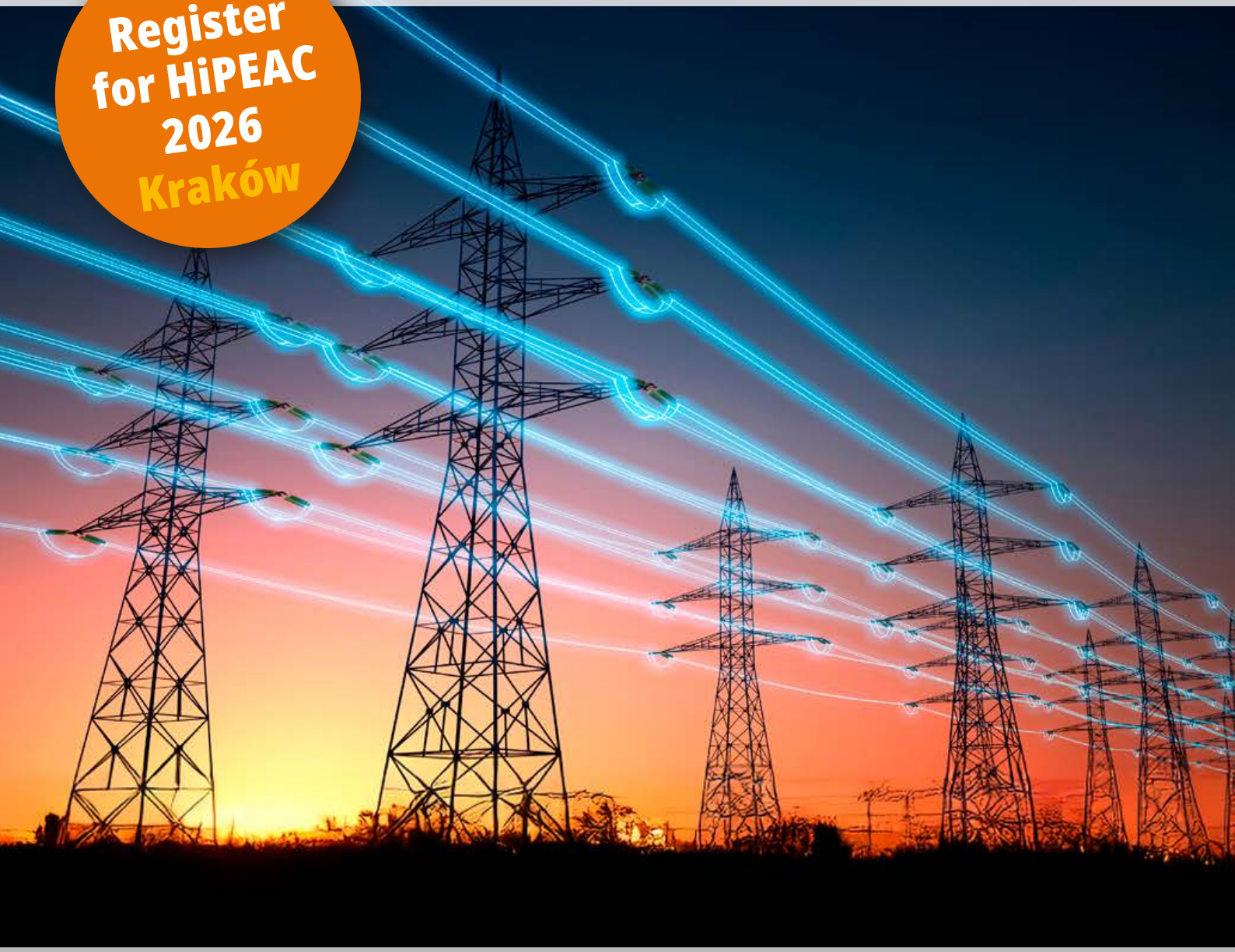
HiPEAC info 76

OCTOBER 2025

Register for HiPEAC 2026 Kraków



How distributed computing, AI and the IoT are driving the energy transformation

Future-proofing the electrical grid with digital twins

Why energy efficiency should be at the heart of computing-system design

Contents

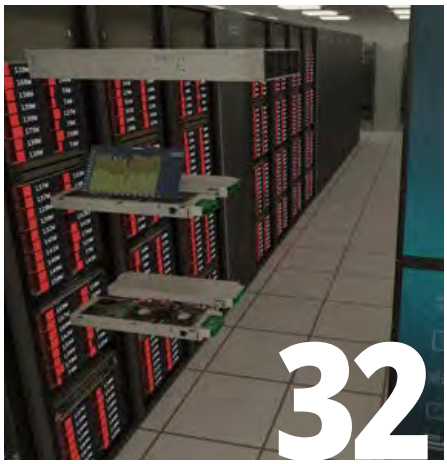
EU distributed-computing initiatives powering the energy transformation

New companies, institutes and funding rounds

Harnessing cloud-edge-IoT systems for energy

How digital twins and swarm intelligence can future-proof the grid

Leveraging the IoT for home-energy management with gridX



MERIC: A decade of energy efficiency in HPC

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: Peter Schreiber Media | stock.adobe.com

Design: www.magelaan.be

Editor: Madeleine Gray

The central theme of this magazine is energy. Energy is very well understood in terms of scientific laws, measurement, and calculation. However, the fundamental nature of energy remains an open question: while we can observe its effects, such as heat and motion, energy itself is invisible to direct observation. Understanding the essence of energy is crucial for interpreting phenomena beyond the human scale, particularly in cosmology and quantum mechanics, where the known laws are insufficient to explain certain observations – prompting theories such as dark energy and the cosmological constant.

In today's world, the significance of energy cannot be overstated. Modern life depends entirely upon reliable, abundant, and affordable energy. Globally, fossil fuels constitute the dominant energy source, but their use must be phased out to halt climate change. Solar and wind power are better for the climate; they are abundant and affordable, but not reliable due to their intermittent nature. Since supply and demand of electricity must always be balanced on the grid, the increasing use of intermittent energy sources forms a difficult technical challenge. Addressing this requires advanced digital technologies in combination with innovative energy-storage solutions.

Observations across Europe illustrate this point: on sunny days, wind turbines are often shut down because of peak solar production. This situation is neither ecologically nor economically justifiable, as cheap renewable energy is wasted at noon while more polluting and expensive energy sources need to be fired up in the evenings. How nice would it be if we could somehow use this energy: by temporarily storing it, for example in car / home batteries, in hydrogen, in pumped storage, … or by using it to extract CO_2 from the air to offset the emissions caused by the fossil fuels needed to generate electricity when renewables are in short supply. The latter could make the electricity production 100% net zero over a period of, for example, one year. Direct extraction and sequestration of CO_2 would in any case be a more effective way to offset emissions than planting trees because it removes and permanently stores the CO_2 when it is generated, not over the lifetime of trees (to release it again when the tree dies).

Hence, there is plenty of room for innovation in the energy sector. This magazine highlights the latest research within the HiPEAC community on computing systems that serve the energy sector. I hope you will find the content both informative and inspiring.

Koen De Bosschere, HiPEAC coordinator

The role of distributed computing and decentralized intelligence in driving the energy transformation

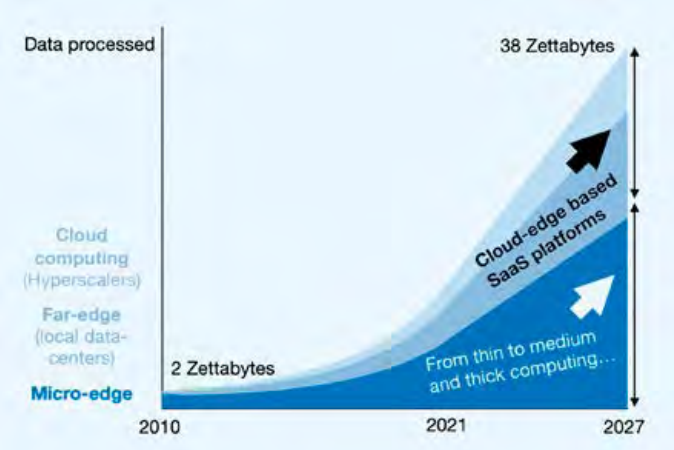
As the global energy sector undergoes a profound transformation towards decarbonization, digitalization, and decentralization, the way we manage, optimize, and secure energy systems is being reshaped. In this article, Rolf Riemenschneider, (Head of Sector IoT, European Commission) explores the implications of this shift, the role new developments in distributed computing and artificial intelligence (AI) can play, and opportunities for funded research to further this vision.

Traditional centralized control models are no longer sufficient to handle the complexity of modern grids, which must integrate renewable sources, electric vehicles, local storage systems, heat pumps and dynamic consumer behaviour. The recent blackout in Spain and Portugal showed that central and rigid control systems cannot deal with fluctuations and disturbances of increasing volatile energy assets.

In addition, today's energy infrastructure is reaching its capacity limits and often operates close to saturation. This results in delays in accommodating new connection requests from solar and wind plants, heat pumps, electric vehicle (EV) charging infrastructure, and industrial sites. Distributed computing and decentralized intelligence are emerging as the backbone of Europe's future decentralized energy system, enabling more resilient, efficient, and adaptive energy systems.

From centralized control to distributed intelligence

Historically, energy systems were designed around a centralized architecture: large-scale power plants generated electricity, which was transmitted and distributed to passive consumers and provided inertia to stabilize the grid. Decision-making and control were concentrated at the top of the hierarchy, with little or no autonomy at the local level.



Credit: DECISION Étude & Conseil, APL Datacenter, 2023 (see 'Further information')

Today, however, the rapid adoption of distributed energy resources (DERs) – such as rooftop solar, community batteries, and microgrids – has fragmented the grid into a highly dynamic ecosystem involving multiple actors and stakeholders. This shift demands new computational and organizational approaches in which computing is distributed closer to where decisions must be made.

Edge computing can be defined as 'the practice of processing data near the source of generation, rather than relying on a centralized data processing cloud infrastructure'. According to a recent study by DECISION (see 'Further information', below), Europe has a major opportunity in industrial internet-of-things (IoT) edge computing, projected to generate nearly €88 billion in global growth by 2027.

European initiatives: MetaOS for energy

The MetaOS project initiative supports the implementation of the European Strategy for Data across the IoT, edge, and cloud continuum. MetaOS research and innovation projects, supported under Horizon Europe with a total of €60 million in funding, have spearheaded the implementation of novel edge-computing paradigms and have established a technology baseline for emerging trends like AI-powered IoT and decentralized intelligence.

The notion of MetaOS emerged as a new concept for software-defined systems. In contrast to a standard operating system (OS), a MetaOS enables an abstraction from underlying computing and physical resources (like electronic control units) and hides the complexity of managing low-level control functions. To capitalize on the new edge paradigm and the next wave of innovation, the MetaOS cluster has brought together the development of IoT-edge nodes with the deployment of next-generation computing components, systems and platforms.

These advances enable the transition to a compute continuum with strong capacities at the edge and far edge, delivered in an energy-efficient and trustworthy manner.

An EnergyOS concept has been developed and predominantly deployed at the edge of the power grid to meet the needs of smart-grid users and achieve secure, efficient, and low-latency data processing. Several companies are now commercially exploiting products based on these developments, for example in substations that serve as critical components of distribution grids. Edge computing introduces local computing resources at the substation and field-asset level in an electrical grid. It is about virtualizing system functions and moving from closed, proprietary hardware, to an open, secure, interoperable software-defined platform, on which utility companies can build their own use cases (to find out more, see the EUCloudEdgeIoT paper in 'Further information').

A digital backbone building on emerging energy-efficient AI models

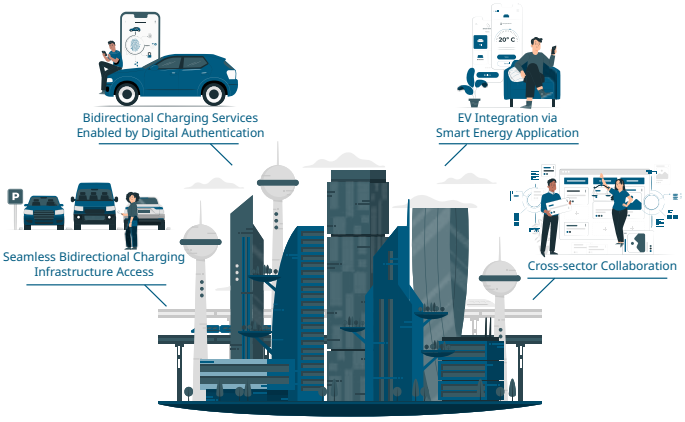
At the heart of the energy transition lies the integration of highly volatile dynamic energy assets like solar, wind, EVs, heat pumps, buildings and other flexible loads, which extend beyond the energy domain. Future energy system operators increasingly need to act as orchestrators of interconnected ecosystems, seamlessly integrating energy flows and managing flexibility of distributed energy resources to manage demand and response.

Asset orchestration goes beyond simple responsiveness and takes into account the dynamic collective behaviour of groups of distributed energy assets. Achieving real-time orchestration requires highly efficient AI models, decentralized computing and AI integrated at the grid edge. Such decentralized, energy-efficient AI is emerging as a digital backbone for transformative change – enabling systems that can anticipate, adapt, and act in real time. This leap in autonomy equips energy companies to manage complexity at scale, from volatile markets to distributed energy systems.

The European Commission (DG CNECT) supports the development of an AI-powered digital spine as part of the Horizon Europe Work Programme 2025 (deadline: January 2026). This aims to manage the uncertainty of volatile energy flows and enable smart energy operations of energy assets such as heat pumps, electric vehicles, home batteries, and solar generation etc.

Future outlook

Distributed computing and decentralized intelligence are pivotal in addressing the energy sector's challenges – rising demand,



grid reliability and resilience, and renewable integration. Agentic AI technologies have the potential to address the energy trilemma of securing supply, reducing emissions, and mitigating infrastructure costs. As energy systems evolve into more participatory, decentralized networks, distributed computing and decentralized intelligence will be essential. Ultimately, the vision is a self-organizing, intelligent energy ecosystem in which millions of actors – consumers, devices, and algorithms – collaborate seamlessly to deliver clean, affordable, and reliable energy for all.

The transition to distributed computing and decentralized intelligence represents more than just a technological shift; it is a structural rethinking of how we design and operate energy systems, and it also embraces innovation from adjacent sectors. Establishing a digital backbone for the energy system will enable the distribution of both computing power and intelligence across the network, and will provide a quantum leap in the digitalization of energy systems, making them adaptable to the needs of the 21st century. Support is provided under Horizon Europe for the development of an AI-powered 'digital spine' as a key element of a digital backbone, expected to break down existing silos between the energy and mobility sectors, creating a seamless, interconnected system.

FURTHER INFORMATION:

DECISION, 2023: *Study on the economic potential of far edge computing in the future smart Internet of Things*
https://bit.ly/DECISION_IoT_Edge_2023

EUCloudEdgeIoT, 2024: *Market Pathways for Cloud Edge IoT in Energy*
zenodo.org/records/13820261

Fraunhofer FIT, 2024: *Leveraging Twin Transformation: Digital Infrastructures to Advance Decarbonisation at the Nexus of Energy and Mobility*
bit.ly/Fraunhofer_FIT_Twin_Transformation_2024

North star for computing-systems navigators: ACACES 2025

'My supervisor told me: "You should attend ACACES once during your PhD." This year, I finally made it and it was truly worth it' *– ACACES 2025 attendee Bijin Elsa Baby*

Computer science and engineering – always fast-changing fields – are moving at a dizzying pace thanks to the disruptive influence of artificial intelligence (AI). This edition of HiPEAC's much-loved summer school, ACACES, helped participants keep up with the latest developments necessary to advance in the field, while reiterating the fundamentals of robust, reliable, trustworthy, and sustainable computing systems.

