliver credible guarantees. At the same time, Europe's regulatory landscape is evolving rapidly, with initiatives such as EUCC and the Cyber Resilience Act pushing towards secure-by-design and lifecycle-aware approaches.

TrustWeave contributes a technical and operational perspective to this evolution. It encourages manufacturers to design products around clear trust boundaries, enable roots of trust, and produce verifiable evidence of firmware and software provenance. It helps system integrators and operators reason about trust in heterogeneous systems assembled from independently sourced components. It also offers policymakers and certification authorities concrete ideas for supporting pilots, recognizing complementary assurance mechanisms, such as attestation services, and gradually evolving frameworks without undermining existing recognition mechanisms.

TrustWeave was initiated in the context of HPC, but its scope deliberately spans cloud, edge, and AI systems. It is intended as an open, collaborative effort bringing together researchers, industry practitioners, certification experts, and policymakers who face the same trust challenges from different angles. Interested readers are invited to engage through the TrustWeave GitHub repository (see 'Further information', below), propose contributions, or join the editorial team.

As Europe invests in the next generation of digital infrastructure, trust can no longer be treated as a static property established once and assumed forever. In complex, composite systems assembled from globally sourced hardware and software components, trust must be continuously justified and recomposed across components, suppliers, and lifecycles. TrustWeave offers a practical way to start addressing this challenge, by enabling verifiable, scalable trust in modern digital infrastructures without abandoning the assurance foundations on which Europe already relies.

FURTHER READING:

Contact:
Alessandro Cilardo acilardo@unina.it

Italian National Centre on HPC, Big Data and Quantum Computing
supercomputing-icsc.it

TrustWeave on GitHub
github.com/trustweave-team

While vast data troves offer rich possibilities for research, ensuring sensitive data is properly handled is essential to ensure trust. In this article, Jano van Hemert, EPCC, University of Edinburgh, explores the ways in which data safe havens may be defined and calls for a harmonized definition of these trusted environments.

Towards a harmonized definition of data safe havens

Data safe havens (DSH) are information systems that enable the processing of sensitive data. They are also referred to as trusted research environments, secure data environments and secure processing environments. A DSH has the following stakeholders:

1. Citizens who are associated with the sensitive data.
2. Data owners who manage these data in an operational context.
3. Researchers who require access to these data.
4. Information governance responsible for providing access to these data.
5. A digital infrastructure provider to store and process these data.
6. Linkage services to enable individual-level record linkage in these data.

Currently, we have no universal definition of what constitutes a DSH. Often a DSH is defined by its **purpose**, such as in the 2025 Scottish Safe Havens Charter:

'Safe havens are secure facilities that provide a controlled environment for accessing and processing personal health and social care data.'

A second approach to defining a DSH is by listing the requirements it must satisfy. For example, in 2022 the UK Government published policy guidelines to define the **requirements** for a secure data environment for the UK's National Health Service (NHS) health and social care data:

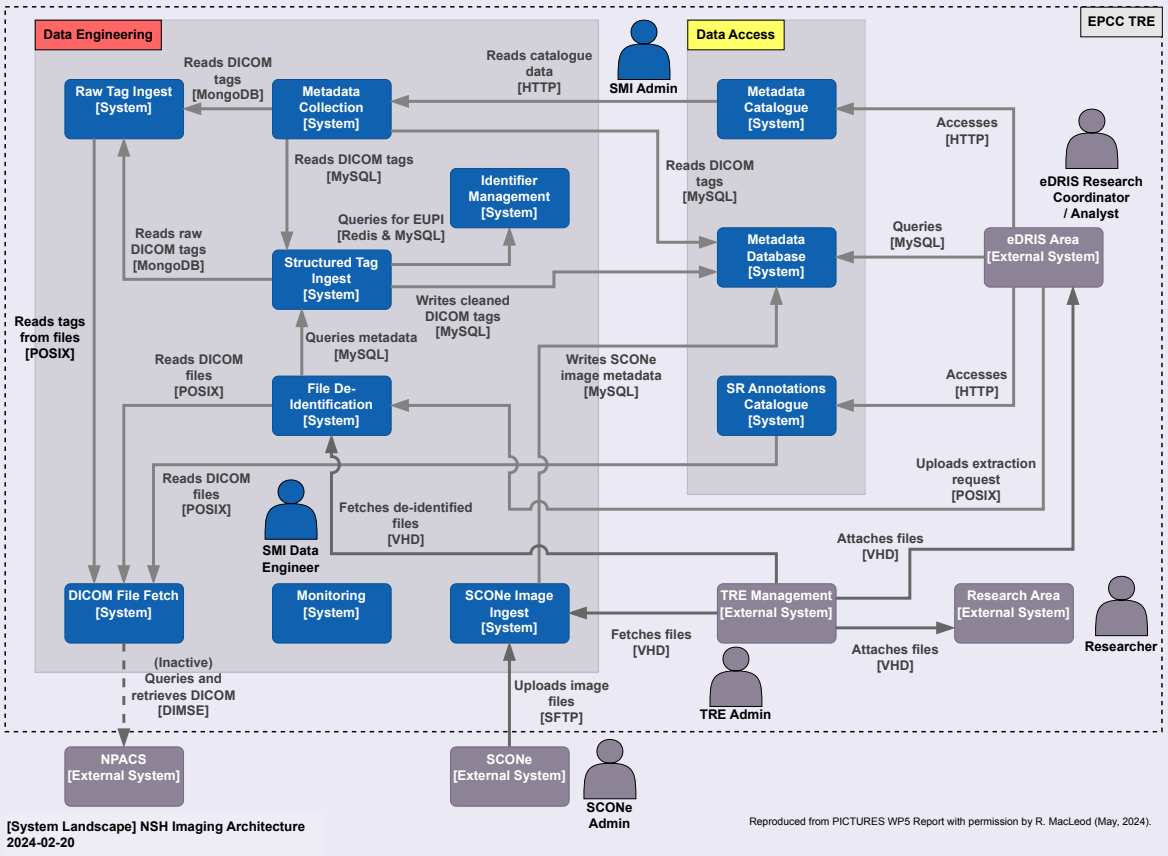
'Secure data environments allow organisations to control who can become a user to access the data, the data that users can access, what users can do with the data in the environment and the information users can remove [meaning: extract].'

