HiPEAC info 74

JANUARY 2025

HiPEAC 2025 Barcelona

HiPEAC turns 20: Reflections by Mateo Valero, Max Lemke, and the HiPEAC community

Twenty years in computing technologies: A retrospective

Margaret Martonosi on trustworthy computing systems, Giovanni De Micheli on EDA and Europe, and Emre Ozer on flexible electronics

contents

Max Lemke on HiPEAC and European computing policy

The HiPEAC story: How it all began

Margaret Martonosi on the power of computer architecture

EDA European style, with Giovanni De Micheli

Emre Ozer's flexible circuits

Twenty years in European tech, from processors to applications

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 YouTube 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: OFC Pictures, Adobe Stock

Design: www.magelaan.be

Editor: Madeleine Gray

The central theme of this magazine is 20 years of HiPEAC, how much our domain has changed over the last 20 years, and what is ahead of us. Twenty years is half a career: some of the HiPEAC founders, early project officers and reviewers have already retired; some, unfortunately, have passed away; others will retire in the coming decade. Mid-life HiPEAC members have never known a time without HiPEAC. For them HiPEAC is one of the certainties in their career.

Over the last 20 years, HiPEAC has changed a lot. It started as a group of 70 people trying to figure out by trial and error how to create a European computing community. Over time, we discovered what works, and what does not work. We have always tried to keep and improve what works, while discontinuing less impactful activities or replacing them by more impactful ones. When we started, there were only two networks (Artist and HiPEAC) and a couple of roadmaps in the community. Today there are tens of industry associations, joint undertakings, coordination and support actions organizing networking events and summer schools, publishing white papers and research agendas and the like.

A key plank of HiPEAC throughout the years has been the HiPEAC Vision, in which we navigate trends in computing and offer guidance for public investment in research. This roadmap would not exist without the steadfast dedication of Marc Duranton, HiPEAC Vision editor in chief since the 2009 edition, whose expert leadership has been fundamental to its development

Some people ask me how HiPEAC has managed to survive for so long. It is not different from biological systems: good genes, nutrients, adaptation and a portion of luck. For HiPEAC, the people represent the genes, and HiPEAC is lucky to have strong genes who want HiPEAC to survive. Funding from the European Union, topped up by industry sponsorship, are the nutrients for the community; without this support, HiPEAC would peter out. Adaptation is key in biology. HiPEAC is continuously reinventing itself to stay relevant both for the community and for European policy makers. And finally, there is the portion of luck that every living organism needs to grow old. There have been times when the future of HiPEAC was unclear, but so far we always found a way to continue. These four factors helped the HiPEAC community to reach adulthood.

I sometimes get the question: what about the future of HiPEAC? The world is changing faster than ever, and there are challenges ahead, but we are committed to staying relevant and to keep serving the community one year at a time.

Koen De Bosschere, HiPEAC coordinator

Twenty years of HiPEAC: Shaping European policy on computing innovation since 2004

Max Lemke, Head of Unit–Internet of Things, European Commission

In 2004, Mateo Valero of the Universitat Politècnica de Catalunya in Barcelona initiated HiPEAC. He gathered esteemed computing professors from across the European Union (EU) to establish a European network of excellence in computing. This collaboration marked the foundation of HiPEAC.

Over two decades, HiPEAC has become the leading platform for Europe's computing research and innovation community. It is respected alongside its counterparts in the United States (US) and Asia. Most prominent professors and experts from European universities, research transfer organizations (RTOs), supercomputing centres, and industry participate in HiPEAC as partners or active members. This network has formed a European think tank covering hardware and software, as well as all computing disciplines from embedded and cyber-physical systems, to cloud computing, to high-performance computing (HPC). HiPEAC's vision complements well the shorter-term industrial research and innovation agendas by the industrial associations of European partnerships such as the Chips and EuroHPC Joint Undertakings and their predecessors.

HiPEAC's vision documents, initially published biennially and now annually, have driven numerous European Union (EU) and national research and innovation projects. These initiatives have helped Europe maintain a leading position in global computing innovation despite industrial challenges.

"HiPEAC's vision documents have driven numerous European Union (EU) and national research and innovation projects"

Europe hosts 30% of the world's top ten supercomputers once again. It leads in foundational research on quantum computing technologies. Europe also holds a strong position in energy efficient, real-time embedded, and cyber-physical systems across sectors like aerospace, energy, health, and automotive. These achievements are all based to some extent on HiPEAC's vision.

In 2023, HiPEAC highlighted that Europe has entered a technological race in six areas driven by competition between the US and China, with Europe squeezed in the middle. These races have intensified since then. They led to divergence of technologies, such as smartphone platforms Android and Harmony. There is a competition for state aid for local industries, exemplified by the US Inflation Reduction Act. Export restrictions affect high-end computing systems, manufacturing technologies, and rare-earth metals. Import restrictions target technologies that may pose cybersecurity threats, like the ones addressed by the EU's 5G Cybersecurity Toolbox, or the proposed rule by the US Department of Commerce on banning certain Chinese and Russian automotive hardware and software.

In his report *The Future of European Competitiveness*, Mario Draghi, former European Central Bank (ECB) president, calls for Europe to close the innovation gap with the US and China. He recommends developing a plan for decarbonization and competitiveness, and increasing security while reducing dependencies.

Below are reflections on some of the races identified in the HiPEAC Vision 2023:

Race for the continuum of computing

Technologies and infrastructures have been converging over the past decades. This convergence has led to the repositioning or formation of new actors in various ecosystems. Examples include fixed and mobile communications, and the integration of embedded, cloud and high-performance computing into a computing continuum.

Henna Virkkunen, Executive Vice-President of the European Commission for Tech-Sovereignty, Security and Democracy during her hearing at the European Parliament. Her mission letter tasks her with, among other things, pushing forward the AI Factories initiative, developing an 'Apply AI Strategy' and an EU Cloud and AI Development Act, following up on the Chips Act (including with an EU quantum chips plan), and investing in frontier technologies, among them supercomputing, quantum computing, semiconductors, and the IoT.
Photo: Lukasz Kobus © European Union, 2024, CC BY 4.0

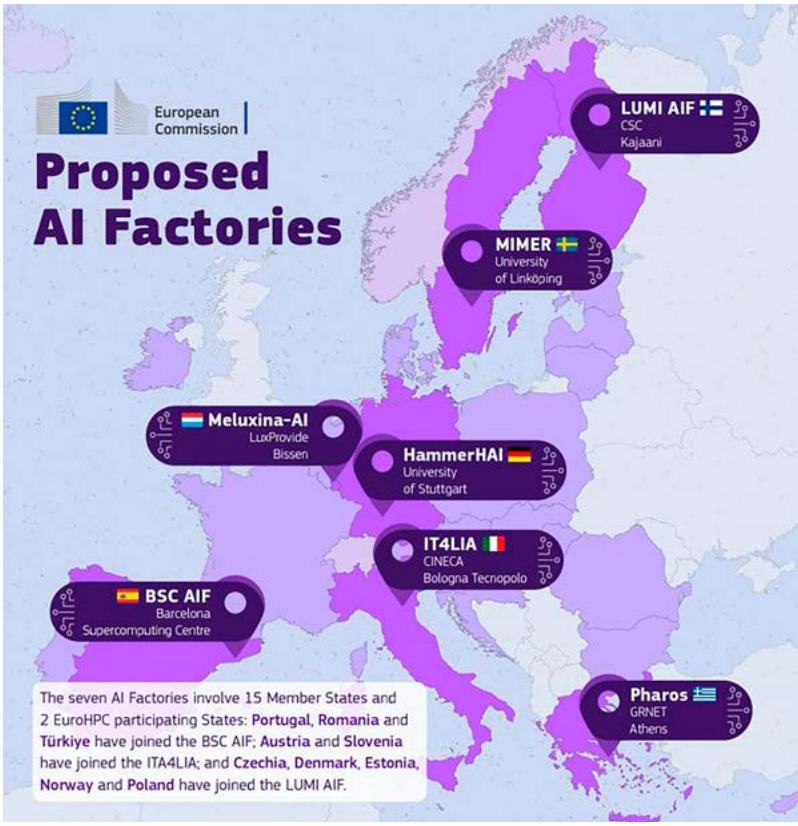
HiPEAC predicted this evolution of computing systems early on. In 2013, it highlighted the data deluge and the need for efficient data processing and increased energy awareness. In 2021, it forecasted the convergence of computing disciplines towards a computing continuum. The European Commission's 2024 white paper on digital infrastructure needs proposes creating the '3C Network' – Connected Collaborative Computing. This network aims to integrate emerging 5G / 6G networking with cloud / edge computing infrastructures. These 3C infrastructures will support a new dimension of decentralized intelligence. European industrial actors and system integrators are well positioned to integrate this emerging infrastructure with digital industrial applications and platforms in sectors such as energy, automotive, aerospace, agri-food, or health.

> ***"There is now an opportunity for European companies – if they act fast enough – to develop and train domain-specific LLMs using industrial data"***

Race for artificial intelligence (AI)

In 2010, HiPEAC's vision was: *'keep it simple for humans, and let the computer do the hard work'*. Programmers would specify functions, and computers would generate software for various platforms. Although HiPEAC did not predict tools like ChatGPT, such tools align with their vision. In 2015, HiPEAC recommended '*Develop[ing] approaches that enable domain specialists to express only what needs to be done, while tools and computers take care of how to transform this knowledge into an efficient computing system representation'*. While they did not explicitly use the word 'prompt', this concept is now reflected in large language models (LLMs).

Most LLMs are developed by US startups and large cloud providers. These models have exhausted public training data and are becoming commoditized. There is now an opportunity for European companies – if they act fast enough – to develop ▶



and train domain-specific LLMs using industrial data. In its January 2024 AI innovation package, Europe is investing in 'AI Factories' with specialized computing resources to support training new models. In addition, its GenAI4EU initiative supports the development of novel use cases and emerging applications in Europe's 14 industrial ecosystems, as well as the public sector. Application areas include robotics, health, biotech, manufacturing, mobility, climate and virtual worlds. The new commission, in office since December 2024, is developing an 'Apply AI Strategy' to promote and support wide AI adoption in key verticals. These efforts strongly support European sectoral actors and integrators in capitalizing on AI.

Race for new hardware technologies

In 2015, HiPEAC recommended developing *'formalisms, methodologies, and tools' for 'adequate precision computing'*, or more generally to deal with *'desired level of Quality of Service'*. This includes addressing power, security, and time constraints using the design-by-contract paradigm. These recommendations are relevant for future autonomous vehicles, which require local availability, energy efficiency, real-time capabilities, and affordability of their hardware.

Current accelerator technologies from large US companies only partially meet these requirements. New approaches in precision computing and advanced memory technologies are needed to overcome the data feeding bottleneck for accelerators. Vice President Virkkunen's mission letter states that the new commission will *'intensify efforts and investments concerning the next wave of frontier technologies, in particular supercomputing, semiconductors, IoT, genomics, quantum computing, space tech and beyond'*. This includes developing a *'long-term EU Quantum Chips plan, as proposed by the Draghi report'*.

Race for sovereignty and reduction of critical dependencies

In 2013, HiPEAC noted that: *'Developing hardware is considered strategic for some countries (e.g. China). Therefore, some governments are directly or indirectly supporting development of microprocessors and other advanced digital circuits to gain strategic independence from external providers.'* In 2017 it stated that: *'A threat to the European computing industry is the rapid development of the computing industry in China, Russia, and Japan. Many countries understand that computing is a key-enabling technology of strategic importance, and invest in their own research, products, and companies. If Europe fails to do the same, it might eventually become dependent on technology which is designed, developed, produced, and controlled outside Europe.'*

Geopolitical tensions and divergent digital platforms from US and Chinese companies pose challenges for European actors. Europe aims to regain strategic autonomy by bringing microelectronics design and fabrication back to the continent. The European Chips Act, which came into force in 2023, marks the beginning of this effort to regain what Europe has lost in recent decades.

Race for sustainability

For over a decade, the HiPEAC Vision has prioritized sustainability in computing systems. Sustainability is a complex, multidimensional issue affecting industries such as mobility, transport, energy, agri-food, and manufacturing. The rise of generative AI and LLMs has resulted in increasing datacentre energy consumption and created irregular electricity loads. Scaling up data centres and reducing their carbon footprint are essential for sustainability. *'Opportunistic computing'*, which manages workloads based on renewable energy availability, is becoming relevant again. Sustainability also involves system-level optimization of cost, performance, loads, energy, and resource management. Reducing the use of rare-earth materials and water consumption and minimizing the carbon footprint of devices throughout their lifecycle are critical. These goals align with the Circular Economy Action Plan (CEAP), the Ecodesign for Sustainable Products Regulation (ESPR), and the Digital Product Passport.

After 20 years, with its relentless efforts to reinforce a vibrant and visible ecosystem, HiPEAC remains integral to the European computing research and innovation community. Its annual vision documents inspire researchers in industry and academia, and guide projects. HiPEAC's annual conferences attract over 500 participants for networking. Its summer schools develop new talent, and its paper and technology transfer awards recognize excellence. With an average annual investment of €1 million from the European Union, HiPEAC demonstrates excellent value for money.

"After 20 years, HiPEAC remains integral to the European computing research and innovation community"

With a rapidly expanding high-performance computing (HPC) ecosystem, Europe is building the capacity for homegrown chips and designed-in-Europe machines. According to Mateo Valero, director of Barcelona Supercomputing Center – Centro Nacional de Supercomputación (BSC-CNS), founder and coordinator of the first HiPEAC project, the HiPEAC network has played a fundamental part in creating this ecosystem. In this interview, Mateo explains how returning to the origins of HiPEAC demonstrates what it takes to create a united Europe capable of making common technology goals a reality.

'In Europe, when we work together, we can achieve anything'

Following a series of unsuccessful projects to build supercomputers in the 1980s, Mateo notes, enthusiasm for European supercomputer projects cooled, with none being funded by the European Commission for some years. Yet it seemed clear that these projects had failed in part because there wasn't a solid base in Europe to support them – a network bringing together the expertise in the relevant fields.

Although it lacked a supercomputing industry, Europe was strong in embedded systems. 'To those of us who started HiPEAC it was clear that chips for embedded systems – which, at that time, were somewhat basic – would converge with the powerful chips needed for supercomputers in less than 20 years,' says Mateo. Hence focusing on high-performance chips for embedded computing could help future-proof the European embedded computing industry, while at the same time building capacity to develop high-performance systems designed in Europe.

To build this capacity, Europe needed a network of excellence in software-hardware co-design, Mateo explained to policy representatives in Brussels. The core of this network, he added, should be the bridge between software and hardware – that is, computer architecture and compilation.

Taking advantage of the 1998 edition of the International Symposium on Computer Architecture (ISCA) in Barcelona, Mateo approached several colleagues 'who immediately bought into the idea' and they started preparing a proposal. Initial academic patrons of HiPEAC, Mateo recalls, were Olivier Temam (France), Stamatis Vassiliadis, Koen Bertels and Georgi

HiPEAC1 kickoff meeting in Juan-les-Pins, 30 September 2004. From left to right: Theo Ungerer, Krisztián Flautner, Michael O'Boyle, Mercè Griera i Fisa, Stamatis Vassiliadis, Bilha Mendelson, Manolis Katevenis, Koen De Bosschere, Pilar Armas, Erik Norden, Mateo Valero, Marco Cornero, Antonio Prete, Jakob Engblom, Per Stenström, Olivier Temam, Josep Llosa

Gaydadjiev (The Netherlands), Manolis Katevenis and Angelos Bilas (Greece), Nacho Navarro, Josep Llosa and Mateo Valero (Spain), Antonio Prete, Pierfrancesco Foglia and Roberto Giorgi (Italy), Theo Ungerer (Germany), Koen De Bosschere (Belgium), Per Stenström (Sweden) and Michael O'Boyle (UK). Industry representatives included Krisztián Flautner (Arm – UK), Erik Norden (Infineon – Germany) and Marco Cornero (STMicro – Italy), Jakob Engblom (Virtutech– Sweden) and Bilha Mendelson (IBM).

HiPEAC1 kicked off on 1 September 2004, partly thanks to the 'unconditional support' of project officer Mercè Griera, Mateo notes. The first project lasted four years and was coordinated by Mateo, with project management by Pilar Armas. Since 2008, HiPEAC has been coordinated by Koen De Bosschere of Ghent University in Belgium. 'Koen has put enormous effort into coordinating subsequent editions of the HiPEAC project. Building on the original idea, he has been very successful in making the network stronger and larger in each sprint, and I am very proud of what it has become,' says Mateo.

A network for Europe: Collaboration, training and dissemination

The main aim of the initial HiPEAC projects was to promote collaboration between European researchers, many of whom, at that time, only crossed paths at American conferences. 'We had funding to do joint research, which allowed us to get to know each other and strengthen our research. From these initial projects, we progressed to preparing bids for joint European projects funded by the framework programmes, with great success,' Mateo recalls.

Education and training also formed fundamental planks of the HiPEAC network from the outset, contributing to excellence in European computing research, with ACACES, the long-running HiPEAC summer school, as the main training event. The HiPEAC conference was established as a key dissemination venue for members' research; it now regularly brings together over 500 researchers in the field from all over the world in different locations around Europe. To strengthen collaborations, HiPEAC ran a series of Computing Systems Weeks for many years; currently, HiPEAC7 is building value networks within the European computing ecosystem, such as the Computing Continuum group of high-level European organizations to align priorities in European Union (EU) research. 'These activities are not only fundamental to making us better researchers; they create a united Europe,' says Mateo.

Indeed, the pan-European nature of the HiPEAC network, as opposed to a group of countries with competing interests, laid the groundwork for strategic high-performance computing projects. 'The Mont-Blanc project, which sought to create supercomputers out of Arm chips, was born in HiPEAC. The European Processor Initiative and its successors have their roots in HiPEAC. The network created a proactive, collaborative research environment that was crucial to building major EU research ventures,' Mateo says.

The evolution of architecture

Since HiPEAC's founding in 2004, the processor landscape has changed significantly; with physical limits challenging Moore's Law, which observed that the number of transistors in an integrated circuit doubles about every two years, engineers have turned to alternative means of increasing performance. Mateo identifies an evolution from multicore chips, to accelerators, to chiplets.

'The emergence of multicores required parallel processing within chips and augmented the importance of on-chip communication, as well as energy efficiency, and had a knock-on effect on compilers and runtime. From the earliest editions of its vision document, HiPEAC foresaw that multicores would be needed for embedded systems,' he adds.

Later, with the waning of Moore's Law, accelerators began gaining prominence. Of these, graphics processing units (GPUs) are currently the most successful in HPC, notes Mateo. Other examples of accelerators for high-performance applications include the Intel Xeon Phi, based on a single architecture devoted to short vectors, and the IBM/Sony/Toshiba Cell processor, based on heterogeneous processing on processor-specific hardware, says Mateo. 'An additional line that we are pursuing at BSC is the long-vector approach, the central subject of our developments around the RISC-V architecture and its vector extension.' Indeed, due to the efficiency gains of vector processors as compared to scalar processors such as GPUs, Mateo sees potential for made-in-Europe chips to compete with giants such as NVIDIA. *(See our interview with Filippo Mantovani on pp.21-23 for further details)*.

Chiplets – mini chips specialized for particular purposes – are the logical evolution of this trend towards heterogeneity, says Mateo, offering the potential for greater density of computing power through 3D integrations.

