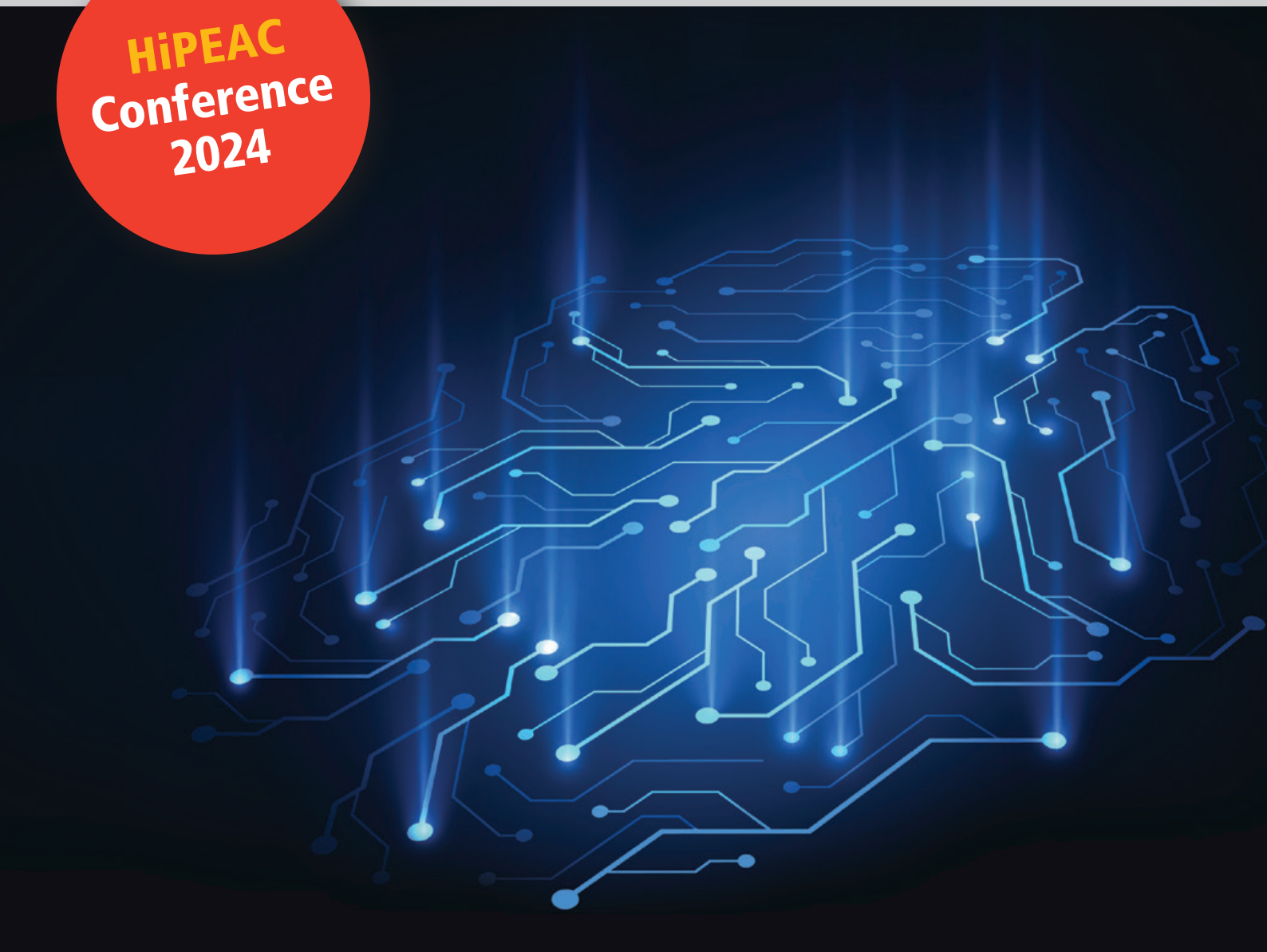
HiPEAC info 71

JANUARY 2024

HiPEAC Conference 2024



Entering the next computing paradigm: The HiPEAC Vision 2024

Spinning out the smarts: Powering edge AI and distributed computing

Lieven Eeckhout on sustainability, Reetuparna Das on data-centric architectures and Mitsuhisa Sato on building a top supercomputer

contents

Lieven Eeckhout on computing and sustainability

Moving to data-centric architectures with Reetuparna Das

Mitsuhisa Sato on the road to zettascale

Entering the next computing paradigm: The HiPEAC Vision 2024

Special feature: Edge AI and distributed computing

Stuart Cording interviews Frontgrade Gaisler's Sandi Habinc

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: Panuwat on Adobe Stock

Design: www.magelaan.be

Editor: Madeleine Gray

First of all, I would like to wish you a healthy and prosperous 2024 – both personally and professionally. I will remember 2023 as the year of the breakthrough of generative artificial intelligence. Almost everybody has heard about the large language model ChatGPT, and millions of people use it regularly.

Unfortunately, 2023 was also an annus horribilis: we are witnessing two horrifying and destructive wars costing tens of thousands of lives at the border of Europe: in Ukraine, and in Israel and Palestine. I hope both will come to an end in 2024.

The year 2024 will also be important at the political level. Globally, 40 countries, representing 41% of the world's population and 42% of global gross domestic product (GDP), according to calculations by Bloomberg Economics, will elect a new government. Among these, the most crucial elections are arguably the Unites States (US) presidential elections, the elections in the European Union (EU) and those in Russia. The outcome of these elections will determine the future leadership of the US, Europe and Russia, and this might lead to a different geopolitical situation in 2025, and to a safer or more dangerous world.

For HiPEAC, the year 2024 will be special too. HiPEAC1 formally started on 1 September 2004, funded by the seventh framework programme (FP7). HiPEAC has since then benefited from uninterrupted funding by the European Commission, in the early years as a network of excellence, and later as a coordination and support action. In 2024 we plan to celebrate 20 years of HiPEAC.

I am also pleased to announce that HiPEAC has been instrumental in the creation of a new project, DISCOVER-US, which aims at strengthening the research collaboration between the US and Europe in the domain of distributed computing and swarm intelligence. In the coming months, you will receive more information about events and about calls to fund EU-US collaboration.

In January 2024 we also launch the HiPEAC Vision 2024 with a new set of recommendations to tackle the many challenges of this time, ranging from the next web, artificial intelligence (AI), accelerators, cybersecurity and sustainability. I encourage you to delve into it, and get inspired by it.

Many of you will read this magazine at the HiPEAC conference, which is our flagship event. I wish you a fruitful conference, and I hope to meet you in Munich or at another HiPEAC event in 2024.

Take care

Koen De Bosschere, HiPEAC coordinator

Willkommen in München! Welcome to Munich!

The third-largest city in Germany, and the capital of Bavaria, Munich is famous for being an engineering hub, as evidenced by the headquarters of companies such as BMW (Bayerische Motoren Werke), Siemens and Infineon. With mountain ranges including the Bavarian Alps in close proximity, and a wealth of excellent museums to visit, it's also well known for its beer, Weisswurst and Brezeln. HiPEAC caught up with HiPEAC 2024 local host Stefan Wallentowitz (Hochschule München University of Applied Sciences) to learn more about this thriving German city.

What makes Munich an ideal location for the HiPEAC conference?

Munich is one of the most thriving technology hubs of Europe. Beside the many well-known large industries it is home to many small / medium enterprises (SMEs) and startups, along with two of the leading German universities and many research institutions. HiPEAC will therefore take place in an innovative environment, and HiPEAC will be a highlight and bring many great minds to Munich.

The quality of life is also great, which many people may even associate more with Munich. I think overall Munich is an excellent package to meet people in a great, innovative technology hub, and have great discussions over a soft drink or a famous Munich beer.

The city is famous for its architecture, including the Rathaus (city hall) on Marienplatz and the Siegestor (Victory Gate)

Munich's famous Hofbräuhaus is the venue for the HiPEAC 2024 social event

Can you tell us about the computing ecosystem, both industrial and academic, in the local area?

The ecosystem around Munich is probably most renowned for the automotive and industrial giants that are headquartered in Munich. There are many other large enterprises here, which have a huge demand on computing specialists across all parts of the stack, and across domains.

Of course, given the presence of four large universities and a supercomputing centre, the academic ecosystem is quite thrilling. Beside the supercomputer, industrial engineering, semiconductor design and many other established fields, there are also interesting new initiatives like Munich Quantum Valley, which brings together a large representation of German industry and academia in building a hub for quantum computer expertise.

What are some of the most interesting initiatives in HiPEAC topics at your university?

As the fourth largest university in Munich we are traditionally focused around practical education and not too much on fundamental research. But we are very strong in applied research, and I have been building a group around my strong background and involvement in RISC-V, where I am on the board of directors. We are actively involved in open-source chip design. As such, we are looking into how to build innovative chips based on RISC-V extensions and verification of those.

My other favourite topic is running WebAssembly as the virtualization technology for future embedded computing platforms. Besides support for runtime systems, programming models and fleet management, we are actively looking into hardware support of bytecode virtual machines.

What should participants do in Munich while they're at the conference?

Munich is a beautiful city. To be honest I prefer the summer much more, because I am not into the cold. But if you bring an extra pair of socks, I would suggest that you enjoy a walk through the city, or even take the quick train ride into the mountains.

If you, like me, prefer to reduce the outside time during winter, you should definitely visit Deutsches Museum, which is a large museum centring around technology and engineering, entirely focused on impressive exhibits of historic and current devices and machines. They are remodelling and have recently re-opened a nice semiconductor exhibition, along with electronics, mathematics and encryption exhibitions. But there are many other things – including planes, optic experiments, large energy machines, robots and much more – than you can stare at and touch in a day. Feel free to get in touch with me during the HiPEAC conference to plan your stay!

Konversationshandbuch

For your stay in Munich, here's our handy guide of key phrases, inspired by famous German thinkers (and non-thinkers).

English	Deutsch
The prudent scientist finds the most important computing innovations at HiPEAC	*Die besten Innovationen in der Datenverarbeitung findet der gescheite Wissenschaftler auf der HiPEAC* (with apologies to Johann Wolfgang Goethe)
No optimization is too small or insignificant that one shouldn't perform it	*Keine Verbesserung ist zu klein oder geringfügig, als dass man sie nicht durchführen sollte* (Theodor Adorno)
Linux or LLVM, as long as it is an operating system	*Linux or LLVM, hauptsache Betriebssystem* (inspired by Andreas Möller)

HiPEAC Technology Transfer Awards 2023 winners announced

Since 2012, the HiPEAC Technology Transfer Awards have recognized examples of academic innovations examples of leading-edge technology being transferred from academia to industry. For the 2023 edition, five candidates were selected, as detailed below. The winners will be recognized in an awards ceremony at the HiPEAC conference, and first-time winners receive a cash prize of €1,000.

On behalf of HiPEAC, congratulations to the winners, who once again have shown the innovative potential of this community.

FPGA-Shell: Rapid Emulation of RISC-V Designs on FPGAs

Behzad Salami, Barcelona Supercomputing Center (BSC), Spain

Emulating chip designs using field-programmable gate arrays (FPGAs) is a crucial step to validate the correctness of register transfer level (RTL) design before undertaking an expensive fabrication process. With the open-source RISC-V instruction set architecture (ISA) democratizing processor design, rapid FPGA validation is gaining even greater significance. However, FPGA emulation can itself be a bottleneck, as it requires thorough understanding of the underlying hardware and different tools and skills to processor design.

Developed by BSC during the EuroHPC project MEEP, FPGA-Shell streamlines the pre-silicon validation of RISC-V custom processors on FPGAs by automating the emulation process. FPGA-Shell automatically creates and compiles the FPGA project by connecting the RTL RISC-V design to commonly used FPGA peripherals, simply by editing a configuration file. Finally, the framework builds the project and creates the FPGA bitstream automatically, all with minimal human intervention and without the need for in-depth FPGA knowledge.

FPGA-Shell has been integrated with several open-source RISC-V systems as well as custom RISC-V designs at BSC. In collaborations with academia and industry, it is constantly evolving with new features. For example, as a part of recent agreement with Lenovo, FPGA-Shell will be supported financially and technically for the emulation of multicore RISC-V designs partitioned onto multiple FPGAs. The FPGA-Shell source code is also publicly available on GitHub for further developments.

github.com/MEEPproject/fpga_shell

meep-project.eu

Enhancing Pavlovian-training experiments with GPU-accelerated machine learning

Christos Strydis, Erasmus Medical Center, Netherlands

In neuroscience, Pavlovian eyeblink conditioning is a crucial experiment for assessing human learning processes. Traditionally, researchers have tracked eyelid movements using potentiometers or electromyography. Recently, computer vision and image processing have offered alternatives, but these require human involvement and lack real-time capabilities. To address this, the neurocomputing laboratory of the neuroscience department at Erasmus Medical Center in Rotterdam joined forces with the startup Blinklab, which turns mobile phones into devices for conducting neurobehavioural evaluations. Researchers evaluated face- and landmark-detection algorithms for automated eyelid tracking, a technique which could help enable closed-loop experiments with promise for insights into neurological disorders.

After evaluating various detection algorithms, histogram of oriented gradients (HOG) and ensemble of regression trees (ERT) algorithms were chosen for eyelid detection and accelerated on graphics processing

units (GPUs) and central processing units (CPUs), yielding substantial speed improvements. The algorithm achieved an application runtime of 0.533 milliseconds per frame, surpassing real-time requirements for human eyeblink conditioning.

ORBIK Cybersecurity: A cybersecurity spinoff for equipment manufacturers

Salvador Trujillo, ORBIK Cybersecurity, Spain

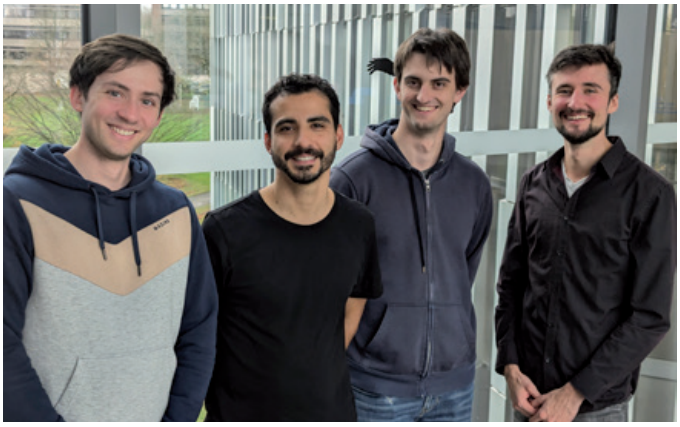
As embedded systems become increasingly interconnected, they are also increasingly susceptible to vulnerabilities. In addition, they are increasingly subject to regulatory requirements, such as the European Union's Cyber Resilience Act, as well as needing to meet customer requirements. Over the last few years, IKERLAN has developed a range of technological assets aimed at enhancing the cybersecurity of equipment manufacturers, with a particular emphasis on those incorporating industrial control systems with dedicated embedded electronics.

Drawing on IKERLAN's expertise and resources, the ORBIK Cybersecurity spinoff was created to provide cybersecurity-assessment services and help clients meet industry standards, such as IEC62443. The fledgling company is already working with clients including in the energy sector, such as equipment providers for power grids.
Find out more about ORBIK Cybersecurity in this video:
vimeo.com/883948062/778891f902

Empowering next-generation processor design: EDA tools for faster time-to-market

Lennart M. Reimann, Niko Zurstraßen, Jose Cubero, Lorenzo Pfeifer, Jan Moritz Joseph – all RWTH Aachen, Germany

In recent years, processor complexity has been growing exponentially, and design tools have been unable to keep pace. Current electronic design automation (EDA) tools are often unable to fully leverage higher core counts in contemporary processors, thus achieving only marginal performance gains, while having to cope with larger, increasingly complicated designs.



Left to right: Niko Zurstraßen, Jose Cubero, Lorenzo Pfeifer and Lennart Reimann

In this technology transfer, the team at RWTH Aachen developed two design tools and transferred these to a major international information and communication (ICT) company as part of joint research projects. The first, a tool for rapid prototyping using abstract modelling of the target hardware, allows users to understand application requirements and derive an initial architecture. The second, a parallel simulation kernel for the open-source project gem5, enables microarchitecture optimization and provides an understanding of the system-level performance impact of low-level design decisions. These tools are being deployed at scale for the company's next generation of processors.