These guidelines are structured according to the 'Five Safes' framework, a set of principles which enable data services to provide safe research access to data developed by the UK Office for National Statistics. The framework categorizes requirements into settings, data, people, projects and outputs. Within data and projects are requirements that refer to the legal basis to allow the processing of the data and to the privacy safeguards needed because of legislation such as General Data Protection Regulation (GDPR).

Requirements guide us to the third approach: to define a DSH through its processes. We know these processes already exist in each operational DSH. For instance, the act of extraction in a DSH is referred to as disclosure control and is a distinctive feature. Other examples are the accreditation of researchers ('safe people') and approval of a project by an ethics committee ('safe projects'). These processes may vary across two DSHs even if designed to satisfy the above-mentioned 'Five Safes' framework.

A fourth approach is to define the architecture of the **digital infrastructure** that supports DSH processes. The combination of infrastructure and processes for a specific DSH is tailored to the digital capability and legal requirements set by the data owners. For example, the Scottish National Safe Haven maintains a copy of 2.5 billion images of Scottish medical imaging data alongside copies of many more types of patient records from the Scottish National Health Service; it derives subsets of these data for research studies within the DSH in its researcher-facing environment. The digital infrastructure is provided through EPCC's Safe Haven Services at the Edinburgh International Data Facility.

The architecture to support the processes of a DSH can become complex, especially if the infrastructure maintains full copies of operational data. In the diagram we show the architecture of the Scottish Medical Imaging (SMI) Archive. The architecture shows how roles and system components interact to manage data in SMI. Much of the complexity in the processing of sensitive data is handled in the 'data engineering' part, which is completely invisible to the researchers. This is necessary as the system and data admin roles in the digital infrastructure must be kept separate from researchers.



The architecture of the Scottish Medical Imaging Archive

The fifth and final definition is through the **data** that is made available in the DSH for its purpose. These data are often of a specific category. For SMI, it is routine healthcare medical imaging data. Furthermore, these data are always bound to people linked to a geographical region for the operational purpose of the data. In the case of SMI, these people are the five million patients that have undergone a radiology imaging session in Scottish NHS healthcare facilities.

With purpose, requirements, processes, digital infrastructure and data 'definitions' in place we can ask an important question: who decides when a DSH is a DSH? In the UK, several organizations provide access to sensitive public administrative data. For instance, the Scottish National Safe Haven is provisioned through Public Health Scotland's Electronic Data Research and Innovation Service (information governance), the University of Edinburgh's EPCC (digital infrastructure) and National Records of Scotland (linkage). These organizations must be accredited by the UK Statistics Authority as stipulated by UK law (Digital Economy Act 2017). However, many DSH operate under different legislations, and that will affect all their 'definitions'.

In the UK, several DSHs have been operational for over a decade, each with their own 'definitions'. With recent changes in legislation and governmental policies, we will see an increase in DSHs in the UK and European Union (EU). For instance, on 11 February 2025, the EU put into force Regulation (EU) 2025/327 to establish the European Health Data Space to harmonize access, in terms of purpose and requirements, to healthcare data about natural persons across the European Union for the purposes of secondary use.

At the same time, we will see demand from researchers to make access across DSH more streamlined. Studies will need to combine data from two or more DSHs. Research will also require access to specific digital infrastructure such as domain-specific software, large storage and accelerated compute resources, which may be provided by organizations where the data is not residing. These rising demands are creating the challenge for DSHs to harmonize their definitions in terms of their purpose, requirements, processes, digital infrastructures and data. Let us start by sharing the 'definitions' of the DSHs we have in operation today.

FURTHER INFORMATION:

bit.ly/EIDF_Safe_Haven_Services


In this article, Grazia Garzo (University of Siena) and Alessandro Palumbo (CentraleSupélec, Inria, CNRS, IRISA) give an insight into their research investigating methods for transparent, accountable AI in the legal domain.

Building trustworthy legal intelligence: From explainable AI to controlled natural language

The growing intersection between artificial intelligence (AI) and the legal domain presents one of the most profound societal challenges of our time. How can we ensure that AI systems assisting judicial reasoning remain transparent, accountable, and under meaningful human control?

This question lies in a research line we are carrying out exploring the path from interpretable machine-learning (ML) models and linguistic formalization, while critically addressing the ethical implications of AI-driven legal systems

The ethical dimension of digital justice

In our paper 'Legal & Ethical Implications of Predictive Digital Techniques in the Judicial Criminal Proceedings' (ISDFS 2025), we examined the ethical, legal, and societal challenges of predictive AI in the Italian justice system. While such systems can improve efficiency, uniformity, and transparency, they also increase the risk of algorithmic bias, judicial conformism, and loss of autonomy.

The paper calls for human-centred AI governance, emphasizing accountability, explainability, and legal responsibility. Key recommendations include the adoption of balanced datasets, interpretable models, clear responsibility frameworks, and continuous monitoring of algorithmic outputs to prevent legal crystallization and social bias.

From Boolean logic to explainable justice

In 'Opening the Black Box: How Boolean AI Can Support Legal Analysis' (CCAI 2024), we proposed a Boolean AI framework to model judicial reasoning in road homicide cases under Italian law. By encoding 44 binary legal features (such as speeding or impairment) into Boolean functions, the model reproduced judicial reasoning with 69% prediction accuracy and 83% feature-extraction precision, offering a transparent, rule-based alternative to opaque statistical learning.

From algorithms to hardware: DAJE

The next step, 'The First Hardware Circuit Emulating Italian Road Homicides Legal Logic (DAJE)' (ES 2025), explored the feasibility of translating legal reasoning into hardware logic. DAJE – Digital Assurance Judicial Enforcer – is a conceptual prototype demonstrating how a subset of judicial rules can be expressed in Boolean form and mapped onto digital circuits. Rather than a full hardware implementation, the work highlights the potential of accelerating reasoning tasks through dedicated architectures, much like how modern systems accelerate networking or cryptographic functions. Beyond its analytical role, this methodology demonstrates how formal, interpretable models can facilitate the structured processing of judicial data, thereby enabling the consistent representation of evidence, facts, and arguments across various cases.