> ***"HiPEAC created a proactive, collaborative research environment that was crucial to building major EU research ventures"***

Located at Barcelona Supercomputing Center, MareNostrum 5 is one of the EuroHPC top-tier machines

To further accelerate complex algorithms, the next step is quantum computing, although, in line with fellow HiPEAC members such as Yale Patt (and the HiPEAC Vision), Mateo's view is that quantum will be limited to accelerating particular regions of code that are too complex to be solved with traditional computing architectures.

European HPC: From acquiring to producing

The importance of supercomputing as a tool for European scientific research was recognized by the European Commission early on, and PRACE, the Partnership for Advanced Computing in Europe, was launched as a means to provide access to top-tier machines. 'Evolving in parallel with HiPEAC, PRACE allows scientists to access the most suitable machine for their needs,' notes Mateo.

In 2017, the EuroHPC JU was launched with twin objectives: to support a pan-European supercomputing infrastructure of top-tier machines while, at the same time, generating a competitive innovation ecosystem in supercomputing technologies, applications and skills. 'PRACE began with a plan to install between three and five tier-0 supercomputers. Currently, in Europe, we have eight tier-0 machines. Not only that, but the variety of design approaches used for these supercomputers is testament to the heterogeneity that has become dominant over the years since HiPEAC was founded,' says Mateo.

Prior to the launch of EuroHPC, Mateo and other HiPEAC founders had been calling for an 'Airbus of HPC' that could deliver a European exascale machine, developed with European processors. In 2018, the European Processor Initiative was launched with the aim of developing a designed-in-Europe Arm-based central processing unit (CPU) and RISC-V-based accelerators. A major international follow-up project, DARE, is due to start in 2025 to continue this work, and MareNostrum 6, the next version of BSC's supercomputer, is to be a demonstration of the technologies developed in this effort. 'HiPEAC and PRACE have been growing along with the software and computing infrastructure in Europe. Now we can focus on the development of chip technology, and the commercialization of research concepts developed by HiPEAC members,' says Mateo. 'The next 10 years will be a period of building a self-sufficient European chip manufacturing and processor-design ecosystem.'

In pursuit of this goal, EU Member States will have to join forces to achieve what they would be unable to do individually. 'The Draghi report on European competitiveness highlighted that most individual member states cannot achieve the scale to deliver world-leading research and technological infrastructures, which is why coordinated initiatives such as EuroHPC and institutions like CERN are essential,' says Mateo.

When HPC met AI

The development of supercomputing technologies in the future, though, will be accompanied by the dominant computing trend of recent years: artificial intelligence (AI). 'Over the last few years, artificial intelligence has been a natural companion of HPC,' says Mateo. 'The data-processing capabilities and compute power of HPC has enabled AI to grow, and AI is now returning the favour in spades: it is changing the way we do research with supercomputers.' In terms of industrial competitiveness, Mateo highlights AI Factories, a new subprogramme within EuroHPC which aims to facilitate access to AI provided by HPC facilities to improve the productivity of small and medium enterprises (SMEs), the backbone of the European economy. 'In the future, EuroHPC supercomputing facilities will also be made available for AI startups, which will be crucial in propelling Europe to be a leader in HPC-AI convergence,' Mateo adds.

Just as Europe needed to fund HPC development, so it must support the development of AI, according to Mateo. 'Leaving the development of AI to the United States or China would be a major error,' he warns. 'We cannot limit ourselves to regulation; to do so would be to accept a role as the referee of a match that we're not playing ourselves. We need to develop chips for AI and, in parallel, develop algorithms to exploit them in an energy-efficient way, through co-design – something which is in HiPEAC's DNA,' he adds. 'HiPEAC offers a model for European politicians and companies of how to unite behind a common objective. In Europe, when we work together, we can achieve anything.'

The H.T. Adams '35 Professor of Computer Science at Princeton University, Margaret Martonosi has been in computer architecture research for over 30 years. HiPEAC spoke to Margaret to find out about why computer architecture is such an important research area, the challenges facing researchers on both sides of the Atlantic, and what the priorities for future research should be.

'We have lost key foundations of durable long-term abstraction that computer architecture was originally based on'

What led you to specialize in computer architecture research?

When I first learned about computer programming as a teenager, I was captivated. It was so much fun to bring programs into existence – it was fun to execute code that I had written, and watch it solve real problems. From that starting point, I realized I was really curious to unpack the various system layers and learn what was really happening down in hardware, when those programs executed. That led me into computer architecture. My incredible experience as a grad student in the Stanford DASH and FLASH multiprocessor research groups really cemented my love for the field.

It has been over 30 years from that starting point, and I still find computer architecture to be exciting and amazingly high leverage – its position in the systems stack means that it has a strong responsibility to respond both to the application trends 'from above' and the technology trends 'from below'. That makes it a very exciting place to work!

How do the respective research landscapes in the United States and Europe compare?

Computing research is changing fast around the world. In particular, the role of academic researchers vs. our industry researcher counterparts is shifting. Given the rise of large-scale industry investments, we all face the challenge of ensuring that academic researchers can have the resources (people, data, hardware, software, etc.) required to do truly cutting-edge long-term research that can complement industry efforts and inspire future industry directions. This is true on both sides of the Atlantic. In Europe, HiPEAC has been an important connector in keeping European Union (EU) architecture research on the cutting edge through these shifts.

What areas should the HiPEAC community focus on over the next five to 10 years?

I see a huge opportunity in working together to advance the trustworthiness of computer systems we are building. I define 'trustworthy' broadly, to include reliability, verifiability, and security. Traditionally, our instruction set architecture (ISA) interfaces have been 'operational'. By that I mean that they have often been designed more for ease of compiler mapping and optimization, rather than with formal specification logic for verifiability. As our processor implementations employ more heterogeneous parallelism, it becomes harder and harder to check that software runs on them as expected. We have lost some of the key foundations of durable long-term abstraction that computer architecture was originally based on. This is particularly disturbing in light of the ever-increasing reliance of computer systems in many safety-critical applications.

To move forward, we need new ways to regain long-term abstractions that work across our full chips, and that support former verification for correctness and security. Such work can, in turn, offer a better foundation for building reliable software and trustworthy artificial intelligence (AI) models and software in the decades to come.

Credited with the invention of the 'network-on-chip' design automation paradigm, HiPEAC 2025 keynote speaker Giovanni De Micheli's stellar career in electronic design automation (EDA) has taken him from his native Milan, to California, and now to EPFL in Lausanne. HiPEAC spoke to Giovanni to learn more about breakthrough hardware innovations, how to build thriving ecosystems, and where Europe should focus its computing research.

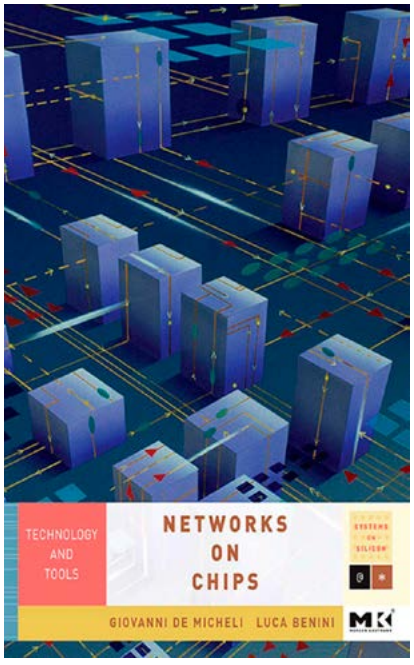
'The strongest European asset is the profound culture that defines and divides us'

How did you end up specializing in EDA?

I grew up in Milan, Italy where I enrolled at the Politecnico in a nuclear engineering programme with an emphasis on data acquisition and control. While there, I became interested into how circuits could be designed in a rational manner. With the help of a Fulbright scholarship, I arrived at UC Berkeley, a fantastic crossroads for research on circuits, systems and their design with computers.

At that time, EDA was in its infancy: it mainly focused on analysing circuits by means of simulation, and was facing the challenge of increasingly larger circuit sizes. It was also the beginning of computer-aided synthesis, with the first tools to create physical layouts of programmable logic arrays (PLAs) from Boolean specifications, thus involving logic optimization and structured layout generation. It was exciting and enriching to work with my adviser, Alberto Sangiovanni-Vincentelli, as well as with Robert Brayton and Richard Newton, and I am grateful to them for the tremendous opportunity of learning this field.

After my graduation, my work at IBM, Stanford and EPFL revolved around circuits and systems design in many flavours. With my graduate students, we researched algorithms and tools for logic and high-level synthesis. We pioneered synthesis from C (and later SystemC) models and provided the community with open-source tools. With Luca Benini, we created the first EDA flow for network-on-chip (NoC) design in 2000; this became the backbone of the INOCS start-up, whose technology was acquired by Arteris, the dominant provider of EDA tools for NoC design. Today, all major chip designs incorporate NoCs of various kinds and it is very rewarding to me to know that most electronic chips embody a network designed by leveraging models, methods, tools based on engineered and improved versions of our original prototype at Stanford.



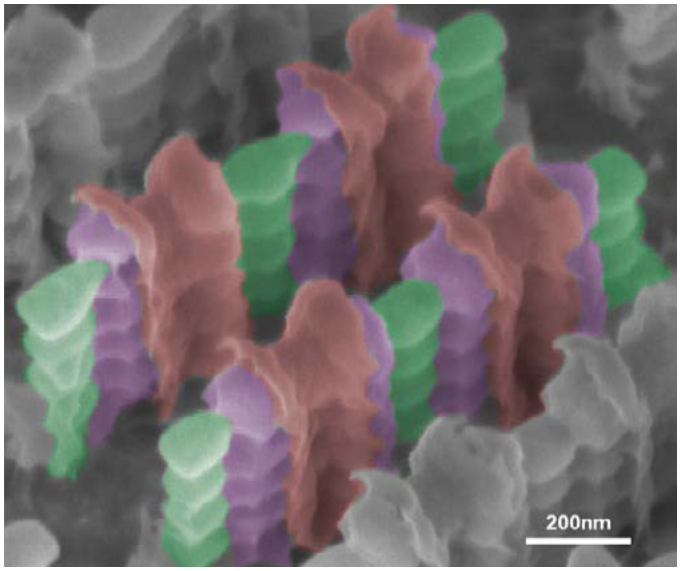
Networks on Chips, by Giovanni De Micheli and Luca Benini

I moved back to Europe, namely to EPFL in Lausanne, Switzerland, after 25 happy and fruitful years in the United States, driven by the desire to taste different aspects of circuit research. I shared the joys and frustrations of the fabrication facility with my colleague Yusuf Leblebici. We embarked on the design of silicon nanowire circuits, the progenitors of today's nanosheets used by advanced commercial fabs. Along a parallel path, I became interested in sensing devices, in particular biosensing applications. With my colleague Sandro Carrara, we developed

transistors that can be turned on by the presence of specific analytes, e.g. glucose, lactose, antigens, etc. This was also an entry point into electronic systems for precision medicine, and thus understanding the wide, untapped opportunities to better society by means of smart biocircuits and systems. As a corollary, we discovered a huge set of applications with heterogeneous materials and devices that need EDA support for design and integration.

How has the EDA landscape evolved since you started your career? How do you see it changing in the future?
In the early days, EDA was known as 'computer-aided design' (CAD), as the collective vision was to have tools around a human designer who is the decision maker at all levels of design. Then synthesis technology came along, with specific flows for the automated design of circuits from high-level specifications, requiring limited human intervention into the design process. Eventually, artificial intelligence (AI) techniques will drive circuit design, leveraging the learning experience from previous chip designs. In essence, the use of AI-driven tools will move up the required level of abstraction by a human designer to specify an integrated system.

On the other hand, I also think that the devil is in the details, and so chips that will outperform their predecessors in power / performance / area (PPA) will require innovation, that is breaking habits and traditions coming from past design experiences. This means that human creativity will be necessary, even though intelligent EDA tools will be needed to cope with the very large number of choices in design. The massive use of AI tools and big-data repositories will also come at the cost of much more computation and energy consumption. Eventually end-users will have to reconcile the desire for intelligent products with the necessity of protecting the planet.

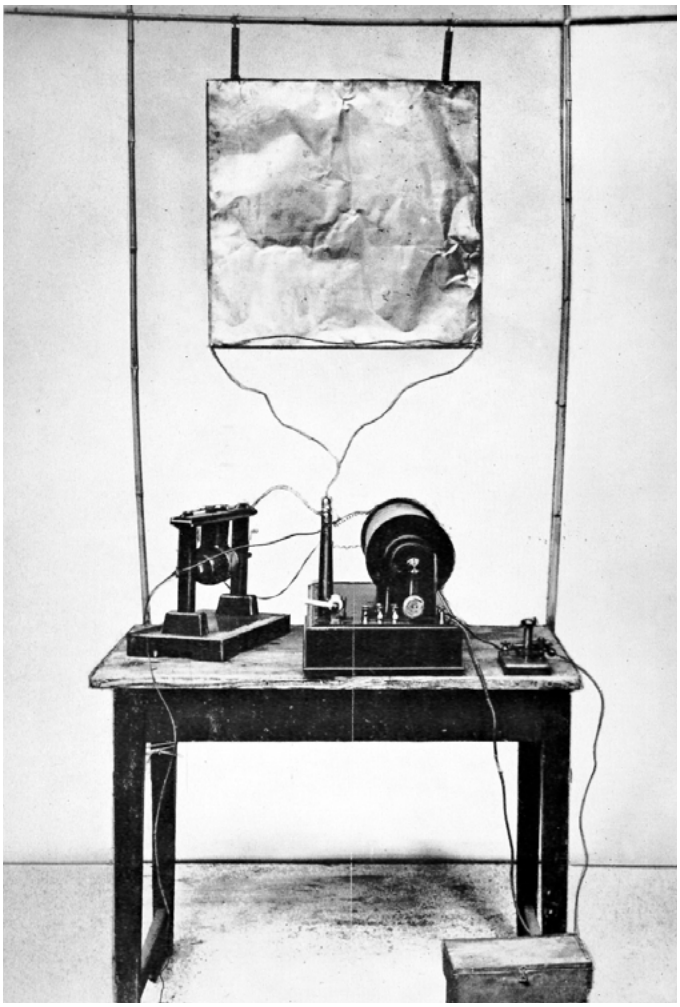


The Polarity-controlled NanoWire Transistor

What, for you, are the strengths and weaknesses of European technology innovation?
New breakthroughs, such as intelligent surfaces and human-computer interaction, stem from the combination of different aspects of science and technology. Europe has multicultural roots and has always been the home of scientists with a renaissance mind and a strong cultural background who are capable of looking far ahead, beyond the limits of short-term returns on investment. I think Europe has the capability of innovating and promoting technologies that can leapfrog the present, just because we educate students to concern themselves with the intrinsic values of engineering products per se, and not just the economic opportunities. In a similar vein, the European Research Council (ERC) grant programme is highly successful because it targets high-risk high-reward research proposals, which are key for the advancement of science and technology.

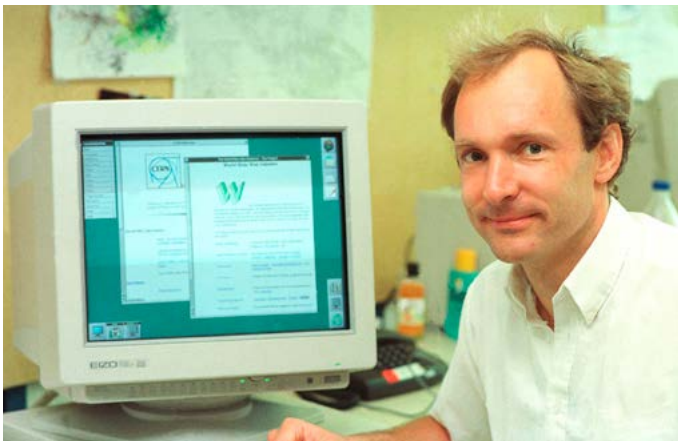
There has been a sequence of inventions and products, from the World Wide Web (developed at CERN) to wireless telephony (the radio was born in Bologna), that originated in Europe and are part of today's fast technological and societal evolution. Unfortunately, Europe has lost a market where it was dominant (i.e. wireless telephony) due to the lack of an appealing interface and a failure to provide related services to clients. This setback is also serious, as we lost the chance of using our inherent multicultural nature as a way to promote communication across different cultural backgrounds. Looking to the future, the strongest European asset is the profound culture that defines and divides us, and that distinguishes us from a melting pot. I believe that leveraging this factor into products that cater to a multicultural society can be a major enabler for our industry to grow and prosper.

What needs to be done to build out hardware ecosystems? What areas do you think should be prioritized in Europe?
An ecosystem involves multiple partners and objectives. Large companies should promote the common good not only by providing financial support to targeted research but also by enabling technical events and participating in person. Startups are often the best instrument to promote technology, eventually benefitting the large corporations. Academic research in Europe has pinnacles of excellence and trains graduates with a solid background. Nevertheless, the hardware ecosystem is still weak in Europe and depends on overseas technology, despite excellence in many domains, such as the telecom and the transportation sectors.



A recreation of Marconi's first radio transmitter. Source: Wikipedia

The visibility of European technology products is limited among university students. We rightly focus our educational principles on teaching the fundamentals, but we could infuse more enthusiasm into youngsters by showing the impactful outcomes of the research. The hardware ecosystem is much stronger in East Asia, where an engineering culture is embraced. We Europeans very often focus on limiting factors such as starting salaries and we don't sufficiently value the pride of achieving top technical solutions. Overall, I think that European leaders in industry and academia should spend much more time and effort in explaining their objectives, in collaborating on educational programmes that value science and technology, and in making sure that quality is valued above all other metrics.



CERN was the cradle of the World Wide Web invented by scientist Tim Berners-Lee as a tool for the high-energy physics community. © 1994 CERN

What do you think the HiPEAC community should focus on over the next five to 10 years?

HiPEAC is unique in its search for bridging the academic and industrial worlds and in its role as a radar for European computing technology. As the influence of artificial intelligence and machine learning is going to be increasingly pervasive in the years to come, it is very important to pursue a path that integrates diversity and inclusion of cultures into AI models. We do not want intelligence to be led by models representing only American and Asian cultures, which, though highly respectable, are not ours.

For this we need to focus on developing our platforms, from edge terminals to the cloud, that put European cultural and scientific values at the forefront. We, as Europeans, need to own both the hardware and software platforms, to avoid the unfortunate fate of becoming a provincial colony. This message has to be passed – and repeated – to our political leaders, who have to be far-sighted by providing long-term growth to our economy, and not just seek to reap short-term opportunities. The scientific community has the responsibility to be vigilant and vocal in passing the message along.

"We do not want intelligence to be led by models representing only American and Asian cultures, which, though highly respectable, are not ours"

A core member of the HiPEAC community for nearly two decades, Emre Ozer has dedicated his career to processor development. HiPEAC caught up with Emre to find out what sparked his interest in computer architecture, the potential of flexible electronics, and why the research community needs to look beyond a silicon-bound perspective.

'We need to filter out silicon-based technology biases to come up with new computing architectures'

Tell us about your career in computing so far.

At high school, I read the *Empire* and *Robot* series of novels by Isaac Asimov. When I was about to finish high school, I was torn between studying astrophysics and computer science, but eventually went for computer science and engineering at Yildiz Technical University in Istanbul. At that time (the late 80s), personal computers (PCs) were not in widespread use, at least in Türkiye, so I learned how to write computer programs in mainframes. The first programming language I learned was Pascal, and then some PL/1 (which I didn't like), and then finally C.

