HiPEAC info 73

OCTOBER 2024

Register for HiPEAC 2025

On processing-in-memory momentum, with Onur Mutlu

HiPEAC does hardware, from design to system integration

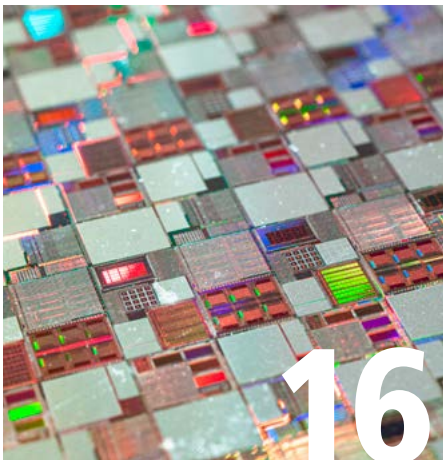
Semidynamics' Roger Espasa on inventing, ourselves

contents

How Chips JU is making EU chips great again

Discovering US: Pete Beckman and Rajesh Sankaran

Disrupting computer architecture with Onur Mutlu



16

How HiPEAC hardwares: Dispatches from the community

28

Frank K. Gürkaynak on expanding the ecosystem with open hardware

32

Roger Espasa on building Semidynamics

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

hipeac.net @hipeac / @hipeacjobs

hipeac.net/linkedin hipeac.net/tv

Funded by the European Union

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

Cover image: Kirill Marakov on Adobe Stock

Design: www.magelaan.be

Editor: Madeleine Gray

The theme of this magazine is hardware. Hardware has always been at the core of HiPEAC, and, for decades, the exponential growth in the number of transistors known as Moore's Law was driving hardware innovations. This fast innovation cycle created the need to regularly replace hardware, resulting in companies with lots of cash to invest in the development of the next generation of chips. This created a virtuous circle that benefitted the community a lot and was generally considered positive. The sky was the limit.

About a decade ago, this situation changed when Moore's Law slowed down, and scientists discovered that the increasingly elaborate means to chase performance also had a dark side, namely that every chip generation had a larger ecological footprint than the previous one because it needed more minerals, had more process steps, and the chips produced were more difficult to recycle. Computing now has an ecological footprint comparable to that of aviation (and it is still growing). From being part of a solution, it became part of a problem.

Then came deep learning that only seemed to be restricted by the amount of compute power available to train large models. These large models are now being integrated into common productivity tools, leading to an increase in compute demand that is faster than Moore's Law can keep up with. The result is that power consumption is currently closely following demand growth, and that hyperscalers are currently building gigawatt-class data centres. One gigawatt is equivalent to the average electricity consumption of around one million households and is comparable to the electricity consumed by the Hydro Aluminium plant in Sunndal, Norway – Europe's largest aluminium smelter.

Some hyperscalers have already announced that they will have to give up their commitments to become climate neutral by 2030, and they have started buying nuclear power plants to power their data centres to limit their carbon emissions. Experts warn that their power consumption might eventually lead to higher electricity prices, and inflation. Regulations and international geopolitical tensions do not help as they lead in some cases to the duplication of data centres, and hence even more electricity consumption. This situation is clearly not sustainable.

However, it does create an opportunity for the HiPEAC community: any technique we can come up with to reduce the embodied or operational emissions will contribute to reducing the environmental impact of computing. I hope that this magazine will inspire you to start doing research in this area.

Koen De Bosschere, HiPEAC coordinator

In 2023, the Chips Joint Undertaking (Chips JU) was launched with the aim of strengthening the European semiconductor industry and promoting innovation. Andreia Martins and Vasiliki Peglidou of the Chips JU communications team brief us on the Chips JU mission and forthcoming calls.

Chips JU: Strengthening the European semiconductor ecosystem

Europe faces significant challenges in the semiconductor industry, including reliance on external chip suppliers that are vital to key sectors such as mobility, healthcare, energy, and manufacturing. To mitigate these risks and foster innovation, Europe needs to develop a robust, local semiconductor ecosystem that can support its strategic goals for energy efficiency and digital sovereignty.

The Chips Joint Undertaking (Chips JU), established under the EU Chips Act in September 2023, plays a key role in strengthening research, development, and manufacturing capabilities in Europe. With nearly €11 billion in funding from the European Union (EU), 32 participating states (see below), and private members (AENEAS, EPoSS, and INSIDE industry associations), the Chips JU has launched several calls for pilot lines, design platforms, competence centres, and research and innovation projects (R+I) projects, including a joint call with the Republic of Korea. These activities aim to bridge the gap between innovation and commercial application, and advance semiconductor technologies, engaging startups, small and medium-sized enterprises (SMEs), universities, and larger companies.

Building a resilient industrial ecosystem

By leveraging programs such as Horizon Europe and the Digital Europe Programme, the Chips JU fosters collaboration from low to high technology readiness levels (TRLs) to advance semiconductor technologies across the semiconductor value chain. This effort is critical to strengthening Europe's autonomy in hardware and software development while securing a leadership role in global technology. To support this vision, the Chips JU has been facilitating collaboration among 1,181 beneficiaries across various R+I projects, including research institutions and SMEs, crucial drivers of digital innovation in Europe.

Chip manufacturers are developing hardware and software simultaneously, enhancing their capabilities in critical fields such as quantum computing, artificial intelligence, and cybersecurity. To support this transformation, the Chips JU is committed to developing a solid and independent semiconductor ecosystem, safeguarding the full technological stack essential for today's society.

Upcoming opportunities

The Chips JU has recently launched two calls on quantum technologies. The first call, QAC-1, seeks to enhance the production of quantum chips, while the second call, QAC-2, focuses on developing scalable infrastructure for chip-based ion-trap technologies in Europe. These calls are a critical step to build technological capacities and advance quantum computing hardware.

Europe's sovereignty in the semiconductor domain is essential not only for technological leadership but also for the digital transformation of our society. By investing in a holistic strategy, thus safeguarding the full technological stack, Europe will be better positioned to navigate the rapidly evolving geopolitical and economic landscape.

The Chips JU is a central mechanism to bridge the gap between cutting-edge research, technological innovation, and industrial production, thus fostering a robust semiconductor ecosystem and enhancing Europe's technological capabilities and resilience against supply chain disruptions.

For further information on open calls and deadlines, visit the Chips JU website.

FURTHER INFORMATION:

Chips JU website chips-ju.europa.eu

QAC-1 call chips-ju.europa.eu/Chips-2024-QAC-1

QAC-2 call chips-ju.europa.eu/Chips-2024-QAC-2

Chips JU participating states: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Israel Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Norway, Poland Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Türkiye, United Kingdom

Scaling new heights as ACACES turns 20

Every edition of ACACES offers a unique and memorable experience, but the twentieth birthday of HiPEAC's annual summer school, which this year was a joint effort between HiPEAC and the DISCOVER-US project, called for an extra special celebration.

Held, as is now tradition, in Fiuggi, Italy, the programme for this year's event was exceptional, featuring internationally renowned tutors including Onur Mutlu (ETH Zurich), Michael Scott (University of Rochester), Hyesoon Kim (Georgia Tech) and Louis-Noël Pouchet (Colorado State University).

Reflecting recent trends, machine learning (ML) and artificial intelligence (AI) dominated the course offering, particularly at the edge of the computing continuum, with courses provided by Omesh Tickoo (Intel), Rajesh Sankaran (Argonne National Laboratory) and Pete Beckman (Northwestern University), Giovanni Ansaloni (EPFL) and Alessio Burrello (University of Bologna). Meanwhile, Cristiana Bolchini (Politecnico di Milano) provided a comprehensive guide to delivering dependable computing systems, Paul Pietrangelo (Lira) set out the path to entrepreneurship and David Bol (UC Louvain) gave participants a thorough, gripping introduction to sustainability in computing.

The keynote speakers were equally high calibre. The opening talk was given by Kunle Olukotun (Stanford University), winner of the 2023 ACM Eckert-Mauchly Award, who briefed participants on computing systems in the foundation models era, drawing on examples from his highly successful companies. For the entrepreneurial keynote, HiPEAC was delighted to be joined by Fabrizio Del Maffeo in his native Italy to hear the inside story of how he launched and scaled the machine-learning acceleration company Axelera AI.

ACACES also provides an opportunity to reflect on the past, present and future of computing systems, and this year's consultations on the HiPEAC Vision's 'next computing paradigm' provided plenty of food for thought.

For many participants, it is also an opportunity to consider their career trajectory, and, as usual, the summer school included a series of career-related events. In addition to the HiPEAC Jobs wall featuring open positions all over Europe, the careers roundtable gave students the chance to discuss possibilities, challenges, skills and more with top researchers and innovators, while the inaugural 'Coffee and Papers' session allowed them to have an informal chat about scientific papers. Meanwhile, a session on the DISCOVER-US project allowed participants to learn how they could boost their CVs and gain invaluable international experience thanks to a transatlantic exchange. *(See p. 42 for more on HiPEAC Jobs activities.)*

To all the tutors, speakers and, most of all, students who made this summer school such a special occasion, HiPEAC would like to extend sincere thanks.

Videos from ACACES, including lecture videos of selected courses, will be made available on the HiPEAC YouTube channel
bit.ly/ACACES24_playlist

Cambridge Computer Architecture Research Centre launched

Addressing the grand challenges in computer architecture and semiconductors with industry support

Timothy Jones, Director, Computer Architecture Research Centre, University of Cambridge

There's a new opportunity for PhD students to work on some of the grand challenges in computer architecture and related fields at the Computer Architecture Research Centre we've just launched at the University of Cambridge.

Computer architecture is a critical area of computing, underpinning today's technologies and driving the next generation of computing systems. But it faces some key challenges. One is sustainability: we need to find ways to reduce the energy consumption and extend the lifecycle of chips without affecting performance. Then there's machine learning (ML) – both architectures for ML and the use of ML in hardware design – and the challenges of going beyond RISC.

But the key issue is people and ensuring a pipeline of future talent in this field. This is where our new research centre comes in.

The PhD students will be based in the Cambridge University Department of Computer Science and Technology, which has very broad areas of expertise ranging from compilers to circuits, through security, performance and reliability, and includes applications and the specialized compute to run them efficiently. We've found this broad approach has generated sharper insights and created very impressive students, as the awards they have won testify. (They include the BCS Distinguished Dissertation Award and the Apple Scholarship, given annually to only a handful of students worldwide.) Our postdocs and faculty have also been recognized by best paper awards and by 10-year, and even 25-year, 'most influential paper' awards.

We've invested further in our research capacity by hiring two new faculty members. Prakash Murali came to us (from Microsoft) to pursue his interests in co-designing quantum computing hardware, software and architecture to realize practical quantum computing. Tobias Grosser joined from the University of Edinburgh and focuses on compilation, programming-language design, and effective-performance programming. Our new research centre will draw on this breadth of supervisor experience across the department, allowing students the freedom to explore the areas they're most passionate about, while addressing industry-relevant research.

We're aiming to have six to eight students a year starting their PhDs. Every six months, we'll hold open events where the students will share their research with the wider world. All our research outputs will be open source. Though there'll be a strong academic focus to the centre, we also want input from industry partners on where they think the challenges lie to help guide the centre's direction. We'll be holding annual 'round tables' where we'll present our vision to partners and they can talk to us about the upcoming issues we should be exploring. This dialogue between academia and industry will help enrich – and improve – the research.

The new centre will be supported by donations. So we're looking to raise funding from industry partners and individuals to support the research leaders of tomorrow and collectively benefit all who work in computer architecture and semiconductors.

For more information, visit the webpage cited in 'Further information', or scan the QR code to see videos of the presentations by faculty and PhD students at our launch event.

timothy.jones@cl.cam.ac.uk

FURTHER INFORMATION:

'Creating a new Computer Architecture Research Centre', University of Cambridge Department of Computer Science and Technology, June 2024
bit.ly/UoC_CARC_launch

Videos from the Computer Architecture Research Centre launch
bit.ly/UoC_CARC_launch_playlist

RWTH Aachen spinoff Roofline enables agile adaptation to new edge AI models and hardware

Rainer Leupers, RWTH Aachen

With the field of artificial intelligence evolving at lightning speed, the agility to adapt to emerging models and disruptive hardware solutions is a significant competitive advantage. Traditional edge artificial intelligence (AI) deployment methods, mainly based on TensorFlow Lite, can't keep up with the pace. Low adaptability, limited performance, and a painful user experience make them barriers to adoption for edge AI.

RooflineAI GmbH, a spinoff from RWTH Aachen University, revolutionizes this process with a software development kit (SDK) that offers unmatched flexibility, top performance, and ease of use. Models can be imported from any AI framework such as TensorFlow, PyTorch, or ONNX. The one-stop AI backend enables deployment across diverse hardware, covering CPUs, MPUs, MCUs, GPUs, and dedicated AI hardware accelerators – all with just a single line of Python.

'Our retargetable AI compiler technology is building on the shoulders of proven technologies to create massive synergies with the open-source community and chip vendors. This approach provides the flexible software infrastructure required to overcome technology fragmentation in the edge AI space,' says Roofline Chief Executive Moritz Joseph, an affiliate member of HiPEAC and winner of a HiPEAC Technology Transfer Award. 'Roofline enables faster to-market times for new architectures, improves the efficiency of existing solutions, and brings the user experience to the next level.'

The compiler is key

AI compiler technology has become mission critical for deploying AI models at scale on edge devices. Today, widely used solutions build on a legacy software stack that interprets AI models but does not allow for compilation, partly relying on handwritten and manually optimized kernels. This limits the application of the technology for state-of-the-art AI models, such as language models, as they are not compatible with the existing technology stack.

AI compilation optimizes model execution at different levels of abstraction – known as 'intermediate representations' – that represent specific features of the execution of an AI workload. The compiler translates the AI model through various intermediate representations to a low level that is close to the target hardware.

Using AI compilation instead of model interpretation allows users to adapt to the constant stream of new AI models. In addition, novel heterogeneous hardware platforms can be targeted as the compiler generates code for each component.

The AI compiler space is driven by open-source technologies that allow the exploitation of synergies beyond the scope of individual players in the market. Roofline is active in the open-source community and committed to providing code that drives innovation from cutting-edge AI models to novel chips.

Roofline is the latest in a series of spinoffs from the Institute of Communication Technologies and Embedded Systems (ICE) at RWTH Aachen. Other examples include Silexica, which was acquired by Xilinx (now AMD) in 2021, and MachineWare, launched in 2022.

Special Roofline licence conditions are available for academic research and teaching purposes. For product demos, contact Moritz Joseph joseph@roofline.ai

FURTHER READING:

Roofline website roofline.ai

linkedin.com/company/rooflineai

Roofline founders (left to right) Maximilian Bartel, Moritz Joseph and Thomas Zimmermann

Funded by the European Union (EU), the DISCOVER-US project is building a research network between researchers in the EU and United States (US) to pioneer new concepts in distributed computing and swarm intelligence. In this interview, the first in the 'Discovering US' series, we caught up with DISCOVER-US members Pete Beckman (Northwestern University) and Rajesh Sankaran (Argonne National Laboratory) to learn about their work and why they are part of the DISCOVER-US community.

Discovering US: Pete Beckman and Rajesh Sankaran

DISCOVER US

What led you to specialize in your research area?

Pete: I've always been interested in fast and big computers. In the first part of my career, new supercomputer architectures and software layers were being invented each year – from new parallel languages to new massively parallel systems. Each year the systems grew faster, and I wanted to be part of that new frontier. However, as I watched instruments and sensors become ubiquitous, providing massive data streams that needed analysis, my interest turned toward taking what I had learned in high-performance computing (HPC) and shrinking it down – joining it with the instruments to build 'edge computing'. In many ways, my career now is focused on HPC embedded with instruments.

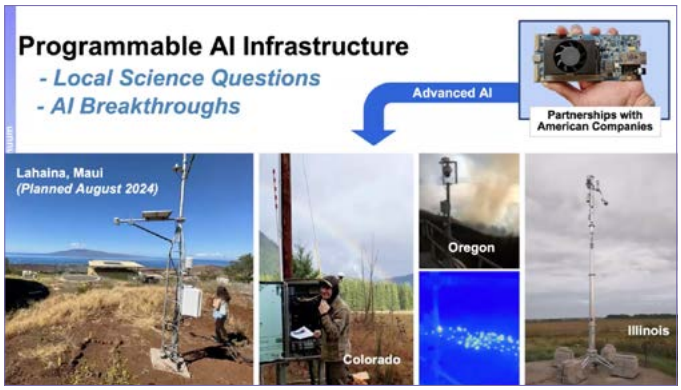
Raj: My interest in digital electronics started at a young age. I inherited a few project books for kids from my cousin and eagerly experimented with various projects, often using whatever components I could find. I was driven by a strong curiosity to build and bring ideas to life. During graduate school, I developed a modular electronics toolkit for creating physical user interfaces using small microprocessors. Later, when faced with the challenge of integrating advanced sensors into real-time workflows, my background in embedded systems and systems science became invaluable. This marked the beginning of my journey into embedding computing along with sensors and instruments in the field and linking them to large central computing infrastructures.

What projects are you currently working on?

Pete: Our team is deeply involved in various projects aimed at advancing edge computing and artificial intelligence / machine learning (AI/ML) at the edge. The challenges include resource sharing at the edge, multi-tenancy, edge-cloud workflows, and programming models for programming the edge-to-HP continuum. We also leverage existing AI/ML methods, including generative AI, to tackle edge inference challenges in practical applications, ranging from understanding urban environments to extreme event studies and security.

Initiated in 2019, the Sage project aims to develop a national-scale cyberinfrastructure tailored for AI at the edge. It is led by the Northwestern-Argonne Institute for Science and Engineering (NAISE) and funded by the National Science Foundation (NSF). The project unites experts from several institutions, including Northwestern University, the University of Chicago, George Mason University, UC San Diego, the University of Illinois in Chicago, the University of Utah, and partners at Argonne.

Raj: Sage empowers researchers by providing adaptable, internet-connected intelligent nodes equipped with various sensors – such as cameras, microphones, and environmental monitors – that can



The Sage infrastructure couples devices at the edge with HPC compute power

be tailored to specific research needs through software. It also integrates with centralized cloud and HPC resources for advanced, global observations, inferences, automated workflows, and AI model development. The Sage nodes employ edge computing, enabling real-time data analysis right where it is gathered. This approach addresses the limitations of traditional sensor networks, which often struggle to manage vast data streams or rely on slow data transfers to the cloud for processing. With Sage, scientists can promptly analyse data on-site, opening up new possibilities for real-time monitoring and measurement of various urban and environmental factors.

What technical challenges have you encountered in building this infrastructure?