BiodAIverse: Smart conservation technologies

Jorge Fernandez-Berni, Institute of Microelectronics of Seville (CSIC-Universidad de Sevilla), Spain

Quite apart from the incalculable loss to the planet, the sixth mass extinction which is currently ongoing also constitutes a critical threat to the future of human civilization due to the associated degradation of ecosystem services. Technology can play a role in helping to bend the curve of mass extinctions by keeping a record of the state of species and ecosystems, identifying causes of extinction and degradation, assessing the effectiveness of mitigation measures, and monitoring the evolution of the environment while collecting data to drive future actions and make informed decisions.

The spinoff BiodAIverse came about from many years of interdisciplinary research on conservation technologies jointly conducted by two groups at the Institute of Microelectronics of Seville and Doñana Biological Station in Andalusia, Spain. Leveraging researchers' extensive know-how in embedded artificial intelligence (AI) and internet-of-things (IoT) systems, as well as in the application of technology to conservation, BiodAIverse provides a scalable, end-to-end technological infrastructure for effective wildlife monitoring, ranging from real-time reporting to long-term habitat assessment.

The key technologies at the core of BiodAIverse's value proposition are:

1. a modular, multi-sensor, low-cost, low-power, IoT platform that can be easily adapted to different customer requirements;
2. AI algorithms tailored for efficient execution on embedded systems; and
3. a cloud platform that collects data coming from systems deployed in the field and aggregates them for proper visualization and analysis through customized panels and notifications.

Further information on BiodAIverse may be found in the 'SME snapshot' article in HiPEACinfo 69 (July 2023).

Have you successfully transformed a research project into industry-ready results? Find out more about the HiPEAC Technology Transfer Awards on the HiPEAC website hipeac.net/awards/#/tech-transfer

SiPearl to equip JUPITER, Europe's first exascale machine

In October, the European semiconductor company SiPearl announced a contract to equip JUPITER, the first European exascale supercomputer. JUPITER's general-purpose cluster module will be based on SiPearl's first-generation microprocessor, Rhea1.

According to SiPearl, this first contract is a major milestone for the company in fulfilling its mission assigned by the European Union through the European Processor Initiative (EPI) consortium: to ensure European sovereignty with the return of high-performance, low-power microprocessor technologies in Europe.

Owned by the EuroHPC Joint Undertaking, JUPITER will be installed at the Forschungszentrum Jülich campus in North Rhine-Westphalia and built by a consortium composed of Eviden (the Atos Group business leading in advanced computing) and ParTec (the German modular supercomputing company). The expected budget is € 273m.

bit.ly/SiPearl_JUPITER_2023

"A major milestone in SiPearl's mission to ensure European sovereignty"

BSC's AccelCom group announces ODOS offloading support for NVIDIA DOCA

Antonio J. Peña and Sergio Iserte, BSC

The AccelCom group at Barcelona Supercomputing Center (BSC), led by HiPEAC member Antonio J. Peña, has developed a solution named ODOS (OpenMP offloading support for NVIDIA DOCA) that provides support for standard OpenMP offloading and semantics for data processing units (DPUs).

A programmable system-on-chip, DPUs are used to accelerate the most demanding workloads. They combine industry-standard, software-programmable processing elements, such as central processing units (CPUs) and graphics processing units (GPU), with networking and other SoC components.

ODOS extends the LLVM compiler infrastructure to enable OpenMP offloading for NVIDIA BlueField DPUs. It allows users to offload computations to GPUs and / or DPUs using the widely adopted OpenMP syntax. This removes barriers for new users, allowing them to leverage the power of DPUs within their existing programming frameworks.

NVIDIA provides assistance and resources for ODOS, funding its development and facilitating its promotion among the user community.

Antonio J. Peña said: 'I expect to soon see many more applications benefiting from the performance advantage of NVIDIA BlueField DPUs thanks to the seamless interface provided by ODOS. We are already collaborating with NVIDIA on new functionalities and enhanced performance.'

'By integrating DPUs seamlessly into the OpenMP ecosystem, we are empowering programmers with greater access to accelerated computing, unlocking new possibilities for high-performance applications,' said Gilad Shainer, senior vice president of networking at NVIDIA.

DISCOVER-US initiates new era in EU-US distributed computing and swarm intelligence research

Launched on 1 January 2024, the DISCOVER-US project is a 30-month project that is set to energize pre-competitive collaborative research between the European Union (EU) and United States (US) National Science Foundation around the computing continuum, distributed computing, and swarm intelligence.

A Horizon Europe coordination and support action, the project will mutually reinforce the research capacity of the EU and US in this area, ultimately contributing to a stronger collaborative transatlantic research infrastructure and strengthening Europe's position in cloud-to-edge computing, the internet of things (IoT) and the tactile internet. This will be achieved by integrating relevant elements of computing, connectivity, IoT, artificial intelligence (AI) and cybersecurity.

'The name DISCOVER-US has two meanings: researchers from the US will discover "us" in Europe, while EU researchers will discover research in the US,' says project coordinator Koen De Bosschere (Ghent University). 'The project will enhance synergies, experience and knowledge sharing, building an EU-US collaborative research ecosystem for pre-competitive research in topics related to the computing continuum, distributed computing and swarm intelligence.'

The project offers an excellent opportunity for HiPEAC members working on research for the computing continuum, distributed computing, and / or swarm intelligence to build partnerships with their counterparts in the US and work on mutual research projects.

DISCOVER-US especially seeks research contributions that:

- Allow the level of abstraction to be raised in the modelling, development, execution and orchestration of complex applications deployed on the computing continuum, especially near and towards the edge.
- Generate new concepts for distributed computing, the computing continuum, swarm intelligence and edge intelligence.
- Allow trustworthy AI-enabled self-organized, dynamic and adaptive management of the resources required for execution near and towards the edge of the compute continuum.
- Deliver collaborative programming frameworks and software development tools.
- Help improve human understanding and control of those complex applications.

Around 20 pre-competitive EU-US research collaborations will be funded, and a Dagstuhl-style workshop will be held to work on key areas. The project aims to result in a transatlantic research community of around 100 senior researchers, and to develop a research vision on the main topics.

The steering board comprises the following HiPEAC members:

- Coordinator **Koen De Bosschere**, leader of the Computer Systems Lab at Ghent University and coordinator of HiPEAC;
- **Rosa M. Badia**, manager of the Workflows and Distributed Computing Group at Barcelona Supercomputing Center (BSC);
- **Marc Duranton**, member of the Research and Technology Department of CEA (French Atomic Energy Commission) and editor in chief of the HiPEAC Vision;
- **Tullio Vardanega**, associate professor in the Department of Mathematics, University of Padova; and
- **Ovidiu Vermesan**, chief scientist at SINTEF.

FURTHER INFORMATION:

discover-us.eu

DISCOVER-US has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement number 101135064. 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.

Funded by the European Union

New era for Czech science as EOSC initiative launches in the Czech Republic

Marie Hostalkova, Masaryk University

The European Open Science Cloud (EOSC) initiative aims to connect scientific communities, establish a robust data-storage and sharing system, and enhance access to research data for scientists and organizations. The Ministry of Education, Youth and Sports (MEYS) of the Czech Republic officially declared support for the initiative on 4 September 2023, signalling a pivotal moment in EOSC's implementation at the national level.

Managed by Masaryk University in collaboration with CESNET and IT4Innovations at VSB – Technical University of Ostrava, the strategic project EOSC-CZ operates under the Jan Amos Comenius Programme, financially supported by the MEYS to advance the scientific environment in the Czech Republic. Alongside the Czech Academic and Research Discovery Service (IPs CARDS) project of the Czech National Technical Library of Technology, EOSC-CZ plays a crucial role in establishing a unified environment for managing and searching for scientific information resources.

With an allocated budget of 450 million CZK over six years, the MEYS of the Czech Republic emphasizes the transformative nature of EOSC-CZ, going beyond infrastructure investments to fundamentally change the system of scientific work.

The recent legal approval marks a critical milestone, allowing the EOSC-CZ project to fully engage in its activities and offer services to the Czech scientific community. This transformative initiative is ready to inaugurate a new era of collaborative and easily accessible scientific research in the Czech Republic.

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

Czech Republic one step closer to quantum computer

Branislav Jansik, IT4Innovations

The international LUMI-Q consortium has signed the hosting agreement in Luxembourg for the acquisition and operation of a quantum computer. It will be installed at IT4Innovations National Supercomputing Center in Ostrava, Czechia, in 2024 and become the first Czech quantum computer, which will also be available to the European research community.

The LUMI-Q consortium, which brings together nine European countries: Belgium, Czechia, Denmark, Finland, Germany, the Netherlands, Norway, Poland, and Sweden, aims to provide academic and industrial users with a quantum computer based on superconducting qubits with a star-shaped topology. Its advantage is that it minimizes the number of so-called swap operations and enables the execution of very complex quantum algorithms. The assumption is that it will contain at least 12 qubits. This quantum computer will be directly connected to the EuroHPC supercomputer KAROLINA, located at IT4Innovations in Ostrava. In addition, the plan is to connect it to other EuroHPC supercomputers, especially those hosted by other members of the LUMI-Q consortium, such as the most powerful European supercomputer LUMI, in Finland, or the Helios supercomputer, which will be located in Krakow, Poland.

'Signing the agreement to host the LUMI-Q quantum computer in the Czech Republic is an important milestone not only for the Czech research community in the field of quantum computers and algorithms but also represents a significant step towards developing European quantum computing resources. Together with other European partners, we are creating an important element of future scientific progress in quantum computing and its applications,' said Vit Vondrak, Managing Director of IT4Innovations.

'We expect that not only the Czech scientific community will gain access to our quantum computer through e-INFRA CZ, but also all consortium members. As 50% of the cost of the LUMI-Q quantum computer is covered by EuroHPC JU, users from all over Europe will also have access to it. Finally, our goal is to make quantum computing available to industrial companies,' says Branislav Jansik, Supercomputing Services Director at IT4Innovations and the LUMI-Q consortium Coordinator.

Italian project RETICULATE tackles real-time, secure AI for CPS

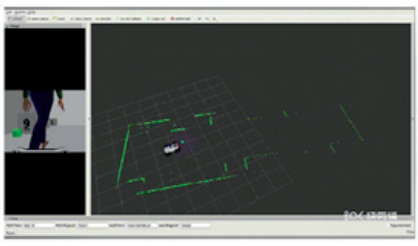
Alessandro Biondi (Scuola Superiore Sant'Anna Pisa) and Alessandro Cilardo (University of Naples Federico II)

Next-generation cyber-physical systems (CPS), are increasingly reliant on artificial intelligence (AI), particularly deep neural networks (DNNs). In CPS, DNNs must (i) execute under real-time constraints, (ii) ensure energy efficiency, as required by most embedded devices, and (iii) ensure confidentiality and integrity of design and user data. DNNs usually need hardware acceleration, for example system-on-chip (SoC) platforms augmented with field-programmable gate array (FPGA) fabrics.

A project supported by the Italian PRIN research funding program, RETICULATE (REal-TIme and seCUre acceLeration framework for Artificial inTElligence) will address the technical challenges for the adoption of AI in future CPS, by delivering an innovative DNN framework for FPGA-based SoCs offering advanced capabilities in terms of real-time performance and cybersecurity. RETICULATE is led by Scuola Superiore Sant'Anna Pisa (principal investigator (PI): Alessandro Biondi), with the University of Naples Federico II as a partner (local PI: Alessandro Cilardo).

The project will develop problem-specific timing analysis methods to predict the performance of DNN accelerators and enable design space exploration, taking into account time predictability. As DNN accelerators are memory intensive, the project also aims at exploring memory-driven optimizations in FPGA platforms. Furthermore, since existing acceleration platforms do not provide any built-in security features, RETICULATE will develop specific mechanisms at the architecture level to increase the security of the CPS as a compute platform. Cybersecurity threats will also be considered at the level of AI algorithms, by providing DNN accelerators with efficient mechanisms detecting adversarial attacks and unsafe inputs in general.

The framework delivered by RETICULATE will be tested and demonstrated in the context of autonomous navigation with real-time AI-based perception for a robotic land rover as well as for confidential remote inference in a smart health application.



IT4Innovations scientists working on ESA project AIOPEN

Barbora Polakova, IT4Innovations

Funded by the European Space Agency (ESA), AIOPEN aims to create a user-friendly, intuitive platform that allows users, regardless of their technical expertise, to use and apply artificial intelligence to process remote sensing and satellite imagery data. The project, which involves IT4Innovations, Space Applications Services, Telespazio, KP Labs, and SERCO, will provide visual workflows and pre-built templates that users can customize with available Earth observation data.

The AIOPEN platform will be created by combining and extending the ASB, EOPEN, and EOEPCA platforms with new, innovative services based on artificial intelligence and machine learning. Two case studies will be implemented on the platform during the project.

The Urban Change Detection case study, for which IT4Innovations is responsible, aims to use machine learning models to monitor urban growth, thereby enabling comprehensive regional development planning and preventing unwanted urban sprawl. Earth observation data and deep neural networks will be used to detect urban change, and a digital twin of Earth's (urban) changes will also be created. Artificial intelligence will allow satellite imagery to be analysed in depth.

Another long-standing problem in many countries is deforestation, which the AIOPEN team will also address under the leadership of KP Labs. Monitoring, segmentation, and analysis of forests will also be carried out using artificial intelligence, algorithms, and publicly available data.

Vaclav Svaton, project investigator for IT4Innovations, commented: 'The AIOPEN project aims to extend existing commercial platforms with AI functionalities. The project's results, i.e. the provision of services enriched with innovative AI functionalities, will be published in the project's network of resources.'

eo4society.esa.int/projects/aiopen

Building the biodiversity component of the Digital Twin of the Ocean

Fabiana De Carlo, Trust-IT Services

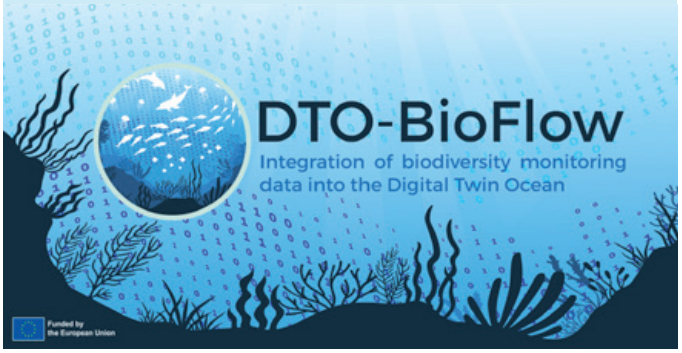
Marine habitats present specific issues when it comes to observing, mapping, and monitoring biodiversity. While significant advancement has been made in Europe to collect, harmonize, and make available data on marine biodiversity, a large portion of the data collected is currently inaccessible; this is referred to as 'sleeping data'.

The DTO-BioFlow project's primary objective is to awaken sleeping biodiversity data, enabling smooth integration of existing and new data into the European Digital Twin of the Ocean (EU DTO).

Over the next four years, the DTO-Bioflow consortium will work on consolidating standards, quality control, communication protocols, harmonization pipelines, data products, data models, ingestion procedures and incentives for sustainable connection to improve the interoperability and digitization of biodiversity data. The project will also test out various technologies to carry out species monitoring on a massive scale. The end-to-end approach will be demonstrated via science-based use cases and via mechanisms to monitor, measure progress and drive community action towards increasing biodiversity data flows.

DTO-BioFlow kicked off on 27 September in Ostend, Belgium. The consortium, led by the Flanders Marine Institute, comprises 30 partners from 14 countries, including research centres, infrastructure hosts, networks, associations, global aggregators, platforms, and others.

FURTHER INFORMATION:
dto-bioflow.eu



Funded by the European Union

EXA4MIND extreme data platform presented at EBDVF 2023

Jan Martinovic, IT4Innovations

Organized by the Big Data Value Association, the European Big Data Value Forum (EBDVF) took place in Valencia on 25-27 October. The EXA4MIND project presented its activity in developing an extreme data platform for advanced data analysis and knowledge extraction integrated with European supercomputing and data spaces at this event.