In my final year at university, I took a class in computer architecture. I loved the topic of central processing unit (CPU) micro-architectures so much that I knew this was the area in which I needed to specialize. I did my MSc on interrupt handling in superscalar processors at the same university, and then went to North Carolina State University to do my PhD on the topic of simultaneous multi-threading in very-long instruction word (VLIW) processors.

Following my PhD, I took a job as a computer architect at a digital-signal processing (DSP) company called StarCore, which was a joint venture between Motorola and Agere Systems. This was followed by a two-year research fellow position working on hardware compilation for field-programmable gate arrays (FPGAs) at Trinity College Dublin, Ireland. In 2005, I joined the research and development (R+D) department at Arm as a senior R+D engineer working in the areas of CPU performance modelling, fault-tolerant CPUs, embedded machine learning, application-specific hardware design, and, most recently, developing processing engines in flexible electronics. I represented Arm in the HiPEAC network, of which Arm was a founding member, from 2007 to 2020. In mid-2022, I left Arm to join Pragmatic to lead the flexible microprocessor development.

Why is the development of flexible electronics important?

Printed / flexible electronics offers an alternative to conventional electronics by developing electronics on low-cost flexible substrates (e.g. polymer, paper, biodegradable materials). It uses low-cost materials and manufacturing, and has physical flexibility – in some cases stretchability – in form factor. Other unique properties of printed / flexible electronics are conformability, comfort and biocompatibility. Researchers and the electronics industry have been developing printed sensors, printed batteries, flexible displays, flexible e-paper displays, printed energy harvesters such as solar cells, flexible near-field communication (NFC) chips and printed antennas.

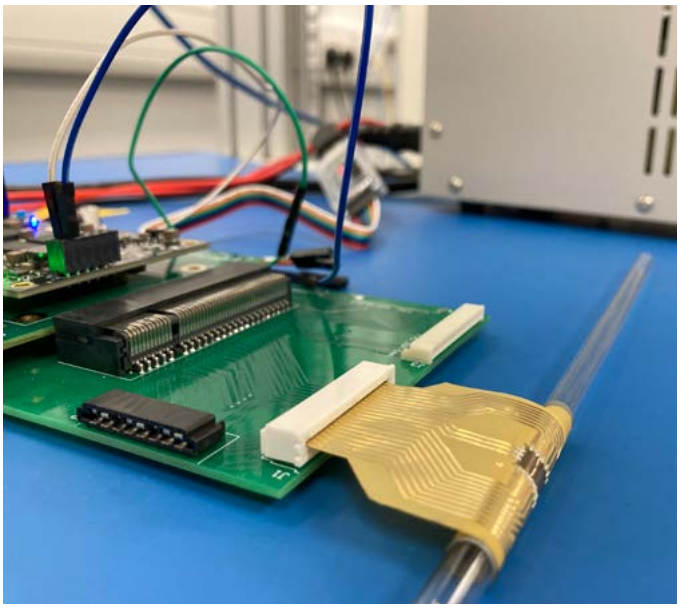
Printed / flexible electronics will not replace silicon-based electronics in conventional market segments. Rather, it is more suitable for applications at what I call the 'extreme edge', which can be defined as the class of applications where the electronics deployed in an end device cannot be based on conventional silicon technology because of one or more of these constraints: form factor, cost, conformability, comfort, biocompatibility, etc. Good examples of extreme-edge appli-

cations are logistics, fast-moving consumer goods, healthcare wearables, implantable / ingestible devices, textiles, agricultural and environmental monitoring.

Pragmatic develops an ultra-low-cost, flexible integrated circuit (FlexIC) technology which is based on indium-gallium-zinc-oxide (IGZO) thin-film transistors (TFTs) fabricated on a flexible substrate. We manufacture FlexICs on industry-standard 200 / 300mm flexible wafers. One of the unique features of our fab technology is the short production cycle; in the context of chip design, tape-out to fab-out time is currently around four weeks, but it can be streamlined to as little as 48 hours. This is significantly faster than a silicon fab, which can take between six and nine months for the tape-out to fab-out process.

What about environmental concerns?
I would say that the fabrication of FlexICs is more sustainable than fabricating silicon-based ICs. According to the lifecycle analysis of FlexICs, our FlexIC fabs use much less energy and water, use fewer harmful chemicals, and have a carbon footprint on the orders of magnitude smaller than a conventional silicon fab.

This is also true for manufacturing other flexible electronic components. Manufacturing components such as printed sensors using screen printing is more efficient in terms of material usage because it is an additive manufacturing process. As for substrates, the polymer substrates on which the electronics sit can be recycled as part of recycled products.



Pragmatic's flexible microprocessor

For example, the substrate of the flexible electronic component embedded into plastic food packaging can be recycled as part of the packaging's standard recycling process. However, there is a lot of research on developing alternative substrates such as paper, bioplastics, silk etc., some of which can be biodegradable. It is just a matter of time before these alternative substrates are in more widespread use.

What areas should computing research focus on over the next five to ten years?
Research in computing is mainly driven by progress in silicon technology, which is itself driven by Moore's Law. However, there are alternative technologies with which compute systems can be built and that can enable new functionalities, and flexible electronics is one of them. The computing research community needs to understand both the capabilities of the alternative technology and the functionalities they enable, instead of bringing in the biases learned from the years of training in silicon-based technology.

For example, the development of a compute system made exclusively of flexible electronic components may not seem any different from the development of a compute system made of silicon components. However, flexible electronics allows low-cost and fast customization, which means that sensors, battery, energy harvesters and the processing engine can be customized to the requirements of the domain in which the compute system will be used, as well as being fabricated quickly and at low cost. This would not be feasible in silicon-based electronics.

How does this affect research in computing systems? Let's take a processor-centric approach, and ask whether we really need a general-purpose microprocessor in the realm of flexible electronics. If a processing engine can be customized to the target domain and fabricated fast and at low cost, how should we architect the processing engine?

This is valid for other alternative technologies such as quantum, optical and molecular electronics: each technology needs to be understood in terms of its capabilities and functionalities. These technologies offer a unique opportunity for the computing research community to innovate. The challenge is for researchers to transfer learning from silicon-based computing to the new realm while filtering out the biases accumulated over the years, in order to come up with new computing architectures.

Twenty(ish) years in computing

1999 > 'Coherent control of macroscopic quantum states in a single-Cooper-pair box' (Yasunobu Nakamura, Yuri A. Pashkin and Jaw-Shen Tsai) published in *Nature*. The results demonstrate that a superconducting circuit can be used as a qubit.

2002 > 'Networks on Chips: A New SoC Paradigm' (Luca Benini and Giovanni De Micheli), IEEE Computer

Giovanni De Micheli, EPFL: 'The network on chip (NoC) breakthrough was related to understanding the synergy between circuit design techniques and network technology. An NoC is very different from the internet, as protocols have to be much simpler in chips to allow for low communication latency. Thus, while NoC has its roots in bridging across different areas and cultures, it is a success because it leverages specific features of circuits and communication architectures.'

2006 > 'A fast learning algorithm for deep belief nets' (Geoffrey E. Hinton, Simon Osindero and Yee-Whye Teh) published in *Neural Computation*

2005-2007 > End of Dennard scaling

Margaret Martonosi, Princeton University: 'An important landmark of the past 20 years has been the end of Dennard scaling and the slowing of Moore's Law. Architects have been at the forefront of responding to these trends by adjusting computer hardware to employ more parallelism and more specialized (and therefore energy efficient) processing units.'

2008 > International Mobile Telecommunications-Advanced (IMT-Advanced Standard) requirements issued for 4G mobile-phone and internet services

2010 > Development of RISC-V (UC Berkeley)

Emre Ozer, Pragmatic: 'An open instruction set architecture (ISA) to develop 32/64/128-bit microprocessors, RISC-V democratized microprocessor development by allowing anyone to design a modern CPU without licensing hurdles. Because it gives so much flexibility to researchers, there has been a great interest in RISC-V ISA in the computer architecture, microarchitecture and compilation research community. Although initially proposed as a research instrument, RISC-V eventually into the embedded and automotive segments in the commercial space, mainly due to the simplicity of the ISA to translate the specification to a CPU microarchitecture implementation and, most importantly, for its fast-growing software ecosystem.'

1999

1999 > NVDIA introduces Geforce 256, the first widely available GPU

2000

2001

2001 > IBM launches POWER4, the first commercially available multiprocessor chip

2002

2003

2003 > First public presentation of Arm multicore strategy

2004

2005

2005 > Launch of Google Maps

2006

2006 > Amazon launches Amazon Elastic Compute Cloud (EC2), which allows users to rent virtual computers on which to run their own computer applications

2007

2007 > First iPhone launched (Apple)

Emre Ozer, Pragmatic: 'The launch of the smartphone led to the boom of mobile computing, where energy efficiency became the first-class citizen. Smartphones had a huge impact, not only on the hardware development – from an application processor to a modem processor, from built-in sensors to display technology and more recently to embedded machine-learning hardware – but also on software, from app stores to social media. With a simple user interface, you do not need to be a technology savvy to use social media on a smartphone, enabling anyone to have a mobile computer.'

2008

2009

2010

2008 > First petascale supercomputer, the Roadrunner, built by IBM, enters operation at Los Alamos National Laboratory

2011 > 'Dark Silicon and the End of Multicore Scaling' (Hadi Esmaeilzadeh et al) published at ISCA

2011 > First EU Mont-Blanc project to develop Arm-based supercomputers begins

Alex Ramírez, Google: 'We designed and deployed the first Arm high-performance computing (HPC) clusters when there was no HPC software stack for Arm. We gathered a super-hot consortium and delivered a successful programme, with massive impact on the media and the community, all with very few resources. Today, Arm supercomputers are in the Top500 and are deployed in the data centres of all the cloud hyperscalers.' (Interview for *HiPEACinfo* 69, June 2023)

2012 > Convolutional neural network (CNN) AlexNet wins ImageNet computer-vision challenge

Emre Ozer, Pragmatic: 'They say that machine learning (ML) was in the Ice Age until the success of training deep convolutional neural networks using GPUs was demonstrated in 2012. The algorithmic evolution in deep learning has been continuing since then, from convolutional neural networks to transformers and finally to large language models. Aside from the algorithmic evolution, an efficient implementation of the inference portion of the ML in hardware and its integration into compute systems including edge devices has breathed fresh air into computer architecture over the last decade.'

2012 > 'A Defect-Tolerant Accelerator for Emerging High-Performance Applications' (Olivier Temam) – ISCA 2012

2012 > 'Neural acceleration for general-purpose approximate programs' (Hadi Esmaelizadeh et al) – MICRO 2012

2017 > 'Attention Is All You Need' – (Ashish Vaswani et al, Google)

2018 > Launch of the European Processor Initiative

2018 > 'Improving Language Understanding by Generative Pre-Training' – (Alec Radford et al, OpenAI)

2019 > 'Quantum supremacy using a programmable superconducting processor' (Frank Arute et al, Google) published in *Nature*

2011

2012

2013

2014

2015

2016

2017

2018

2019

2020

2021

2022

2011 > IBM Watson computer wins Jeopardy! TV quiz show

2011 > D-Wave announces first commercially availalble quantum computer

2016 > Google publicly announces TPU at Google I/O

José Cano, University of Glasgow: 'The development of application-specific integrated circuits (ASICs) and other specialized hardware accelerators for ML such as the TPU (tensor processing unit), NPU (neural processing unit), LPU (language processing unit), etc., have allowed us to drastically improve training and inference procedures for both cloud and edge environments.'

2018 > Rollout of 5G begins

2020 > Fugaku supercomputer – the first to be based on the Arm architecture – unveiled by RIKEN and named the fastest supercomputer in the world by the TOP500

2022 > First exascale supercomputer, Frontier, begins operation at Oak Ridge National Laboratory (ORNL)

2022 > ChatGPT launched to the public

Margaret Martonosi, Princeton University: 'Clearly, the huge inflection points in AI's capabilities and usage have been important technology landmarks. Architects are responding by exploring AI accelerator designs and by developing hardware-software interfaces for seamlessly incorporating accelerators into the richly heterogeneous and parallel architectures that dominate the field today.'

‘It was a collaborative European effort based on the idea that multicore could do more than mobile handsets’

Advanced RISC Machines Ltd (ARM) was founded in 1990 as a joint venture between Acorn Computers, Apple Computer (now Apple Inc.), and VLSI Technology (now NXP Semiconductors N.V). Through an innovative model of licensing its intellectual property (IP), the company’s silicon designs powered early Nokia mobile phones. Thirty years later, Arm states that more than 99 percent of the world’s smartphones are based on Arm technology, and the company has expanded into the high-performance computing (HPC) and datacentre spaces, partly thanks to projects funded by the European Union (EU). HiPEAC spoke to John Goodacre (University of Manchester / UK Research and Innovation), an early advocate for multicore at Arm, to learn about the architectural advances and international collaborations that helped Arm-based technology become so pervasive.

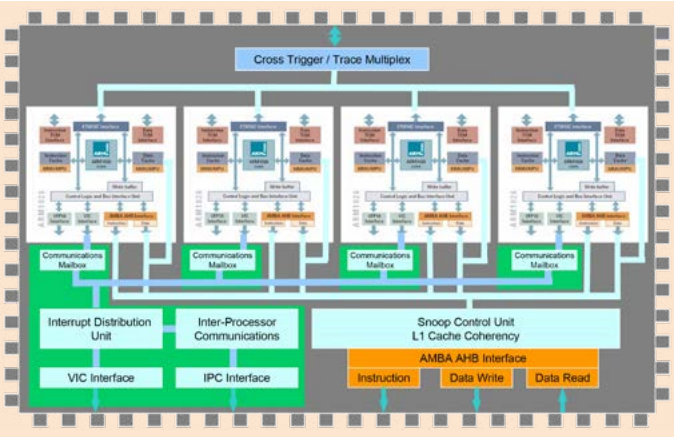
How Arm multicore grew from mobile to HPC, with a little help from the EU

When John Goodacre joined Arm in 2001, multicore processors hadn’t yet made it into mainstream desktop computers, let alone mobile devices. Yet even in the pre-smartphone era, when the main applications were SMS (short message service) text messaging and basic games, there was a need for greater compute power. ‘Having joined Arm from Microsoft, I’d come from the enterprise software multiprocessing background, so I understood the benefits of multicore processing for diverse workloads,’ he says.

Although Arm was pushing towards its first gigahertz-level central processing unit (CPU) at that time, John could see that single-core processors were simply not going to deliver the scalable performance – and low power usage – required for next-generation mobile computing. In his first public presentation of the proposed multicore strategy for Arm in 2003, he outlined the initial concept: up to four CPU clusters connected with a ‘snoop control unit’ (SCU) for memory coherency and an integrated interrupt controller. This was implemented on a stack of field-programmable gate array (FPGA) boards, a visual stimulus to the chiplet model and 3D stacking of IP, which have been gaining traction in industry for some years since.

In early 2004, the Arm11 MPCore was first released as a single intellectual property (IP) core, capable of implementing between one and four CPUs. The first customers for the Arm multicore concept were in image processing – producers of digital cameras, etc. – who had already hit the limit of what a single processor could do. ‘Initially, this concept was mostly used as an easy way to build a system-on-chip (SoC) with multiple processors. Selling the concept was pretty easy—I would run a workload, typically playing a video on one CPU, dynamically turn on the other three CPUs and the power consumption would halve, without any dynamic voltage and frequency scaling (DVFS), while still playing the video at the same quality and speed,’ John notes.

By 2008-9, it was becoming increasingly clear that smaller transistors and more pipeline stages would not continue to deliver increasing gigahertz. Despite continued market pushes for single-core gigahertz designs, multicore had become a clear way forward. However, John was aware that there would be significant challenges and software requirements when scaling up the multicore approach, not least the complete absence of fundamental software. ‘Implementing a scalable



A pre-product illustration of what an Arm multicore design might look like, from John's presentation at MPSoC 2003. This was eventually productized with an Arm11

multicore roadmap was always going to involve building out the ecosystem, particularly the system software. In the early 2000s there was no operating system, real time or otherwise, for Arm multicore devices, so there was a lot of work ahead.'

From multicore to chiplets, across mobile, data centres and HPC: Chapters in EU-funded research

To build out the ecosystem, it was necessary to work with silicon manufacturers, third-party IP providers, operating system vendors, and device and software manufacturers. As part of this effort, a series of EU-funded projects developed different aspects of the ecosystem.

'Arm scalable multiprocessing was the root of various EU-funded projects from 2008 onwards,' explains John. 'Three big questions were tackled across these projects. The first was how do you make it an architecture that is suitable for HPC and cloud. One of the primary questions there was the virtualization, which was addressed primarily through the ICT-eMuCo, VIrtical and SAFE projects,' says John. 'Then there was the physical deployment: how do we use the chiplet architectures to build a system architecture with the memory system for HPC? Then there was the question of what had to be changed in the instruction set to make the chips suitable for HPC.'

Proof of concept through virtualization

The first project, ICT-eMuCo (Embedded Multi-Core Processing for Mobile Communication Systems, 2008-2010, grant agreement no. 216378) investigated the scalability multicore approach for mobile devices. The eMuCo consortium developed an operating-system infrastructure, virtual hardware platforms – including a reference platform which developed the big.LITTLE approach for maximum performance and energy efficiency – and a methodology for efficient application programming on multicore architectures. 'We demonstrated the adaptive power-management advantage of the multicore approach to big.LITTLE, which is now an accepted power / performance technology,' says John.

A later project, VIrtical (SW / HW extensions for virtualized heterogeneous multicore platforms, 2011-2014, grant agreement no. 288574) explored the multicore model for embedded devices through virtualization, which offered a way to uncouple applications from hardware, thereby allowing the flexibility required for multicore operation.

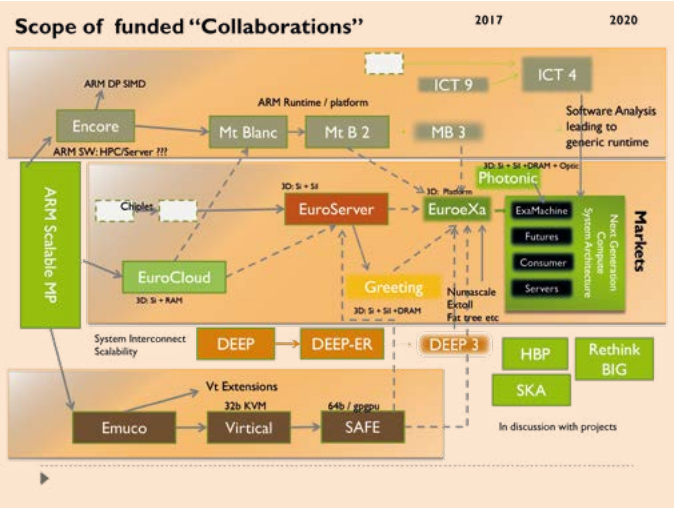
Chiplet approach to system architecture

The EuroCloud project (Energy-conscious 3D Server-on-Chip for Green Cloud Services, 2010-2013, grant agreement no. 247779) investigated RAM stacking and developed cloud-native servers with Arm cores. As discussed in detail in 'Scale-out processors' on pp. 26-27, this research was commercialized by the semiconductor company Cavium, resulting in the Cavium ThunderX.