Raj: First, the advanced AI systems we are developing rely on electronic components that were not specifically designed for harsh, remote, outdoor environments. For instance, the central NVIDIA graphics processing unit (GPU) is intended for stable office environments with consistent electrical power, rather than the intense heat of Texas. Consequently, every design phase required meticulous considerations and rigorous testing from the outset.

Second, once deployed, these systems often cannot be physically accessed for maintenance or repairs. Similar to satellites in orbit, all diagnostics, updates, and repairs must be conducted remotely. Thus, ensuring system redundancy is crucial to prevent a single failure from compromising the entire system.

Additionally, the extreme cybersecurity risks associated with sensors and actuators, such as moving cameras, necessitate the implementation of extensive safeguards. These measures address a range of risks not typically encountered in conventional computing infrastructure.

Pete: Another challenge is multi-tenancy, where edge-computing resources are shared among various applications competing for these resources. It is essential to ensure that all applications maintain a minimum acceptable quality of output despite the rapidly changing environment. Weather conditions and the phenomena under study affect the availability of resources at the edge and the computational and memory needs of AI/ML applications.

Furthermore, we required a method to schedule jobs on the nodes remotely, allowing them to continue executing tasks despite communication loss. This involves making decisions about job scheduling based on available resources and context, all while striving to meet data and inference goals for the scientific objectives of the scheduled jobs. Each node may pursue multiple scientific objectives simultaneously.

Finally, the challenge of managing data security – at the edge, in transit to cloud and HPC resources, and within trusted computing enclaves operating on secure data – has become increasingly critical.

Why is international research collaboration important?

Pete: International collaboration plays a crucial role in driving research and innovation forward. Programmes such as DISCOVER-US allow EU researchers to spend several weeks in the US, working closely with host research groups. Such international experiences provide valuable exposure to diverse research cultures, methodologies, and perspectives, which enhances creativity and adaptability. They also support personal development and offer a broader view of global issues.

The benefits of these collaborations are several, including access to a broader range of expertise and perspectives, shared resources, and the mobility of students and faculty. This fosters a vibrant, multicultural research environment and encourages the cross-pollination of ideas, leading to new solutions and approaches. Additionally, these collaborations facilitate the transfer of knowledge between teams. When researchers return to their home institutions, they can continue their collaborations, leveraging the progress made during the visit to achieve breakthroughs that might be challenging for either group in isolation.

Why did you join DISCOVER-US? How can it support EU and US researchers?

Raj: Joining the DISCOVER-US community was a natural extension of our established connections with scientists in European HPC centres. With AI now increasingly accessible on edge devices, these researchers and their students can explore innovative distributed systems, driving progress in both the EU and the US.

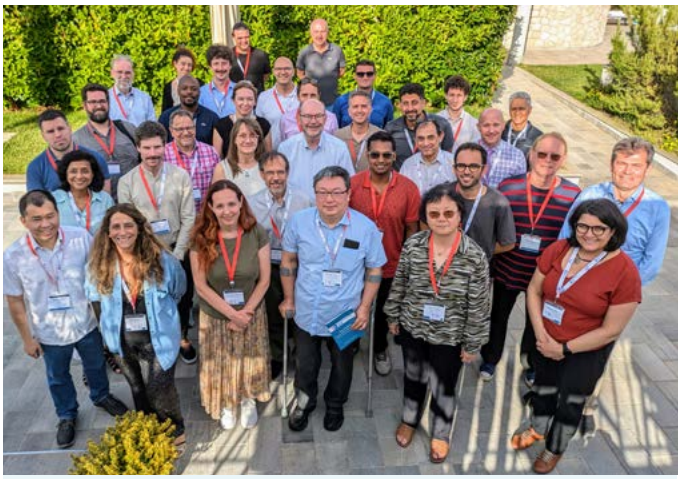
Pete: Our efforts with Sage and related projects are highly interdisciplinary, involving collaborations with researchers across various fields. This diverse research landscape offers ample opportunities to address complex problems that are not only theoretically intriguing but also hold the potential to solve significant challenges within application domains. There are numerous prospects for joint publications and collaborative grant proposals. Additionally, the NSF's commitment to funding exchanges between the US and the EU provides a vital pathway to nurture and sustain these collaborative efforts.

FURTHER INFORMATION:

DISCOVER-US website discover-us.eu

DISCOVER-US webinar:
'Sage: A software-defined sensor network'
youtu.be/-vxul94TNlo

Sage website sagecontinuum.org



DISCOVER-US shapes vision of AI and distributed computing at transatlantic gathering

On 24-26 June, the first DISCOVER-US vision workshop was held in Sabaudia, Italy. Bringing together top scientists from the United States (US) and European Union (EU) in fields such as distributed computing, the compute continuum, swarm intelligence and edge AI, the workshop provided a space for brainstorming future concepts, identifying challenges and defining priorities in these key areas.

Through a dynamic, varied programme of invited talks, breakout sessions, fishbowl sessions and plenary gatherings, the workshop explored key themes across the computing continuum. Following talks setting the policy context from Rolf Riemenschneider (European Commission) and Jason O. Hallstrom (National Science Foundation), dedicated sessions focused on computing paradigms, managing complexity through abstraction, emerging concepts and challenges, self-organizing and collective systems, and programming frameworks. The event concluded with an analysis of the respective strengths, weaknesses, opportunities and threats (SWOT) of the EU and US, and a distillation of research priorities in this field.

Thanks to contributions from the excellent researchers who took part, participants learned about inspirational projects on both sides of the Atlantic, including edge-to-cloud-to-HPC scientific infrastructure, innovative agricultural applications, and future immersive environments.

A white paper synthesizing the results will be made available on the DISCOVER-US website in due course. The DISCOVER-US team would like to thank all the participants for their valued contributions, and look forward to working on future concepts with them.

discover-us.eu

Putting the intelligence into HPC: PUMPS+AI ACM Europe Summer School

Building on Europe's strengths in high-performance computing (HPC), the Programming and Tuning Massively Parallel Systems + Artificial Intelligence Summer School (PUMPS+AI) empowers attendees to harness HPC for compute-intensive applications such as artificial intelligence (AI), an application area which has taken the computing systems community by storm.

With demand for graphics processing units (GPUs) reaching an all-time high, the 14th edition of PUMPS+AI, an official ACM Europe Seasonal School since 2023, was a particularly timely opportunity to get up close and personal with GPUs and other accelerators.

'AI is currently the dominant trend in computing, and GPUs are currently the star AI processors,' said local organizer and HiPEAC member Antonio J. Peña. 'PUMPS is the summer school for all those looking to master the dynamic, fast-growing area of HPC-powered AI.'

Taking place on 8-12 July 2024 at the Universitat Politècnica de Catalunya-Barcelona Tech (UPC), the summer school aimed to enhance the skills of researchers, graduate skills and instructors. It introduced cutting-edge techniques and offered hands-on experience in developing applications for manycore processors with massively parallel computing resources like GPU accelerators, with a focus on AI.

Featuring lectures by NVIDIA luminaries Wen-mei Hwu, fresh from being awarded the ACM Eckert-Mauchly Award, and Juan Gómez-Luna, the school also featured courses by Izzat el Hajj (American University of Beirut) as well as local lecturers Antonio J. Peña, Marc Jorda, Leonidas Kosmidis, Xavier Martorell and Xavier Teruel. The summer school is co-directed by HiPEAC co-founder and first coordinator Mateo Valero, the director of Barcelona Computing Center (BSC). As is now tradition, this year's edition featured a HiPEAC Jobs careers session coordinated by HiPEAC Jobs' Laura Menéndez Gorina (BSC).

PUMPS+AI will return to Barcelona in 2025. Check out the website for further information: pumps.bsc.es

Second RISC-V Summit Europe demonstrates increasing momentum around the standard

Teresa Cervero (Barcelona Supercomputing Center) and Christian Fabre (CEA)

The RISC-V Summit Europe is the premier event that connects the European movers and shakers – from industry, government, research, academia and ecosystem support – that are building the future of innovation in RISC-V. The event is designed to help attendees to explore, learn and share commercial and research applications and solutions.

Following the first RISC-V Summit Europe on 5-9 June 2023 in Barcelona, the community has once again demonstrated its engagement with RISC-V with the resounding success of the second edition, on 24-28 June 2024 in Munich. The core event (from Tuesday to Thursday) received more than 720 attendees (a 30% increase from the previous year) and 40 sponsors (+73%), offering more than 24 technical talks, 11 keynotes and invited talks, and three panels. In addition, the expo area gave delegates the opportunity to explore more solutions by visiting booths, interacting with more than 130 posters (+12% compared to Barcelona), attending demo or university theatre talks.

The day before the core event, 330 people gathered together, either to contribute to the RISC-V Technical Working Groups (TWG) or to attend RISC-V tutorials. Another successful initiative was the hackathon, in which undergraduate students and / or early career researchers were challenged to solve problems related to RISC-V. In general, these Monday activities were a mechanism to strengthen the community's skills, promote engagement and help RISC-V adoption for newcomers.

RISC-V has matured and evolved from 2023 to 2024 by moving from demonstrators to real products, ratifying several specs and profiles, and gaining traction in new markets. This was palpable in the 'Development Area' included in this year's event, where many RISC-V devices were on display.

In line with global trends, artificial intelligence (AI) was a hot topic in Munich. The industry seems to be rallying around two main platforms for AI: NVIDIA, the leader in the cloud, and a vivid set of RISC-V based accelerators for various market segments and applications. The goal is to solve more and more complex problems faster and more efficiently. Technically, the interest in AI for RISC-V focuses on executing matrices in a more efficient way, and that has prompted vendors to work together on a specification, which is still under discussion.

Another hot topic worth mentioning is automotive, which is experiencing considerable momentum thanks both to public investment by the European Commission, but also to private investment led by Infineon, Quintauris and Bosch. While RISC-V is having a considerable impact in the embedded domain, it is also gaining traction in other domains, from the internet of things (IoT) to high-performance computing (HPC). A clear example of this is that now there are the first real full-scale RISC-V microcontroller units (MCUs) available, such as those produced by Renesas.

It is important to remember that RISC-V is a community driven standard, which means that nurturing the open-source ecosystem around RISC-V is critical to continue pushing innovation forward. As Frank G. Gürkaynak said in his talk at the summit, 'there is no European / American / Chinese open source, but there can be European / Chinese / American support for open source'.

We thank all those who helped make the RISC-V Summit Europe in Munich such a success, and we expect the next RISC-V Summit Europe to be equally exciting.

Videos, abstracts and slides from the June 2024 Summit Europe, together with the upcoming announcement for the 2025 RISC-V Summit Europe 2025, can be found on the RISC-V Europe Summit website: riscv-europe.org/summit/2024

Wen-mei Hwu receives ACM-IEEE CS Eckert-Mauchly Award

In June, it was announced that HiPEAC associate member Wen-mei W. Hwu, a senior distinguished research scientist at NVIDIA and professor emeritus at the University of Illinois, Urbana-Champaign, was the recipient of the ACM-IEEE CS Eckert-Mauchly Award. Professor Hwu was selected for 'pioneering and foundational contributions to the design and adoption of multiple generations of processor architectures', with ACM noting that 'his fundamental and pioneering contributions have had a broad impact on three generations of processor architectures: superscalar, VLIW, and throughput-oriented manycore processors (GPUs)'.

One of the original architects of the high-performance substrate (HPS) model that pioneered superscalar microarchitecture, several papers co-authored by Professor Hwu have received 'test of time' and/or 'most influential paper' awards. He is the recipient of multiple awards, including the IEEE Computer Society B.R. Rau Award, the IEEE Computer Society Charles Babbage Award, the ACM Grace Murray Hopper Award, and the ACM SIGARCH Maurice Wilkes Award.

On behalf of HiPEAC, congratulations!

Computing community mourns MIT professor Arvind

In June it was announced that the MIT professor Arvind Mithal had died aged 77. Known by his mononym, Arvind, who had worked at MIT for nearly 50 years and who was a renowned computer scientist, made significant contributions to dataflow computing. In later years he worked on tools for formal modelling, high-level synthesis, and formal verification of devices such as microprocessors and hardware accelerators, as well as memory models and cache-coherence protocols for parallel computing architectures and programming languages.

Arvind was a teacher at the 2013 HiPEAC summer school, ACACES, and was well known by many in the HiPEAC community. Our thoughts are with his wife and family.

Photo credit: M. Scott Brauer

Dates for your diary

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

ISC High Performance
Workshop deadlines: 21 October and 11 December 2024
10-13 June 2025, Hamburg, Germany
isc-hpc.com

OCX 2024: Open Community Experience
22-24 October 2024, Mainz, Germany
ocxconf.org/event/2024

NorCAS 2024: IEEE Nordic Circuits and Systems Conference
29-30 October 2024, Lund, Sweden
events.tuni.fi/norcas2024
Further information: Jari Nurmi, Tampere University
norcas@tuni.fi

LOCO 2024: 1st International Workshop on Low Carbon Computing
3 December 2024, Glasgow, UK
locos.codeberg.page/loco2024

EFECS 2024
5-6 December 2024, Ghent, Belgium
efecs.adriacongrex.it

HiPEAC 2025: High Performance, Edge And Cloud computing
20-22 January 2025, Barcelona, Spain
hipeac.net/2025/barcelona
Sponsorship opportunities available
sponsorship@hipeac.net.

An idea whose time has come

Onur Mutlu on processing in memory, holistic architecture design and fundamentally better computing systems

In 2018, HiPEAC spoke to Onur Mutlu (ETH Zurich) about the role of memory in computing systems ['"It's the memory, stupid": A conversation with Onur Mutlu', *HiPEACinfo* 55, 2018, pp. 13-15] just after Onur's course at the HiPEAC summer school, ACACES. Six years later, fresh from a new edition of ACACES, we caught up with Onur to find out what's changed in the field of processing in memory, how disruptive hardware concepts get taken up, promising directions in computer architecture, and more.

How much has changed in this field over the last few years?

A lot, and in a very positive direction. When I first taught at ACACES in 2013, I described our initial ideas on processing in memory (PIM), including RowClone, a simple and effective mechanism for in-memory bulk data copy and initialization. However, at that time, we did not even have our first works published – some had been submitted but were rejected as being unrealistic or niche.

By the time I returned to teach at ACACES in 2018, we had a relatively large body of published work on the topic, so my course included much more material on processing in memory. Our works generated a lot of excitement, but there were still no industrial products or prototypes and there was still large scepticism, even among academics. For example, our Tesseract paper, published at ISCA 2015, was rejected twice before being accepted, while our Ambit work, published at MICRO 2017, was rejected four times before finally being accepted. Both works led to significant follow-on research after publication and have had a lasting influence; in fact, the Tesseract paper was selected for inclusion in the *ISCA@50 25-Year Retrospective: 1996-2020*.

Six years later, in 2024, we have a completely different landscape. Processing in memory was an even bigger part of my ACACES 2024 course. My lectures covered various aspects of the field, from the near-term 'processing-near-memory' approaches, which place conventional logic elements near memory structures, to the longer-term 'processing-using-memory' approaches, which fundamentally exploit the analogue operational principles of memory structures for logic computation.

As for industry, with Google we have demonstrated the benefits of processing-near-memory for mobile workloads and machine-learning accelerators, while with NVIDIA we showed the benefits of PIM on graphics processing unit (GPU) workloads. There is now at least one commercial product (from a European company, UPMEM) that places a general-purpose programmable multithreaded processor next to each bank in a dynamic random-access memory (DRAM) chip. We experimentally evaluated the benefits and trade-offs of such a PIM product in both my ACACES course and our research papers. Meanwhile, the best paper at HPCA 2024 was based on the UPMEM chip.

One thing is clear: having such hardware has enabled significant progress in the research community by enabling software development, benchmarking, and much better understanding.

There are prototype DRAM chips from multiple large companies with similar near-bank processing capabilities specialized for accelerating machine-learning workloads, e.g. the Samsung FIMDRAM, SK Hynix AiM, and Alibaba's prototype. There are also other prototypes that perform processing near DRAM chips, such as the AxDIMM module from Samsung and Meta. Many startups are also trying to create systems that perform some sort of PIM.

Furthermore, in work published in 2024, we were able to show that commodity DRAM chips, without any modification to the chip itself, just by modifying the memory controller, are able to execute bulk-bitwise operations in a reasonably robust manner. These fascinating findings demonstrate the fundamental computation capability of DRAM, even when DRAM chips are not designed for this purpose.

What does it take for a disruptive hardware concept to be taken up?
Demonstrating results showing that PIM hardware provides large improvements in both performance and energy efficiency helps industry to pay attention. If workloads where the benefits are high are critical workloads to users, the attention level increases.

Disruptive hardware should address a serious need in industry, and the risk involved in changing paradigm should be compensated by significant benefits. For PIM, this is increasingly the case: data access and movement are causing an increasingly worse bottleneck for essentially all metrics we care about (performance, energy, scalability, sustainability, robustness), applications are increasingly bottlenecked by data movement and access, and we are extremely power and energy constrained in our systems (see cutting-edge machine-learning / artificial-intelligence (ML/AI) workloads and systems, for example). At the same time, memory technology is not getting better.

Yet it is not enough to introduce disruptive hardware, even if it is outstanding; we also need to address barriers to adoption. A large software stack needs to be built to make sure the hardware is usable, efficient, and effective – and that is where most of the work needs to be done to make a disruptive hardware concept successful. We therefore focus a lot on software programming, tools, methodologies and compilers in our research. We have multiple works in this area that make life easier for the programmer, by enabling them to decide what to execute where (e.g. PIM-enabled instructions and DAMOV), compile easily into PIM hardware (e.g. SIMDRAM and MIMDRAM), and program using high-level frameworks that enable better productivity and better automated optimizations (e.g. SimplePIM and Dappa).

What are the main memory technologies to keep an eye on?
It is unclear when we will find a new technology that could replace DRAM for main memory and NAND flash for storage – bear in mind that it took at least three decades for NAND flash memory to become widespread.

In my view, DRAM and flash memory will continue to be very strong in the general-purpose domain, despite all the scaling challenges that are plaguing them today (e.g. RowHammer and RowPress) – challenges that would be better handled with more system-memory co-design. For example, industry is finally moving to having intelligence (i.e. logic) in DRAM chips and memory controllers to avoid RowHammer and RowPress bitflips. This is a step in the right direction, but we should also be rethinking the rigid interfaces we have to memory chips today, which give them little breathing room to perform management or computation functions (see, for example, our upcoming work on self-managing DRAM).

For future DRAM, emerging 3D and ferroelectric technologies are promising. While true 3D DRAM may take some time, 3D stacking of DRAM and logic, with a high-quality (i.e. logic process) logic layer, will happen in the shorter term. This will enable much better processing-near-memory.

