Spiker+ Simplifying custom hardware for spiking neural networks

HiPEAC info 75

JUNE 2025

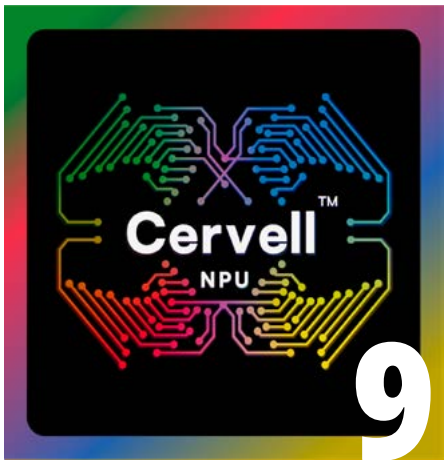
ACACES 2025 Fiuggi

Tools of the trade: Enabling technologies for next-gen computing systems

How DARE is spearheading digital autonomy in Europe

The HiPEAC companies energizing the European deeptech scene

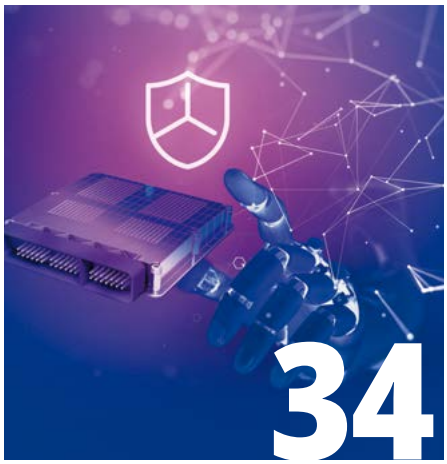
Contents



How DARE is taking EU sovereignty ambitions to a new level

Ecosystem developments and product launches

Tools of the trade: The latest tools for computing systems



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 / @hipeacjobs

hipeac.net/linkedin 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: InfiniteFlow on Adobe Stock

Design: www.magelaan.be

Editor: Madeleine Gray

The theme of this magazine is tools. Given the increasing complexity of modern digital systems, powerful tools are essential to design them, operate them, secure them, etc. These days, there are clear signs that artificial intelligence (AI) is going to change the tools world. I became aware of this when I learned that an AI trained in translating between natural languages can also be trained to translate between programming languages. The result was not always perfect, but it turned out to perform better than expected, and it worked in both directions. Since then, many more experiments have been done, not only in compilation, debugging, performance optimization but also in electronic design automation (EDA) tools.

Experts expect that the industry will demonstrate artificial general intelligence (AGI) before the end of this decade. It is not easy to predict what will happen when tools are able to improve themselves without human intervention. This might turn out to be the beginning of a new era in computing, one in which the role of computer scientists is going to change. If an AI can train itself to be the best chess player in the world in just a couple of hours today, why wouldn't it be able to develop and improve advanced tools by the end of this decade? Who needs a compiler expert if AI builds compilers cheaper, better and faster than humans can?

We are entering an era in which humans will no longer compete with other humans, or companies with other companies, but one in which AIs will compete with AIs, and the winner will probably be the one with the most compute power. The fact that the United Arab Emirates recently committed to investing $1.4 trillion in AI, and that it will build a 25 km^2 AI campus and 5GW-capacity AI data centres, shows what is at stake.

The question is what Europe will do to stay relevant. We have neither the money nor the energy to compete with such investments, which means that we will have to be creative to compete. To help the European computing community deal with this new reality, we will submit a project proposal to continue HiPEAC until 2029. One of the goals of that project will be to help the HiPEAC community to transition to the AGI era. I am very committed to continuing the coordination of HiPEAC.

Koen De Bosschere, HiPEAC coordinator

Part of the European Union's strategy to reach European autonomy in strategic hardware technologies, the DARE (DARE (Digital Autonomy with RISC-V in Europe) project is a major new research and innovation effort in chips for high-performance computing (HPC) and artificial intelligence (AI). Bringing together 38 partners from 15 different countries across Europe, the project will develop chiplets for HPC / AI, designed and implemented in Europe. HiPEAC caught up with DARE's principal investigator, Osman Ünsal, and Vanessa Iglesias, EU projects innovation manager (both Barcelona Supercomputing Center) to find out more.

'DARE takes the ambition for self-sufficiency in software-hardware technology to a new level'

Following a tumultuous period of intensifying geopolitical competition – with numerous reports, from Draghi to Heitor, casting a critical eye over the state of European competitiveness – Europe's need for technological autonomy has come into ever-clearer focus. 'It has become increasingly apparent that Europe's dependence on external supply chains for critical technology such as semiconductors poses a series of risks,' Vanessa Iglesias, EU projects innovation manager at BSC, notes. 'These risks range from restricted access to the latest generation of technologies to consequences for cybersecurity, data protection, trade secrets, and data ownership, for example.'

The quest for European technological sovereignty is not new, however. As John Goodacre (UKRI) and Filippo Mantovani (BSC) noted in their respective interviews for *HiPEACinfo* 74, prior to 2016's acquisition of Arm by Softbank, European efforts focused on Arm for technological sovereignty. In the domain of HPC, the Mont-Blanc project, which started in 2011, successfully demonstrated that Arm cores could be used for HPC. Almost a decade after Mont-Blanc began, the Fugaku supercomputer in Japan provided real-life proof of the viability of this approach.

Members of the DARE consortium at the kickoff meeting in Barcelona, March 2025

Complementing work undertaken on Arm architectures, RISC-V emerged as an attractive route for sovereign European technology development, given its open-source nature. Multiple projects have been funded by the European High-Performance Computing Joint Undertaking (EuroHPC JU) to create a designed-in-Europe processor and accelerators, and to build out the ecosystems underpinning the deployment of this hardware, including the European Processor Initiative (EPI), EUPILOT, eProcessor, EUPEX, MEEP, and DEEP-SEA.

With €120 million in funding from the EuroHPC JU, an amount which will be matched by co-funding provided by participating states, DARE is a major initiative to further advance made-in-Europe hardware for HPC and AI. 'DARE is the continuation of an effort for Europe to be self-sufficient in software-hardware technology from top to bottom that started many years ago,' explains Osman Ünsal, manager of the Computer Architecture for Parallel Paradigms at BSC and the principal investigator of DARE. 'Building on work undertaken in previous projects, DARE takes the ambition for self-sufficiency to a new level.'

Chiplet design

One of the things which marks out DARE from its predecessors is the incorporation of chiplet research. 'There have been chiplet research projects in the past, but DARE is novel thanks to its combination of RISC-V and a go-to-fab agenda,' says Osman. The project will develop chiplets for HPC / AI, designed and implemented in Europe.

While NVIDIA is the acknowledged leader in semiconductors for high-performance computing (HPC), the RISC-V HPC ecosystem is being built out and momentum is growing, says

Osman. 'The strength of RISC-V for Europe is that, as an open instruction set architecture (ISA), it promotes inclusivity: anyone can come in and play.'

The project aims to deliver, in Osman's words, 'the first mix-and-match chiplet design with designed-in-Europe technology'. Participants in DARE will design, develop and tape out three industry-standard RISC-V-based chiplets, to be fabricated at advanced technology nodes (TSMC 7nm or below). The design of each of these chiplets will be led by an industry partner, as follows:

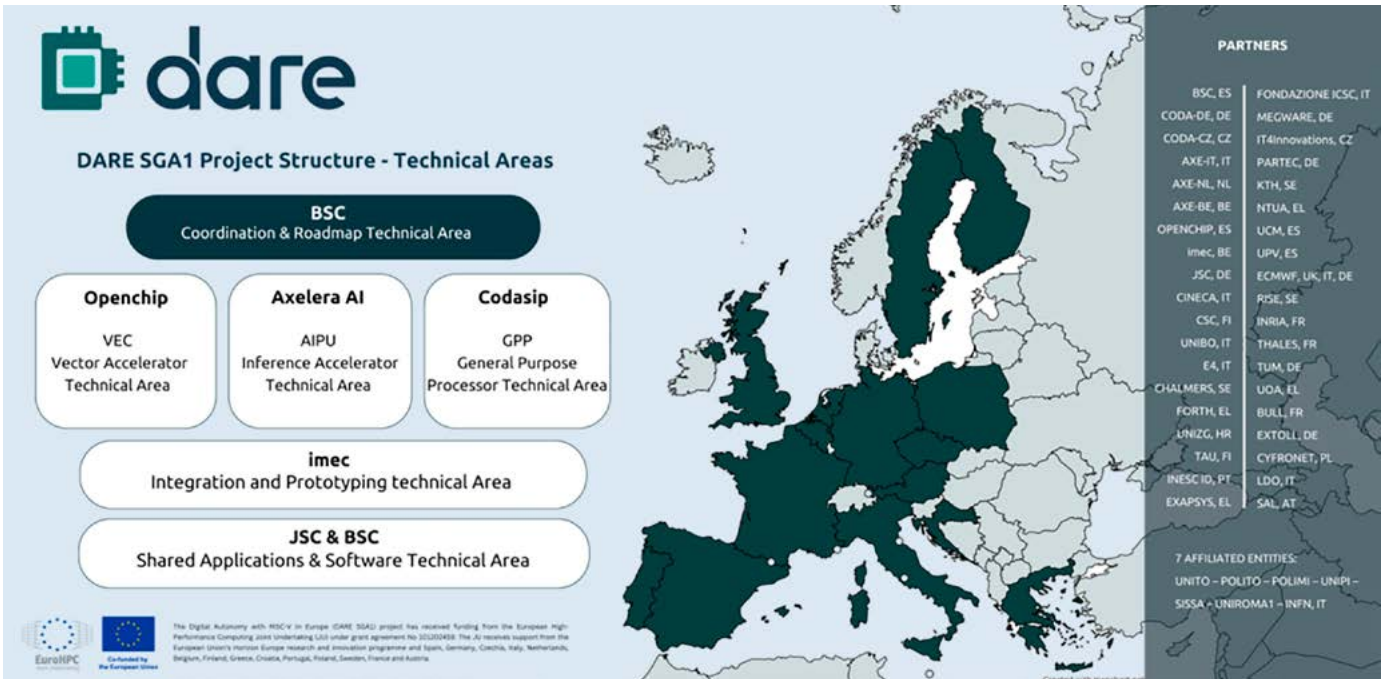
- A general-purpose processor, optimized for priority workloads for Europe, including in the areas of health, climate modelling, and energy solutions. This work will be led by processor-solutions provider Codasip, which is headquartered in Munich with design centres in Brno, Prague, Villeneuve-Loubet, Munich, Heraklion, Thessaloniki, Marousi, Barcelona, Bristol, Cambridge, and London.
- A vector accelerator for high-precision HPC tasks and applications merging HPC and AI. This will be led by Openchip, which is headquartered in Barcelona, with offices in Gdańsk, Rome, and Ghent. 'Unlike graphics processing units (GPUs), this accelerator can self-host – it doesn't need a core central processing unit (CPU) to boot the operating system and so forth,' says Osman. 'It is a more general processor than a GPU and, as such, will require less effort on the part of HPC domain experts to modify their applications.'
- An AI processing unit specifically designed for accelerating AI inference within HPC applications. This work will be led by Axelera AI, whose chief executive is HiPEAC member Fabrizio Del Maffeo and whose chief technology officer is HiPEAC member Evangelos Eleftheriou. Axelera's advisory board includes HiPEAC members Luca Benini and Marian Verhelst, and its board of directors includes HiPEAC member Massimo Vanzi. Axelera AI is headquartered in Eindoven, with offices in Leuven, Zürich, Bristol and Milan and operations in 15 European countries.

The project will also develop a full HPC and AI software stack (compilers, libraries, operating system, tools), optimized for DARE hardware, to support applications which are viewed as a priority for Europe – for example in the areas of energy, health and climate modelling. By using a hardware-software co-design methodology – something long championed by Osman – DARE will ensure that the hardware is exploited to the maximum while ensuring programmability. Jülich Supercomputing Centre and BSC will act as technical leads for shared applications and software, while imec will serve as technical lead for integration and prototyping.

The development of these processors and accelerators, along with their associated software, aims to strengthen the European HPC ecosystem and define a roadmap for future HPC systems based on made-in-Europe technology. This goes to the heart of the project's aim to address Europe's deficit in digital autonomy.

Technical challenges

The project team is closely involved with the development of the RISC-V standard, for example by participating in the working groups on security, vectorization and fault tolerance. 'The whole DARE landscape is standardized; for example, partners are working on error reporting with the new standard on fault



tolerance,' Osman notes. 'Others are working on the vector standard, and the RISC-V vector C intrinsics has recently been approved.' With members of the DARE team, such as Roger Ferrer at BSC, having already worked on this as part of the EPI, the project is therefore helping to shape the standards, not just following them. 'Because the standard is open, everyone can see these updates immediately and can collaborate or compete on an equal footing,' Osman adds. 'We need to be champions of the RISC-V standard, and it is crucial that Europe continues to maintain it in future.'

Along the way, DARE will need to overcome a number of technical challenges. As always, delivering high performance on a low power budget is paramount for ever-more demanding HPC and AI systems. The project team will also have to tackle chiplet integration while building out an emerging software ecosystem. They will have to address advanced node manufacturing and tape-out, an area in which Europe has less experience than other world regions, while keeping pace with the frantic pace of AI.

Coordinating partners and IP

With a consortium of 38 partners and seven affiliated organizations, DARE clearly poses enormous challenges on the project management front as well as the technical side. Yet there are advantages to a large, geographically distributed consortium in a major EU project, according to Osman: 'Having so many partners in the project builds resilience into the project. It also allows the knowledge to be spread across different European countries: the project facilitates knowledge exchange and teams can grow together, propagating what they have learned in their local institutions and ecosystems.'

'Coordinating a project of this scale, with projects across different countries, sectors and areas of expertise, is inherently challenging,' adds Vanessa Iglesias. 'However, the DARE consortium involves partners who have worked together on previous projects, which facilitates communication and cooperation. Moreover, governance structures were put in place from the outset, including clear work packages for management and communication. It is also worth mentioning that BSC has extensive experience coordinating projects with large consortia, with a well-qualified, experienced team.'

Osman Ünsal presenting DARE at ISC 2025

The ambition to create made-in-Europe technology and build a successful chip industry also poses the question of how open the research should be: is sharing research and roadmaps openly the way forward, or should Europe start guarding its intellectual property (IP) more closely? 'Commercialization in DARE is being driven by the three chiplet owner companies, while the startups and small / medium enterprises will develop their own IPs for future phases,' says Osman.

'We take the European Commission's mantra "as open as possible, as closed as necessary" as our starting point. From the early stages of the project, knowledge and IP issues will be managed by the innovation managers, who will identify and monitor, as early as possible, results that have the potential to be exploited and who will support their owners to define strategies to protect and make concrete use of these results,' adds Vanessa. 'DARE will follow two approaches: one that is primarily industry-oriented, with a strong focus on commercialization, and the other taking a more exploratory path, embracing higher-risk research and innovation.'

DARE's legacy

What kind of impact will DARE have in the long run? The main goal is for Europe to shift from being a consumer to a producer of HPC / AI technology, according to Osman. 'In 10 years' time, I'd like to see a RISC-V system based on European technology among the top three or four players in the top 500 lists – whether HPC or AI,' he says. Yet the collaboration between European partners is also a key part of this effort. 'Beyond the concrete results, I would like DARE to prove the success of a federated approach to building European technology, demonstrating how bringing universities, research centres and industry players can work together to achieve our technical sovereignty.'

Along the way, the project will build much-needed capacity in European design-to-fab expertise, verification skills, and in-depth knowledge of state-of-the-art architectures, building on work started in EPI.

FURTHER INFORMATION:

DARE website dare-riscv.eu

HiPEAC at the Open Source AI Strategy Forum

Gaël Blondelle and Ovidiu Vermesan at GOSIM AI Paris 2025

On 5 May, HiPEAC participated in the Open Source AI Strategy Forum workshop, which was co-located with the GOSIM - Global Open-Source Innovation Meetup AI Paris 2025 Conference.

Representing HiPEAC were Gaël Blondelle, VP of the Eclipse Foundation (workshop organizer) and Ovidiu Vermesan, chief scientist at SINTEF, who gave a presentation titled 'Empowering the Edge. Open-Source Edge AI Accelerating Research and Innovation'. Gaël and Ovidiu both participated in the panel discussions in the 'Open-Source AI and Research' session, as did Pierre Gaillard of HiPEAC partner CEA.

The workshop brought together thought leaders, researchers, and pioneers from the open-source AI community around the world, who engaged in dynamic discussion on the challenges and opportunities in the open-source AI domain.

Key themes explored

During the workshop, participants exchanged perspectives and debated strategies around several critical themes shaping the future of open-source AI. These included:

- **AI models and innovation**: The importance of transparency, reproducibility, and open accessibility in fostering trustworthy AI model development.
- **Openness and collaboration**: How to define and implement practical standards and best practices for thriving open AI ecosystems.
- **Sustainability and resources**: Key considerations for securing long-term viability through sustainable funding, robust infrastructure, and responsible energy consumption.
- **Governance and ethics**: Navigating urgent regulatory questions and ethical responsibilities, and establishing robust trust mechanisms within open AI development.
- **Sovereignty and strategic autonomy**: Striking a balance between promoting global cooperation and maintaining national or regional sovereignty in AI infrastructure and policy-making.