This exploratory design thus opens the perspective that, in principle, dedicated architectures could one day support or accelerate computational modules for legal reasoning, embedded within trustworthy and tamper-resistant environments, reinforcing the European vision of transparent and sovereign computation.

Evaluating AI's limits: Does ChatGPT understand the law?

For our paper 'Does ChatGPT Understand the Law?' (JURIX 2025), we evaluated the legal literacy of GPT-4o. Testing the model on 60 legal definitions and ten real rulings, we found that, while it accurately reproduces definitions (with an average of 1.95/2), it fails to establish causal and normative relations, often hallucinating citations. This confirms that generative models mimic legal reasoning linguistically but lack a structured, normative understanding, highlighting the need for hybrid architectures that combine logic-based and language-based AI.

Human in the loop: formalizing the law in ACE

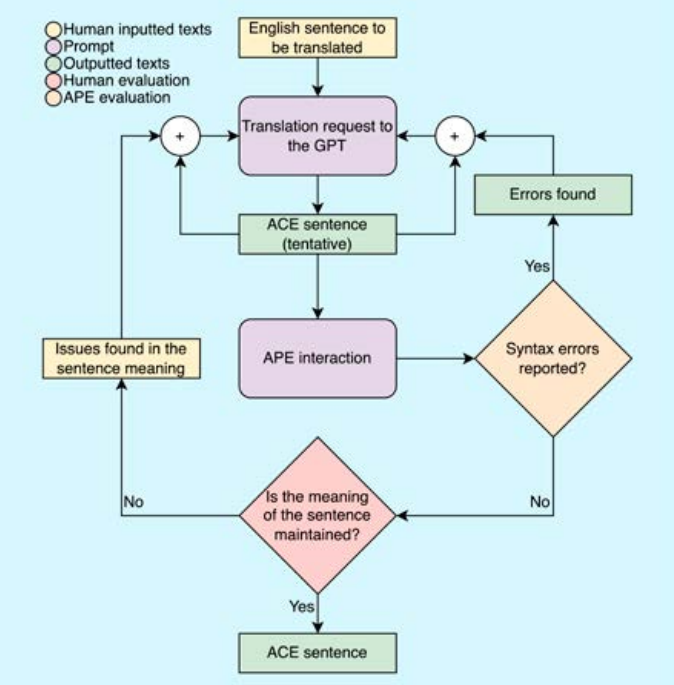
Our study 'Human-in-the-Loop: Legal Knowledge Formalization in Attempto Controlled English (ACE)' (ISDFS 2025) introduces a semi-automated workflow to formalize legal texts through an iterative and verifiable process. Attempto Controlled English (ACE) is a controlled natural language with a simplified grammar and vocabulary, designed so that each sentence is unambiguous for humans and automatically interpretable by

machines. Because ACE closely resembles standard English, even non-legal experts can follow complex legal reasoning and understand normative statements more easily.

The workflow methodology reported in the paper, shown in the figure below, begins by providing the legal text to a large language model (LLM), asking it to rewrite the passage according to ACE syntax rules. The resulting sentence is then processed by the Attempto Parsing Engine (APE), which checks grammatical and syntactic correctness. If APE reports errors, these are sent back to the LLM, which reformulates the sentence accordingly. This loop continues until APE validates the ACE sentence and produces its discourse representation structure (DRS): a formal logical representation (machine computable) of the inputted sentences.

Once syntactic and grammar validation are achieved, the final step is the semantic verification. Experts check whether the ACE translation preserves the legal meaning, intent, and consistency of the original provision. If the translation aligns semantically with the source text, the ACE version is accepted as correct; otherwise, the process restarts from the beginning.

Through this approach, the authors translated sections of EU Directive 2009/103/EC and CJEU rulings (Vnuk, Linea Directa), obtaining 45 validated ACE sentences that maintain both syntactic correctness and legal fidelity. Writing legal texts directly in ACE could therefore make laws clearer to citizens and experts alike, while simultaneously enabling automatic reasoning and interoperability in AI-assisted legal systems – a crucial step toward transparent, human-aligned digital justice.



Towards machine-understandable law

Would it be possible to design an automatic compiler that integrates the strengths of both LLMs and the Attempto Parsing Engine (APE)? Can we build a system able to automatically verify not only the syntax but also the semantic consistency of legal sentences?

Such a tool could autonomously translate normative texts into unambiguous, machine-understandable representations while preserving their original legal meaning. The ultimate goal of this research direction is to build fully automated yet transparent pipelines capable of converting legislative texts into formal logic structures interpretable by machines and verifiable by humans. These frameworks could enable AI systems that assist rather than replace judges, ensuring that automation serves justice without compromising its ethical and legal foundations.

We are open to collaborations and interdisciplinary exchanges, and we invite readers to get in touch if they share an interest in exploring these ideas further. Suggestions for further reading, discussing more in detail the approaches and vulnerabilities, follow.

g.garzo@student.unisi.it
linkedin.com/in/grazia-garzo-661685166/

alessandro.palumbo@inria.fr
palessumbo.github.io

FURTHER READING:

G. Garzo, S. Ribes, A. Palumbo, 'Opening the Black Box: How Boolean AI Can Support Legal Analysis', CCAI 2024
bit.ly/Garzo_Ribes_Palumbo_CCAI_2024

G. Garzo, A. Palumbo, 'The First Hardware Circuit Emulating Italian Road Homicides Legal Logic' *(DAJE)*, ES 2025
bit.ly/Garzo_Palumbo_ES2025

G. Garzo, A. Palumbo, 'Does ChatGPT Understand the Law?', JURIX 2025
bit.ly/Garzo_Palumbo_JURIX2025

G. Garzo, A. Palumbo, 'Human-in-the-Loop: Legal Knowledge Formalization in Attempto Controlled English', ISDFS 2025
bit.ly/Garzo_Palumbo_ISDFS2025

G. Garzo, A. Palumbo, 'Legal & Ethical Implications of Predictive Digital Techniques in the Judicial Criminal Proceedings', ISDFS 2025
bit.ly/Garzo_Palumbo_ISDFS2025_2

With computing now integrated into everything from pacemakers to power grids, while the complexity of component supply chains increases, robust protection against cybersecurity vulnerabilities is more important than ever. In this article, the consortium of the European-funded RESCALE project explain how the project is addressing cybersecurity vulnerabilities and providing a platform with verifiable guarantees.