Attendees were treated to inspirational lectures from world-class computing experts including Ahmad-Reza Sadeghi (TU Darmstadt), Subhasish Mitra (Stanford University) and Timothy Roscoe (ETH Zürich). Topics included boundary-pushing research areas such as quantum computing with Poulami Das (University of Texas at Austin), dataflow modelling with Shuvra Bhattacharyya (University of Maryland) and prompt engineering with Maximilian Moundas (Vanderbilt University). The summer school also included in-depth classes on the transversal topics of entrepreneurship (with Wim De Wispelaere, Ghent University) and sustainability (with Adrian Friday, Lancaster University).

As in 2024, the 2025 edition of ACACES included a DISCOVER-US track focusing on topics related to distributed computing and swarm intelligence. Nicola Ferrier (Argonne National Laboratory–ANL) walked students through the SAGE cyber-infrastructure linking edge AI to high-performance computing (HPC) facilities, while Kate Keahey and Michael Sherman (ANL / University of Chicago) demonstrated how the Chameleon computing testbed is being used to power experiments in areas from network fingerprinting to autonomous vehicles. Irem Boybat (IBM Research) guided participants through in-memory computing, while Davide Schiavone (EPFL / OpenHW Group) showed how RISC-V accelerators can be deployed at the edge.

Keynote talks by Jan Rabaey (UC Berkeley) and Fabio Violante (Arduino) delved into the evolution of the computing continuum and the story of how Arduino grew to become a household name, respectively.

As usual, the summer school provided a host of activities aimed at providing students with personal and careers development. The dynamic poster session allowed them to share their research with peers and senior researchers, while the careers night and HiPEAC Jobs Wall gave insights into possible careers paths.

Videos from ACACES, including lecture videos of selected courses, can be viewed on the HiPEAC YouTube channel bit.ly/ACACES25_videos

Pumped for HPC-powered AI: PUMPS+AI

The Programming and Tuning Massively Parallel Systems + Artificial Intelligence summer school (PUMPS+AI) provides essential training for anyone interested in programming advanced, complex systems based on powerful processors such as graphics processing units (GPUs).

Organized by Barcelona Supercomputing Center, the Universitat Politècnica de Catalunya-Barcelona Tech, the University of Illinois at Urbana-Champaign and HiPEAC, the 15th edition of the summer school took place in Barcelona on 14-18 July. This year, attendees got a sneak peak of new material to be released in the upcoming fifth edition of the textbook *Programming Massively Parallel Processors*. Topics covered included large language models, matrix-multiplication optimization techniques, wavefront algorithms, and the latest multi-GPU libraries.

'The AI Factories initiative is an example of how high-performance computing (HPC) can power incredible AI applications,' commented local organizer and HiPEAC member Antonio J. Peña (BSC). 'At PUMPS+AI, our expert tutors guide participants to harness the immense computing power of GPUs, putting an array of powerful tools at their disposal.'

As usual, PUMPS+AI lectures were given by NVIDIA's Wen-mei Hwu and Juan Gómez-Luna, along with Izzat el Hajj (American University of Beirut) and local lecturers Antonio J. Peña, Marc Jordà, 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 in previous years, the 2025 edition of PUMPS+AI featured a HiPEAC Jobs careers session, coordinated by HiPEAC Jobs' Laura Menéndez Gorina (BSC).

PUMPS+AI will return to Barcelona in 2026. Places are limited, and those wishing to attend are advised to apply early. Check out the website for further information: pumps.bsc.es

TechNexus Interim Communities Exchange

On 8 October, TechNexus, a value network catalysed by HiPEAC, held its Interim Communities Exchange, one of a series of workshops focusing on the cross-disciplinary topic of technology integration for complex and critical systems.

Without a dedicated focus on the integration and uptake of new technologies by complex systems, such as in national infrastructure, the capacity to use such new technologies is limited. In addition, increasing inter-connectivity brings new risks, particularly relating to safety and security, in the absence of holistic approaches.

By building an interaction point and supporting research, TechNexus aims to help industry remain competitive by enabling technology uptake, which then promotes greater uptake of home-grown technologies.

The Interim Communities Exchange featured presentations divided into three main areas:

- assurance and safety-critical engineering
- cloud / edge, internet of things (IoT) and monitoring
- governance, standards and interoperability

The next TechNexus workshops will be held in person at the 2026 HiPEAC conference in Kraków, both focusing on supporting cross-disciplinary bridging technologies in complex and critical systems. On 26 January, the workshop titled 'STEADINESS for TechNexus' will explore system engineering domains, while on 27 January, 'FORECAST for TechNexus' will take an in-depth look at functional engineering domains.

FURTHER INFORMATION:
TechNexus ICE Workshop, October 2025
bit.ly/TechNexus_ICE_Oct25
STEADINESS at HiPEAC 2026 bit.ly/HiPEAC26_STEADINESS
FORECAST at HiPEAC 2026 bit.ly/HiPEAC26_FORECAST

Casimir Institute for semiconductor and high-tech systems launched at TU/e

At the Future of Chips Event on Tuesday 30 September, Eindhoven University of Technology (TU/e) officially opened the Casimir Institute, the university's new research institute for future chips and high-tech systems. In launching the institute, TU/e stated that it aims to reinforce its ambition to become Europe's leading chip university, while also contributing to the technological sovereignty of the continent. The Casimir Institute brings together three existing TU/e units – the Eindhoven Hendrik Casimir Institute, the High Tech Systems Center and the Future Chips Flagship – into one entity.

Named after the renowned physicist Hendrik Casimir, the new research institute brings together over 700 scientists focusing on semiconductors, quantum technology, photonics, advanced materials, high-tech systems and fundamental research. 'This integrated approach is unique,' said TU/e Rector Magnificus Silvia Lenaerts. 'It enables researchers from different disciplines to connect more quickly, share insights more easily and collaborate across fields more effectively. This way, we can speed up the process of translating knowledge into economic and societal impact.'

TU/e is developing a new €200 million laboratory and cleanroom facility, partly supported by ASML, and is working with TNO on a pilot factory for photonic chips at the High Tech Campus. The university plays a role in various pilots and initiatives arising from the European Commission's Chips Act, which was adopted in 2023, such as the ChipNL Competence Center and the Europese Chip Design Platform.

The Casimir Institute management team. Left to right: Erwin Kessels, Víctor Sánchez Martín, Aida Todri-Sanial, Bart Smolders, Joost Kok and Olaf van der Sluis. Credit: TU/e, Bart van Overbeeke

Panel discussion at the Future of Chips Event, featuring representatives from ASML, NXP, TSMC, imec, and Axelera AI, moderated by Aida Todri-Sanial. Credit: TU/e, Bart van Overbeeke

The new institute builds upon TU/e's track record in developing new materials for chips, including so-called '2D' materials, with collaborations with major chip companies including Intel, TSMC, ASM and ASML. It will also have a significant focus on socially responsible and sustainable solutions, including energy-efficient chips, self-healing materials, and privacy-preserving edge solutions.

'Our goal is clear,' commented Bart Smolders, Scientific Director of the Casimir Institute. 'We want to be the place where the future of chips and high-tech systems is shaped for the Netherlands, for Europe and even the world.'

HiPEAC member Aida Todri-Sanial, part of the management team of the Casimir Institute, noted its relevance to the HiPEAC community: 'TU/e has launched the Casimir Institute for Future Chips and High-Tech systems to drive innovation across design and integration, materials and processes, equipments and foundational technologies. This new institute is highly relevant to the HiPEAC community as we share a common vision and face similar challenges to address energy efficiency and sustainability of future chips and systems. I encourage students, industry partners and academic colleagues to engage with us and explore opportunities for partnerships and synergies.'

FURTHER INFORMATION:
tue.nl/en/research/institutes/casimir-institute

KU Leuven spinoff Belfort raises € 5 million for secure, real-time processing of encrypted data

Belfort, a spinoff of KU Leuven, has launched a hardware accelerator for encrypted compute. According to the announcement by KU Leuven, this innovation – the first available hardware accelerator built specifically for encrypted compute – makes it possible to compute directly on encrypted data without ever decrypting it, ensuring both security and privacy. This means that, for the first time, encrypted compute becomes practical for real-world use cases such as fraud detection, genomic analysis, and secure government operations.

'AI is transforming everything, but the infrastructure to keep sensitive data and models secure has not caught up,' said Michiel Van Beirendonck, co-founder and CEO of Belfort. 'Encrypted compute is the answer, but without hardware acceleration, it cannot be scaled up. The company that cracks that challenge could become the next billion-dollar company.'

Belfort builds on cryptographic research by the COSIC ('Computer Security and Industrial Cryptography') research group at KU Leuven. HiPEAC member Ingrid Verbauwhede, a global authority in cryptographic hardware, is a co-founder of Belfort and its head scientist. The research matured through several grants and projects including two European Research Council (ERC) grants and participation in a competitive US government contract with DARPA.

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

'For decades, we have explored how to make computation secure by design. With Belfort, we are finally bringing those ideas into practice,' commented Ingrid Verbauwhede. 'For the first time, we see cryptographic theory translate so directly into usable, real-time systems. It is a game-changer for anyone who needs to protect sensitive data.'

Belfort has raised €5 million in seed funding to bring encrypted data processing into practical, real-time use. The funding round is led by Vsquared Ventures, with participation from Anagram, Protocol VC, Inovia Capital, Syndicate One, Prototype, Credibly Neutral, and high-profile angel investors including Jeff Dean (Google) and Naval Ravikant.

FURTHER INFORMATION:
belfortlabs.com
esat.kuleuven.be/cosic

Axelera AI announces Europa processing unit

Hardware-acceleration technology provider Axelera AI has announced Europa™, an artificial intelligence processor unit (AIPU) designed for multi-user generative AI and computer vision applications. According to the company's announcement, the Europa AIPU features eight second-generation AI cores, each incorporating Axelera's advanced digital in-memory compute (D IMC) technology and large vector engines, delivering up to 629 tera operations per second (TOPS) at INT8 precision. The AI cores are complemented by two clusters of eight dedicated RISC V vector processing cores for non-AI pre- and post-processing, achieving a peak performance of 4915 giga operations per second (GOPS). An integrated H.264/H.265 decoder further offloads media tasks.

Axelera CEO Fabrizio Del Maffeo with the Europa AIPU

According to Axelera, Europa also integrates 128MB of on-chip L2 SRAM and 256-bit LPDDR5 interface to provide 200 GB/s bandwidth, thereby addressing memory bottlenecks.

HiPEAC member and Axelera AI Chief Executive Fabrizio Del Maffeo commented: 'Europa makes enterprise-class AI processing power available to nearly anyone. From manufacturing automation to intelligent surveillance to autonomous systems, Europa enables the next generation of breakthrough AI applications without compromise, complexity or massive budgets.'

bit.ly/Axelera_AI_Europe_2025

Chipmind raises $2.5 million to develop AI agents for chip design

Chipmind, a startup developing AI agents for chip design and verification tasks using electronic design automation (EDA) tools, has raised $2.5 million in a pre-seed funding round led by Founderful. Based in Zürich, the company was founded by Harald Kröll, the chief executive, and Sandro Belfanti, the chief technology officer, who met while studying at ETH Zürich. HiPEAC member Luca Benini is a scientific advisor to the company.

'In the semiconductor industry, deep customization and data protection are fundamental, but true design awareness is what separates a generic tool from an intelligent partner. Each company's chip is a complex hierarchy with unique constraints, surrounded by a proprietary environment of tools and workflows,' said Harald Kröll in a Chipmind press release. 'Our "design-aware" agents are engineered to holistically understand the entire chip context, not just the surrounding tools.'

'Anyone who's spent time in chip development knows how much of the work is repetitive and time-consuming, demanding precision but not necessarily creativity,' said Sandro Belfanti. 'I've often wished for a solution that could magically take care of those tedious tasks so I could focus on solving real engineering challenges. With Chipmind Agents, we're finally bringing that solution to life: AI agents that can autonomously handle the boring parts, letting engineers focus on what truly matters: innovation.'

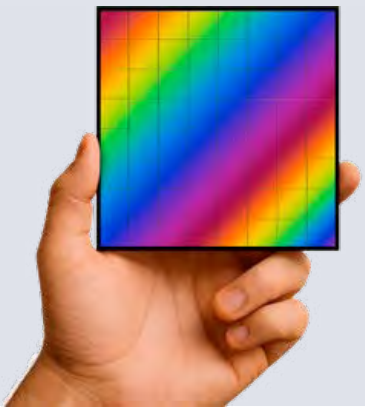
chipmind.ai

Euclyd announces inference architecture CRAFTWERK

European startup Euclyd has announced its CRAFTWERK 'system-in-package' (SiP), which combines custom processors, custom memory, and advanced 2.5D/3D packaging. According to Euclyd, the chiplet-based architecture features 16,384 custom SIMD processors delivering up to 8 PFLOPS (FP16) or 32 PFLOPS (FP4), paired with 1 TB of custom ultra-bandwidth memory (UBM) offering an 8,000 TB/s bandwidth.

CRAFTWERK powers Euclyd's flagship rack-scale system, CRAFTWERK STATION CWS 32, which integrates 32 SiPs to deliver 1.024 exaflops of FP4 compute and 32 TB of UBM. In multi-user mode, Euclyd claims that CWS 32 is projected to achieve 7.68 million tokens per second at 125 kW. According to Euclyd, this represents a 100x improvement in power efficiency and cost per token over leading alternatives.

euclyd.ai



ASML and Mistral form strategic partnership

Semiconductor-equipment manufacturer ASML and artificial intelligence (AI) model developer Mistral AI have announced a strategic partnership. The partnership is based on a long-term collaboration agreement to explore the use of AI models across ASML's product portfolio, as well as research, development and operations, with the aim of enabling ASML to bring products to market faster and empowering the company to build higher performance holistic lithography systems.

ASML also announced in September that they are investing €1.3 billion in Mistral AI's Series C funding round – which saw the AI model developer valued at €14 billion – as lead investor, in order to support its development and reinforce the long-term partnership benefits. This results in ASML holding an approximately 11 percent share on a fully diluted basis in Mistral AI.

asml.com
mistral.ai

QuiX Quantum secures € 15 million in Series A funding

In July, the Dutch photonic quantum computing company QuiX Quantum announced that it had secured €15 million in Series A funding to deliver a single-photon-based quantum computer, designed to implement a universal gate set enabling any quantum operation, in 2026. The Series A round, co-led by Invest-NL and EIC Fund, with participation from existing investors, PhotonVentures, Oost NL, and FORWARD.one, was preceded by the award of the European Innovation Council (EIC) Accelerator programme.

Built on silicon-nitride chips designed for high-volume manufacturing, QuiX Quantum says that its systems are highly scalable, operate primarily at room temperature, and are fully compatible with datacentre environments.

quixquantum.com

Scintil Photonics closes $58 million Series B round

Scintil Photonics, a Grenoble-based company developing photonic system-on-chip (PSoC) solutions for artificial intelligence (AI) infrastructure, has announced that it has secured $58 million in Series B funding round to ramp up the commercial development of its LEAF Light™ technology.

Built on its proprietary process, named SHIP™ (Scintil Heterogeneous Integration Photonics), Scintil's technology integrates lasers, photodiodes, and modulators onto a single chip, replacing dozens of discrete components to deliver higher performance, efficiency, and density for next-generation co-packaged optics (COP). The company's technology draws on 15 years of research at CEA-Leti.