The EUROSERVER project (Green Computing Node for European micro-servers, 2013-2017, grant agreement no. 610456) sought to leverage Arm-based designs for low-power data centres. Chiplet prototyping was one of the main themes of EUROSERVER, while the memory model, UNIMEM, was matured within the project, allowing the software architecture to be reused. 'If you can define how chiplets interact, in a coherent, scalable manner, with shared memory, this provides the logical means for chiplets to plug into your architecture and address map for a given device. UNIMEM was a significant step towards providing the necessary modularity and interoperability for this concept,' explains John.

Two start-up companies emerged from EUROSERVER: Kaleao Ltd, launched to commercialize a server solution, and ZeroPoint Technologies AB, which introduced memory compression as a means of delivering up to 50% more performance per watt. Along with Numascale hardware and OnApp virtualization, Kaleao's cloud-server tecnologies were used for a later project, ACTiCLOUD (ACTivating resource efficiency and large databases in the CLOUD, 2017-2019, grant agreement no. 732366), which developed a more efficient, scalable cloud architecture.

A trio of projects running from 2015-2019 – ExaNoDe (European Exascale Processor Memory Node Design, 2015-2019, grant agreement no. 671578), ExaNeSt (European Exascale System Interconnect and Storage, 2015-2019, grant agreement no. 671553) and ECOSCALE (Energy-efficient Heterogeneous COmputing at exaSCALE, 2015-2019, grant agreement no. 671632) built on the previous projects' work, such as the UNIMEM memory model, to develop HPC systems



Slide taken from a 2014 EUROSERVER presentation by John Goodacre, showing the scope of EU-funded projects building out the Arm multicore ecosystem

for the exascale era. These projects culminated in EuroEXA (Co-designed Innovation and System for Resilient Exascale Computing in Europe: From Applications to Silicon, 2017–2021, grant agreement no. 754337), which 'would then prototype the physical architectural machine with that memory model', John explains.

Ensuring an architecture fit for scalability

EnCore (ENabling technologies for a programmable many-CORE, 2010–2013, grant agreement no. 248647) provided a programming environment to simplify software development for manycore systems, increasing portability and scalability. 'We asked: what is it in this architecture that doesn't work?' says John. 'We had to fix things in the microarchitectural approach to a scalable chip: for example, we found out that not doing SIMD double-precision float meant that the floating point was absolutely useless, and that back-to-back copies of memory were essential.'

While these issues were being addressed in EnCore, the first Mont-Blanc project (European scalable and power efficient HPC platform based on low-power embedded technology, 2011–2015, grant agreement no. 288777), launched with the aim of delivering an Arm-based HPC prototype. With its promise of building a supercomputer from 'mobile' chips, it captured the popular imagination and was continued in a series of follow-up projects.

Today, the Fugaku supercomputer, the most powerful the world when it was first deployed, is testament to Arm's success in the HPC sector. The addition of the Arm Scalable Vector Extension (SVE) provided the final computing muscle, but John credits this suite of European projects with delivering the architecture, communications and memory necessary for Arm-based designs to handle HPC. 'It all comes back to the initial multicore strategy, followed by our investigations into why Arm multicore couldn't, at that time, do HPC workloads – the instruction set, the architecture, the memory system. All we had to do was fix all of those, and work out how to deploy it cost effectively,' he says.

While global events including the UK's exit from the European Union meant that Arm-based designs for HPC were no longer IP within the EU, John maintains that the original technology strategies were a success. 'It was a collaborative European effort based on the idea that multicore could do more than mobile handsets,' says John. 'When you look back and see how far the ecosystem has come since my initial multiprocessing presentation, it is quite extraordinary.'

John presenting at the EuroEXA workshop during HiPEAC 2018

Europe's control over its technological destiny has become an increasingly fraught topic. Sandwiched between superpowers, lagging in many areas of computing technology, and reliant on an increasingly unstable global supply chain, the European Union (EU) has laid foundations for homegrown computing through numerous research projects investigating potential avenues for processor intellectual property (IP).

One of the most exciting research areas is high-performance computing (HPC), where initial experiments with Arm chips have recently been joined by a flurry of projects focusing on the RISC-V instruction set architecture (ISA). HiPEAC spoke to Filippo Mantovani, an established researcher at Barcelona Supercomputing Center (BSC), to find out more.

RISC-V for HPC
The journey in Europe so far

Given that processors are the basis of computing, the ability to produce chips is a key plank of technological sovereignty. The American chip giant Intel was, for many years, the dominant force in consumer chips, but it was blindsided by the emergence of the mobile market, in which IPs developed by the British company Arm gained dominance.

For many years, the European Commission supported the development of Arm-based IP through research projects that helped pave the way for the expansion of the low-power chip architecture into the high-performance computing (HPC) and datacentre sectors. With Arm's acquisition by Softbank in 2016 and the UK's exit from the European Union, it became clear that the EU might need to diversify strategies to strengthen its technological autonomy.

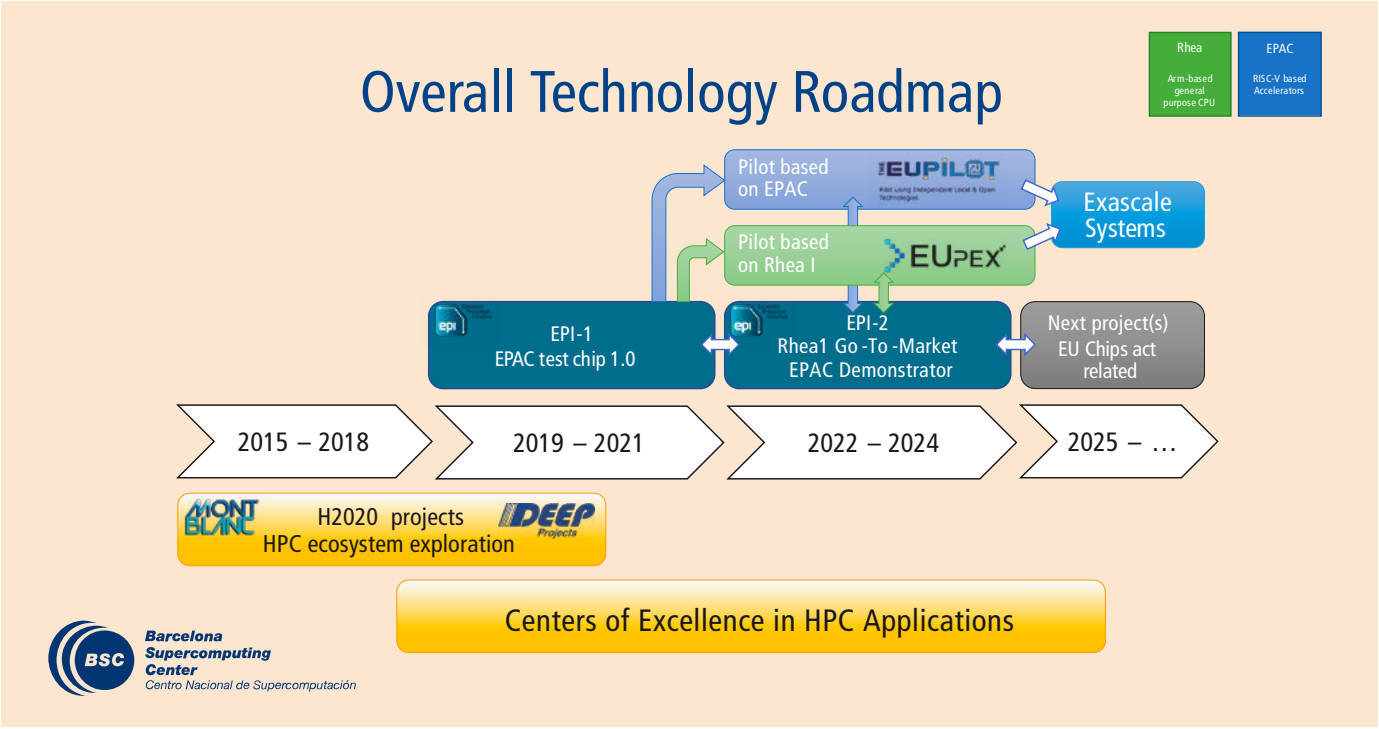
Enter RISC-V

The emergence of the open RISC-V instruction set architecture (ISA) in 2010 offered a new possibility: an avenue by which Europe could develop its own technological ecosystem and, in the process, build much-needed hardware experience in the bloc. Early champions of the RISC-V ISA in Europe included ETH Zürich and the University of Bologna, whose joint PULP Platform collaboration started taping out chips in 2013.

This opened up the intriguing possibility that RISC-V could be the basis of future, designed-in-Europe high-performance systems. In fact, thanks to the lack of incumbents, Europe could even become a trailblazer in this area, as Mateo Valero of Barcelona Supercomputing Center (BSC) explained to HiPEAC in 2022 (see HiPEACinfo 66): 'By expanding the RISC-V ecosystem, Europe can leapfrog into new technology areas and overtake the current leaders.'

In 2018, the European Processor Initiative (EPI) was launched to develop the fundamental European technologies for a high-performance computing node based on European technology, explains Filippo Mantovani, an established researcher at BSC. 'The European Processor Initiative has two main development strands. The first is the development of a general-purpose processor, based on the Arm architecture, and its natural successor is the EUPEX project. The second strand involves the development of a set of RISC-V-based accelerators, and this strand is being built upon by the EUPILOT and eProcessor projects,' he notes.

The EPI HPC node comprises an Arm-based central processing unit (CPU), which acts as a host, coupled with RISC-V based accelerators. Within the accelerator strand, known as EPAC, there are three pillars: the VEC, a vector central-processing unit



(CPU), is being developed by Barcelona Supercomputing Center and Semidynamics; the VRP (VaRiable Precision) accelerator is being developed by CEA; and the STX (Stencil Processor) accelerator is a joint effort between ETH Zurich and Fraunhofer.

Why RISC-V?

The advantage of RISC-V is that it offers a standard among the various proprietary ISAs on the market today,– rather as the USB (universal serial bus) standard means you don't have to buy a different cable for each device. According to Filippo, the ISA is the most important interface in computing as it regulates how software interacts with hardware. Examples of ISAs include Intel's X86-64, SPARC (developed at Berkeley), MIPS (developed at Stanford), IBM Power, and Arm.

Developed at Berkeley in 2010, the open RISC-V ISA offers several advantages over the proprietary ISAs that have dominated for decades. First, it is open and royalty free, allowing anyone to develop chips based on the RISC-V architecture. This flexibility enables diverse business models where both companies and research entities can adopt RISC-V, deciding independently how to commercialize or protect their final products. Second, having a universal ISA significantly benefits system software. Essential tools, such as compilers, can rely on a robust, shared core that is developed collaboratively and supported by companies with vested interests in the RISC-V ecosystem. Finally, RISC-V is modular, with a base set of core instructions that can be expanded through extensions. These extensions add specialized functionality, tailoring hardware to specific needs. One noteworthy example is the vector extension, ratified in 2021, which has been adopted within EPAC's VEC to enhance computational capabilities.

Milestones in European processor design

For the VEC, the EPI team opted for a vector architecture, using long vectors. Such architectures can be programmed like standard scalar processors using a standard programming language allowing for portability and eschewing the complexity of the host-device model embraced by NVIDIA, according to Filippo. 'With NVIDIA's graphics processing units (GPUs), programming requires a different binary for the accelerator, plus you have the overhead of offloading to the GPU and the overhead of moving data to and from the GPU, all of which makes this model complex for the programmer and less portable among different HPC systems.'

RISC-V is a modular ISA, with a mandatory base ISA which can be extended for different applications. The RISC-V Vector (RVV) extension of RISC-V extends this ISA with vector instructions. The EPI project realized the first European RISC-V vector CPU implementing the RVV extension defined by the RISC-V committee which booted Linux and supported a vector unit with very large vectors of 256 double-precision (DP) elements; for comparison, other CPU vendors, such as Intel, used a maximum vector length of eight DP elements. 'The project gave us the opportunity to explore this design point in comparison with other RISC-V architectures, CPUs, GPUs and other long-vector architectures such as the NEC Aurora SX. In other words, it allowed us to experiment with a design point that is somewhat outside the mainstream and companies would be wary of exploring,' says Filippo.



Bringing up the EPAC

Drawing on lessons learned in the EU-funded Mont-Blanc project, which helped expand the Arm ecosystem for HPC applications, the EPI has recognized the need to solidify the RISC-V software base to contribute to a solid HPC software stack. Compilation and system software are a key area of focus, with one highlight being the support for the RISC-V vector extension v1.0 in the LLVM compilation infrastructure.

More generally, EPI and the other EU projects adopting RISC-V have provided European teams with the experience of validating a full technological chain needed to produce a chip, allowing them to gain expertise in which Europe was severely lacking. 'The EPI has produced European IPs which have been validated through the tapeout process,' says Filippo. 'In addition to building capacity in processor design in Europe, it also produced physical chips, which act as an inspiration for others considering this path.'

This work also helped nurture an environment of 'development driven by measurement', Filippo notes, breaking the traditional 'hardware vs. software' conflict. 'Thanks to the open standard, having access to the architectural internals, and having the tools to inspect them, allows us to measure and understand in depth the effect of design points, so we can then report back to hardware engineers, compiler developers and even scientists who are early adopters of our systems,' he says. Filippo notes that GPUs, for example, while powerful, are also far from fully exploited, which means that the computational efficiency of many supercomputers tends to be low. While there will always be a compromise between specialized hardware and the software that can exploit it, close cooperation between hardware designers and software developers helps deliver a more efficient product.

RISC-V workshop at HiPEAC 2023

To ensure its applicability in real-life settings, the technology developed as part of EPI is also being tested on real scientific codes with the European Centres of Excellence and other scientific communities. These include:

- CEEC, which focuses on computational fluid dynamics (CFD), and which has tested Alya, Flexi, Neko, NekRS, WaLBerla, and SOD2D codes on EPI technology;
- Plasma-PEPSC, which focuses on plasma physics simulations and which has tested the BIT1, GENE, GENE-X, PIConGPU, and Vlasiator codes;
- ChEESE, which developed exascale-ready solid earth simulation codes and will test the codes ExaHyPE, Salvus, SPECFEM3D, HySEA, SeisSol, AWP-ODC, FALL3D, ASHEE;
- and MultiXscale, which promotes the award-winning EESSI code for multiscale simulations.

The future of RISC-V-based HPC in Europe

The near future will see the tapeout of the VEC and its integration into boards as part of EUPILOT. This will be followed by a project to develop large-scale European HPC and artificial intelligence (AI) technologies based on RISC-V.

Could future RISC-V based accelerators challenge the dominance of GPUs in the HPC market? While Filippo acknowledges that competing with NVIDIA is far-fetched, he believes they could attract profiles of research institutions and companies that care about efficiency. 'However, the development of this technology is a long process, as our experience with Arm-based processors for HPC at BSC has shown. In 2011, for the Mont-Blanc project, we started experimenting with Arm chips using Android development kits for HPC installations. It took another 10 years for the Arm-based supercomputer Fugaku to see the light – and that was with Arm and Fujitsu driving this work in the background. It's easy to see how widespread adoption of RISC-V accelerators for HPC could take even longer,' he adds.

One of the founding members of HiPEAC, Olivier Temam was at Inria for almost 10 years before joining Google in 2014. Now a director at Google DeepMind, he saw the potential of neural networks at an early stage. HiPEAC caught up with Olivier to find out about milestones in machine learning, finding the right path in neuromorphic computing, and the potential for chip development in Europe.

'My first paper on neural networks was rejected many times'

What, for you, have been the main milestones in machine learning (or computing systems more generally) since HiPEAC began in 2004?
Primarily, the advent of deep neural networks (DNNs) around 2006, and in 2012, the very significant progress marked by AlexNet which won the Large Scale Visualization Challenge by a large margin. From that point on, it became increasingly clear that we were on a path of exponential progress of the capacity of neural networks (NNs), which carries to this day through large language models (LLMs).

From a computing systems point of view, it became also clear back then that (1) the application scope of such NNs was so broad that it would warrant specialized/optimized architectures, and (2) that graphics processing units (GPUs) at the time, while great high-performance chips, did not pack enough compute capacity to cost effectively support the broad deployment of NN algorithms.

These notions led to a series of academic work on artificial intelligence (AI) accelerators, the first one being an ISCA paper in 2012 which laid the foundations for AI accelerators, followed by the academic work of the DianNao series of AI accelerators. Then to industrial work, especially tensor processing units (TPUs) at Google, which was considering broadly deploying NN algorithms since 2012 or so.

Your paper 'ShiDianNao: Shifting Vision Processing Closer to the Sensor' was recognised by the ISCA Retrospective 1996-2020. What do you think made this paper (and the DianNao accelerator work more generally) special, and what has the lasting impact been?
Thank you for kindly acknowledging this work. The real foundational articles were the 2012 ISCA article which explored how efficient NN accelerators could eventually become one day (and as an industry, we are trying to get as close as possible to that limit), and the first DianNao paper published at ASPLOS in 2014 which proposed the first NN accelerator design for inference applications (it won the best paper award at ASPLOS, and the test of time award in 2024). The other significant paper was the follow up DaDianNao paper on an NN supercomputer for training applications, published at MICRO in 2014 (and which also won the best paper award).

Given the initial scepticism about neural networks, has their stratospheric rise since 2012 taken you by surprise or did you see this coming?
Until 2006 or so, while I was interested in neural networks, their algorithmic progress didn't seem significant enough, compared to alternative image analysis algorithms. But the progress of DNNs since 2006 started to be quite significant and made a lot of waves, albeit in a small community (at the time).

Because of my personal interest in neural networks, I got advance signals that something profound was happening with NNs, and it is true that, especially in the computing systems community, initial propositions for NN accelerators were met with tremendous scepticism. When I gave a keynote at ISCA in 2015, I gently reminded the community that my first ISCA paper on NNs (2012) was rejected many times before I could publish it. Even though the programme committees of conferences in computing systems state that they welcome innovative

publications, the reality is that the consensus decision process of program committees leads to conservatism and filters out innovative ideas.

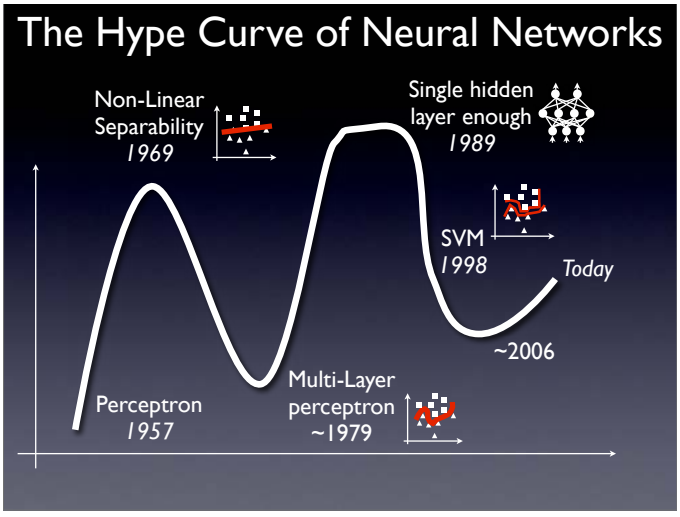
While no one really predicted the progress of AlexNet in 2012, once the regular progress of NN algorithms became clear (around that date), their potential application scope was so broad that it was fairly clear that they would lead to massive industrial applications and investment.