What do you think are promising directions in computer architecture more generally?
I think we should freely explore creative ideas that have high potential to enable fundamentally better – i.e. more efficient, higher performance, robust, and sustainable – computing systems. Interdisciplinary research to enable systems for bioinformatics, algorithm-hardware co-design, and research that optimizes workloads (e.g. graph analytics, AI, genomics) across the stack, all the way from algorithms to devices, is very valuable. These directions require a more holistic way of thinking about the computing stack and co-designing across the transformation hierarchy, which I call the 'broader view of computing architecture'. The HiPEAC Vision touches on this nicely by calling for 'global co-design'.

One major direction is designing algorithms and systems for biological sequence analysis and information processing (e.g. for genomics). This is going to be even more important in the future as fast, efficient analysis of data is critical for many medical, public health, and personalized-medicine use cases.

Examples from my group include in-storage computing accelerators for genome filtering and in-storage computing for metagenomics. In-storage computing and specialized accelerators placed near flash chips greatly improve the performance and efficiency of many genome analytics workloads by reducing the data movement bottleneck from the storage system and specializing the computation to the primitives needed by genome-analysis workloads.

Processing in memory is another promising direction to accelerate such data-intensive workloads with algorithm-architecture co-design and co-optimization. We are also examining an exciting new paradigm for genome analysis, raw-signal analysis, that operates directly on electrical signals generated by modern nanopore sequencing devices, without requiring the translation of such signals to the genomic alphabet – a very costly process in modern systems.

Another exciting direction is architectural controllers, such as memory controllers, prefetchers, memory and thread-management mechanisms, that are designed such that their policies are not dictated by humans but are data driven, based on observed data in the field via machine-learning techniques. A data-driven architecture enables the machine itself to learn the (best) policies for managing itself and executing programs. Prime examples are reinforcement-learning-based, self-optimizing memory controllers and prefetchers.

We believe an intelligent architecture will consist of a collection of such intelligent controllers that perform automatic data-driven online policy learning, including learning how to best coordinate with each other to make decisions that benefit the overall system. Such machines learn the best policies over time and thus become better as they learn, adapting, evolving, and executing far-sighted policies.

To enable such a machine, we need to revisit the design of all controllers (e.g. caching, prefetching, storage, memory, interconnect) and turn them into data-driven agents. Some example works in this area apply data-driven design principles to memory controllers, prefetchers, memory hierarchy management policies, and hybrid storage systems.

A third direction that is critically important is building fundamentally robust (i.e. safe, reliable, and secure) systems, particularly since computing infrastructure is increasingly used in all scenarios that affect human life today. Robustness should become a goal from the beginning of the design to the end of the lifetime of a computing system. We have a long way to go to achieve this goal and technology scaling does not help us: denser

chips have more robustness problems (e.g. the RowHammer and RowPress phenomena in modern DRAM chips used in essentially all computers), which can be exploited for security attacks or which can manifest themselves as safety problems (think bitflips in self-driving cars, planes and spacecraft).

A final direction I would highlight is architecture-technology co-optimization. Rethinking how modern architectures should be designed for emerging technologies, like tightly integrated packaging and interconnection techniques (e.g. 3D stacking of logic and DRAM, monolithic 3D stacking, 3D DRAM / NVM, promising and unconventional emerging memory technologies) is an exciting and important direction.

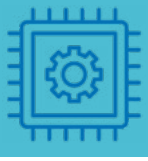
What about other considerations, such as sustainability?

As our systems are not designed to be energy efficient and sustainable from the get-go, we are recklessly building huge data centres, mainly to try to satisfy a particular type of growing ML models, that waste the world's most important energy and power resources. Fundamentally rethinking how we can make the hardware extremely efficient and sustainable is necessary.

With any fundamental rethinking, we should explore various options, including the extremes. A good extreme option for researchers to explore is to take a clean-slate approach and ask how one would design everything in a computing system to be energy efficient, sustainable and high performance. I believe processing everywhere, including in memory and storage, is the right choice if one thinks this way.

A longer version of this interview and all research papers referenced are available on the HiPEAC website:
bit.ly/Onur_Mutlu_interview_2024

Videos of Onur Mutlu's ACACES24 course are available on the HiPEAC YouTube channel: bit.ly/ACACES24_playlist



With data-intensive applications – such as those powering artificial intelligence – exposing the limitations of the von Neumann architecture, in- and near-memory computing paradigms are becoming an increasingly attractive alternative. However, to become a viable prospect, these paradigms need to overcome a series of challenges, among them the lack of programming support. In this article, Asif Ali Khan and Jeronimo Castrillon (TU Dresden) explain how they have tackled the programmability issue.

Memory as the new CPU

Programming paradigms for in-memory computing

The recent surge in data-driven applications has exposed the limits of conventional computing systems and highlighted their inefficiencies. Data movement in von Neumann computing systems has been shown to consume more than 60% of the total system energy (see Boroumand et al in 'Further reading', below). As a result, a series of domain-specific and general-purpose, fixed-function and emerging architectures have been proposed, including compute-near-memory (CNM) and compute-in-memory (CIM) architectures. These systems have demonstrated orders-of-magnitude improvements in performance and energy efficiency compared to traditional accelerators.

In CNM systems, specialized complementary metal-oxide-semiconductor (CMOS) logic is integrated onto or near the memory chips, reducing but not eliminating data movement. CIM systems, in turn, perform computations directly within the memory devices. Key categories of CIM systems include:

- crossbars for analogue dot-product operations, used in machine learning, for example;
- content-addressable memories (CAMs) for search operations, and
- CIM accelerators for bulk-bitwise logic and arithmetic operations

CIM systems have been demonstrated on both traditional dynamic random-access memory (DRAM) / static random-access memory (SRAM) and emerging nonvolatile memory (NVM) technologies, with some even commercially available today. This includes implementations in resistive NVM technologies and more recently in SRAM, magnetic random-access memory (MRAM), and ferroelectric field-effect transistor (FeFET). The CIM / CNM computing market is rapidly growing, with a value of US$ 15.5 billion in 2022 and a projected compound annual growth rate (CAGR) of 17.5% over the next decade.

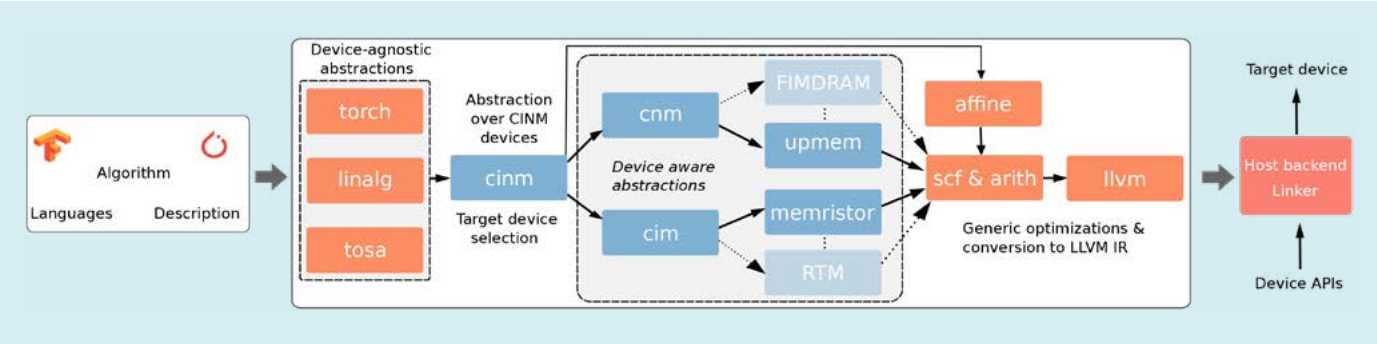
At the Chair for Compiler Construction TU Dresden, we have been working with CIM and CNM systems for over half a decade at various layers in the stack, in collaboration with researchers from academia and industry around the world. Recently, we conducted a comprehensive survey on the development of these systems, both academically and commercially, both to position our research and to identify critical gaps and challenges for future research (see Khan et al in 'Further reading').

One such key challenge is the lack of programming support for CIM / CNM systems. While sharing aspects of the general programmability challenge in heterogeneous computing, CIM and CNM computing requires a rethink of the traditional von Neumann computing paradigm and its well-established programming abstractions. Most existing solutions offer low-level device libraries, leaving the tasks of pattern matching, operator mapping, synchronization, and optimization to the programmer.

Compilation frameworks for CIM systems

At TU Dresden, we have been studying these challenges for different technologies and have developed specialized compilers for different CIM systems, details of which can be found in 'Further reading', below. For analogue crossbars, we developed an MLIR-based compiler known as the Open CIM Compiler (OCC), which abstracts hardware details from the programmer and generates optimized code considering the fundamental properties of the memory devices, such as slow-write operations and fixed-memory array dimensions. OCC is one of the first high-level compilers for CIM crossbars and is openly available.

For CAM accelerators, we proposed a compilation framework called C4CAM, which identifies CAM-amendable search patterns in high-level input applications and rewrites operations when necessary to make them suitable for CAM offloading. C4CAM also facilitates rapid design space exploration by inputting system specifications and optimization objectives along with the high-level application. For logic-CIM accelerators, we have



The CINM compiler. The abstraction lowers from left to right

developed compiler frameworks that generate efficient codes by taking device properties into account and optimizing for them.

We have an ongoing effort to generalize from our previous research to develop a single framework capable of targeting heterogeneous targets. A high-level overview of our framework, named CINM (Cinnamon), is presented in the figure above. Leveraging the hierarchical abstractions of MLIR, Cinnamon consists of high-level device-agnostic and low-level device-specific abstractions, each implementing a series of analysis and transformation passes. CINM is already used by some of our academic partners and has received special interest from startups developing CIM systems. The core part of CINM is released and can be accessed via GitHub (see 'Further reading'). The framework is still under active development; if you are a company or research group interested in contributing or using it, please feel free to reach out to discuss the next steps.

Contact:
Asif Ali Khan asif_ali.khan@tu-dresden.de
Jeronimo Castrillon jeronimo.castrillon@tu-dresden.de

FURTHER READING:

A. Boroumand, S. Ghose, Y. Kim, R. Ausavarungnirun, E. Shiu, R. Thakur, D. Kim, A. Kuusela, A. Knies, P. Ranganathan, and O. Mutlu, 'Google workloads for consumer devices: Mitigating data movement bottlenecks', in Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS'18, (New York, NY, USA), p. 316–331, Association for Computing Machinery, 2018.

A. A. Khan, J. P. C. De Lima, H. Farzaneh, and J. Castrillon, 'The landscape of compute-near-memory and compute-in-memory: A research and commercial overview', arXiv preprint arXiv:2401.14428, 2024.

C. Lattner, M. Amini, U. Bondhugula, A. Cohen, A. Davis, J. Pienaar, R. Riddle, T. Shpeisman, N. Vasilache, and O. Zinenko, 'MLIR: Scaling Compiler Infrastructure for Domain Specific Computation', in IEEE/ACM International Symposium on Code Generation and Optimization (CGO), pp. 2–14, 2021.

A. Siemieniuk, L. Chelini, A. A. Khan, J. Castrillon, A. Drebes, H. Corporaal, T. Grosser, and M. Kong, 'OCC: An Automated End-To-End Machine Learning Optimizing Compiler for Computing- In-Memory', IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2021.

Open CIM Compiler github.com/adam-smnk/Open-CIM-Compiler

H. Farzaneh, J. P. C. de Lima, M. Li, A. A. Khan, X. S. Hu, and J. Castrillon, 'C4CAM: A Compiler for CAM-Based In-Memory Accelerators', in Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, ASPLOS'24, (New York, NY, USA), Association for Computing Machinery, 2024.

H. Farzaneh, J. P. C. D. Lima, A. N. Khelejani, A. A. Khan, M. Mayahinia, M. Tahoori, and J. Castrillon, 'SHERLOCK: Scheduling Efficient and Reliable Bulk Bitwise Operations In NVMs', in Proceedings of the 61st Annual Design Automation Conference (DAC'24), ACM, June 2024.

J. P. C. de Lima, A. A. Khan, L. Carro and J. Castrillon, 'Full-Stack Optimization for CAM-Only DNN Inference', in Proceedings of the 2024 Design, Automation and Test in Europe Conference (DATE), DATE'24, pp.1–6, IEEE, March 2024.

A. A. Khan, H. Farzaneh, K. F. A. Friebel, C. Fournier, L. Chelini, and J. Castrillon, 'CINM (Cinnamon): A Compilation Infrastructure for Heterogeneous Compute In-Memory and Compute Near-Memory Paradigms', in (to appear) Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'25), ASPLOS '25, Association for Computing Machinery, March 2025.

CINM Compiler github.com/tud-ccc/Cinnamon

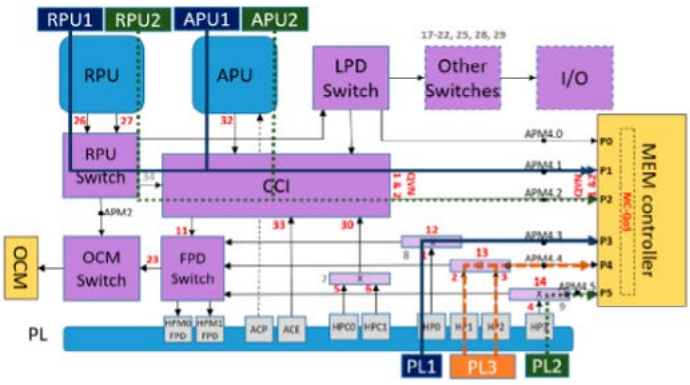
With the performance demands of modern safety-critical systems growing, engineers are puzzling out how to use more complex chips while complying with stringent safety requirements. Here, Jaume Abella and Francisco J. Cazorla (Barcelona Supercomputing Center) explain their group's approach.

Performing under pressure

Enabling the use of high-performance processors in safety-critical applications

As systems used in safety-critical settings, such as cars and robots, become increasingly automated and autonomous, their performance demands increase. High-performance (HP) processors are used to deliver the necessary functionalities in real time. However, the complexity of HP processors poses challenges for the development of safety-critical systems.

'The challenges of using HP processors in safety-critical systems may be divided into three main areas,' explains Jaume Abella, one of the directors of the Critical and AutOnomous Systems (CAOS) group at Barcelona Supercomputing Center (BSC). 'First, HP processors may have millions of possible configurations. Second, HP processors can cause applications to experience arbitrary timing interference due to their interactions with other applications in the use of shared resources. Third, HP processors offer limited observability channels to identify the cause of the interference. This is crucial because, as well as affecting performance, interference can prevent a system meeting the real-time guarantees which underpin safety-critical applications, such as those responsible for controlling steering and braking in autonomous vehicles.'



Systematic exploration of processor configurations

To tackle these challenges, the CAOS group at BSC has developed a number of technologies over the years. These range from low-level software libraries to closely manage commercial off-the-shelf (COTS) devices to hardware modules in the form of open-source intellectual properties (IPs) offering observability and controllability support to master timing interference in HP processors. According to Francisco J. Cazorla, also a director of the CAOS group, 'HP processors can only be adopted for safety-critical, real-time applications if they are as user-friendly and cost-effective to use as the much simpler processors that have already been used for decades in these domains'.

Mastering HP processor configuration

Analysis by the CAOS group shows that the number of potential configurations in modern HP processors relevant for domains such as automotive and avionics may easily go into hundreds of millions. 'Not only is the number massive: many of the combinations of parameters that lead to these configurations are simply invalid because they may prevent a task running in a core from accessing the memory, for example,' says Jaume. 'Even worse, different hardware IPs in HP processors have been designed independently, often by different IP providers, so the way their parameters interact is neither controlled nor documented.'

In this context, BSC has developed a number of methodologies showing how configurations can be explored systematically to identify the best one for given performance requirements. The CAOS group has also proposed an approach, outlined in the ECRTS 2023 paper cited in 'Further reading', whereby skilled engineers generate a set of configurations in response to a given set of abstract requirements. 'Using this approach, end users just have to match their requirements with those of one of the configuration templates before selecting the corresponding configuration, hence abstracting end users away from platform complexity,' explains Francisco, the last author of the ECRTS paper.

Performance observability

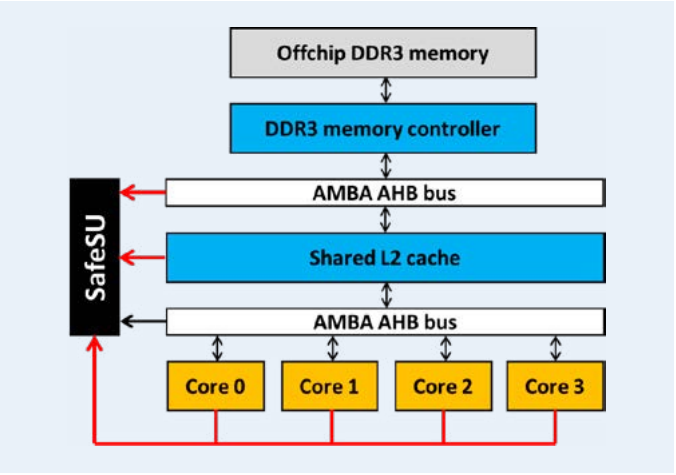
Through analysis of multiple HP processors from different vendors, the CAOS group reached two conclusions regarding performance observability in HP processors. 'The first conclusion we came to was that the number of event monitors in these devices was far from sufficient for monitoring timing interference across applications accurately,' explains Jaume. 'The second conclusion was that these event monitors are often last-minute additions for the verification of processors. Hence their operation may be unreliable, due to a lack of verification of the monitors themselves; their functionality is also poorly described, and may be completely counterintuitive.'

In response, the CAOS group has developed a number of methodologies aimed at validating the real functionality of relevant event monitors, i.e. those providing information that is useful for characterizing the use of shared hardware resources. 'Using these methodologies we can validate the behaviour of individual monitors, such as whether an access counter properly counts all types of access, or only specific subsets. We can also validate the relationship between event monitors, such as whether a set of monitors offers a breakdown of events counted by another monitor, allowing greater confidence in the value measures of those monitors,' says Francisco.

The CAOS group has also developed platform-specific libraries to abstract the complexities of using these monitors for platforms in the automotive and avionics domains. Some of these libraries have been licensed to Rapita Systems SL (formerly Maspatechnologies, a spinoff of BSC created by Francisco and Jaume). Along with other software-validation tools also licensed by CAOS, this has allowed Airbus to achieve certification for one of their avionics systems; see Maspatechnologies news release in 'Further reading', below.