The funding will support the commercial ramp of LEAF Light™, which Scintil describes as 'the industry's first PSoC DWDM-native light engine, aligned with next-generation co-packaged optics (CPO)'. DWDM stands for 'dense wavelength division multiplexing'; according to Scintil, being DWDM-native means the single-chip device can output many precisely spaced wavelengths, increasing bandwidth while lowering energy use.

scintil-photonics.com

Optical computing startup Arago raises $ 26 million in seed funding

Paris-based fabless chip company Arago has raised $26 million in seed funding to accelerate the commercialization of its hybrid AI accelerator, codenamed 'JEF'. Co-founded by former researchers at ETH Zürich, Arago's technology combines electronic and photonic elements to reduce energy consumption. According to the company, Arago's technology can run AI models using industry-standard software frameworks, while remaining compatible with existing compute infrastructure and semiconductor manufacturing processes, hence facilitating real-world adoption.

arago.inc

Vertical Compute recruiting range of technical roles

Having closed a successful €20 million seed round in January, Belgian startup Vertical Compute is working on chiplet-based solutions which use vertical lanes on top of computation units to reduce data movement and overcome the memory wall. Founded by Sylvain Dubois (ex-Google) and Sebastien Couet (ex-imec), the company is recruiting for a number of roles based in Europe, including both full-time positions and internships.

'Memory technologies face limitations in both density and performance scaling, while processor performance continues to surge. The extreme data access requirements of AI workloads exacerbate this challenge, making it imperative to overcome the memory wall to enable the next wave of AI innovations,' commented Sébastien Couet at the time of the seed round. 'We believe going Vertical is the path to 100x gains.'

verticalcompute.com

EUPILOT announces tape-out of two HPC / AI chips

Romana Konjevod,
Barcelona Supercomputing Center (BSC)

In August, the EUPILOT consortium successfully completed the tape-out of EUPILOT's VEC and MLS chips on GF12 LP+ and TSMC7. VEC is optimized for HPC workloads, while MLS is optimized for AI workloads.

Following the successful arrival of the project's first 12nm test chips in July 2024, designed using GlobalFoundries' FinFET technology, this tape-out is an important step towards demonstrating next-generation accelerator platforms built on European research and design. By combining open-source principles with innovative hardware-software co-design, EUPILOT is laying the foundations for future European exascale systems while strengthening collaboration across Europe's computing community.

Commenting on the news, EUPILOT Principal Investigator **Carlos Puchol** (BSC) said: 'This tape-out represents a major milestone in EUPILOT's mission to extend the RISC-V ecosystem into high-performance computing (HPC) and high-performance data analytics (HPDA) applications – one which reflects the hard work and collaboration of everyone involved.'

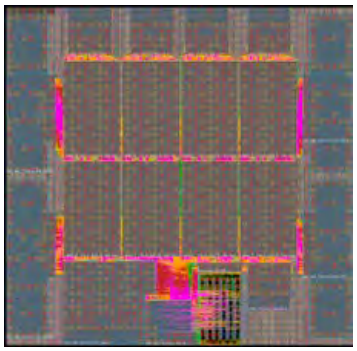
Vanessa Iglesias (BSC), who leads EUPILOT's exploitation efforts, commented: 'Each phase of EUPILOT has brought us closer to reducing our dependencies and building on the technological sovereignty the EU aims for. More than just a technical milestone, the true value lies in the knowledge and expertise gained.'

Technical Overviews

• VEC

As described by **Jordi Fornt** (BSC), the VEC chip is designed to support vectorized and memory-intensive workloads:

- 8 cache-coherent 64-bit RISC-V cores, each with a 16-lane vector unit
- 4 LPDDR4 channels (68 GB/s total bandwidth)
- 5 chip-to-chip links (40 GB/s total bi-directional bandwidth)
- 1 PCIe channel (16 GB/s bi-directional bandwidth)



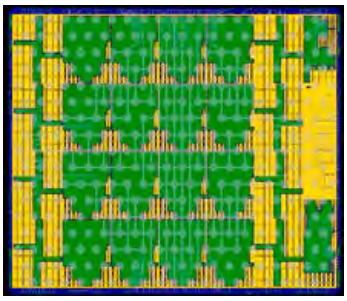
VEC chip die, courtesy of Fraunhofer

• MLS

At ETH Zurich, **Tim Fischer** highlighted progress on MLS (codenamed Picobello), highlighting the many-core architecture underlying the chip:

- based on the Snitch RV32 core
- 144 cores arranged in 16 clusters with 9 cores each, plus an additional STX cluster
- more than 10 MB on-chip memory running at 1 GHz

Developed through the TSMC Academic Program and supported by Europractice as part of the EUPILOT project, the chip is expected to return from fabrication in early 2026, with results to be published after initial measurements.



Picobello, courtesy of Thomas Benz – ETH Zurich

LOOKING AHEAD

With the tape-out completed, EUPILOT now enters a critical stage focused on silicon bring-up, validation, software enablement, system-level integration, and datacentre deployment. These steps are essential to turning the designs into fully functional accelerator platforms for HPC and AI. While challenges remain, the consortium's expertise and collaboration provide a strong foundation for what comes next.

Carlos Puchol stressed that tape-out is not the finish line but an important step toward demonstrating working accelerator platforms: 'The coming months will bring challenges, but the collective effort that got us here provides the stepping stones to continue forward. Every contribution – whether in design, verification, or physical implementation – was essential in reaching this point.'

Vanessa Iglesias added that the know-how gained as part of this process is already a major step forward: 'The true value lies in the workforce, and in the knowledge and expertise gained, which will accelerate future developments and strengthen Europe's technological position.'

Building on this momentum, the team looks forward to sharing further progress with the broader community, advancing Europe's position in the global HPC and AI landscape.

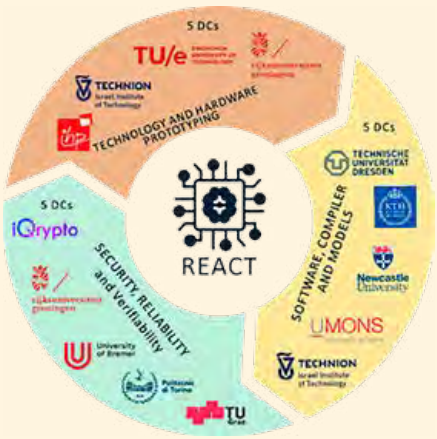
eupilot.eu

REACT MSCA to train 15 early-stage PhD researchers to create a neuromorphic platform

Farhad Merchant, University of Groningen

Self-awareness in humans is a natural and deeply ingrained cognitive tendency, enabling individuals to perceive, interpret, and evaluate their own actions, emotions, and states of being. This capacity arises from complex processes within the human brain, which integrates a wide array of signals originating from the body's sensory systems – vision, hearing, touch, proprioception, and internal physiological feedback. These inputs allow humans not only to respond to the surrounding environment but also to adapt their strategies and behaviours based on an internal model of themselves.

Translating this remarkable biological capability into the realm of electronics, through what is known as neuromorphic computing, opens the door to creating machines capable of sophisticated decision-making and adaptive learning. Neuromorphic systems aim to replicate, at least partially, the efficiency and resilience of the human brain by using hardware and software architectures inspired by neural networks and synaptic processing. Such systems have the potential to enhance a wide variety of applications, ranging from autonomous vehicles that can navigate safely under changing road conditions to intelligent personal assistants that learn user preferences and anticipate needs in real time.



However, achieving this vision is not without obstacles. Neuromorphic computing must address key challenges in energy efficiency – since processing large amounts of data can be highly power-intensive – as well as in reliability and security, ensuring that systems perform consistently and are protected against faults and malicious attacks.

The REACT MSCA Doctoral Network, coordinated by Farhad Merchant at the University of Groningen, The Netherlands, addresses these challenges. The project aims to develop a neuromorphic platform that embodies self-awareness in the domains of energy efficiency, security, and reliability. It will train 15 early-stage PhD researchers across a broad set of interdisciplinary fields, including materials science, device physics, computer architecture, hardware prototyping, compiler design, simulation and emulation frameworks, system security, fault tolerance, and formal verification – equipping them to advance the next generation of intelligent, trustworthy computing systems.

project-react.eu

Launch of Phortify, a network to power Europe's photonics talent pipeline

Phortify

Launched in September, Phortify ('Photonics education network for next-gen innovation and digital skills excellence for industry and society') is a new photonics education and training network that links Europe's leading classrooms and laboratories to boost digital skills and innovation.

Coordinated by Vrije Universiteit Brussel (VUB) with a dozen partners across Europe, Phortify is looking to cover the full photonics value chain – from fundamentals to design, fabrication, packaging, testing, and deployment. The project will create a Europe-wide photonics curriculum across seven master's programmes and modules co-designed with industry. Phortify will enable students and professionals to earn recognized credentials and access top labs, instructors, and digital resources anywhere in Europe.

Professor Heidi Ottevaere, Project Coordinator, VUB, Belgium, said: 'By building a harmonized photonics education network, we are equipping the next generation with the innovation mindset and digital skills excellence that industry and society urgently need. This launch marks the beginning of a powerful collaboration across borders, disciplines and sectors.'

phortify.eu

BSC launches ODOS MPI, enabling transparent SmartNIC offloading of computation and communication kernels in OpenMP

Muhammad Usman, Mariano Benito, Sergio Iserte and Antonio J. Peña, Barcelona Supercomputing Center (BSC)

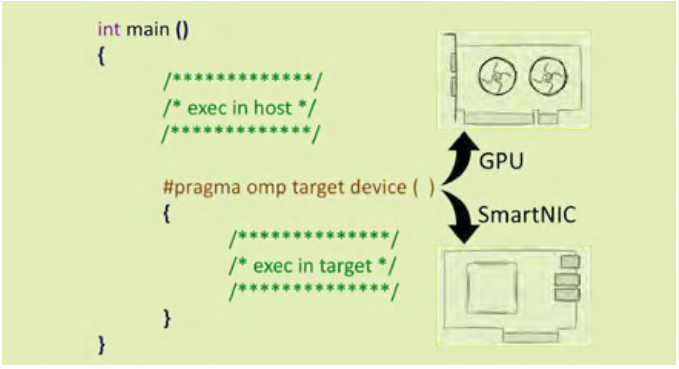
BSC has announced a major advance in high-performance computing (HPC) programmability with the release of ODOS-MPI, an innovative framework that brings SmartNIC acceleration into the mainstream by integrating it seamlessly into widely used programming models.

SmartNICs, network interface cards with embedded processors, are transforming data centres by enabling operations such as input / output (I/O), encryption, and security tasks directly at the network edge. Now, researchers at BSC have extended this potential to HPC by offloading communication-intensive tasks directly to these devices. Leveraging NVIDIA BlueField data processing units (DPUs), ODOS-MPI allows developers to take advantage of SmartNICs for both computation and communication without rewriting their applications in low-level languages.

'Our goal was to reduce the entry barrier for domain scientists and HPC developers who want to explore SmartNIC acceleration,' said Antonio J. Peña, principal investigator of the project. 'By integrating MPI kernels offloading into OpenMP (the de facto standards for parallel programming) ODOS-MPI allows users to exploit SmartNICs without learning complex device-specific application programming interfaces (APIs).'

At its core, ODOS-MPI extends the MPI+OpenMP programming model to enable offloading of computation and communication kernels to SmartNICs while preserving OpenMP's syntax and execution model. This means developers can annotate critical regions of their code for offloading, relying on the ODOS runtime to manage communication, memory transfers, and synchronization transparently. The framework builds on the LLVM compiler infrastructure and NVIDIA's DOCA software development kit (SDK), ensuring highly efficient interactions with BlueField DPUs.

The team validated ODOS-MPI by porting point-to-point, one-sided, and collective communication operations of the Ohio State University (OSU) Micro-Benchmarks suite. In tests with osu_ialltoall, one of the most communication-intensive benchmarks, ODOS-MPI achieved up to 10x improvement in execution time for large messages compared to standard Open MPI.



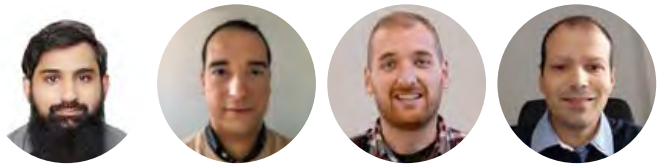
To demonstrate real-world applicability, the researchers integrated ODOS-MPI into miniWeather, a simplified atmospheric simulation code. By offloading computation and communication to SmartNICs, they achieved over 18% execution time improvement, showcasing the potential for accelerating scientific applications that combine heavy computation with frequent inter-node communication.

'This is more than an optimization; it's a new way to think about using SmartNICs in supercomputing,' said Muhammad Usman, main author of the paper. 'We are enabling concurrent computation and communication in a way that simplifies development and improves efficiency at scale.'

The implications for HPC are significant: ODOS-MPI opens the door to widespread SmartNIC adoption for workloads like stencil codes, molecular dynamics, and distributed deep learning, where overlapping computation and communication can dramatically boost performance.

ODOS-MPI is part of the Open UCX initiative and is now available to the research community via BSC's public repository (see 'Further information').

The work will be presented at SC25, the premier international conference for high-performance computing, networking, storage, and analysis.



BSC researchers involved in this effort (left to right): Muhammad Usman, Mariano Benito, Sergio Iserte and Antonio J. Peña

FURTHER INFORMATION:
gitlab.bsc.es/accelcom-public/odos-mpi

New AIOTI paper on edge IoT immersive and spatial computing applications

The Alliance for IoT and Edge Computing Innovation (AIOTI) Working Group on Research and Innovation has published a paper titled *Edge IoT Industrial Immersive and Spatial Computing Applications*. The paper aims to provide a comprehensive overview of the convergence between the internet of things (IoT), artificial intelligence (AI), edge, and spatial computing, and how they are applied to various immersive applications across different industrial sectors.

Immersive technologies, including virtual reality (VR), augmented reality (AR), mixed reality (MR), and extended reality (XR), alongside advanced concepts such as digital twins (DT), immersive triplets (IMT), the metaverse, omniverse, and spatial computing, are becoming increasingly relevant in a number of industrial applications.

According to the AIOTI WG, the convergence of these immersive technologies with edge IoT, AI, and advanced intelligent connectivity infrastructure is shaping an industrial real-digital-virtual continuum, termed the 'phygital' world. By combining real-world interactions and virtual simulations, the WG finds that industries can improve operational efficiency, reduce downtime, enhance safety protocols, and gain superior decision-making capabilities.

Exploring the ways in which these technologies are converging, and examining how they are applied in industrial settings, this position paper details the impact of these applications across industrial sectors including culture and heritage, manufacturing, automotive, energy, construction, mobility, logistics, healthcare, agriculture, and tourism.

FURTHER INFORMATION:
bit.ly/AIOTI_edge_IoT_immersive_2025

Pragmatic wins Best Paper at MICRO 2025

The paper 'Flexing RISC-V Instruction Subset Processors to Extreme Edge', by Alireza Raisiardali, Konstantinos Iordanou, Jedrzej Kufel, Kowshik Gudimetla, Kris Myny and Emre Ozer was named best paper at MICRO 2025, which took place in October.

According to the announcement by Pragmatic Semiconductor, the paper explores a new way to automatically generate Pragmatic's FlexICs based on a 32-bit RISC-V instruction subset. Utilizing open-source RISC-V architecture, it breaks down instructions into small 'building blocks' which are stitched together to create a custom processor. Pragmatic states that, as each building block is already pre-tested and verified, this approach reduces design and testing time, making it quicker and cheaper to create processors tailored to specific tasks. It also notes that the chips use about 30% less power and space compared to normal RISC-V processors and are up to 30 times more energy efficient than the smallest existing 32-bit RISC-V chip.