The other scientific element that gave a lot of confidence in that continued progress is the simple but powerful realization, in the ML community, that increasing the size of NNs seemed to consistently lead to significantly improved capacity. That spectacular trend continues to this day with LLMs.

Do you think HiPEAC has helped nurture the development of neuromorphic computing in Europe? What priorities do you think the community should focus on over the next few years?

As explained above, the computing systems community at the time was quite sceptical about neuromorphic systems, and HiPEAC was no different. More broadly, Europe made a strategic mistake about neuromorphic computing by choosing the wrong path. It invested in brain-inspired architectures (with the Human Brain Project) in 2013, and most European funds went to such architectures, in spite of clear signs since 2006, and especially since 2012, that artificial neural networks were extremely promising. That European choice is what led me, personally, to seek environments more focused on industrial applications and artificial neural networks.

But I do see very interesting and encouraging signs that Europe wants to reclaim its computing systems future with the European Chips Act. I would also encourage European investments in fabs for high-performance chips, and the HiPEAC community to contribute to that growing ecosystem. Also, because of the massive computing needs of AI, I would encourage the HiPEAC community to not just focus on chips, but the whole hardware stack, from chips to data centres and even energy production. Innovation must happen and be integrated across the whole hardware stack.



A slide from Olivier's presentation on neural networks at ISCA in 2010

What, for you, has been the main impact of HiPEAC more generally over the last 20 years?

I am very fond of the role of HiPEAC in Europe. As a co-founder of HiPEAC, I think that the European Commission helped us establish a powerful academic community in Europe, and I am also grateful for the steadfast leadership of Koen De Bosschere over so many years. Prior to HiPEAC, we were joking that Europeans would mostly meet in US conferences. I also praise the HiPEAC community for its tight interactions with the European industry, and making such academic / industry relationships a priority from its inception.

Chips (and computing systems as a whole) are a special technology in that they are used by virtually every industry on the planet; only energy may surpass chips in its industrial ubiquity. So it is extremely important for Europe's future that the continent invests in that technology, and HiPEAC can be one of the powerful engines to help make it happen. My only concluding advice, echoing some of the aforementioned comments, would be to carefully focus on industrially relevant applications and innovations.

FURTHER READING

Olivier Temam, 'A Defect-Tolerant Accelerator for Emerging High-Performance Applications', ACM/IEEE International Symposium on Computer Architecture (ISCA), June 2012 PDF pages.saclay.inria.fr/olivier.temam/files/eval/T12.pdf
BibTex pages.saclay.inria.fr/olivier.temam/homepage/bibtex/T12.txt

Tianshi Chen et al., 'DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning', ASPLOS '14: Proceedings of the 19th international conference on Architectural support for programming languages and operating systems, pp.269-284, doi: 10.1145/2541940.2541967

Y. Chen et al., 'DaDianNao: A Machine-Learning Supercomputer', 2014 47th Annual IEEE/ACM International Symposium on Microarchitecture, Cambridge, UK, 2014, pp. 609-622, doi: 10.1109/MICRO.2014.58

Z. Du et al., 'ShiDianNao: Shifting vision processing closer to the sensor', 2015 ACM/IEEE 42nd Annual International Symposium on Computer Architecture (ISCA), Portland, OR, USA, 2015, pp. 92-104, doi: 10.1145/2749469.2750389

Concerns about datacentre energy consumption are nothing new. Noting a wasteful mismatch between the desktop central processing units (CPUs) and datacentre requirements, researchers set out to deliver a chip that was fit for purpose. The result was commercialized and sparked a trend for Arm-based server chips for the cloud. HiPEAC caught up with Babak Falsafi (EPFL) to find out more.

Scale-out processors

How an EU-funded project pioneered cloud-native chips

In the early 2010s, data centres were expanding rapidly, and their energy requirements were spiralling. This bloated energy consumption was partly due to the fact that the hardware and operating systems used for servers were in fact originally designed for desktop computers, Babak Falsafi, professor in the School of Computer and Communication Sciences at EPFL, explains. 'Server chips were being built with x86 (Intel / AMD) technologies. These chips were primarily power-density limited: the cores were huge, and they had functionality that was not needed at the time by a lot of applications in the data centre, such as data serving, web serving, data caching, and analytics.'

Not only were the CPU cores a complete mismatch for the workloads, Babak notes; other aspects pushed up energy demands: 'The server chips also integrated a lot of on-chip memory to cool down the chip (memory has a much lower power density) which was at best not used by the applications but in the worst case slowed down lookups for information that resided in larger on-chip memory.'

From EU-funded project to Arm-based server chip

An EU-funded project, EuroCloud Server, whose consortium comprised Arm, EPFL, imec, Nokia and the University of Cyprus, was launched in 2010 to develop cloud-native servers with 64-bit out-of-order Arm cores, derived from the Cortex A-15 and 3D-stacked DRAM running cloud services. The project resulted in 'scale-out processors' that organized silicon resources into physical servers called 'pods' which shared I/O and memory ports and pins with other pods.

'We showed for the first time (even before Arm cores could be used for servers) that designing optimal chips from a performance and power density perspective would give a 10x improvement from silicon and the electricity used for popular open-source software stacks representing datacentre workloads. This 10x improvement came partially from having a lot more cores for a given electricity budget and having a faster instruction supply for deep software stacks in the cloud,' says Babak.

This work laid the foundation for an Arm-based CPU server chip for the cloud, the Cavium ThunderX, which employed 48 in-order Arm cores offering an order-of-magnitude larger core-to-cache silicon ratio than conventional server CPUs. However, Babak notes that, at the time, the software ecosystem (Linux) for Arm servers was not available; as a result, Cavium pivoted to the high-performance computing (HPC) – as opposed to cloud – market.

This has changed since: 'A few years later, Amazon Web Services and HiSilicon started not only building Arm servers but also building a community to create the Linux server ecosystem which is now mature,' says Babak. 'NVIDIA is benefitting from this effort and their future chips with tightly integrated graphics processing units (GPUs), CPUs, memory and the network using Arm cores. With liquid cooling, which is expensive but needed for GPUs, NVIDIA can also achieve much higher power densities for their Arm CPUs, allowing them to shift much of the power to computing rather than less productive on-chip memory.'

The future of server design

With unprecedented energy demands projected to meet the requirements of artificial intelligence (AI) workloads, datacentre efficiency is once again in the spotlight. However, Babak notes that this should be kept in perspective. 'Datacentre electricity consumption has been flat for some time worldwide, thanks to consolidation and improvements in efficiency at roughly 200 TWh per year. Goldman Sachs projects that this number will grow at 16% per year starting at 2019 until 2030 and AI will contribute a bit to it,' he says. 'But one should not forget that server chips have also seen a small exponential growth in TDP (thermal design power) – a measure for the maximum amount of heat a chip can dissipate – since 2019 in CPUs and a doubling of TDP per generation in AI chips.'

What does this mean for the design of future servers? 'We are at the limits of power density today and there are no simple silver bullets to improve efficiency. The latest Arm server chips by Ampere and the latest server chips from Intel are now resorting to dramatically more efficient cores (both in terms of area and electricity budget) to improve performance density and performance per watt,' says Babak. 'We are witnessing the post-Moore era of computing where the next waves of efficiency will come from a tighter integration of technologies from algorithms to housing infrastructure. The next big change in the data centre will most likely come from integration at the rack scale, not just for AI but also for the cloud.

Babak Falsafi taught a cloud-native server architecture course at ACACES 2023, the HiPEAC summer school

FURTHER READING

Pejman Lotfi-Kamran, Boris Grot, Michael Ferdman, Stavros Volos, Onur Kocberber, Javier Picorel, Almutaz Adileh, Djordje Jevdjic, Sachin Idgunji, Emre Ozer, and Babak Falsafi, 'Scale-out processors', Proceedings of the 39th Annual International Symposium on Computer Architecture (ISCA '12). IEEE Computer Society, USA, 2012, pp.500–511 doi: 10.5555/2337159.2337217

Pejman Lotfi-Kamran, Boris Grot, Michael Ferdman, Stavros Volos, Onur Kocberber, Javier Picorel, Almutaz Adileh, Djordje Jevdjic, Sachin Idgunji, Emre Ozer, and Babak Falsafi, 'RETROSPECTIVE: Scale-Out Processors, in José F. Martínez and Lizy John (eds), ISCA@50 25-Year Retrospective: 1996-2020, ACM SIGARCH and IEEE TCCA, June 2023 bit.ly/isca50_retrospective

Babak Falsafi, Michael Ferdman and Boris Grot, 'Server Architecture From Enterprise to Post-Moore', in IEEE Micro, vol. 44, no. 5, pp. 65-73, Sept.-Oct. 2024, doi: 10.1109/MM.2024.3418975

The first lecture from Babak's 'Cloud-native server architecture' course at ACACES 2023 is available to view on HiPEAC TV: youtu.be/se2ceP-vbPA

In this interview, Rainer Leupers, RWTH Aachen University, explains how HiPEAC acts as a crucible for technology transfer from academia to industry, and details activities to expand the network's reach across Europe.

Bridging industry and academia

How HiPEAC has supported technology transfer and widened the European network

How did the HiPEAC network contribute to the TETRACOM and TETRAMAX Innovation Actions?

Like its predecessor project TETRACOM, TETRAMAX was deeply rooted in the HiPEAC community. The idea was to conceive and to perform a project fully devoted to technology transfer, based on the powerful infrastructure already built by HiPEAC over many years. This would result in a clear win-win. Several project partners and staff had dual roles within both projects, which enabled perfectly seamless and synergetic cooperation. For instance, HiPEAC coordinator Koen De Bosschere served on the TETRAMAX steering committee (SC), while I had several roles in the HiPEAC SC over time. Other TETRAMAX consortium members were recruited from the HiPEAC community as well, including several partners whom Koen and I got to know during our joint workshop series for new European Union (EU) member states (see below).

Altogether, while other information and communication technology (ICT) tech transfer projects often waste some time in ramping up their own communities and client base, the tight links to HiPEAC permitted TETRAMAX to operate effectively from day one. Similarly, TETRAMAX reconfirmed the value of HiPEAC as an ideal tech-transfer platform.

As one of its key activities, TETRAMAX issued a series of open calls for co-funded Technology Transfer Experiments (TTX). HiPEAC had a key role in propagating these calls to a broad pan-European client community, thereby maximizing the number of high-quality TTX proposals received. We also leveraged HiPEAC's proven dissemination channels, such as *HiPEACinfo*, the newsletter, and the annual conference, to report about TETRAMAX activities. We organized special tech-transfer sessions, to highlight and promote several of our TTX clients' success stories. The large poster exhibits at several HiPEAC conferences, with a line-up of dozens of uniform TTX posters with their key results, were definitely among the most visible and successful dissemination events in TETRAMAX. Finally, TETRAMAX also organized an entrepreneurial track at the HiPEAC summer school, ACACES, allowing the project to take advantage of an established and well-attended event.

How has HiPEAC supported tech-transfer efforts at RWTH Aachen?

My ICE institute at RWTH Aachen has a long track record of tech-transfer projects with industry across the globe. Some of

these were actually ignited by spontaneous meetings at HiPEAC events, where the main actors from international academia and industry come together and generate new project ideas. Interestingly, this often happens within informal meetings besides the official sessions. We also highly appreciate the ACACES summer school, where PhD students sharpen their profile by participating in lectures and keynotes from both leading scientists and seasoned entrepreneurs.

The launch of startups itself, at least in our case, is not immediately related to EU-level projects like HiPEAC, since the formation of founder teams mostly follows other, more locally oriented, mechanisms and funding schemes. However, once a new startup has been born and is getting out of stealth mode, HiPEAC provides a wealth of opportunities and support. For instance, Silexica (which had a successful exit to Xilinx / AMD) exhibited and recruited via booths at HiPEAC events. Our recent spin-offs, MachineWare and RooflineAI, also benefit from news articles in *HiPEACinfo* magazine and press releases widely promoted by HiPEAC. This type of support clearly helps startups to obtain broad visibility, which is a key prerequisite for new market actors for successfully generating international business in the long term.

Why did you initiate the expansion of the HiPEAC network in new EU member states? What was the broader impact of this work?
Some years ago, the European Commission highlighted that there was an undesirable imbalance across the EU in European research frameworks and projects. While it appears natural that large countries carry out more projects than smaller ones, the participation of EU new member states (NMS) was clearly under-proportional, especially concerning the so-called 'EU13 countries' in Central and Eastern Europe.

Drawing on my connections to Poland, I initiated a couple of individual research visits to the leading technical universities in Warsaw, Wroclaw and Krakow, sometimes even supported by the German Embassy. This resulted in valuable new partnerships, but I soon realized this wouldn't yield the critical mass, so I brought the issue up within HiPEAC. Looking at our members list, we found that EU13 countries were also underrepresented, and we decided to take action. We found that local, one-day workshops in the capital EU13 cities would be the way to go. Each workshop was jointly organized with a local host capable of assembling the major local academic and industry stakeholders for an in-depth exchange on local technical expertise and, of course, to promote HiPEAC and its benefits to the community.

In the course of that campaign, Koen and I (often accompanied by other HiPEAC partners) visited all EU13 states and got a lot of insight into their local expertise and working constraints. In a later round, we expanded that campaign towards the current EU candidate states like Albania, Bosnia-Herzegovina, and Montenegro. I'm proud to say that today we have HiPEAC members from virtually all European countries, thus connecting their local champions much better to the traditional Western European communities. In many cases, e.g. in TETRAMAX, NMS country entities have become invaluable partners in our projects. The HiPEAC NMS campaign is still going on now on an informal basis, so we may expect continuous gains in our efforts towards true pan-European collaboration and networking.

Rainer Leupers and Koen De Bosschere visited all EU13 states, including this workshop at Epoka University in Tirana, Albania

Over the twenty years of its existence, HiPEAC has had an impact on the careers of legions of researchers. From attending the ACACES summer school as students, to academic exchanges and internships, to finding partners for research projects, to building an ecosystem for their technology, HiPEAC has always strived to support computing researchers at every step. In this article, we hear personal stories from the community about how HiPEAC was instrumental in their research goals.

A network built to last

HiPEAC's impact on the European computing community

Building a research community with HiPEAC ResilientHPC workshops

Manolis Marazakis (FORTH) and Petar Radojković (BSC)

From 2018 to 2023, the authors of this note had the opportunity to organize a series of thematic workshops on the broad topic of systems resilience, with particular emphasis on high-performance computing (HPC) systems. Each of the workshops included a keynote presentation, three to four contributed talks, and a panel session on systems-resilience topics.

Resilience for HPC systems encompasses a wide spectrum of fundamental and applied research and development, including theoretical foundations, fault detection and prediction, monitoring and control, end-to-end data integrity, enabling infrastructure, and resilient computational algorithms. Moreover, facility operations and cost management concerns need also to be weighed in, in the context of a systematic risk-management framework.

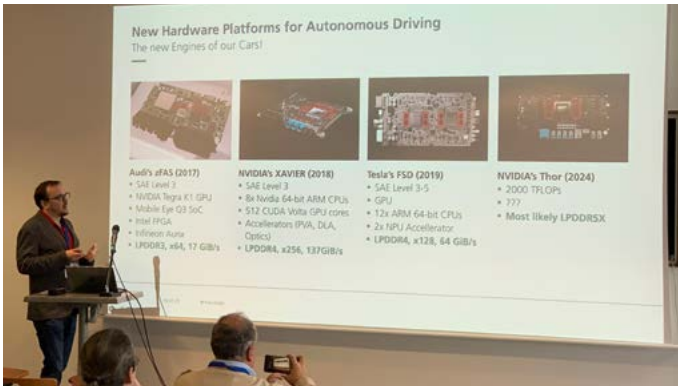
Our aim was to bring together experts and practitioners from the broad spectrum of computing-systems technologies to further research and development in HPC resilience, and to foster exchanges and collaboration across diverse communities. HiPEAC has served over the years as a convenient meeting place, and enabled us to form the nucleus of a special-interest community in Europe, the European HPC Resilience Initiative. This community also created a blueprint, 'Towards Resilient EU HPC Systems', which is one of the starting points for the recently-released RISC-V RAS 'Error Record Register Interface' (RERI) specification (see 'Further information', below).

Our experience is in line with HiPEAC's stated mission to advance computer architecture and computing systems research and development as a discipline in Europe. Our particular topic might seem, at first take, to be at the boundary between more clearly defined domains. Nevertheless, we have greatly

HiPEAC22 - Bianca Schoeder

HiPEAC23 - Chris Temple

HiPEAC23 - Dimitris Gizopoulos



HiPEAC23 - Matthias Jung

benefitted from the collaboration and networking opportunities that HiPEAC stimulates. Our workshop series was hosted and promoted by HiPEAC, and we will always remain grateful for the effective and courteous support that we consistently received over time. We are privileged to attend this year's conference and workshops, to interact with colleagues from several disciplines, and celebrate with them HiPEAC's 20th anniversary.

Acknowledgements

We would like to extend our heartfelt thanks to all contributors and attendees to our workshop series, and particularly express our gratitude for the contributions by our keynote speakers. In chronological order: Franck Cappello (Argonne National Laboratory, USA), Onur Mutlu (ETH Zürich, Switzerland), Christian Engelmann (Oak Ridge National Laboratory, USA), Bianca Schroeder (University of Toronto, Canada), Dimitris Gizopoulos (University of Athens, Greece), and Matthias Jung (Fraunhofer IESE, Germany).

Our effort over time has been supported by several EU-funded project activities, including ExaNeSt, EuroEXA, EuroLab4HPC, DEEP-SEA, eProcessor, EPI-SGA2, and EUPILOT.

Contact: maraz@ics.forth.gr / petar.radojkovic@bsc.es

FURTHER INFORMATION:

European HPC Resilience Initiative resilienthpc.eu

Petar Radojković, Manolis Marazakis, Paul Carpenter, Reiley Jeyapaul, Dimitris Gizopoulos, Martin Schulz, Adria Armejach, Eduard Ayguade, François Bodin, Ramon Canal, Franck Cappello, Fabien Chaix, Guillaume Colin de Verdiere, Said Derradji, Stefano Di Carlo, Christian Engelmann, Ignacio Laguna, Miquel Moreto, Onur Mutlu, Lazaros Papadopoulos, Olly Perks, Manolis Ploumidis, Bezhad Salami, Yanos Sazeides, Dimitrios Soudris, Yiannis Sourdis, Per Stenstrom, Samuel Thibault, Will Toms, Osman Unsal. 'Towards Resilient EU HPC Systems: A Blueprint', April 2020. https://resilientHPC.eu/blueprint2020

Petar Radojković, 'Towards resilient EU HPC systems: a blueprint. In Proceedings of the 16th ACM International Conference on Computing Frontiers (CF '19). Association for Computing Machinery, New York, NY, USA, 339, 2019 doi.org/10.1145/3310273.3323434

RISC-V RERI Architecture Specification. RERI Task Group, 2024 bit.ly/RISC-V_RERI_GitHubz

From HiPEAC reviewer to project coordinator and conference chair

Cristina Silvano, Politecnico di Milano

My journey with the HiPEAC community began when I was selected to be an expert reviewer for the HiPEAC Network of Excellence in 2005. This gave me a unique insight into the early stages of building the HiPEAC community. Drawing from my career in industry and academia, I contributed to shaping the first edition of the HiPEAC 'roadmap', which has since evolved into the HiPEAC Vision.