EBDVF serves as a platform for the exchange of knowledge and the exploration of novel perspectives on big data and AI to industry experts, developers, researchers, and policymakers. The three-day event featured 250 speakers who presented topics related to big data and artificial intelligence over a series of plenary and special sessions, workshops and panel discussions.

The EXA4MIND was represented by its team and management, showing its commitment to the European big data community. The project co-sponsored EBDVF 2023 and had a booth in the exhibition.

'The forum gave me very valuable insights into the landscape of European data spaces and big data. Besides presenting the project, we got key tips on, for example, the Dataspace Protocol from the Data Spaces Support Centre and the International Data Spaces Association, enabling us to make connections,' said Stephan Hachinger (Leibniz Supercomputing Centre), EXA4MIND science and co-design coordinator.

EBDVF 2023 attracted 650 attendees; its next edition will take place in Budapest.

FURTHER INFORMATION:
exa4mind.eu
european-big-data-value-forum.eu

Industry 4.0 conference returns to Bosnia and Herzegovina

Isak Karabegović, Academy of Sciences and Arts of Bosnia and Herzegovina

On 5-6 October 2023, the Academy of Sciences and Arts of Bosnia and Herzegovina (ANU BiH) organized a conference named 'Basic technologies and models for the implementation of Industry 4.0' in Sarajevo. The conference was organized in collaboration with the Society for Robotics of Bosnia and Herzegovina, the Foreign Trade Chamber of Bosnia and Herzegovina, and the Chamber of the Economy of the Federation of Bosnia and Herzegovina.

The aim of the conference was to share knowledge about developing and implementing technologies in industry 4.0, and to adapt to the challenges while taking advantage of the opportunities this new paradigm offers.

Industry 4.0 brings a change of business paradigms and production models that will be reflected at all levels of production processes and supply chains. These changes are caused by technological advances such as robotics and automation, the internet of things (IoT), big data, 3D printing, smart sensors, radio frequency identification (RFID), virtual and augmented reality, artificial intelligence, advanced security systems, etc.

Benefits of the industry 4.0 approach include improved product quality, reduced working hours, enhanced safety, minimized costs, and ongoing maintenance. However, there are many challenges to the implementation of industry 4.0, such as changing business bias, resource planning, legal issues, security issues, and standardization.

While many small / medium enterprises (SMEs) in the Western Balkans intend to gradually introduce smart solutions, methods and technologies, the costs of implementation and high complexity of the technologies present major challenges. It is therefore necessary to create a framework that describes the key issues and emphasizes possible answers, while also providing a platform for private-public cooperation and partnership on emerging issues.

In this context, and building on previous events such as the industry 4.0 workshop organized in October 2022, as reported in *HiPEACinfo* 68, this conference brought together experts from all over the country. Topics included edge AI, the role of software engineering in industry 4.0, engineering skills for intelligent manufacturing, advanced robotics, and digital twins.

Conference proceedings are available for download from the ANU BiH digital library
https://bit.ly/BiH_Industry-4-0_Oct23_proceedings

Next-generation solutions for sexual health with bitsxlamarató

Organized by the Facultat d'Informàtica de Barcelona (FIB) at the Universitat Politècnica de Catalunya-Barcelona Tech (UPC), Hackers@UPC, Barcelona Supercomputing Center and the Institut de Recera Sant Pau, bitsxlamarató is an annual hackathon which contributes to the 'Marató' fundraising initiative of Catalan broadcaster TV3. Bringing together programmers, academics and health professionals, the hackathon seeks to find technology solutions for healthcare.

The 2023 edition, which took place in Barcelona on 16-18 December, focused on sexual and reproductive health. In total, 150 people took part, including 125 hackathon participants, mentors and organizers. There were a total of 35 projects split among four different challenges, covering different aspects of sexual and reproductive health. For the participating health organizations, the hackathon provided solutions which will help them continue their vital work.

As in previous years, HiPEAC supported the hackathon, providing welcome packs and setting up a HiPEAC Jobs wall.

fib.upc.edu/ca/la-marato

Antonio J. Peña receives Agustín de Betancourt award

On 21 November 2023, HiPEAC member Antonio J. Peña (Barcelona Supercomputing Center) was awarded the Agustín de Betancourt y Molina Prize for young researchers by the Spanish Real Academia de Ingeniería. The award recognizes professionals who have made original and relevant contributions in engineering, specifically valuing aspects related to technology transfer.

The jury noted the impact of technology transfer of Antonio's research, with contributions to the MPICH implementation of the message passing standard 'MPI', whose derivatives are used in the largest supercomputers in the world, thus contributing to much of the progress in the fields of computational science. They also valued his collaborations with Intel and NVIDIA in software research and development (R+D) for the democratization of the use of heterogeneous resources in supercomputing, such as different accelerators and memory subsystems.

On behalf of HiPEAC, congratulations to Antonio!

ACACES 2023 and HiPEAC 2024 on HiPEAC TV

Videos of lectures from ACACES 2023 are now available on the HiPEAC YouTube channel, HiPEAC TV. You'll also find interviews, keynote talks, social media shorts, animations, and more.

More HiPEAC 2024 videos will be available in due course – stay tuned for further information.

hipeac.net/tv

Aida Todri-Sanial awarded AiNed fellowship grant

HiPEAC member Aida Todri-Sanial (Eindhoven University of Technology – TU/e) has been awarded an AiNed Fellowship Grant for her research project 'AI-on-ONN: Online Learning for Sense-to-Compute Edge AI with Oscillatory Neural Networks'. This project will investigate energy-efficient computing and processing of signals from the ever-growing number of sensors in smart systems, estimated to reach 45 trillion by 2030.

Aida's research focuses on a novel computing paradigm whereby programs (functions) are directly mapped onto the computational hardware, based on the principle 'let physics do the computing'. The problem she will address in this fellowship is to investigate whether oscillatory neural networks (ONN) can allow a seamless sense-to-compute paradigm.

AiNed Fellowship Grants help Dutch academic knowledge institutions to attract talented researchers in artificial intelligence (AI). Awarded by the Dutch National Growth Fund programme, they aim to consolidate the AI knowledge and education base in the Netherlands and strengthen the national AI ecosystem.

'Over the last few years, I have been working with my team and international collaborators on developing ONN computing covering aspects from computing models, algorithms and its physical implementation in hardware. This fellowship is a great boost to allow to develop this novel sense-to-compute paradigm with ONNs and assess its pros and cons in edge AI,' Aida commented.

Congratulations on behalf of HiPEAC!

bit.ly/AiNed_Aida-Todri_Sanial

ETH Zurich students scoop multiple awards at SC23

A team of students from ETH Zurich, going by the name of RACKlette, have become the first European winners of the Student Cluster Competition at the annual Supercomputing conference, SC23, which was held in Denver, Colorado, United States, in November 2023.

The Student Cluster Competition at Supercomputing was started in 2007 with the aim of introducing the next generation of students to the high-performance computing (HPC) community, bringing together undergraduate teams from around the world. Sponsored by various hardware and software vendors, each team is tasked with designing, building, and operating a small computer cluster.

For the 2023 edition, 11 teams took part, including six from the United States, three from China, one from Singapore and one from Switzerland.

The RACKlette team was supervised by HiPEAC member Torsten Hoefler (ETH Zurich / Swiss National Supercomputing Centre–CSCS), with significant contributions from Hussein Harake, a senior systems engineer at CSCS. The team was sponsored by E4 Computer Engineering, NVIDIA, SPCL at ETH Zurich, HaslerStiftung, and CSCS.

In related news, Marcin Chrapek, Mikhail Khalilov and Torsten Hoefler won the SC23 Best Student Paper and the Best Reproducibility Advancement Award for their paper 'HEAR: Homomorphically Encrypted Allreduce'.

Congratulations on behalf of HiPEAC!

Photo credit: Jo Ramsey, SC Photography

Top: Team RACKlette at SC23; Bottom: Marcin Chrapek accepting the Best Student Paper Award, with Torsten Hoefler

Dates for your diary

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

ARITH 2024: 31st IEEE International Symposium on Computer Arithmetic
10-12 June 2024, Málaga, Spain
Abstract deadline: 18 January 2024 | Paper deadline: 25 January 2024
arith2024.arithsymposium.org

DATE 2024: Design, Automation and Test in Europe
25-27 March 2024, Valencia, Spain
HiPEAC Jobs activities with the Young People Programme
date-conference.com

EuroSys 2024: European Conference on Computer Systems
20-24 April 2024, Athens, Greece
Includes MECC 2024: MetaOS for the Cloud-Edge-IoT Continuum (22 April) / Workshop by FLUIDOS, aerOS, ICOS, NebulOuS, NEMO, and NEPHELE
meccworkshop.github.io

GraphSys'24: 2nd Workshop on Serverless, Extreme-Scale, and Sustainable Graph Processing Systems
Co-located with ICPE 2024
7-8 May 2024, London, UK
Paper submission: 22 January 2024
sites.google.com/view/graphsys24

HPDC 2024: 33rd ACM International Symposium on High-Performance Parallel and Distributed Computing
3-7 June 2024, Pisa, Italy
Abstract submission: 18 January 2024
hpdc.org/2024

Euro-Par 2024: 30th International European Conference on Parallel and Distributed Computing
26-30 August 2024, Madrid, Spain
Abstract submission: 5 March 2024
2024.euro-par.org

ISC High Performance
12-16 May 2024, Hamburg, Germany
isc-hpc.com

HEART 2024: 14th International Symposium on Highly-Efficient Accelerators and Reconfigurable Technologies
19-21 June 2024, Porto, Portugal
Submission deadline: 18 March 2024
fe.up.pt/heart2024

HiPEAC 2024 keynote speaker Lieven Eeckhout (Ghent University) is an ACM and IEEE Fellow, and the recipient of multiple awards including the Maurice Wilkes Award, OOPSLA Most Influential Paper Award, and MICRO and ISPASS Best Paper Awards, as well as five European Research Council (ERC) grants. HiPEAC caught up with Lieven to learn about his background in computer architecture and his recent focus on sustainability.

'By 2030, ICT could be responsible for 7-20% of all electricity demand. It's our moral duty to act'

How did you end up specializing in your field?

I started studying engineering because I loved math. In my first year at the university, during a course on numerical techniques where we had to program optimization problems, I realized that I loved programming and I found it fascinating that I could experiment at home behind a computer. Over time, I realized that I was most interested in how computer systems work, how to design them and how to program them. During my PhD I became particularly interested in performance modelling and workload characterization. Later, I also got interested in microarchitecture and computer-system resource management. My latest focus in on sustainability.

What prompted you to start researching sustainability? Why should it be on the HiPEAC community's radar?

I've always loved nature. Watching how humanity is exhausting the planet, and witnessing young people protesting for climate change, I started wondering how big the impact of information and communication technology (ICT) on the environment was. I proposed to my department that we should start a new course on sustainable computing, and they approved. It was while I was working on the course material that I became really interested in the topic. At the time, two to three years ago, there was no course material available elsewhere, and so I had to compose my lecture material myself. I did a lot of research trying to make sense of the limited, disparate data that is available.

I believe the HiPEAC community should indeed focus on sustainability. ICT today is responsible for 2-4% of greenhouse gas emissions. This is on par with the aviation industry, and it is growing. Some projections state that by 2030, ICT will be responsible for 7-20% of all electricity demand. We should do something. I would even state that it's our moral duty to act! And I believe there is lots we can do, as hardware designers, system integrators, and software developers.

What, for you, are the main issues concerning sustainability?

It turns out that when looking at the total environmental footprint of a computing device, the embodied emissions – the emissions for manufacturing, assembly, transportation, and end-of-life processing – are dominant compared to the operational emissions during a device's lifetime. This is the case for personal mobile devices such as smartwatches, smartphones, tables, laptops as well as for servers in hyperscale data centres. Operational emissions seem to dominate for always-connected devices.

My own analyses indicate that embodied emissions will soon dominate for nearly all computing devices. The reason is twofold. On the one hand, devices are becoming increasingly energy efficient, so this reduces the operational footprint. On the other hand, the demand for computer chips continues to increase – by around 9% per year – and, in addition, semiconductor manufacturing is becoming increasingly energy demanding – an increase of around 11% per year. Hence we can expect that the embodied footprint will continue to increase.

As computer engineers and scientists, we therefore need to start designing computer systems with sustainability and especially the embodied footprint as a primary design goal. We need to start designing smaller chips to reduce the embodied carbon

Cartoon by Arnout Fierens from the HiPEAC Vision 2024

footprint. To pick one example: dark silicon – or providing tens of accelerators on chip that are powered on only when needed – is considered the way forward to continue scaling performance as progress in chip technology is slowing down, but it is harmful for sustainability. The reduction in operational footprint is unlikely to outweigh the embodied footprint that these accelerators incur. Sustainability requires us all as a community to start thinking differently about how to design computer systems. We need to design computer systems more holistically considering the overall environmental footprint.

What are the main challenges in researching and teaching this topic?

A major challenge is that there is little data available pertaining to sustainability, and when data is available, there is quite a bit of uncertainty about the data. Companies are starting to publish lifecycle assessment reports about the products they bring to market, and these reports also acknowledge that some numbers are based on industry averages or estimates.

My take on it is that we should embrace the uncertainty and go back to first principles using proxies for the embodied and operational footprint. We can use chip area as a proxy for the embodied footprint and energy / power as a proxy for the operational footprint. Reasoning about chip area and energy / power and performance in a holistic manner enables computer engineers and scientists to make reasonable design decisions despite the uncertainty. So, while we should continue to aim for high-quality data and better understand the environmental footprint of computing, there is no time to waste, and I believe we can make a difference today.

Teaching sustainability topics is challenging but very rewarding and interesting. Sustainability is a fundamentally multidimensional problem, and there are so many stakeholders. Sustainability is also a much broader problem than global warming and carbon emissions: it covers raw material extraction, end-of-life repurposing or recycling, water consumption, business models, legislation, etc. This oftentimes leads to very interesting discussions in class where we analyse the pros and cons of sustainable developments.

For example, while we as engineers and scientists can try to design computing devices with a lower carbon footprint and less material use, there is always the risk of a rebound effect, also called Jevons' paradox. It is well known that making devices more efficient oftentimes leads to increased usage and deployment. Given our linear economy, which is based on selling stuff, I believe that we need new business models towards a circular economy. Also, we need regulation and legislation to temper the demand for new devices and the increase in the environmental footprint of ICT.

What are your career highlights and future plans?

My whole career has been a highlight in my view. It was great to grow as a researcher during my PhD and postdoc period, and later as a professor. Mentoring students, seeing how they learn new stuff in class, seeing how they grow as researchers is all very rewarding. I always enjoy working on a new project with collaborators and students. In the beginning there are lots of unknowns and you don't really understand what the problem is and how to tackle it. But over time you start understanding what's going on, and the most wonderful moment is when all pieces of the puzzle come together.

Of course, it's also very rewarding to be able to publish your research in the field's top-tier conferences and then also receive international recognition through awards – I've been lucky enough to receive a few. As for the foreseeable future, I'm planning to focus my research on making our computer systems more sustainable.

HiPEAC keynote speaker Reetuparna Das is an associate professor at the University of Michigan. She has previously worked at Intel Labs and the Center for Future Architectures Research, and has co-founded a precision-medicine start-up, Sequal Inc.

'Move to data-centric architectures and leave behind compute-centric designs'

How did you get into computer architecture research?
Discovering computer science in middle school ignited my love for programming. Growing up, my curiosity about how computers work deepened, focusing on their 'brain' – the processors. Taking an undergraduate computer architecture class intensified this interest; I was particularly fascinated by concepts like speculation, prediction, and caching. This experience confirmed my decision to pursue a PhD in computer architecture, and I have not looked back since.

What are the most pressing research topics in the field?
The global datasphere is experiencing unprecedented growth, with the data generated in the next five years expected to surpass the cumulative amount created since the inception of digital storage. Propelled by advanced technologies like artificial intelligence (AI), the internet of things (IoT), 5G, and industry 4.0, this immense influx of data, often left unexamined, poses both challenges and opportunities.

If we could rise to the challenge of processing at least the bulk of the data flood, we can unleash many momentous societal benefits. I would encourage the next generation of computer architects to move to data-centric architectures and leave behind compute-centric designs. There are several compelling paradigms, including in-memory computing and domain-specific customization, that can be leveraged to accomplish this.