Commenting on the award, HiPEAC member and 2025 HiPEAC conference keynote speaker Emre Ozer said: 'Our goal is to enable fast and reliable design of custom processors, fabricate them as FlexICs, and deploy them to the extreme edge with speed and confidence.'

Test of Time Award for HiPEAC members' CASES 2010 paper

During the 21st ACM/IEEE Embedded Systems Week (28 September to 3 October), the paper 'Practical Aggregation of Semantical Program Properties for Machine Learning-Based Optimization' (CASES 2010) was named winner of the 2025 CASES Test of Time Award. The paper was authored by Mircea Namolaru, Albert Cohen, Grigori Fursin, Ayal Zaks, and Ari Freund. Albert Cohen, Grigori Fursin and Ayal Zaks are all HiPEAC members.

ICE and TalTech explore the future of computing in joint seminar

Nils Bosbach and Rainer Leupers, RWTH Aachen

In June 2025, researchers from the Institute for Communication Technologies and Embedded Systems (ICE) at RWTH Aachen University visited the Tallinn University of Technology (TalTech) for a joint scientific seminar on the latest developments in computing systems. The event featured technical talks addressing key topics such as artificial intelligence (AI) on edge devices, neuromorphic computing, secure hardware, RISC-V customization, and system-on-chip reliability. The seminar provided valuable opportunities for both HiPEAC-community research groups to exchange knowledge.

Opening presentations introduced the research focus areas of both TalTech and ICE, with welcoming remarks by Professor Maksim Jenihhin (TalTech) and Professor Rainer Leupers (RWTH Aachen). These were followed by a series of in-depth technical sessions. ICE researchers presented innovations in hardware security, logic locking, RRAM-based neural networks, and flexible AI deployment on edge devices. TalTech contributions highlighted open-source tools for hardware reliability of deep neural networks (DNNs), fault tolerance of DNN hardware accelerators, RISC-V verification using assertions, lightweight cryptographic algorithms to address overproduction of integrated circuits, and system-on-chip fault management.

The ICE team thanks Professor Maksim Jenihhin, Professor Gert Jervan, and the TalTech community for their warm hospitality and engaging discussions.

André Seznec receives ACM-IEEE CS Eckert-Mauchly Award

This summer, HiPEAC founding member André Seznec, Fellow Research Director at INRIA / IRISA and Fellow at SiFive, received the ACM Eckert-Mauchly Award during a ceremony at ISCA 2025, the IEEE /ACM International Symposium on Computer Architecture. The award was made in honour of André's 'extensive impact on computing, most notably pioneering contributions to branch prediction and cache memories'.

According to the announcement by ACM, André's inventions can be found in billions of central processing units (CPUs) worldwide. These include the TAGE branch predictor and skewed-associative cache. His work has served as a gold standard of branch prediction for the last 15 years, with most current structures in industrial designs rooted in his contributions.

His early contributions were in vector architectures, particularly the memory system. Since the early 1990s, his main research activity has been focused on the architecture of microprocessors, including caches, pipelines, branch predictors, speculative execution, multithreading, and multicores.

Inaugurated in 1979, the ACM-IEEE Eckert-Mauchly award was named for John Presper Eckert and John William Mauchly, who collaborated on the design and construction of the Electronic Numerical Integrator and Computer (ENIAC). It is known as the computer-architecture community's most prestigious award.

Photo credit: Inria / H. Raguet

Carole-Jean Wu wins ACM SIGARCH Maurice-Wilkes Award

HiPEAC associate member Carole-Jean Wu, a research scientist at Meta AI Research, was named winner of the ACM SIGARCH Maurice-Wilkes Award during ISCA 2025 for 'pioneering computer architecture research to enable efficient, scalable, and sustainable machine learning'. The award is made annually in recognition of an outstanding contribution to computer architecture made by an individual whose computer-related professional career started no earlier than 1 January of the year that is 20 years prior to the year of the award.

An active HiPEAC associate member, Carole taught at the HiPEAC summer school, ACACES, in 2023.

Read our 2023 interview with Carole in HiPEACinfo 69 bit.ly/HiPEACinfo69

Joel Emer wins ACM SIGARCH Alan D. Berenbaum Distinguished Service Award

During ISCA 2025 in June, HiPEAC associate member Joel Emer, a former teacher at the ACACES summer school, was presented with the ACM SIGARCH Alan D. Berenbaum Distinguished Service Award, a recognition of outstanding service in the field of computer architecture and design. A senior distinguished research scientist at NVIDIA and professor at Massachusetts Institute of Technology (MIT), Joel was selected for this award 'for creating the 'Meet a Senior Architect' programme and more broadly for decades of exemplary mentoring and visionary leadership to build a culture of inclusion and technical excellence across the computer architecture community'.

EIC Pre-Accelerator call

Paul Pietrangelo, Lira

The pilot European Innovation Council (EIC) Pre-Accelerator call is a joint scheme between the EIC and the 'Widening participation and strengthening the European Research Area' (WIDERA) programme. It represents a significant opportunity for deep-tech companies in widening countries to access both funding and strategic support.

To be eligible, applicants must be located in one of the following member states and associated countries: Bulgaria, Croatia, Cyprus, Czechia, Estonia, Greece, Hungary, Latvia, Lithuania, Malta, Poland, Portugal, Romania, Slovakia, Slovenia, Albania, Armenia, Bosnia and Herzegovina, Faroe Islands, Georgia, Moldova, Montenegro, Serbia, Tunisia, Turkey or Ukraine.

The overall budget for 2025 is € 20 million, and the maximum project duration is 24 months. With a 70% funding rate and a grant funding range of € 300,000 to € 500,000, the EIC expects to fund at least 40 companies in 2025.

The main goal of this funding scheme is to help deep-tech companies to become investor-ready and / or a good candidate for the EIC Accelerator at the end of the project. Applicants have to demonstrate they can cover 30% of project costs since the funding rate is 70%.

All activities at technology readiness level (TRL) 3 must be completed; the scheme is designed to fund activities at TRL 4 and TRL 5. The goal is to raise the TRL to 6 by the end of the project.

The deadline for submitting a 22-page proposal is 18 November 2025. Results are expected by February 2026, with the grant starting date being in June 2026.

eic.ec.europa.eu/eic-funding-opportunities/eic-pre-accelerator_en

Credit: Lira

Dates for your diary

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

HiPEAC 2026: High Performance, Edge And Cloud computing
26 - 28 January 2026, Kraków, Poland
hipeac.net/conference Sponsorship opportunities available
sponsorship@hipeac.net.

NorCAS 2025: IEEE Nordic Circuits and Systems Conference
28 - 29 October 2025, Riga, Latvia
events.tuni.fi/norcas
Further information: Jari Nurmi, Tampere University
norcas@tuni.fi

EFECS 2024
3 - 4 December 2025, St Julian's, Malta
efecs.eu

Open Source Experience
10 - 12 December 2025, Paris, France
opensource-experience.com
Visit the Eclipse booth to catch up with HiPEAC news

ARITH 2026:
33rd IEEE International Symposium on Computer Arithmetic
28 June - 1 July 2026, Fulda, Germany
Abstract deadline: 23 January 2026
arith2026.org

ISC High Performance 2026
22 - 26 June 2026, Hamburg, Germany
Calls for submission ongoing
isc-hpc.com


Computing services leveraging the full continuum from cloud to edge to the internet of things (IoT) are pivotal to enabling the transition from traditional centralized energy systems to distributed systems fed by intermittent energy sources. Two major projects funded by the European Union are rolling out large-scale pilots and working on the cloud-edge-IoT (CEI) technologies necessary to power this transition. To find out more, we spoke to Ignacio Lacalle Úbeda (Universitat Politècnica de València), member of the coordination team of the O-CEI project, and Ioanna Drigkopoulou (Netcompany), the coordinator of COP-PILOT.

Piloting energy systems in Europe

How large-scale pilots are putting the cloud-edge-continuum at the service of energy

'The biggest challenges facing Europe's energy system today are tied to the transition from a centralized energy model to a decentralized one, which relies heavily on renewable energy sources (RES). This creates significant fluctuations in supply that the current grid struggles to handle,' explains O-CEI representative, Ignacio Lacalle Úbeda. 'To keep pace with renewables, Europe's power-management flexibility needs to double by 2030.'

The need for greater power-management flexibility stems from a radical shift in energy production and distribution. 'Future energy systems will be much more dynamic and complex than in the past. Instead of a one-way flow from large power plants, the grid will need to manage energy from diverse, decentralized sources like solar and wind,' says Ignacio. 'This new system requires a high degree of responsiveness and real-time data sharing to anticipate and adapt to sharp, unpredictable fluctuations in demand and supply.'

Harnessing the cloud-edge-IoT (CEI) continuum is essential to respond to this new energy reality. As COP-PILOT coordinator Ioanna Drigkopoulou says: 'CEI systems are essential for enabling the energy transition by transforming passive electricity networks into active ones with two-way power flows.' This is enabled, says Ignacio, through 'real-time, decentralized monitoring and information processing, which is crucial for managing energy flexibility'.

Due to the 'sheer volume of data and the need for immediate action', cloud services alone are insufficient. 'CEI architectures enable intelligence to be pushed closer to where the data is generated, at the "edge" (such as vehicles, chargers, and smart buildings), thereby reducing latency and improving responsiveness. This approach helps break down data silos and allows for seamless cooperation between different devices and systems, creating a functional, decentralized network,' explains Ignacio.

Large-scale pilots enabling the energy transition with CEI

As highlighted in the 2024 Draghi report on EU competitiveness, a reliable supply of clean, affordable energy is essential for Europe to flourish, while decarbonization offers significant potential to galvanize the European economy. Given this context, both O-CEI and COP-PILOT include dedicated plans to put next-generation computing at the service of the energy sector.

O-CEI is developing an open CEI continuum platform to promote energy flexibility, with a focus on interoperability, security, and energy efficiency. The project's solutions are being validated through eight large-scale pilots in key sectors, including electromobility, logistics and electrical grids. The pilots aim to demonstrate how CEI can, for example:

- **Optimize grid performance** by integrating renewable energy sources.
- **Implement intelligent charging strategies** for electric vehicle fleets to reduce emissions and minimize grid stress.
- **Manage energy demand in challenging environments** such as maritime ports, adapting power needs based on real-time data.
- **Increase social acceptance of energy flexibility options** in urban areas by enabling prosumers to participate in the energy market.

'By providing blueprints, utilities, and a federated marketplace, O-CEI is building the digital backbone needed for a more sustainable and resilient CEI ecosystem, which will foster a smoother transition to a cleaner energy future,' says Ignacio. 'Being able to exploit distributed computing and energy-related assets, and having them cooperate in matters of anticipation and prediction in energy flexibility, is pivotal in O-CEI and in Europe.'

For its part, COP-PILOT is building a collaborative open platform (COP) framework to enable end-to-end orchestration across service domains. COP-PILOT includes a cluster focusing on energy, in which partners will use IoT and smart metering to gather energy data from distributed energy resources, including biogas and photovoltaic units, as well as from commercial and residential customers.

Through its energy cluster, COP-PILOT aims to accomplish the following:

- **Enabling edge intelligence**: the project's platform will perform real-time data analytics at the edge to create flexibility and adjust production or consumption.
- **Increasing grid reliability**: The work will focus on fostering real-time flexibility from active electricity distribution grids. This helps solve operational issues like line congestion and node overvoltages that arise from increased reliance on distributed energy resources.
- **Ensuring an uninterruptible power supply for EV charging** with predictive-maintenance and monitoring systems, optimizing resource utilization and maximizing charger uptime to support the widespread adoption of EVs.
- **Reducing outages and ensuring reliability at a biogas plant** thanks to a predictive maintenance system.

Technical challenges in creating a seamless CEI infrastructure

Implementing CEI systems for the energy sector will require a number of technical challenges to be overcome. Both Ignacio and Ioanna cite interoperability as a major issue. 'We need to overcome the fragmentation of the current cloud ecosystem, where many platforms exist in silos,' explains Ignacio. 'A key challenge is achieving interoperability and trustworthy data exchange across different domains. This involves developing common standards, ontologies, and standardized interfaces for data sharing.'

In this area, O-CEI is creating an open platform and providing blueprints that enable seamless data exchange across computing ecosystems. 'This framework will support a dynamic marketplace for digital assets, including services, datasets, and AI models,' adds Ignacio.

The platform being developed by COP-PILOT also aligns with open standards. 'The project's vision is centred on a secure integration fabric (SIF): an encrypted, programmable layer that facilitates flexible, trust-based data integration across NGSI-LD-compatible brokers,' explains Ioanna. 'This mechanism enables the creation of private data spaces and allows organizations to leverage existing infrastructure, without compromising on security or interoperability.'

Members of the O-CEI consortium (above) and COP-PILOT consortium (below) at their respective kick-off meetings

Security and privacy are also integral to the development of the O-CEI platform, says Ignacio. 'We will implement zero-trust environments with federated identity-management and authorization mechanisms,' he says. 'We're also working on integrating AI-powered cybersecurity mechanisms to detect threats.'

Another major challenge is orchestration, given the need to manage a host of computing elements from IoT devices to cloud nodes. 'The rise of 5G / 6G technologies is leading to an immense increase in the number of connected IoT devices, requiring intelligent management and automation,' notes Ioanna. 'The COP-PILOT platform is designed to handle this complexity through a hierarchical orchestration approach. In addition, the project is developing AI-based extensions to allow the orchestrator to intelligently analyse complex data to drive actionable insights and manage multiple services.' COP-PILOT's end-to-end orchestrator and programmable integration fabric will provide enhanced visibility and granularity for effective system management.

The O-CEI approach to orchestration, meanwhile, revolves around easy-to-deploy, ready-to-adopt technology solutions, known as 'blueprints', says Ignacio: 'Managing diverse and heterogeneous infrastructures requires solutions that can dynamically deploy workloads in the most suitable locations to minimize energy consumption and latency. Our "blueprints" – pre-configured technological solutions – will simplify deployment and operation.'

Finally, the projects aim to deliver user-friendly, standardized services to maximize take-up. 'We have integrated measures such as guidelines, secure sharing schemas, and successful implementation examples into the core of our activities so that the innovation / adoption gap will be optimized,' says Ignacio.

In the case of COP-PILOT, users will be shielded from the complexity of low-level orchestration application programming interfaces thanks to a user interface based on AI models. The project platform also features a plugin that uses a closed-loop service chain to ensure service level agreements are met through AI forecasting, zerotouch automation, and secure actuation.

Long-lasting impact

Beyond the immediate scope of their three-year terms, both large-scale platforms aim to have a longer-lasting impact on the energy sector, as well as in the other sectors addressed as part of the projects.

By contributing to standards relevant to the energy sector, including the ETSI-OSL/TFS/ZSM/OSM frameworks and extensions to TMF interfacing standards, COP-PILOT 'will help enable a flexible and reliable energy supply for Europe through the use of an open-source, vendor-agnostic platform', Ioanna says. Similarly O-CEI aims to create a pre-standard for CEI deployments, which will accelerate adoption and ensure interoperability.

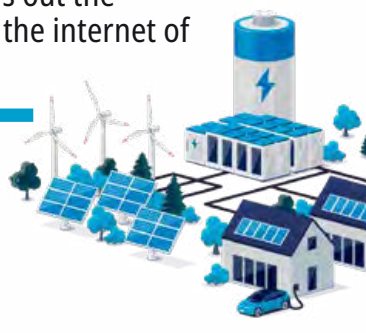
The pilots also aim to improve efficiency in the energy sector. O-CEI's customized CEI blueprints will help reduce businesses reduce operational costs (OPEX) and energy consumption, says Ignacio. For its part, the COP-PILOT project aims to reduce the provisioning and deployment time of complex, cross-sector applications by at least an order of magnitude (10x faster) compared to current configurations and reduce OPEX by 30%, spurring takeup by at least 400 existing businesses for their smart IoT-based applications.