During these early stages, HiPEAC launched impactful dissemination and networking initiatives including the HiPEAC conference, the ACACES summer school, and 'Computing Systems Weeks'. These served as foundation seeds for successful careers and building a robust, influential European community on computer architectures and compilers, spanning domains from high-performance computing (HPC) to cloud and edge computing.

Over the last 20 years, the HiPEAC conference has grown in importance for the community, and I was therefore honoured

to be invited to serve as programme chair for HiPEAC 2021. Although the COVID-19 pandemic forced us to hold that edition of the conference online, I was able to reprise my role as general chair at the conference in Budapest in 2022 – for many of us, the first in-person conference since the start of the pandemic.

Building European project consortia

Thanks to my experience as a HiPEAC project reviewer, I also gained invaluable insights into the process of building research consortia and promoting multidisciplinary research collaboration. Motivated by this, in 2007 I led a project proposal focusing on design-space exploration for multiprocessor system-on-chip (SoC) architectures. The project, MULTICUBE (2008-2010), was approved, and was followed by 2PARMA (2010-2013), which focused on parallel programming models, design-space exploration and runtime resource management techniques for manycore architectures.

As a member of the HiPEAC community, I had the opportunity to meet numerous experts at European supercomputing centres such as CINECA, Barcelona Supercomputing Center and IT4Innovations. With CINECA and IT4Innovations, along with other major European research institutions and industry partners, we formed the ANTAREX project consortium. The main goal of ANTAREX (2015-2018) was to define a breakthrough approach for application autotuning and runtime management for green, heterogeneous HPC systems.

In this context, in 2019 we demonstrated the EXSCALATE virtual screening platform, developed by the pharmaceutical company Dompé, which targeted Zika virus pandemics. At the time, the ANTAREX4ZIKA HPC experiment was the largest virtual screening experiment ever undertaken in terms of computational threads (up to one million) and size of the compound database screened (one billion molecules).

This experiment demonstrated how supercomputers could act as technology enablers to explore the vast space of potential drug compounds and rapidly identify candidates for new drugs.

Cristina Silvano with João Cardoso at HiPEAC 2019

In 2020, the search for therapies to combat COVID-19 gave this work new urgency. The EXSCALATE4COVID project, coordinated by Dompé with the participation of Politecnico di Milano and CINECA, developed a fully integrated pipeline of in silico modelling and simulation, followed by in vitro experimental validation to speed up the search for active compounds to be tested in clinical trials.

More recently, I was invited to join the consortium of EUMaster4HPC, a project funded by the EuroHPC Joint Undertaking to design and implement the first pan-European HPC master's degree programme. I've also promoted the MSc in HPC engineering at Politecnico di Milano; starting in 2022-2023, this programme is the first of its kind in Italy.

To conclude, being an HiPEAC member has been very valuable for my career, introducing me to brilliant scientists and allowing me to build strong research collaborations and lasting friendships.

From ACACES student to company founder: An ode to HiPEAC

Anton Lokhmotov, KRAI

On the 20th anniversary of HiPEAC, I am pleased to share how HiPEAC has had an outsized impact on my career as a researcher, engineer and entrepreneur.

I joined HiPEAC in 2005 as a PhD student at the University of Cambridge. Attending the second ACACES summer school in 2006 was truly inspirational, with lectures by prominent professors such as Wen-mei Hwu, Margaret Martonosi, Ruby Lee and Rainer Leupers. I also met my future collaborators Grigori Fursin and Albert Cohen from Inria for the first time.

In 2008, as a post-doc at Imperial College London, I received a HiPEAC collaboration grant to visit Grigori in Paris. I was fascinated to learn about Grigori's innovative approach to software performance optimization called Collective Tuning (cTuning),

which originated from the FP6 MILEPOST project. We continued discussing cTuning for many years, until in 2014 we received funding from the FP7 TETRACOM project to develop from scratch a new cTuning framework called Collective Knowledge (CK).

At that time, I had been working at Arm for five years as a lead general-purpose graphics processing unit (GPGPU) compiler engineer and experienced firsthand many of the reproducibility and optimization issues that CK aimed to address. Recognizing the potential of the CK technology, Grigori and I co-founded dividiti to focus on its commercialization. Attracting early customers such as Arm and General Motors provided additional funding and use cases for CK, which eventually led to its wider adoption in MLPerf, the industry-leading competition of machine-learning (ML) systems (often called 'The Olympics of ML benchmarking'). An ISCA 2020 paper on the MLPerf Inference benchmark received a HiPEAC 2020 Paper Award.

In 2012, Albert Cohen and I started a collaboration on GPGPU compilation through the FP7 CARP project. My own experience at Arm pointed to the vast predisposition of industry towards permissive software licences such as MIT ('do what you want, just don't sue us') vs. 'copyleft' licences such as GPL. To minimize any barriers to adoption of our research software in industry, we agreed to release our PENCIL compiler under the MIT licence. This also triggered an aggressive relicensing of previously released polyhedral libraries such as isl, marking a shift away from copyleft licences in other research projects too. In 2016, Alastair Donaldson from Imperial College London, the coordinator of the CARP project, received a HiPEAC technology transfer award for marrying the CLsmith compiler fuzzing tool originating from the project with the CK framework. Alastair's further research in compiler fuzzing led to the founding of GraphicsFuzz, which was acquired by Google in 2018.

While during my PhD I declined a HiPEAC-sponsored internship opportunity at Arm, I made full use of the internship facility later, having hosted excellent research students at Arm and dividiti, including Dominik Grewe, Richard Membarth, Tobias Grosser, Diego Caballero, Cedric Nugteren, Ulysse Beaugnon, Georgios Pinitas, Marco Cianfriglia, Emanuele Vitali. Thanks to support from HiPEAC, I have managed to maintain a healthy balance between my industrial and research interests over the years.

In 2020, I founded KRAI to continue pushing the boundaries of research and development (R+D) towards creating ultra-efficient and cost-effective ML Systems. My involvement in HiPEAC has remained essential to this goal from recruiting exceptional talent to forming long lasting collaborations. I would like to congratulate the whole HiPEAC community on its 20th anniversary, and hope to be around for its 50th anniversary too!

Building a research career with HiPEAC

George Dan Mois, Technical University of Cluj-Napoca

I became part of the HiPEAC community in 2015, when my PhD advisor, Liviu Miclea, added me as an affiliate member. At that time, I didn't know the impact it would have on my career, both in industry and academia. Having graduated in 2011, my PhD defence was relatively recent, and I felt lucky to be able to interact with this world of outstanding researchers, top professionals and policy makers in the computing domain.

The first surprise was the availability of travel grants for HiPEAC members in my part of Europe to attend HiPEAC events – a significant benefit since research money was scarce in Romania at that time. I participated in the 2016 HiPEAC conference in Prague and attended inspiring presentations and sessions. One that I still remember, and that confirmed to me that I was in the right place, was a keynote talk by Krisztián Flautner on the internet of things.

Many inspiring events followed over the years. I particularly remember Alberto Sangiovanni-Vincentelli's keynote talk at the 2019 HiPEAC conference in Valencia, and Rainer Leuper's lectures promoting collaboration between academia and industry, a practice not that common at my university 10 years ago. I gained greater confidence in my skills and subsequently participated in several projects involving academic and industrial partners, including a TETRACOM grant that our team discovered through HiPEAC. Two of the projects I contributed to were honoured with a HiPEAC Technology Transfer Award.

These are just a few examples of the experiences that helped me advance in my career as an embedded systems engineer and researcher. I would like to express my gratitude to the HiPEAC team for the incredible opportunity I was given and for their outstanding efforts, particularly in supporting and facilitating access to the network for the younger generation.

HiPEAC's impact on…

Industry-academia relations

'Before HiPEAC, the research and development (R+D) collaboration between the computing research community and the relevant industry was sporadic and relied on personal relationships across different institutions. HiPEAC changed this dramatically with their networking instruments such as the conference, seasonal brainstorming workshops, internships, secondment programmes, etc. HiPEAC was the catalyst in creating many European projects, some of which were led by industry. In addition to being a networking platform, HiPEAC has also been producing a technology vision for computing, which predicts the technologies in computing that will have a huge impact on society and makes recommendations for directions in which investments should be made.

I have always said that HiPEAC has been the most successful European project funded by the European Commission.'
Emre Ozer, Pragmatic

International research collaboration

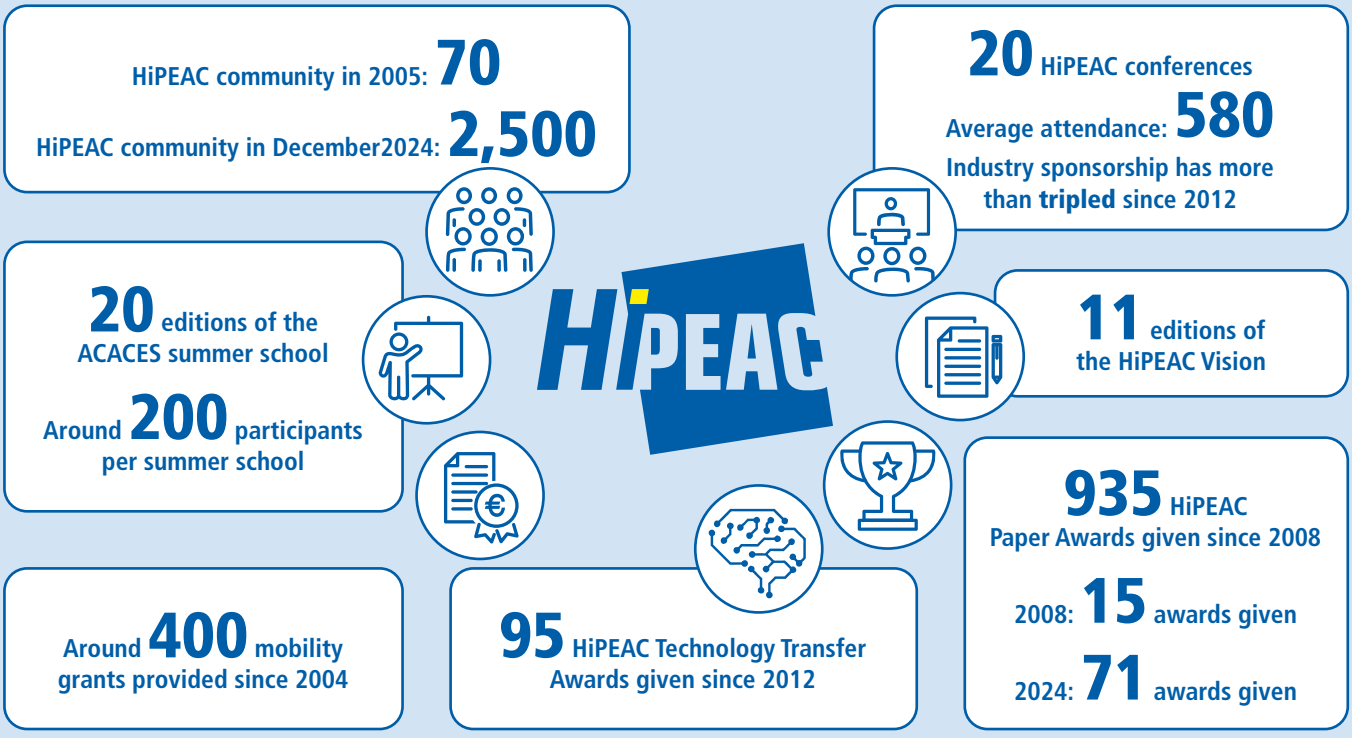
'It has been exciting to watch how HiPEAC has helped European computer architects – both in industry and academia – come together with unified overall goals and attack long-term opportunities at scale. HiPEAC has also been impactful in terms of student and workforce development – so many graduate students in the EU are able to move between HiPEAC-affiliated research labs in many countries, to broaden their research experience and fill gaps.

The net result has been a dynamic research community with lots of people well prepared (and well funded!) to go after important problems.'
Margaret Martonosi, Princeton University

Student-senior networking

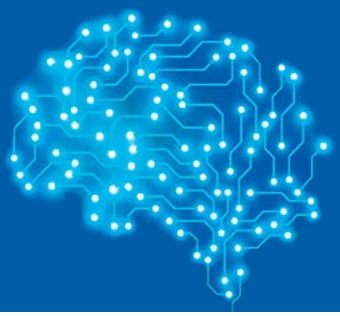
'ACACES is a great educational and networking tool, super powerful for both students and teachers.'
Filippo Mantovani, BSC

HiPEAC in numbers



HiPEAC Technology Transfer Awards 2024: The winners

The HiPEAC Technology Transfer Awards were launched in 2012 to recognize and encourage the transfer of leading-edge technologies from academia to industry. This year, eight candidates were selected, as detailed below. The winners will be recognized in an awards ceremony at the HiPEAC conference, and first-prize winners receive a cash prize of €1,000.



Graph of Thoughts: Solving elaborate problems with large language models

Maciej Besta, ETH Zürich, Switzerland

Graph of Thoughts (GoT) is an advanced framework that enhances large language models (LLMs) like GPT, PaLM, and LLaMA by allowing their reasoning to form a flexible graph structure, surpassing prior methods such as chain-of-thought and tree-of-thought. This approach enables more sophisticated problem-solving by modelling thoughts as vertices and their dependencies as edges, mimicking human cognitive processes. GoT has shown up to 62% improvement in tasks like sorting and document merging while reducing costs by 31%, supported by a strong GitHub presence with over 2,100 stars.

Industry applications include aiding a major chemical producer in scalable data analytics, enhancing event detection for a rare disease treatment leader, and aiding in semantic content indexing and legal compliance monitoring for media broadcasters, highlighting the scalability and robustness of the framework across sectors.

github.com/spcl/graph-of-thoughts

Roofline AI

Jan Moritz Joseph, Roofline GmbH, Germany

RooflineAI GmbH aims to make artificial intelligence (AI) deployment at the edge flexible, efficient and easy. The foundation of Roofline's development was established at RWTH Aachen University, where Roofline was spun off and successfully completed an IP transfer. Now, Roofline offers an innovative deployment solution as a software development kit (SDK) that utilizes a retargetable AI compiler toolchain.

roofline.ai

GenoGra: Pangenomic analysis platform

Marco Santambrogio, Politecnico di Milano, Italy

The GenoGra team

Pan-genomics is a novel approach to analysing genomic data. Instead of a single reference genome, it involves the usage of a composite reference of multiple genomes sampled from across a wider range of variability within a species. However, while pan-genomics ensures maximum accuracy and simplicity, it is computationally demanding and slow.

To overcome this problem, the 'GenoGra - Genome Analysis, Unleashed' research project was born in 2020 at the NECSTLab at Politecnico di Milano, aiming at developing high-performance solutions for pan-genomic analysis. The project's main outcome was a methodology to accelerate the sequence-to-graph alignment process on heterogeneous computing systems equipped with graphics processing units (GPUs), protected by an Italian patent application (2022) and its PCT extension (2023).

In June 2023, the research team founded GenoGra s.r.l., an innovative Italian startup that aims to bring the outcomes of the research project to the market. To achieve its goal, the company is developing a user-friendly platform where bioinformaticians can effortlessly run accurate pan-genomic tools that ensure high performance, thanks to the patented technology.

genogra.com

FAiNDER: FAiND the best hardware for your AI application

Petar Radojković, Barcelona Supercomputing Center (BSC), Spain
Team: Mariana Carmin, Javier Beiro, Victor Xirau, Pouya Esmaili, Petar Radojković, Eduard Ayguadé (all BSC), Emanuele Confalonieri, Rishabh Dubey and Jason Adlard (all Micron, United States)

In the context of a bilateral sponsored research agreement with Micron Technology, Inc., Barcelona Supercomputing Center (BSC) developed the FAiNDER online platform, which is designed to transform how researchers and developers navigate the rapidly evolving AI landscape. The platform is designed to help researchers and developers to explore and enhance the most relevant AI models and their performance.

FAiNDER offers up-to-date, comprehensive, filterable tables and interactive charts, making it easier to analyse vast amounts of information. It covers all major AI models: deep neural networks, convolutional neural networks, deep learning recommendation systems, transformers, graph neural networks and recurrent neural networks. The FAiNDER analysis provides detailed information on all phases of the AI software pipeline, their main architecture, arithmetic operations, data types, memory footprint and the target hardware platforms. It provides reliable information collected from the relevant scientific publications.

FAiNDER.eu

Interconnection network and memory subsystem IPs for high-performance, scalable SoC solutions

Madhavan Manivannan, InfiniNode Technologies AB, Sweden

RISC-V is rapidly gaining adoption due to its customizability, openness and its potential to drive innovation in future computing systems. As RISC-V–based microprocessors scale to meet the demands of AI and high-performance computing (HPC) applications, the performance of the memory subsystem and network-on-chip (NoC) becomes crucial. To support this scaling, careful optimization of the memory subsystem — especially the last-level cache — and the NoC is essential.

Left to right: Ahsen Ejaz, Mehrzad Nejat, Bhavishya Goel, Patrik Millsjo, Madhavan Manivannan and Ioannis Sourdis

A team of researchers from Chalmers University of Technology recognized that existing interconnect and caching solutions targeting RISC-V do not scale effectively for manycore processors. To address this issue, the team embarked on developing a set of intellectual property (IP) blocks as part of several national and EU-funded projects, including the European Processor Initiative. These IPs include a scalable coherent home node, a high-performance network-on-chip, and an optimized last-level cache.

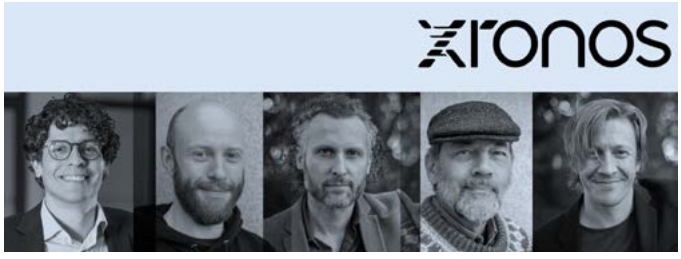
The team, composed of researchers and faculty from the computer architecture group at Chalmers University of Technology, is set to receive support from Chalmers Ventures, the venture-building arm of the university. Additionally, the team has formed a start-up company, InfiniNode Technologies, aimed at development and commercialization of these technologies and will be led by an experienced business developer. InfiniNode Technologies will collaborate with key academic and industrial partners through various national and European programs to develop the IPs and bring them to market. The company plans to engage with the broader RISC-V community, offering these IP blocks to accelerate the development of next-generation RISC-V–based high-performance system-on-chip (SoC) solutions.

infininode.com

LinguaFranca: A language to bring intelligent systems to life

Jeronimo Castrillon and Christian Menard, TU Dresden, Germany

Society and industry increasingly rely on software-defined intelligent systems that interact with the physical world. These systems are programmed using mainstream programming



models and software frameworks that lack abstractions to guide how software components synchronize, react to inputs, and manifest physical behaviour. As a consequence, such systems rarely provide safety guarantees and suffer from concurrency bugs that are hard to track and can be extremely costly if left unchecked.

Lingua Franca (LF), a large collaborative project originated at UC Berkeley, addresses this challenge with a concurrent programming model with explicit semantics for physical time. LF can run on embedded microcontrollers, multicore systems, and cloud systems, offering the same programming experience across these different tiers of computing platforms, automatically exploiting available parallelism, and delivering deterministic behaviour.