Fortifying the foundation
RESCALE's revolution in secure supply chains

As the computing landscape becomes increasingly complex, reliance on integrated hardware, firmware and software components sourced from expansive supply chains has introduced systemic integrity gaps. Supply-chain attacks and accidental vulnerabilities have become a defining risk for modern computing.

Whether a cloud service built from open-source packages or an embedded platform combining firmware and silicon intellectual property (IP), today's 'product' is a moving graph of dependencies. High-profile incidents, such as the SolarWinds attack, have underscored the urgent need for verifiable security and trust across the entire ecosystem.

It is against this backdrop that the Revolutionised Enhanced Supply Chain Automation with Limited Threats Exposure (RESCALE) project (grant agreement 101120962) has developed a comprehensive, secure-by-design assurance framework, making supply-chain assurance automated, evidence-based and continuously updated.

Ahead of our dedicated workshop at the HiPEAC 2026 conference in Kraków, RESCALE is moving from a conceptual framework to a fully integrated and validated platform. Our mission is to deliver verifiable audit procedures and guarantees, replacing reliance on unverified traditional documentation with the project's central innovation: the trusted bill of materials (TBOM). This article outlines the project's purpose and the substantial architectural and technical progress achieved in creating a resilient foundation for the next generation of computing security.

From SBOMs to TBOMs: Trust with proof

As already stated, RESCALE's central concept is the trusted bill of materials (TBOM). A software bill of materials (SBOM) answers the question 'what is included?', but not whether components were assessed, how, or whether the evidence is current.

RESCALE binds components to verifiable security evidence and trust metadata via two artefacts: a static supply-chain component guarantee (SSCG), derived from static analysis (including formal techniques), and a dynamic supply-chain component guarantee (DSCG), derived from runtime testing (including fuzzing, firmware analysis, and other dynamic assessments). Together, these enable a traceable path from tool output to a structured, auditable trust record.

Trust as a living property

Rather than being a one-off assessment, trust is dynamically evaluated in RESCALE. TBOMs are linked to a bill of vulnerabilities (BOV) that is updated when new vulnerability intelligence appears or when new analysis results are generated. The TBOM lifecycle tracks vulnerability posture across severity categories (instead of an 'average' score) and supports an 'unknown' class for newly detected issues that are not yet in public databases. This matters in practice: supply-chain security must reflect both known common vulnerabilities and exposures (CVEs) and issues discovered through project-specific analyses.

Members of the RESCALE consortium

A deployment-oriented architecture

RESCALE organizes the platform into three cooperating domains. The Assessment Domain hosts the testing toolchain and the SSCG / DSCG generators. The Management Domain orchestrates workflows and stores TBOM-related documents in repositories. The Security & Trust Domain provides continuous monitoring, certificate-based identity and distributed-ledger integration for tamper-evident traceability across organizations.

At the centre sits the Trust Orchestrator (TrustOR). TrustOR generates and validates TBOMs by ingesting SBOMs, SSCGs and DSCGs, managing TBOM lifecycle state and coordinating re-evaluation when new security signals arrive. The ledger stores cryptographic proofs (hashes and relationships), while full documents remain in managed repositories, enabling verifiability without unnecessary disclosure.

Progress: integrating a broad security toolchain

RESCALE has deliberately advanced an assessment toolbox which is apt for continuous integration (CI). On the static side, SASTer-CLI provides containerized, modular scanning for common languages (e.g., C / C++ and Python) with standardized outputs. SAVE-ME applies static vulnerability detection based on machine learning to Erlang using a fine-tuned CodeBERT approach, while DetectEr adds runtime / formal verification techniques for concurrent systems. Where source code is unavailable, IVEE contributes binary-level analysis using symbolic execution.

On the dynamic side, RAISE extends stateful representational state transfer application programming interface (REST API) fuzzing with optimization to prioritize higher-risk request sequences, while EvoMaster-based approaches (including BRUTE) support evolutionary testing with runtime feedback. For embedded firmware, FATex combines extraction and emulation-based analysis to produce actionable findings. RESCALE also includes hardware-relevant assessment capabilities, including side-channel leakage evaluation techniques, to better capture risks that matter for real deployments.

Pilots: GRiSP.io and SkyFlok in the loop

Two pilot tracks help ensure RESCALE works in practice. The pilot led by Peer Stritzinger centres on the GRiSP.io ecosystem, integrating static and dynamic testing into continuous integration / continuous development (CI / CD) for Erlang and C components and producing guarantees that can be published to the platform or executed in 'dry run' mode for developer-only feedback. The pilot led by Chocolate Cloud focuses on the SkyFlok platform, applying analysis to Python microservices and a shared in-house C++ library (Rlnclib), enabling realistic evaluation of dependency propagation across services.

Beyond producing evidence, the pilots exercise supply-chain behaviours: validating TBOMs in the dashboard for end users, generating TBOMs for shared dependencies and triggering update-and-notify flows when new vulnerabilities appear or when new dynamic results change the TBOM / BOV.

RESCALE tracks progress systematically against earlier state-of-the-art gaps. In the latest consolidation, 14 of the initial 30 gaps are fully addressed, 12 are partially addressed and progressing through final integration and validation and four remain open with defined resolution actions. Pilot integration has also surfaced six new gaps, reinforcing RESCALE's focus on operational practicality.

RESCALE at HiPEAC 2026

The RESCALE workshop at HiPEAC 2026 will showcase the end-to-end flow: automated assessment producing SSCG / DSCG evidence, TBOM generation and validation via TrustOR, ledger-backed traceability and continuous trust updates as vulnerability intelligence evolves. For the HiPEAC community, where embedded platforms, hardware / software co-design and performance-critical systems meet, RESCALE offers a practical blueprint for supply-chain assurance that spans source code, binaries, firmware and hardware.