The pilots also aim to foster an ecosystem of small / medium enterprises and startups around the technology developed. 'O-CEI will fund up to 32 new projects through its open calls, enabling applicants to create and monetize innovative edge solutions,' says Ignacio. For its part, the COP-PILOT consortium predicts that 30 new businesses will be created based on the platform, and also aims to lay the foundation for a vibrant, open-source research ecosystem. 'The project's results are expected to reach a technology readiness level (TRL) of 7 by the end of the project, with a path to TRL 9 within five years of completion,' says Ioanna. 'This high-TRL foundation is intended to inspire over 100 researchers to extend the open-source work of the COP-PILOT toolset, with at least five future research and development projects capitalizing on the platform's assets to advance research.'

O-CEI's first call for companies who will develop and test edge solutions in the form of CEI utilities (apps and services) is open until 20 November. Further information can be found on the O-CEI website o-cei.eu/open-calls

O-CEI and COP-PILOT have received funding from the EU's Horizon Europe programme under grant agreement number 101189589 and 101189819 respectively. 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. COP-PILOT has also received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

How can Europe ensure a smooth transition to next-generation energy systems while maximizing opportunities for the continent's computing sector? In this article, adapted from a recent CEI-Sphere market brief, Maria Giuffrida (Trust-IT) lays out the opportunities and challenges for the continuum of cloud, edge and the internet of things (CEI continuum) in the energy sector.



Market brief: Cloud-edge-IoT systems in the energy sector

The European energy sector is undergoing significant transformation due to macroeconomic and geopolitical uncertainties, high energy costs, and the urgent need for decarbonization. Companies in the European Union face electricity prices two to three times higher than in the US – hence the need for a transition to more efficient and sustainable energy systems.

The energy transition represents a shift from centralized to distributed energy systems characterized by the integration of distributed energy resources (DERs) like solar and wind, the expansion of electrification, and decarbonization. Featuring novelties such as bidirectional power flows and active consumer participation, this transition presents both challenges and opportunities for innovation and new revenue streams.

Underpinning the shift to decentralized energy systems is a digital transformation: utilities are increasingly adopting the internet of things (IoT), edge computing, artificial intelligence (AI), and cloud computing to enhance grid stability, optimize energy management, and improve operational efficiency.

Key use cases include:

- **IoT-enabled smart grids**, which gather real-time data on energy consumption and distribution, optimizing electricity flow and reducing transmission losses.
- **AI-driven predictive maintenance of grid infrastructure**, which minimizes downtime and enhances reliability.
- **Real-time data processing to enable demand-response strategies** that balance supply with real-time usage, improving grid efficiency.

In 2024, the total CEI solution spending in the European energy sector was estimated at €17.6 billion. However, the sector is divided, with some companies leading in IT architecture modernization while others lag behind.

Major players in the CEI ecosystem within the energy sector include:

- **System integrators and service providers**: Accenture, Capgemini, CGI, Cognizant, Deloitte, EY, HCLTech, IBM, Infosys, PWC, TCS, Wipro.
- **Platform providers**: AWS, Microsoft, Google
- **Enterprise software vendors**: Aveva, Microsoft, SAP, Oracle, IBM, IFS, Kraken, Dassault Systèmes, Schneider Electric.
- **Operational technologies software vendors**: ABB, Schneider Electric, Siemens Energy, GE Vernova, Rockwell Automation, Emerson.
- **Network equipment and telecoms**: Orange, Deutsche Telekom, BT, Verizon, Vodafone, Telefonica, Equinix, Telstra, Ericsson, Nokia, Qualcomm, Intel.
- **Smart metering vendors**: Landis+Gyr, Itron, Siemens, Schneider Electric, Aclara, Kamstrup, Honeywell.

Major challenges that will need to be addressed include:

- strategic planning for CEI adoption and development;
- overcoming organizational resistance;
- integration with legacy systems;
- cybersecurity threats, as the distributed nature of modern energy systems increases vulnerability to cyberattacks; and
- a shortage of professionals with expertise in CEI solutions.

To enable the digital transformation of the energy sector, several development needs should be addressed. For example, there should be improved collaboration across business departments to leverage collected data for enhanced business value. Data should be uncoupled from applications and technologies to make a stronger business case for CEI investments. Finally, developing robust security frameworks to protect distributed energy systems is essential.

CEI-Sphere supports the development of an open, interoperable, and competitive CEI ecosystem to support the large-scale pilots O-CEI and COP-PILOT and is funded by the EU's Horizon Europe programme under grant agreement number 101189683.

ceisphere.eu/events/cei-market-talks-focus-energy

The incorporation of renewable energy sources – which are often distributed and dynamic – into the electrical grid poses new challenges for systems originally designed for centralized energy generation from stable sources. To avoid incidents like the Iberian Peninsula blackout on 28 April, swift and coordinated decision making is required. In this interview, Eduardo Iraola, Francesc Lordan, Xavier Casas and Rosa Badia (all Barcelona Supercomputing Center) explain how their research – including the HP2C-DT and COLMENA projects, funded by the Spanish government – harnesses digital twins and swarm computing for a smooth transition to an electrical grid fit for the future.

Futureproofing the electrical grid

Using digital twins and swarm computing for decision making and local action in real time

Why are digital twins important for the future management of the electrical grid?

Rosa: While traditionally the electrical grid has been managed by robust monitoring and control systems, these are often rigid and rely heavily on manual intervention by operators. As the grid has evolved, it has integrated renewable energy sources it was not originally designed to handle and has become more complex.

Digital twins are software components that mirror the state and behaviour of physical systems in real time, using both sensed data and simulations. In a broad sense, they are not just visualization tools but closed-loop systems that take part in decision-making by reading from the physical world and, when necessary, acting on it. Digital twins offer a way to extend and adapt existing control frameworks with greater flexibility and accuracy.

What are the different roles of edge, cloud, and high-performance computing in the context of electrical-grid management?

Rosa: High-performance computing (HPC) resources provide a complementary but powerful way to support the needs of digital twins. Edge computing handles urgent tasks close to the data source, cloud platforms enable orchestration and centralized coordination, and HPC clusters add the ability to run heavy workloads using specialized non-virtualized hardware such as graphics processing units (GPUs), RISC-V architectures, or even quantum processors, along with high memory bandwidth. This makes them especially useful for complex workflows, AI-model training or large-scale parallel simulations that may be needed for forecasting and decision-making in demanding scenarios.

What are the main challenges in the transition to renewable energy sources? What role can digital twins, swarm computing, and other advanced computing concepts play in enabling this transition?

Eduardo: The electrical grid is transitioning from a classical mix of stable, predictable energy sources like coal and gas to a mix dominated by renewables such as solar and wind, which are variable and weather dependent. This shift reduces system inertia – unlike large spinning turbines, solar panels connected via inverters do not naturally dampen fluctuations – making the grid more sensitive to sudden changes. At the same time, many consumers are becoming small-scale producers, like households with rooftop solar, causing power to flow in both directions and increasing the chance of local faults spreading. Managing all this to avoid incidents, such as the Iberian Peninsula blackout on 28 April, requires fast, localized, and coordinated decisions.

Digital twins can simulate scenarios in real time – even integrating data from sensors, legacy systems, and models – to produce a high-accuracy forecast. For example, a digital twin can simulate the impact of sudden cloud cover on solar output and trigger control actions at the neighbourhood level within milliseconds.

Xavier: Swarm computing can further enhance this management by coordinating decision making across a network of decentralized agents – such as substations, smart meters, and edge devices – that act locally and adapt collectively, without requiring centralized control. For instance, if a node detects abnormal frequency or voltage, nearby nodes can rapidly negotiate load-shedding, rerouting, or storage activation to contain the issue. This distributed intelligence supports fault tolerance, scalability, and faster response times. ▶

How is HP2C-DT addressing these issues?

Eduardo: HP2C-DT (see Iraola et al in 'Further reading') addresses the challenge of balancing proximity to physical systems for a fast response through edge nodes with the need to handle heavy computational workloads for simulation and optimization using HPC clusters. The project tackles this by designing and implementing a reference architecture that integrates both aspects.

The edge layer provides a function-as-a-service infrastructure, allowing the deployment of configurable computation units called 'functions' that interact directly with hardware and support flexible event triggering and data aggregation. The HPC layer contributes the computational muscle, serving as the target for offloading the most demanding tasks from other nodes in the architecture.

The cloud layer sits between them as a monitoring hub, centralizing data flow and providing visualization for the operator. All layers are connected through COMPSs agents (see the 2021 paper 'Colony' by Lordan et al in 'Further reading'), deployed on each node, which assess local workloads against available resources and redistribute tasks to less loaded or more capable nodes when needed.

How can COLMENA be applied to electrical grid management?

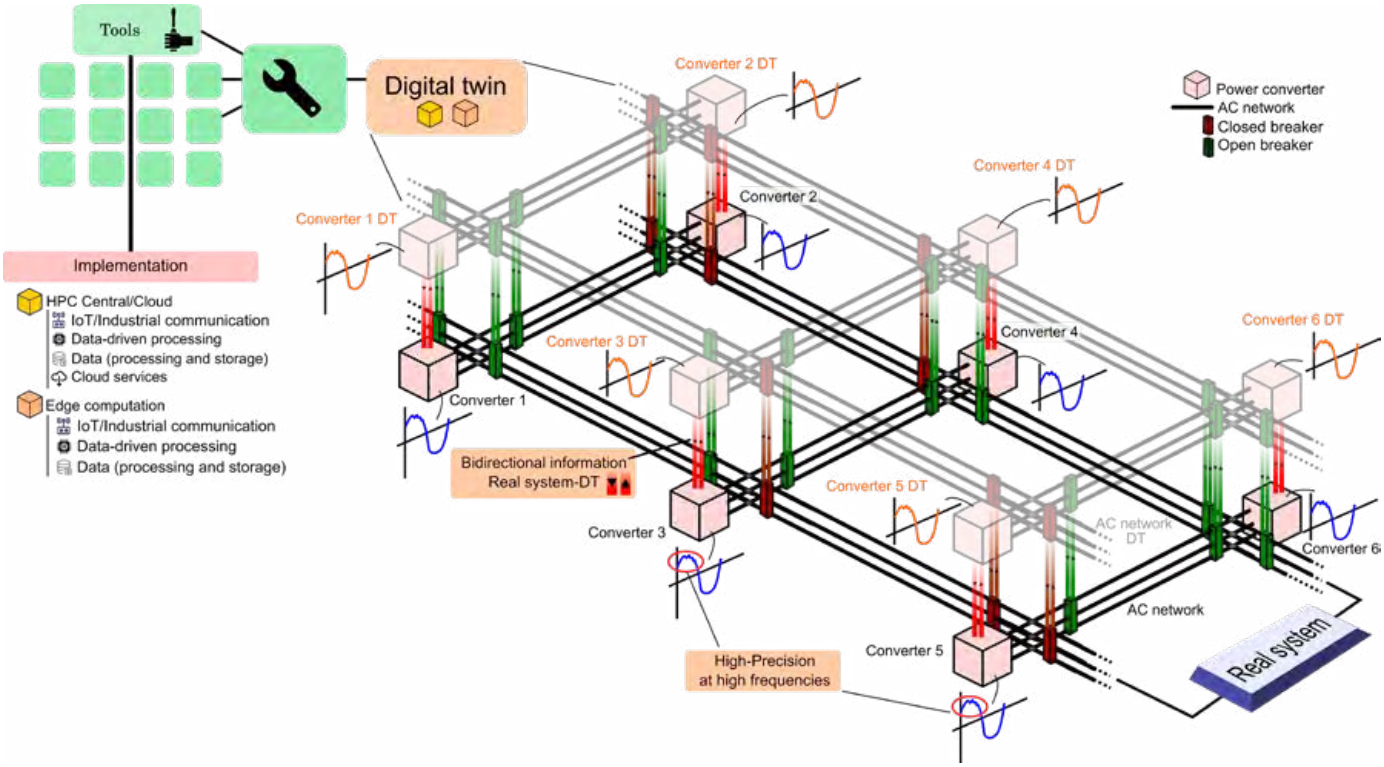
Francesc: COLMENA's software (see the 2024 paper 'Taming the swarm' by Lordan et al in 'Further reading') provides a programming environment where each device becomes an agent capable of making autonomous decisions to ensure the proper functioning of the system. These agents can be embedded in the different components of the power grid, and each agent can carry out multiple functionalities of the system (known as 'roles' in the terminology used by COLMENA).

Xavier: These roles may involve modifying the configuration of the device itself or its internal operation, or they may pursue collaboration with other agents to maintain system-wide stability and efficiency. For instance, an agent might offer a forecasting or storage-balancing service that is consumed by others within a nearby area or across the broader network. This decentralized approach allows agents to dynamically adapt to changing conditions and contribute to collective decision-making processes in a flexible and decentralized way. This enhances the resilience, scalability, and adaptability of the network, particularly as more variable renewables producers join the grid.

Are there real-world examples where HP2C-DT and COLMENA have helped the electrical grid become more resilient and efficient?

Rosa: Neither HP2C-DT nor COLMENA has yet been deployed in a real-world electrical grid. However, both have been tested and validated in controlled laboratory environments.

Eduardo: For HP2C-DT, the testing environment involved a full real-time simulation of the power grid using OPAL-RT hardware (OP4512) running Hypersim (see 'Further reading'), which models electromagnetic transients (EMT) at kilohertz resolution – critical for accurately capturing fast dynamics in renewable-heavy systems. In this setup, the OPAL-RT acts as



the physical grid, while edge nodes are deployed on dedicated compute hardware connected directly to it. The cloud layer is implemented using virtual machines hosted at BSC, and high-performance computing tasks – such as large-scale simulations or AI model execution – run on MareNostrum 5. This integration across edge, cloud, and HPC layers demonstrates the feasibility of HP2C-DT to support advanced grid management functions like forecasting, anomaly detection, and control under realistic conditions.

Xavier: In the case of COLMENA, we have collaborated with eRoots Analytics (see 'Further reading'), a company developing breakthrough technologies for power-grid analysis, control and operation. In the test, autonomous agents monitor grid frequency and respond to disturbances by executing control algorithms that help stabilize the system. To that purpose, COLMENA was integrated with the ANDES power system simulator. The setup used the IEEE 39-bus topology, a standard power system testing benchmark for research and analysis, and we introduced disturbances such as load changes (to simulate consumption variability) and transmission line disconnections. These events were triggered at different times (t = 3s, 18s, and 33s), and the results showed that with COLMENA's decentralized agent-based control, the frequency could be successfully restored after each incident. This demonstrates the platform's potential to coordinate responses in complex, dynamic grid scenarios.