Hardware support for performance monitoring

The CAOS group's third set of solutions aims to make software development easier by providing explicit, unambiguous information needed to design, verify and validate safety-relevant applications on HP processors. The group's 'safe statistics unit' (SafeSU) module is a portable module that is compatible with the most popular on-chip interconnect interfaces available in numerous HP processors. The SafeSU offers specialized and validated event monitors providing information about the use of shared hardware resources and the interference caused by different tasks within the system. In addition, it provides end users with pre-program access quotas for the different tasks running so that, if any of the quotas is violated, an interrupt is raised and system software can react accordingly, with short latency and zero cost, to monitor access counters. SafeSU is publicly available (see the GitHub repository in 'Further reading', below); earlier versions have already been integrated in the context of a RISC-V multicore chip for the space domain, as set out in the De-RISC paper cited in 'Further reading'.



The SafeSU module

Overall, the CAOS group at BSC has developed several complementary hardware and software technologies paving the way towards the adoption of the HP processors required for performance-hungry, safety-critical applications. These technologies are currently being further matured as part of different Chips JU initiatives and other Horizon Europe projects. Some of these initiatives, such as ISOLDE, aim to deliver RISC-V HP processors for domains such as automotive, in initiatives relating to the development of the 'software-defined vehicle'; others, such as SAFEXPLAIN and EdgeAI-Trust, are focused on integrating artificial intelligence (AI) components as part of safety-critical systems.

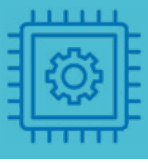
FURTHER READING:

Sergio Garcia-Esteban et al., 'Quasi Isolation QoS Setups to Control MPSoC Contention in Integrated Software Architectures', 35th Euromicro Conference on Real-Time Systems (ECRTS 2023). Leibniz International Proceedings in Informatics (LIPIcs), volume 262, pp. 5:1-5:25, Schloss Dagstuhl – Leibniz-Zentrum für Informatik (2023)
doi.org/10.4230/LIPIcs.ECRTS.2023.5

'Maspatechnologies Helps Airbus Pass Multicore Certification on a NXP T2080' bit.ly/Maspatechnologies_Airbus_LinkedIn

SafeSU on GitHub github.com/bsc-loca/SafeSU

N.-J. Wessman et al., 'De-RISC: the First RISC-V Space-Grade Platform for Safety-Critical Systems,' 2021 IEEE Space Computing Conference (SCC), Laurel, MD, USA, 2021, pp. 17-26
doi.org/10.1109/SCC49971.2021.00010



Neural networks are fundamental to many modern computing systems, but their computation and energy demands pose a challenge. In this article, Alexander Lehnert, Marc Reichenbach (both University of Rostock) and Ralf Müller (Friedrich-Alexander-Universität Erlangen-Nürnberg) explain how their approach, named 'computation coding', could be a game changer for accelerator design.

Computation coding

Revolutionizing ANN representation for optimal hardware design



Artificial intelligence (AI) and artificial neural networks (ANNs) are common building blocks in most systems today. However, the current trend of replacing every algorithm with AI models comes with challenges. At the edge (e.g. mobile devices), and particularly at the extreme edge (e.g. highly integrated sensors), computational resources and energy are immensely restricted. In response, system designers are coming up with dedicated accelerator circuits to improve the efficiency of running inference in AI models. In this domain, European innovation and research is providing some promising solutions.

While most AI accelerator architectures perform well, the optimal accelerator is yet to be designed. Finding the optimal representation of ANNs for both the application and the processor is anything but trivial. Traditional approaches to speed up computation or to reduce the computational cost of algorithms, such as quantization of weights or pruning, focus on compressing memory aspects of AI models. However, little research aims at computational compression of the operations at stake.

In the paper 'Linear Computation Coding' (see 'Further reading'), Müller et al. present a novel representation of linear operations, which dominate the computational cost of inference in ANNs, called 'computation coding'. This patented method performs exceptionally well at finding the representation of a linear function which requires the least number of operations for any desired precision. For traditional matrix-vector multiplications, a reduction of up to 84% is not untypical. Hardware accelerators directly benefit from this with reduced resource and power requirements, as Lehnert et al. have demonstrated. Müller et al. have also demonstrate the viability of this approach for convolutional neural networks (see 'Further reading' for all papers cited). From this we conclude that computation coding is the basis for breakthroughs in hardware design for ANN accelerators.

Computation coding from edge to cloud

The world of AI accelerators is divided into two paradigms. At the edge, small-scale processors which take up minimal resources are required. At the other end of the continuum, the processing requirements of data centres call for rolled-out dataflow architectures with maximal throughput. Both use cases benefit from the computation-coding representation of constant matrix vector multiplication. An in-depth review of dataflow architectures using this approach is presented in the 2023 IEEE Access paper by Lennart et al. (see 'Further reading'). The dataflow architecture leverages the reconfigurability of field-programmable gate arrays (FPGAs) to implement certain operations without resource overhead. All of this is possible thanks to the particular representation of multiplications in the computation-coding approach.

At its core, computation coding produces an approximate multiplicative matrix decomposition. It performs exceptionally well at low quantization rates and large problem dimensions, where traditional approaches fail. Expensive multiplications are represented using bit-shifts and additions only, all combined in a fixed computation structure. This enables optimal usage of reconfigurable logic such as FPGAs. Routing is used to implement bit-shifts leaving additions as the only cost for resource requirements, while the fixed processing structure enables dense pipelined hardware designs with minimal total slack.

All of this is possible with a low energy footprint and no impact on throughput. In a 2023 paper published in Applied Reconfigurable Computing (see 'Further reading'), Lehnert et al. show that even highly unconventional three-input adders can be leveraged on FPGAs with great effectiveness. Meanwhile, ongoing work is investigating a programmable accelerator design based for processing computation coding decompositions (see the patent application in 'Further reading'). It is our

contention that with this accelerator European system design will compete on the international stage.

We developed a Python-embedded hardware description framework to map dataflow architectures, especially computation coding decompositions, to FPGAs. Rather than embedding a hardware description language (HDL) into Python, we provide a library to leverage Python programming techniques directly.

Our approach comes with distinct benefits over existing HDLs. First, hardware designs are modelled using the high-level programming concepts of Python. Second, architectures are mapped to the low-level VHSIC hardware description language (VHDL), which guarantees optimal compatibility with a wide range of hardware design tools. At the same time, and in contrast to concepts such as high-level-synthesis (HLS), the semantic gap between the description of hardware and the actual mapping to logic is bridged deterministically. This allows fine-grained control over the architecture, while still permitting automated pipelining of architectures. Our framework will be open source and available to hardware architects across Europe.

The new area of computation coding is still in its infancy, and it is expected that many more algorithms and codes will be discovered. This is true for linear, but also for nonlinear multi-dimensional functions. It will significantly reduce both the chip area and the power consumption of many future neural networks.

REFERENCES

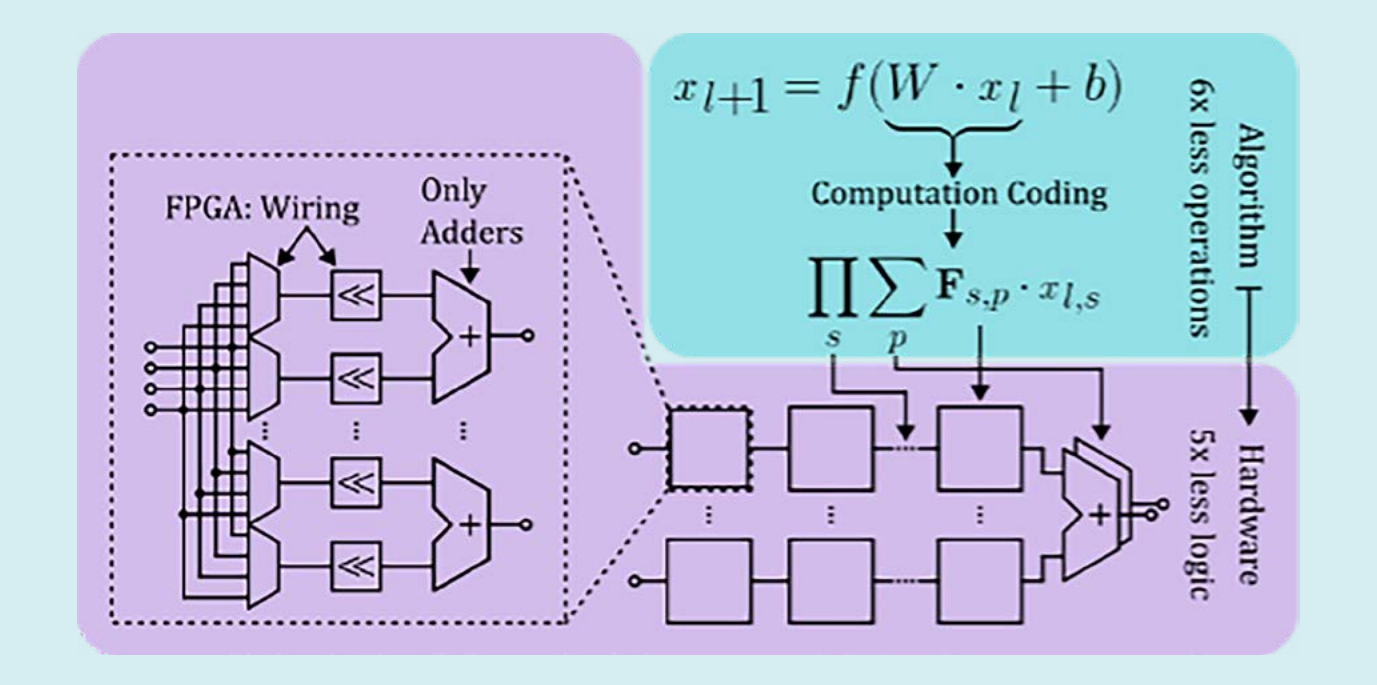
Lehnert, P. Holzinger, S. Pfenning, R. Muller, and M. Reichenbach, 'Most Resource Efficient Matrix Vector Multiplication on FPGAs', IEEE Access, vol. 11, pp. 3881-3898, 2023, doi: 10.1109/ACCESS.2023.3234622
ieeexplore.ieee.org/document/10007836

Lehnert, H. Rosenberger, R. Müller, and M. Reichenbach, 'More Efficient CMMs on FPGAs: Instantiated Ternary Adders for Computation Coding', in Applied Reconfigurable Computing. Architectures, Tools, and Applications, vol. 14251, F. Palumbo, G. Keramidas, N. Voros, and P. C. Diniz, Eds., in Lecture Notes in Computer Science, vol. 14251., Cham: Springer Nature Switzerland, 2023, pp. 275-289. doi: 10.1007/978-3-031-42921-7_19.
link.springer.com/chapter/10.1007/978-3-031-42921-7_19

R. Müller, B. Gäde, and A. Bereyhi, 'Linear Computation Coding', 2021, doi: 10.48550/ARXIV.2102.00398.
arxiv.org/abs/2102.00398

R. R. Müller, H. Rosenberger, and M. Reichenbach, 'Linear Computation Coding for Convolutional Neural Networks', in 2023 IEEE Statistical Signal Processing Workshop (SSP), Hanoi, Vietnam: IEEE, Jul. 2023, pp. 562–565. doi: 10.1109/SSP53291.2023.10207943.
ieeexplore.ieee.org/document/10207943

R. Müller, H. Rosenberger, and M. Reichenbach, 'Apparatus and method for computing a matrix vector product of a certain matrix and a vector', US20240028665A1, 25 January 2024



Mapping computation coding decompositions to FPGAs


Existing hardware-design tools are automated but still rely significantly on human guidance to reach an efficient hardware design. In response, Sam Coward (a PhD student at Imperial College London and Intel) and Jianyi Cheng (an assistant professor at the University of Edinburgh, formerly a research intern at Intel) propose an approach based on equality saturation to automatically discover optimal hardware designs. In this article, they explain how equality saturation can help optimize hardware design and reduce the need for human intervention.

Equality saturation

A new approach to optimal hardware design

While hardware is designed using software, there are major differences between hardware design and software programming, according to Jianyi Cheng (University of Edinburgh). 'Designing digital circuits is fundamentally different from software programming, for two main reasons. First, traditional software development mainly focuses on performance, while hardware development focuses on more complex metrics. Typically, customers demand high-performance computer chips that sip power and are price competitive, which means they must fit within a reasonable silicon area budget,' explains Jianyi. 'Second, the development cycle of digital circuits is usually years, significantly longer than software development. This leads to a long feedback loop to evaluate the quality of the design and verify its correctness. Although hardware design tools today provide automation and simulations, they still require significant human guidance to produce an efficient design.'

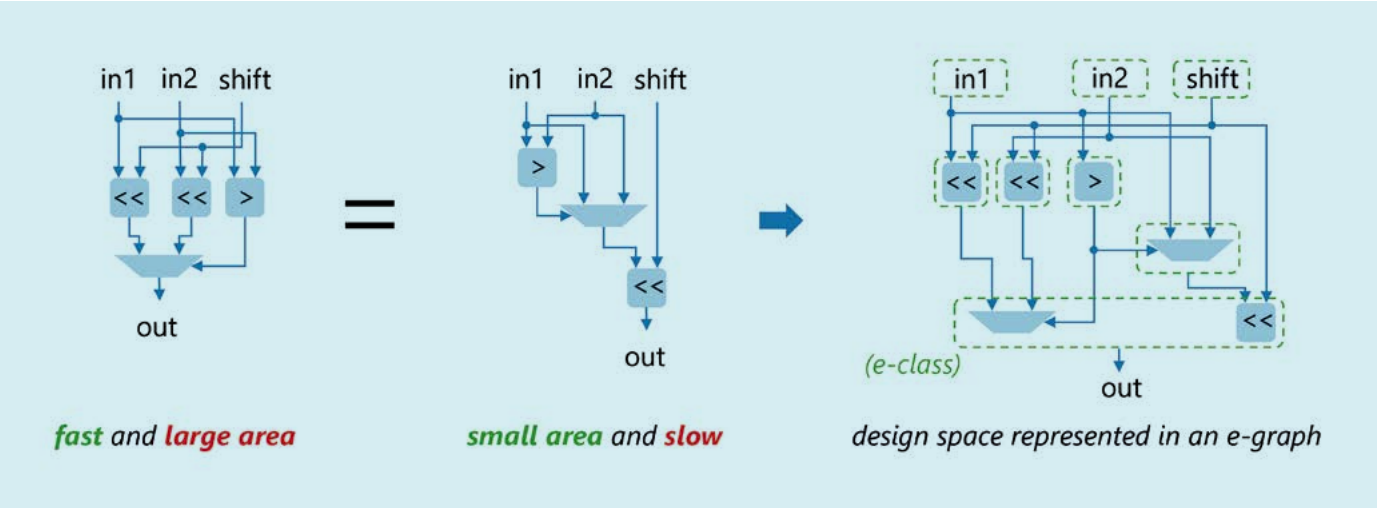
One of the issues with existing design tools is that they apply a fixed sequence of optimizations for all inputs. This is problematic because 'a fixed sequence only explores a restricted design space for a particular program, potentially making the optimal design unreachable – an issue known as the "phase-ordering problem" in compilers,' says Sam Coward (Imperial College / Intel). 'Finding the correct order to apply optimizations is challenging, because applying one optimization prevents the application of another. In addition, the long feedback in the hardware design flow makes an exhaustive search infeasible.'

To help designers explore such a huge hardware design space efficiently, Sam and Jianyi have developed two hardware optimizers named ROVER and SEER. 'These tools express hardware design metrics at the source level, facilitating evaluation with fast feedback, and exploit equality saturation to explore all optimization sequences concurrently. Given the design specifications and requirements, ROVER and SEER automatically determine a sequence of optimizations for efficient hardware synthesis,' explains Jianyi.

Already established in the compiler field, equality saturation is a technique that constructively applies optimizations and analyses to infer a set of equivalent design candidates. In 2022, Sam published a paper on applying equality saturation to hardware design at ARITH 2022. 'This paper provided a way to encode digital circuits in the data structure at the core of equality saturation, the equivalence-graph (e-graph),' says Sam. 'The resulting optimization tool, ROVER, explored different functionally equivalent digital circuits of the same arithmetic expression.'

An example is shown in the left and the middle of the figure on the facing page, which are functionally equivalent, says Sam. 'The left implementation executes all the operations in parallel, using two logical shift circuits. In contrast, the middle circuit executes the comparison and logical shift in serial, using only a single shift circuit. Both implementations are functionally equivalent, forcing the designer to make a difficult choice. Note that existing hardware design tools do not automate these choices.'

To the right of the figure is an e-graph, which Sam uses to model the hardware designs. 'An e-graph is not a single digital circuit, but a set of its equivalent representations. In the given example, the e-graph contains both the left and middle designs, where equivalent expressions are grouped into equivalence classes,' he says. 'Equality saturation is the process of growing an e-graph by applying a set of user-defined rewrites, adding new equivalent expressions to the data structure. The key idea is to retain all candidates, deferring the final implementation selection to a later extraction phase.' This is the basis for the ROVER tool, a fully automated datapath optimizer, which the researchers note is capable of reducing circuit area by up to 63%.



Aware that a limitation of ROVER is that it only handles data paths, while most hardware designs also contain control paths, the researchers published joint work at ASPLOS 2024 that extends the datapath abstractions to high-level synthesis (HLS). 'HLS automatically translates a software program into a functionally equivalent hardware design. An HLS program expresses both the control and data paths in its software abstraction,' says Jianyi. 'In this work, we connected the software abstraction to the e-graph. In particular, we express MLIR, a popular software intermediate representation, in an e-graph.'

'We developed automatic orchestration methods to reuse existing software optimizations in MLIR passes for hardware optimizations. This significantly expands the design space and explores potential optimal designs at scale,' adds Sam. 'For example, optimization techniques for hardware control paths, such as loop unrolling and loop tiling, are explored using existing loop transformation passes in MLIR. In addition, the multi-level abstractions in MLIR allow concurrent exploration of optimizations at different granularities. We observed that the combination of control path and data path optimizations in a certain order in our e-graph could potentially lead to a better design than existing tools and even human hardware experts.'

Finally, the paper also proposes a hardware evaluation model to evaluate these designs at the source level, avoiding the long-feedback in hardware design tools and identifying efficient designs in an e-graph in seconds, according to the researchers. This work resulted in an HLS optimzer named SEER at Intel. Given a software program, SEER automatically determines a sequence of optimizations for efficient hardware synthesis. 'From this work, we concluded that equality saturation provides a path to efficiently reach optimal digital circit designs,' says Jianyi. 'Our future work will explore the potential of hardware-software co-design via equality saturation, raising our optimization exploration to the system level.'