What are some of the main challenges for in-memory computing?
The key problem in realizing 'practical' in-memory computing is building a software ecosystem. This is not a unique problem: graphics processing units (GPUs) faced and overcame it by building CUDA programming frameworks and a customized system-software stack. For in-memory computing there are several programming models, and our community needs to converge.

Another challenge is data staging, i.e. the problem of placing and aligning data. Von Neumann architectures simply solve this problem by moving data from any memory location to registers via loads, and compute units work out of registers. In-memory computing is unique since the memory units are themselves compute units, thus operand data must be placed in the correct memory locations and moved around minimally between computing.

On the technology side, exciting opportunities exist in architecting emerging memory devices such as ferroelectric memories, magnetic memories, and new genres of resistive memories.

How is your research applied?
Our prior work repurposes thousands of existing cache memory arrays into massive vector compute units, providing parallelism several orders of magnitude higher than a contemporary GPU. Additionally, it saves energy spent shuffling data between storage and compute units – a significant concern in big-data applications. Caches that compute can be a game changer for AI: they can add accelerator capabilities to general-purpose processors, avoiding the significant die-area cost of a dedicated accelerator. For example, we showed that compute-enabled caches in Intel Xeon can improve processor efficiency by 629x for convolutional neural networks (CNNs).

Our research also extends into precision health and genome sequencing, with certain projects requiring a meticulous co-design approach, incorporating wet-lab procedures, computational biology algorithms, and hardware design. The significance of genetic molecular markers cannot be overstated, particularly for surgical decision-making, cancer diagnosis, and clinical trial enrolment. One notable accomplishment of our research is the demonstration of ultrarapid sequencing of brain-tumour tissue in under 40 minutes; typically, this takes several days to weeks for a lab sendout. Our breakthrough has promising implications for advancing surgical procedures for tumour removal and diagnosis.

HiPEAC 2024 keynote speaker Mitsuhisa Sato has been a deputy director of RIKEN Center for Computational Science since 2018, and is the research team leader of the programming environment research team at the RIKEN Center for Computational Science (R-CCS). He was involved in the project to develop the flagship Japanese supercomputer, Fugaku, which held the number one spot on the Top500 list of the world's most powerful computers from June 2020 until November 2021, and which was number 4 at the time of writing. HiPEAC asked Professor Sato about his work at RIKEN, choosing a supercomputer architecture, and why we need more computing power.

'The effective use of silicon is becoming more important'

Could you tell us a bit about your career so far?

I worked for Real World Computing Partnership, Japan, from 1996 to 2001 as the head of the parallel and distributed system performance laboratory. In this project, I was leading computer-cluster technology and the Omni OpenMP compiler project. I worked as director of the Center for Computational Sciences at the University of Tsukuba from 2007 to 2013, leading the development of supercomputers using computer-cluster technology.

In 2010, I joined RIKEN as the research team leader of the programming environment research team at the RIKEN Center for Computational Science (R-CCS). From 2014 to 2020, I was working as a team leader of architecture development team in the FLAGSHIP 2020 project to develop the flagship Japanese supercomputer, Fugaku. Since 2023, I have been working as division director of the Quantum HPC Hybrid Computing Platform Division at R-CCS.

Why was an Arm-based architecture chosen for Fugaku?

Our vender partner was Fujitsu. They used the SPARC instruction set architecture (ISA) for the K supercomputer, the predecessor to Fugaku, but it is now obsolete. At the beginning of the development of Fugaku, the choice of ISA was Intel or Arm. Although Arm was mainly used for embedded processors at that time, we made the decision to use Arm since the company was interested in the market of high-performance computing (HPC) and high-end servers. Fujitsu contributed to the design of the SVE (Scalable Vector Extension) instruction set for HPC.

Supercomputers are already so powerful – why do we constantly need more computing power? Which research directions (e.g. quantum) do you think will get us there in the face of a slowing Moore's Law?

New application areas such as artificial intelligence (AI) are emerging. New information technology (IT) infrastructure – such as cyber-physical systems (CPS) and digital twins – demands more computing power. Although the progress of silicon technology is slowing, the effective use of silicon is becoming more important thanks to novel architectures such as application-specific accelerators.

Quantum computing cannot be a general solution because its applications areas are limited. But it will be revolutionary, and it will lead to ground-breaking breakthroughs in some applications.

What about making supercomputers more efficient?

Looking at several supercomputer centres, the power capacity limit seems about 50MW from an economic standpoint. The capacity will gradually increase, but the computational demands will increase more rapidly, so it is important to improve efficiency.

Fugaku broke new ground in HPC thanks to its Arm-based architecture

Combining technologies from the web, cyber-physical systems (CPS), the cloud, the internet of things (IoT), digital twins, and AI into a coherent, federated ecosystem, the "next computing paradigm" is the main focus of the HiPEAC Vision 2024. In this article the HiPEAC Vision editorial board picks out the highlights.

Making the "next computing paradigm" a reality: The HiPEAC Vision 2024

For some years now, it has been obvious that technology is evolving faster than humans can adapt – witness, for example, the dizzying progress of artificial intelligence (AI) – while, at the same time, geopolitical and environmental pressures, such as global warming, are intensifying. Given this context, it is important that the European ecosystem of computing systems reacts quickly and efficiently to improve its place within the competition and proposes solutions that conform to 'European' ethics.

This HiPEAC Vision 2024 sets out our vision of how Europe could do this: through the **'next computing paradigm' (NCP)**, a convergence of key technologies from the web, cyber-physical systems (CPS), the cloud, the internet of things (IoT), digital twins, and AI into a coherent, federated ecosystem. This ecosystem emphasizes spatial computing, (generative) AI at the edge, dynamic web integration, user-centric models and intelligent, efficient and trustable orchestration of distributed services. Europe should work on building a coherent NCP and promote it as a worldwide approach.

Although based on existing technologies, the NCP will be highly disruptive. As the main highlight of the HiPEAC Vision 2024, it will rely on a synergy of the results of recommendations in the other topics outlined in the document.

Recommendations of the HiPEAC Vision 2024

As the HiPEAC Vision is now updated every year, many of the recommendations of the previous HiPEAC Vision necessarily still hold, and the same principle of 'leadership races' – in artificial intelligence, hardware, cybersecurity, and sustainability – is still used to structure them. However, this edition focuses on how these elements will contribute to and shape the NCP.

Promote the next computing paradigm and develop the technologies that will make it happen

- **Develop 4D-aware implementation technologies**, standardizing representation and protocols for encoding physical objects and spaces, supporting mobile computation, and powering 4D-enabled operations.
- **Augment APIs for interoperability**, enhancing APIs with specifications for non-functional properties and dynamic service composition.
- **Enable the mobility of computation**, relocating data and processes as needed.
- **Adopt (generative) AI at the edge** for greater efficiency and privacy, and reduced latency.
- **Develop AI-powered edge orchestrators** that can dynamically combine services based on user needs.
- **Encourage non-proprietary integration** by supporting open standards and platforms.
- **Initiate proof-of-concept efforts**, creating demonstrators to showcase the advantages of NCP technologies.

Make the EU a strong player in artificial intelligence, particularly for widespread use at the edge

- **Support EU growth in AI domains** by investing in AI research and infrastructure.
- **Develop foundation models based on 'European values'** and reflecting regional needs.
- **Promote open-source AI models** to enable access to shared AI resources.
- **Develop local AI solutions** and **specialized accelerators** for edge devices.
- **Use AI for software and hardware development**, upskilling engineers and researchers.

- **Develop policies around AI accessibility and societal impact**, ensuring equitable benefits from AI.
- **Ensure 'correctness by construction'**, automating AI output verification for trustworthiness.

Develop innovative and efficient new hardware solutions, from architecture to technology

- **Continue to improve performance and energy efficiency**, including by exploring and integrating new technologies.
- **Promote interdisciplinary research**, driving new hardware paradigms.
- **Explore innovative architectures** for data-intensive computing.
- **Develop a full European ecosystem**, promoting fast prototyping and specialized architectures and developing the chiplet / interposer European ecosystem.
- **Ensure sustainable hardware development**, focusing on energy and emission reduction, and recyclability.

Make cybersecurity a major upfront concern in every computing system

- **Address vulnerabilities**, building systems with early consideration of cybersecurity and privacy.
- **Reduce dependence on external ICT**, using trusted European companies or open-source solutions instead.
- **Address security challenges in large language models (LLMs).**

Make sustainability lifecycle assessment a requirement for all new computing systems

- **Create validated lifecycle models** that model environmental impact.
- **Develop sustainability-focused design**, accounting for environmental costs in product design.
- **Create viable sustainable business models** for the ICT industry.
- **Create ICT solutions for green applications** in other industrial sectors.

In all domains, foster global thinking and promote cross-domain / cross-topic collaborations

- **Promote collaboration:** encourage teamwork across European research and technology groups.
- **Cross-domain project calls:** foster interdisciplinary research for innovative solutions.
- **Competence centres:** create centralized European expertise hubs.
- **Build open-source ecosystems** to **accelerate innovation** and **accessibility**.
- **Multi-dimensional tooling:** develop tools supporting capacity to address complex and critical challenges.
- **Adopt a holistic approach to efficiency**, emphasizing global co-design and system thinking.

It has become increasingly obvious that digital technology in general, and the web in particular, constitute critical infrastructure, and that control over these technologies has major implications for geopolitics. Given the cross-cutting importance of technological autonomy for the EU, the topic of sovereignty is now more focused and is included in each key area.

FURTHER READING

For this edition of the HiPEAC Vision, the main recommendations are in one document, while a second document, titled 'Rationale', brings together the supporting material. In addition to the main recommendations, we encourage readers to look at the more detailed 'Rationale' document, which is structured according to the main themes of the HiPEAC Vision 2024, helping readers understand the current situation, the forecast of things to come, and the reasons why this set of recommendations was formulated.

If you would like a print copy of the 'Rationale' document, please contact us: info@hipeac.net

vision.hipeac.net

The NimbleAI project, funded by the European Union's Horizon Europe programme, is developing a sensing-processing neuromorphic 3D chip that is inspired by the detection of light in eyes and the processing of visual information in brains. Looking to biological systems for inspiration, the project aims to deliver significant efficiency gains compared to mainstream processors. HiPEAC caught up with NimbleAI coordinator Xabier Iturbe (IKERLAN) to find out more.

Eye spy

NimbleAI's quest for a bio-inspired 3D chip for computer vision

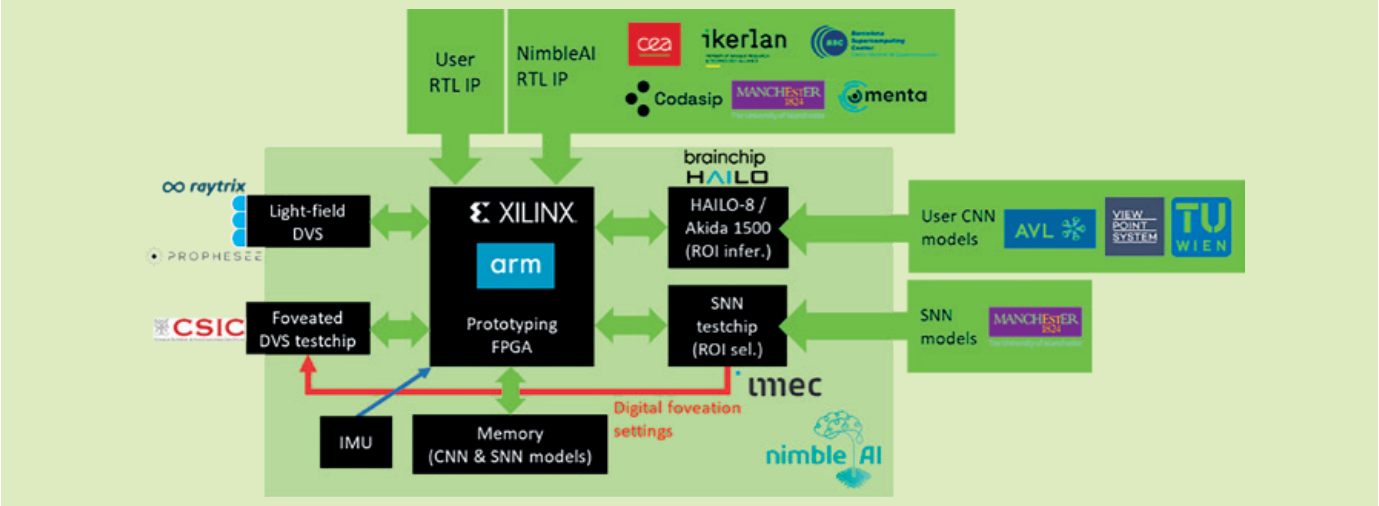
Which biological systems provide inspiration for this project?

We are taking inspiration from both insect and vertebrate visual systems. We are designing a novel foveation mechanism – that is, a mechanism to boost resolution in selected regions or objects – for dynamic vision sensors (DVS). This will allow to dynamically adjust DVS resolution based on the value of visual information each sensor region brings to the application: a visual scene is sensed in low resolution to detect regions of interest to be foveated, that is, sensed in high resolution for better accuracy. The foveation is driven by selective-attention algorithms inspired by central and peripheral vision in vertebrates, which together provide an extremely efficient visual information-gathering mechanism.

We have also built the first-ever light-field DVS that captures directional information of incoming light to enable event-driven, ultra-low-latency and energy-efficient depth perception. This is inspired by the structure of compound insect eyes, which allow for effective depth and motion detection using very simple neural networks in the insects' tiny brains. To some extent, NimbleAI's neuromorphic vision is inspired by the sort of foveated compound eyes in dragonflies and mantises, which allow them to combine high-resolution vision of their prey with 3D perception to coordinate precise, swift aerial attacks. The NeuroEdge workshop, co-located with HiPEAC 2024, will elaborate on these concepts.

How do you plan to translate this biological inspiration into a chip?

We are adopting a time- and cost-effective approach that relies on reusing and extending the capabilities of existing technology (for example, using Prophesee DVS to capture light-fields). We are producing standalone testchips to implement new capabilities that are not commercially available, especially those related to the novel sensing mechanisms envisioned in the project, such as foveation. In parallel, we will be demonstrating how the different NimbleAI sensing and processing components work together at



The NimbleAI PCB prototype to be delivered in Q4 2024 (some components will be available later)

the logical and physical levels. We are advancing towards an actual 3D physical implementation of key components related to the novel NimbleAI vision modalities: the foveated DVS, light-field DVS and spiking neural network (SNN) engine that runs selective-attention algorithms. The latter components are to be partitioned and laid across at least three layers in the NimbleAI 3D silicon stack, allowing for the combination of different process technologies with specific support for features like low-noise image pixels and advanced high-performance nodes for neural-network processing. Although the project budget doesn't stretch to manufacturing and assembling such a complicated 3D chip, we will carry out a physical implementation including physical verification and sign-off to ensure technology feasibility.

In parallel, we are building a miniaturized printed circuit board (PCB) prototype that integrates NimbleAI testchips, commercial AI chips and a field-programmable gate array (FPGA), to emulate the functionality of the intended 3D chip. The expectation is that this prototype will help attract early adopters of the NimbleAI technology by allowing user applications to be tested on the emulated NimbleAI chip using live input data captured by NimbleAI DVS sensors and running users' convolutional neural networks (CNNs) on commercial AI chips, or on NimbleAI components prototyped on the FPGA. Both technology feasibility and adoption interest are key to continue developing the NimbleAI concept into an integrated 3D chip, and we expect to fulfil these two requirements within the project timeframe.

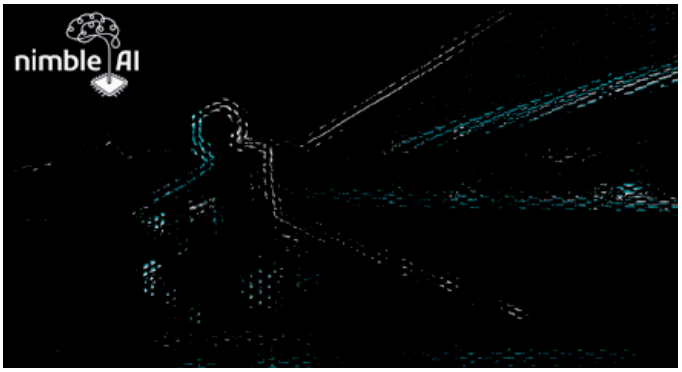
How is NimbleAI progressing in its objectives?