FURTHER READING:

E. Iraola, M. García-Lorenzo, F. Lordan-Gomis, F. Rossi, E. Prieto-Araujo, and R. M. Badia, 'HP2C-DT: High-Precision High-Performance Computer-enabled Digital Twin', 12 June 2025, arXiv: arXiv:2506.10523
doi: 10.48550/arXiv.2506.10523

F. Lordan, D. Lezzi, and R. M. Badia, 'Colony: Parallel Functions as a Service on the Cloud-Edge Continuum', in Euro-Par 2021: Parallel Processing, vol. 12820, L. Sousa, N. Roma, and P. Tomás, Eds., in Lecture Notes in Computer Science, vol. 12820. , Cham: Springer International Publishing, 2021, pp. 269-284
doi: 10.1007/978-3-030-85665-6_17

F. Lordan, X. Casas-Moreno, P. Cummins, J. Conejero, R. M. Badia, and R. Sirvent, 'Taming the Swarm: A Role-Based Approach for Autonomous Agents', in Euro-Par 2024: Parallel Processing Workshops, vol. 15386, in Lecture Notes in Computer Science, vol. 15386. , Cham: Springer Nature Switzerland, 2025, pp. 15–25
doi: 10.1007/978-3-031-90203-1_2

Hypersim, Opal-RT Technologies
opal-rt.com/software-platforms/hypersim

eRoots Analytics
eroots.tech

HP2C-DT (High-precision, high-performance digital twin with computer for modern electrical system applications) and COLMENA (COLaboración entre dispositivos Mediante tecnologia de ENjAmbre) are funded by the Spanish Ministry of Science and Innovation and European Union Next-Generation EU programme.



Funded by the European Union, the P2CODE project is building an open platform for the dynamic management of end-user applications over distributed, heterogeneous and trusted internet-of-things to edge (IoT-edge) node infrastructures. In this article, Eleftherios Mylonas, Alkiviadis Louridas (both of the Independent Power Transmission Operator, Greece) and Kevin Keyaert (Martel Innovate) delve into the project's use cases for a high-voltage substation.

Making utilities more resilient with P2CODE's approach to smart infrastructure

As global energy demands rise and grids become more complex, high-voltage (HV) substations are at the forefront of a digital revolution. Traditionally reliant on legacy monitoring systems, these critical infrastructures have suffered from limited real-time data analysis, leaving them exposed to inefficiencies and security vulnerabilities. Enter P2CODE, a project that is bringing edge intelligence, artificial intelligence (AI), and IoT-powered analytics into the world of utilities.

From reactive to predictive: The role of AI in grid maintenance

Historically, utility companies have had little choice but to react to failures as they happen. But breakdowns in HV substations can have widespread consequences, leading to power outages, operational losses, and costly emergency repairs. P2CODE flips this model on its head by enabling predictive maintenance.

By integrating high-frequency IoT sensors and real-time AI-driven analytics, substations can now detect subtle shifts in power flows and equipment performance before they lead to failures. This means maintenance teams can act before problems escalate, reducing downtime and extending the lifespan of expensive infrastructure.

Developed by the University of Patras, P2CODE's predictive-maintenance use case provides a testbed integrating heterogeneous devices in the form of playback emulators with data collected from a real HV substation, combining legacy substation sensors with modern phasor measurement units (PMUs). Legacy sensors communicate over Modbus, while PMUs provide synchrophasor data. Both data types are collected, decoded and formatted in a standard message format at the gateway app at the edge of the network and then ingested into the message broker, an intermediary program module that provides translation between messaging protocols, which ensures interoperability and orchestration of the telemetry flows.

The key innovation comes from the smart data analytics (SDA) modules, which run a variety of anomaly-detection techniques. Basic threshold-based detection monitors measurements against predefined safe operating limits. Statistical methods compare live values with historical baselines, flagging unusual deviations. Most importantly, the system incorporates AI algorithms based on recurrent neural network (RNN) technology and more specifically long short-term memory (LSTM)-based models, which are especially effective for analysing time-series data from substations. These models learn temporal patterns, enabling them to predict equipment behaviour and detect anomalies that rules alone might miss.

When an anomaly is identified, alarms are raised and sent to operators through a Grafana-based dashboard. The dashboard visualizes both real-time and historical data streams, providing a comprehensive overview of the substation's operational state. This human–machine interface allows engineers to track patterns over time, validate alarms, and schedule interventions before anomaly events escalate into failures.

During the initial testing cycle, the complete workflow of this use case was validated. Data was successfully acquired from Modbus sensors and PMUs, aggregated and formatted in a standard message structure, published to the message broker and stored in a database for long-term storage, analysed by SDA modules, and visualized via Grafana. Alarms were tested under simulated abnormal conditions, and the pipeline operated seamlessly from end to end.

This use case highlights how predictive maintenance in digital substations can be achieved by integrating existing legacy devices with new digital solutions, orchestrated through P2CODE. By embedding AI models based on a combination of statistical methods and RNNs / LSTMs into the analytics layer, the use case shows the potential of combining traditional monitoring with advanced machine learning to achieve real predictive intelligence in the energy sector.

Drones and AI: New eyes on substation security

Beyond operational efficiency, security is a growing concern for power infrastructure. Many utilities sites are remote, with limited on-site personnel, making them vulnerable to intrusion and sabotage. P2CODE is addressing these threats with AI-powered surveillance systems and drone patrols.

The P2CODE intruder-detection use case integrates CCTV cameras and unmanned aerial vehicles (UAVs), both of which stream video to the P2CODE platform. Video feeds are processed by computer-vision modules that perform two core tasks: human detection and face recognition. Detected faces are compared with a database of authorized personnel. If a match is not found, an intrusion alarm is generated.

To support accurate identification and reduce false alarms, a dedicated app for Android, named 'Worker', was developed. Through the Worker app, staff register and store their biometric information, ensuring that authorized personnel are correctly recognized during operations. This integration reduces false positives and ensures alarms are generated only for genuine intrusions.

The operator of the intruder-detection application interacts with the system through a dedicated interface in the substation's security cabin. The interface visualizes real-time video feeds, detected events, and intrusion alarms. When an alarm is triggered, UAVs can be dispatched automatically to the suspicious area, providing dynamic situational awareness and additional video data.

The first testing cycle validated the workflow: video streams were ingested from cameras and UAVs, detections were performed, identities were cross-checked with the UWS database, and alarms were generated in case of unauthorized presence. The coordination between fixed cameras, UAVs, and operator dashboards was demonstrated under realistic conditions.

This use case also involved a proof of concept demonstrating the viability of AI-powered offloading of workloads, whereby computations are dynamically distributed between IoT devices, edge servers, and the cloud. Lightweight tasks, such as basic motion or presence detection, may run locally on the camera node. More demanding tasks, such as face recognition, may be offloaded to the edge, while deeper analysis can be escalated to the cloud when bandwidth and latency allow.

This adaptive strategy ensures a balance between low latency, high accuracy, and efficient resource usage. It also avoids unnecessary transmission of raw video, saving bandwidth – a crucial factor for substations located in remote areas with constrained connectivity. While applied initially as part of the intruder-detection use case, the same task offloading approach can be extended to other digital substation applications, such as predictive maintenance analytics or real-time monitoring of power quality.

Digital backbone: The P2CODE platform

In all the applications described above, the P2CODE platform is the backbone of the system. Via P2CODE, devices are monitored and registered to the system via device-attestation techniques, assuring the integrity of the devices' firmware and configuration. Additional techniques such as identity management are used for device authentication, while the software is also attested for potential vulnerabilities or malware behaviour.

The inherent services and resource orchestrators make sure that the apps are running smoothly across the edge-cloud continuum based on their requirements, while preparing dedicated 5G network slices for fast and reliable connectivity. All these features, along with other automation tools and dedicated user interfaces for platform users, ensure that P2CODE can enable end-to-end, secure, automatic service deployment to the edge-cloud continuum while addressing application requirements for critical infrastructures as in HV substations.

p2code-project.eu


In overcoming the challenges of orchestrating diverse energy sources, the internet of things (IoT) plays a pivotal role. In this article, Greta Mayr and Georgia Knapp show how gridX leverages the IoT to provide the flexibility necessary for the clean energy transition, with targeted use of artificial intelligence and strong cybersecurity protections.

IoT-driven flexibility in smart energy grids

Designing secure, software-centric control

gridX

The clean energy transition is transforming the structure of energy systems. Instead of centralized fossil-fuel plants, modern grids increasingly rely on millions of distributed energy resources (DERs), such as rooftop solar systems, home batteries, electric vehicle (EV) chargers and heat pumps. While these assets promote energy independence, they also bring with them complexity. For their full potential to be realized, each asset must be monitored, controlled and coordinated in near-real time. However, different asset types from different manufacturers do not typically communicate with one another. There is no universal language to enable orchestration of these assets as a smart single unit.

Computing is the key to overcoming this challenge. IoT-enabled data collection lays the foundation of holistic coordination, software stacking opens up multiple use cases for maximum cost savings, and secure cloud platforms provide scalability. In essence, computing technologies provide the backbone for smart energy management – and thus our smart, clean, decentralized energy future.

The IoT: The nervous system of modern energy grids

For the energy transition to succeed, DERs must be active participants in the grid. To achieve this, they need connectivity and intelligence – and this is where the IoT provides the foundation.

Devices vary widely in capabilities, protocols and reliability. Orchestrating them requires sophisticated middleware that abstracts away heterogeneity while still delivering low latency, security and resilient control. IoT infrastructure lays the foundation to intelligently connect energy assets to harmonize consumption and generation, enable remote monitoring and diagnostics, and integrate external control signals, such as forecasts, electricity prices and grid conditions. This distributed computing fabric essentially transforms passive isolated assets into an integrated, flexible resource that grid operators and energy providers use to balance supply and demand and minimize costs.

Flexibility as a computational challenge and a huge opportunity

gridX's *Flexibility Report 2025* analysed how flexible energy assets are activated across Europe, providing a data-driven view of the sector's maturity. The findings underline how flexibility is, fundamentally, a computational problem. Fragmented standards, constantly changing regulation and varying original equipment manufacturer (OEM) interfaces make it difficult to unlock flexibility at scale. Home energy management systems (HEMS) must provide the basis for seamless integration of energy assets, upon which more sophisticated software-driven functionalities can be added to pool the flexibility of residential energy assets and make these assets available for trading. Doing so can unlock up to €800 of financial value per year for end users, while also opening up new revenue streams for utilities.

The ability to aggregate the flexibility of DERs and participate in energy markets requires fast response times (often within seconds), as well as scalable infrastructure to deal with the millions of varied devices that are expected to be installed in the coming years. Managing this complexity requires algorithms that can aggregate different asset types dynamically, predict their behaviour and dispatch them in coordination with grid needs and market signals. Unlike centralized generation, which can be modelled relatively easily, DERs are heterogeneous, user-dependent and often unpredictable.

This is where computing power becomes essential: scalable architectures, efficient scheduling algorithms and predictive modelling are all necessary to ensure flexibility is not only technically feasible but economically valuable as well.

AI forecasting: One tool among many

Artificial intelligence (AI) has been widely promoted as a game-changer in energy. At gridX, we view AI forecasting as an important, but not exclusive, tool.

Forecasting the behaviour of flexible assets – such as EV charging demand, household consumption or solar generation – is inherently uncertain. Traditional statistical models can capture many patterns, but AI can accelerate and improve accuracy by learning from vast amounts of historical and real-time data. Machine learning enables more accurate forecasting and increased efficiency by removing manual processes, such as inputting individual preferences. It also helps efficiently train installers and automate customer support. In the future, AI will play an increasingly important role in helping end users to better understand asset behaviour, such as why storage systems charge or discharge at certain times, and in breaking down complex front-of-the-meter scenarios with multidimensional optimization problems.

That said, AI has its limits. Training and running large AI models is energy intensive and sometimes unnecessary for smaller-scale problems. Many behind-the-meter optimizations (for example within the household) run faster and more robustly without AI – especially when they occur frequently at intervals of seconds. Therefore, we at gridX embed AI only where it adds clear value – as one of several tools in a broader optimization toolkit. In many cases, hybrid approaches (e.g. rule-based control combined with lightweight machine learning models) are more effective and sustainable.

This pragmatic approach ensures that AI forecasting enhances efficiency without becoming a costly or over-hyped dependency.

Cybersecurity: Protecting an interconnected future

As grids become more digital and interconnected, the attack surface expands. Each IoT device, communication channel and software module is a potential vulnerability. Ensuring security is therefore not optional – it's a prerequisite for trust in the energy transition.

In energy management, cybersecurity principles should be embedded throughout the entire architecture. At the device level, cryptographic identities, secure boot processes and tamper detection ensure that assets cannot be easily compromised. Communications between devices and the cloud are protected through end-to-end encryption, robust certificate management, and regular key rotation, ensuring that sensitive data cannot be intercepted or manipulated.

The cloud infrastructure itself is safeguarded by secure software development lifecycles, penetration testing, and automated vulnerability scanning. These measures ensure that the platforms orchestrating thousands of assets remain resilient even under evolving threat landscapes. On top of this, continuous operational monitoring – including intrusion detection, logging and incident response processes – provides an additional layer of defence.

One of the toughest challenges lies in managing software updates. With thousands of devices deployed in the field, keeping them secure requires remote, authenticated, and fail-safe update mechanisms. This demands robust computing pipelines that preserve both security and operational continuity, ensuring that devices remain protected without interrupting their ability to provide flexibility to the grid.

In effect, cybersecurity in energy management is not a supporting feature but a computational discipline in its own right, spanning cryptography, distributed systems and real-time monitoring at scale.

The intersection of computing and energy

The energy transition requires more than renewable generation. It requires intelligent, secure orchestration of distributed assets. IoT provides the nervous system, software architectures deliver control, AI forecasting enhances optimization and cybersecurity ensures resilience. Together, these computing-driven tools enable the flexibility needed to keep our grids stable, efficient, and sustainable. As we move toward a decentralized, digital energy future, computing will not just serve the energy sector – it will define it.

Download the *Flexibility Report 2025* from the gridX website
gridx.ai

Making the transition to a low-emission household is a step many people consider, although they may be deterred by uncertainty about the financial or practical implications. In this article, HiPEAC coordinator Koen De Bosschere (Ghent University) shares his personal experience of moving to domestic renewable-energy generation and lower-emissions devices.

Adventures in low-emissions energy
A personal story

Several years ago, I made the decision to comply with the Paris agreement and try to reduce my personal emissions by 7% per year. I installed solar panels, a home battery, and a heat pump, and I bought an electric car. In parallel, I also tried to reduce electricity consumption at home by replacing old appliances with more energy-efficient ones.

Today I can compare my life before (2015) and after (2025). Here are my conclusions.

Installing a heat pump in a 25-year-old house was technically challenging. Fortunately, my whole house had underfloor heating, which is ideal for a heat pump. The 30 kW gas boiler was replaced by a 6 kW heat pump with a coefficient of performance (COP) of three. The yearly gas boiler consumption was 30 MWh, the heat pump produces 18 MWh of heat per year. The gap of 12 MWh is closed by technological improvements: more intelligent controllers, and a smart thermostat which contains a digital twin that models the thermal behaviour of the house. The last yearly energy bill for the gas boiler was €2,345; the first electricity bill for the heat pump was €1,875, which is 20% lower.

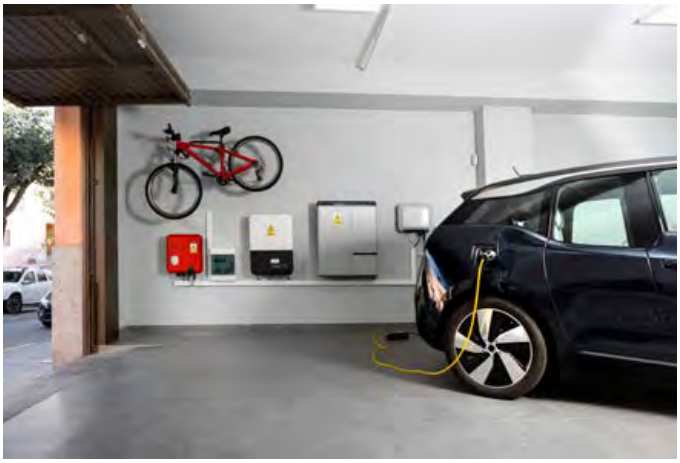
I installed 6.5 kW solar panels and added a home battery of 10 kWh to optimize domestic use of the electricity generated. Of the total yearly electricity production, 34% is immediately used by the house, 31% is stored in the battery for later use, and 35% is sold to the grid. During the summer months, around 49% of the domestic consumption is directly provided by the solar panels and 42% is provided by the battery. Only 9% comes from the grid (during cloudy days). During the winter months, 15% is direct consumption and 23% is provided by the battery, while 62% is provided by the grid – increasing to 82% in December.