'One review of our paper noted that, while there is an obvious need for better hardware design tools, there is a danger of worsening already slow compile times. It suggested that equality saturation, coupled with heuristics for approximating code quality, could be the right answer to this problem,' adds Sam.

FURTHER READING

S. Coward, G. A. Constantinides and T. Drane, 'Automatic Datapath Optimization using E-Graphs', 2022 IEEE 29th Symposium on Computer Arithmetic (ARITH), Lyon, France, 2022, pp. 43-50, doi: 10.1109/ARITH54963.2022.00016
ieeexplore.ieee.org/document/9974492

J. Cheng, S. Coward, L. Chelini, R. Barbalho and T. Drane, 'SEER: Super-Optimization Explorer for High-Level Synthesis using E-graph Rewriting', ASPLOS '24: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2, pp. 1029-1044, doi:10.1145/3620665.3640392
dl.acm.org/doi/10.1145/3620665.3640392

"ROVER and SEER automatically determine a sequence of optimizations for efficient hardware synthesis"

Coordinated by Barcelona Supercomputing Center (BSC), the European PILOT (EUPILOT) project will deliver the first open-source and open-standard-based software- and hardware-integrated high-performance computing (HPC) system by creating a set of accelerators designed, implemented, manufactured, and deployed fundamentally in Europe.

EUPILOT progresses towards advanced computing accelerators made in Europe

Carlos Puchol and Romana Konjevod, Barcelona Supercomputing Center (BSC)

EUPILOT aims to create and validate two EU-based accelerator platforms, one for HPC and another for compute-intensive machine learning (ML) and artificial intelligence (AI) applications. The implementation will cover the full range of required technologies, from HPC and high-performance data analytics (HPDA) applications to a complete software stack on the software side. The accelerators are based on highly efficient and high-performance RISC-V cores, integrated into high-density accelerator boards and chassis with innovative immersion cooling.

RISC-V, an open-source instruction set architecture (ISA), is central to EUPILOT. The project leverages RISC-V to drive innovation in processor design and promote a flexible computing ecosystem. EUPILOT will develop new RISC-V implementations, showcasing their performance through applications like GROMACS (molecular dynamics), EC-EARTH (climate modelling), AI video processing, and AI-based drug discovery.

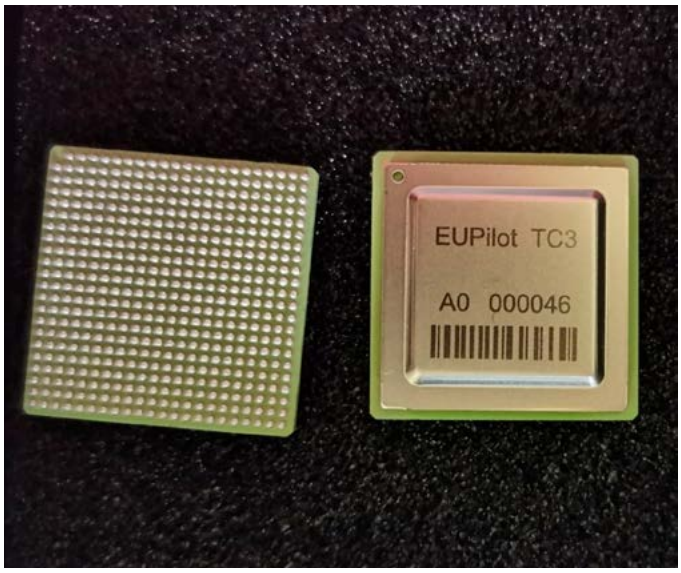
In so doing, EUPILOT is helping to push the boundaries of what RISC-V can achieve in terms of performance, efficiency, and application scope. In addition, by incorporating RISC-V into its frameworks, EUPILOT is supporting the growth of an open-source ecosystem. This promotes collaboration and innovation across the industry, reducing reliance on proprietary technologies. The project involves extensive testing and validation of RISC-V technologies, which helps refine the architecture, ensuring it meets performance and reliability standards for various applications.

EUPILOT is also advancing academic research and education by offering a practical platform for studying and developing RISC-V technologies to researchers and students interested in open-source hardware. By showcasing practical applications and demonstrating the benefits of RISC-V, EUPILOT is contributing to increased adoption of RISC-V technologies in both commercial and research settings.

Pre-exascale accelerators

On the hardware front, the EUPILOT project aims to build an end-to-end demonstrator of accelerators intended for pre-exascale systems. Making full use of European and open-source technologies and standards, the project will produce three chip tapeouts.

The first tapeout is a test chip for validating the 12nm technology node, improving from previous 22nm technology node tapeouts. In early July 2024, EUPILOT received its first 12nm test chips, designed with GlobalFoundry's 12nm FinFet technology. Testing is ongoing, with initial results indicating expected performance and successful data transmission across SerDes/PHYs. This milestone shows strong advancement in the project, as well as promising further progress in future. The second and third tapeouts, developed concurrently, will include a vector accelerator with up to eight cores and a machine learning and stencil accelerator with up to four groups of cores.



EUPILOT's first tapeout: 12nm test chips

"This European end-to-end demonstrator project showcases technology trends for the future of high-performance computers"

Each EUPILOT chip will be combined with LPDDR memory chips onto accelerator modules that will be installed into standard accelerator boards, leveraging the Open Compute Project Accelerator Module (OAM) standard. These, put together in groups, comprise the accelerator systems in OCP's Universal Base Boards (UBBs). These accelerator systems are paired with host servers and are deployed in liquid immersion tanks to efficiently support ultra-high-power densities at very high power-usage effectiveness (PUE) levels. In this way, this European end-to-end demonstrator project showcases technology trends for the future of high-performance computers.

From a European and global level, EUPILOT will provide a full accelerator technology as a a comprehensive accelerator technology foundation for European exascale systems. The capabilities of the RISC-V ecosystem will be showcased and expanded into the HPC and HPDA domains, thus contributing to European digital sovereignty and technology independence in HPC.

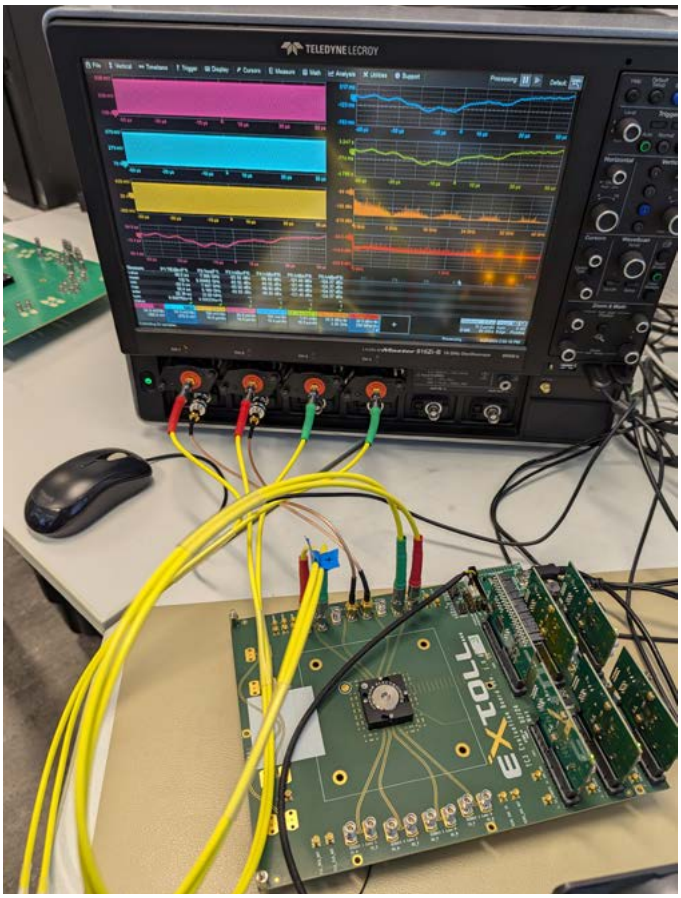
The project is about two-thirds complete and collaborates with the EUPEX project, which focuses on co-designing a European modular exascale-ready pilot system into which the EUPILOT accelerators can fit as offloading engines for HPC and AI. This successful collaboration will culminate in a demonstration of interoperability at the project's conclusion, which will both mark a significant milestone in advancing Europe's position in high-performance computing and contribute towards innovation on the wider global tech landscape.

FURTHER INFORMATION:

The European PILOT (EUPILOT)
eupilot.eu

EUPILOT hardware
eupilot.eu/hardware

European Pilot for Exascale (EUPEX)
eupex.eu



EUPILOT accelerator boards, from partners EXAPSYS (left) and EXTOLL (right)

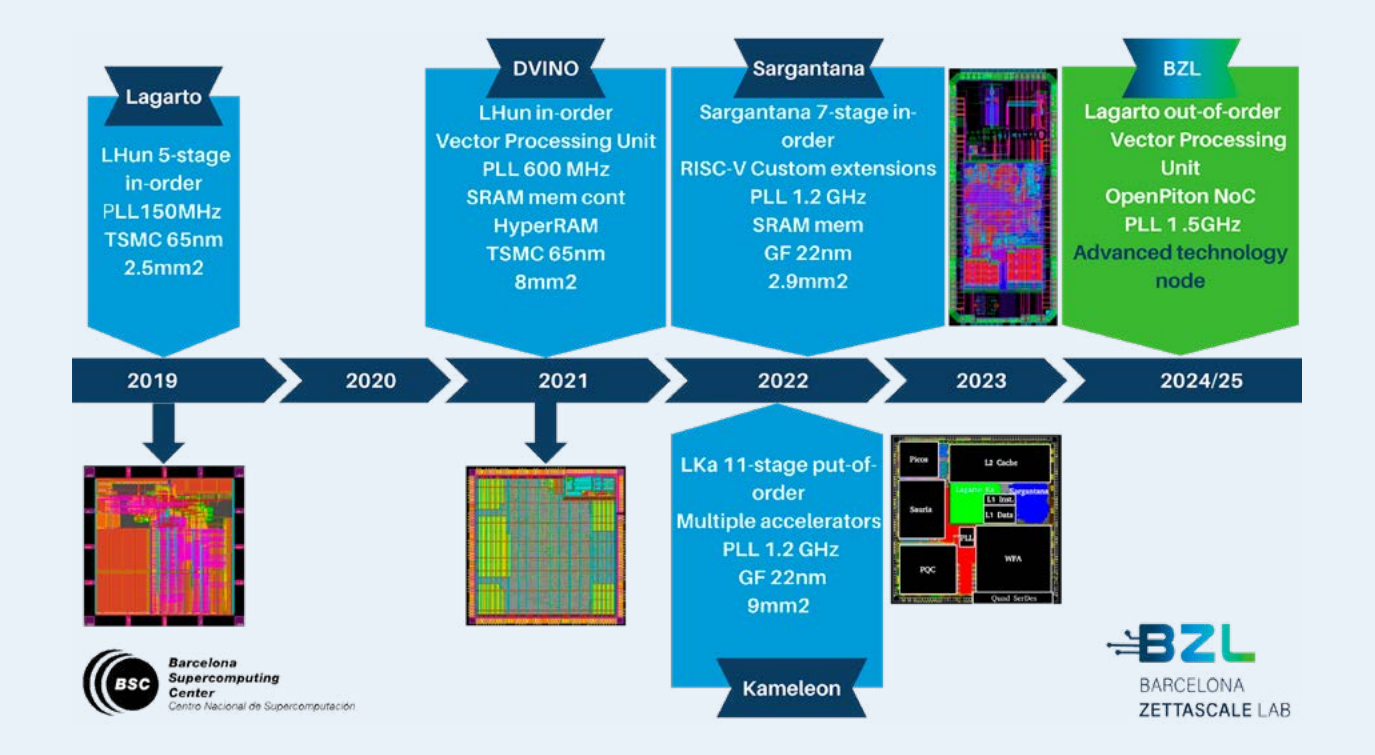
Building on the growing momentum in RISC-V, the new Barcelona Zettascale Lab will be a development hub for high-performance chip prototypes. In this article, Mateo Valero, Rafael Gomà, Miquel Moretó and Xavier Teruel (Barcelona Supercomputing Center) tell us how this project will promote innovation and take a further step towards European hardware sovereignty.

Barcelona Zettascale Lab:
Innovative HPC chip solutions made in Europe

RISC-V is gaining popularity across application segments. Although the most successful market for RISC-V so far has been embedded computing and the internet of things (IoT), comprising 90% of RISC-V-based processors shipped annually, companies are already developing RISC-V products in other markets, such as aerospace, automotive, and data centres.

In this context, Barcelona Supercomputing Center (BSC) is leading the Barcelona Zettascale Lab (BZL) project with the aim of improving Europe's capacity in the design of high-performance computing (HPC) chips based on RISC-V, as another step towards achieving European technological sovereignty. Co-funded by the Spanish Ministry of Digital Transformation and of Public Services, within the framework of the Resilience and Recovery Facility and the Next Generation funds from the European Union, BSC will continue leading the way in the design of high-performance chip prototypes to enable alternative technology pathways as a mechanism to influence future European supercomputers.

In addition to hardware, the BZL project covers the software stack and promotes hardware and software co-design activities to enhance the productivity of both teams, in line with the objectives of the project. On the software side, BZL aims to contribute to the ecosystem, including the operating system and device drivers, compilers, runtimes supporting parallelism in shared-memory and distributed environments, and basic numerical libraries. This allows the execution both of representative applications in the HPC domain and of emerging applications that combine aspects of artificial intelligence (AI) with simulations based on the traditional principles of HPC.



Timeline of BSC RISC-V SoC tapeouts

The size and scope of the BZL project has posed challenges in many areas, ranging from the organizational perspective on balancing the investigation and delivery focus to the need for a solid strategy of software development vehicles (SDVs) to support design, verification and emulation across the different phases of the project. A key resource in this context is BSC's field-programmable gate array (FPGA) emulation capability. This takes advantage of the infrastructure developed by the European MEEP and European Processor Initiative (EPI) projects, creating an environment to support hardware emulation of large RISC-V systems by partitioning the design and addressing observability challenges.

'The BZL project will develop over several steps of increasing complexity, with a last tapeout in 2025 of a multicore system-on-chip (SoC) on advanced technology nodes,' says Mateo Valero, BSC director and BZL coordinator.

In-depth expertise in RISC-V architectures

To build the BZL prototypes and its software toolchain, BSC will draw upon its extensive experience in RISC-V architectures and components arising from various research projects. This co-design approach between hardware and software has been demonstrated to be effective in former contributions to the RISC-V community.

Within the framework of the EPI, EUPilot and eProcessor projects, BSC has developed a vector processing unit (VPU) with long vectors of 256 double-precision elements. This VPU has eight floating-point functional units and can perform 16 floating point operations per cycle. BZL's objective is to develop the current VPU to support more floating-point operations per cycle, follow the RISC-V vector extension 1.0 and enable manufacturing with advanced technology nodes.

As part of the Catalan-funded project DRAC (Designing RISC-V-based Accelerators for next generation Computers), BSC has already designed and manufactured four scalar SoCs with increasingly advanced features: Lagarto Hun (2019 on TSMC 65nm), DVINO (2021 on TSMC 65nm), Sargantana (2022 on GF 22nm) and Kameleon (2022 on GF 22nm). At BZL, BSC experts aim to advance the design of scalar cores by developing a new core, Lagarto Ox, that has enough performance to feed the advanced VPU. Lagarto Ox will be able to execute instructions out of order and issue up to four instructions per cycle.

In terms of the on-chip cache hierarchy and an outcome of MEEP, BSC designed a two-level cache memory hierarchy based on the OpenPiton manycore processor project. BZL researchers aim to develop the family of Lagarto cores compatible with OpenPiton, add a third cache level, increase the number of in-flight cache requests and maximize the bandwidth of the interconnection network.

With respect to the BZL's software stack, different BSC teams are working on several software components. Within the Operating System (OS) Group, the project has several Linux distributions (leveraging results from EPI, MEEP, and EUPilot) booting on the proposed platforms. They are also developing specific new drivers and services for such systems, a continuation of the work carried out in MEEP. The SDV, compiler (i.e. LLVM) and performance analysis (i.e. Paraver / Extrae) teams are in charge of creating a RISC-V tool-chain to enable the exploitation and performance awareness of the underlying hardware. Finally, the frameworks and libraries teams will strengthen several HPC (e.g. MPI) and AI (e.g. Pytorch) components to support the most representative applications in the scientific domain.

FURTHER INFORMATION:

Barcelona Zettascale Lab bzl.es

NextGeneration EU next-generation-eu.europa.eu

MareNostrum Experimental Exascale Platform (MEEP) meep-project.eu

European Processor Initiative (EPI) european-processor-initiative.eu

EUPilot eupilot.eu

eProcessor eprocessor.eu

Designing RISC-V-based Accelerators for next generation Computers (DRAC) drac.bsc.es/en

OpenPiton parallel.princeton.edu/openpiton

Paraver tools.bsc.es/paraver

Extrae tools.bsc.es/extrae

The director of the Microelectronics Design Center at ETH Zurich, Frank K. Gürkaynak is a vocal advocate for open hardware, including through the Parallel Ultra-Low Power (PULP) Platform, a collaboration between ETH Zurich and the University of Bologna. With a significant portfolio of tapeouts under its belt, PULP has been pioneering open hardware in Europe. HiPEAC caught up with Frank to learn more about why openness is central to his hardware philosophy.

'End-to-end openness potentially lowers the barriers to getting ideas turned into silicon'

What does 'end-to-end openness' refer to in hardware development? Why is it important?
End-to-end openness involves open-source principles at all levels of the design of integrated circuits (ICs). Everything in the design process should be accessible to anyone, from the manufacturer's process design kit (PDK), which contains all the information allowing an IC to be reliably manufactured in a particular technology, to the electronic design automation (EDA) tools used to transform a design idea into a manufacturable circuit, to the description of the design itself.

By dramatically increasing accessibility to all aspects of IC design, teaching and training can be made available to a broader audience, which has the potential to significantly lower the barriers to getting ideas turned into silicon. In addition, there are several applications where transparency is of the utmost importance, and having this level of access to all design stages supports independent audits. Lastly, it removes black boxes in the design process and allows innovation at all levels of the IC design flow.

OK, great! Let's get started on my next open-hardware idea...
Not so fast. IC design is still a costly endeavour with very high non-recurring engineering (NRE) costs that can only be compensated with volume. Openness lowers the barriers but doesn't turn the cost structure upside down.

Besides, open-source EDA is still in its infancy: it's possible to make viable designs, but there is still a gap between open-source EDA and commercial solutions in terms of the features, the usability and the quality of the results, which is to be expected. Currently, designs in mature technologies in the 180nm–130nm range would be feasible with EDA tools; some more work is required to move to designs in the 65nm–28nm range, and there is still a gap for the most advanced nodes in 10nm and under.