In January 2024, the team at TU Dresden (Christian Menard and Jeronimo Castrillon) together with researchers at UC Berkeley (Marten Lohstroh and Edward A. Lee) and Jeff C. Jensen (ex-Blackrock Neurotech, ex-Creator) co-founded Xronos Inc. to commercialize the LF technology, with the goal of accelerating the pace of innovation in robotics and automation – enabling the creation of intelligent systems we can depend on.

lf-lang.org
xronos.com

BlueSign: Translating Italian Sign Language into avatar signs

Roberto Giorgi, University of Siena

According to the World Health Organization (WHO), more than 1.5 billion people (nearly 20% of the global population) live with hearing loss, with 430 million of these having disabling hearing loss. Most of these individuals, particularly those who are prelingually deaf, rely on sign language for communication.

Sign language is crucial for deaf individuals to access most of the information needed for daily life, making it indispensable for their full social integration. For deaf people, sign language is often the first language learned and the preferred way of communication. Sign language is recognized by law in many nations worldwide and should be present in many formal contexts; however, human translators are not enough.

The Computer Architecture Laboratory at the University of Siena, Italy, developed a special encoding technique to integrate Italian Sign Language (LIS) into a larger number of devices. Recent developments involved collaborating with Quest-It, an Italian company specializing in AI solutions, to create an artificial human (displayed as an avatar) able to produce LIS on a web page, making it available in many contexts.

From DaCe to Daisytuner: Performance optimization as a service

Lukas Trümper, Daisytuner

Compilers are essential to software development, converting high-level code into low-level instructions that processors can execute. This process ensures that applications are optimized to leverage a processor's specific features. However, the optimization techniques employed by modern compilers rely on peephole algorithms and static cost models embedded within the compiler's code.

The Data-Centric Programming (DaCe) framework, conceived by the Scalable Parallel Computing Lab (SPCL) at ETH Zürich a decade ago, shifts the focus towards optimizing dataflow, which has become increasingly critical as the gap between memory speed and computation widens. By explicitly representing programs as dataflow and control-flow graphs with symbolic memory accesses, DaCe enables sophisticated analyses and performance optimizations via metaprogramming. This facilitates the development of efficient software across heterogeneous computing systems.

Building on this foundation, Daisytuner introduces DOCC, a novel framework for automatic performance optimization with an extensible database of performance optimizations, which can be automatically queried by the compiler. Utilizing neural networks, the compiler generates code embeddings for specific code regions and identifies optimal transformations in the database.

DOCC allows developers to:

- Optimize and transpile applications seamlessly across processor architectures, all within the integrated development environment (IDE) of the developers.
- Integrate with compilers via plugins, enabling fully automated performance improvements in autopilot mode.

Thanks to Daisytuner, developers can harness automatic, real-time optimizations without the need for specialized in-house resources, allowing them to meet the increasing performance demands of many applications.

Find out more about the HiPEAC Technology Transfer Awards on the HiPEAC website: hipeac.net/awards/#/tech-transfer

The EIC Work Programme 2025: Opportunities to innovate

Paul Pietrangelo, Lira

With a budget close to €1.4 billion in 2025, the European Innovation Council (EIC) presents many exciting opportunities for deep-tech innovators working in academia or industry.

EIC Pathfinder

The EIC Pathfinder funds research of radical and visionary ideas (technology readiness levels (TRLs) 1 to 4). Although high-risk, these projects should nevertheless have the potential for high reward, such as a significant technological advancement that could transform industries or create entirely new markets. The programme offers single applicants or consortia grants of up to €4 million to achieve proof of principle and validate the scientific basis of breakthrough technologies.

In 2025, the EIC Pathfinder has a budget of €262 million, with €142 million allocated to the open call where breakthrough ideas from any field of research can apply, and €120 million allocated to four challenge calls, each with €30 million. Specifically, for computing researchers, there is a challenge call titled 'Generative-AI based Agents to Revolutionize Medical Diagnosis and Treatment of Cancer'. While the focus of this call is generative artificial intelligence (genAI), other advanced AI technologies, such as topological and geometric deep learning, neural fields, graph neural networks, etc., may also apply. In 2025, the deadline for the open call is 21 May, while the deadlines for the challenge calls are 29 October.

EIC Transition

The EIC Transition is designed to fund the maturation and validation of a novel technology and its accompanying business idea from the lab to the relevant application (TRLs 3/4 to 5/6). This could involve prototyping, formulation, models, user testing, and validation tests, as well as exploring and developing sustainable business strategies for commercializing the technology in high-potential markets. For undertaking these activities, grants of up to €2.5 million are offered to single applications or consortia. In 2025, the EIC Transition has a budget of €98 million, and this is organized into an open call only. However, it is worth noting that the EIC Transition is not open to all applicants; the project must build on the results from an EIC Pathfinder, Future and Emerging Technologies (FET), European Research Council (ERC) Proof of Concept, or European Defence Fund (EDF) project. In 2025, the deadline for the EIC Transition call is 17 September.

EIC Accelerator

The EIC Accelerator funds single start-ups and small / medium enterprises (SMEs) that are developing high-impact, deep-tech solutions that can create new markets or disrupt existing ones. It offers these high-risk, high-potential companies grants of up to €2.5 million and equity investments of up to €10 million to fund the later stages of technology development as well as scale up (TRLs 6 to 9).

In 2025, the EIC Accelerator has a budget of €634 million, with €384 million allocated to the open call for companies from any industry, and €250 million allocated to five challenge calls, each with €50 million. Again, there is a challenge call of particular relevance to computing researchers: 'GenAI4EU: Creating European Champions in Generative AI'. This focuses on supporting startups and SMEs that are further developing and validating new genAI models, adapting existing models to specific sectors or data types, or integrating and testing genAI solutions in existing workflows. In 2025, there are two deadlines for both the open call and the challenge calls, 12 March and 1 October.

Lastly, the EIC Work Programme 2025 includes the launch of a new funding programme: EIC Strategic Technologies for Europe Platform (STEP) Scale Up. This supports the development and manufacturing of game-changing innovation with significant economic potential in strategic technology areas. This includes digital technologies and deep-tech innovations, clean and resource-efficient technologies (including net-zero technologies), and biotechnologies. It does so by offering scale-ups in line with the EU's key strategic priorities equity funding up to €30 million. In 2025, the EIC STEP Scale Up programme has a budget of €300 million and, interestingly for HiPEAC members, there is a specific focus on semiconductor and quantum technologies.

Paul Pietrangelo delivered the entrepreneurial course at the HiPEAC summer school, ACACES, in 2020, 2021 and 2024.

LOCO 2024: First International Workshop on Low Carbon Computing

Wim Vanderbauwhede, University of Glasgow, on behalf of the LOCO 2024 organizing committee

The first International Workshop on Low Carbon Computing took place on 3 December 2024 in Glasgow and online. There were 55 attendees in person and 60 online. It was organized by the Low Carbon and Sustainable Computing Theme of the School of Computing Science of the University of Glasgow.

The aim was to bring together researchers and practitioners with a keen interest in low-carbon and sustainable computing. The workshop provided a forum for sharing new ideas, for presenting ongoing work and early results, and for bringing forward well-founded criticism.

LOCO was designed to be a low-carbon, diverse, inclusive and accessible event: we encourage online participation and low-carbon travel, vegetarian food, diversity and inclusion on all levels (e.g. encouraging non-academic practitioners to participate); online participation was free, and the in-person fee was very low.

The themes were energy efficiency, carbon awareness, frugal computing, sustainable software engineering, computing for climate science and measurements, testbeds, and simulation. Most of the talks and discussions focused on energy efficiency, carbon-aware computing, embodied carbon and the rebound effect.

We received 40 submissions for regular papers, of which 21 were accepted, and 10 lightning talks, all of which were accepted. We use a post-proceedings model with submission of extended abstracts for the talks and full papers after the conference.

We want to make LOCO a recurrent event and have created a steering committee for that purpose. We will issue a call for bids to organize LOCO 2025 in April.

FURTHER INFORMATION

sicsa.ac.uk/loco/loco2024

bit.ly/Uni_of_Glasgow_low_carbon

Advancing ICT sustainability with HiPEAC

With six out of the nine planetary boundaries breached – while, at the same time, hyperscalers are ratcheting up energy demands – sustainability in information and communication (ICT) technologies is an ever-more urgent topic. HiPEAC has been pushing for sustainable ICT for many years, for example in the HiPEAC Vision and at the ACACES summer school. Initiatives within the HiPEAC community, such as LOCO (see above) and the Doctoral Summer School on Sustainable ICT (SICT) co-organized by ACACES 2024 tutor David Bol, complement these efforts.

HiPEAC is putting together a working group to focus on ICT sustainability. Further information will be available in due course. In the meantime, if you are interested in joining this group, contact info@hipeac.net.

David Atienza and Giovanni De Micheli win ESWEEK 2024 Test of Time Award

Along with fellow authors Srinivasan Murali, Almir Mutapcic, Rajesh Gupta and Stephen Boyd, HiPEAC members David Atienza and Giovanni De Micheli (both EPFL) were honoured with the Embedded Systems Week CODES+ISSS Test of Time Award for their 2007 paper 'Temperature-aware Processor Frequency Assignment for MPSoCs Using Convex Optimization'.

The paper featured an early adoption of temperature-aware frequency control in multiprocessor systems-on-chips. This technology explored optimization methods to solve processor frequency assignments, in a bid to reduce energy consumption and make temperature gains in small devices.

CODES+ISSS, the International Conference on Hardware/Software Codesign and System Synthesis, is the premier conference in system-level design, hardware / software co-design, modelling, analysis, and implementation of modern embedded systems, cyber-physical systems, and the internet of things.

On behalf of HiPEAC, congratulations!

Dimitris Gizopoulos wins the ACM SIGMICRO Distinguished Service Award 2024

Research work and publications is only one side of HiPEAC members' contribution to the computing systems community. Active participation on a voluntary basis in the organization of major international ACM and IEEE conferences from different roles is another side which has a significant impact to our community.

One of the three flagship symposia in the field of computer architecture (and the oldest one), the ACM/IEEE International Symposium on Microarchitecture (MICRO) is co-sponsored by the ACM Special Interest Group on Microarchitecture (SIGMICRO) and the IEEE Technical Community on Microprogramming and Microarchitecture (TCuARCH). The 57th edition of the symposium was held on November 2024 in Austin, Texas, USA. The 2024 ACM SIGMICRO Distinguished Service Award was presented to HiPEAC member Dimitris Gizopoulos, Professor of Computer Architecture at the University of Athens, Greece 'for his contributions to the SIGMICRO community, especially for the organization of MICRO during the pandemic period (2020 and 2021), two of the most pivotal years in the long history of the conference'. Dimitris Gizopoulos served as general chair of MICRO 2020 and MICRO 2021 which were initially planned to be held in the city of Athens, but which were held virtually during the pandemic years over Zoom.

HiPEAC would like to congratulate Dimitris on this award and thank him for his service to the community.

Petar Radojković and colleagues win Best Paper Runner Up at MICRO 2024

A paper which HiPEAC member Petar Radojković co-authored with colleagues at Barcelona Supercomputing Center and Micron, was selected as 'Best Paper Runner Up' at MICRO 2024, the ACM/IEEE International Symposium on Microarchitecture. Titled 'A Mess of Memory System Benchmarking, Simulation and Application Profiling', the paper was also presented during a panel at MEMSYS 2024, the 10th International Symposium on Memory Systems, which received the award for 'the most controversial, truth-telling panel discussion'.

Many congratulations on behalf of HiPEAC!

Once thought to be a myth, silent data corruptions (SDCs) have been recognized as a real problem – and finding the solution is not trivial, as Dimitris Gizopoulos (University of Athens) explains.

Why we need to get to grips with silent data corruptions

Incorrect results produced by a program executed on a computing system have always been a known issue for systems designers. Silicon chips and the computing machines built upon them are not perfect; they contain defects and occasionally produce wrong results even after thorough testing during manufacturing and in mission mode. But when an incorrect result exists silently – without any visible indication to the user – it severely damages our trust in the machine.

The first formal use of the term 'silent data corruptions' (SDCs) was during a panel at DSN 2008, the IEEE / IFIP International Conference on Dependable Systems and Networks. For decades, we told ourselves that SDCs were a myth, or, in the worst case scenario, an SDC might affect the unlucky user of one in a million – or even one in a billion – computers. Alas, the reality is very different.

Over the last four years, datacentre hyperscalers (Meta, Google, Alibaba) have disclosed an unexpectedly high number of central processing units (CPUs) – around one in 1,000 – that produce SDCs, i.e. program executions that produce wrong results without any observable indication. SDCs from other chips with data-level parallel architectures have also been reported, including graphics processing units (GPUs) and other accelerators. Reports are in agreement that the root cause of such SDCs are silicon chips which are born defective (i.e. that slipped through manufacturing testing), become defective through ageing, or simply differ from each other (timing variability).

Defining the problem

SDCs are, by their nature, neither detected nor corrected. What the community needs to do is to minimize or, ideally, eliminate the rate of SDCs. In a perfect world, the result of a computation produced by chips with or without defects should be either correct (i.e. the defect does not matter, or a correction worked), or incorrect but known to be so (i.e. detection worked but no correction is possible – just don't use the result!).

In terms of the scale of the issue, while the 'one in 1,000' disclosure is very useful, it only tells us that one chip in 1,000 has been identified as having a defect that could generated SDCs under certain circumstances. It is much harder to estimate the actual rate of SDCs. What we do know is that when silicon is deployed at scale, with huge numbers of devices running programs 24 hours a day, statistics are not on our side.

Until tight and validated bounds for SDC rates in the wild are measured or estimated, it is hard to take decisions for the protection – and the cost this protection incurs – of large-scale systems from defects leading to SDCs. While some industry players have recognized that the problem exists, failure rates in computing systems are valuable trade secrets; after all, who would risk stating that their computer chips fail, given the dependence of society upon them? Rather than the vague statements about SDC rates, more accurate estimates of SDC rates in the field are essential before cost-effective mitigations can be decided.

In the meantime, the research community should focus on computing systems resilience, developing solutions at all levels of the stack, from device physics and circuit levels all the way up to cloud system software.

This article was adapted by a post on the SIGARCH blog. For a fuller exploration of the topic and references, see sigarch.org/sdcs-a-b-c

Funded by the European Union

HIGHER: TOWARDS ALL-EUROPEAN PROCESSOR AND MANAGEMENT MODULES FOR CLOUD / EDGE INFRASTRUCTURES

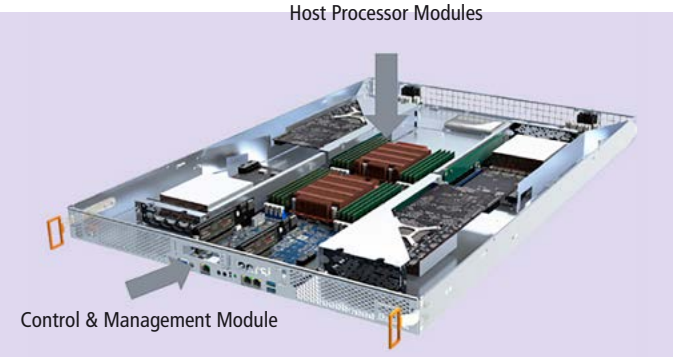
HIGHER aims to prototype and demonstrate the first all-European datacentre-ready processor and management modules and integrate them into cloud and edge infrastructures using European technologies and Open Compute Project (OCP) standards. HIGHER will adhere to the OCP Data Center Stack (DC-Stack) family of standards, aiming to provide datacentre-ready integrated systems for edge, private cloud, and large-scale data centres.

Following the OCP Data Center – Modular Hardware System (DC-MHS) specifications for modules compatible across servers, chassis and vendors, HIGHER will develop the following hardware modules:

1. Processor modules: Two OCP-compliant processor modules, based on the OCP Host Processor Module (HPM) or the OCP Universal Baseboard (UBB) standard, one hosting an Arm chip (SiPearl's Rhea2, derived from the European Processor Initiative project) and the other hosting a RISC-V chip (VEC, from the EUPILOT project).
2. Management module: A datacentre-ready secure control module (DC-SCM), hosting a RISC-V processor inside a field-programmable gate array (FPGA)-based board for server management, security, and control functionality.

The hardware modules developed in HIGHER will be integrated into the OCP-compliant HIGHER chassis (see image on the right), utilizing commercial off-the-shelf OCP-compliant modules for power distribution, network connectivity, and storage. The mechanical aspects of the server chassis will be developed within the project. The resulting platform will provide a modular hyperconverged infrastructure, offering distributed resources interconnected directly in hardware for efficient utilization and management. The modular infrastructure will be easily configurable, and by integrating different HIGHER processor modules we will be able to provide cloud systems incorporating multiple OCP sleds with high-end processors and accelerators, or smaller edge systems with weaker processors, based on required needs and the position of the server in the cloud-edge continuum.

HIGHER will provide Linux OS and driver support for the OCP compute modules (T4.1 and T4.2, for the Arm and RISC-V modules respectively) described above, and will create appropriate software sets for them. Moreover, it will provide firmware and system software support for the management module. The



OCP server chassis, with host processor modules and secure control module

management module will be built and expand OpenTitan, an open-source root-of-trust project. The management module will incorporate a RISC-V core together with hardware and software elements for secure measurements, signature checks, and key management. The system software stack advocated by HIGHER will also incorporate a meta-OS, i.e. middleware to facilitate efficient data processing, storage, and networking across a distributed computing continuum comprising internet-of-things (IoT), edge, and cloud platforms.

Evaluation of HIGHER platforms will be driven by use cases focusing on accelerated data analysis (incl. use of disaggregated memory) and cloud service models (IaaS, PaaS).

CONTACT: Manolis Marazakis, FORTH maraz@ics.forth.gr
Stelios Louloudakis, FORTH slouloudak@ics.forth.gr
PROJECT NAME: HIGHER: European Heterogeneous Cloud/Edge Infrastructures for Next Generation Hybrid Services
START / END DATE: 01/01/2025 – 31/12/2027
GRANT AGREEMENT: 101189612 / **DG / AGENCY:** HaDEA
KEY THEMES: cloud, edge, Arm, RISC-V, computing continuum, computing for servers, data centres, open-source software
PARTNERS: **Greece**: Foundation For Research And Technology – Hellas (FORTH - Coordinator), Exascale Performance Systems (EXAPSYS); Institute Of Communication & Computer Systems (ICCS) **Spain**: Semidynamics Technology Sevices S.L. (SMD); Barcelona Supercomputing Center (BSC); **France**: Sipearl SAS; 2CRSI; **Germany**: Extoll GMBH; **Sweden**: Kungliga Tekniska Hoegskolan (KTH); RISE Research Institutes Of Sweden AB (RISE); **Switzerland**: Cloudsigma AG.
BUDGET: €5,996,250
higher-project.eu

HOW SPECTRUM IS SHAPING THE FUTURE OF DIGITAL INFRASTRUCTURES FOR DATA-INTENSIVE SCIENCE

Large-scale scientific challenges are pushing the boundaries of what is currently possible with computing technologies. The amount of data gathered, shared and processed in frontier research is set to increase steeply in the coming decade, leading to unprecedented data processing, simulation and analysis needs.

The EU-funded SPECTRUM project brings together leading European science organizations and e-infrastructure providers to formulate a Strategic Research, Innovation, and Deployment Agenda (SRIDA) defining the vision, overall goals, main technical and non-technical priorities, and investment areas for next-generation scientific infrastructure. The project is developing a research, innovation and deployment roadmap for data-intensive science and infrastructure, set out in a technical blueprint for a European compute and data continuum. This collaborative effort is set to create an exabyte-scale research data federation and compute continuum, fostering data-intensive scientific collaborations across Europe.