The design of the foveated DVS sensor is almost complete and will soon be sent to the foundry for manufacturing. The light-field DVS prototype has already been assembled and the first real-world datasets collected in an automotive setting. This allows us to validate the algorithms that we have been designing during the first year of the project using synthetic datasets. Likewise, CEA's computational memory has been optimized to run CNN inference on highly sparse DVS data.

What results can we expect over the next few months?

Once the light-field DVS algorithm is validated using real-world datasets, we will move from simulation to register transfer level (RTL) design and FPGA prototyping. This will be key to ensuring that the algorithm is not only useful in terms of efficiency and accuracy, but also delivers results with ultra-low latency. In fact, we expect ultra-low latency to be a key advantage of NimbleAI technology as, to the best of our knowledge, there is no passive 3D perception solution in the sub-millisecond range.

By the end of 2024 we will have a clearer view of the performance of our 3D perception algorithm using live real-world data. By the second quarter of 2024, Raytrix will have completed the adaptation of its software development kit (SDK) to support light-field DVS data, allowing potential adopters of this technology to experiment with neuromorphic 3D perception. In early 2024, we will also launch the 3D silicon integration activities using the electronic design automation (EDA) tool and sensing and processing components designed in the first half of the project. Finally, the first functional prototype of the NimbleAI 3D chip, including platform software and hardware, will be available by the fourth quarter of 2024. We would encourage anyone wishing to test their vision pipelines on this prototype to contact us, so that they can harness the biological advantage of NimbleAI technology.



A motorcyclist captured in the first-ever light-field DVS dataset

What kind of applications would you see this technology being used for?

Any application that requires energy-efficient and low-latency vision, especially those in which 3D perception is also important. Partner use cases in NimbleAI are good examples:

- automated and autonomous driving by AVL (where the focus is on ultra-low latency to react quickly to unexpected obstacles)
- eye tracking by Viewpointsystem (which focuses on energy efficiency for use in lightweight glasses)
- portable medical devices by ULMA Medical (where again the focus is on energy efficiency)

Beyond these, we target drone navigation and robotics, as well as space applications. In fact, we are discussing collaborations with the European Space Agency (ESA) and other key stakeholders in the space industry to explore the benefits of NimbleAI technology for space rendezvous and landing manoeuvres. Finally, we are also exploring the use of NimbleAI technology as an enabler for novel AI algorithms which have not yet made their way to industry, including neural circuit policy networks by TU Wien.

The **NeuroEdge** workshop at **HiPEAC 2024** will expand on topics discussed in this article, in particular the talk titled 'I spy with my little insect eyes: Combining DVS and light-fields' neuroedge.eu

NimbleAI has received funding from the EU's Horizon Europe research and innovation programme (grant agreement 101070679), and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant agreement 10039070).

Launched in 2020, the VEDLIoT (Very Efficient Deep Learning in IoT) project set out to bring high-performance deep learning to the edge and IoT. As the project comes to an end, VEDLIoT coordinators Carola Haumann and Jens Hagemeyer (Bielefeld University) summarize some of the project's main achievements.

The VEDLIoT project

Next-generation accelerated IoT

VEDLIoT tackles the complex task of integrating deep learning into internet-of-things (IoT) devices, which often have limited computational capabilities and strict power consumption limits. Addressing the problem from different angles, the project has delivered results across the computing stack, as follows:

- A scalable artificial intelligence of things (AIoT) **hardware platform**, spanning edge to cloud, equipped with optimized hardware components and specialized accelerators tailored for IoT applications, which takes advantage of microserver technology and heterogenous computing for increased performance and efficiency.
- **Advanced middleware** that simplifies the programming, testing, and deployment of neural networks across diverse hardware environments.
- A specialized **architectural framework** for requirements engineering to define, synchronize, and coordinate the requirements and specifications of AI components alongside traditional system elements.
- Integrated **safety and security by design** principles and components throughout the entire framework.

These strategies have been rigorously tested and refined through demanding use cases in critical industry sectors, including automotive, automation, and smart-home technologies. Furthermore, ten additional use cases have been included through an open call, expanding the scope of potential applications.

The figure below provides a detailed view of the VEDLIoT project's achievements and developments. It illustrates the progression from the foundational microserver hardware platform to the toolchains and application use cases, with security considerations and requirements engineering being integral throughout the development process.

The different components of the VEDLIoT project are detailed in the following sections. These components are integral to the project's success and contribute significantly to its overall functionality and effectiveness.

Customizable hardware platforms and flexible AI accelerators

Unlike traditional hardware platforms that only support uniform devices, the RECS platform, powered by AI-enabled microserver technology, enables the integration of a variety of technologies. This adaptability allows for precise customization of the platform to suit specific applications, creating an extensive cloud-to-edge solution. All RECS variants adhere to a design philosophy centring on a densely packed, highly integrated heterogeneous microserver system, featuring a high-speed, low-latency communication framework.

There are three distinct RECS platforms, each tailored for different environments:

- the RECS|Box for cloud / datacentre applications
- t.RECS for edge computing
- u.RECS for IoT applications

These servers employ industry-standard microservers that are interchangeable, allowing for easy updates to the latest technology by simply swapping out a microserver.

VEDLIoT also offers a diverse array of accelerators suitable for applications ranging from small embedded systems with minimal power requirements to robust high-power cloud platforms. Choosing the right accelerator from this broad spectrum is a complex task. VEDLIoT has addressed this challenge by conducting in-depth evaluations of various architectures, including graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs). The project meticulously analyses these accelerators' performance and energy efficiency to ascertain their appropriateness for specific applications.

Hardware-aware pruning and quantization

Trained deep-learning models can be compressed by several factors with minimal accuracy loss because of their inherent redundancy. However, theoretical speed gains obtained

by model compression often don't match actual hardware performance due to a lack of consideration for specific hardware. VEDLIoT addresses this by focusing on hardware-aware model optimization and co-design in deploying models on edge devices, encompassing training, optimization, compilation, and runtime.

The EmbeDL toolkit provides a suite of tools and methods for tailoring deep-learning models to run efficiently on devices with limited resources. It accounts for the unique constraints and characteristics of different hardware, enabling developers to effectively compress, quantize, prune, and optimize models. This approach ensures minimal resource use while preserving high accuracy in inference.

To promote interoperability, VEDLIoT uses Kenning, a tool that translates neural networks into a unified format using ONNX. The project also utilizes Renode, an open-source simulation framework, to test FPGA accelerator prototypes. Renode acts as a functional simulator for complex heterogeneous systems, capable of simulating complete systems-on-chips (SoCs) and running the same software that is used in actual hardware environments.

Safety and security

In its quest to integrate deep learning with IoT, VEDLIoT strongly emphasizes the principles of security and safety. The project integrates trusted execution environments (TEEs) such as Intel SGX and Arm TrustZone, along with open-source runtime platforms like WebAssembly. TEEs create secure zones that isolate crucial software components, providing protection from unauthorized access and alterations. By employing WebAssembly, VEDLIoT establishes a consistent execution environment throughout the entire range, from IoT devices to edge computing and cloud services.

Focusing on TEEs, VEDLIoT has developed Twine and WaTZ as trusted runtime environments for Intel SGX and Arm TrustZone, respectively. These runtime environments simplify the process of developing software in secure areas by leveraging the modular nature of WebAssembly. This strategic integration forges a connection between TEEs and AIoT, facilitating the seamless integration of deep learning frameworks. VEDLIoT's use of WebAssembly within TEEs ensures a hardware-neutral, strong defence against external threats, thereby preserving the privacy and integrity of both data and deep-learning models.

Architectural framework for the design of AIoT systems

This framework is composed of various architectural views that cater to specific design needs and quality attributes of the system, with a special focus on security and ethical issues. By employing these architectural views as guiding templates and completing them, it becomes possible to identify the relationships and dependencies between the architectural views that determine

The VEDLIoT consortium

quality and other design choices, such as the construction of AI models, the selection of data, and the architecture of communication systems. This comprehensive approach ensures that security and ethical considerations are thoughtfully woven into the overall system design. Such integration underscores VEDLIoT's dedication to creating robust solutions and addressing the evolving challenges in AI-driven IoT systems.

VEDLIoT is hosting the '**DL4IoT – Workshop on Deep Learning for IoT**' at **HIPEAC 2024** on Friday, 19 January 2024, presenting results from VEDLIoT and related projects.

FURTHER INFORMATION

VEDLIoT website vedliot.eu

'Teaching the IoT to learn with VEDLIoT'
Interview with Jens Hagemeyer (Bielefeld University) by Wisse Hettinga (eeNews Europe)
HiPEACinfo 67, pp. 18-19, November 2022
hipeac.net/magazine/7163.pdf#page=18

'Using WebAssembly for a more interoperable, secure cloud-edge continuum.' Jämes Ménétrey, Pascal Felber, Marcelo Pasin and Valerio Schiavoni (Université de Neuchâtel)
HiPEACinfo 68, pp. 24-25, January 2023
hipeac.net/magazine/7164.pdf#page=24

Talks by the VEDLIoT consortium in the Computing Systems Week Tampere 2022 playlist
bit.ly/CSWSpring22_HiPEACTV_playlist

HiPEAC interview with Jens Hagemeyer at the 2022 HiPEAC conference
youtu.be/ACNSBQT6WsM

VEDLIoT (Very Efficient Deep Learning in IoT) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 957197.

The internet of things (IoT) has been a game changer thanks to its capacity to gather real-time data from sensors and other devices, powering a wide range of applications. In parallel, by bringing computational and data-storage capabilities closer to devices, edge computing promotes reduced latency and enables insights in real time. However, implementing IoT and edge-computing infrastructure can be challenging, particularly in harsh environments. In this article, Domenico Garlisi (University of Palermo), Tiziana Cattai, Redemptor Jr Laceda Taloma, Ioannis Chatzigiannakis and Francesca Cuomo (all Sapienza University of Rome) explore some of the key issues in IoT edge computing in the context of water distribution.

How data streams plug water leaks, and other stories

Advances, challenges and applications of IoT edge computing for water-distribution networks

With water scarcity affecting communities and ecosystems across the globe, appropriate water management is a key plank of sustainability. 'Escalating demand for clean, drinkable water, coupled with the repercussions of climate change, has spurred research focusing on improved water management,' explains Domenico Garlisi. 'According to the European Commission, average water wastage in the European region is around 26%, spiking to a staggering 45% in certain areas. This is a clear indication of the urgent need for effective water-management strategies and solutions to minimize waste and ensure sustainable use of this vital resource.'

To address this, Domenico and his colleagues are working towards integrating an IoT infrastructure with edge-computing capabilities in water-distribution networks – an integration which they say will enhance efficiency, reliability and sustainability. The functionality provided by edge computing in the IoT is a boon for the water and wastewater treatment sector, says Tiziana Cattai. 'Edge computing is inaugurating a paradigm shift whereby data is stored and processed closer to IoT devices. This minimizes latency and allows immediate insights to be derived,' she notes. 'Edge computing can be used to monitor water flow, detect potential contamination, and provide alerts in the event of a potential problem, including the detection of leakages. It also offers improved scalability, enabling more efficient expansion of operations. This helps with the rapid growth of water and wastewater treatment operations, as new nodes can quickly be added to the network as required.'

Real-time data analytics is another area where edge computing is key, says Tiziana. 'Real-time processing facilitates quicker decision making, enhances operational efficiency, and leads to substantial cost savings. In addition, edge computing helps reduce latency: by processing data locally, rather than having to wait for a server to return the data, latency is significantly reduced. This is especially important in the water and wastewater treatment industry, allowing faster responses to changing conditions,' she adds.

Challenges and technical choices

However, the implementation of edge computing and the IoT in the water sector does pose challenges, notes Redemptor Jr Laceda Taloma. 'Challenges include the need for a robust, secure, and reliable communication infrastructure capable of handling data collection through diverse data sources installed in harsh environments – which is often not the case in rural, semi-rural and even urban environments,' he says.

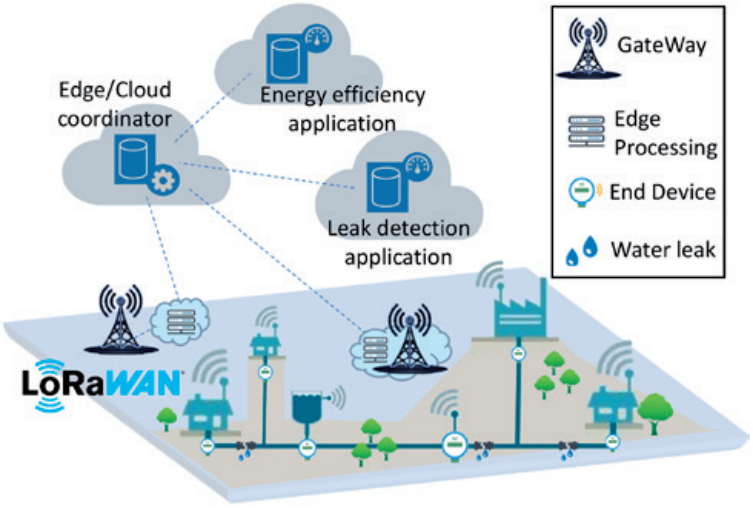
Specialized hardware and software, such as sensors, gateways, analytics, and control algorithms, are also needed. In addition,

From left to right: Domenico Garlisi, Tiziana Cattai, Redemptor Jr Laceda Taloma, Ioannis Chatzigiannakis and Francesca Cuomo

multiple distributed IoT devices and edge nodes need to be managed and coordinated. 'All of this means that we need to revisit the way we manage the network and deploy computation resources across the network, to address the specifics of the water-treatment system. This requires significant resources and expertise, including experienced personnel for system development and maintenance, secure data storage, and compliance with various regulatory and safety requirements,' adds Redemptor.

Choosing a sensor technology for this work is complicated by the fact that there is no universal technology that could fit the purpose of water-distribution network monitoring, Ioannis Chatzigiannakis points out. 'IoT chips with low-power consumption and long-distance wireless communication capabilities – such as wireless MBUS, cellular networks or LPWAN networks – are ideal for these purposes,' says Ioannis. 'LPWAN devices are expected to dominate the field as they can accommodate deployments in underground environments or in remote locations, with different infrastructure requirements, including autonomous LPWANs such as LoRaWAN. The literature suggests that LoRaWAN is the best candidate, thanks to benefits including low power use, extensive coverage, simplicity, and ease of management, although it does face potential scalability issues.'

To enable real-time analysis and processing of the vast amounts of data generated by connected devices within the network edge, the researchers are therefore targeting LoRaWAN technology. 'We propose Edge2LoRa (E2L) a new LoRaWAN-based architecture to support edge processing that builds upon the over-the-air activation (OTAA),' explains Ioannis. 'In E2L, each sensor is serviced by one E2L GW that carries out the data processing tasks on the received sensor data streams. Moreover, E2L employs elliptic-curve cryptography for generating cryptographic keys and enables secure encryption and decryption of the data stream at the edge.



Smarter, more efficient water management

The main focuses of the research are on a) energy-efficient monitoring and b) leak detection. 'For the former, we are harnessing the power of graph theory and graph-signal processing to represent water flow and simultaneously minimize the number of IoT sensors communicating those measurements,' explains Francesca Cuomo. 'The proposed approach significantly reduces energy consumption while ensuring precise flow estimation.'

In the case of leak detection, the researchers are exploring the benefits of machine-learning (ML) algorithms at the network edge to detect leaks in water-distribution networks using real-time streaming data, allowing prompt intervention. 'We are implementing a stream-based clustering technique that profiles nodes in the water distribution network in terms of their spatial proximity and hydraulic characteristics,' says Francesca. 'The core of the approach is the development of a robust model that can analyse the streaming data generated from each cluster and automatically detect potential leakages by identifying patterns of deviation from expected behaviour. Results attain an average accuracy of detecting and localizing the zone of the leakages of about 98.6% when leakages are not present during the training of the ML models. Finally, by moving the process from cloud to edge, the proposed approach reduces the latency by about 86%.'

This research builds on work done in the ELEGANT Horizon2020 project, which is building a unified software programming paradigm for IoT and big data frameworks, and the European Union (EU)-funded RESTART-WITS project, which focuses on analysing data collected by IoT infrastructure in WDN, particularly in massive IoT scenarios.