As RESCALE moves into its final phase, the focus is on tightening end-to-end integration, completing validation of partially covered requirements and packaging the platform for adoption, on-premises where needed, CI-native where possible.

FURTHER INFORMATION:

RESCALE D2.5: High Level Architecture (first version) zenodo.org/records/16971227

RESCALE: Contribution to standardization | RESCALE White Paper 2025
zenodo.org/records/17896135

RESCALE Workshop at HiPEAC 2026
hipeac.net/2026/krakow/#/program/8263

The RESCALE project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101120962.

In this article, Ariadna Rodríguez, Davide Cirillo and Daniele Lezzi (Barcelona Supercomputing Center – Centro Nacional de Supercomputación) brief us on a European success story in which an artificial intelligence (AI) edge-to-cloud research prototype was transformed into a market-ready solution for lifelong stroke prevention.

From research to market: How a European AI breakthrough became a scalable solution for stroke prevention

Every year, more than 12 million people worldwide experience a stroke, causing between 6.5 and 7 million deaths and leaving over 100 million survivors with long-term consequences. In Europe, stroke stands as the leading cause of adult physical disability, placing an annual economic burden of €60 billion on healthcare systems, families, and society. As populations age and risk factors increase, global incidences of stroke are projected to grow rapidly, creating an urgent need for effective, scalable, and personalized prevention technologies.

Electrocardiograms (ECGs), routinely recorded in hospitals and increasingly via smartwatches, often alongside continuous pulse sensors, remain central to diagnosing atrial fibrillation (AF), a major cause of cardioembolic stroke. Recent advances in artificial intelligence (AI) enable highly accurate AF detection from these signals, yet current AI tools still rely almost exclusively on electrophysiology data. Multi-omics information, including genomic and proteomic profiles, provides complementary insights into an individual's inherited and environmentally influenced risk. Integrating multi-omics data with ECG-based monitoring could significantly enhance personalized AF detection and stroke risk management, opening new opportunities for AI-driven, data-intensive healthcare solutions.

To address this need for multi-omics information, the EU-funded project AI-SPRINT set a scientific challenge for personalized medicine. The research carried out during the project is now being commercialized by OneCareAI, a spinoff of the Barcelona Supercomputing Center – Centro Nacional de Supercomputación (BSC-CNS). The work behind these advancements is jointly led by two different departments at the BSC, represented by Daniele Lezzi (Computer Science Department) and Davide Cirillo (Life Sciences Departments), who were principal investigators in AI-SPRINT.

OneCareAI founders Davide Cirillo (left) and Daniele Lezzi

AI-SPRINT: Edge-to-cloud prototype

The story of OneCareAI begins with the AI-SPRINT project, funded by the European Union's Horizon 2020 Research and Innovation Programme. Within this initiative, the Computer Science Department and the Life Sciences Department of the BSC worked together to lead the Personalized Healthcare Use Case with the goal of creating an edge-to-cloud AI prototype for cardioembolic stroke risk assessment. The prototype focused on detecting AF from single-lead ECGs, while integrating clinical questionnaire data such as lifestyle, high-risk conditions, age, and sex.

The first AF detection model, built using the BSC-developed PyCOMPSs/dislib library and trained on public datasets, was recognized by the European Commission's Innovation Radar. Its performance has since been enhanced using a deep transformer architecture, achieving over 95% accuracy, matching state-of-the-art models tested on equivalent data.

The model was tested through two pilot studies, both compliant with the General Data Protection Regulation (GDPR):

- **Spain (2023)**: In collaboration with Freno al Ictus and Nuubo, the system analysed over 9,000 hours of ECG data from 11 stroke survivors and 14 healthy individuals aged 31–78. These tests revealed clinically relevant sex differences in post-stroke behaviour.
- **Italy (2024)**: At Niguarda Hospital in Milan, the model was applied to 30 patients with giant cell arteritis, evaluating stroke predisposition within a different pathology.

With these validations, the prototype reached technology readiness level (TRL) 4, demonstrating solid technological and clinical potential.

Innostroke: Building the infrastructure for scalability

Following the success of AI-SPRINT, the story continued with the launch of the Innostroke project, funded by funded by the Spanish Ministerio de Transformacion Digital y de la Función Publica within the framework of the Artificial Intelligence Strategy 2024 and the Plan de Recuperación, Transformación y Resiliencia, funded by the European Union – NextGeneration EU.

The project delivered a comprehensive infrastructure that includes a cross-platform mobile app for both iOS and Android, support for Fitbit, Apple Watch, LifeVit, and other consumer wearables, and a centralized REST application programming interface (API) that connects system components. It also provides a clinician-oriented web dashboard and a co-designed patient registry developed in collaboration with Hospital Sant Pau and Hospital del Mar in Barcelona. In addition, the system features an administrative interface for overall management and a secure, GDPR-compliant data architecture integrated with an AI inference engine.

'Innostroke provided the architectural foundation that transformed the AI-SPRINT research prototype into a scalable digital health platform ready for real-world integration,' states Daniele Lezzi, principal investigator of the project.

The full system is deployable across cloud and high-performance computing (HPC) environments, such as private clouds (BSC virtual instances), commercial clouds (Google Cloud Platform) and OVHCloud (currently funded through the OVH Startup Program).

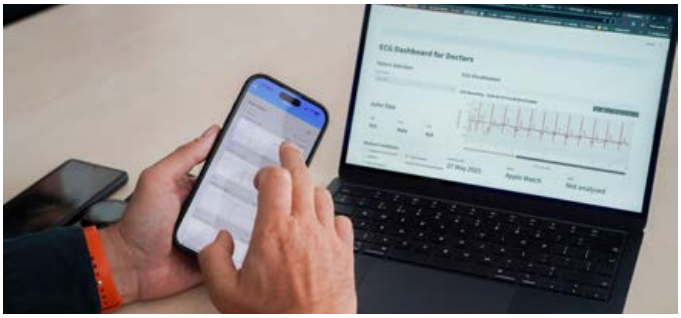
Thanks to Innostroke, researchers were able to develop a fully functional infrastructure and a validated AI model, demonstrably at TRL 6, which provided the foundation for a spinoff.