A significant outcome of the workshop is the ongoing development of a strategic white paper. This document will synthesize the event's discussions, highlight actionable recommendations, and summarize the key points of consensus and debate. The white paper aims to serve as a reference and roadmap for stakeholders in open-source AI.

Gaël Blondelle presenting at GOSIM AI Paris 2025

Ovidiu Vermesan (far right) participating in panel discussions at the workshop

ECLIPSE white paper on open source in Europe

In September 2024, the Eclipse Foundation's Gaël Blondelle and Philippe Krief chaired a consultation on open source as part of their work in the HiPEAC project. At this event, stakeholders from across Europe convened to discuss critical challenges and opportunities related to open source in the European technology ecosystem.

Eclipse has now released a white paper, titled 'The Vital Role of Open Source in Europe', which summarizes the main findings of the meeting. The white paper covers six topics, as detailed below.

Open source:
- for EU sovereignty and competitiveness
- and artificial intelligence (AI)
- and the new regulations on hardware
- and the emerging fourth sector
- in Europe amid geopolitical tensions
- and the skill shortage

The white paper concludes by noting that the European Union needs to find a path that balances government leadership, commercial innovation and collaborative technological development. While acknowledging that open source is not a silver bullet, the paper states that it is a strategic instrument of digital sovereignty, reducing dependence on foreign technologies, promoting interoperability and fostering collaboration across sectors by supporting ethical, inclusive and competitive digital development.

DISCOVER-US sets out vision for distributed computing and swarm intelligence

Integrating expert contributions from leading researchers in the EU and the US, the DISCOVER-US vision paper provides an in-depth exploration of future distributed computing systems, swarm intelligence, and edge-intelligent things.

The vision paper focuses on the four main research areas of the project, each of which has a dedicated chapter exploring challenges, trends, the state of the art and future research recommendations.

These four focus areas are:

- **Managing complexity through high levels of abstraction**
  High levels of abstraction support the computation process and allow the design and implementation of systems that can address intricate scenarios efficiently.
- **Distributed computing, swarm intelligence (SI), and edge-intelligent things**
  In distributed computing models, resources are shared seamlessly across the computing continuum, from cloud infrastructures to IoT devices, enhancing the performance and agility of applications.
- **AI-based concepts for self-organized, dynamic, and adaptive management**
  AI's capacity for self-organization and adaptability and the integration of swarm intelligence leads to innovative management strategies essential for optimizing dynamic systems across the computing continuum.
- **Collaborative programming frameworks and software development tools**
  The development of these tools fosters an environment where developers can innovate rapidly and collaboratively, enabling them to produce robust applications that can operate in the increasingly interconnected distributed computing ecosystem.

The paper reflects EU and US researchers' collective knowledge, experience, and insights, offering a framework for building a robust transatlantic research ecosystem in distributed computing and swarm intelligence.

discover-us.eu/vision

TechNexus white paper on complex and critical systems in Europe

The TechNexus Programme has published a white paper setting out key insights on the development of complex and critical systems in Europe. Drawing on discussions at workshops at the 2025 HiPEAC conference, the paper considers aspects such as integrating innovations while ensuring dependability, building on European strengths and shoring up sovereignty, and rethinking European projects for maximum impact.

Titled 'Bridging Domains in Large-scale Complex and Critical Systems', the TechNexus Programme white paper includes actionable recommendations from the FORECAST and STEADINESS workshops at HiPEAC 2025.

Key takeaways from the white paper include:

- The challenges associated with increasing complexity of technology integrations into mission-critical applications.
- The need for robust methodologies to ensure system safety, security and interoperability.
- The need for innovative validation and verification techniques to ensure functional dependability in rapidly evolving application domains.
- How the European regulatory environment could be turned into a competitive advantage for cyber-physical systems.
- The need to manage overarching system complexity while taking into account the autonomy of individual systems.
- The need for a clearer distinction between basic research and innovation impact projects.

The white paper's authors are Spyros Lalis, Danh Le Phuoc, Alois Zoitl, Octavian Fratu, and Charles Robinson. This work is supported by the following projects: MLSysOps, ARROWHEAD, SMARTY, ENERGY ECS, SmartEdge, ISOLDE and HiPEAC.

FURTHER INFORMATION:

Robinson, C., Lalis, S., Le Phuoc, D., Zoitl, A., & Fratu, O. (2025). 'Bridging Domains in Large-scale Complex and Critical Systems', 2025 Ed. HiPEAC 2025, Barcelona. Zenodo doi.org/10.5281/zenodo.14920027

Semidynamics announces Cervell™ RISC-V NPU

Semidynamics, a company providing customizable RISC-V processor intellectual property (IP), has announced Cervell™, a scalable and programmable neural processing unit (NPU) built on RISC-V. Cervell combines central processing unit (CPU), vector, and tensor capabilities in a unified architecture, enabling low-latency computations for artificial intelligence (AI) across applications from edge AI to large language models (LLMs).

Delivering up to 256 TOPS (tera operations per second) at 2GHz, Cervell scales from C8 to C64 configurations, allowing designers to tune performance to application needs, from 8 TOPS INT8 at 1GHz in compact edge deployments to 256 TOPS INT4 in high-end AI inference.

HiPEAC member Roger Espasa, chief executive of Semidynamics, commented: 'As an NPU, [Cervell] delivers the scalable performance needed for everything from edge inference to large language models. But what really sets it apart is how it's built: fully programmable, with no lock-in thanks to the open RISC-V ISA, and deeply customizable down to the instruction level. Combined with our Gazillion Misses™ memory subsystem, Cervell removes traditional data bottlenecks and gives chip designers a powerful foundation to build differentiated, high-performance AI solutions.'

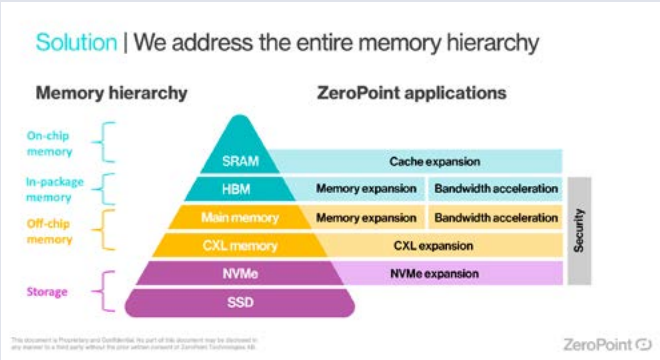
In their announcement, Semidynamics noted that Cervell NPUs are purpose-built to accelerate matrix-heavy operations, enabling higher throughput, lower power consumption, and real-time response. By integrating NPU capabilities with standard CPU and vector processing in a unified architecture, designers can eliminate latency and maximize performance across diverse AI tasks, from recommendation systems to deep-learning pipelines.

bit.ly/Semidynamics_Cervell_announcement

ZeroPoint announces AI-MX, a solution to increase foundation model addressable memory by 50%

ZeroPoint Technologies has announced AI-MX, a hardware-accelerated memory optimization product that enables nearly instantaneous compression and decompression of foundation models, including leading large language models (LLMs).

According to ZeroPoint, this product, which will be available in the second half of 2025, will enable datacentre operators to realize a 1.5x increase in addressable memory, memory bandwidth, and tokens served per second for applications that rely on large foundational models. AI-MX works across a wide variety of memory types, ensuring that the memory optimization benefits apply to nearly every possible artificial intelligence (AI) acceleration use case.



'Foundational models are stretching the limits of even the most sophisticated datacentre infrastructures. Demand for memory capacity, power, and bandwidth continues to expand quarter-upon-quarter,' said Klas Moreau, chief executive of ZeroPoint Technologies. 'With the announcement of AI-MX, we introduce a first-of-its-kind memory optimization solution that has the potential to save companies billions of dollars per year related to building and operating large-scale datacentres for AI applications.'

In April, ZeroPoint also announced a strategic partnership with AI chip company Rebellions to develop the next generation of memory-optimized AI accelerator solutions, with a focus on delivering increased performance for inference workloads.

ZeroPoint's founders are HiPEAC founding member Per Stenstrom and HiPEAC member Angelos Arelakis. The company was launched in 2016 and is headquartered in Gothenburg, Sweden.
Full technical specifications of AI-MX bit.ly/ZeroPoint_AI-MX_spec

Daisytuner launches in beta following pre-seed funding round

The team behind Daisytuner, the winner of a HiPEAC Technology Transfer Award, have announced the launch of the startup's continuous benchmarking solution following a successful pre-seed funding round. Founded by Lukas Trümper and Adrian Schmitz, Daisytuner empowers developers to automatically analyse the performance and bottlenecks of a software application directly from GitHub.

The platform is the first step towards the company's ambitious goal: making AI-assisted coding applicable to performance-critical codes and large-scale HPC applications. Currently, coding tools such as GitHub's CoPilot and ChatGPT lack the necessary performance data about an application in their reasoning context. Daisytuner's platform fills the missing gap, as it supports benchmarking across diverse architectures, including x86, Arm, graphics processing units (GPUs), and the collection of a variety of relevant information such as profiler data and hardware utilization. As such, it will allow engineers to port high-performance applications much faster to a wide range of chips, including those developed by European companies such as Axelera AI.

'Tools liks GitHub Copilot and ChatGPT don't understand what happens when code hits real hardware. That's where things break – memory bottlenecks, CPU stalls, low GPU occupancy – all invisible from just reading code. To tune performance, you need profilers and hardware metrics, and you need to iterate based on the data,' explains Lukas, a HiPEAC member. 'Integrating profiler and compiler output into AI coding copilots is essential for their success.'

The launch of the beta site follows a successful €1 million pre-seed funding round led by LEA Partners, which was joined by Angel Invest, Jens Lapinski, Christian Stiebner, Sara Schneider and Elias Schneider, Hans-Juergen Schmidtke, Jessica Holzbach and Mario Götze.

daisytuner.com

Mosaic SoC awarded CHF 150,000 in funding from VentureKick

Mosaic SoC, the company founded by HiPEAC students Moritz Scherer and Alfio Di Mauro (see *HiPEACinfo* 73 p.43), has secured CHF 150,000 in funding from the VentureKick acceration fund. The funding aims to accelerate the development of the Mosaic SoC microcontroller architecture, which is designed to eliminate critical memory bottlenecks in real-time data processing.

As VentureKick noted in their announcement of the funding, augmented reality (AR), virtual reality (VR), robotics, and autonomous systems increasingly rely on instantaneous environmental mapping to enable seamless interactions with the physical world. Achieving these capabilities demands that microcontrollers rapidly process immense volumes of data. Conventional microcontrollers, even those equipped with neural processing units (NPUs), typically fall short due to memory constraints and limited processing speeds, severely impacting overall application performance.

Addressing these challenges, Mosaic SoC's innovative microcontroller architecture fundamentally redefines how accelerators access memory. Traditional chip designs require processors and accelerators to share data pathways to memory, severely restricting performance due to sequential data access and limited bandwidth. Mosaic SoC disrupts this model by introducing a pioneering memory architecture that allows multiple accelerators to simultaneously read from and write to memory at high throughput without interference. This design significantly enhances data processing speed, improves energy efficiency, and boosts overall system performance.

'The additional funding provided by Venture Kick will help our transition from academic research and technology development into creating market-driven solutions that improve our customer's edge processing pipelines on all levels between applications and hardware,' said Moritz, the company's chief technology officer.

mosaic-soc.com

Openchip continues European expansion with Ghent office opening

In March, Openchip announced the opening of an office in the Wintercircus at Ghent, in conjunction with a strategic partnership with imec. The young company, which is integrating system-on-chips (SoCs), designed in-house, and full-stack software for artificial intelligence (AI) and high-performance computing (HPC) with plans for a chip by 2027, is headquartered in Barcelona, and has offices in Gdańsk, and Rome in addition to the Ghent office, as well as operations in Ireland.

Steven Latré, formerly the head of AI at imec, has been appointed as head of AI at Openchip. He commented: 'We want to develop faster, better, and safer artificial intelligence – to compete with American and Chinese tech giants. Today, AI has a major societal impact, which will only continue to grow. So we need a strong European AI ecosystem, now more than ever.

Commenting on the Ghent office opening and partnership with imec, Openchip Chief Executive Francesc Guim said: 'This partnership strengthens our ability to develop next-generation AI solutions, and Steven's expertise will be instrumental in shaping our future in AI and semiconductor technology.'

Openchip also recently announced a strategic partnership with Kalray, a company based in Grenoble that develops high-performance, low-power data processing unit (DPU) solutions.

TASKING acquires LDRA to boost safety-critical toolchain for embedded development

Florian Süssmair, TASKING

TASKING has acquired LDRA, a long-standing provider of static and dynamic code analysis tools, in a move that significantly enhances its offering for engineers developing safety- and mission-critical embedded systems.

LDRA's tools are widely used in regulated industries to enforce coding standards, ensure traceability, and support certification workflows. With a strong user base across the aerospace, automotive, industrial, and defence sectors, LDRA brings deep expertise in functional safety and security compliance – key challenges in modern embedded development.

For software engineers, the acquisition means broader access to integrated tooling that supports the entire development lifecycle: from early requirements tracing and static analysis to unit testing and system-level verification. TASKING's compiler technology and LDRA's testing capabilities are highly complementary, especially for teams working on multicore platforms and high-assurance applications.

The two companies have a long history of technical cooperation. Now, as part of the same organization, they aim to streamline toolchains and provide a more cohesive workflow for developers working under stringent safety standards such as ISO 26262, DO-178C, and IEC 61508.

This acquisition continues TASKING's strategy of expanding its ecosystem to support the growing complexity of embedded systems development – while reinforcing its position as a trusted partner for engineers building tomorrow's safety-critical software.

Launch of DARE SGA1 marks major step towards digital autonomy in Europe

Romana Konjevod, Barcelona Supercomputing Center (BSC)

In March, the Digital Autonomy with RISC-V in Europe, Special Grant Agreement 1 (DARE SGA1) project was launched, with the aim of strengthening Europe's sovereignty in high-performance computing (HPC) and artificial intelligence (AI). Supported by the EuroHPC Joint Undertaking, and coordinated by Barcelona Supercomputing Center (BSC-CNS), DARE SGA1 unites 38 leading partners from across Europe to develop next-generation European processors and computing systems, including an optimized software ecosystem, designed for research and industry applications.

With a budget of €240 million, this ambitious three-year project marks the first phase of a six-year DARE initiative. DARE SGA1 is set to build a fully European supercomputing hardware / software (HW / SW) stack for HPC and AI, featuring high-performance and energy-efficient processors designed and developed in Europe. The initiative is a direct response to Europe's strategic need for digital sovereignty, ensuring that the continent has full control over its critical computing infrastructure.

Commenting on the launch, EuroHPC JU Executive Director Anders Jensen stated: 'I am proud to announce the launch of the DARE project which marks a significant milestone for European digital sovereignty. This ambitious initiative will drive innovation in both hardware and software technologies and leverage the full power of HPC and AI to develop secure, efficient, and European-led solutions for the future.'

The project is coordinated by BSC, and the companies Openchip, Axelera AI and Codasip will lead work to develop three RISC-V-based chiplets – a vector accelerator (VEC) for high-precision HPC and emerging applications, an AI processing unit (AIPU) for AI inference acceleration in HPC applications, and a general-purpose processor optimized for HPC workloads in European supercomputers. Imec and Jülich Supercomputing Center will serve as technical leads

dare-riscv.eu/

Czech organizations integral to the DARE SGA1 push for homegrown European processors

Markéta Dobiašová, IT4Innovations

The next few years will see the development of a fully European hardware and software ecosystem for supercomputing and artificial intelligence, with high-performance, energy-efficient processors and accelerators designed and developed in Europe. One of the key steps in this effort is the DARE project, the aim of which is to develop three different chiplets based on the RISC-V architecture. These chiplets are smaller specialized blocks that can then be used to build highly complex, powerful processors for next-generation supercomputers. Chiplets overcome the limitations of traditional monolithic chips through greater efficiency, scalability, and lower cost.

Organizations from the Czech Republic play a key role in the development of this ecosystem. The Brno-based company Codasip will lead the development of a universal processor designed for highly complex computational tasks. Meanwhile, scientists from IT4Innovations National Supercomputing Center at VSB – Technical University of Ostrava will also be involved in its development. 'Primarily, we will be optimizing power consumption and thermal management using the MERIC suite software, which we have been developing for ten years. MERIC will also be used to monitor power consumption and debug hardware parameters of DARE chiplets,' explain Lubomír Říha from the Infrastructure Research Lab, principal investigator of the project at IT4Innovations, and Ondřej Vysocký, the head of the MERIC suite development.