SPECTRUM will enable the scaling and planning of the compute infrastructures needed for major experiments that will produce scientific advances over the next decade. Understanding the gaps in infrastructure provision is an opportunity to optimize advanced computing research to address the real needs of user communities, as well as to foster the adoption of future advanced computing technologies to maximize the impact of European scientific research.

The project is carrying out activities in a range of areas:

Community of practice

The SPECTRUM Community of Practice (SPECTRUMCoP) connects scientists and infrastructure managers working towards a mutual understanding of future needs, associated challenges and their possible solutions. The SPECTRUMCoP serves as a discussion forum across relevant scientific domains to identify common directions for the design and operations of future systems, processing models and solutions, and to provide feedback on investment to funding agencies and policy makers.

Use cases

SPECTRUM is identifying relevant use cases, related challenges and opportunities, relating to future compute needs in high-energy physics, radio astronomy and other domains. A compendium of use cases, including gaps and requirements related to technical and policy aspects, is due for publication in March 2025.

Landscape analysis

An analysis is being undertaken to better understand the e-infrastructure landscape and best practices. A compendium of existing approaches, services, technical solutions and policies for the federation of data and compute installations will be created.

Access policies

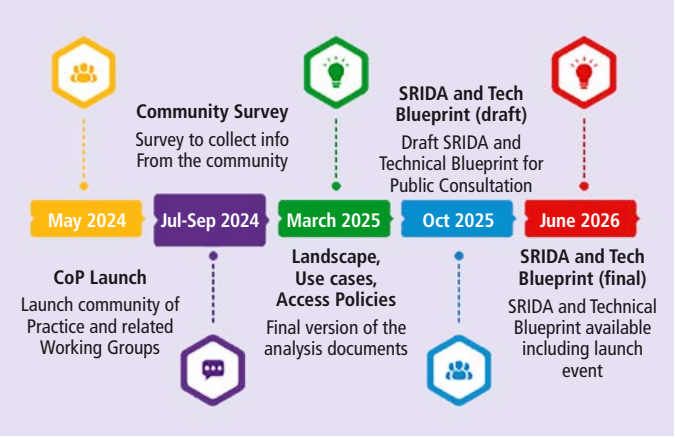
To increase collaborative e-infrastructure service delivery at the national, European and international levels, a set of recommendations will be compiled for a corpus of interoperable access policies.

Strategic action

SPECTRUM is working towards consensus on strategic action paths, specific actions and policy recommendations for next-generation e-infrastructure. This work will result in two major documents:

- a technical blueprint of a European compute and data continuum
- SRIDA: a strategy and plan for the implementation of exascale research data federation and compute continuum for data-intensive science

As part of this work, SPECTRUM and the JENA Computing Initiative are conducting a survey to understand current best practices and future needs in large-scale and data-intensive scientific computing. The survey is directed at researchers, managers of scientific initiatives and infrastructure managers.



GET INVOLVED:

Join the SPECTRUM CoP bit.ly/SPECTRUM_CoP
Complete the SPECTRUM survey bit.ly/SPECTRUM-JENA_Survey1
Follow SPECTRUM on LinkedIn bit.ly/SPECTRUM_LinkedIn
Subscribe to the project newsletter bit.ly/SPECTRUM_newsletter
Contact: Sergio Andreozzi / Xavier Salazar info@spectrumproject.eu

FACILITATING THE DEVELOPMENT OF DIGITAL TWIN APPLICATIONS ACROSS THE COMPUTE CONTINUUM WITH INTERTWIN

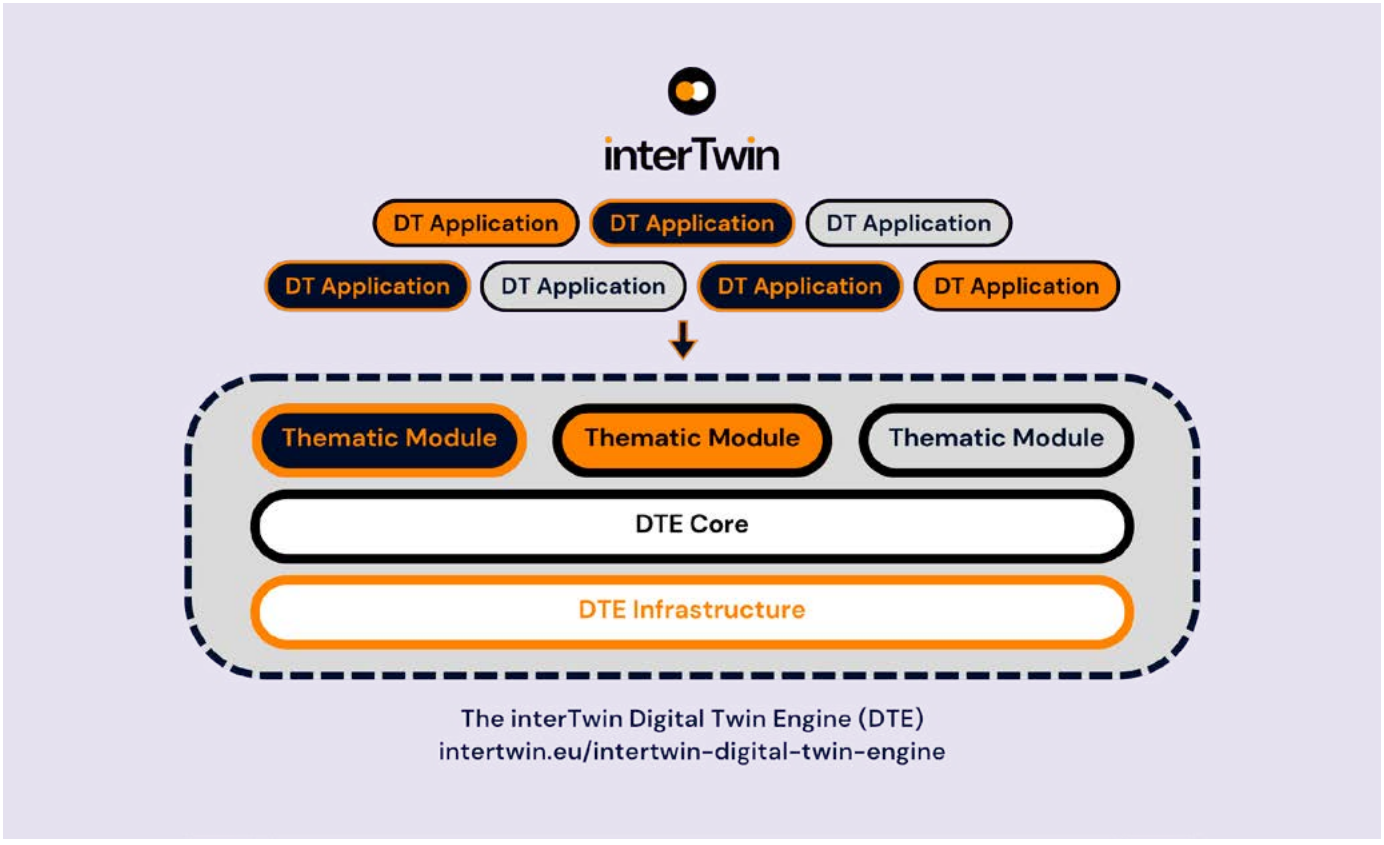
A digital twin (DT) is a virtual representation of a physical object, process, or system. It is created and sustained with information derived from one or many sources of data, such as sensors or models, taking into account both historical and real-time observations.

Digital twinning is a well-established market for industrial applications. Commercial adoption of DTs is in progress across various verticals, including manufacturing, healthcare, aerospace, automotive, smart cities, energy and utilities, and construction. DTs are also becoming increasingly useful in science because they allow researchers to simulate complex systems in changing environments and predict how they will behave under different conditions, while offering the ability to adapt depending on data gathered in real time. This can help researchers optimize their experiments, improve their understanding of complex phenomena, and develop more effective strategies for managing environmental systems.

On the infrastructure side, digital twinning requires an operational compute infrastructure able to operate across the whole compute continuum. While commercial initiatives provide end-to-end solutions – edge, IoT, cloud and high-performance computing (HPC) are well-established markets – within Europe there are many compute resources used for scientific purposes that need to be adapted to tackle the challenges of DTs. interTwin aims to promote and enhance the usage of those systems in order to tackle the technical challenges of DTs, ensuring that the investments in publicly available infrastructure are fully leveraged.

Interoperable, interdisciplinary platform for digital twins

interTwin is co-designing and implementing the prototype of the Digital Twin Engine (DTE): an open-source, interdisciplinary software platform that offers the capability to integrate with application-specific DTs. Its functional specifications and implementation are based on a co-designed interoperability framework and conceptual model of a DT for research, the DTE blueprint architecture. The ambition of interTwin is to create consensus on a common approach to the implementation of DTs that is applicable across the whole spectrum of scientific disciplines and that will facilitate interoperability across different DTs.



The interTwin Digital Twin Engine (DTE)
intertwin.eu/intertwin-digital-twin-engine

The main results from interTwin will be as follows:

Interdisciplinary Digital Twin Engine: A software platform that provides generic and tailored functional modules for modelling and simulation to facilitate the development and deployment of DTs that address scientific problems in different domains. The first release of the Digital Twin Engine (DTE) encompasses 39 components that are fully described here: intertwin.eu/intertwin-digital-twin-engine

Interoperability framework: guidelines, specifications, and blueprint architecture: The interTwin interoperability framework aligns technical approaches and fosters collaboration in modelling and simulation application development across scientific domains. The final version of the blueprint architecture and design specifications is available on Zenodo: zenodo.org/records/14034231

Toolkit for artificial intelligence (AI) workflow and method lifecycle management: AI-based methodologies to extract application sector-specific information from research data at the exabyte-scale level in real time and increase the efficiency and accuracy of simulation and modelling outputs. One of the results is the itwinAI component, a Python library that streamlines AI workflows while reducing coding complexity: intertwin.eu/article/core-dte-module-itwinai

Quality framework: Tools for automated quality measures and trust, development of standard quality mapping and indicators for appropriately communicating differences in the quality of inputs and outputs from digital twins, addressing issues such as data and model pedigree, accuracy, and lack of knowledge. The platform for quality assessment and awarding of multiple digital objects (source code, services, data (SQAaaS) has implemented a set of tools that enable automatic triggering of the service. The full platform is available here intertwin.eu/article/core-dte-module-sqaaas-platform

DTE federated infrastructure integrated with the European Open Science Cloud (EOSC) and EU Data Spaces: Federated distributed compute platform providing access to distributed data and integrating high-throughput computing (HTC), high-performance computing (HPC), cloud and quantum computing capabilities for processing. The DTE federated infrastructure has provided the necessary software solutions to:

1. enable resource provisioning on a wide range of compute providers to implement a digital continuum;
2. support data access, data management and to support real time data processing, in a federated environment; and
3. provide services and tools to enable the automated storage and compute resources orchestration enhancing the automatic and intelligent identification of the best provider either based on static and dynamic metrics.

One of the flagship components developed is interLink, an open-source service to enable transparent access to heterogeneous computing providers, from HPC to edge. It provides an abstraction for the execution of a Kubernetes pod on any remote resource capable of managing a container execution lifecycle: intertwin.eu/article/infrastructure-component-interlink

interTwin Open Source Community: The community of DT application developers, users and operators who are responsible for the design,

development and maintenance of the DTE code base. All components and modules are available via the interTwin open-source community repository on GitHub: github.com/interTwin-eu

The main users of the DT platform are as follows:

- DT developers interact with interTwin DTE, seen as a PaaS ('platform as a service'), developing DT applications and occasionally thematic modules tailored to the needs of specific user communities.
- DT infrastructure providers provide computational resources and storage, to build and run the DTs and eventually provide connectivity with the physical twin existing in the real world.
- DT users access the DTE as SaaS ('software as a service') via the DT applications developed by the DT developers. An end user can choose an 'out-of-the-box' DT application and connect it to its use case (physical twin) or configure the parameters required for their experiments.

The Digital Twin Engine modules developed have been prototyped by the implementation of several real-life use case in multidisciplinary domains covering environmental digital twins (e.g. climate change, extreme weather impact, etc) and high-energy physics, radio astronomy, and gravitational waves. The result of the feedback loop will be included in the second release of the DTE modules expected by April 2025.

FURTHER INFORMATION:

intertwin.eu
Linkedin.com/company/intertwin
youtube.com/@interTwin
github.com/interTwin-eu
Zenodo zenodo.org/communities/intertwin
Contact: Andrea Manzi / Xavier Salazar info@intertwin.eu

Now a tradition of the Design Automation and Test in Europe Conference (DATE), the Young People Programme is back for 2025 with more activities than ever. Sara Vinco (Politecnico di Torino) and Christian Pilato (Politecnico di Milano), two of the Young People Programme chairs, tell us more.

Seven reasons to take part in the DATE Young People Programme

The largest electronic design automation (EDA) conference in Europe, DATE will be held on 31 March – 2 April 2025 in Lyon, France. The DATE Young People Programme (YPP) features several events for young researchers to promote their skills and find their path in their future career.

1 Career Fair – Industry

During a company presentation session, PhD students, master's students and early career researchers can discover the latest trends, working environment, and open positions at the companies sponsoring DATE, helping them expand their network and find job opportunities. A speed-dating session will allow them to talk to HR representatives to ask questions, leave their CV or plan interviews.

2 Brainstorming career perspectives

Attend our panel on career perspectives to hear about various career paths to help young researchers get a clearer idea about possibilities, necessary skills, and typical job profiles either in the academic world or in the industry. Afterwards, you can have a coffee and share thoughts with mentors from academia or industry during the mentoring coffee breaks.

3 PhD Forum

The DATE PhD Forum is a great opportunity for PhD students to present their work to a broad audience from both industry and academia. The forum may also help students to establish contacts for entering the job market and for future activities.

4 Student Teams Fair

The Student Teams Fair brings together university student teams with sponsoring companies willing to provide support (such as free tool licences and financial support) for future activities.

5 Career Fair – Academia

It's not just industry representatives who are looking for top-talent graduates; academic institutions also need high-potential candidates. Academic programme leaders will promote new and upcoming academic research initiatives and open positions, also advertised on a "Jobs Wall".

6 University Fair

The DATE University Fair provides a platform to disseminate mature projects with live demonstrations and potentially applied to commercial products. As there are often follow-up projects, which need a new generation of researchers, this closes the loop with the Academic Careers Fair.

7 HackTheSilicon and SoC Labs

During HackTheSilicon, teams will receive a Google's OpenTitan SoC design with implanted security vulnerabilities, and will have to provide exploits, and propose mitigations. The SoC Labs event aims making SoC design more accessible even with no prior experience, through presentations of successful projects and an education section.

Sponsorship of attendance

To enable students to attend DATE, sponsoring companies and the IEEE Council on Electronic Design Automation (CEDA) will fund registration for the full conference for YPP participants. In addition to the numerous YPP events, participants will have the opportunity to attend the full DATE program.

FURTHER INFORMATION:

date-conference.com/call-for-papers#ypp-students

For further details about individual events, please contact:

Academic Careers Fair: ypp-academia@date-conference.com

Industry Careers Fair: ypp-industry@date-conference.com

PhD Forum: ypp-phd@date-conference.com

In this thesis presentation, Jude Christudas Haris explains how his research supports hardware-software co-design for efficient neural network inference at the edge.

Three-minute thesis

NAME: Jude Christudas Haris
RESEARCH CENTRE: University of Glasgow
SUPERVISOR: José Cano
THESIS TITLE: Hardware-Software Co-Design of FPGA-based Neural Network Accelerators for Edge Inference

The demand for efficient deep neural network (DNN) acceleration on resource-constrained devices has grown as edge-based artificial intelligence (edge AI) applications have become more prominent. Due to the slowdown of transistor scaling, homogeneous computing using general-purpose central processing units (CPUs) or graphics processing units (GPUs) either struggles to keep up with the computational demand of new AI-based applications or is power inefficient. Hence, there is a need for specialized hardware accelerators that can efficiently process the workload presented by DNNs. These accelerators can be purpose-built to address the computation required within DNNs, enabling better latency, throughput, power and area performance.

To tackle the design challenge of these accelerators, one approach involves using field-programmable gate arrays (FPGAs), which can be reprogrammable to enable different hardware architectures, allowing iterative design and test of new accelerators without the need for fabrication. However, current methods for designing and integrating new FPGA-based hardware accelerators with DNN frameworks (e.g., TensorFlow) are complex, error-prone and time-consuming.

Enabling efficient hardware-software co-design

The first contribution of this thesis is the SECDA methodology (SystemC Enabled Co-design of DNN Accelerator), which enables hardware-software co-design of hardware accelerators for DNN inference on edge FPGAs. SECDA provides a clear path in designing hardware accelerators for DNNs. Integrating the design process within the target DNN application framework enables the hardware designer to fully co-design and test not only the hardware components but also the software (e.g. host-driver code) and algorithmic components of the system from early in the development cycle.

To expand upon the SECDA methodology, we developed the toolkits SECDA-TFLite and SECDA-LLM to enable developers to adopt the SECDA design methodology within TensorFlow Lite and llama.cpp, two popular frameworks for DNN inference on edge devices.

Accelerating generative-AI workloads

Generative-AI (genAI) workloads present a unique challenge compared to traditional DNN models as they require 'up-sampling' to enable the generation of new data. The key operation within the genAI models is transposed convolution (TCONV). Using the SECDA methodology and the SECDA-TFLite toolkit, we designed and evaluated our specialized accelerator architecture, MM2IM, to target TCONV operations within genAI models. The design considers key observations of the computational patterns present within TCONVs to skip ineffectual computations.

Automated generation of host driver code

A key aspect of the optimal performance of a specialized accelerator is the host-side driver code, which acts as a bridge between the host CPU and the accelerator. However, developing the driver code is time-consuming and error-prone.

The final contribution of this thesis is AXI4MLIR, an extension to the MLIR compiler framework that enables efficient host-accelerator communication by automatically generating host driver code that is aware of the accelerator architecture and capable of performing efficient data transfers. AXI4MLIR can generate host-drive code based on simple JSON-like configurations, which summarize the properties of the new accelerator. With AXI4MLIR, we can target any linear algebra operation supported by our accelerator, represented in MLIR.

Jude's supervisor **José Cano** commented: 'Designing specialized accelerators for emerging workloads such as DNNs requires familiarity with and exploitation of different levels of the systems stack (i.e. algorithms, systems software, hardware).

Jude proposed a design methodology and associated toolkits to create specialized FPGA-based accelerators for different types of DNNs (convolutional neural networks, generative adversarial networks and large language models). The toolkits developed can help the community explore and develop new hardware-software co-designed accelerators for current and emerging workloads, especially on resource-constrained edge devices.'

Thanks to all our sponsors for making #HiPEAC25 such a success!

AMD arm doitnow Google HUAWEI

semidynamics SMILIES SERICS redpitaya ASML AXELERA AI

E4 COMPUTER ENGINEERING egi Hewlett Packard Enterprise MAXELER UK Research and Innovation ZeroPoint

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación dAIEDGE ECSPEC epi European Processor Initiative KRAI monetdb

Springer TASKING THALES ECLIPSE FOUNDATION

Sponsors correct at time of going to print. For the full list, see hipeac.net/2025/barcelona

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

Funded by the European Union

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