OneCareAI: European deep-tech spinoff

In this context, BSC officially launched OneCareAI, a spinoff dedicated to translating years of research into a commercial product capable of improving stroke prevention and long-term heart-health management. OneCareAI's mission is to harness HPC, multimodal AI, and wearable data to deliver personalized, continuous, and clinically reliable stroke risk prediction.

OneCareAI interprets ECG-derived electrophysiological patterns as surrogate markers of multi-omics indicators, thereby addressing a gap in current market solutions, which typically analyse each data stream in isolation. This multimodal integration is essential for capturing early warning signals and adapting risk profiles dynamically over time.

'OneCareAI bridges a critical gap in today's technologies by using ECG-based patterns as proxies for multi-omics insights.



OneCareAI's infrastructure and model allows the use of ECG-based patterns as proxies for multi-omics insights

This unique multimodal integration enables us to catch early warning signals and continuously refine stroke-risk profiles,' explains Xavier Guillem, the chief executive of OneCareAI.

With a strong multidisciplinary team, an established advisory board, and clear commercial routes through hospitals, insurers, and device manufacturers, OneCareAI is preparing for full-scale market deployment.

Next steps

OneCareAI will shortly launch a pilot study with Hospital Sant Pau that involves the recruitment of stroke patients to validate the accuracy of stroke-risk prediction models, paving the way for broader market introduction. The patients will wear a smartwatch to collect the ECGs through the OneCareAI app while the doctors at the hospital will have access to a rich set of data analysis via a web dashboard. The pilot has been designed to ensure adherence to the highest regulatory and ethical standards, facilitating market access and building user trust.

This marks an important milestone for the project, as it will allow the team to test the technology in real clinical conditions, refine its performance based on real-world data, and accelerate its path toward meaningful impact for patients and healthcare professionals.

Conclusion

From its origins in EU research to its transformation into a promising digital health startup, the story of OneCareAI exemplifies how Europe's innovation pipeline can successfully guide scientific breakthroughs toward clinical and commercial impact.

'Our journey illustrates the full arc of scientific innovation, showing how fundamental research, when properly validated and scaled, can evolve into solutions that meaningfully improve public health,' states Davide Cirillo, principal investigator of the project.

onecareai.com
ai-sprint-project.eu
innostroke.com


Managing the increasing complexity of diverse computing hardware is an ongoing challenge for the computer-architecture community. In this article, Manuel de Castro and Arturo Gonzalez-Escribano (University of Valladolid), introduce us to the EPSILOD framework, which uses abstraction to tame the complexity of heterogeneous distributed systems.

EPSILOD: Open-source framework for iterative stencil loops on highly heterogeneous distributed systems

Computing is becoming increasingly heterogeneous, as today's challenges demand specialized devices to achieve higher performance and lower energy consumption. Nowadays, supercomputers can easily comprise thousands of distributed nodes, with multiple central processing unit (CPU) cores and graphics processing units (GPUs), while embedded and edge devices combine low-power CPUs, GPUs, field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs) into a single system-on-chip. Artificial intelligence (AI) is also contributing to this heterogeneous explosion, as AI accelerators such as tensor processing units (TPUs) are being included in customer-grade CPUs and GPUs. This trend is expected to continue in the next years.

Programming efficient applications for heterogeneous platforms is significantly delicate, as each computing device presents its own programming paradigms, application programming interfaces (APIs), software-development frameworks, and optimization techniques. Additionally, all the computing devices used in a system have to be properly managed, coordinated, and synchronized by the runtime. This is especially complex when the devices are distributed across multiple nodes.

Iterative stencil loops (ISLs) are a representative class of scientific applications with high scalability potential. In each iteration, all the elements of a multidimensional grid are updated using a fixed computation, based on the values of a set of neighbouring elements in the previous iteration, described by a stencil pattern. As an example, convolutions are a subset of stencil computations. Each element is computed independently of the others, but all elements present dependencies on the previous values of their neighbours.

Many computing devices, including multicore CPUs, GPUs, and FPGAs, can leverage different optimization techniques for efficient ISL execution. However, when executing ISLs over massive grids on multiple computing devices concurrently, the data dependencies introduce data communication requirements between them. Even if the programming and data movement patterns are regular, mixing different programming models and techniques, and coordinating efficient data communication across different programming and execution technologies, makes developing such solutions cumbersome and error prone.

The Trasgo research group at the University of Valladolid has developed EPSILOD, a parallel skeleton for developing efficient ISL applications on heterogeneous distributed systems. EPSILOD abstracts the details of work sharing and computation / communication patterns to simplify development. It is presented as a C11 library with the execution engine enclosed in a highly abstract function that can target different computing devices, such as multicore CPUs, NVIDIA or AMD GPUs, and Intel FPGAs, using their native runtimes and compilers in the toolchain. The user can provide device-optimized implementations of the core stencil kernel for each device. EPSILOD automatically executes the appropriate kernel version on the corresponding device. In addition, it abstracts the inter-device, inter-node data communications and device synchronizations, whatever the combination of devices or the number of distributed nodes used.