'The Czech Republic's participation in this activity places us among the European technological leaders in IT. At the same time, we will have access to cutting-edge technologies already at the stage of their development, which will give us a certain technological advantage,' adds Vít Vondrák, managing director of IT4Innovations.

The DARE SGA1 project has received funding from the European High-Performance Computing Joint Undertaking (EuroHPC JU) under grant agreement no. 101202459. The JU receives support from the European Union's Horizon Europe research and innovation programme and Spain, Germany, Czechia, Italy, Netherlands, Belgium, Finland, Greece, Croatia, Portugal, Poland, Sweden, France and Austria. 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.

O-CEI and COP-PILOT to deliver cloud-edge-IoT platforms

Giulietta Ravera and Maria Giuffrida, Trust-IT

January 2025 saw the launch of two large-scale pilots, O-CEI and COP-PILOT, both funded by the European Union (EU), that aim to increase European competitiveness in cloud-edge-IoT platform solutions. The two major projects are supported by the coordination and support action CEI-Sphere.

Building on the foundational efforts of the EU-funded EUCloudEdgeIoT initiative, the large-scale pilots will focus on internet-of-things (IoT) and decentralized edge intelligence, with the goal of facilitating the shift to edge computing as an alternative to centralized cloud services. In the process, they will empower European stakeholders to play a leading role in the data economy, contributing to Europe's strategic autonomy and competitiveness.

The projects are also intended to accelerate the adoption and integration of clean-energy technologies, such as electric-vehicle (EV) chargers, heat pumps, solar panels and residential batteries, through open IoT platforms and European CEI solutions. Partners of the pilots will collectively develop solutions to link energy and European transportation. These solutions could be used in the sectors of logistics and agrifood systems, for example.

O-CEI

With EU funding of over €23 million and a total budget of nearly €28 million, O-CEI brings together 58 partners from 20 countries. It aims to develop an open, interoperable, and sustainable platform, addressing key challenges in decentralized CEI networks such as rising energy consumption, costs and carbon emissions.

O-CEI's eight large-scale pilots include optimizing electricity grids with renewable energy integration, advancing software-defined vehicles, enhancing smart charging for electric postal fleets, and improving energy management in maritime ports.

Project representatives at the 'Advancing Cross-Domain Standardisation for IoT and Edge and Edge Computing' workshop, held in Brussels in November 2024

COP-PILOT

Meanwhile, COP-PILOT, which will receive around €22,500,000 in EU funding and has a total budget of almost €28 million, has a consortium of 47 partners from 12 countries. The project will create a collaborative open platform that is standards-aligned and market-oriented, enabling end-to-end orchestration across service domains.

Activities are organized into four piloting clusters, as follows:

- **industry** (mining, manufacturing, recycling)
- **smart buildings** / smart cities
- **agriculture**
- **energy management**

Cross-sector scenarios will focus on mobility and logistics.

COP-PILOT services will manage data flows to applications and repositories, while software components will enable configuration and resource orchestration across different infrastructures. Finally, COP-PILOT platform services will provide end-to-end service orchestration across geo-distributed domains.

CEI-Sphere

CEI-Sphere works with the large-scale pilots and European industry players to translate use cases into business models, promoting replicability and scalability of the solutions developed. The project is mapping pilots and actors, developing a use-case catalogue, promoting interoperability through a minimum interoperability mechanisms (MiMs) approach, and supporting the development of secure and compliant CEI systems.

To maintain market focus, CEI-Sphere will engage networks of industry actors via events and via its CEI Tech Backbone Toolkit. To maintain the quality, security, and interoperability of cloud-edge infrastructures throughout Europe and to promote the adoption of these diverse and complex systems, CEI-Sphere will work towards a trust-label framework.

o-cei.eu
cop-pilot.eu
ceisphere.eu

O-CEI, COP-PILOT and CEI-Sphere have received funding from the EU's Horizon Europe research and innovation programme under grant agreement no. 101189589, 101189819 and 101189683 respectively.

Launch of TerraDT digital-twin project to plan shipping routes and green areas

Roberta Fabrizi, Trust-IT

Funded by the European Union (EU), TerraDT is a digital-twin project designed to enhance our understanding of how glaciers, sea ice, vegetation, and aerosols influence the Earth's climate. By delivering high-resolution climate impact assessments, it will provide decision-making tools for local planning, such as helping determine optimal locations for shipping routes, parks, and other infrastructure. The project will leverage Europe's most advanced supercomputers to achieve unprecedented modelling accuracy.

Closely linked to the European Commission's Destination Earth (DestinE) Climate Change Adaptation Digital Twin (Climate DT), TerraDT expands the climate twin's modelling capabilities by integrating new components that more precisely capture the role of glaciers, sea ice, and aerosol particles. These enhancements will significantly improve the reliability of climate projections, supporting better adaptation and mitigation strategies.

In TerraDT, climate projections are made at the very high resolution of 10 km. This requires a significant amount of computing resources. The main computing platforms in the projects are CSC – IT Center for Science's LUMI supercomputer in Kajaani, Finland, and Barcelona Supercomputing Center's MareNostrum 5, located in Barcelona. Both of these are pan-European EuroHPC supercomputers.

Information to support decision-making

'The aim is to obtain information to support decision-making, for example, on how sea ice conditions impact shipping routes. What kind of effects will sea level changes or extreme weather conditions have on coastal and offshore construction and urban planning? The model also provides us with important information related to forest biodiversity and carbon sink assessment,' says Jenni Kontkanen, Development Manager at CSC – IT Center for Science, who leads the TerraDT project.

An interactive user interface is being developed for TerraDT, through which users can ask concrete 'what if?' questions. For example, how does building a park in a particular city impact temperatures or carbon sequestration? TerraDT's high resolution and accuracy make it a practical tool for local decision-making: the effects of various phenomena can be viewed and predicted locally at a precise level.

The TerraDT project started at the beginning of 2025 and will last for four years. Its total budget is about €15 million. TerraDT is a Horizon Europe-funded research project that involves 18 organizations from all over Europe.

terradt.eu

DRAMSys project powers open-source exploration for DRAM memory systems

Matthias Jung, University of Würzburg JMU / Fraunhofer IESE

Launched in May 2024, the DI-DE:RAMSys project is funded by the German Ministry of Research, Technology and Space (BMFTR) with a budget of approximately €1 million. The consortium includes HiPEAC members Matthias Jung (University of Würzburg – JMU / Fraunhofer IESE) and Norbert Wehn (University of Kaiserslautern-RPTU). The project's aim is to develop an open-source exploration framework for DRAM memory systems, based on the DRAMSys simulator, DRAMPower, and gem5, allowing for highly automated DRAM subsystem configuration.

DRAMSys, a widely used open-source simulation tool in industry, which won a HiPEAC Transfer Award in 2021, is being expanded to support new memory technologies and standards such as CXL, UCIe, HBM4, LPDDR6, and DDR6. Additionally, DRAMPower, the power and energy simulator for estimating DRAM memory power consumption, is being enhanced to improve both execution speed and accuracy in energy consumption modelling. In January 2025,

the project team was awarded the Best Paper Award at the HiPEAC conference RAPIDO workshop for DRAMPower 5.

The project embraces open-source principles to foster collaboration with other developers and companies. Furthermore, the integration of DRAMSys into the gem5 open-source processor simulator will be extended.

The 'Design Instruments for Sovereign Chip Development with Open Source (DE:Sign)' initiative is supported by ththe German Federal Ministry of Research, Technology and Space (BMFTR) to promote projects focused on the research of new design tools and methods, as well as innovative chip designs, with an emphasis on open-source processes for new microelectronics.

DI-DE:RAMSys website dramsys.de
Official repository on GitHub
github.com/tukl-msd/DRAMSys

PROTECT: Proving Next-Generation Secure Systems

Lennart M. Reimann and Rainer Leupers, RWTH Aachen

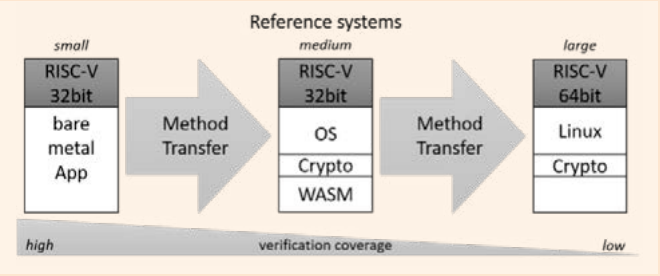
As cyberattacks on companies and institutions increase worldwide, there is a growing need for more secure information technology (IT) systems. The PROTECT project aims to address this challenge by developing innovative solutions that make IT systems more resistant to threats.

Funded by the German Agentur für Innovation in der Cybersicherheit GmbH (Cyberagentur), PROTECT brings together experts from academia and industry. The project's goal is to employ new methods of formal verification to improve IT security and build trust in digital systems.

The project consortium comprises the following partners:

- German Research Center for Artificial Intelligence (DFKI), which coordinates the project
- RWTH Aachen University
- Cryspen SARL (Paris, France)
- Gesellschaft für Informatik e.V. (Berlin/Bonn)
- RPTU Kaiserslautern
- UBIS EDA GmbH (Kaiserslautern)
- University of Lübeck

RWTH Aachen University is a key partner in this effort. As part of the project team, RWTH researchers are working on developing advanced verification techniques. They are also contributing to reference system architectures based on RISC-V.



RWTH is advancing the use of virtual prototypes in combination with formal verification techniques for RISC-V-based systems. The university is leveraging the SAIL language, which provides a formal specification of the RISC-V instruction set architecture (ISA).

SAIL allows for precise specification of RISC-V processors, enabling RWTH researchers to create accurate virtual prototypes. These prototypes can be used to test and verify designs before physical implementation.

The project takes a unique approach by combining existing strengths of partners instead of enforcing a one-size-fits-all solution. All work is published as open source or open access, and it is driven by well-established researchers in their fields.

PROTECT is part of the 'Ecosystem verifiably secure IT – Provable Cybersecurity' (EvIT, German acronym: 'ÖvIT'). It started in January 2025 and will run for four years. By working on these challenges, RWTH Aachen and its partners in PROTECT aim to create more secure IT systems for the future.

This project shows RWTH's ongoing commitment to cutting-edge research in cybersecurity. It is an important step in making digital systems more trustworthy in an increasingly connected world.

PROTECT

EDAA Achievement Award 2025 presented to Subhasish Mitra

In March 2025, HiPEAC associate member and ACACES 2025 teacher Subhasish Mitra was recognized with the European Design and Automation Association (EDAA) Achievement Award. This award is given to individuals who made outstanding contributions to the state of the art in electronic design, automation and testing of electronic systems. To be eligible, candidates must have made innovative contributions that impacted how electronic systems are being designed.

Subhasish, who also gave a keynote talk at HiPEAC 2023, holds the William E. Ayer Endowed Chair Professorship in the Departments of Electrical Engineering and Computer Science at Stanford University. According to the EDAA announcement of the award, results from his research group have influenced almost every contemporary electronic system and have inspired significant government and research initiatives in multiple countries. In addition to his role at Stanford, he has held several international academic appointments, including at CEA-LETI and EPFL, and has consulted for major technology companies.

His honours include the IEEE Computer Society Harry H. Goode Memorial Award, the ACM SIGDA and IEEE CEDA Newton Technical Impact Award in EDA, the University Researcher Award by the Semiconductor Industry Association and Semiconductor Research Corporation, the Intel Achievement Award, the Humboldt Research Award, and the Distinguished Alumnus Award from the Indian Institute of Technology, Kharagpur.

Onur Mutlu receives IEEE Computer Society Harry H. Goode Memorial Award

Earlier this year, HiPEAC member and ACACES 2024 teacher Onur Mutlu, a professor at ETH Zürich, received the IEEE Computer Society Harry H. Goode Memorial Award 'for seminal contributions to computer architecture research and practice, especially in memory systems'.

According to the IEEE summary, many techniques that Onur has invented with his group and collaborators have largely influenced industry and have been employed in commercial microprocessors and memory and storage systems used by billions of people.

Onur is an IEEE Fellow, an ACM Fellow, and an elected member of the Academy of Europe. He has received numerous honours for his research, including the 2024 IFIP Jean-Claude Laprie Award in Dependable Computing, 2021 IEEE High Performance Computer Architecture Conference Test of Time Award, 2022 Persistent Impact Prize of the Non-Volatile Memory Systems Workshop, 2021 Intel Outstanding Researcher Award, 2020 IEEE Computer Society Edward J. McCluskey Technical Achievement Award, and the 2019 ACM SIGARCH Maurice Wilkes Award.

Onur's lectures from ACACES 2024 are available to view on HiPEAC TV
bit.ly/ACACES24_playlist

Michael O'Boyle elected fellow of the Royal Society of Edinburgh

HiPEAC founding member Michael O'Boyle, chair in computer science at the University of Edinburgh's School of Informatics, has been elected fellow to the Royal Society of Edinburgh in recognition of his compiler research.

Michael's research interests include adaptive compilation, machine learning-based optimization, auto-parallelizing compilers, and heterogeneous general-purpose graphics processing unit (GPGPU) multicore platforms. His is director of the Arm Centre of Excellence and has been awarded multiple best paper, distinguished paper, and test of time awards.

Commenting on the announcement, Michael said: 'I am delighted that the compiler research at Edinburgh has been recognized. I am honoured to be elected a Fellow of the Royal Society of Edinburgh and would like to thank my colleagues in ICSA and Informatics more broadly, for their generous support over the years. I am excited about the future and how machine learning continues to transform our domain.'

Sergi Abadal receives Agustín de Betancourt y Molina medal

In November 2024, HiPEAC member Sergi Abadal received the Agustín de Betancourt y Molina medal, which is presented annually by the Spanish Real Academia de Ingeniería (Royal Engineering Academy) to the top five engineers under 40 in Spain.

Sergi is a distinguished researcher in the Department of Computer Architecture at the Universitat Politècnica de Catalunya-Barcelona Tech (UPC). His areas of specialization are graphene antennas, nanocommunications and integrated networks, among others. He has been awarded a European Research Council (ERC) Starting Grant and a Proof of Concept Grant. He also received the ACM NanoCom Outstanding Milestone Award in 2022 and the Young Investigator Award of the Nano Communication Networks Journal in September 2019.

Giovanni De Micheli wins IEEE Gustav Robert Kirchhoff Award

HiPEAC 2025 keynote speaker and IEEE Life Fellow Giovanni De Micheli, director of the Integrated Systems Laboratory and of the EcoCloud Centre at EPFL, has been honoured with the 2025 IEEE Gustav Robert Kirchhoff Award 'for fundamental contributions to the design of systems and networks-on-chip'.

According to the IEEE summary, Giovanni pioneered methods that map algorithmic descriptions of computation and communication into digital circuits and networks. His most impactful contribution is the automated design of networks on chips (NoCs). In his visionary approach, he conceived integrated circuit (IC) components communicating through data packets, and he created the technological path to make NoCs energy efficient and competitive on silicon.

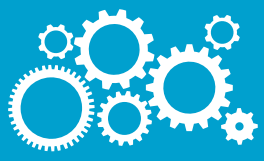
View Giovanni De Micheli's full keynote talk, plus interviews with eeNews and EE Times, in the HiPEAC 2025 playlist on HiPEAC TV
bit.ly/HiPEAC25_videos

In memoriam: Sally McKee

In February, we were saddened to hear of the death of HiPEAC member Sally McKee. Sally, who coined the term 'memory wall', was a professor at Chalmers University of Technology, and was one of the original founders of the Computing Frontiers conference.

Paying tribute to Sally at the Computing Frontiers conference in May, HiPEAC member Francesca Palumbo said: 'Sally was more than a brilliant researcher. She was a mentor, a leader, a friend, and a source of inspiration for many in our community. Her passion, intellect, warmth, and tireless support shaped not only the field of memory systems research, but also the personal and professional lives of countless colleagues and students.'

Our thoughts are with her family and friends.



Developed by researchers at the Politecnico di Torino, Spiker+ allows users to automatically generate descriptions of hardware accelerators optimized for their spiking neural networks and deployable to field-programmable gate arrays (FPGAs). In this article, Alessio Carpegna, Alessandro Savino and Stefano Di Carlo (all Politecnico di Torino) explain why spiking neural networks are important and how Spiker+ helps translate these innovative architectures to custom hardware.

Spiker+

Simplifying custom hardware for spiking neural networks

How did Spiker+ come about?

Stefano: For some time, we had observed the rapidly growing interest in spiking neural networks, with researchers trying to match the performance of more established architectures like convolutional neural networks (CNNs) or deep neural networks (DNNs). Alessandro and I recognized a unique opportunity to combine our group's expertise in computer architectures, reliability, fault-tolerant systems, and approximate computing in this exciting new domain. We decided to focus on the development of efficient computational units capable of deploying SNNs in environments with extremely constrained computational resources. Around this time, we met Alessio Carpagna, who subsequently joined our team as a PhD student and became the driving force behind Spiker+.

What are the advantages of using spiking in neural networks (as opposed to coding for neural activity with floating point numbers, for example)?