The important part is the open PDKs being made available. At the moment these are limited to mature technologies, such as the 130nm from the Leibniz Institute of High Performance Microelectronics (IHP) in Germany. A technology in the 65nm–28nm range, in particular, would allow many industrial designs to be realized efficiently.

When we last spoke [see HiPEACinfo *66 p.23], you mentioned that the EDA tools market was dominated by proprietary tools. Is this changing?*
This still holds true: at least 99% of the market uses proprietary tools. However, EDA design tools allow design flows to be created with a mix of tools from different vendors, both proprietary and open source. While I don't expect a wholesale change from completely proprietary to completely open tools, I do expect a gradual penetration of tools from the open-source community into design flows.

One important difference is that, while proprietary tools' licensing fees place a limit on the number of concurrent users / runs you can have, open-source tools can be used without limitation. This makes open-source tools very attractive for automated tasks such as those in continuous integration and early evaluation, where the design is still being refined.

In these discussions, people often think of high-end designs like machine-learning accelerators destined for data centres. However, this market constitutes a relatively small proportion of ICs designed. For many applications you can afford to have a design that's a bit larger, a bit slower, but it would help if you could get started right away without negotiating licensing agreements, especially if it's not 100% sure that you will actually commit to an IC design process.

For this reason, I expect that the availability of open-source tools will increase the number of people who get some familiarity with design flows, increasing innovation in this area. This could, eventually, even lead to more business for commercial EDA vendors, as more people venture into more ambitious projects after their initial foray into IC design on open-source EDA tools and technologies.

All right, I'm ready to embark on my first design. What resources are available for people starting out in open hardware?

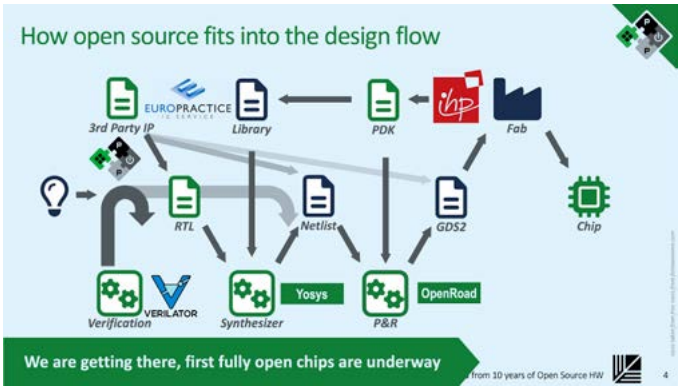
At ETH Zurich, we have a long tradition of IC design courses, dating back to 1986. From spring 2025, this course, which has led to hundreds of application-specific integrated circuits (ASICs) being designed and manufactured, will use mostly open-source tools, making exercises and lecture material available in the process. We have already managed to tape out a larger, Linux-capable, RISC-V-based system based on our Cheshire platform in IHP130 technology recently (see Basilisk information in 'Further reading'), and we will use the experience from this tapeout to adapt our course material.

In parallel, Tiny Tapeout continues to support completely open-source designs for everyone and they are in the process of adapting their tapeouts to the IHP130 process. The Tiny Tapeout approach allows enthusiasts and students at all levels to get acquainted with the IC design process and receive chips that include their design at very affordable prices.

Are there any downsides to open hardware?

Some people point out that open-source EDA tools can't be used to tape out a high-end graphics processing unit (GPU), for example. Comparisons have also shown that, so far, open-source EDA tools don't deliver the same quality of results as proprietary tools, although the difference is less than some people believe. Others are concerned about the risks involved in relying on 'free' tools that do not come with professional support for their IC projects that costs millions.

These are all valid points, but they do not really diminish the impact of open-source EDA tools. There are many designs and projects that can be realized with open-source tools that are completely sufficient for teaching and training purposes. The tools are also getting progressively better; recently I experienced this first hand, where within a six-month period the logic synthesis results progressively improved to produce circuits 1.6x smaller and 2.3x faster than they were at the beginning.



The other discussion I hear centres around national sovereignty, which either sees open source as a threat (allowing certain geographies access to technologies that could otherwise be limited), or a solution (achieving independence from companies from certain geographies). One needs to realize that open source is open for everyone, and not limited to political boundaries. I don't think we're having the same discussion abouts GCC or LLVM, so I think these discussions will also with time disappear.

FURTHER READING

Philippe Sauter et al. 'Insights from Basilisk: Are Open-Source EDA Tools Ready for a Multi-Million-Gate, Linux-Booting RV64 SoC Design?' arxiv.org/abs/2405.04257

The IIS Chip Gallery: Basilisk (2024) asic.ethz.ch/2024/Basilisk.html

Tiny Tapeout tinytapeout.com

The Basilisk chip, taped out using the IHP130 PDK

Frank teaching at the inaugural EFCL Summer School on Open Source IC Design and Computer Architectures

Faced with bottlenecks in transferring data and lagging memory capabilities, computer architects are searching for innovative solutions to ensure continued performance scaling. In this article, Joel Minguet Lopez and François Andrieu (CEA Leti) outline how the ERC-funded My-CUBE project leveraged 3D monolithic dense co-integration between high-performance, resistive random-access memory (RRAM) devices and logic transistors to shape future in-memory computing platforms.

My-CUBE harnesses CEA Leti's multidisciplinary research for memory-based computing

In today's digital era, technology has transformed our lives through the widespread proliferation of connected edge electronic devices. Along with other sources, this is inducing an exponential increase in the amount of data generated worldwide, leading to alarming estimations of the data that will be generated in the near future.

Despite tremendous progress in scaling complementary-metal-oxide-semiconductor (CMOS) based computing systems, computing performance in distributed von Neumann architectures remains seriously limited by the power consumption required for data transfer between logic and memory blocks, rather than for the computation itself (commonly referred as the 'von Neumann bottleneck'). In addition, the progress in memory and processor technologies has not been symmetrical, leading to an increasing performance gap between the two (commonly referred as 'memory wall'). Overall, finding solutions for these two challenges remains key for the development of next-generation computing platforms.

In this context, a new paradigm based on non-von Neumann systems to perform in-memory computing (IMC) with non-volatile memory (NVM) technologies is gaining importance. It was this observation that motivated My-CUBE project, funded by the European Research Council (ERC) under agreement 820048. The project aimed to develop a powerful viable technological solution for intensive memory-based computation, taking advantage of the wealth of multidisciplinary research skills at CEA, ranging from technology integration and development to application benchmarking, via device- and component-specific design.

To address this challenge, the project followed two main objectives, both targeted towards the development of 'non-von Neumann' computing systems:

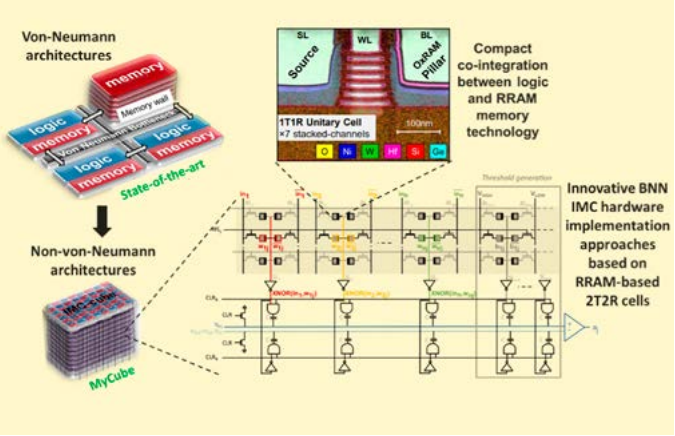
- **Memory device technology development**
  First, non-von Neumann computing systems call for a revolution in memory technologies, requiring innovative approaches that enhance tight integration of logic and memorization. To provide an answer to this challenge, the My-CUBE project coupled the high-density capabilities provided by 3D monolithic-integration approaches with the high performance provided by non-volatile resistive random-access memory (RRAM) technologies. With this strategy, My-CUBE developed the base building block of an IMC computing platform able to downscale towards very dense systems.

- **Applicative-oriented circuitry development**
  Second, non-von Neumann computing systems demand a revolution in the approaches to implementing computing in hardware, moving towards in situ IMC information processing with analogue components. However, analogue components suffer from intrinsic non-idealities, i.e. undesirable effects, which may degrade the accuracy of computations. In response, the My-CUBE project developed specific circuitry with enhanced immunity to the imperfections of these analogue components. The My-CUBE project also proposed coupling this optimized circuitry with low-precision binarized neural network (BNN) algorithms. This type of neural network exploits binary weights and activations, alleviating memory requirements while preserving high accuracy and enhanced resilience to analogue noise and variability issues. With this strategy, CEA researchers developed an ultra-low power IMC accelerator for BNN inference on-chip hardware implementation.

:

The research carried out during My-CUBE delivered a number of results:

- **A compact one-transistor – one-resistor (1T1R) memory cell**, where the RRAM active material (HfO2) is deposited inside the transistor drain contact, has been demonstrated. This approach has successfully been implemented for both gate-all-around stacked nanosheet and 28nm fully depleted silicon-on-insulator (FDSOI) logic transistor technologies. Notably, this 1T1R architecture benefits from 3D integration scalability and the co-location of logic and memory, allowing to reach memory density capabilities that are competitive with crossbar ultra-dense architectures while preventing sneak-path current issues. Remarkably, >106 endurance cycles, 2h data retention at 150°C and >109 read disturb with multi-level cell (MLC) capabilities have been achieved, which is competitive with the state-of-the art.
- **An IMC-oriented accelerator** based on RRAM-based two-transistor – two-resistor (2T2R) cells with a capacitive output neuron to implement low precision BNNs has been fabricated above CMOS technology, highlighting excellent robustness to noise and analogue variability. An excellent peak energy efficiency of 96 TOPS/W has been achieved while promising high BNN accuracy on image-classification tasks. Moreover, this solution promises a state-of-the-art 449.3 TOPS/W for an equivalent hardware implementation in 22nm FDSOI technology.
- In addition to the aforementioned neural-network accelerators, other applications have been explored, from 3D NOR non-volatile-memory to high-performance computing, both exploiting the high-density integration flow developed in My-CUBE.



Schematic illustration of the IMC computation approach developed in the framework of My-CUBE project

FURTHER READING

S. Barraud et al, '3D RRAMs with Gate-All-Around Stacked Nanosheet Transistors for In-Memory-Computing', 2020 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2020, pp. 29.5.1-29.5.4, doi: 10.1109/IEDM13553.2020.9371982

M. Ezzadeen et al., 'Ultrahigh-Density 3-D Vertical RRAM With Stacked Junctionless Nanowires for In-Memory-Computing Applications', in IEEE Transactions on Electron Devices, vol. 67, no. 11, pp. 4626-4630, Nov. 2020, doi:10.1109/TED.2020.3020779

D. Bosch et al., 'All-Operation-Regime Characterization and Modeling of Drain Current Variability in Junctionless and Inversion-Mode FDSOI Transistors', 2020 IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 2020, pp. 1-2, doi: 10.1109/VLSITechnology18217.2020.9265036

T. Dubreuil et al., 'A novel 3D 1T1R RRAM architecture for memory-centric Hyperdimensional Computing', 2022 IEEE International Memory Workshop (IMW), Dresden, Germany, 2022, pp. 1-4, doi: 10.1109/IMW52921.2022.9779306

T. Dubreuil et al., 'Fabrication of Low-Power RRAM for Stateful Hyperdimensional Computing', 2023 International VLSI Symposium on Technology, Systems and Applications (VLSI-TSA/VLSI-DAT), HsinChu, Taiwan, 2023, pp. 1-2, doi: 10.1109/VLSI-TSA/VLSI-DAT57221.2023.10134182

T. Dubreuil et al., 'Integration of HfO2-based 3D OxRAM with GAA stacked-nanosheet transistor for high-density embedded memory', ESSDERC 2023 – IEEE 53rd European Solid-State Device Research Conference (ESSDERC), Lisbon, Portugal, 2023, pp. 117-120, doi: 10.1109/ESSDERC59256.2023.10268513

M. Ezzadeen et al., 'Low-Overhead Implementation of Binarized Neural Networks Employing Robust 2T2R Resistive RAM Bridges', ESSCIRC 2021 – IEEE 47th European Solid State Circuits Conference (ESSCIRC), Grenoble, France, 2021, pp. 83-86, doi: 10.1109/ESSCIRC53450.2021.9567742

M. Ezzadeen, et al., 'Implementation of binarized neural networks immune to device variation and voltage drop employing resistive random access memory bridges and capacitive neurons', Commun Eng 3, 80 (2024), doi.org/10.1038/s44172-024-00226-z

M. Ezzadeen et al., 'Binary ReRAM-based BNN first layer implementation', 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE), Antwerp, Belgium, 2023, doi: 10.23919/DATE56975.2023.10137057

With a growing intellectual property (IP) portfolio and a rapidly expanding team, the Barcelona-based chip-design company Semidynamics, a sponsor of HiPEAC 2025, is on a roll. HiPEAC caught up with Roger Espasa, the founder and chief executive of Semidynamics, to learn more about building a European hardware company from the ground up.

'We must create products that are useful for Europe, and the rest of the world, in Europe'

semidynamics

After 12 years at Intel and two at Broadcom, Roger Espasa decided it was time to launch his own venture. 'I'd had a few projects cancelled by high-level executives on the [US] West Coast while I slept, and decided I didn't want to do that anymore,' he says wryly. Semidynamics started life as a design service hub, initially providing designs to specifications requested by Esperanto Technologies. Realizing that the company should expand its customer base, and keen to work on their own designs, in 2019 the Semidynamics team started developing their own IP.

"Combining Atrevido or Avispado with the vector unit and tensor unit provides a powerful all-in-one solution that you can scale up or scale down"

'We spent two years designing our first IP, and we were the first to produce a RISC-V-based out-of-order core and an out-of-order vector unit. By that point we were keen to sell but we didn't even have a sales department; we were a company of engineers,' explains Roger. During 2022, a customer approached Semidynamics and 'took a leap of faith', according to Roger; by mid-2023, with the technology polished and the first customer secured, it was time to go to market. 'We hired a sales team, hired a marketing team, started going to conferences and having booths at trade shows. Our first trade show in China was something of a reality check, though, as people didn't know who we were,' Roger says.

Since then, customer feedback has been pivotal to shaping the direction of Semidynamics' technologies, according to Roger. 'Our customers have their own customers, and their requirements change over time, particularly in terms of increasing demands. Every year your solution has to get better.' In addition, unlike a large technology company, with large teams to absorb different aspects of the business, in a small company the chief executive is exposed to this necessity, he notes.

Atrevido (left) and Avispado (right)

All in one solution

Semidynamics' solutions include two RISC-V cores:

- **Atrevido** (Spanish for 'bold'): 64-bit 2/3/4-wide out-of-order RISC-V core
- **Avispado** (Spanish for 'witty'): 64-bit 2-wide in-order RISC-V core

These are complemented by a **vector unit**, which complies with the RISC-V International vector extension RVV 1.0. 'The vector unit is highly configurable: customers can choose the vector length

to be scaled from 128 bits to 4,096 bits, meaning that it can be deployed anywhere from the edge to high-performance computing,' Roger says. The company also recently announced a **tensor unit**, necessary to keep up with the matrix multiplications which form the basis of many artificial intelligence (AI) programs.

'Combining these three elements – Atrevido or Avispado plus the vector unit and tensor unit – provides a powerful all-in-one solution. Compared to having to acquire three different technologies from three different vendors, this offers several advantages,' says Roger. 'First, it massively simplifies things for the software team, who just have to work with one RISC-V based software stack, rather than being forced to get software stacks from different vendors to talk to one another. Second, it removes the need for direct memory access, as data doesn't need to be copied right and left, removing another burden from the software team. Finally, as data doesn't need to be copied from the tensor unit to the vector unit to the core, latency and power use are reduced.'

Customers who need significant computational muscle, for example for AI applications, can scale up the number of these all-in-one processing elements as required, up to hundreds of tera operations per second (TOPS), Roger adds. 'It's an all-in-one building block that you can scale up or scale down.'

The configurability and flexibility of the all-in-one IP means that customers can optimize their implementation and tailor it to their specific application needs, Roger notes, 'be it in medical image processing, security cameras and or deployments in the area of software-defined vehicles, to name a few'. 'Privacy and security are other aspects that become more and more important as the 'intelligence' really lies in the model parameters that are loaded into the all-in-one processor. Hence enabling private or confidential computing approaches via RISC-V features such as hypervisor support and crypto instructions is crucial as well as ensuring that the IP we deliver is robust against the latest attacks,' he adds.

The role of RISC-V and European funding

A high-profile figure within the RISC-V community, Roger is categorical about the central role of RISC-V in building Semidynamics. 'Without RISC-V, we would have been faced with two options: either to take out a proprietary IP licence, under which we would not have been able to invent instructions, or to invent something from scratch, including all the software – compilers, debuggers, libraries, etc.,' he says. 'RISC-V is an open-source standard that explicitly says you can change the standard in any way you want. So when customers come with a need for specific instructions that aren't currently available, you can respond to this request.' This ability to customize allows Semidynamics to offer what Roger calls 'open-core surgery', adapting their cores in response to customer requests.

Another crucial element in the early stages of Semidynamics' IP design was European funding. 'European-funded projects, including the European Processor Initiative (EPI) and EUPILOT, have allowed Semidynamics to develop our IP,' says Roger. 'They also allowed us to meet other partners in Europe and connect IPs within the European ecosystem.' One major benefit was the development of the test chips for the EPI, which Roger celebrates as a tangible output of a complex, joint research effort, meaning that European silicon can now be physically displayed at trade shows.

Towards a thriving European silicon ecosystem

The subject of technology transfer – specifically, how to ensure that more of the excellent research in Europe becomes commercialized – has been a key driver behind European Commission policy in recent years. Roger believes that, in order to take a product to market, an industry-led, focused model is the way forward, which should run in parallel to funding basic research at the lower technology readiness levels.

A lack of venture capital (VC), particularly in comparison to the United States, is also often cited as an issue for European companies wishing to scale. 'Silicon ventures are very capital intensive, often requiring tickets beyond European VCs' usual targets. At the same time, we have only a handful of silicon venture successes in Europe, reinforcing VCs' reluctance to invest in this area. We need silicon startups and VCs to grow in scale hand-in-hand,' Roger says. 'We also have to acknowledge the reality that we are not living in the US, and Europe is more fragmented – but the trend is towards greater integration.'