'This infrastructure offers numerous benefits, including real-time monitoring and decision making, reduced latency, scalability, improved data security, energy efficiency, predictive maintenance, optimized resource use, and cost savings,' adds Domenico.

elegant-h2020.eu
fondazione-restart.it/projects/f13-wits

ELEGANT has received funding from the European Union's Horizon2020 research and innovation programme under grant agreement no. 957286. RESTART-WITS has received funding from the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART") in the WITS (Watering IoTs) focused project.

How SAFEXPLAIN is working to deliver safety-critical AI

Artificial intelligence (AI) will form an essential part of the safety-critical systems – like cars, trains and satellites – of the future. However, AI using deep-learning (DL) methods lacks transparency: the algorithms are powerful, but the underlying decision process is hard to understand. While autonomous systems that use AI demonstrate accurate perception and decision-making, integrating AI components into safety-critical systems still requires a process to ensure that they function transparently.

SAFEXPLAIN
Safe and Explainable Critical Embedded Systems based on AI

The European Union (EU)-funded project SAFEXPLAIN seeks to incorporate AI components into critical systems in a way that is traceable and explainable, thus allowing them to be certified for use in different safety-critical scenarios. In this article, Robert Lowe of SAFEXPLAIN partner RISE (Research Institutes of Sweden) gives us an update on the project.

What steps has SAFEXPLAIN taken to incorporate AI into safety-critical software systems development for autonomous vehicles?

The SAFEXPLAIN consortium has jointly developed a first draft of a new DL-based functional safety management (FSM) lifecycle. This DL-FSM lifecycle maps the functional safety requirements of traditional software development processes (based on the current FSM) to the required steps and phases of DL development and deployment.

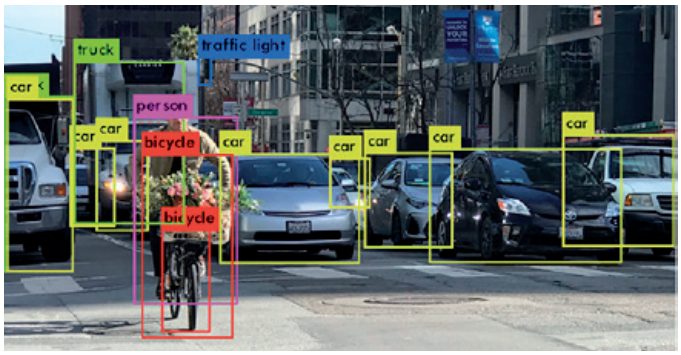
Within these DL phases, we have identified measures for ensuring traceability and explainability within the operational domains of interest, for example for specific automotive, railway and space scenarios. These measures seek to clarify when DL components produce predictions that exhibit the required quality and transparency (e.g. traceable, explainable) for ensuring domain-specific safety.

Explainable artificial intelligence (XAI) is a subfield of AI that is still in its infancy; however, much work has been done in this area over the last few years. The SAFEXPLAIN project explores state of the art XAI methods with respect to our domain-relevant scenarios. XAI methods can help inform the machine (about the automated process regarding whether to accept or reject the DL prediction), the user (to make a safety-critical decision) and AI developers (to improve the DL algorithms used).

Examples of such methods include the use of algorithms that address critical uncertainties in data management and model development / deployment. These uncertainties are as follows:

i) domain specific (uncertainty about the domain, i.e., the scenarios being tested)
ii) epistemic (uncertainty about the model performance)
iii) aleatoric (uncertainty about the degree of randomness in the domain)

Our research identifies how to leverage existing XAI methods to mitigate (during data management and model development) and identify (during model deployment) hazards that stem from these three sources of uncertainty.



What have been the main challenges in implementing the work?

One of the biggest challenges has been using state-of-the-art AI methods in the face of the rapid evolution of knowledge in this field. Something that was state of the art a couple of years (or even a couple of months) ago may now be outdated. This means that identifying DL algorithms that can be used in critical autonomous systems and that require XAI to ensure their safe application needs a degree of foresight into what is likely to stay, or to become relevant, in the coming years. Other challenges include the complexity of the safety architecture involving several DL components. For example, how different predictions and sources of explanation are synchronized / weighted among multiple DL-XAI components.

A trade-off intrinsic to XAI is that interpretability (to users or developers) comes at a cost of accuracy (in DL predictions) or processing speed. For example, something provided by the XAI that is highly intuitive and interpretable to a human (e.g. part of the DL algorithm that learns a specific part of an object) might only approximate what the DL algorithm has actually learned and predicted for a specific situation. Minimizing this trade-off can entail additional processing costs (i.e. more computationally intensive XAI usage).

What are the main results of the project so far?

A first draft of a DL(XAI)-FSM safety lifecycle maps safety requirements for software development to relevant phases of DL lifecycle. Several candidate explainable DL techniques have been identified and evaluated in relation to critical safety requirements throughout the XAI lifecycle. These techniques provide, for example, intuitive explanations, hierarchical functional decompositions of the DL algorithms and automated data labelling of sub-explanations of the DL predictions.

What results can we expect to see over the next few months?

Over the next few months, we will provide the first release of the proposed DL(XAI)-FSM lifecycle specifications. Further iterations of this set of specifications will be reported by the end of the project and will serve as recommendations for adapting existing functional safety standards so that they can certify aspects of software that incorporate DL components. This document will detail how XAI can be used to make the DL components of the lifecycle explainable, traceable (from DL prediction through inner workings of the DL algorithm to the data upon which its predictions depend), and robust.

The specifications will also detail our recommendations on how to deploy 'supervisor' architectures that receive input from:

1) DL components
2) XAI explanations
3) information related to the data, operational domain and hardware (e.g. car lidar / camera / radar sensors)

These inputs will allow for the identification and control of anomalous and artefactual DL predictions.

Innovations in AI are coming thick and fast at the moment. Do you expect any of these developments to have an impact on SAFEXPLAIN?

The rapidly evolving discipline of AI presents opportunities and challenges. The game-changing technology that consists of large (pre-trained) language models (LLMs) has emerged since the inception of the SAFEXPLAIN project. LLMs themselves are not intrinsically explainable due to the billions of parameters that the state-of-the-art models use. However, the potential for combining human-centred language explanations with visual explanations of DL predictions for the benefit of functional safety in the automotive industry is great.

Moreover, increasingly powerful XAI algorithms are being developed that focus on causal and concept-based attribution to DL predictions. These approaches appeal to human-focused reasoning whilst being more invariant to conditions (e.g. sensor noise, lighting, object orientation and location) that have the potential to lead to anomalous output or be subject to malicious adversarial manipulation. By keeping abreast of these developments, we feel confident that the SAFEXPLAIN project will deliver relevant XAI-based safety recommendations rooted in core human-centred AI modelling foundations, now and in the future.

FURTHER INFORMATION

safexplain.eu

Members of the SAFEXPLAIN project consortium

A collaboration between the United States, Ireland and Northern Ireland, the SWEET (Hardware and Software Sustainable Wearable Edge InTelligence) project is delivering smarter, more efficient wearables for healthcare applications – even with limited connectivity. In this article, SWEET researchers Deepu John (University College Dublin), Dimitrios S. Nikolopoulos (Virginia Tech), Bo Ji (Virginia Tech) and Hans Vandierendonck (Queen's University Belfast) tell us more.

Smarter, stronger wearables for all

How the SWEET project is enabling robust, efficient sensing technologies for healthcare applications

Given that real-time monitoring of physiological indicators, coupled with early intervention, can save lives, it's little wonder that wearables are becoming a vital tool in the healthcare sector. However, health services based on machine learning often require wearables with strong predictive abilities, along with enormous data stores, fast networks and fast servers to extract insights from the data collected. These technologies simply may not be available to communities living in areas with limited broadband connectivity, and with few resources to invest in computing and communication infrastructure. This means that machine-learning healthcare services are currently unavailable to large parts of the world's population.

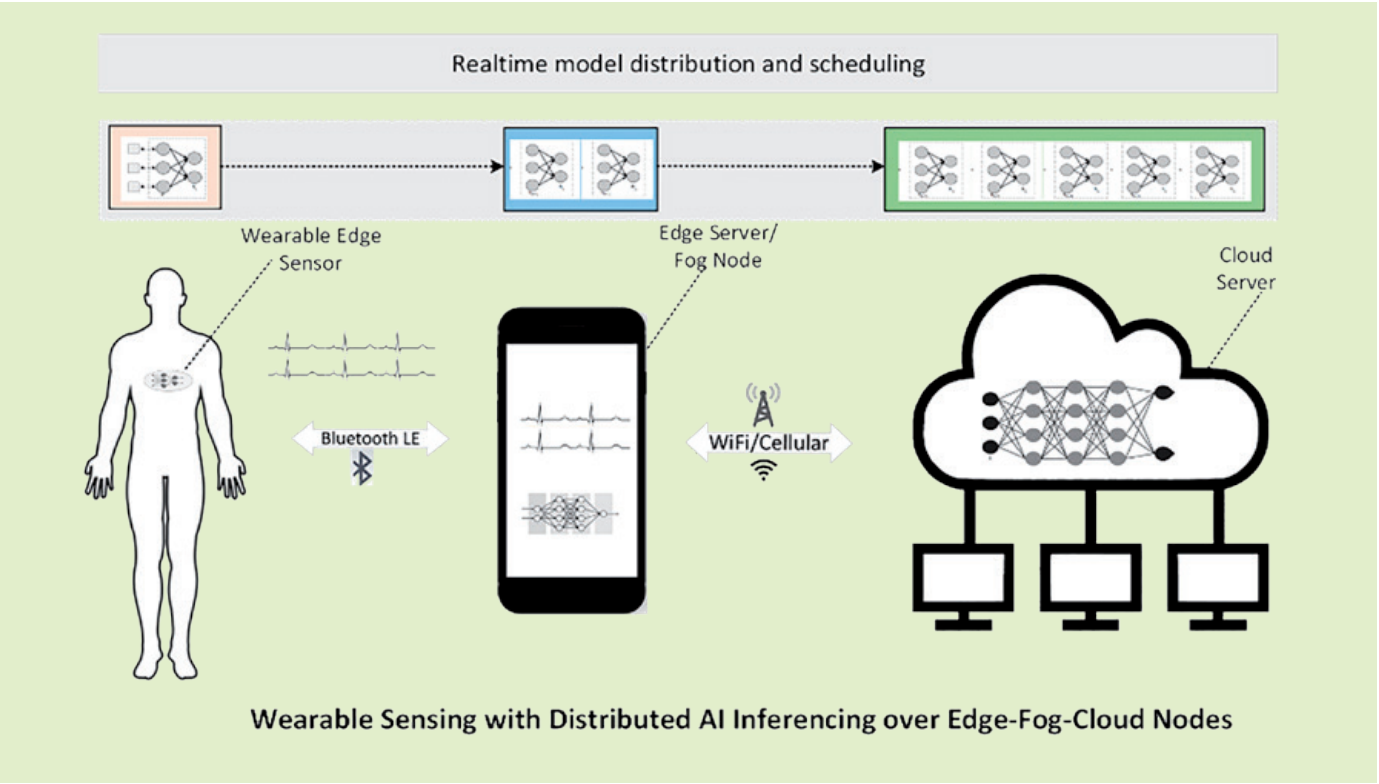
It was this observation that led to the SWEET project, described by the researchers involved as 'a sustainability- and accessibility-focused view of computer systems research'. 'SWEET addresses issues relating to systems aspects of deploying machine learning (ML) models in distributed computing environments, such as edge computing and the internet of things (IoT),' explains Hans Vandierendonck. 'The project takes a cross-stack approach to designing, implementing and evaluating the sustainable and efficient operation of wearable edge intelligence,' he adds. Sustainability in this context is interpreted both as the sustainable operation of the wearable devices and as sustainable access to machine-learning services on the part of underserved communities.

Transatlantic cooperation

The project brings together multiple research teams with complementary expertise – spanning wearable sensors, hardware, software, systems and algorithms – to address this multidisciplinary healthcare challenge. 'We were able to get support through the US-Ireland scheme, a collaborative scheme between the National Science Foundation (NSF) in the United States, the Science Foundation Ireland and the Department for the Economy in Northern Ireland,' says Dimitrios S. Nikolopoulos. 'The NSF has recently established international academic and industry partnership prorgammes to address the US national chip shortage and pave a roadmap for the future or semiconductor design and manufacturing. These initiatives advocate cross-stack approaches to the relevant challenges.'

The project's goal of developing sensing and computing technologies to improve the affordability and accessibility of healthcare monitoring and intervention in underserved populations is relevant across the world. However, access issues are particularly acute in the US, Ireland and Northern Ireland due to the limited connectivity of rural populations, notes Deepu John. 'The so-called "rural broadband gap" in Ireland is among the worst in Europe, according to a 2021 Eurobarometer report. Similarly, almost 30% of the US population faces connectivity challenges, compounded by severe barriers to accessing even basic healthcare services,' he notes.

Left to right: SWEET researchers Deepu John, Dimitrios S. Nikolopoulos, Bo Ji and Hans Vandierendonck



Wearable Sensing with Distributed AI Inferencing over Edge-Fog-Cloud Nodes

More efficient, longer-lasting wearables

One of the major challenges the project has to overcome is prolonging the lifetime of wearable devices that perform biomedical signal acquisition and processing, while expanding their computational and processing capabilities. 'State-of-the-art wearable sensors typically have a battery life of less than 48 hours, which limits their diagnostic yield. A significant portion (>90%) of the power consumed by such sensors is simply used for continuous radiofrequency (RF) transmission (e.g. via Bluetooth) to a handheld device, which then transmits the data to the cloud,' explains Dimitrios.

What's more, the sensors are heavily dependent on connection to the cloud. 'The sensors may have no feedback when a reliable cellular connection to the cloud is not present, and incur exceedingly high response times when the connection to the cloud is interrupted. These problems are particularly prominent and detrimental in rural and low-income communities with limited broadband connectivity,' adds Bo Ji.

The project is therefore also investigating how to perform more efficient, robust and trustworthy machine learning in computing devices outside the cloud, while also finding scalable and sustainable development and deployment models for distribute machine-learning (ML) services, without having to depend on the robustness and availability guarantees of the cloud. To do so, the research team is taking a cross-stack approach, with activity on both the hardware and software sides.

'We propose new hardware accelerators for wearable sensing that enhance wearables' predictive capabilities while dramatically increasing sensor lifetime,' says Deepu. 'We also propose a new transprecise, serverless computing paradigm and associated system software that removes the dependence on real-time physiological monitoring from the cloud. The system will provide real-time ML task scheduling for scaling ML tasks in the IoT-edge continuum, and lightweight service deployment and data caching techniques for running ML services using serverless frameworks on edge devices,' adds Hans. The project will build on recent work published by the consortium on topics including reducing the power use of edge sensors, transprecise computing, serverless computing, and network systems optimization.

While the SWEET framework has broad application potential, it will be evaluated on two specific use cases, chosen for their high impact in terms of saving lives and improving quality of life. These use cases are as follows:

1. monitoring for sudden cardiac arrests
2. onset detection and classification of cardiac arrhythmias

FURTHER READING

SWEET: Hardware and Software for Sustainable Wearable Edge Intelligence

bit.ly/NSF_SWEET_project

Bridging horizons: Exploring the European Cloud, Edge & IoT Continuum initiative

Catarina Pereira (Martel Innovate)

In the dynamic realm of digital transformation, the European Cloud, Edge & IoT Continuum initiative (EU-CEI) plays a key role, propelled by the concerted efforts of two coordination and support actions (CSAs): Open Continuum and UNLOCK-CEI. These CSAs strategically address the supply and demand sides of the CEI continuum, navigating the complexities of cloud, edge, the internet of things (IoT), artificial intelligence (AI), and connectivity.

Open Continuum: Nurturing the supply side

Albert Seubers, director of Martel Innovate BV and leader of Open Continuum, highlights the drive behind the CSA: 'By allowing open and secure collaboration between technologies from different vendors and different ages, the reduction of scarce resources is supported.' Open Continuum promotes European strategic autonomy and interoperability through an open ecosystem, guided by open source and open standards.