Alessandro: The biggest benefit of spiking neural networks is that they're much closer to the way our brains work. Instead of dealing with numbers all the time, these networks use quick pulses – spikes – to communicate. That makes them extremely efficient because they don't waste energy processing signals unless there's something meaningful happening. Another cool thing is that they're super compatible with neuromorphic hardware, special chips designed to imitate brain functions. These chips use very little power, making spiking networks ideal for devices that run on batteries or that need to process data in real time, like robots, sensors or wearables.

Stefano: Spiking also naturally handles time-based information better. Since spikes encode timing precisely, they are great for tasks involving patterns over time – think audio, video, or anything else that is dynamic. And because spikes are sparse – meaning that they're not constantly firing – these networks usually have a smaller memory footprint and scale better, which is great when you're building larger systems. Plus, they're robust: biological neurons handle noise and faults well, and spiking neural networks inherit some of this resilience. If part of the system fails, it doesn't necessarily break the entire model.

Alessandro: Of course, there's still work to do – training spiking networks isn't always straightforward, and the software and tool support isn't as mature as it is for traditional deep learning yet. But overall, for applications where energy, real-time responses and biological realism matter, spiking neural networks can really shine.

What are the advantages of field-programmable gate arrays (FPGAs) for accelerating spiking neural networks?

Stefano: There are two main reasons for targeting FPGAs to accelerate spiking neural networks. The first reason is practical: FPGAs are ideal platforms for hardware prototyping, widely accessible in academic labs and small companies that might not have the resources to tape out entire custom chips.

Second, we believe that FPGAs have enormous potential for edge and internet-of-things (IoT) applications. They can enable the design of custom hardware accelerators that can be precisely tailored for specific tasks without needing to completely redesign the underlying hardware architecture. They can even be multiplexed in time thanks to their in-field reconfigurability.

Tell us about Spiker+. How can it help designers navigate the complexity of hardware accelerators for spiking neural networks?

Alessandro: Spiker+ is a comprehensive framework designed to simplify the deployment of custom hardware accelerators for SNNs. When we first started exploring SNNs, we noticed that, although various hardware accelerators were being proposed in research papers, there was a lack of tools to support designers in the process of translating a high-level SNN model into deployable hardware. This was what led us to create Spiker+.

Alessio: Unlike traditional hardware accelerators, which typically provide a fixed hardware architecture that must be programmed to support specific SNN configurations, Spiker+ offers a more flexible approach. With Spiker+, users start directly from their dataset, train an appropriate spiking neural network tailored to their specific task, and automatically generate a full VHDL description of a hardware accelerator optimized for that network and deployable to commercial FPGAs, such as those produced by AMD. This approach ensures that only the specific hardware necessary for the trained network is deployed, significantly improving efficiency in terms of both area and power consumption.

Spiker+ is continuously evolving, and it now includes additional tools such as SpikeExplorer, which allows users to perform comprehensive design-space exploration and network optimization before hardware deployment, and SpikingJet, a fault-injection tool for assessing the resilience of SNNs against hardware faults – especially useful for safety-critical applications. (See 'Further information', below).

What kind of applications could Spiker+ be used for?

Alessio: Spiker+ is specifically designed for compact, embedded, or edge applications that require real-time processing of temporal data, particularly data originating from analogue sensors. These sensors typically measure continuous signals from the environment such as light intensity, sound-pressure levels, tactile feedback, or EEG signals. Although Spiker+ does not handle the analogue-to-spike conversion itself, it can be paired with any of the numerous lightweight and efficient analogue-to-spike conversion methods that exist in the literature to deploy customized, highly efficient hardware accelerators for SNNs.

This approach makes the framework ideal for low-power and highly responsive sensing applications, such as visual sensing using standard photodiodes, audio applications using basic microphones, or tactile sensing using conventional pressure sensors, all without the complexity involved in directly interfacing specialized neuromorphic sensor formats.

Stefano: Spiker+ is an open-source project, and is designed to help support the open-source hardware community. We encourage anyone interested to read our research papers outlining the Spiker+ framework itself and the neuron architectures it supports, to explore Spiker+ via our GitHub repository, and to watch our video tutorial guide on how to get started (see 'Further information', below).

Politecnico di Torino, Department of Control and Computer Engineering | DAUIN | SMILIES reSilient coMputer architectures and LIfE Sciences

FURTHER INFORMATION:

Spiker+ on GitHub github.com/smilies-polito/Spiker

Spiker+ video tutorials on YouTube bit.ly/Spiker_tutorials

A. Carpegna, A. Savino and S. Di Carlo. 'Spiker: an FPGA-optimized Hardware accelerator for Spiking Neural Networks'. 2022 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Nicosia, Cyprus, 2022, pp. 14-19 doi.org/10.1109/ISVLSI54635.2022.00016

D. Padovano, A. Carpegna, A. Savino, & S. Di Carlo (2024). 'SpikeExplorer: Hardware-Oriented Design Space Exploration for Spiking Neural Networks on FPGA'. Electronics, 13(9), 1744. doi.org/10.3390/electronics13091744

A.B. Göğebakan, E. Magliano, A. Carpegna, A. Ruospo, A. Savino and S. Di Carlo (2024). 'SpikingJET: Enhancing Fault Injection for Fully and Convolutional Spiking Neural Networks'. 2024 IEEE 30th International Symposium on On-Line Testing and Robust System Design (IOLTS), Rennes, France, 2024, pp. 1-7 doi.org/10.1109/IOLTS60994.2024.10616060

SMILIES research group smilies.polito.it

Neurobench benchmarks for neuromorphic computing neurobench.ai

Tools of the trade

Given the complexity of modern computing systems, tools are an essential part of the design and maintenance of efficient systems. Bridging hardware and software, the HiPEAC community has tools in its DNA. In this special feature, we hear about some of the tools being developed within the ecosystem, from tuning graphics processing units (GPUs) and mapping neural networks to field-programmable gate arrays (FPGAs) to leveraging data-science tools to build solutions for artificial intelligence.

Kernel Tuner: An open-source tool for optimizing GPU applications

Kernel Tuner

Ben van Werkhoven, Leiden University

In modern high-performance computing (HPC), achieving optimal performance on graphics processing units (GPUs) is essential for driving scientific discovery and innovation. However, writing high-performance GPU kernels can be challenging, due to the many ways of mapping computations to GPU architectures.

This is why we created Kernel Tuner. Kernel Tuner is designed to systematically explore and optimize the vast configuration space of GPU kernels. By automating the tedious process of tuning parameters like thread block sizes, tiling sizes, and shared memory usage, Kernel Tuner helps developers identify the most efficient configurations that deliver peak performance on target hardware. Since its inception nine years ago, the developer team has grown to nine core developers working at four different research institutes in the Netherlands, with many more open-source contributors.

Kernel Tuner provides a flexible approach to performance optimization. A user supplies the kernel code, a set of tuning parameters (for instance, thread-block dimensions), and a dataset to measure the performance of the kernel. The tool then compiles the kernel for each combination of parameters, executes it, and measures the resulting performance. The best configuration is reported back to the developer.

This process helps ensure that GPU codes can fully leverage the massive parallelism available on modern hardware. Moreover, because Kernel Tuner systematically searches the tuning space, developers gain insights into how different combinations of code optimizations affect performance, enabling more informed decisions when writing high-efficiency kernels.

Compatibility with programming models

One of the strengths of Kernel Tuner is its compatibility with multiple programming models, such as HIP, OpenCL, and CUDA. Recently, we have extended this support to directive-based programming models in Fortran and C such as OpenACC, and support for OpenMP offloading is underway. This means it can be integrated into various projects and workflows with minimal friction, providing a versatile solution for researchers, engineers, and data scientists working on GPU-accelerated applications. Whether you are an HPC developer pushing the limits of supercomputer performance or a domain scientist optimizing GPU codes, Kernel Tuner's automation can significantly reduce the time spent on trial-and-error performance tuning.

Recognizing the importance of user education, the Kernel Tuner team has developed substantial training materials to help both novice and experienced users get the most out of the tool (see 'Further information', below). These include hands-on tutorials and presentations that have been delivered at some of the most prestigious HPC conferences, such as SC21, ISC-HPC22, and SC23.

The tutorials cover everything from getting started with Kernel Tuner to advanced usage and best practices for performance tuning on the latest GPU architectures. These in-depth sessions demonstrate real-world examples, ensuring that participants come away with a practical understanding of how to apply Kernel Tuner in their own work. The SC23 tutorial focused specifically on how to use Kernel Tuner to optimize the energy efficiency of GPU applications.

Left to right: Ben van Werkhoven, Floris-Jan Willemsen, Stijn Heldens and Alessio Sclocco at Supercomputing 2023 (SC23), where they gave a tutorial on energy-efficient GPU computing

Improving HPC application performance

Beyond its broad applicability, Kernel Tuner is also a key component in collaborative efforts to improve HPC application performance in critical areas of research. Through the ESiWACE3 EuroHPC-JU project, Kernel Tuner is being used to support Europe's weather and climate modelling community in adapting their codes to the latest supercomputer architectures. The weather and climate domain has for a long time relied on central processing unit (CPU)-based architectures and MPI parallelism, but as most of the performance in modern supercomputers now comes from GPUs it is necessary to restructure and port these applications. By tapping into Kernel Tuner's capabilities, researchers can fine-tune their GPU kernels to run more efficiently on modern supercomputers, enabling more accurate and faster weather forecasts and climate projections.

Whether you are tuning an existing GPU code or creating a new HPC application, Kernel Tuner offers an invaluable, user-friendly solution for uncovering optimal configurations. With comprehensive training materials and a growing user community, now is an exciting time to adopt Kernel Tuner to maximize performance on modern GPU-based systems.

FURTHER INFORMATION:

Kernel Tuner on GitHub
github.com/kerneltuner/kernel_tuner

Esclapez, L., Soucasse, L., Jungbacker, C., Jansson, F., de Roode, S.R., Costa, P., van den Oord, G., Sclocco, A (2025). 'Accelerating the Dutch Atmospheric Large-Eddy Simulation (DALES) model with OpenACC'.
arxiv.org/abs/2502.20412

Kernel Tuner tutorials on GitHub
github.com/kerneltuner/kernel_tuner_tutorial

ESIWACE project website
esiwace.eu

FPG-AI: AUTOMATING DNN ACCELERATION ON FPGAS FOR SPACE APPLICATIONS AND BEYOND



Tommaso Pacini, Pietro Nannipieri and Luca Fanucci, University of Pisa

As artificial intelligence (AI) workloads grow increasingly complex, researchers and engineers face significant challenges in optimizing deep neural networks (DNNs) for field-programmable gate array (FPGA) implementation, balancing performance, power efficiency, and design time. A novel DNN-to-FPGA automation toolflow, FPG-AI streamlines this process by automating key steps, from model quantization to hardware mapping, leveraging hardware / software (HW/SW) optimizations to enhance efficiency and adaptability.

By reducing the need for extensive manual tuning, FPG-AI enables both experts and non-specialists to harness the power of FPGAs for AI acceleration. The toolflow also plays a role in education and training, lowering the barrier to entry for students and researchers interested in hardware-aware AI development.

FPG-AI takes as its input a pre-trained DNN, the application dataset, and the target FPGA device. The first step involves applying model compression, where a post-training quantization algorithm converts the model from floating-point (FP) to fixed-point (FXP) arithmetic. Next, a design space exploration (DSE) best configures the tunable hardware accelerator for the compressed model. The hardware design is a modular deep learning engine (MDE), a handcrafted HDL-based architecture that supports various devices and resource budgets.

▶

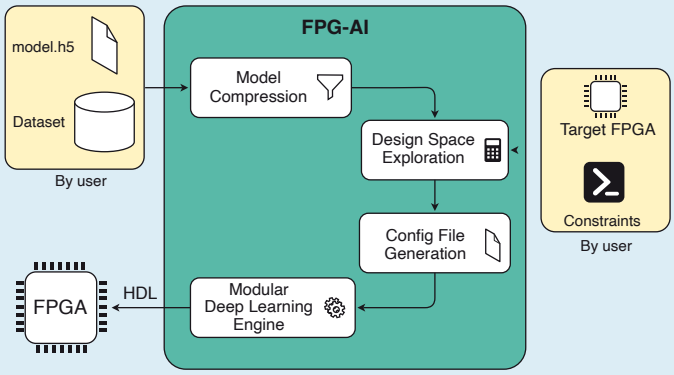
At present, FPG-AI supports the widest range of FPGAs among state-of-the-art toolflows, targeting devices manufactured by AMD Xilinx, Intel, Microchip, and NanoXplore. The framework outputs the hardware description language (HDL) sources of the customized accelerator for the given DNN-FPGA pair. The generated intellectual property (IP) is an AXI memory-mapped block that can be easily integrated into a user-defined system-on-chip (SoC), also allowing additional FPGA resources to be allocated for other tasks.

The framework has gained the interest of research centres and companies in the space field. Notably, FPG-AI is the first toolflow to enable AI acceleration on NanoXplore FPGAs, one of the few radiation-hardened (rad-hard) FPGAs designed specifically for space, aerospace, and defence applications. NanoXplore FPGAs are also the only European solution for space systems.

Additionally, the code generated by FPG-AI is human-readable, facilitating a comprehensive space qualification campaign while ensuring high explainability and reliability.

The development of this toolflow has been supported by two European Space Agency (ESA) projects (see 'Further information', below), further underlining its strategic importance in advancing AI for space applications.

Overall, FPG-AI provides an end-to-end solution that automates the critical steps of the DNN-to-FPGA process and significantly reduces the time and expertise required to port AI models to FPGA platforms. This makes it an invaluable tool not only for hardware designers but also for AI practitioners who seek efficient inference without deep FPGA expertise. By simplifying FPGA-based AI acceleration, FPG-AI paves the way for broader adoption of reconfigurable hardware in modern AI applications.



FURTHER INFORMATION:

Pacini, T., Rapuano, E., Fanucci, L. 'FPG-AI: A Technology-Independent Framework for the Automation of CNN Deployment on FPGAs', IEEE Access, March 2023
doi.org/10.1109/ACCESS.2023.3263392

'A novel design framework for rapid and efficient Artificial Intelligence deployment for on-board space applications', Open Space Innovation Platform (OSIP) Activity sponsored by the European Space Agency (ESA), Contract Number: 4000129792
activities.esa.int/4000129792

'FPG-AI: a Technology Independent Framework for Edge AI Deployment Onboard Satellite, and its Characterisation on NanoXplore FPGAs', Open Space Innovation Platform (OSIP) Activity sponsored by the European Space Agency (ESA), Contract Number: 4000141108
activities.esa.int/4000141108

How JupyterHub can help build AI / ML tools

Marc Michalke, TU Braunschweig

While the use of artificial intelligence (AI) and machine learning (ML) tools is seeping into almost all aspects of our professional lives, a less highlighted aspect is which tools can be used to help build these AI tools. While the choice of the best platform, pipeline or development environment is never trivial, having an effective solution can be the deciding factor in meeting project deadlines, collaborating effectively, and generating impactful results. In this article, we would like to introduce you to the open-source project JupyterHub: a containerized development environment well known in the data-science community but not necessarily used for development of AI and ML solutions, despite its benefits mostly translating to this area as well.

Background

First, we need to go over the core concept of Jupyter notebooks. Usually running in a container, a web interface is served that the user can interact with and use, not only as an integrated development environment (IDE) but also to execute either the whole file ('notebook') or just selective code snippets in the browser. The beauty here lies in the ability to separate the programming environment from the execution environment, especially when the software and browser are running on different machines, as well as allowing for easy and quick code tests and visualization.

This separation not only ensures that the execution environment for the code remains stable even through long development cycles; it also eliminates a variety of compatibility and platform concerns. Changes and updates to the user's operating system do not hinder development, as long as their browser works. This execution environment, the so-called kernel, can also be switched on demand, allowing for multiple consistent environments for different projects.

The hub

JupyterHub now enhances this concept significantly by expanding these scenarios into a powerful multi-user environment, a feature that can be a game-changer for team collaboration. Here, all members share a consistent environment, which drastically reduces time spent troubleshooting 'it works on my machine' problems. Notebooks, data, and experiment results can be instantly shared among teammates, making it easier to do peer reviews, explore new ideas, and iterate quickly; all core needs in constructing robust solutions.

In addition, JupyterHub is offered as a Helm chart, allowing for deployment on Kubernetes clusters including all its benefits of scaling and maintenance. For a quick test, 'The Littlest JupyterHub' provides a minimal setup that can easily be deployed even on a local machine.

Usage in ICOS

In the ICOS project, JupyterHub is included in the 'AI Support' container to provide a multi-user environment for interactive data analysis and development of AI models that can then be used to train on and predict metrics. For additional security, models then have to be exported manually and imported on the ICOS controller by an administrator, and then be considered by the policy manager, allowing for reaction to bounds being crossed, e.g. via rescheduling workloads across the IoT-edge-cloud continuum through the ICOS MatchMaker. This setup supports flexible experimentation and collaboration within the broader ICOS intelligence framework, allowing users to quickly test code and models.