However it is achieved, for Roger the importance of technological sovereignty is clear. 'As an EU citizen, I see sovereignty as essential. Growth comes from creating value, which comes from innovation, new ideas, new technology. In Spain, we have painful memories of the saying "Let others invent!" [***"¡Que inventen ellos!"***, attributed to the Spanish writer Miguel de Unamuno and used to describe the marginality of science in Spain]. Giving up on sovereignty is as foolish as giving up on innovation.' He also notes that creating domestic technology gives Europe oversight over the technology and allows Europe to enforce its own rules: 'We must create products that are useful for Europe, and the rest of the world, in Europe.'

Semidynamics is hiring. Find out more on the company website:
semidynamics.com/en/hiring

SME snapshot

Offering design and intellectual property (IP) services for advanced nodes, the German company Racyics has delivered over 100 chip designs in the 15 years since it was established. Design Service Director Florian Bilstein tells us more.

Racyics, your go-to design partner

COMPANY: Racyics GmbH
MAIN BUSINESS: turnkey application-specific integrated circuit (ASIC) solutions, mixed-signal system-on-chip (SoC) design service, custom intellectual property (IP), foundry access, packaging services
KEYWORDS: ASIC, SoC, design partner, advanced nodes
LOCATION: Dresden, Frankfurt am Main and Duisburg
WEBSITE: racyics.com

Racyics® is Europe's leading design partner for mixed-signal SoC design and turnkey ASIC solutions in advanced nodes. With Germany-wide locations in Dresden, Frankfurt am Main and Duisburg, and over 130 employees, the company offers design services for analogue, mixed-signal, and digital integrated circuits (ICs) including foundry access, custom IP and turnkey ASIC solutions.

Having worked for leading semiconductor companies for many years, the Racyics team has contributed to more than 100 successful chip designs, down to 3nm feature size, for automotive, consumer and communication applications.

Racyics ABX Platform and makeChip

Over more than 15 years of dedication in SoC design and turnkey ASIC solutions, the Racyics team has always supported its customers with design enablement and unique IP offerings.

Body biasing is a disruptive 22FDX® feature enabling on-the-fly adaptation of transistor threshold voltages. One of Racyics key IP offerings is our ABX Platform, providing reliable and predictable ultra-low voltage (ULV) operation down to 0.4V, compensating process, supply voltage and temperature variations (PVT) to guarantee timing and power with high yield. For automotive applications, Racyics® ABX Platform enables significant leakage reduction at 0.8V high-temperature corners.

Since 2017, Racyics has also offered makeChip, a cloud-based design platform enabling startups, small / medium enterprises (SMEs) and academia to successfully design their own chip designs in a clone of Racyics professional design environment.

By combining the power of the ABX IP platform with the flexibility and accessibility of makeChip, Racyics offers a comprehensive ecosystem that caters to university and research institutes as well as emerging startups and established companies.

Valued partnerships

As an Arm Approved Design Partner, Racyics offers deep expertise and proven capabilities in designing and implementing Arm-based solutions. This partnership includes tailored customer solutions covering full turnkey solutions, frontend design, design verification, backend design, post-silicon validation and supply-chain management based on the latest Arm IP.

In addition, Racyics has been a GlobalFoundries Channel Partner® since 2013. This partnership builds on Racyics' ability to provide customers with access to GlobalFoundries' advanced manufacturing technologies, including the 12LP+ and 22FDX platform. As GlobalFoundries channel partner with a focus on advanced technologies, Racyics provides access to 28nm, 22nm and 12nm prototyping runs (MPWs).

Vision and goals: Shaping the future of semiconductors

Looking ahead, Racyics is participating in national and European research projects, for example in the field of cryoelectronic and edge AI applications.

In addition, Racyics supports current and future customers with access to our global and European supplier network. Depending on our customers' needs, we offer complete turnkey solutions including a true European supply chain (wafer, assembly and test).

Addressing the need for high yield and low cost for mid- and low-volume, high-end products, Racyics is offering chiplet package solutions based on organic substrates for heterogeneous technology integration. This will enable our customers to scale their system at low risk while developing IP on the best process node and giving them the possibility to reuse individual chiplets in different systems.

FURTHER INFORMATION

Racyics ABX Platform racyics.de/products/abx-ip-platform

makeChip makechip.design

Achieving optimal tradeoffs for neural networks involves a delicate balancing act between the model and the hardware. In this article, Francesco Ratto, Francesca Palumbo (both University of Cagliari) and Claudio Rubattu (University of Sassari) argue that adapting data precision at runtime using field-programmable gate arrays (FPGAs) is a promising way forward.

Want to fully exploit neural-network tradeoffs? Enable runtime-adaptive hardware

Without doubt, machine-learning applications are the main target for current hardware development. They offer many competing metrics, i.e. tradeoffs, to be exploited. The model's performance, e.g. the mean squared error for regression tasks or the accuracy for classification tasks, and the hardware performance, e.g. latency, throughput, and energy consumption, are two strongly correlated aspects.

As an example, the data precision that is adopted influences both the application performance and the achievable hardware performance. Reduced precisions tend to reduce the application performance while increasing the achievable hardware performance. When the data precision is converted from floating to fixed point we refer to it as quantization or binarization. Other techniques try to reduce the number of operations and parameters of a model by skipping unnecessary calculations, e.g. zero multiplications, or increasing the model sparsity. These techniques are called pruning techniques. Even in this case, more hardware-efficient computation is traded off with reduced model accuracy.

One possible way to tackle this problem is to design hardware in a way that we believe best optimizes these tradeoffs. However, this approach falls short in a dynamic environment when the constraints and priorities evolve over time. As an example, a video surveillance system may wish to increase model accuracy when a threat is detected or signalled by another source of information. This can be achieved by increasing the data precision at runtime, resulting in an increase in the model accuracy but also in higher energy consumption for the hardware that executes the inference of the model. Also, this adaptive behaviour is not easy to achieve: you need hardware support to efficiently switch among different precision, and you need to co-design the model and the hardware in a way that the reduced precision does not degrade performance.

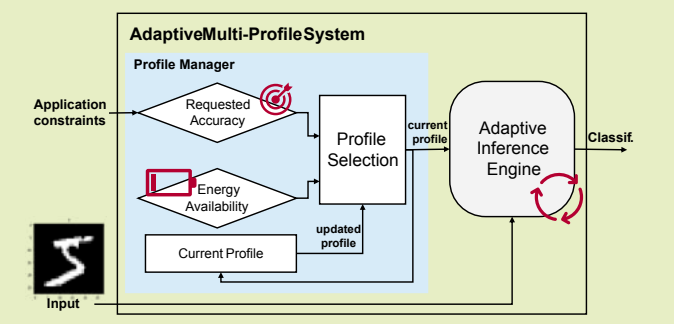
FPGAs offer an ideal computing platform to adapt the data precision at runtime and lend themselves to co-design solutions. Nevertheless, as the design space is wide, significant effort is required in developing design automation solutions that close the gap between the machine-learning model and its execution on FPGAs. This effort has already been made for non-adaptive solutions, with many mature frameworks that allow the designer to select the model precision at design time and then derive the corresponding inference engine. We believe that enabling runtime adaptivity constitutes a crucial step to fully exploit FPGAs' reconfiguration capabilities.

Supporting precision adaptivity on hardware opens a new set of research challenges. Once the working points have been selected and the adaptive support is available, a smart profile-selection mechanism must be put in place to optimize the performance of the system, e.g. minimizing the energy consumption, while meeting the variable accuracy requirements imposed by the application.

To conclude, we believe runtime-adaptive hardware is a promising solution to cope with evolving requirements and competing tradeoffs. To achieve this, we need not only reconfigurable hardware, but also effective design-automation frameworks and efficient runtime managers.

FURTHER READING:

F. Manca, F. Ratto and Francesca Palumbo, 'ONNX-to-Hardware Design Flow for Adaptive Neural-Network Inference on FPGAs' arxiv.org/pdf/2406.09078



System architecture to exploit a runtime adaptive inference engine

With escalating demands on hardware, developing hardware and software isolation is no longer an option, argues Giuliano Sisto (imec). In this article, he sets out why hardware designers and software developers need to embrace a 'shift left' mentality for optimal results.

Towards true hardware-software co-design

An EDA perspective

Our times are unequivocally characterized by an artificial intelligence (AI) boom, with large language models (LLM) becoming a well-established tool in countless fields and companies like NVIDIA skyrocketing to the top of this market-driven world.

However, it is well known that different paces of evolution have characterized the parallel development of hardware and software, a fundamental problem which the computer scientist Sara Hooker describes as the 'hardware lottery'. The term 'hardware lottery' refers to the fact that, irrespective of the scientific value of a research idea, the main factor determining its success is the compatibility of the idea with existing hardware resources. This stems from hardware and software being historically independent, rather than two complementary cogs of the same machine.

With this in mind, the current landscape of the semiconductor industry paints a worrying picture. While software is thriving due to a variety of novel applications becoming increasingly mature, hardware is struggling to provide the necessary support for those applications, due a variety of obstacles, above all the undeniable slowdown of Moore's Law. Guaranteeing competitive power and performance numbers while limiting area and cost (PPAC) with every transistor generation is proving to be extremely challenging. The need for close collaboration between hardware, semiconductor technology, and software is, therefore, more evident than ever.

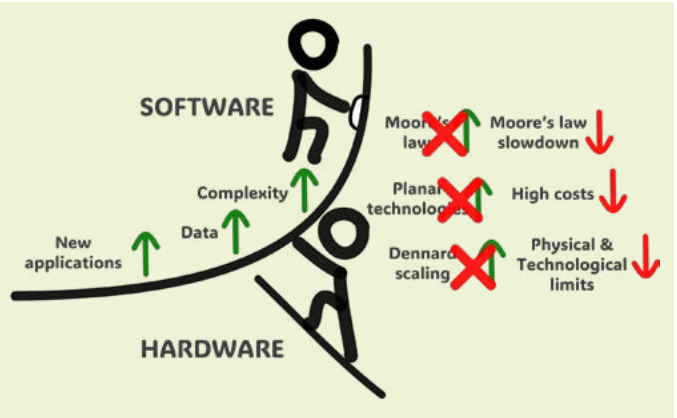
Bringing these fields together, however, is not an easy task. The cross-cutting, collaborative nature of this process is a source of many challenges, one of which is the need for an entirely new design ecosystem with tools to support it. For example, to support the hardware to become increasingly heterogeneous and specialized, holistic compiler and mapper tools are often required. This approach may lend itself to promote over-specialization for one application and hence offers limited reuse.

Advancing hardware and software in parallel

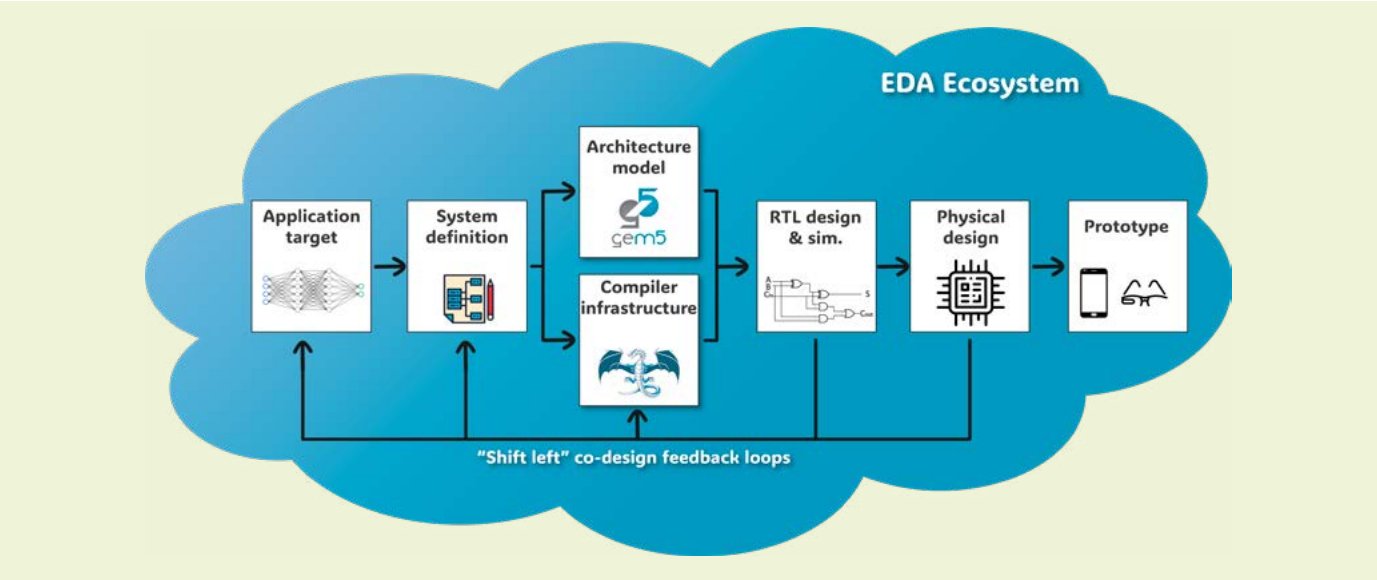
Specific choices at the workload level ripple down the entire integrated circuit (IC) design flow, influencing not only the programming model and hardware interface but also the microarchitecture of the computing system, as well as its physical implementation and packaging. Similarly, disruptive CMOS devices and technologies, such as complementary field-effect transistor (CFET), chiplets, and 3D integration, provide performance enhancements at the physical level, thus enabling unprecedented functionality exploitable at the software level. They also imply additional, non-trivial constraints on the mapping and execution of the program.

Going forward, the importance of joint improvements at distinct levels of abstraction is bound to grow even further, becoming key to the success of a research idea. Therefore, the inclusion of electronic design automation (EDA) tools at the intersection between hardware and software is essential for true co-design.

Examples of this combined optimization of technology and algorithms already exist both in research and industry; see for example the papers by Yang et al. and Wuu et al. cited in 'Further reading'. However, they often target a specific end-product, for which they also rely on significant manual



The previously independent tasks of software and hardware development now need to work together to achieve continued success



Visualization of the so called "shift left" design paradigm, in which physical hardware, software, and architecture are developed concurrently, rather than sequentially. EDA tools effectively act as glue among all these different stages

effort from the designers. The EDA ecosystem currently lacks an automated solution able to receive any workload and any microarchitecture definition as inputs and map the former on the latter taking into account its specific hardware resources.

Ensuring that all the high-level improvements are reflected in the physical design world requires the use of a variety of already existing independent tools. Additionally, completely new products, such as compiler infrastructure and architecture simulators like LLVM and gem5, are also required to bridge current gaps and provide a unified environment.

While new tooling is a key component to enable future research in the semiconductor industry, it also needs to be exploited through a paradigm shift in the way the digital IC design flow is perceived. Traditionally, register-transfer level (RTL) and physical implementations are seen as tasks whose goal is to meet certain specifications while factoring in physical limitations. To maintain PPAC scaling on the right track in the age of artificial intelligence (AI) and with Moore's Law winding down, a more collaborative mindset, through which technology, architecture, and software can mutually influence each other, is paramount. This new way of perceiving the design flow is sometimes referred to as 'shift left', referring to the integration of tasks traditionally performed at the end of the flow into earlier stages.

Large language models (LLMs) and chiplets are examples of modern algorithmic and technological innovations requiring closer collaboration between hardware and software, but they are far from being the only ones. A complete workload-to-architecture-to-technology co-design loop is likely going to be part of the future of the semiconductor industry, with EDA tools as key enabler and as the main means to prevent the hardware lottery from dictating computing-system design.

FURTHER READING:

S. Hooker, 'The hardware lottery', Communications of the ACM 64, 12 (December 2021), pp. 58-65, doi.org/10.1145/3467017
cacm.acm.org/research/the-hardware-lottery/

L. Yang, et al., 'Three-Dimensional Stacked Neural Network Accelerator Architectures for AR/VR Applications' in IEEE Micro, vol. 42, no. 06, pp. 116-124, 2022, doi: 10.1109/MM.2022.3202254
bit.ly/Yang_et_al_IEEE_Micro_2022

J. Wuu et al., '3D V-Cache: the Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU', 2022 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2022, pp. 428-429, doi: 10.1109/ISSCC42614.2022.9731565
ieeexplore.ieee.org/document/9731565

The LLVM compiler infrastructure llvm.org

The gem5 simulator system gem5.org

J. Ferguson, 'IC designers: let's talk about shift left strategies', Siemens, 28 September 2023
bit.ly/Siemens_EDA_IC_shift_left_blog

Innovation Europe

In this edition, we learn how NEUROPULS draws on photonics to deliver ultra-low-power acceleration, while QUADRATURE is overcoming quantum scalability issues by integrating quantum cores with wireless communication and control capabilities.

Funded by the European Union

OVERCOMING QUANTUM SCALABILITY CHALLENGES WITH QUADRATURE

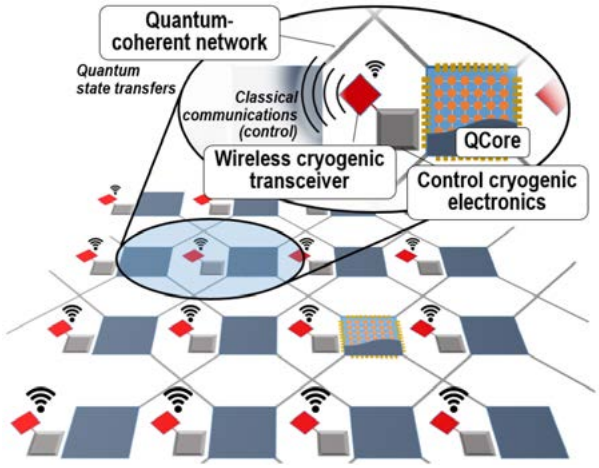
The QUADRATURE project aims to revolutionize quantum computing by addressing the scalability challenges of quantum processors. It seeks to create a quantum computing architecture that integrates coherently coupled multiple quantum cores (Qcores) with wireless communication and control capabilities, enabling unprecedented levels of optimization and efficiency in quantum computing systems.

The primary objectives of QUADRATURE include developing the first integrated all-radio frequency (RF) qubit state transfer link between multiple Qcores using a cryogenic superconducting cavity waveguide. This innovative approach will enable high fidelity and long coherence times, which are critical for scalable quantum computation. Another significant objective is the design and characterization of antennas that operate at microwave and mm-wave frequencies at deep-cryogenic temperatures. This effort includes developing the first-ever mm-wave wireless transceiver specifically for quantum computing applications.

In addition to these technical advancements, QUADRATURE seeks to innovate wireless networking within the quantum processor package to achieve extreme levels of optimization. This involves leveraging the static and noise-free environment of the quantum system to facilitate highly efficient communication between Qcores. Furthermore, the project proposes the development of a flexible multi-Qcore platform capable of dynamic reconfiguration. This adaptability will allow the system to meet various algorithmic requirements, thereby enhancing the overall efficiency of quantum computations.