There is room for optimization. On average per year, I sell 35% of my generated electricity to the grid (2.2 MWh in the summer, €80), and I buy 31% of my consumed electricity (1.9 MWh in the winter, €630). It would be better to find a purpose for the 2.2 MWh in the summer than to sell it for almost nothing to the grid company.

I bought an electric car with a driving range of 500 km to avoid range anxiety. Charging at home is the cheapest option: €0.33/kWh in the winter, and almost free in the summer (€0.033/kWh which is what the grid company offers me). For an average daily distance travelled of 20 km in an urban environment (approximately 8,000 km per year; the rest is done by e-bike), the car needs around 3 kWh per day. The average cost to drive the car 20 km per day is €0.5 (compared to €2.5 for a comparable petrol-based car). This saves me €600 per year compared to a petrol-based car. On top of this, the government adds a tax break of €250 per year, which gives me a saving of €850 per year. The car was €7,000 euro more expensive than a petrol-based car, and I hope to keep the car for at least 15 years. That means a return of €12,750 on an investment of €7,000, which corresponds to a compound yearly interest rate of 7%.

Solar panels on a Dutch house. Credit: Hilda Weges | stock.adobe.com

A smart electric meter in the UK. Credit: Ming | stock.adobe.com



Charging an electric car in a garage. Credit: BasilicoStudio Stock | stock.adobe.com

Charging station in Germany. Credit: Anne Czichos | stock.adobe.com

My conclusions:

1. Some people don't like the quality of the fossil-free alternatives like electric cars or heat pumps, but I did not experience any inconveniences. The new heating system works as well as the old one, while electric cars are nice to drive, and there are enough public charging points in the part of Europe where I live. The only disadvantage is that one must own a house and have a well-oriented roof to install solar panels, a home battery and a heat pump. To charge a car at home, it helps to have a driveway.

2. I was surprised to discover that my electricity bill in 2025 so far is 20% lower than my combined electricity + gas + car fuel bill in 2015 (in absolute terms). Considering the inflation of 33.7% over the same period, the reduction is 40%, which corresponds to a compound yearly reduction of 5%. This means that, in addition to being good for the climate, the initial expenditure has turned out to be a sound financial investment. The total cost of the solar panels and home battery was around €14,000, but they save me about €1,400 per year, which means the investment has a return of 10% – higher than the return on a risk-free investment at the bank. There is no reason not to do it if one has the money.

3. The biggest challenge is to manage the different subsystems of the home energy system.
   a. Solar power that cannot be used or stored locally is basically worthless on the energy market. Hence it is important to find a use for it. Obvious applications would be producing hot, sanitary water, or cooling the house.
   b. Batteries lose 10–20% of energy in the charging/discharging process. Charging one battery from another (like charging the car battery from the home battery or vice versa) does not make sense because the losses add up.
   c. Home and car batteries should not only be able to store energy from solar panels, but also from the grid when the electricity price is very low. This would also make them useful in the winter and can help balance the grid.
   d. It is worth analysing domestic energy consumption. As an example, the standby power of my house is about 100 W, which is 0.8 MWh per year, or 10% of household energy consumption. It would be good if appliances consumed less standby power.

Currently, the components of a home energy system often do not cooperate (e.g. a car draining a home battery when charging), and managing them can be complex for typical homeowners, both technically and in terms of monitoring. A more professional approach would be to delegate the management of the entire home-energy system to a service company that optimizes its use and makes suggestions of how to invest the savings into additional optimizations.

To sum up, my experience of investing in electrification and power generation at home has been positive, both financially and for the environment. Pre-requisites include owning a property with the right characteristics to house the necessary devices and having the capital for an initial upfront investment.

In the future, I would like to see more intelligent, user-centred services for managing home-energy systems to fully optimize them – which seems an obvious opportunity for computing startups and energy providers.

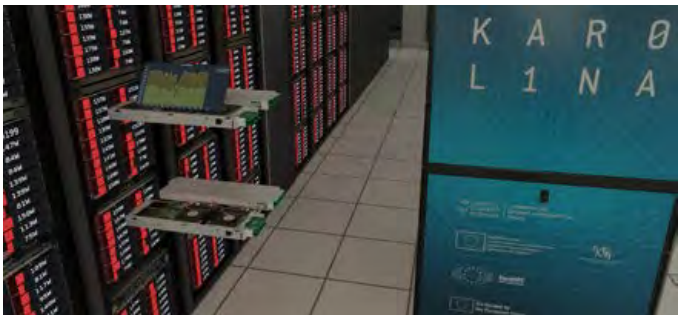
High-performance computing (HPC) drives scientific progress, but its rapidly growing energy demand poses major sustainability challenges. Over the past decade, the MERIC Software Suite, developed at IT4Innovations National Supercomputing Center (Czechia), has evolved into a comprehensive solution for datacentre power monitoring, management, and optimization. In this article, Ondřej Vysocký (IT4Innovations) explains how MERIC enhances energy efficiency and sustainability across heterogeneous HPC environments.

MERIC: A decade of energy efficiency in HPC

In 2015, the READEX project was launched with the ambition of developing an automated tool to optimize the energy consumption of parallel HPC applications through dynamic tuning of power management knobs. As part of this initiative, the MERIC runtime system was created to measure energy usage and perform manual tuning of Intel processor parameters. New lines of research and development led to the establishment of the Energy Efficiency Research Group at IT4Innovations, which continues to develop methods and tools that help HPC systems to deliver top performance while remaining energy efficient.

Today, the MERIC runtime system, together with its power-management knob control services, provides a unified interface for users to monitor and optimize energy consumption across diverse hardware architectures, including Intel and AMD central processing units (CPUs), Nvidia and AMD graphics processing units (GPUs), Fujitsu A64FX, and others. Users can enable applications to measure both energy consumption and power-management-related metrics per application regions, applying region-based tuning for maximum power savings while respecting user-defined performance constraints.

The MERIC software suite consists of many additional components. The MERIC suite extends standard datacentre monitoring infrastructure with power and power management-related daemons, providing instantaneous and averaged measurements, including power, temperature, frequency, floating-point operations per second (FLOPs), and memory bandwidth. These metrics allow the evaluation of hardware power management behaviour under real workloads. Job-level energy and carbon accounting combines monitoring data with scheduler information, delivering scheduler-independent reporting. Users can retrieve energy consumption for CPUs, GPUs, nodes, and for the whole system, which includes energy consumed by the infrastructure, and respective carbon emissions. This approach enables precise assessment of energy usage and associated emissions, supporting both operational decisions and sustainability reporting. An additional feature of the MERIC suite is 3D visualization of data-hall infrastructure enhanced by real-time and historical data from the system monitoring, offering deep operational insight and situational awareness.



On the administrative side, MERIC implements dynamic, rule-based power capping, enforcing cluster-wide limits on nodes, racks, or system partitions. Running-window analysis ensures that average power consumption respects thresholds while avoiding unnecessary performance throttling.

The MERIC suite modular architecture integrates seamlessly with existing HPC monitoring and scheduler systems, making it scalable for large HPC and artificial intelligence (AI) facilities. The suite is deployed in EuroHPC systems Karolina and Deucalion, enabling the wide range of its energy efficiency capabilities for both administrators and system users.

The development and ongoing evolution of the MERIC have been supported by the Czech national research and development e-infrastructure e-INFRA CZ, and several European projects, including EUPEX, POP3 Center of Excellence, SEANERGYS, and DARE.

After a decade of continuous development, MERIC provides a proven, scalable pathway towards sustainable HPC, combining advanced monitoring, user-space control, and policy-driven energy management to address the operational and environmental challenges of modern HPC systems.

FURTHER INFORMATION:
MERIC Energy Efficiency HPC SW Suite
code.it4i.cz/energy-efficiency/meric-suite

SECDA-TFLite v2: Open-source toolkit for HW-SW co-design of FPGA-based AI accelerators


Jude Haris and José Cano, University of Glasgow

As the use of artificial intelligence (AI) grows, with services such as ChatGPT reaching over 700 million users weekly, there is a need for efficient computational devices for AI workloads. In addition, the complexity and scale of large language models (LLMs) used within the latest AI tools and services have significantly increased. As such, researchers have been focused on designing new and efficient accelerators for resource-constrained hardware devices using field-programmable gate arrays (FPGAs).

Reprogrammability and fine-grained control that FPGAs allow for innovative architecture design and space exploration. Unfortunately, the design, integration, and deployment of these FPGA-based accelerators require hardware-software co-design experts to integrate within an existing stack of software, develop hardware designs, and implement driver code to connect the accelerators with existing host devices.

Using the SECDA-TFLite toolkit (see 'Further information', below), we reduce the time and effort taken from idea formulation to hardware evaluation. The toolkit uses the SECDA design methodology, which allows the designer to iteratively improve the hardware-software solution through a simulation-based design loop that avoids lengthy logic synthesis while eliminating the adaptation phase usually needed to go from a simulated design to real hardware design. SECDA-TFLite specifically focuses on enabling integration of new AI accelerators for Google's TensorFlow Lite (also known as LiteRT). This integration helps us concentrate on edge-based resource-constrained hardware where the amount of compute, memory, and critically, the power budget is limited.

With the initial release of SECDA-TFLite, we garnered interest from fellow academics, worked within the dAIEDGE European project, and developed accelerators for varying topics such as generative AI and power-of-two-based quantization. Over the last few years, we have been active in development, and we have finally released SECDA-TFLite v2.

With the new release of SECDA-TFLite v2, we have included many new features and additional tooling that allow for easier and faster design and evaluation of new FPGA-based accelerators. Our new benchmarking suite allows researchers and hardware designers to quickly evaluate new TFLite-based AI models with their accelerator, using easily configurable JSON files that provide both simulated and real hardware performance results.

To complement this, we have also developed a hardware automation tool, which takes accelerator-design source files and automatically generates an FPGA bitstream for the target FPGA device using high-level synthesis (HLS) and logic synthesis (HLX). We have tested and confirmed support for PYNQ-based FPGA boards, and will add more official support as development continues.

Other additional features include a SECDA profiler to enable fine-grained profiling and analysis of new designs, SECDA delegate generation (supporting 11 new operations / layers), remote HLS / HLX server support, and VSCode DevContainers-based installation.

The project is in active development along with other SECDA-related projects. We aim to have a community-driven research platform around our tools, and we encourage colleagues from both academia and industry to reach out to us for questions, requests, feedback, or potential points of collaboration.

FURTHER INFORMATION:

J. Haris, P. Gibson, J. Cano, N. Bohm Agostini, and D. Kaeli, 'SECDA-TFLite: A toolkit for efficient development of FPGA-based DNN accelerators for edge inference'. Journal of Parallel and Distributed Computing, March 2023.
doi: doi.org/10.1016/j.jpdc.2022.11.005

J. Haris, P. Gibson, J. Cano, N. Bohm Agostini, and D. Kaeli, 'SECDA: Efficient hardware/software co-design of FPGA-based DNN accelerators for edge inference'. Proc. 33rd IEEE International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD), Belo Horizonte, Brazil, Oct. 2021.
doi: doi.org/10.1109/SBAC-PAD53543.2021.00015

Github repository – SECDA-TFLite
https://github.com/gicLAB/SECDA-TFLite

Contact: Jude.Haris@glasgow.ac.uk

Squeezing more performance out of computing systems has traditionally been the primary goal of computer architects. However, while performance shouldn't be sacrificed, it's essential that computer architecture adequately addresses energy efficiency and reliability, argues George Papadimitriou (University of Patras).

Towards energy-conscious and resilient computer architectures

Emerging computing paradigms, particularly in artificial intelligence (AI) and large-scale distributed environments, expose the limitations of performance-centric design, emphasizing the need for holistic optimization strategies. Today's architectures must operate under strict energy constraints and deliver reliable outcomes in unpredictable environments.

As architectural complexity increases, so does the opportunity to rethink our design priorities, not just focusing on performance, but also on how we integrate reliability and energy efficiency from the ground up. This means that these aspects must be addressed at the microarchitecture level, not merely through software-level optimizations or power-management techniques. This is especially relevant in emerging domains such as edge computing, autonomous systems and safety-critical cyber-physical systems, in which hardware faults, defects, energy variability, and timing violations can no longer be treated as peripheral concerns.

To this end, simulation and modelling become critical enablers. By leveraging timing-accurate simulators and microarchitecture-level resilience analysis frameworks, researchers can quickly analyse how architectural and microarchitectural decisions propagate through the computing stack, impacting both system-level reliability and energy profiles. These tools help evaluate design choices early on, allowing for smarter and more reliable co-design of hardware and software for future processors and AI accelerators. Open and flexible architectures like RISC-V are becoming more important, as they make it easier to experiment with virtualization features and build custom accelerators tailored to specific needs. Combining RISC-V with microarchitectural fault models and energy-aware simulations provides a foundation for building domain-specific architectures that are tailored to the needs of edge AI, or embedded control systems, all under constrained power envelopes.

Reflecting these priorities, at the Technology and Computer Architecture Lab (TCAL) in the Computer Engineering and Informatics Department (CEID) of the University of Patras, Greece, we explore, among other topics, how modern hardware, ranging from lightweight processors to domain-specific and edge accelerators, can be co-designed with resilience and energy efficiency in mind. We focus on key challenges in analysing resilience and modelling reliability at the microarchitecture level, working across different instruction set architectures and specialized accelerators, using both detailed simulators and real hardware systems. A key research direction, which we strongly believe aligns with European trends toward technological autonomy and the growing emphasis on green computing, focuses on the cross-layer integration of reliability and energy awareness.

This shift does not mean performance is being sacrificed. It highlights the challenge of making trade-offs that keep things efficient, given the growing complexity and constraints of modern computing environments. Achieving this vision demands systematic analysis, reproducible experimentation, and transparent evaluation methodologies; principles we need to pursue by accurately extending and enhancing existing simulation platforms and integrating fault, power, and sustainability models from the ground up.

Academic research has a clear role to play, not just in proposing new techniques, but in building the tools, frameworks and methodologies that industry and system designers can trust. As we move towards a more energy-constrained and reliability-critical computing landscape, our challenge is to ensure that computer architecture research evolves in parallel, remaining measurable, explainable, and aligned with the real demands of future systems.

FURTHER INFORMATION:

Technology and Computer Architecture Lab (TCAL), Computer Engineering and Informatics Department (CEID), University of Patras, Greece

tcal.ceid.upatras.gr

Credit: Tri | Adobe Stock (generated with AI)

RWTH Aachen spinoffs special

MachineWare's virtual platforms and Roofline's edge AI compiler

In this special edition of our regular look at deep-tech companies, we learn about two startups from Aachen, focusing on simulation solutions and edge AI compilation, respectively.

HOW MACHINEWARE GMBH DELIVERS TANGIBLE BENEFITS THROUGH VIRTUAL HARDWARE

Lukas Jünger, MachineWare

Founded in 2022 and headquartered in Aachen, Germany, MachineWare GmbH is a rising innovator in the field of high-performance simulation solutions for embedded software development and testing. As a spin-off from the renowned Institute for Communication Technologies and Embedded Systems (ICE) at RWTH Aachen University, MachineWare builds on decades of academic excellence to bring cutting-edge simulation technologies from research into industrial practice.