Conclusion

JupyterHub is a robust platform that simplifies collaboration, enhances resource management, and streamlines development workflows. By removing barriers around software setup and version control, JupyterHub enables teams to focus on building, training, and refining cutting-edge machine learning models. Whether you're working on breakthrough research, developing production-grade systems, or instructing the next generation of data scientists, JupyterHub offers a powerful, flexible, and secure environment that caters to today's demanding AI / ML needs.

FURTHER INFORMATION:

JupyterHub website jupyter.org/hub

The Littlest JupyterHub tljh.jupyter.org

ICOS project icos-project.eu

As we enter an era of 5GW data centres, with energy use rocketing to serve the demands of artificial intelligence (AI) workloads, technology companies are turning to extreme solutions, such as powering up disused nuclear power plants. But what if you don't have access to your own personal power plant? In this article, Luigi Brochard and Julita Corbalán (Energy Aware Solutions) explain how their Energy Aware Runtime (EAR) tool allows users to monitor and optimizes the energy use of their systems.

'On average, Energy Aware Runtime can increase performance per watt by up to 50%'

While reducing power use has long been a key concern for the computing systems community, the energy issue isn't going away; in fact, it's getting increasingly critical, according to Luigi Brochard, chief executive of the Spanish company Energy Aware Solutions (EAS). 'Data centres' collective energy consumption is already huge, and it is increasing: by 2030, it could reach up to 13% of global power consumption.'

Power-hungry processors, particularly for AI applications, are a large part of the issue: 'In the last decade, central processing unit (CPU) power demands have more than doubled (from 150W to 350W per socket), while those of graphics processing units (GPUs) have quadrupled (from 400W to 1200W). To serve the requirements of AI, we are seeing routine configuration with four and eight GPUs per server, leading to 10KW servers,' says Luigi. 'This has dramatic implications for how the servers are cooled and how to deliver enough power to a system with hundreds or thousands of servers.'

Taking these figures into account, it's easy to see how the energy consumption of data centres has risen so dramatically, often surpassing 100MW, says Luigi: 'This explains why some AI companies are now looking at dedicated nuclear plants to power their data centres.'

Optimization through understanding

In this context, monitoring the energy use of HPC / AI data centres is a critical first step to being able to control and reduce it. 'The first step for optimization is understanding,' explains Julita Corbalán, the chief technology officer at EAS. 'Knowledge of how much energy is being used is key to be able to optimize, both statically by users and app developers, but also to improve tools for automatic optimization. This knowledge will also help data centres and vendors better understand the real requirements of applications, and then innovate based on accurate information.'

While tools for monitoring the power use of servers and other datacentre infrastructure have existed for some time, the EAR tool is different in that it takes a holistic approach, says Julita.

Photo credit: Aashish Yadav on Unsplash

'Existing solutions mostly addressed power use, temperature, etc., from the hardware point of view. There was also some basic accounting included in the schedulers, but specific energy- and power-management tools were lacking. For example, SLURM includes some basic reporting on energy consumption, but there is no correlation with performance, or any optimization,' she explains.

In contrast, EAR monitors both the hardware and the workloads running on the servers, Luigi says. 'Thanks to this systemic approach, EAR can show not only how much power a workload is consuming, but also what this workload is doing at runtime. In addition to allowing EAR to report the energy consumption and CO_2 emissions of each workload, this means that it can take action to limit the energy consumption of each job, to trigger actions if the power use of the whole data centre is going over limits, or to issue alerts if workloads or servers demonstrate "abnormal" behaviour.'

In addition to monitoring, EAR also offers analytics tools to analyse workloads or the system as a whole, as well as a smart power and thermal cap to make sure the system does not exceed predefined limits. It also offers dynamic energy optimization. 'As such, EAR can increase the performance per watt of workloads running on CPUs and GPUs. On AI workloads, we measure, on average, a 20% performance-per-watt improvement on GPUs with no user intervention, which increases to 50% with user intervention,' says Luigi.

'By "user intervention", we mean that EAR provides information and hints so that the user just has to accept and apply our recommendations,' adds Julita. 'Beyond individual applications, the benefit can be extrapolated to an entire data centre, because it would be effective for all executions of the same use case. Dynamic and static optimizations complement one another, which is a win-win strategy for users and data centres.'

Balancing energy use and performance

Does optimizing for EAR have a knock-on effect on performance? 'EAR has a controlled impact on performance. In some cases, EAR will detect that hardware resources are idle or poorly used, and the energy optimization will not have any performance impact. In other cases, where EAR optimizes energy while the CPUs or GPUs are computing, EAR is able to predict the energy impact of a potential CPU or GPU frequency change. In these situations, the system administrator can tell EAR to optimize energy within a predefined limit on performance impact, such as 3% or 5%. Of course, the higher the performance tolerance, the higher the energy gain,' explains Luigi.

Photo credit: Mangolovemom on Adobe Stock

'Rather than seeing it from the perspective of a performance penalty, you can view it from the perspective of making more performance available,' adds Julita. 'If you increase the execution time of an application by 5% but you reduce power consumption by 25%, in a power-limited data centre that frees up power to start another application. Doing relatively simple maths, you can therefore calculate how optimizing energy offers an opportunity for increased performance – at least in those data centres that don't have the option of building nuclear power plants to power their systems.'

While EAR is currently designed to serve centralized computing facilities, Julita notes that it is distributed software that is designed to be dynamically configurable and extensible. 'Most of EAR's features are implemented with a plug-in mechanism. That means that, even though in its current form it doesn't target edge computing, new policies, models, reporting mechanisms, monitoring application programming interfaces (APIs), etc., can be developed to fit new scenarios without affecting the core of the software. Moreover, EAR components are not limited to interacting with each other, so we can not only configure features but also configure which components are used, and hence minimize the number of components deployed.' Luigi agrees: 'We're always working on new features, so look out for future developments.'

FURTHER INFORMATION:

Energy Aware Solutions website eas4dc.com

Using multi-devices with OpenMP


Xavier Teruel and Roger Ferrer (Barcelona Supercomputing Center)

For some years, the advantages of the cloud computing model have been clear. Cloud services enhance business performance through agile deployment, robust security, efficient data management, and utility-based sharing models. The total cost of ownership is low, and they can offer access to high-performance computing (HPC). Cloud computing supports quicker adaptation and smoother operations in dynamic market environments, simplifying technology adoption.

Data centres play a fundamental role in enabling cloud computing, serving as the physical backbone that powers a vast array of digital services and platforms. These facilities house the servers, storage systems, and networking infrastructure required to process, manage, and store enormous volumes of data generated and accessed through the cloud. As demand for HPC continues to grow, driven by technologies such as artificial intelligence (AI), machine learning (ML), and large-scale data analytics, modern data centres are increasingly built with systems that incorporate an increasing number of graphics processing units (GPUs).

GPUs are particularly well-suited for handling computationally intensive tasks due to their ability to perform parallel processing at high speeds, making them ideal for workloads that involve deep learning, simulations, and big-data processing. At the same time, there is a growing emphasis on energy efficiency, as data centres consume significant amounts of energy and have a considerable environmental impact.

Maximizing GPU utilization

A common challenge in programming systems equipped with multiple GPUs is maximizing their utilization to achieve high performance. To tackle this, developers rely on programming models that are specifically designed to expose and exploit the massive parallelism offered by GPUs. Algorithms that demonstrate significant data parallelism tend to benefit the most from GPU acceleration. Several programming models support this effort, including OpenMP, which offers directive-based parallelism and has been extended to target GPUs. Other widely used models include CUDA, developed by NVIDIA for fine-grained control over GPU programming, and OpenCL, which provides a platform-independent approach for writing code that runs across heterogeneous systems.

In this article, we will delve into the opportunities presented by OpenMP. Our exploration will focus on the current limitations of this programming model in terms of the multi-device scenario. We will also examine how developers can effectively annotate their applications to maximize expressiveness, providing a comprehensive overview of the capabilities and advantages of OpenMP in this scenario.

OpenMP, HPC and GPUs

OpenMP originated within the specific realm of HPC. Unlike other programming models that necessitate the use of routine invocation via an extended application program interface (API), OpenMP simply requires developers to mark sections of their applications that can benefit from task- and data-parallelism.

In addition, OpenMP (since version 4.0) provides directives that allow the offloading of specific parts of applications to any attached accelerator, leveraging both types of parallelism mentioned above. Task-parallelism in this context roughly corresponds to the offload process itself and data-parallelism exploits the parallel execution of the accelerator. In the context of GPUs, exploiting data-parallelism is key to achieving good performance and speeding up applications. OpenMP, when it comes to accelerators, is not linked to any specific vendor, making it portable across different acceleration technologies.

OpenMP is flexible, and it can be used in clusters where each node has many GPUs, such as the ones in data centres as stated above. However, the programming model does not provide enough expressiveness in these scenarios. For instance, the programmer needs to explicitly distribute data and execution over the different GPUs in the node. We can see this in the left-hand side of the figure on the next page. The programmer needs to explicitly compute the bounds of the different chunks of the data-parallel loop in order to distribute it among the different GPUs. Computing these bounds adds to the burden of accelerating an application and is not flexible in more complex scenarios such as those exposing imbalance.

At Barcelona Supercomputing Center (BSC), as part of the RISER project funded by the European Union, we have been working on providing an easier way to program systems with several

```
int n = <Number of iterations>;
int numDevices = <N>;
int blockSize = <Block size>;


#pragma omp parallel num_threads(numDevices)
{
    int chunkSize = (n + numDevices - 1) / numDevices;
    int numBlocks = (chunkSize + blockSize - 1) / blockSize;
    int tid = omp_get_thread_num();
    int start = tid * chunkSize;
    int end = min((tid + 1) * chunkSize, n);

#pragma omp target teams distribute parallel for    \
        map(from: X[start:chunkSize]) device(tid)   \
        num_teams(numBlocks) thread_limit(blockSize)
    for (int i = start; i < end; i++) { [...] }

}
```

```
int n = <Number of iterations>;
int numDevices = <N>;
int numBlocks = <Number of blocks>;
int blockSize = <Block size>;

#pragma omp target spread teams distribute parallel for     \
        map(from: X[omp_start:omp_size]) devices(1,2,…, N) \
        num_teams(<Blocks>) thread_limit(<BlockSize>)
    for (int i = 0; i < n; i++) { [...] }

}
```

Source-code comparison: manual distribution (left) and target spread (right)

accelerators, with a specific focus on a scenario with homogeneous accelerators, like systems with many GPUs found in data centres.

To this end, we have extended OpenMP with a new directive that we call spread. We can see an example in the right-hand side of the figure above. In the example, the spread construct extends the target directive when applied to a parallel loop. That creates an additional level in the hierarchical parallel decomposition of the attached loop. The first level divides the loop iteration steps into chunks (i.e. pieces) to later assign them across the set of participating devices. The user can now specify a set of devices instead of a single device. The proposed devices clause allows to include a list of comma-separated device identifiers that will become the associated set of participating devices. For each device, the programmer can also use special variables, such as omp_spread_start and omp_spread_size which will allow them to compute the boundaries of the different chunks offloaded to the accelerator device. These variables, implemented by the compiler, help map the data to the different device memories.

The main objective of this proposal is not just to boost performance; above all, we aim to enhance the expressiveness and productivity of programmers who want to utilize multiple GPU systems using the OpenMP programming model.

Currently, we have a prototype implementation of the spread directive on top of Clang / LLVM compiler infrastructure. We are evaluating its usefulness and investigating the performance impact in codes that benefit from this new directive. In most cases, we expect performance results comparable to the manual approach. However, in certain instances, performance improvements may occur due to the runtime system's more effective decision-making, particularly through load imbalance correction techniques. The selection of codes being evaluated includes numerical applications, machine learning, cryptography and image processing. These codes are representative of the applications typically run on cloud platforms.

The RISER accelerated platform presents itself as a system with many accelerators, and the spread directive provides a tool to exploit this scenario, thanks to the portable nature of OpenMP.

The RISER project has received funding from the European Union's Horizon Europe research and innovation funding programme under grant agreement number101092993. 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.

"Above all, we aim to enhance the expressiveness and productivity of programmers who want to utilize multiple GPU systems using OpenMP"

Created to design the processor whose roadmap was developed by the European Processor Initiative, the European company SiPearl is seeing increased demand for its Arm-based central processing units (CPUs) for servers, thanks in large part to the demands of artificial intelligence (AI). In this article, Roberto Mostallino shows how SiPearl is building out the Arm ecosystem and contributing to European autonomy in the semiconductor market.

The role of SiPearl's CPUs in the Arm ecosystem and the RISER Project

Market opportunities for Arm-based CPUs in the cloud

The semiconductor industry, which powers a wide range of modern technologies from automotive microcontrollers to supercomputers, is forecasted to grow significantly over the next few years, despite challenges such as supply-chain disruptions and geopolitical tensions. By 2028, the market size is projected to exceed $600 billion, with a compound annual growth rate (CAGR) of 8%, as shown in Figure 1, below.

The compute segment, which includes server CPUs and accelerator cards, represents approximately 26% of the total market and is expected to grow by 7%, driven by the increasing need for efficient and scalable server solutions due to the explosion of data generation.

Arm CPUs are particularly well suited to meet this demand, thanks to their power efficiency and performance per watt. Market analysts predict that Arm-based CPUs will capture 20% of the datacentre market share by 2028 (source: Gartner Research), driven by adoption from major cloud service providers such as AWS, Azure, and Google Cloud. Geopolitical and supply chain considerations have also accelerated the adoption of Arm-based CPUs, as companies and governments seek to diversify their supply chains and reduce dependency on single suppliers. The push for technological sovereignty, especially in Europe, has further spurred investment in Arm-based development projects, positioning Arm CPUs as a key component in creating a resilient and autonomous technology infrastructure.

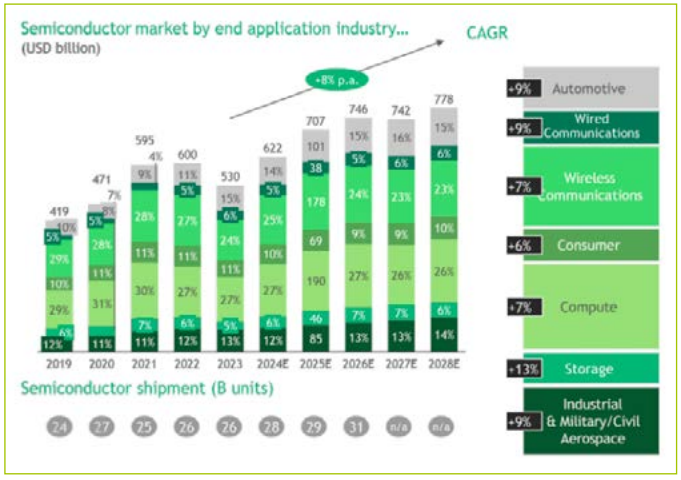
With the growing demand for more powerful and efficient chips in the semiconductor industry, the need for energy-efficient solutions has never been more urgent.

The Arm ecosystem and SiPearl's role in it

The significant growth in data, primarily driven by the requirements of AI and the increasing reliance on cloud services, has led to a surge in datacentre power consumption, which is forecasted to rise to about 1,000 TWh by 2030, as shown in Figure 2. This increase has highlighted the need for energy-efficient solutions, with Arm CPUs emerging as a compelling alternative due to their superior power efficiency and performance per watt compared to traditional x86 processors, as mentioned above.

The Arm ecosystem has steadily grown, providing a robust alternative to x86 architecture, especially in server environments. SiPearl, a European innovator, is distinguished by its high-performance, low-power processors optimized for supercomputing and AI. These processors, designed for memory-bound workloads, are backdoor-free and fully auditable, making SiPearl a strong contender in the European AI market.

Since 2008, Arm-based server CPUs have matured across various software ecosystems, now robustly supported by hypervisors like Hyper-V and KVM, operating systems including Linux and Windows, and containerization technologies like Docker and Kubernetes. The server market is predominantly controlled by original equipment manufacturers (OEMs) such as Dell Technologies, HPE, and Lenovo, which are primarily based in the USA and Asia. Major suppliers of CPUs and graphics processing units (GPUs), such as Intel, AMD and NVIDIA, are



Source: Gartner research and Boston consulting group

Figure 1: Semiconductor market by end application industry

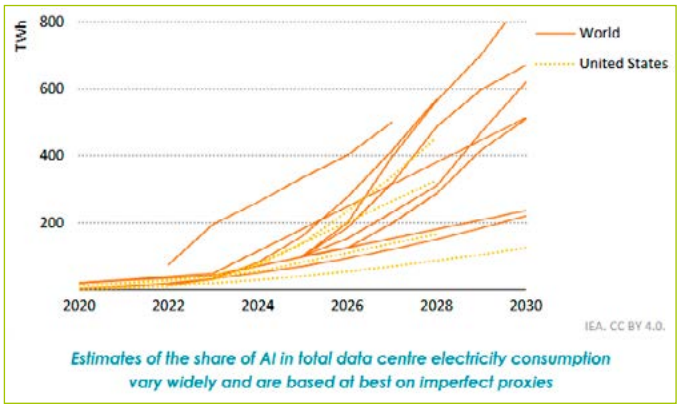


Figure 2: Datacentre power consumption forecast to 2030

Source: IEA analysis based on data from Deloitte (2024), Gartner (2024), Goldman Sachs (2024), Schneider Electric (2024), SemiAnalysis (2024), and Shehabi, et al., (2024).