The work plan and methodology of QUADRATURE follow a virtual-loop methodology encompassing several work packages. Key components include the development of quantum-coherent cavity channels, the design and implementation of cryogenic antennas, and the establishment of wireless networking solutions within quantum packages. Additionally, the project focuses on creating reconfigurable quantum architectures and conducting full-stack simulation and design-space exploration.

Drawing on a combination of cutting-edge research and cross-disciplinary expertise, QUADRATURE is hence tackling key scalability bottlenecks in quantum computing. By addressing the fundamental challenges of scalability and integrating multiple Qcores with advanced wireless communication capabilities, QUADRATURE aims to pave the way for scalable, high-performance quantum systems. This could lead to significant advancements in scientific and technical fields, enabling the resolution of real-world problems that are currently intractable with conventional computers.



PROJECT NAME: QUADRATURE: Scalable multi-chip quantum architectures enabled by cryogenic wireless / quantum-coherent network-in-package
START / END DATE: 01/06/2023 - 31/05/2027
KEY THEMES: scalable quantum computing systems, wireless network-on-chip, quantum-coherent communication, cryo-CMOS, quantum computing architectures, quantum algorithms
PARTNERS: **Spain**: Universitat Politècnica de València (coordinator), Universitat Politècnica de Catalunya, Barcelona Supercomputing Center; **Netherlands**: Delft University of Technology; Germany: University of Siegen; **Italy**: University of Catania; **Ireland**: Equal1 Labs, University College Dublin; **Switzerland**: EPFL.
BUDGET: €3,420,513.75
quadrature-project.eu
linkedin.com/company/quadrature-eu
x.com/QUADRATURE_EU

QUADRATURE has received funding from the European Union's Horizon Pathfinder programme, under grant agreement no. 101099697.

NEUROPULS: SILICON PHOTONICS TO ACCELERATE MACHINE LEARNING, FROM MATERIALS TO SYSTEMS

Silicon photonics (SiPh) is a rapidly maturing CMOS-compatible technology that allows us to realize chip-scale optical circuits that vastly outperform conventional free-space and discrete optics in terms of size, weight, and power (SWaP) metrics. While the conventional domain of photonics is communications, SiPh enables us to bring the benefits of ultra-high bandwidth, minimal latency, and low-loss signal propagation without Joule dissipation or impedance matching constraints to on-chip interconnects and networks-on-chip (NoCs). Over recent years, advances in SiPh technology have led to emergence of optical computing hardware, where information, encoded onto light beams by high-speed optical devices, can be both transferred and processed while benefitting from the aforementioned advantages. Furthermore, new approaches such as in-memory or neuromorphic computing can also be efficiently realized in photonics.

The NEUROPULS project aims to develop ultra-low-power photonic hardware for efficient acceleration of machine learning workloads, addressing edge computing applications. High computing throughput and energy efficiency are enabled by using application-specific integrated SiPh circuits as the accelerator core of the architecture.

In NEUROPULS, a standard silicon-on-insulator (SOI) integrated photonic platform is further augmented with III/V materials and optical phase-change materials (PCMs). This augmented technology stack allows us to realize both spiking laser-based neurons (using III-V materials) and non-volatile optical memory elements that significantly reduce the accelerator power requirements (thanks to their non-volatile character) while also allowing for more unconventional learning methods based on emergent synaptic plasticity. More specifically, the SiPh integrated circuit enables highly efficient matrix-vector operations acceleration with ultra-low latency and PCM-minimized power consumption.

Furthermore, SiPh will be leveraged to implement hardware security layers, using primitives such as physical unclonable functions (PUFs) to further harden the computing system against several types of malicious attacks, including advanced machine learning (ML)-based modelling attacks.

NEUROPULS explores the proposed electronic-photonic hardware approach from the bottom up, i.e. by carrying out basic research at a material and device level up to the system level by building a fully functional prototype. To aid in the development of the accelerator and to investigate tradeoffs while taking into account both the electronic interfaces and photonic hardware, system-level simulators based on the open-source gem5 framework are being developed in the project. These simulators integrate a custom-made extension called gem5-MARVEL that enables support for RISC-V instruction set architectures (ISAs) to support open-source technology interfacing with the photonic hardware.

The NEUROPULS accelerator will be benchmarked with three distinct, edge-focused use cases, namely:
1) anti-jamming of global navigation satellite systems,
2) anomaly detection in the scope of network security, and
3) trajectory prediction in the context of autonomous vehicles.

Functional and system-level modelling as well as hardware-software co-design are pursued to optimize and maximize the usage of the photonic core and harness its high performance for the selected use cases. We believe this comprehensive system-level implementation in the NEUROPULS project is key to demonstrating a practical, photonic neuromorphic accelerator.

PROJECT NAME: NEUROPULS: NEUROmorphic energy-efficient secure accelerators based on Phase change materials aUgmented siLicon photonicS
START/END DATE: 1/1/2023 – 31/12/2026
KEY THEMES: neuromorphic computing, integrated photonics, security, simulations, RISC-V, CMOS-compatible platforms
PARTNERS: **France**: Centre nacional de la recherche scientifique (CNRS) (coordinator), Ecole Centrale de Lyon (ECL), Université de Bourgogne, Commissariat à l'Energie atomique et aux energies alternatives (CEA); **Belgium**: Ghent University, Hewlett Packard Enterprise (HPE); **Czechia**: Argotech; **Germany**: Ludwig-Maximilians-Universität Muenchen (LMU), Technische Universität Berlin (TUB); **Greece**: National and Kapodistrian University of Athens; **Italy**: Politecnico di Torino, Universita degli Studi di Verona; **Portugal**: Instituto de Engenharia de Sistemas e Computadores, Investigacao e Desenvolvimento Em Lisboa (INESC-ID); Spain: Barcelona Supercomputing Center (BSC), Albora Technologies
BUDGET: €8.3 million

NEUROPULS

neuropuls.eu
linkedin.com/company/92966082
x.com/neuropuls
Coordinator: Fabio Pavanello
fabio.pavanello@cnrs.fr

NEUROPULS has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101070238.

Interested in experimenting with different platforms and expanding your skillset? There are a range of university outreach programmes and courses available to help you get up to speed on the latest hardware. In this article, we find out about initiatives from Arm, AMD and RISC-V.

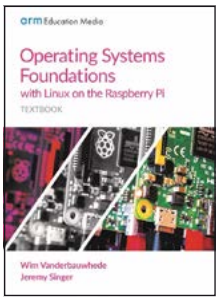
Expand your hardware horizons

University outreach programmes, courses and more

University outreach activities at Arm

Kieran Hejmadi, Software and Academic Ecosystems Manager, Arm

Arm's university outreach programme has a strong focus on system-on-chip (SoC) education, providing students and educators with access to advanced tools and resources to develop their skills in semiconductor and computing technologies. A key component of this outreach is Arm's partnership with hardware providers, like Raspberry Pi, to create practical courses that incorporate real-world tools into higher education curricula. For example, the free-of-charge *Operating Systems Foundations* textbook, authored by HiPEAC members Wim Vanderbauwhede and Jeremy Singer, uses Raspberry Pi boards to teach students about embedded hardware and software integration, giving them hands-on experience with Arm-based microcontrollers and development environments.



Arm's Academic Access programme further supports research-driven projects by providing approved universities with the intellectual property (IP) needed to design custom accelerators and other advanced hardware projects free-of-charge. For example, the University of Southampton has leveraged Arm's IP to develop their own research accelerators and created a community of practice, SoCLabs, that aims to foster a network of other universities creating SoC design around Arm IP.

In addition to physical development boards, Arm provides access to fixed virtual platforms (FVPs), which are simulation tools that model Arm hardware with high accuracy. This allows students to design, code, and test their projects on virtual models of Arm processors, which is particularly valuable for institutions that want to benchmark their application across different microarchitectures without being tied to hardware. FVPs are widely used in university programmes to teach concepts like computer architecture, operating systems, and hardware-software co-design.

These efforts are supported by Arm's educational courses on EdX, which include modules on hardware design, system-on-chip (SoC) development, and digital signal processing, ensuring that students receive comprehensive instruction on the latest hardware technologies. Overall, Arm's hardware-focused university outreach initiatives provide essential tools, resources, and industry partnerships to foster innovation and practical learning in semiconductor technologies.

An Arm Insight Day in Warwick, UK. The event targets those who are about to go to university

FURTHER INFORMATION:

Arm Academic Access bit.ly/Arm_Academic_Access

Arm Education courses on EdX edx.org/school/armeducation

W. Vanderbauwhede and J. Singer, *Operating Systems Foundations with Linux on the Raspberry Pi*
arm.com/resources/ebook/operating-systems-ty

SoCLabs soclabs.org

AMD University Program: Empowering the next generation of innovators

Cathal McCabe, AUP Director Americas & EMEA, AMD Ireland

At AMD, we are leaders in high-performance and adaptive computing, driving the products and services that address the world's most critical challenges. Our technologies advance the future of data centres, embedded systems and edge artificial intelligence (AI), gaming, and PCs.

The AMD University Program (AUP) is building that future by equipping the next generation of innovators with the tools and skill they'll need to succeed. We are part of the Research and Advanced Development Team (RAD) and bridge the gap between academic research and the AMD research team.

The AUP provides access to AMD technologies and resources via our donation program and AI & HPC Fund, along with research support and opportunities for grant funding. Researchers can request AMD hardware, such as central processing units (CPUs), graphics processing units (GPUs), field-programmable gate arrays (FPGAs) and adaptive system-on-chips (SoCs), and neural processing units (NPUs) to enhance their work. Our AI & HPC Fund offers compute time on our academic HPC cluster, providing global researchers with the resources and leading-edge compute power to tackle the world's hardest problems.

Our Heterogeneous Accelerated Compute Clusters (HACC) research initiative drives advancements in adaptive computing and high-performance computing. These cutting-edge clusters, equipped with the latest AMD hardware, enable researchers at esteemed institutions and universities around the world to explore novel areas of heterogeneous computing.

To help researchers stay ahead in the rapidly evolving tech landscape, we also offer free in-person and virtual training sessions. These sessions cover topics like AI, high-performance computing (HPC), and FPGA/SoC development, ensuring researchers have the knowledge to leverage our latest innovations. In addition, AUP provides educational resources, teaching materials, and courseware to support the development of advanced curricula in universities.

We invite early stage HIPEAC researchers to contact us and discuss how AUP can support your research needs and advance your projects. We are particularly interested to hear from researchers interested in working with AI PCs and NPUs.

Contact: Cathal McCabe, AUP Director Americas & EMEA, AMD Ireland
cathal.mccabe@amd.com

FURTHER INFORMATION:

AMD University Program amd.com/AUP
AMD AI & HPC Fund amd.com/en/corporate/hpc-fund.html
AMD Heterogeneous Accelerated Compute Clusters amd-haccs.io

RISC-V opportunities to learn and contribute

Megan Lehn, Community Director, RISC-V International

RISC-V International offers many ways for students and the community to get educated from home and become active participants in the open standard.

Free courses are taught by RISC-V experts on a wide array of subjects, including the updated 'Introduction to RISC-V' with community-driven improvements, the hands-on 'Computer Architecture with an Industrial RISC-V Core [RVfpga]', and our most popular course, 'Building a RISC-V CPU Core', plus many more. Whether you're a software developer or a hardware enthusiast, you can build your expertise at your own pace and gain a competitive edge.

After taking courses, you can showcase that you've acquired the skills needed to ensure peak performance in RISC-V. The RISC-V Foundational Associate Certification (RVFA), along with the RISC-V Fundamentals Course, were developed by experts from the community to help you succeed and stand out in this rapidly expanding ecosystem.

RISC-V also offers a paid mentorship programme, with applications opening on 9 January 2025. Mentorships run three times a year with paid opportunities in spring, summer and fall. The mentorship programme is an opportunity to solve real-world problems with the help of RISC-V mentors available to answer questions and provide guidance. See the 2023 Mentorship Showcase in 'Further information', below, for mentees' testimonials.

Questions about RISC-V education and training opportunities? Contact the RISC-V team: learn@riscv.org

FIND OUT MORE:

RISC-V certifications and courses riscv.org/certifications-and-courses
RISC-V Foundational Associate bit.ly/RISC-V_Foundational_Associate
RISC-V mentorship programme riscv.org/risc-v-mentorship-program/
RISC-V Mentorship Showcase - 2023 Projects
bit.ly/RISC-V_Mentorship_Showcase_2023
RISC-V careers fairs riscv.org/career-fairs

Since 2016, HiPEAC Jobs has been organizing talks, information sessions, roundtables and more for those seeking careers guidance. In this article, HiPEAC Jobs' Laura Menéndez (Barcelona Supercomputing Center) briefs us on recent and future activities for anyone thinking about their career options.

Considering your next step? HiPEAC Jobs has got your back

With a wealth of careers opportunities available to computer science and computer engineering graduates, it can be hard to work out which path to take. In computing systems research, your career may take you to academia but also to industry, or even set you on the path of launching your own company.

For some years now, HiPEAC Jobs has been offering activities to help you decide on the next step in your career. The HiPEAC Jobs online portal is a snapshot into the variety of full-time positions, PhD options and internships in the field, while its physical counterpart, the HiPEAC jobs wall, displays these opportunities at HiPEAC events and external events such as womENcourage or local hackathons. At the HiPEAC conference and sister events such as DATE, the HiPEAC virtual jobs fair allows you to upload your CV prior to the event, and even arrange interviews in advance.

The STEM Student Day at the HiPEAC conference is a great opportunity to meet people from top research institutions and companies large and small. At the interactive HiPEAC Jobs 'Inspiring Futures' sessions you can learn firsthand from people working in industry, academia and innovation, while the HiPEAC Jobs careers roundtables are a chance to hear personal stories of career journeys.

HiPEAC Jobs also offers numerous opportunities to develop as a researcher. These include the HiPEAC Student Challenge, a great opportunity for undergraduate, master's and early PhD students to flex their coding muscles and tackle the problems they care about most. The poster sessions at the HiPEAC conference and particularly at the HiPEAC summer school, ACACES, are a great opportunity to share your research with your peers. Meanwhile, this year, the inaugural 'Coffee and Papers' sessions at ACACES provided the chance to discuss scientific papers with fellow researchers.

Check out the HiPEAC Jobs portal for more information:
hipeac.net/jobs/#/recruitment

Machine learning (ML) at the edge is a strategic research area for Europe. In this article, Moritz Scherer (ETH Zurich / Mosaic SoC) explains how his PhD took tinyML energy efficiency to new levels and laid the foundation for a start-up.

Three-minute thesis

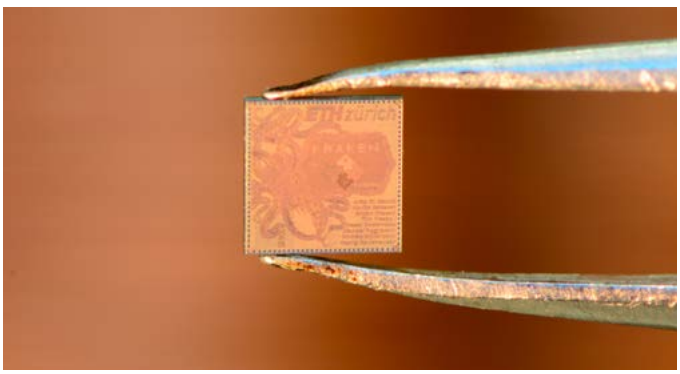
NAME: Moritz Scherer
RESEARCH CENTRE: ETH Zurich
SUPERVISOR: Luca Benini
DATE DEFENDED: 08/07/2024
THESIS TITLE: Hardware-Software Co-Design for Energy-Efficient Neural Network Inference at the Extreme Edge

Advancements in machine learning (ML) in the past decade have opened the door to extracting high-level information from embedded sensor data. Applications for such high-level information range from using low-power, low-resolution embedded cameras for distributed surveillance to driving the augmented-reality user experience through gesture and voice recognition.

While processing these increasingly complex algorithms on microcontroller-class devices directly at the data source promises to reduce their energy cost and latency massively, embedded deployment remains a challenging problem, requiring innovation across the hardware-software stack.

My thesis tackles the challenges of near-sensor energy-efficient 'tiny machine learning' from all angles. The first part of my thesis studies how ML algorithms can be put on a diet to fit the computing and memory resources available on microcontrollers. We developed several broadly applicable recipes for audio and gesture recognition on microcontrollers, achieving highly accurate predictions with models using only a few hundred kilobytes.

The second part of my thesis studies solutions to enable the deployment of ML algorithms on devices with tiny power budgets. We propose a convolutional neural network accelerator, CUTIE, that improves the energy efficiency achievable with microcontrollers by several orders of magnitude to 1 petaop/s/W by leveraging aggressive ternary neural network quantization techniques and designing the datapath for maximum data reuse. We integrated this in a full system-on-chip, Kraken, which has been integrated in several demonstrator systems.



The manufactured Kraken chip including the CUTIE accelerator

The third and last part of my thesis bridges the gap between innovations in ML algorithms and computer architecture by developing a specialized compiler framework for embedded ML applications called Deeploy. Rather than lowering code into pre-determined building blocks, Deeploy lets users control which abstractions best fit their hardware, while coordinating memory allocation and data movement under the hood. We demonstrated the viability of this approach by achieving state-of-the-art results on various benchmarks, and by executing cutting-edge transformer algorithms on Siracusa, a microcontroller featuring nine cores and a state-of-the-art convolutional neural network accelerator, using on-chip memory exclusively.

The start-up Mosaic SoC, which I founded with my colleague Alfio Di Mauro, aims to combine these innovations in a single device. We strive to develop microcontrollers optimized for the most demanding tinyML scenarios, focusing on augmented reality applications.

Moritz' PhD supervisor, **Luca Benini**, commented: 'Moritz' work addresses key challenges in the domain of tiny ML and proposes new designs and optimizations in the hardware and software domains. His research spans across the machine-learning stack, and the insights of his work are highly practical. For instance, his work on ternary neural networks (TNNs), and more specifically, the CUTIE TNN accelerator, has achieved 1fJ/ternaryop (corresponding to 1ternaryPOPS/W), a record still not broken. All results achieved in Moritz' thesis are thoroughly validated in silicon, and his work is available open source for the research community to build upon.'

HiPEAC

HiPEAC conference

20-22 January 2025

Barcelona

artificial intelligence · neuromorphic computing · processors · accelerators · next-generation communications · memory · edge · cloud · internet of things · high-performance computing · cyber-physical systems · energy efficiency · cybersecurity · robustness … and much more

hipeac.net/conference

#HiPEAC25

Funded by the European Union