The four main objectives of Open Continuum encapsulate the essence of its mission. It seeks to promote the establishment of a European industrial open ecosystem, map and analyse the supply-side landscape, engage European Union (EU) industrial and research actors, and coordinate existing EU projects for an open European ecosystem for the cloud-edge-IoT continuum.

UNLOCK-CEI: Unleashing market potential

Golboo Pourabdollahian, consulting manager at IDC – European Government Consulting, and leader of UNLOCK-CEI, emphasizes the crucial role of the CEI computing continuum: 'Creating this CEI environment – a computing continuum of data collection, storage and processing from edge to cloud – will be an essential component of a globally competitive, secure and dynamic data-agile economy in Europe.' UNLOCK-CEI focuses on demand-side drivers and challenges, aiming to identify technology-driven innovations and business opportunities.

UNLOCK-CEI's objectives encompass a systematic assessment of the European CEI demand landscape, defining market scenarios, building and activating the CEI industry constituency, and creating a productive interface between the demand and supply sides. Pourabdollahian underscores the importance of understanding the needs and requirements of European adopters and fostering collaboration among key stakeholders.

Harmonizing synergies and task forces for a digital future

The collaboration between Open Continuum and UNLOCK-CEI forms a harmonious synergy that addresses both the supply and demand sides of the CEI continuum. Open Continuum lays the foundation for an open ecosystem, emphasizing interoperability and collaboration among diverse technologies. Simultaneously, UNLOCK-CEI explores market opportunities and technological solutions to drive demand, fostering a productive interface between the demand and supply constituencies.

The EUCloudEdgeIoT.eu initiative extends its impact through dedicated task forces aimed at coordinating and disseminating information across stakeholders in the cloud, edge, and IoT ecosystems. These task forces, covering strategic liaisons, open-source engagement, architecture, ecosystem engagement, market and sectors, and communication, serve as multipliers by creating common strategies, approaches, and methodologies within the CEI Ecosystem.

As Europe positions itself at the forefront of shaping the CEI ecosystem, these CSAs pave the way for a globally competitive, secure, and dynamic data-agile economy. Through collaborative efforts and the legacy of existing EU projects, the EU-CEI initiative strives to realize a unified cloud, edge, and IoT continuum that aligns with the evolving needs of businesses, governments, and citizens alike. In a world increasingly shaped by digital technologies, the EU-CEI Continuum emerges as a beacon guiding Europe towards a connected and resilient future.

FURTHER INFORMATION
eucloudedgeiot.eu

Innovating across the edge-AI computing continuum

Ovidiu Vermesan, Chief Scientist, SINTEF, Norway

Edge AI converges edge computing, artificial intelligence (AI), and the internet of things (IoT), which allows machine-learning (ML) and deep-learning (DL) algorithms to run in real time on local edge devices with constrained computing capacity. Edge AI brings intelligence and high-performance capabilities to the edge, from micro- to deep- and meta-edge, where sensors, actuators and IoT devices are located.

By combining the effectiveness of AI, the adoption of IoT devices, and the distributed capability of edge computing, the use of federated learning further advances edge AI applications. In federated learning, AI models are trained on edge devices using their local data, with updates to the model sent to a server, while the model updates are merged into a consolidated model, which is pushed back to client edge devices.

Deploying edge AI patterns across various locations and applications presents specific challenges such as data gravity (an opportunity for edge processing, as minimizing the space between data and processing results in lower latency for applications and faster throughput for services), heterogeneity, scale, and resource constraints. Distributed AI can address edge AI's challenges by integrating intelligent data collection, automating the data and AI life cycles, adapting and monitoring boundaries, and optimizing data and AI pipelines.

Distributed edge AI allows the allocation, coordination, and forecasting of tasks, objectives, or decision performance within a multi-agent intelligent environment. This enables the ability to scale edge AI applications to many devices and entitles AI algorithms to autonomously process across multiple systems, domains, and devices on the edge.

Edge AI technology advances have opened prospects for machines and edge IoT devices to intelligently operate and perform advanced tasks in real time under different circumstances.

The efficiency and increased performance of deploying AI models at the edge emerge from advanced computing architectures (graphics processing units (GPUs), tensor processing units (TPUs), intelligence processing units (IPUs), etc.) with AI-based accelerators and mature neural-network topologies, real-time capabilities with low-latency connectivity capabilities and increased processing capabilities at the edge combined with advanced sensing and actuating of edge IoT devices.

Edge AI at the heart of the Chips JU

Europe is investing in breakthrough edge AI technologies to boost collaborative research and development across the continent, advance the state of the art and implement embedded and edge AI across information technology (IT) and operational technology (OT) systems to unlock untapped value for real-time industrial applications to achieve sustainability and operational efficiency.

The Chips Joint Undertaking (JU) is the leading implementer of the Chips for Europe Initiative (expected total budget of €15.8 billion until 2030). Its goal is to strengthen Europe's semiconductor ecosystem and economic security by managing an expected budget of nearly €11 billion by 2030, provided by the EU and participating states.

In October, the European edge AI ecosystem gathered in Athens for the European Conference on EDGE AI Technologies and Applications – EEAI 2023. With more than 130 delegates, the event was co-organized by the Chips JU EdgeAI project in collaboration with CLEVER, REBECCA, TRISTAN, NeuroKit2E, LoLiPoP IoT projects, ECSEL JU ANDANTE, AI4CSM, AI4DI, TEMPO and InSecTT projects to deliver a platform to exchange knowledge and ideas among experts interested in advances in edge AI circuits and device design, edge AI hardware architectures, industrial edge AI technologies, tool-chains, and applications.

In late autumn 2023, the EdgeAI project also participated in the Chips JU Launch Event exhibition in Brussels, which brought together over 800 participants representing experts and practitioners in the electronic components and systems sector.

FURTHER INFORMATION

EdgeAI: Edge AI Technologies for Optimised Performance Embedded Processing
edge AI-tech.eu

The Edge AI project at the recent Chips JU launch in Brussels

HiPEAC 2024 sponsor Frontgrade Gaisler designs the toughest systems for the harshest environments. Technology journalist Stuart Cording ('The Electronics Reporter') interviewed the company's general manager, Sandi Habinc, to find out about intellectual property (IP) for space, the rise of RISC-V, how European research expands boundaries, and more.

'Our electronic systems and components are crucial in space-exploration missions'

Tell me about Frontgrade Gaisler.

Frontgrade Gaisler is a world leader in embedded computer systems for harsh environments. Our personnel have extended design experience and have been involved in establishing European standards for application-specific integrated circuit (ASIC) and field-programmable gate array (FPGA) development for space applications. We are located in Gothenburg, Sweden, and we also have employees working remotely from other European countries.

We are part of Frontgrade Technologies, a US company that develops and manufactures radiation-hardened (rad-hard) microelectronics, radiofrequency (RF) transmission and antenna solutions, high-power amplifiers, motion-control products, and power-management solutions.

What are your products?

At Frontgrade Gaisler, we provide a complete framework for the development of processor-based system-on chip (SoC) designs. This framework is centred around the LEON (SPARC) and NOEL-V (RISC-V) processor cores, and it includes a large IP library named GRLIB and related software development tools – all the necessary components to create high-quality and high-performance products. We also use the same framework internally to design rad-hard microprocessors and SoCs, like the GR740, our quad-core high-performance rad-hard microprocessor.

Who are your customers and what are their key requirements?

Our customers primarily include government space agencies, commercial space companies, satellite manufacturers, and other entities involved in space missions. I would say that the top requirements are:

The GR740 Quad-Core LEON4FT SPARC V8 Processor

- **Radiation hardening and reliability**: Space environments expose electronic components to high levels of radiation, which can cause errors or failures. Our customers require processors and IP cores that are specifically designed and tested to be radiation hardened and reliable in these challenging conditions.
- **High-performance processing**: Space missions often require advanced processing capabilities for tasks such as data processing, navigation, communication, and scientific analysis. Frontgrade Gaisler customers seek processors that offer high computational power while meeting stringent power and thermal constraints.
- **Space-qualified and flight-tested solutions**: Our customers demand space-qualified components that have been thoroughly tested and validated for space missions. They look for products with a proven track record supporting successful spaceflight missions.
- **Low power consumption**: Power efficiency is crucial for space applications as it impacts the overall mission duration and thermal management. Customers prioritize processors and IP cores that offer low power consumption without compromising performance.
- **Scalability and flexibility**: Space missions vary in complexity and requirements, so our customers need solutions that are scalable and adaptable to different mission profiles. Processors and IP cores like our NOEL-V can be easily integrated into various space systems and can be customized to meet specific mission needs.

What types of space applications are your customers tackling?

Our electronic systems and components have been crucial in missions like the Mars rovers, interplanetary probes, asteroid missions, and space telescopes, as well as in many satellite constellations. For example, a recently launched spacecraft known as Juice is a European Space Agency (ESA) mission to explore Jupiter and its icy moons. Frontgrade Gaisler's dual-core microprocessor was employed in seven out of the ten science instruments for the Juice mission.

Do your customers build their own silicon or do they integrate your IP into an FPGA?
Both options are possible! Frontgrade Gaisler's library uses a consistent method for simulation and synthesis, making it easy to use with different third-party electronic design automation (EDA) tools.

Moreover, GRLIB is designed to be used in digital system designs, independent of the target technology. It supports various FPGA technologies including AMD/Xilinx, Intel/Altera, Lattice, Microchip, and NanoXplore, and it can also be employed for ASIC design.

What makes your RISC-V IP suitable for space?
We have extensive experience in developing fault-tolerant SoCs. In space, radiation-induced errors can occur, and detecting such errors is just the beginning. The real challenge lies in effectively handling these errors in a way that maintains the system's integrity and functionality.

For instance, in deep-pipeline processors where multiple actions take place within a single clock cycle, we employ techniques to prevent error propagation and ensure fault containment. When the error is detected, the correction process is handled transparently to software, ensuring uninterrupted execution. Moreover, our fault-tolerance mechanisms are designed to complete error-handling functionalities within a timeframe that does not adversely affect the system's performance or introduce delays that could be detrimental to critical space missions.

We also test our processors in relevant operational environments. Our team travels to various radiation testing facilities across Europe to validate the IP through accelerated ground testing to demonstrate performance and functionality.

Why is RISC-V interesting for space customers? What about alternative processor architectures?
We have experienced significant success with SPARC V8, particularly with LEON processors, which have been adopted in space missions worldwide. SPARC's triumph can be attributed to its spin-in of commercial technology, non-proprietary architecture, availability of flight silicon from various vendors, and global acceptance.

RISC-V has emerged as a compelling alternative with open-source benefits similar to SPARC in the 1990s. ESA-funded research and development (R+D) activities explore RISC-V's potential, including fault-tolerant concepts for space applications and porting of hypervisors. Aligning on one open architecture between Europe and the US would help accelerate hardware development and create a shared market for niche space software.

Has your RISC-V processor been adopted in European Union (EU) projects?
Yes, Frontgrade Gaisler's RISC-V processor has been chosen to be a part of initiatives such as De-RISC, SELENE, and ISOLDE.

As part of the De-RISC project, which addressed computer systems in the space domain, an international consortium introduced a hardware and software platform based on our RISC-V core. The project output was a multicore RISC-V SoC design running the XtratuM hypervisor from fentiSS, together with example applications provided by Thales Research and Technology.

With the ISOLDE project, which has just recently begun, we're moving a few more steps forward. The consortium will develop high-performance RISC-V processing systems and platforms based on NOEL-V, and it will bring various building blocks to at least technology readiness level 7. The aim is to create industrial-grade, open-source support for development, verification, and maintenance to promote the use of ISOLDE's high-performance components in industrial-quality products.

In addition to FPGA development boards, what tools do you offer to help with the selection and development process?
Frontgrade Gaisler offers a range of hardware and software tools designed to facilitate and streamline the development of space-grade electronic systems. The GRMON debugger is an essential tool for debugging and monitoring the operation of our processors in real time. It provides valuable insights into system behaviour, helping to identify and resolve potential issues.

Frontgrade Gaisler provides TSIM, a high-performance behavioural simulator of the LEON processors. Using the simulator, we can develop and debug target applications before the real flight hardware is available, thereby shortening the product development cycle.

We do also provide several software packages such as board-support packages (BSPs), device drivers and compiler toolchains for various operating systems, such as Linux, RTEMS, Zephyr and VxWorks.

In short, Frontgrade Gaisler offers a complete development environment, from the processor IP to the flight-qualified bootloader and the operating system. This allows us to support customers throughout the development process.

You can watch the video version of this interview on HiPEAC TV:
bit.ly/HiPEAC24_sponsor_interview_Frontgrade_Gaisler

Catch up with the Frontgrade Gaisler team in the industry exhibition and industry session at HiPEAC 2024: bit.ly/HiPEAC24_industry_session

Find out more about the GR740 gaisler.com/gr740

Innovation Europe

In this edition of Innovation Europe, we learn about MLSysOps' self-managing systems, how INCODE is reimagining the IoT and edge computing, and how dAIEDGE is building the European edge AI community. Plus how FPG-AI is taking accelerators space forward and how ACROSS has delivered a cross-stack platform for HPC and AI workloads.

Funded by the European Union

SELF-MANAGING SYSTEMS FROM EDGE TO CLOUD: THE MLSYSOPS PROJECT

In response to the escalating deluge of data processed by computing systems, there has been a shift towards processing data as close to its origin as possible, known as edge computing. The emergence of cloud-edge computing exacerbates the already complex task of managing diverse and dispersed resources, this time on an immense scale, rendering human-in-the-loop management entirely impractical. To achieve a system and application management approach that is dynamic, flexible, and requires minimal user intervention, the concept of autonomic computing systems was introduced long ago, referring to systems capable of self-management based on high-level objectives set by administrators.

The Horizon Europe project MLSysOps project will extend the autonomic paradigm by introducing a control framework powered by machine learning (ML) that interfaces with available management mechanisms. Running for three years, and with a consortium of 12 partners from eight countries, MLSysOps will also introduce hierarchical, distributed, explainable, and adaptable ML models for autonomous system operation of the cloud-edge-IoT continuum. To achieve adaptability, MLSysOps incorporates continual ML model learning with intelligent retraining concurrently with application execution. The project prioritizes openness and expandability, making use of explainable ML techniques and providing an application programming interface (API) for interchangeable ML models.

MLSysOps considers crucial aspects like energy efficiency (including the use of sustainable / green energy sources), performance optimization, minimizing latency, efficient and robust storage, devices with limited resources, and network connectivity. It employs ML models to co-optimize such objectives in a challenging computational environment. The framework architecture of MLSysOps separates management from control and seamlessly integrates with popular management frameworks at various layers of the spectrum.

The MLSysOps consortium

MLSysOps will be evaluated through two well-defined use cases utilizing cloud, smart, and deep-edge infrastructures:

- **Precision agriculture**: one of the MLSysOps partners is providing a fully automated on-tractor system for monitoring and managing the complete lifecycle of various crops. This system will be enhanced to seamlessly integrate with a real-time scanning system-equipped autonomous unmanned aerial vehicle (UAV), demonstrating the cooperative, ML-driven, power of multiple intelligent edge devices.
- The project's **smart-city application** will leverage smart lampposts equipped with NVIDIA Jetson edge nodes. These lampposts will enable tracking solutions, crisis detection, and identification of traffic congestion. Machine learning (ML) will play a crucial role in optimizing operations to prevent unnecessary usage of energy, storage, and processing resources.

PROJECT NAME: MLSysOps: Machine Learning for Autonomic System Operation in the Heterogeneous Edge-Cloud Continuum
START/END DATE: 01/01/2023 – 31/12/2025
KEY THEMES: computing continuum, machine learning, heterogeneity, resource management, continual ML, explainable ML, accelerators, energy efficiency, green energy, trust
PARTNERS: Greece: University of Thessaly, Nubis, Augmenta; Italy: Universita della Calabria, NTT Data Italia; Netherlands: Technische Universiteit Delft; Ireland: University College Dublin; France: Institute National de Recherche en Informatique et Automatique (INRIA); Portugal: Fraunhofer Portugal, Ubiwhere; Israel: NVIDIA; Denmark: Chocolate Cloud
BUDGET: € 5,711,250
mlsysops.eu
linkedin.com/company/mlsysops
@mlsysops

MLSysOps has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement ID 101092912.