American. Both OEMs and CPU / GPU suppliers manufacture their products in the USA or Asia, either internally or with companies such as TSMC and Samsung. This geographical concentration underscores the predominance of high-tech companies outside of Europe.

Ensuring Europe's technological sovereignty involves developing and supporting European players to mitigate the risks associated with supply-chain disruptions and geopolitical tensions, thereby driving innovation and economic growth. The Atos business Eviden plays a critical role in the European server market, focusing on high-performance computing solutions. SiPearl designs CPUs, collaborating with Arm, Synopsys, and Alphawave, and partnering with OEMs, original design manufacturers (ODMs), and end customers to develop optimal architectures. As part of the Arm ecosystem, SiPearl also provides comprehensive software development services, further bridging the gap between design and implementation to ensure innovative and market-aligned solutions.

SiPearl's contribution to the RISER Project

SiPearl leverages advancements from the European Processor Initiative (EPI) and EU Pilot projects. The company's first product, Rhea1, is an Arm-based server processor that leverages RISC-V cores for embedded functions.

The RISER project, which is funded by the European Union (EU), aims to develop and validate open-source designs for standardized form-factor system platforms suitable for supporting cloud services and deploying cloud applications (see *HiPEACinfo* 68 p.38). This initiative reinforces Europe's strategic autonomy in the semiconductor market.

SiPearl's contributions to RISER include developing the Seine Reference Platform for the Rhea1 processor and integrating it with the RISER PCIe acceleration card. This platform addresses the requirements and use cases defined within the RISER project, and is designed for applications in cloud and AI environments. SiPearl's ongoing efforts include developing drivers for the RISER accelerator and evaluating various use cases, with active contributions to the Data Center Security and Control Module (DC-SCM) and the Host Processor Module (HPM) being created as part of the project.

SiPearl is also actively engaged in the follow-up project to RISER, the EU-funded HIGHER project. This initiative focuses on developing open-source designs for high-density rack-scale systems that support cloud and edge services, utilizing Arm CPUs and RISC-V EPAC processors (as reported in *HiPEACinfo* 74 p.42). HIGHER is adopting the Open Compute Project (OCP) Server family of standards to build processor modules, aiming to provide modular rack systems with reusable, standards-based infrastructure. This collaborative effort includes 11 partners from industry and academia, with the aim of driving innovation and securing Europe's position in the global semiconductor landscape.

As part of the HIGHER project, SiPearl aims to develop a dual socket HPM blade based on Rhea2 that could leverage the accelerator developed within the RISER project (see Figure 3).

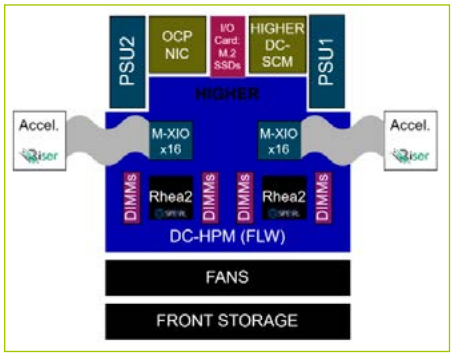


Figure 3: Example of the Host Processor Module with Rhea2 in a double socket configuration

FURTHER READING:

RISER: RISC-V for Cloud Services riser-project.eu

HIGHER: European Heterogeneous Cloud/Edge Infrastructures for Next Generation Hybrid Services higher-project.eu

Thangam, D. et al. Impact of Data Centers on Power Consumption, Climate Change, and Sustainability, March 2024, Computational Intelligence for Green Cloud Computing and Digital Waste Management (pp.60-83) Publisher: IGI Global Publishers, USA

RISER and HIGHER have received funding from the European Union's Horizon Europe research and innovation funding programme under grant agreement numbers 101092993 and 101189612. 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.

As 'vibe coding' is hyped as the future of programming, artificial intelligence (AI) is disrupting both software and hardware development. Analysing the state of the art and future directions, the HiPEAC Vision 2025 has a dedicated chapter by Paul Carpenter (Barcelona Supercomputing Center) covering the evolution of tools to develop computing systems. In this article, we summarize the chapter's main findings.

The HiPEAC Vision 2025 on tools

The history of hardware and software development has been one of increasing abstraction, shifting the focus from *how* things are implemented to *what* needs to be achieved. The primary factors that have fuelled this progression have been improvements in processor speed and compiler optimization. Today, we are witnessing a major leap forward: AI-powered development tools are widely used to generate and refine hardware or software directly from a natural-language description.

The HiPEAC Vision 2025 has two broad recommendations for Europe:

1. promote the use of AI in software development, and
2. promote the use of AI in hardware development.

However, these come with a caveat: robust measures should be implemented to ensure correctness, safety, security, confidentiality, and regulatory compliance, and human oversight should remain part of the process.

It is likely that all aspects of code development, debugging, optimization and maintenance will shift to using natural language as the bridge between the human and machine. However, it remains uncertain whether the ambitious vision of an AI-based model directly transforming 'natural language to transistors or machine code' will ever be feasible. One or more levels of abstraction between natural language and machine-level code will remain necessary; abstraction not only helps mitigate ambiguity and complexity but also provides modularity and structure.

In addition, while AI-based tools are on the ascent, there is still a place for more 'traditional' tools. Traditional optimization algorithms will still have a place at the lowest level, and there is a significant opportunity for AI-driven tools to drive developer tool use through natural language interaction, automate tool integration within a larger AI controlled workflow, and translate cryptic error messages into higher-level code suggestions.

A European toolbox for next-generation computing systems

Europe's universities, research centres and companies must be at the forefront of basic research in AI, pursuing important research topics such as the following:

- Correctness, safety and security, using formal methods.
- Programming languages and abstractions.
- Open-ended problems in hardware design, such as architecture search and exploration.
- Multi-criteria optimization of neural networks, particularly in the face of increasing complexity.

AI development tools will serve as the backbone of the digital economy, facilitating the creation of chips, communication networks, cloud infrastructures, middleware, and applications. The race to build AI development tools is, in essence, a competition for leadership in the global tech economy. If Europe lags behind in this area, it will become dependent on foreign suppliers whose interests may not align with European priorities. This dependency could even lead to hardware and software being compromised or containing hidden backdoors, creating national security risks.

The increased use of AI-based tools also raises questions about the future of the workforce in the hardware and software industries, as well as how education and training can best prepare developers. While technologies have speeded up the development of hardware and software, the need for experienced developers and designers is unlikely to disappear any time soon, although the skills required may evolve. In any event, AI is likely to serve as a powerful assistant to, rather than replacement for, human developers.

vision.hipeac.net/chapters--tools.html

The DISCOVER-US vision paper, a collaborative effort between scientists in Europe and the United States, covers four research streams, each with a dedicated chapter. In this article, in keeping with the theme of this magazine, Ovidiu Vermesan (SINTEF), the vision paper's coordinator, highlights some of the key findings from the fourth research stream: collaborative programming frameworks and software development tools.

DISCOVER-US: Focus on tools

Collaborative programming frameworks and software development tools for distributed computing and swarm intelligence

In today's rapidly evolving technological landscape, software development is undergoing a fundamental transformation. The increasing complexity of software, coupled with its deep integration into distributed and intelligent systems, has rendered traditional development methods insufficient. This is highlighted in the recent vision paper, *Distributed Computing and Swarm intelligence*, launched this year by the DISCOVER-US consortium, which is a collaboration between Europe and the United States, fostering precompetitive research in the areas of the compute continuum, distributed computing, artificial intelligence (AI), and swarm intelligence.

The authors argue that a new generation of collaborative programming frameworks and advanced software tools has emerged, empowering developers to work together seamlessly, regardless of their physical location or level of expertise. This new paradigm is not without its challenges. Enabling real-time collaboration among distributed teams introduces significant hurdles, including synchronization issues, network latency, and the complexities of version control. Maintaining consistency across a constantly changing codebase, resolving conflicting updates, and ensuring robust security for sensitive intellectual property are critical concerns. Furthermore, the cognitive load on developers managing these intricate systems can be immense, often limiting productivity.

The industry is actively moving towards integrated, intelligent software development platforms that consolidate various tools into cohesive frameworks. This shift addresses the fragmented nature of older toolchains, providing a more unified experience across the entire software development life cycle, from initial planning and design to testing, deployment, and ongoing maintenance. A growing emphasis on compliance and trust is also shaping this landscape as the legal and security risks associated with software tools, particularly those leveraging AI, become more pronounced.

A key trend in programming for these next-generation systems is the adoption of task-based models. This approach abstracts the core logic of an algorithm from the specific hardware on which it runs, allowing developers to define tasks with associated metadata and constraints. This model is particularly well-suited for the 'computing continuum', a fluid environment of cloud, edge, and local devices, as it enables intelligent scheduling and resource allocation.

Artificial intelligence, especially generative AI, is playing an increasingly pivotal role in this evolution. AI is no longer just suggesting snippets of code; it is now capable of generating entire functions, optimizing existing codebases for better performance and energy efficiency, and even translating code between different programming languages. By analysing vast datasets of open-source code, large language models (LLMs) can automate many of the more tedious aspects of programming. AI is also being used to predict the best deployment configurations, manage resources dynamically, and enhance the reliability of continuous integration and deployment (CI / CD) pipelines.

Looking ahead, the future of software development is poised for even greater abstraction and automation. The rise of low-code and no-code platforms is set to democratize development, enabling individuals without formal programming skills to build and deploy applications. Other concepts, such as digital twins and immersive triplets (virtual models of physical objects with spatial information) and 'infrastructure as code', are further blurring the lines between software and the physical world, allowing for increased levels of automation and control.

AI agent-based frameworks and swarm intelligence represent another exciting frontier. These systems use decentralized, self-organizing AI agents that can collaborate to perform complex ▶

tasks, adapting dynamically to changing conditions. Future research will likely focus on creating modular frameworks that integrate these AI and swarm paradigms, paving the way for highly adaptive, resource-efficient, and inclusive development ecosystems where AI-assisted 'developers' collaborate with their human counterparts.

Read the full vision paper
bit.ly/DISCOVER-US_vision_pdf
Get up to date on the latest tools in the DISCOVER-US webinar videos
bit.ly/HiPEACTV_DISCOVER-US

Developing robust tools and platforms: DISCOVER-US exchanges

DISCOVER-US is funding research exchanges at top institutions in the US, awarded through competitive calls. Five exchanges have taken place so far, as follows:

Daniel Balouek (Inria) visited the SCI Institute at the University of Utah, led by Manish Parashar. The scientific objective was to design software abstractions to facilitate the programming of urgent analytics workflows –specifically, to extend the capabilities of existing middleware to provide a single system image for managing a computing continuum platform. The driving application is an earthquake early warning system.

Luigi Pomante (Università degli Studi dell'Aquila) visited Edward Lee's group at the University of California Berkeley to study the feasibility of the adaptation of HEPSYCODE, a system-level HW/SW co-design methodology, to the Lingua Franca framework, in order to address distributed systems and the computing continuum. The focus was on a feasibility analysis of the adoption of Lingua Franca as a specification language for HEPSYCODE.

Delia Velasco Montero (Estación Biológica de Doñana-EBD-CSIC) visited Rajesh Sankaran at Argonne National Laboratory. The exchange focused on developing an integrated system for remote environmental monitoring using embedded devices and LoRaWAN technology on nodes in the Sage cyberinfrastructure. The work included building a data pipeline for encoding and transmitting data from remote areas, and automatic decoding and cloud publishing by a nearby Sage node.

Dimitrios Spatharakis (National Technical University of Athens) visited the PEARL Lab at Virginia Tech, headed by Dimitrios S. Nikolopoulos, where he worked on scheduling for edge AI. His research during this time focused on two main areas: speculative decoding techniques for large language models (LLMs) and emerging strategies for the efficient deployment of multi-modal models at the edge.

Germán Moltó (Universitat Politècnica de València) visited Kate Keahey's group at Argonne National Laboratory. During this visit, the team integrated Chameleon computing testbed managed by the University of Chicago as a cloud backend to the Infrastructure Manager developed by UPV. This allows the deployment of transatlantic computational testbeds that aggregate computational resources from large-scale distributed infrastructures, including the EGI Federated Cloud. The transatlantic testbed was then used to execute computationally intensive use cases such as AI-based fish detection and flood hazard and impact modelling, from the European projects iMagine and interTwin, respectively.

In this article, we learn about cutting-edge cybersecurity research at Bosch, from honeypots to fuzz testing to reliable software bills of materials.

Securing tomorrow's vehicles, today
Cybersecurity updates from Bosch

HONEYPOTS FOR AUTOMOTIVE THREAT INTELLIGENCE AND INTRUSION DETECTION


Niclas Ilg and Dominik Germek (Robert Bosch GmbH)

The rapid shift towards software-defined vehicles – driven by connectivity, high data rates, over-the-air updates, and advanced driver-assistance systems – presents significant security challenges. While the 2015 Jeep hack showcased the devastating potential of remote attacks, a similar large-scale incident hasn't recurred.

An assumed reason for this is the unfamiliar technology in vehicles. Thus, increased connectivity isn't the only upcoming risk; easier-to-access technology also lowers the barrier for malicious actors. The adoption of familiar technologies like Ethernet/IP and Linux within vehicles makes them more attractive targets than ever, potentially eclipsing the need for complex electronic control unit (ECU) manipulation and proprietary bus analysis.

The automotive attack surface currently unfolds in two ways. On one hand, the backend systems and infrastructure that are mostly reachable via the internet are targeted. In 2023, researchers exploited backend systems of multiple different original equipment manufacturers (OEMs) to track vehicles, gain remote code execution, and even receive full access to the cars (see 'Further reading'). Web interfaces of electric-vehicle (EV) charging stations are also repeatedly targeted by attackers. On the other hand, attackers target the vehicle itself; visible interfaces like USB, short-range wireless technology like Bluetooth and Wi-Fi, infotainment applications, as well as remote connectivity to the infotainment and telemetry units, are just some entry points.

Example of a honeypot application that could be deployed on the high-performance unit

Much of the current understanding of the attack surface's exploitability comes from white-hat researchers. Studies from 2021 and 2024 (see 'Further reading') provide a comprehensive list of attack vectors against vehicles and automotive Ethernet. The high number of remote attacks (41%) demonstrated illustrates the need for threat-intelligence as well as prevention measures inside the vehicle.

Using honeypots to deflect attacks

For both threat intelligence and prevention measures, honeypots – decoy resources aiming to attract adversaries in order to learn more about their approach – can help. In the past, the automotive domain was unwelcoming territory for honeypot systems for various reasons. It was difficult to take advantage of threat intelligence on the internet because updating systems was highly expensive: customers needed to visit a garage for every update. Furthermore, deeply embedded systems quickly raise suspicion on the internet, and only a few people know how to manage them. In the vehicle, ECUs lack resources for honeypot deployments and the ability to react quickly to findings.

However, this is now changing: emerging electrical / electronic (E/E) architectures are introducing centralized computing, virtualization, and over-the-air updates. These features provide a technological basis for honeypot deployments inside the vehicle. They could even be activated in an adaptive manner when threat intelligence indicates an increased level of risk.

Additionally, honeypot deployments on the internet can provide early warning signals to defenders. Due to the increased adoption of technologies from internet-native domains, honeypot deployments no longer stand out and can add to the threat-intelligence effort. The acquired knowledge can then be patched onto the vehicle over the air. For more on the challenges and possibilities of automotive honeypots , we invite you to read our paper, cited in 'Further reading', below.

FURTHER READING:

Miller, C. and Valasek, C. 'Remote Exploitation of an Unaltered Passenger Vehicle', IOActive Technical White Paper, 2015
bit.ly/Jeep_hack_paper_2015

'Web Hackers vs. The Auto Industry', 2023
samcurry.net/web-hackers-vs-the-auto-industry

Pekaric, I. et al. 'A taxonomy of attack mechanisms in the automotive domain', Computer Standards & Interfaces, Volume 78, 2021
doi.org/10.1016/j.csi.2021.103539

De Vicenzi, M., et al. 'A Systematic Review on Security Attacks and Countermeasures in Automotive Ethernet', 2024, ACM Computing Surveys, Volume 56, Issue 6, Article No. 135, pp.1-38
doi.org/10.1145/3637059

Ilg, N. et al. 'Work-in-Progress: Emerging E/E-Architectures as Enabler for Automotive Honeypots'. 2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Vienna, Austria, 2024, pp. 361-366
doi: 10.1109/EuroSPW61312.2024.00046.