Bridging the gap between hardware and software

At the heart of MachineWare's offering are virtual platforms (VPs) – comprehensive, functional simulations of embedded systems that run on standard computers. These platforms empower software engineers to begin development well before physical hardware is available, dramatically reducing time-to-market while improving software quality, performance, and security. Their scalable architecture enables teams to tackle the complexity of modern software stacks with confidence and precision.

Powered by next-generation simulation models

MachineWare's virtual platforms are driven by two proprietary processor simulation models:

- SIM-V™: An ultra-fast RISC-V simulator
- SIM-A™: A high-speed Arm simulator

These simulators leverage advanced technologies such as MachineWare's FTL just-in-time dynamic-binary-translation engine for accelerated processor simulation and the Arm-on-Arm simulation hypervisor for native execution. The result: faster test cycles, earlier insights, and significant cost savings for development teams.

Open-source innovation with VCML

To streamline the creation of virtual platforms, MachineWare offers VCML (Virtual Component Modeling Library) – an open-source library built on the IEEE SystemC TLM-2.0 standard. VCML provides a rich set of modelling primitives and ready-to-use models of industry-relevant hardware components, making it an ideal starting point for both professionals and newcomers to SystemC TLM.

MachineWare actively fosters the VCML community by curating a list of open-source projects that leverage VCML's capabilities. VCML also serves as MachineWare's integration hub for other tools, including:

- Arm FastModels: Arm's high-fidelity simulation models
- MachineWare QBox: A co-simulation solution combining QEMU with SystemC
- Vector SIL Kit: An open-source framework for building compound simulations that integrate multiple system simulators provided by Vector Informatik GmbH

Seamless integration across industries

MachineWare's solutions are designed for versatility and compatibility. They integrate smoothly with a wide range of third-party tools and frameworks, including Lauterbach Trace32, IAR Embedded Workbench, TASKING Toolchains & Debugger, SEGGER emStudio, tracetronic ecu.test, Rapita Systems RapiCover, dSpace VEOS, Vector CANoe and more.

This broad ecosystem support makes MachineWare a valuable partner across industries – from automotive and aerospace to industrial automation and beyond.

FURTHER INFORMATION:

MachineWare website machineware.de

VCML community projects machineware.de/vcml-community

HOW ROOFLINE'S MODEL-TO-CHIP MATCHMAKING POWERS LIGHTWEIGHT EDGE AI

Thomas Zimmerman, RooflineAI GmbH

Located in Aachen, Germany, RooflineAI GmbH is a startup on a mission to enable edge AI products in the disruptive area of physical AI. With its innovative, retargetable edge AI compiler, it expands the range of chips innovators can harness, opening up hundreds of potential use cases.

Physical AI relies on local, low-power and low-latency execution of complex AI models. Currently, however, building edge AI products is still painful. While AI developers want to run models with minimal effort – just like in the cloud – the deployment process for edge AI systems is often unpredictable. The model might not be supported, performance might be slower than expected, and the tool flow is often complicated.

Roofline aims to solve this with a flexible, efficient, and simple software development kit (SDK) that takes any trained AI model and matches it with the best-suited chip. Developers can directly access Roofline's SDK via the machine-learning library PyTorch, with models taken and adopted from Hugging Face or self-trained. This is enabled by Roofline's deep technical innovation, an edge AI compiler. On the application side, it is compatible with all developer frameworks and any model architecture, from conventional vision networks to the latest large language models (LLMs). On the hardware side, it supports all central processing units (CPUs), all mobile graphics processing units (GPUs), and is onboarding the first two dedicated AI accelerators (NPUs), together with customers.

Founded in 2024 by Jan Moritz Joseph, Maximilian Bartel, and Thomas Zimmermann, the company was spun off from RWTH Aachen University and successfully raised multiple funding rounds. The team consists of experienced experts from AMD, Intel and Vector Informatik, and is growing swiftly. In 2024, Roofline's success in transforming research into a commercial venture was recognized by a HiPEAC Technology Transfer Award.

As a major milestone, Roofline has recently been selected by the European Innovation Council (EIC) Accelerator, receiving €2.5 million in grant funding and a pre-committed equity investment for their next funding round. The EIC Accelerator is Europe's most competitive deep-tech funding program. It supports high-impact innovations across Europe, providing substantial financial backing and business acceleration services. Roofline's successful EIC submission, titled 'Retargetable AI Compiler Technology for Scalable Edge Deployment of Next-Generation AI Models', will boost Roofline's core technology development and enable faster growth and better market entry. In this EIC round, only 40 startups were selected from nearly 1,000 applicants. Winning this award just a year after foundation is an endorsement of the company's strong technology achievements.

Roofline is actively hiring for junior and senior roles in C++/compiler development, DevOps, and sales, and welcomes partners for publicly funded collaborations. Learn more by visiting the company website, or reach out directly to the founders to explore the product, the roadmap, and your opportunity to contribute.

FURTHER INFORMATION:

RooflineAI website
roofline.ai

RooflineAI on LinkedIn
linkedin.com/company/rooflineai

"Roofline's SDK takes any trained AI model and matches it with the best-suited chip"

As many readers will know, HiPEAC organizes a range of activities aimed at early-career researchers wishing to find out about career opportunities and supporting companies to identify the right talent to take their innovation forward. In this article, HiPEAC Talent Manager Rosana Cortés (Barcelona Supercomputing Center) updates us on HiPEAC Jobs activities and looks ahead to the HiPEAC conference in Kraków on 26-28 January.

HiPEAC's got talent

How the pan-European network helps match candidates to careers

New-look HiPEAC Jobs Portal

If you haven't visited the HiPEAC Jobs Portal recently, it's worth a look! With a new layout and refined search and filtering tools, finding opportunities across Europe is now easier than ever, whether you're looking for a funded doctoral programme, an internship or a full-time position.

For companies, research institutes and universities, the portal is an excellent way to reach highly qualified candidates. In a recent survey, the portal achieved an overall recommendation score of more than 8 out of 10. Users particularly valued its centralized access to European jobseekers (71%), reach and visibility among early-career researchers (67%), and its effectiveness in attracting suitable candidates (48%).

hipeac.net/jobs

ACACES careers roundtable

The careers roundtable at HiPEAC's annual summer school has become a staple, attracting numerous students wishing to hear from renowned computer scientists and engineers in industry and academia, as well as entrepreneurs. In a relaxed setting, panellists offer advice on key priorities in different sectors, drawing on their personal stories to show how careers paths are rarely linear.

hipeac.net/acaces

HiPEAC conference: STEM Day and Student Challenge

For several years, the STEM Day at the HiPEAC conference has allowed students to make meaningful connections with key industry representatives, find out about the latest recruitment opportunities and learn first-hand what it's like to work in different settings. Over the years, STEM Day activities have been refined to offer a highly productive event for participants. These activities include:

- attending the keynote talk
- joining an interactive industry tour
- uploading CVs to the jobs portal and arranging interviews with participating companies
- taking part in the 'Inspiring Futures' careers session, where students can learn about working in different sectors from experts in the field

In a survey sent after the 2025 STEM Day in Barcelona, students gave the event an overall rating of 4.7 out of 5. Respondents highlighted benefits such as finding out about innovative products and technologies, improving communication skills with business representatives, and learning about companies and career opportunities.

Complementing the STEM Day, which mainly targets students from the region hosting the conference, the annual HiPEAC Student Challenge brings together international teams of students to tackle a technical problem. For the 2026 edition, teams can choose to define their own problem, reproduce a scientific paper, or contribute to the RISC-V ecosystem. Selected teams will present their results in Kraków and receive feedback from senior researchers and professors.

The STEM Day is taking place on 27 January during the HiPEAC conference in Kraków.
For further information, visit
hipeac.net/conference

Students visiting the HiPEAC 2025 exhibition (left) and discussing the visit (right)

The poster session at ACACES 2025 was once again a triumphant display of early-career research. In this article, Salvatore Bramante (IMT School for Advanced Studies Lucca) summarizes his poster, which explored how model learning algorithms can expose hardware vulnerabilities.

My ACACES poster

Inside the black box: Unveil hardware vulnerabilities with model learning algorithms

Authors: Salvatore Bramante (IMT School for Advanced Studies Lucca), Matteo Busi (Ca' Foscari University of Venice), Alessandro Cilardo (University of Naples Federico II), Riccardo Focardi (Ca' Foscari University of Venice), Flaminia Luccio (Ca' Foscari University of Venice)

Modern computing architectures have become more susceptible to side-channel attacks that leak information based on physical properties. Cache side-channel attacks are a very insidious category of vulnerabilities: they can steal secret information by analysing the timing of memory access patterns in an execution. Due to increasing computer architecture complexity, these types of attacks are becoming stealthy and difficult to identify. To address this, hardware security researchers increasingly rely on formal verification automated tools to identify potential vulnerabilities.

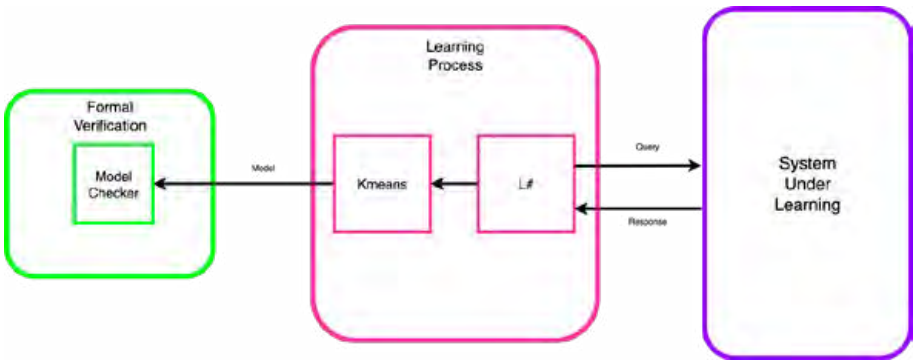
While custom tools tailored to specific hardware are often developed for this purpose, a general technique called active automata learning (AAL) can be used to extract a finite-state machine (FSM) from an actual microcontroller, hereafter called system under learning (SUL). More specifically, an AAL algorithm repeatedly interacts with the SUL to execute assembly instructions, obtaining different execution traces needed to reconstruct global system behaviour.

For our project, we ran an AAL algorithm on the PULPino, an open-source RISC-V microcontroller of the PULP platform, to identify cache-based timing attacks. We ensured that the learning algorithm could observe each operation's duration via the rdcycle instruction.

But here there is a catch: the AAL algorithm fails when the system response contains information about execution time. The AAL algorithm we used only works for deterministic automata, and caches introduce non-determinism in timing because of the hit / miss mechanism. To avoid non-determinism, we implemented a k-means clustering algorithm that automatically decides whether a transaction is a hit or a miss transaction, without relying on a static threshold. Once the learning succeeds, we have a (probably) correct and deterministic FSM modelling the timing behaviour of the system. We then use a model checker to verify whether the system's specifications have been respected. This kind of verification is very powerful since it can lead to an exhaustive discovery of unknown vulnerabilities, looking at the evolution of the SUL behaviour.

Looking at the whole workflow, is it just like a fuzzing technique? While both methods repeatedly test a system with different inputs, the real difference is that AAL adapts its inputs to discover the SUL machine state, giving precise guarantees on the learned FSM.

This project demonstrates how formal methods can systematically unveil hardware vulnerabilities, providing an explainable representation of information leakage patterns. The open-source nature of our toolchain makes it accessible to the hardware security research community. We are planning to enable reproducible vulnerability analysis across different ISAs and make formal verification easier and comprehensible for all embedded system engineers, making future processors more resilient.



In this edition of our thesis series, Gorka Nieto explains how he used reinforcement learning to determine where best to execute computations on the computing continuum, balancing latency, energy efficiency and resource availability.

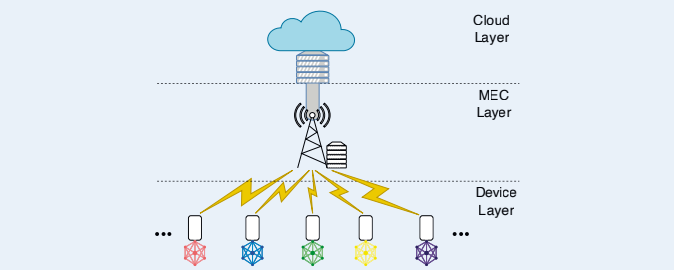
Three-minute thesis

NAME: Gorka Nieto
RESEARCH CENTRES: Ikerlan Technology Research Centre, Basque Research and Technology Alliance (BRTA) and University of the Basque Country (UPV/EHU)
SUPERVISORS: Idoia de la Iglesia (Ikerlan/BRTA), Unai Lopez-Novoa (UPV/EHU)
THESIS TITLE: Reinforcement learning enabled offloading of applications in the 5G edge-cloud continuum

We are entering an era of pervasive connectivity, where technologies like the internet of things (IoT), 5G, and multi-access edge computing (MEC) enable real-time interaction among countless devices. This infrastructure brings extraordinary potential, but also a fundamental challenge: how can we decide, in real time, where to execute each computing task that composes an application, balancing latency, energy efficiency, and resource availability?

This PhD work addresses that challenge by framing task offloading as a core optimization problem in distributed intelligent systems. Our approach relies on deep reinforcement learning (DRL), a technique that learns directly from the environment, without labelled data, to make adaptive, real-time decisions in dynamic contexts. We perform this decision-making in a decentralized manner directly on edge devices, rather than relying on centralized servers. We embrace lightweight DRL agents that enable real-time, adaptive decision-making with minimal overhead, enhancing scalability and reducing decision latency.

We began by developing a lightweight actor-critic model that selects where to execute each task, based on different parameters such as central processing unit (CPU) load, or wireless channel conditions, among others. This model makes offloading decisions based on minimizing the energy consumption of the device, while guaranteeing incoming tasks' latency requirements. We considered two scenarios with multiple devices, each running an instance of the DRL model and competing for shared resources, such as the wireless channel or the server's CPU. We evaluated the algorithm through the simulation, and results showed that it outperformed other baselines in terms of success and / or energy consumption in dynamic environments.



The next step was to compare DRL algorithms with control theory algorithms such as Lyapunov optimization. In a different set of scenarios, simulation results show that, while Lyapunov optimization excels in static environments, DRL approaches are more effective in dynamic settings, and better adapt to changing environments. This experimentation was carried out using the simulation platform ITSASO: github.com/tlmat-unican/ITSASO

The final step of the dissertation was to evaluate our proposal in a real-world scenario, by deploying DRL models on different edge devices, like a Raspberry Pi or a Jetson AGX Orin. We analysed the feasibility of running this kind of algorithm in constrained devices, using a real-world testbed where channel conditions and servers' status change over time. We also explore the trade-offs between training locally versus offloading the training to a server depending on the device, focusing on latency and energy.

The main conclusion of this work is that DRL agents can be implemented on edge devices successfully, and are able to make decentralized real-time decisions that reduce latency, and enhance system resilience on real-world deployments.


Gorka's supervisors **Idoia de la Iglesia** and **Unai López-Novoa** commented: 'Gorka's thesis provides a practical and forward-thinking solution to a key challenge in the modern edge-cloud continuum: how to make intelligent decisions about where computation should occur in complex environments, in a distributed and decentralized manner. It's a timely and relevant contribution that has applications in many domains like industry or autonomous systems, where responsiveness and efficiency are critical.'

HiPEAC conference

26-28 January 2026

Kraków

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

hipeac.net/conference

#HiPEAC26

Funded by the European Union