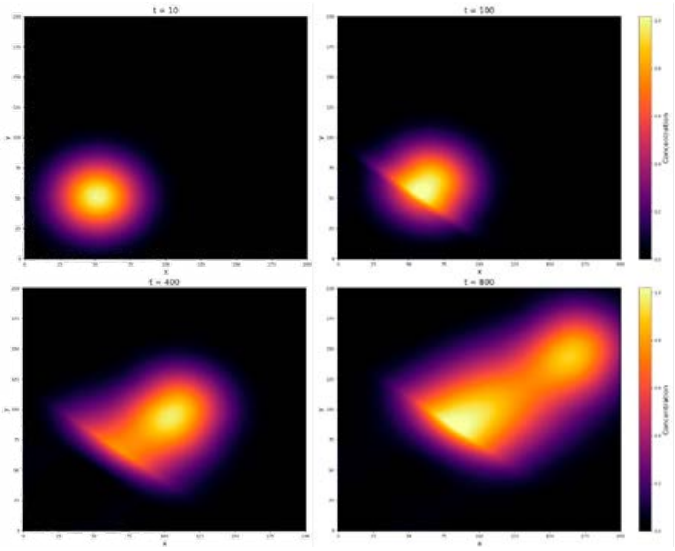
EPSILOD has been built on top of two function libraries also developed by our group: the Controllers library, which provides a heterogeneous portability layer; and the Hitmap library, which provides an abstraction for load distribution and communication in distributed systems. EPSILOD automatically distributes the global application domain or multidimensional grid across the different nodes and devices, generates the data communication patterns, handles intra- and inter-process communications, and overlaps computation and communication whenever possible to achieve high performance. Internally, EPSILOD differentiates three logical parts on the computation grids: the inner part, which is local to each device and is not communicated; the borders, which comprise the elements of the local domain that are computed and communicated to other devices; and the halos, which are the elements received from other

devices, that are not computed locally. EPSILOD overlaps halo and border communications with inner computation, which efficiently hides the data movement latencies and enables good scalability.

EPSILOD can execute hand-optimized kernels for different devices, but we also provide a domain-specific language (DSL) to express general stencils in terms of the neighbourhood description of a generic cell (relative positions and weights) and optional local computations. The translator provided generates baseline implementations of the kernel for specific devices.

In our first publication about EPSILOD, we showed its weak and strong scaling capabilities with 2D stencils in a distributed cluster with tens of NVIDIA GPUs. We also showed its adaptability to heterogeneous environments by mixing NVIDIA and AMD GPUs. Recently, through the EuroHPC Joint Undertaking programme, we also tested its weak scaling capabilities with more complex stencils, such as a 3D lattice-Boltzmann method application, in a massive cluster with up to 256 NVIDIA GPUs on 64 distributed nodes. In our experiments, EPSILOD outperforms state-of-the-art solutions based on other parallel skeletons such as Muesli, or more general solutions programmed with Celerity / SYCL.

We also extended EPSILOD to support Intel FPGAs. With this addition, EPSILOD can leverage device combinations that are unsupported by common heterogeneous frameworks. We conducted an experimental study in a highly heterogeneous test system, which showed that, by using static load balancing techniques, EPSILOD can efficiently leverage a whole system comprising a multicore CPU, an NVIDIA GPU, and an Intel FPGA. Compared to only using the GPU, EPSILOD could compute the same domain in 37.6% less time, and a 1.78× bigger domain in the same time.



Example of an iterative stencil-loop simulation representing the temporal evolution of smoke concentration in a fire. An initial deflagration leads to the formation of a front that propagates downwind, driven by the wind field that transports the smoke plume.

Our current research lines with EPSILOD include testing its scalability capabilities in other pre-exascale computers, enabling more FPGA-specific optimizations, and integrating and testing automatic load balancing techniques.

EPSILOD and its underlying libraries are fully open source. We also plan to create tutorials aimed at assisting researchers and scientists in utilizing EPSILOD.

FURTHER INFORMATION:

M. de Castro, I. Santamaria-Valenzuela, Y. Torres, A. Gonzalez-Escribano, and D. R. Llanos, 'EPSILOD: efficient parallel skeleton for generic iterative stencil computations in distributed GPUs'. *The Journal of Supercomputing*, January 2023. DOI: https://doi.org/10.1007/s11227-022-05040-y

D. Díez, S. Alonso Pascual, and A. Gonzalez-Escribano, 'Easy programming of stencil applications for large-scale systems with a parallel skeleton'. Under review for the *International Journal of Parallel Computing*.

M. de Castro, S. Alonso Pascual, R. Gran Tejero, Y. Torres, and A. Gonzalez-Escribano, 'Exploiting highly heterogeneous systems with stencil applications'. To appear in the *Proceedings of The 23rd International Workshop on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Platforms (HeteroPar 2025)*. Preprint available at: doi.org/10.5281/zenodo.17142750

GitLab Repository – EPSILOD
gitlab.com/trasgo-group-valladolid/epsilod

Contact:
trasgo@infor.uva.es
arturo@infor.uva.es
manuel@infor.uva.es

trasgo

Boosting careers at the HiPEAC STEM Student Day

Every year at the STEM Student Day, students get the chance to connect with companies working at the frontier of computing systems. In this article, Ezgi Sena Karabacak and Clàudia Pàmies explain how student activities at HIPEAC 2025 gave them valuable experience and contacts.

From HiPEAC hackathon to RISC-V vector research

Clàudia Pàmies, Universitat Autònoma de Barcelona

At HiPEAC 2025, I had the opportunity to take part in a RISC-V hackathon organized in partnership with Openchip and Barcelona Supercomputing Center. Our three-person team attempted to build an internet-of-things (IoT) platform to monitor local air pollution and train a short-term machine-learning (ML) predictor that, using a seven-day window, estimated the next 24 hours of pollution levels. We focused the prototype on schools so playground schedules could be adapted to air-quality forecasts.

The week was planned around teamwork as the key to understanding and developing the project. In parallel, informal conversations with professionals from Openchip and BSC provided valuable insight into the semiconductor industry and eventually led me to an interview for a summer internship at Openchip, where I joined the hardware-architecture team.

At Openchip I moved from the ML project to instruction-level work. My main task was to add a matrix multiply-accumulate operation into Gem5 RISC-V instruction set architecture (ISA) decoder so it could be issued and exercised as a vector instruction. This required implementing the instruction in the simulator, creating a 'golden' model for verification, writing inline-assembly, unit tests and benchmarking the new vector instruction against a reference scalar implementation to characterize its performance. The development process was collaborative, guided by mentors and colleagues, and validated through testing and simulation.

Both experiences reinforced my decision to pursue a career in the semiconductor and high-performance computing (HPC) fields. For students, initiatives like HiPEAC hackathons are invaluable: working in teams on real problems, meeting industry professionals and gaining experience can make it much easier to decide where to focus future studies and career goals.