GENERATION OF RELIABLE SOFTWARE BILL OF MATERIALS FROM BINARIES

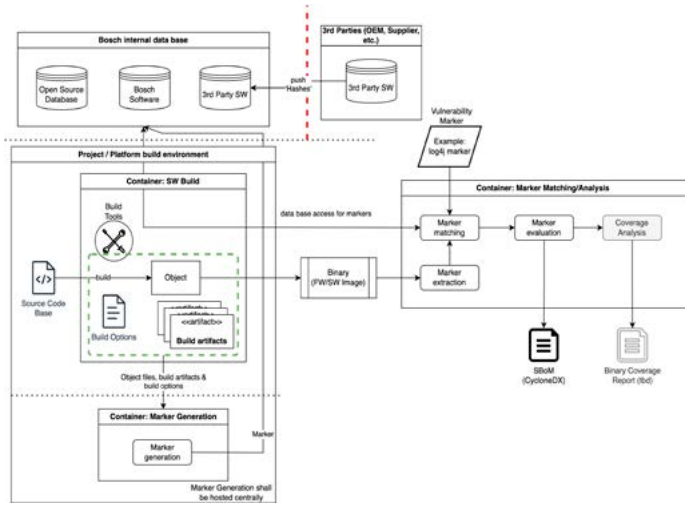
Martin Ring and Lars Vogel (Robert Bosch GmbH)

New laws, orders, regulations, and upcoming requirements (including the US Executive Order 14028, Cyber Resilience Act (CRA), and Digital Operational Resilience Act (DORA)) mandate the availability and constant updates of bills of materials (BoM) not only for hardware but also for software (internal, operating-system software (OSS) and software from customers and suppliers). This necessitates a detailed overview of all libraries and functions that are part of a product placed on the market. This is also necessary to improve and enable vulnerability handling.

There are tools available that alleviate this problem for certain programming languages, such as npm or PyPI as package managers, or for specific software types, like the Kubernetes BoM Tool, which allow for the generation of a software bill of materials (SBoM).

However, since companies like Bosch are not limited to these programming languages and utilize a significant amount of bare-metal proprietary software, these tools are unable to generate a satisfactory SBoM. To help resolve this, we ventured in two directions: first, we researched what is available in adjacent fields and found interesting approaches for finding intellectual-property (IP) infringements in software. We implemented a proof of concept that yielded good results with certain limitations, mainly the needed time to generate an SBoM. Our second approach was a survey of commercially available tools. To properly evaluate our own and commercially available solutions, we conducted multiple tests. As mentioned above, we are interested in finding internal, OSS and proprietary packages and libraries with patch-level granularity (semantic versioning: x.y.z – x, major version; y, minor version; z, patch level). First tests were conducted using the publicly available firmware of the LTE modem from Bosch Motorsports. We also examined IoTGoat, a deliberately vulnerable firmware. Both of these are based on OpenWRT and were used to test for the capabilities to find OSS components. The planned pipeline how this approach will be rolled out is depicted in the figure below.

To further assess the applicability regarding internal and proprietary software, we picked an exemplary deeply embedded system with bare-metal Bosch software. For the evaluation we built three integration levels (software versions) for this system. Some tools allow the buildup of own databases that can be used for identifying proprietary libraries. When this possibility was not given, the commercial tools either found



Planned pipeline for the software bill of materials

nothing or only found false positives. With the possibility to build up a database, patch-level granularity was possible – for example, from one integration to the next one we only changed one variable type (uint32 to uint64) and the corresponding pre-compiler directives, and this was successfully detected.

With our approach we were able to identify the software parts included: although we were not able to get down to patch-level granularity, we were still able get a first step to vulnerability management for example, we could find out if we used the vulnerable log4j library for all our proprietary binaries.

As we now have ways of generating SBoMs for arbitrary software, what is missing is information how much of the binary is covered by the generated SBoM. This is still an open research question and the next step we plan on tackling.

SMART FUZZ TESTING FOR AUTOMOTIVE SYSTEMS: ESCRYPT CYCURFUZZ


Marcos Cardoso (ETAS GmbH) and Christopher Huth (Robert Bosch GmbH)

Automotive systems are becoming increasingly complex, and with this complexity attack vectors increase, too. International regulations and standards, such as UN R-155 and ISO/SAE 21434, require cybersecurity testing as part of such systems' development and validation process. Fuzz testing, or fuzzing, is explicitly recommended by ISO/SAE 21434 to evaluate the resilience and robustness of automotive systems.

Fuzz testing is a method of repeatedly sending semi-valid input to a test target for processing, while observing the target's behaviour to fine-tune the test input. While fuzz testing can automatically test software, it usually requires effort to set up and to customize workflows, especially for cyber-physical systems such as automotive control units.

Enhancing the quality of automotive software

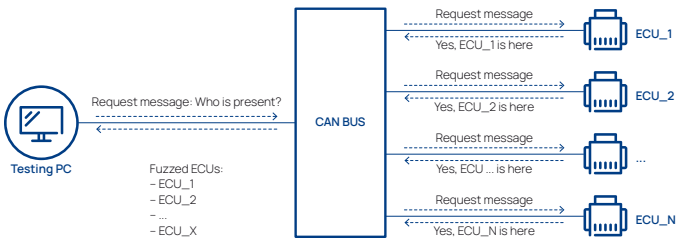
A smart automotive fuzz-testing tool allows for customization, automation, and acceleration of the test procedure. It constantly improves its testing and is efficiently embedded in your development process. ESCRYPT CycurFUZZ is a state-of-the-art fuzz-testing tool that helps users meet current regulations and standards. Automotive cybersecurity testing knowledge is built in and is used to assess the security maturity of automotive systems, thus improving the software quality of products throughout development. Generally, fuzz testing tries to trigger all different paths in a test target; once unwanted behaviour is triggered and observed, the root cause can be found and fixed.

Exceptional defect-detection rate and execution speed

ESCRYPT CycurFUZZ addresses the main automotive protocols and supports various fuzzing modes. With an exceptional defect-detection rate, it also features a dynamic timing feature, which boosts test performance, as well as the ability to add customized scripts; together, these features increase overall execution speed and greatly extend the fuzz-testing ability. Full-headless mode is supported and compatible with all common bus access tools (DB9, USB connector) and REST API for external integration.

CycurFUZZ offers versatile fuzzing capabilities at the component, system, and vehicle levels. Test targets can be either physical or virtual ECUs, including hardware-in-the-loop and software-in-the-loop setups. Thanks to ESCRYPT's professional security testing services, the product also benefits from continuous enhancement of automotive security testing knowledge.

In summary, ESCRYPT CycurFUZZ represents a cutting-edge solution for automotive cybersecurity testing, ensuring compliance with essential regulations while enhancing software quality and resilience against cyber threats.



One of the EuroHPC supercomputers, Karolina – hosted by IT4Innovations National Supercomputing Center in Ostrava, Czechia – offers significant potential to turbocharge business processes. In this article, Kateřina Fraisová (IT4Innovations, National Centre of Competence in HPC Czechia) explains how Karolina has been adapt simulations of paper machines so that they can run efficiently on the supercomputer.

Adapting simulations for paper machine development to the Karolina supercomputer

Bellmer is a company specializing in manufacturing machines for the paper industry. It provides complete fibre-processing solutions, including raw-material processing machines, paper technology, and separation and dewatering systems.

In the paper-machine development process, they use complex numerical simulations to optimize their designs and production processes. Bellmer uses the software OpenFOAM to simulate the behaviour of mainly non-Newtonian fluids, which behave significantly differently and non-intuitively compared to conventional ones.

But how can this tool be used correctly on supercomputers to achieve the desired results quickly and efficiently?

Bellmer decided to seek an answer to this question with experts from IT4Innovations National Supercomputing Center. In the future, they would like to use the Karolina supercomputer operated by IT4Innovations for their calculations. However, this requires optimal setup and proper configuration.

A team of high-performance computing experts have managed to design modifications to their existing simulations so that they can run efficiently on supercomputing systems. 'The proposed modifications were designed with an emphasis on the correct distribution of computational tasks and ensuring maximum use of available resources,' explains Tomáš Karásek, Head of the National Centre of Competence in HPC Czechia.

The configuration of OpenFOAM codes has also been modified so that they can perform complex simulations using the maximum computing power. This included advice on adequately setting up the parameters and task structure for optimal results for simulations of different sizes.

Thanks to this collaboration, Bellmer now has a better understanding of how to harness the power of supercomputers for their demanding calculations and how to set up its simulations for the Karolina supercomputer in Ostrava as efficiently as possible.

Our goal is to continue to support their transition to high-performance computing and ensure that their future paper machine development is supported by state-of-the-art computing facilities.

Petr Fryčák, head of research and development at Bellmer, comments: 'At the beginning of our collaboration, numerical simulation of the wastepaper-pulping process in a vertical pulper was solved. In wastepaper-processing technology, the vertical pulper is the first machine that processes the raw material. Wastepaper is loaded into the pulper tub together with water as the process medium. The rotor agitates the paper and water mixture by spinning and generating a shearing force in the pulper tub, causing the wastepaper to be pulped. This process can be described as breaking down the paper into smaller particles, i.e. fibres.

The pulping process efficiency depends on many factors, the most important of which are the shape and size of the tub and rotor, the rotor speed, the type of wastepaper (or pulp) to be processed, and the proportion by mass of paper (fibres and impurities) in the water suspension.

A numerical simulation of the pulping process was used to investigate how changing the size and shape of the guide vane located in the pulper tub affects it. At the same time, a calculation was performed to verify the effect of rotor-blade height and rotor speed on the pulping intensity.'

While new technologies are needed to help keep European industry competitive, deploying robotics and artificial intelligence too rapidly can lead to increased occupational risk, including accidents and stress. In this article, Isabelle Dor (CEA) introduces the EARASHI project, which is tackling this issue head on.

Funded by the European Union

CENTAUR TEAMS: EARASHI'S HUMAN-CENTRED APPROACH TO AI-ENABLED WORKPLACES

According to the European Commission, 'Industry 5.0' 'places the well-being of the worker at the centre of the production process and uses new technologies to provide prosperity beyond jobs and growth while respecting the production limits of the planet'.

The rapid development and implementation of mobile and / or collaborative robots in industries can result in more dangerous situations for operators, with accidents being more likely where knowledge of the actual risks is lacking. The implementation of automation and robotization solutions can induce work-related stress, especially in older generations, and it also requires new knowledge for current and future generations at the workplace. In order to ensure the wellbeing of workers, especially in manufacturing, a human-centred approach is critical.

The Horizon Europe-funded EARASHI project aims to help workers in their daily activities and improve their working conditions, thereby contributing to an increase in productivity. It explores the shift from Industry 4.0 to Industry 5.0, emphasizing the importance of human-centric approaches in manufacturing. It also delves into Operator 4.0, stressing the optimization of human-machine interaction for success, focusing on technology acceptance, human factors, and human-robot interaction.

In line with the Industry 5.0 vision of collaboration between technology and humanity, EARASHI prioritizes human-centric approaches, enhancing job satisfaction and performance, ensuring equitable technology access, and addressing the digital divide. To this end, EARASHI provides insights into methods, tools and equipment for inclusive design.

As part of the project, 10 challenges were defined to foster the uptake of advanced-digital and eco-responsible technologies provided by small and medium companies. Ten companies were selected through two open calls. Among the successful candidates were projects to develop products to provide robotic assistance, projects related to the digitalization of their environment, and projects that aim to enhance human wellbeing and workplace safety.

By investing early in competitive emerging technologies, EARASHI is contributing to a sustainable and human-oriented EU industry, addressing the challenges posed by the digital transition while maximizing its benefits for workers and businesses alike.

PROJECT NAME: EARASHI: Embodied AI / Robotics Applications for a Safe, Human-oriented Industry
START / END DATE: 01/09/2022 – 28/02/2026
GRANT AGREEMENT: 101069994
KEY THEMES: collaborative embodied AI (robotics systems), empowering end-users and workers, occupational health and safety, manufacturing
PARTNERS: **France**: Commissariat à l'énergie atomique et aux énergies alternatives (CEA), Minalogic, STMicroelectronics, Blumorpho; **Belgium**: Flanders Make, AMS Belgium, CECIMO; **Germany**: Steinbeis Innovation; **Portugal**: INEGI; **Spain**: Mondragón, Ikerlan, Aldakin.
BUDGET: EU contribution € 5 million
€ 2 million financial support to third parties
earashi.eu

In 2018, Alberto Ros (University of Murcia) was awarded an ERC Consolidator Grant for his project ECHO (Extending Coherence with Hardware-Driven Optimizations). In this article, Alberto explains how the project helped shape his own career and that of his PhD students, while former PhD students Ashkan Asgharzadeh and Sawan Singh detail their thesis work carried out as part of ECHO.

Multiplying ECHO

How an ERC grant sparked new collaborations and shaped research careers

The ERC grantee: Alberto Ros

'We are pleased to inform you that your proposal has been retained for funding…' Reading these words after years of applying for an ERC grant brought a tremendous sense of fulfilment. Yet I could not have anticipated how much my working style and habits would change. Almost overnight, I transitioned from having no PhD students to advising five early-stage candidates. With this came a greater sense of responsibility, and my focus quickly shifted from the project itself to the people behind it. In essence, when the team is happy and motivated, the project naturally advances toward its goals.

The core idea of ECHO is to achieve non-speculative execution while retaining the benefits of speculation – essentially, enjoying the performance of traditional speculative execution but with the guarantee of success in certain cases. Our main focus is on memory instructions such as load, store, and read-modify-write.

A key outcome, presented at ASPLOS 2025, was a method to identify immutable critical sections, allowing them to execute non-speculatively yet concurrently with other critical sections. This approach guarantees successful execution of atomic regions, eliminating the costs and risks of speculation and retries. The result is greater parallelism and significantly reduced execution time for applications.

We also proposed new prefetching techniques, showing they can be trained non-speculatively (at commit) without performance loss, thus protecting prefetchers from speculative side-channel attacks. This work led to an ERC Proof of Concept Grant (Berti-Chip), aimed at testing these prefetchers in RISC-V cores in collaboration with the Barcelona Supercomputing Center and Semidynamics.

ECHO has fostered new collaborations, including regular research visits from Professor Biswabandan Panda that led to fruitful work on microarchitectural predictions. Two ERC-funded postdocs, Rubén Titos and Juan M. Cebrian, are now faculty members in our department, further strengthening our research group. Three PhD students are expected to graduate in the present year. Additionally, the first two ECHO PhD graduates, Ashkan Asgharzadeh (now at the Barcelona Supercomputing Center) and Sawan Singh (now at AMD), completed their degrees in November 2024. Ashkan, co-advised with Stefanos Kaxiras, focused on optimizing the execution read-modify-write operations, while Sawan, co-advised with Alexandra Jimborean, leveraged compiler support for non-speculative detection. Below, they detail their PhD theses.

Looking back to 2018, I could not have foreseen exactly how the ERC would shape my career, but I am immensely proud of both the technical achievements and the growth of our team over the last few years.

The PhD theses

Ashkan Asgharzadeh Donighi: 'Unfencing atomic operations: optimizations and applications', November 2024

Atomic read-modify-write instructions (atomic RMWs) are hardware instructions that hardware vendors provide to implement atomicity. This thesis observed that the traditional implementation of atomic RMWs impose unnecessary cost. To improve the performance of atomic RMWs, it proposes an efficient implementation while respecting the atomicity and memory consistency of these instructions. The thesis also suggests a new application for atomic RMWs beyond providing atomicity.

Traditional implementation of atomic RMWs disabled instruction and memory-level parallelism using memory fences. This means that atomic RMWs had to execute in isolation of other memory instructions. The first contribution is 'Free Atomics', a fence free (unfenced) implementation of atomic RMWs that not only removes the costs imposed by memory fences, but also enables speculative execution of atomic RMWs. Free Atomics is a collaboration between three different European universities that was realized thanks to a HiPEAC collaboration grant.

Ashkan Asgharzadeh's thesis defence. Ashkan is third from the left, Alberto second from the right

Our paper on Free Atomics was published at ISCA 2022, and nominated for the best paper candidate.

The second contribution is hardware cache-locking, a mechanism used by atomic RMWs to provide atomicity, which can solve the false-sharing problem between non-atomic RMWs. The proposal was published at ICCD 2024.

The last contribution analysed the impact of contention over unfenced atomic RMWs. It makes the observation that Free Atomics as a modern atomic implementation might achieve sub-optimal performance if it always initiates 'eager' execution. If an atomic is contended – i.e. the cacheline used by an atomic is requested by more than one core – to achieve optimal performance it should use 'lazy' execution, i.e. delay execution while certain conditions are met. A dynamic mechanism was proposed to predict contention over cachelines used by atomic RMWs and decide eager or lazy execution per atomic at runtime. This proposal was published at HPCA 2025.

Sawan Singh: 'Microarchitectural Optimizations for an Efficient Utilization of Processor Resources', November 2024

Modern processors achieve high performance through deep pipelines, out-of-order (OoO) execution, and optimized memory hierarchies. Among these techniques, OoO execution is particularly crucial as it allows instructions to execute as soon as their operands are ready rather than strictly following program order. However, ensuring correctness in OoO execution requires complex hardware structures such as the reorder buffer (ROB), load queue (LQ), store buffer (SB), and store queue (SQ). These structures, while essential, can become bottlenecks if they are too small, leading to stalls, or too large, causing inefficient searches and increased delays. Additionally, frontend inefficiencies, particularly with the micro-op (µ-op) cache, can further hinder performance, especially in workloads with large instruction footprints, such as those found in data centres and servers.

This thesis explores several innovative techniques to improve processor efficiency by optimizing these structures and mitigating frontend stalls. One approach, 'Regional Out-of-Order Writes' (ROOW), presented at PACT 2020, enables safe out-of-order writes within data-race-free regions, reducing processor stalls by 7.11% and improving execution time by 8.13%. Another contribution, 'Compiler-Assisted Efficient Load-Load Ordering' (CELLO), presented at PACT 2023, reduces unnecessary load queue searches by 47%, leading to a 2.8% improvement in execution time while cutting LQ energy expenditure by 33%.

To further enhance performance, an instruction fusion mechanism called 'Helios', presented at MICRO 2022, fuses non-consecutive instructions, reducing stalls in key processor structures and improving overall execution by 8.2%. Lastly, to address frontend stalls, a novel micro-op cache prefetching mechanism (UCP), presented at ISCA 2024, is introduced, which prefetches µ-ops for hard-to-predict branches, yielding a 2% performance improvement in server workloads. Helios and UCP were both developed partly in conjunction with Arthur Perais (TIMA), during a research stay facilitated by a HiPEAC collaboration grant.

By refining the utilization of key processor structures and improving instruction delivery, this work demonstrates significant execution efficiency gains, with overall performance improvements ranging between 2% and 8%. These optimizations help reduce stalls and maximize processor throughput, paving the way for more efficient and scalable microarchitectures.

Sawan Singh (third from left) at his thesis defence

Industry or academia? Early-career researchers are often faced with a choice between applying their skills practically in an industry setting or continuing their studies in academia. But what if you could have both? In this personal account, Sam Coward (currently working on a postdoc at the University of Cambridge) relates how integrating a full-time job at Intel with his PhD at Imperial College, London, worked to everyone's advantage.

A foot in both camps

How an unconventional industry-sponsored PhD led to a new research direction

Each research community can point to figures that have successfully walked the line between academic research and industrial application. For example, at AWS, Byron Cook, who is also a UCL professor, has added substantial value by applying formal methods to almost every product the company offers. But how do you learn the necessary skills to balance on this tightrope? This article describes my experience as an early-career researcher trying to span this divide by integrating a full-time PhD with a full-time industrial position. We informally called this model for doctoral studies an 'acadustrial' PhD. To many this sounds unsustainable, but if the right common understanding is in place, it can work to everyone's advantage.

My 'acadustrial' PhD opportunity came about after a 2019 internship at Intel with Theo Drane's Numerical and System Level Design Group (NSD). At the conclusion of this internship, we met with Theo's former PhD supervisor, Professor George Constantinides, where the PhD proposal was put together. This was followed by more than a year's worth of legal infrastructure development, establishing the intellectual property (IP) agreement between the two organizations. In April 2021, my PhD and my Intel job started simultaneously, with Theo setting out the results that Intel were interested in: 'an order of magnitude development time improvement to numerical register-transfer level (RTL) optimization and numerical hardware quality'.

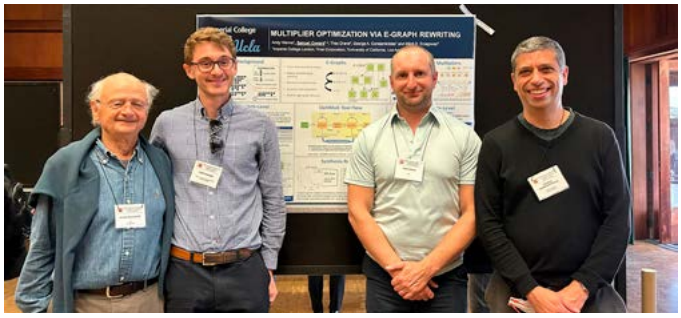
The inception highlights two key points. First, given the multi-year commitment to support an individual, company internship programmes serve as a natural pathway to recruit candidates in collaboration with the academic supervisor. Second, the legal infrastructure must be in place before the PhD begins. This establishes clear IP ownership from the outset, commits the company to timely publishing reviews and guarantees access to industrial data and testcases.

Complementing research with practical application

During the PhD, I was physically located in the university but was fully integrated into both the academic and industrial groups. In a typical week, 75% of my time was dedicated to my core PhD research, which pursued a novel approach to RTL optimization and verification. With the remainder, I contributed to wider Intel projects, typically optimizing a given circuit design, exposing me to real-world challenges faced by industrial engineers. These frequent interactions across Intel significantly influenced my PhD, providing countless testcases and opportunities to apply my research to valuable problems.

This connection, where my research sought to automate my manual production work, was critical to being able to maintain two roles simultaneously. By construction, the PhD model ensured continual knowledge transfer and frequent quality feedback on the research direction. The model also helped Intel

George Constantinides (middle) and Sam (left) visiting Theo Drane (right) and the NSD team at the Intel Folsom campus in 2023



George (right), Theo (middle right) and Sam (middle left), presenting joint work with collaborator Milos Ercegovac (UCLA) in ASILOMAR in 2023

NSD to develop wider academic connections that inevitably aided recruitment. According to Theo, 'NSD's seven current university collaborations owe their existence and / or strength to this sponsored research', while 'half of NSD's interns come from the network Sam Coward built'.

The weekly three-way meetings with both supervisors were of immense value, and were based on a common understanding of what quality research looks like. These frequent technical discussions guaranteed research and impact alignment, generating quality IP to the benefit of Intel. As George observed, 'having the right three people in the room' is critical for the success of the model and is best founded on a pre-existing relationship between industrial and academic supervisor. During the PhD on a small number of occasions, research alignment was stressed by diverging interests, but thanks to the regularity of the contact these were quickly resolved to everyone's satisfaction. While some may question the overt influence of industry on academic research directions, Theo's comments serve as testament to the sponsored research: it 'founded the field of E-graph based hardware design and won an outstanding PhD award'.

My hope is that this article provides an example of how industry and academia can closely align to jointly develop a student and valuable research agenda. Compared to standard industrial PhD sponsorship, it requires a greater investment for the company, but the potential returns make it a compelling model. Systematizing such an approach is not easy given the dependence on individual human relationships. As George noted, 'perhaps the key lesson here is to create the environments where such relationships and trust can grow, rather than rush into sponsored research'.

HiPEAC 2025 STEM Day students get careers inspiration at BSC

On 12 May, winners of the HiPEAC 2025 STEM Day challenge were invited to Barcelona Supercomputing Center (BSC) for a day of inspirational careers presentations and a tour of the MareNostrum supercomputer, as well as the newly installed Quantum Spain computer.

Researchers representing each department at BSC (Life Sciences, Earth Sciences, Computer Applications for Science and Engineering, Computer Sciences, and the Computational Social Sciences and Humanities Lab) gave insights into their research careers and the paths they took to get there. Laura Deshelms of the HR department also gave a presentation on what to expect from working at the research centre.

Participating students – some in person, some remotely – also gave presentations on their research, interests, and career plans.

The discussions were followed by a tour of the MareNostrum supercomputer, one of the most powerful computers in the world, and the Quantum Spain computer located in the emblematic chapel at BSC.

Abdollah Rezagholi, a student at the Universitat Politècnica de Catalunya – BarcelonaTech, commented: 'Being able to step inside one of Europe's leading HPC facilities was a truly motivating experience.

I was especially fascinated by the chance to see quantum computing devices. Seeing the hardware in person gave me a real sense of how these emerging technologies are becoming an active part of research infrastructure.

The visit gave me a broader understanding of how high-performance computing and related technologies are applied in real-world research, and it has definitely motivated me to learn more and stay engaged with developments in this space.'

HiPEAC would like to thank everyone who made this event possible, and the students who participated.

The next HiPEAC conference will take place in Kraków in January 2026, and will feature a range of student activities. Check the website for further information hipeac.net/conference

Centre of Excellence workshops at HiPEAC 2025 shift the needle in gender balance

Maria Arista-Romero, Rosa Rodríguez-Gasén and Marta García-Gasulla, Barcelona Supercomputing Center

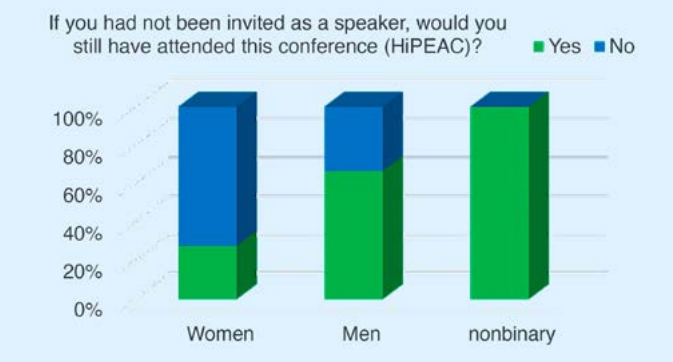
During the HiPEAC Conference in Barcelona, two workshops were organized to discuss the current status of high-performance computing (HPC) applications in Europe and ongoing collaborative efforts to enhance their scalability. The workshops featured contributions from different European Centres of Excellence (CoEs) and other relevant projects. Special attention was given to gender balance when inviting speakers, reflecting a commitment to inclusivity and diversity in the HPC community. These sessions showcased contributions from 11 of the 12 active CoEs, as well as representatives from Destination Earth, EPICURE, ENES, CASTIEL-2, and the EuroHPC Joint Undertaking. The agenda featured technical presentations, live demonstrations and a round table of experts.

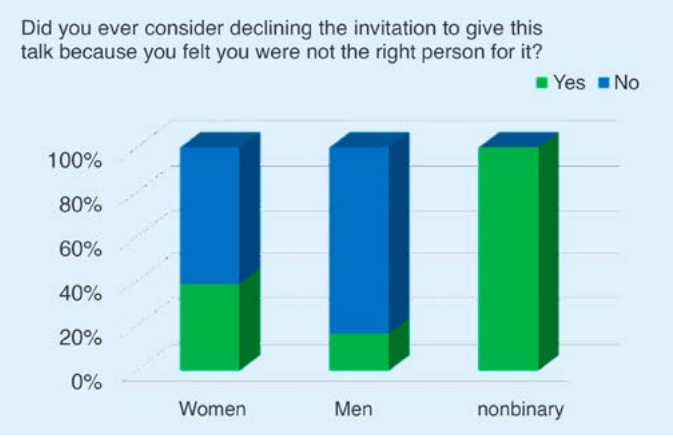
Presentations covered a range of high-impact topics, including performance benchmarking across European HPC codes, cross-CoE collaboration and training experiences, and co-design strategies for exascale systems.

Early-career researchers featured prominently among the speakers, bringing fresh perspectives and driving discussions on application readiness for exascale computing. Several talks also addressed strategies for increasing female participation in HPC.

The speaker distribution was 74% female, 22% male, and 4% non-binary, while the gender balance of the participants was 56% women and 44% men – particularly significant considering that the typical gender distribution at the HiPEAC conference tends to between 10% women / 90% men and 20% women / 80% men.

Following their participation, speakers at the workshop were invited to complete a short questionnaire. This highlighted some interesting results; for example, only 23% of the female speakers said they would have attended the HiPEAC conference, compared to 66% of the male speakers. In addition, 39% of the female speakers said they had considered declining the invitation as they felt they weren't suitable, compared to 16% of the male speakers.





FURTHER INFORMATION:

ESIWACE report on the workshop bit.ly/HiPEAC25_CoE_ESIWACE_news

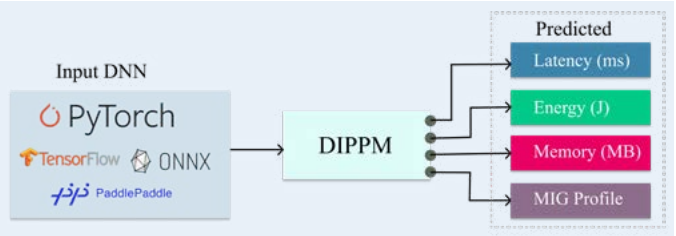
From Petascale to Exascale and Beyond: The Centres of Excellence Challenge bit.ly/HiPEAC25_CoE_workshop_1

Tackling Software Exascale Challenges: The Centres of Excellence in High-Performance Computing Perspective bit.ly/HiPEAC25_CoE_workshop_2

In this article, we learn how Karthick Panner Selvam developed predictive performance models to estimate key metrics for deep-learning (DL) workloads across different hardware resources. This research was conducted as part of the EuroHPC MAELSTROM project, funded by the European Commission and the Luxembourg National Research Fund.

Three-minute thesis

NAME: Karthick Panner Selvam
RESEARCH CENTRE: University of Luxembourg, Interdisciplinary Centre for Security, Reliability and Trust (SnT)
SUPERVISOR: Dr Mats Brorsson
THESIS TITLE: Performance Prediction Models for Deep Learning: A Graph Neural Network and Large Language Model Approach



Deep learning (DL) has revolutionized numerous fields, such as healthcare, finance, and autonomous vehicles. However, the explosive growth of these models also comes with a significant environmental and economic cost due to their immense computational demands. To mitigate these impacts, accurate performance prediction of DL models, estimating how efficiently they utilize hardware resources, is essential.

My PhD thesis addresses this critical issue by developing predictive performance models that estimate essential metrics, such as latency, memory usage, and energy consumption, for diverse DL workloads across various hardware configurations. The research was crucial because existing methods could not effectively manage the rapidly evolving complexity and diversity of DL models and hardware architectures.

One significant challenge was the need to predict performance across multiple DL frameworks like TensorFlow and PyTorch. I developed a universal tool, DIPPM, leveraging Graph Neural Networks (GNNs) to provide accurate, framework-agnostic predictions by representing DL models as generalized computational graphs. Additionally, I developed a method to optimize resource allocation using NVIDIA Multi-Instance GPU (MIG) technology, enabling efficient GPU partitioning tailored to diverse deep learning workload requirements.

Another obstacle was the scarcity of labelled performance data, a critical requirement for precise predictions. To overcome this, I introduced TraPPM, a semi-supervised learning method that utilizes unlabelled data to significantly improve the accuracy of performance predictions.

For large language models (LLMs), I faced unique challenges due to their complex structures and intensive computational demands. My solution was a tree-based prediction model that significantly improved inference speed and accuracy.

Finally, to rapidly adapt performance predictions to new hardware environments, I collaborated with researchers from Google DeepMind and Google Research to develop a multi-modal approach combining the strengths of GNNs and LMMs. This integration enables quick and accurate adaptation to new hardware, even when labelled data is sparse. This work was presented at the ICML 2024 WANT workshop.

Overall, my research uniquely contributes to creating a greener, more efficient AI landscape by optimizing computational resources and significantly reducing environmental impacts. By accurately predicting DL model performance, developers can better match hardware configurations to their specific needs, enhancing efficiency, cutting costs, and minimizing carbon emissions, thus paving the way towards sustainable AI deployment.

Together with my supervisor, Dr Mats Brorsson, we are commercializing this technology through InfraTailors.ai, a company aimed at helping organizations optimize their hardware resource selections, with the help of a FNR (Luxembourg National Research Fund) JUMP Grant.

Karthick's supervisor, **Mats Brorsson**, commented: 'This thesis tackles crucial deep-learning deployment challenges. Karthick has developed innovative, framework-agnostic performance prediction models, cleverly using GNNs, LLMs, and semi-supervised learning. His work dramatically improves accuracy, resource efficiency, and adaptability, showing deep insights into the optimization and deployment of deep learning models. I am truly looking forward to seeing what the future brings.'

FURTHER INFORMATION:
MAELSTROM project maelstrom-eurohpc.eu
InfraTailors.ai infratailors.ai

SPONSORSHIP OPPORTUNITIES

HiPEAC conference

26 - 28 January 2026

Kraków

Europe's largest computing systems research event (600+ participants, 80+ companies)

Expand your business network and meet new clients

Excellent recruitment opportunity: HiPEAC Jobs Fair and STEM Student Day

Media coverage and enhanced visibility options

Tailored sponsorship plans

	PLATINUM €10,000	GOLD €5,000	SILVER from €2,500	BRONZE €1,500
Logo on HiPEAC website and communications	✓	✓	✓	✓
Conference passes	6	3	2	1
Industry session presentation	30 min	20 min	10 min	✗
Industry exhibition booth	✓	✓	✓	✓
Privileged booth location and customized options	✓	✓	✗	✗
Eligibility to sponsor specific activities	✓	✓	✗	✗
Year-round HiPEAC Jobs support	✓	✓	✗	✗
Additional promotional opportunities	✓	✓	✗	✗

sponsorship@hipeac.net

Funded by the European Union

bit.ly/HiPEAC26_sponsorship