How the HiPEAC STEM Day helped cement my internship

Ezgi Sena Karabacak, Universitat Politècnica de Catalunya-Barcelona Tech

I first heard about the STEM Student Day when I encountered the HiPEAC conference. I enrolled in both and looked forward to a day dedicated to connecting students with leading technology companies.

After an introduction to the STEM Student Day, we were grouped into teams with students from all over the world and rotated through discussions with various companies. The diversity and quality of the companies were far beyond my expectations. I had the chance to meet professionals from companies such as Huawei, Google, ASML, and many other technology leaders.

Most importantly, I connected with Semidynamics, a company I was already aware of but had not previously had the chance to explore in depth. The conversation quickly developed into engaging discussions with several team members. One team leader asked technical questions, both to capture our attention and to assess our problem-solving approach. This interaction left a strong impression on me.

After the conference, I discovered that Semidynamics were offering internship opportunities. I decided to apply and joined a team for a three-month internship. During this time, I mostly worked within the verification team, contributing to tasks focused on ensuring the hardware design was correct and reliable prior to tape-out. It was very inspiring to see people from different teams working together, shaping the future of chip design.

I am extremely grateful for the STEM Student Day experience. It opened the door to a valuable learning opportunity and helped shape my professional journey. I am sure the connections I made and the things I learned will continue guiding my career.

In this edition of our thesis series, Veronia Iskandar explains how she developed tools to make near-memory computing more accessible, targeting field-programmable gate arrays (FPGAs) with high-bandwidth memory (HBM).

Three-minute thesis

NAME: Veronia Iskandar
RESEARCH CENTRES: Chair of Adaptive Dynamic Systems, Institute of Computer Engineering, Technische Universität Dresden
SUPERVISORS: Diana Göhringer
THESIS TITLE: Architectures and Frameworks for Near-Memory Computing using 3D-Stacked Memories

Modern computing systems are increasingly driven by data-intensive applications such as machine learning, data analytics, signal processing, and scientific computing. Recent advances in artificial intelligence, especially transformer-based models, have further increased the demand for fast access to massive volumes of data. However, while processors continue to improve, memory systems have become a critical bottleneck. This growing imbalance, commonly referred to as the memory wall and bandwidth wall, limits the performance and energy efficiency of modern systems and motivates the need for new computing paradigms.

This PhD research addresses these challenges by exploring near-memory computing (NMC), an architectural approach that moves computation closer to where data is stored. In particular, the thesis focuses on FPGA platforms integrated with HBM, which offer enormous memory bandwidth but also introduce new design and programming complexities. An intuitive analogy is a busy kitchen: instead of repeatedly walking ingredients across the room to a single workstation, tools and chefs are placed next to storage areas, drastically reducing unnecessary movement and improving efficiency.

A major challenge of this research was the lack of accurate and fast simulation tools for FPGA-HBM systems. Existing tools often model memory systems too generically or fail to capture the specific interconnect structures of HBM platforms. To overcome this, a system-level simulation framework was developed that accurately models HBM-based FPGA architectures. This simulator enables early design space exploration, allowing architects to evaluate performance trade-offs without costly and time-consuming hardware implementations.

Another challenge lies in identifying which parts of an application benefit from near-memory acceleration. Developers may lack deep hardware knowledge, making it difficult to reason about memory behavior and bandwidth utilization. To address this, the thesis introduces automatic application characterization and performance prediction techniques using machine learning. By learning from architectural features and kernel properties, these models estimate performance on HBM-based systems and guide kernel selection without extensive profiling.

Once suitable kernels are identified, they must be efficiently mapped to hardware. This work proposes a compiler-based framework that detects linear algebra kernels in legacy code and offloads them to FPGA accelerators. On the hardware side, a novel FPGA architecture was designed to support both dense and sparse matrix-vector multiplication, exploiting HBM bandwidth to overcome memory bottlenecks. Experimental results demonstrate improved performance and bandwidth efficiency compared to central processing units (CPUs), graphics processing units (GPUs), and prior FPGA solutions.

The originality of this research lies in its holistic, system-level approach. Rather than optimizing a single accelerator or application, the thesis integrates simulation, compiler support, accelerator design, and workload mapping into a unified framework. An optimization framework using meta-heuristic algorithms further enables efficient mapping of multi-kernel and multi-application workloads onto FPGA-HBM platforms.

Overall, this work provides practical tools and methodologies that make near-memory computing accessible. By reducing data movement and improving bandwidth utilization, the proposed solutions contribute to scalable computing systems that meet the demands of next-generation applications.

Veronia's supervisor **Diana Göhringer** commented: 'Veronia's thesis addresses the growing memory and bandwidth requirements of data-intensive applications, such as artificial intelligence, by providing a holistic framework for near-memory computing systems based on modern FPGAs with HBM. Her hardware-software-framework consists of a simulator, compiler support for identifying and mapping linear algebra kernels to FPGA-HBM accelerators, and a hardware accelerator for executing dense and sparse matrix-vector multiplications on FPGA-HBM systems.'

HiPEAC

Thanks to all our sponsors for making #HiPEAC26 such a success!

CYFRONET
arm
HUAWEI
STREAM HPC
ABB
AMD
ASML
Collins Aerospace An RTX Business
doitnow
ECLIPSE FOUNDATION
Google
HPE
POLYMAGE LABS
SPECTRUM
BZL BARCELONA ZETTASCALE LAB
CODECO
EDGELESS
ETP4HPC SHAPING THE FUTURE OF EUROPEAN ADVANCED COMPUTING
ECSPEC
EVIDEN an atos business
Springer
THALES
FONDAZIONE links PASSION FOR INNOVATION
NET4EXA

Sponsors correct at time of going to print. For the full list, see https://www.hipeac.net/2026/krakow/#/

Join the community @hipeac hipeac.net/linkedin YouTube hipeac.net/tv hipeac.net

Funded by the European Union

The HiPEAC project has received funding from the European Union's Horizon Europe research and innovation funding programme under grant agreement number 101069836. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.