BREAKING THE EDGE COMPUTING STATUS QUO: THE INCODE PROJECT

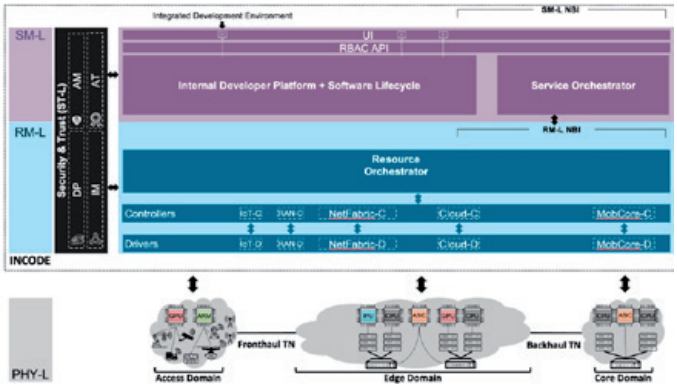
The rise of cloud computing technologies and the shift of processing intelligence to the network edge has made the private use of the edge at scale more accessible. However, increasing edge capacity is not enough to unlock the full potential of edge computing systems. The INCODE project is a pioneering initiative which aims to address this challenge.

Funded by the European Commission's Horizon Europe programme for research and innovation, INCODE brings together nineteen partners committed to breaking the edge computing status quo by creating a wide-open, secure and trusted IoT-to-edge-to-cloud compute continuum that will realize the true potential of edge intelligence. To this end, over a three-year period, the project consortium will design and develop an open platform for the deployment and dynamic management of end-user applications across distributed, heterogeneous and trusted internet-of-things (IoT) edge node infrastructures, with enhanced programmability features and tools. The platform will do so by implementing innovative design approaches and will constitute a fully integrated infrastructure under the cloud-managed INCODE architecture.

How it works

Programmability is at the heart of INCODE's unique approach to reimagining the IoT and the potential of edge computing. In the project's vision, all smart devices – from everyday sensors to sophisticated industrial machines – work together seamlessly, driven by the ability to be fully programmable. Imagine each device having its own unique identity, a digital fingerprint secured by blockchain and advanced hardware certification. This authentication ensures not only the legitimacy but also the integrity of each device.

A novel dynamic orchestration system then distributes tasks across a grid of infrastructure-management instances. This intelligent coordination ensures efficient workload management, a crucial aspect in a world teeming with diverse devices and locations. The core innovation lies in the creation of a fully programmable data plane. This means that the platform adapts and responds to diverse smart IoT devices, edge nodes and powerful servers, creating a collaborative ecosystem.



Thanks to this novel approach to programmability and orchestration systems, INCODE unlocks the potential of applications with a harmonized approach across the IoT, edge and cloud computing, crafting customized applications that seamlessly navigate through this diverse technological terrain. This would not be possible without the use of blockchain technology, which ensures secure data sharing while providing an auditable record of datasets and software changes. With INCODE, programmability is not just a feature but rather a catalyst that propels us into an era where the IoT is intelligently programmed for a dynamic and responsive future.

INCODE's architecture is being validated in four application areas, as follows:

a. **Smart logistics**, to evaluate scenarios at terminal stations through application-level programmability;
b. **Utilities inspection**, to create a digital prototype high voltage substation;
c. **Smart worker assistant**, to achieve effective management and improve adaptive human-machine interaction in smart factories, while monitoring operators' health and wellbeing;
d. **Smart PPDR**, to evaluate scenarios for smart public protection disaster relief with drones and ground robots through collaboration and coordination.

PROJECT NAME: INCODE: Programming Platform for INtelligent COllaborative DEployments over Heterogeneous Edge-IoT Environments
GRANT AGREEMENT NUMBER: 101093069
START/END DATE: 01/01/2023 – 31/12/2025
KEY THEMES: IoT, edge computing, swarm intelligence, programmability, computing continuum
PARTNERS: Luxembourg: Uni Systems (coordinator); Israel: Red Hat; Germany: NEC Labs Europe, FIWARE Foundation, AgentsCape, Romania: Siemens; Italy: MADE Competence Centre 4.0, Politecnico di Milano; Cyprus: Ubitech, Suite5 Data Intelligence Solutions; Bulgaria: K3Y; UK: University of West Scotland, University of Manchester, Greece: IPTO - Independent Power Transmission Operator, iLink, ArĐs Developments Hellas, Axon Logic, University of Patras; Switzerland: Martel Innovate
BUDGET: € 7,153,035.38

incode-project.eu
@INCODE_eu
linkedin.com/company/incodeproject

BUILDING THE EUROPEAN EDGE AI COMMUNITY: ANNOUNCING THE DAIEDGE NETWORK OF EXCELLENCE

By combining edge computing and artificial intelligence (AI) to process data directly at the point of origin, devices can make decisions in milliseconds, without insecure connections, high latency, large energy overheads or transmission costs. Edge AI is therefore a pathfinder and accelerator for many new applications in areas such as autonomous driving, personalized digital assistance and intelligent service robots.

Under the umbrella of the European AI Lighthouse, the dAIEDGE network of excellence promotes the application of AI on edge computing platforms. Led by DKFI, the German Research Center for Artificial Intelligence, experts in AI, embedded computing, microprocessors, distributed hardware and software, computer science, and computer engineering will work closely together to:

- mobilize the AI and edge community
- connect AI-on-demand platforms, digital innovation centres, and AI / edge projects with relevant stakeholders
- initiate European partnerships and projects
- provide ideas, tools, services, guidelines and identify trends to support the next generation of edge AI technologies

Advanced edge AI technologies for different industries

To accelerate the digital and green transformations through advanced AI technologies, applications and innovations, dAIEDGE builds on the existing assets and strengths of European industry. The main objective is to support and ensure rapid development and market adoption of distributed AI technologies, including hardware, software, frameworks and tools. The applications of dAIEDGE are expected to be used in a wide range of fields, from the internet of things (IoT) and robotics to transportation systems and healthcare.

dAIEDGE will work closely with major European AI initiatives such as HumanE-AI-Net, CLAIRE, ELLIS and AI4EU. To promote the mobility of scientists through research exchanges and industrial research projects, the network of excellence will also support 30 projects via three open calls, offering a total of €1.8 million in funding.

PROJECT NAME: dAIEDGE: A network of excellence for distributed, trustworthy, efficient and scalable AI at the Edge
START/END DATE: 01/09/2023 - 31/08/2026
KEY THEMES: edge computing, artificial intelligence (AI)
PARTNERS: Germany: Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) (coordinator), Deutsches Zentrum für Luft- und Raumfahrt e.V., Fraunhofer Gesellschaft; Switzerland: Aegis Rider, Bonseyes Community Association, Centre Suisse d'Electronique et de Microtechnique (CSEM), ETH Zürich, Haute école spécialisée de Suisse (HES-SO); Sweden: Blekinge Institute of Technology; France: Commissariat à l'Energie Atomique et aux énergies alternatives (CEA), Institut national de recherche en informatique et automatique (Inria), SAFRAN Electronics and Defense, Sorbonne Université, CNRS; Belgium: Centre d'excellence en technologies de l'information et de la communication (CETIC), imec, Katholieke Universiteit Leuven; Poland: FundingBox Accelerator SP; Greece: Foundation for Research and Technology – Hellas; Italy: HIPERT SRL, ST Microelectronics, University of Modena and Reggio Emilia; Bulgaria: INSAIT - Institute for Computer Science, Artificial Intelligence and Technology, Spain: IoT Digital Innovation Hub, University of Castilla-La Mancha, University of Salamanca, Vicomtech; Norway: SINTEF AS; Ireland: Synopsys International Limited, Thales, Ubotica Technologies Limited; UK: University of Edinburgh, University of Glasgow; Finland: Varjo Technologies; Netherlands: VERSES Global B.V.
BUDGET: € 14.4 million (of which €10.7 million is funded by the European Union)

daiedge.eu
@dAIEDGE
linkedin.com/company/daiedge

Partners at the dAIEDGE kick-off meeting

FPG-AI: ESA AND THE UNIVERSITY OF PISA JOIN FORCES TO ACCELERATE AI IN SPACE

Recent years have seen an upsurge of interest in artificial intelligence (AI) within the space community, driven by system miniaturization and intensifying commercial competition. Field-programmable gate arrays (FPGAs) have emerged as potent accelerators for AI algorithms, and automating their design is becoming more and more important for the future.

To take FPGA AI acceleration in space forward, the European Space Agency (ESA) and the University of Pisa have begun an ambitious project, FPG-AI, aimed at transforming the landscape of AI deployment on satellites. FPG-AI is designed to empower a wide range of users, regardless of specific skills, to accelerate AI models on FPGAs while reducing development time.

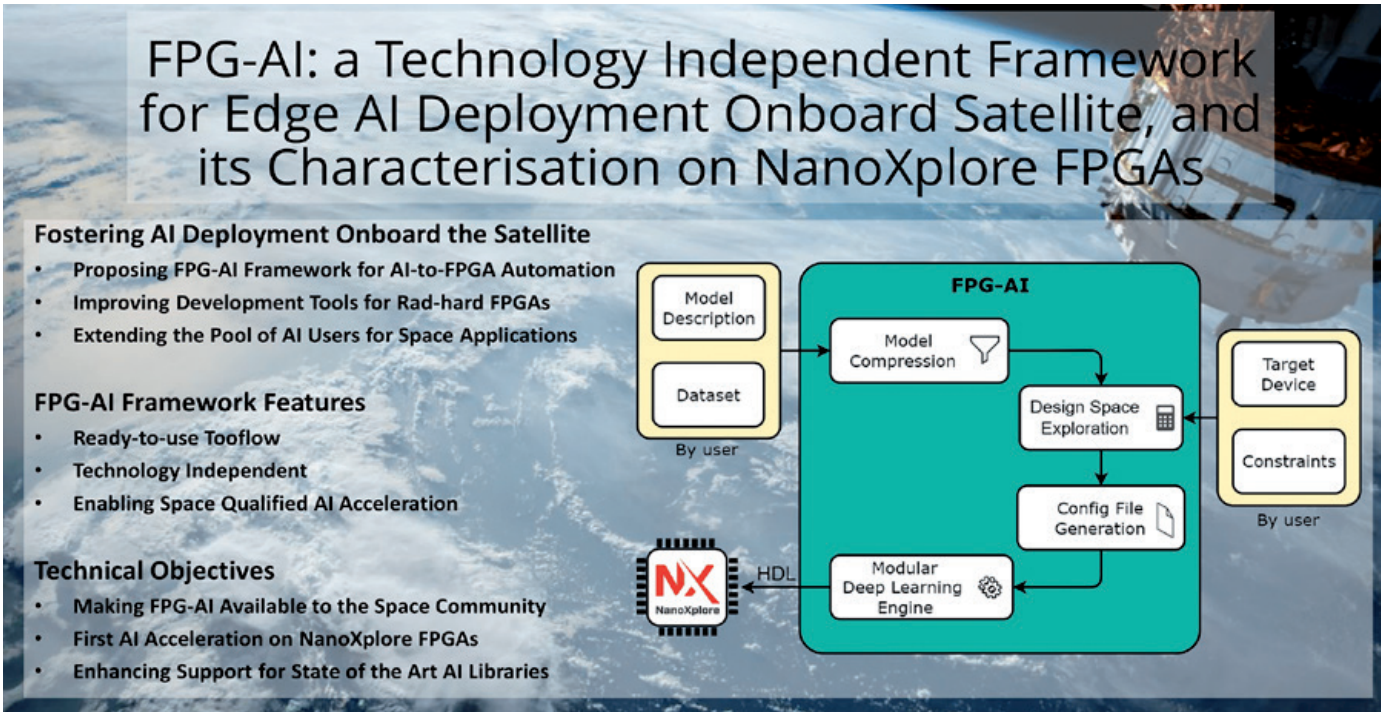
What sets FPG-AI apart from other solutions is its non-vendor-specific nature, eliminating technology limitations on AI accelerator characterization in terms of inference time, resources, and power. FPG-AI also uses a fully handcrafted and human-readable hardware description language (HDL) without third-party intellectual properties, thereby enhancing code explicability, reliability, and space qualification. The framework outputs HDL sources of the accelerator rather than the final bitstream, allowing users to maximize the unused portion of the FPGA for supplementary tasks.

The primary objectives of the project are to elevate FPG-AI to technology readiness level 4 (TRL4) and make it accessible to the global space community. These objectives will be achieved through:

1) Extending and consolidating the framework to support a broader range of AI algorithms, including recurrent neural networks (RNNs) and providing hardware characterization for AI-based space applications such as fault detection, isolation, and recovery (FDIR) and telemetry forecasting.
2) Ensuring compatibility with all state-of-the-art devices, with special focus on NanoXplore FPGAs, thereby enabling the utilization of these devices for AI applications and advancing European sovereignty in space technology.
3) Assessing the tool's capabilities through a prototype hardware demonstrator, a move that has already garnered keen interest from industry leaders.

FPG-AI represents a significant leap in AI deployment on satellites, paving the way for advanced space missions, enhanced efficiency, and more comprehensive data analysis. With the support of pioneering organizations and dedicated research, the FPG-AI project is poised to redefine the future of AI in space exploration.

PROJECT NAME: FPG-AI: A technology independent framework for edge AI deployment onboard satellite, and its characterisation on nanoxplore FPGAs
Contract number: 4000141108
START/END DATE: 31/03/2023 – 30/04/2024
KEY THEMES: artificial intelligence (AI), hardware accelerators, space, satellite applications
PARTNERS: European Space Agency (funder), University of Pisa (contractor)
BUDGET: € 100K
activities.esa.int/index.php/4000141108



A CROSS-STACK PLATFORM FOR HPC AND AI WORKLOADS: THE ACROSS PROJECT

Launched in March 2021 and delivered by a consortium of 13 interdisciplinary organizations, the ACROSS project set out to deliver a platform for hybrid high-performance computing (HPC) / artificial intelligence (AI) / big data (BD) workloads on state-of-the-art supercomputers. As the project draws to an end, HiPEAC caught up with ACROSS dissemination manager Alberto Scionti (LINKS Foundation) to reflect upon its achievements.

What were the main objectives of the ACROSS project?

The ACROSS project had the ambitious goals of developing a software stack to support the efficient execution of hybrid workflows on current petascale and pre-exascale machines, and of optimally exploiting heterogeneous computing architectures. Interleaving traditional HPC tasks (such as complex physical simulations) with the training and inference of machine learning models, hybrid workflows are gaining momentum. These workflows often have to deal with large data sets, and often include high-performance data analytics (HPDA). This situation is reflected by the architecture of recent supercomputers, which comprises hardware accelerators that support double-precision floating-point operations (HPC), along with lower-precision arithmetic (machine learning and deep learning).

ACROSS intended to fill the gap between the frameworks and libraries generally used in these domains, which are quite different and often require different setups of the underlying infrastructure. ACROSS also aimed at simplifying the way users express the various steps of their workflows and their specific execution requirements, as well as smoothing the data and execution dependencies between steps using cloud resources. By unifying the orchestration of such different tasks and heterogeneous resources, we have enabled major innovations in sectors like aeronautics, weather and climate simulations, and energy and carbon sequestration simulations. The figure below shows the architecture of the devised software stack.

What are the main results of the project?

ACROSS successfully co-designed and implemented a coherent, modular software stack supporting the definition and execution of hybrid workflows, targeting the use of modern petascale and pre-exascale machines, while being ready to exploit upcoming exascale ones. A major part of the effort was devoted to optimizing algorithms towards low-level frameworks. To achieve such an ambitious goal, many innovations have been introduced at all the levels of the stack. These include:

- **Fast Machine Learning Engine** (FMLE): this has been leveraged to efficiently make use of modern supercomputer nodes to train complex machine learning and deep learning models.
- **The Workflows-aware Advanced Resource Planner** (WARP) software component, which was created to overcome some of the limitations exposed by state-of-the-art batch schedulers. Working in tandem with a dedicated batch scheduler plugin, WARP allows computing resources to be reserved on demand, matching the right number of computing cores, nodes, and duration of the allocation.
- **StreamFlow**: a modular and flexible workflow engine used to facilitate the management of dependencies when executing hybrid tasks.
- **HyperQueue**, a scheduler which allows the management of compute resources with very fine granularity and with more flexibility than that allowed by batch scheduler jobs.

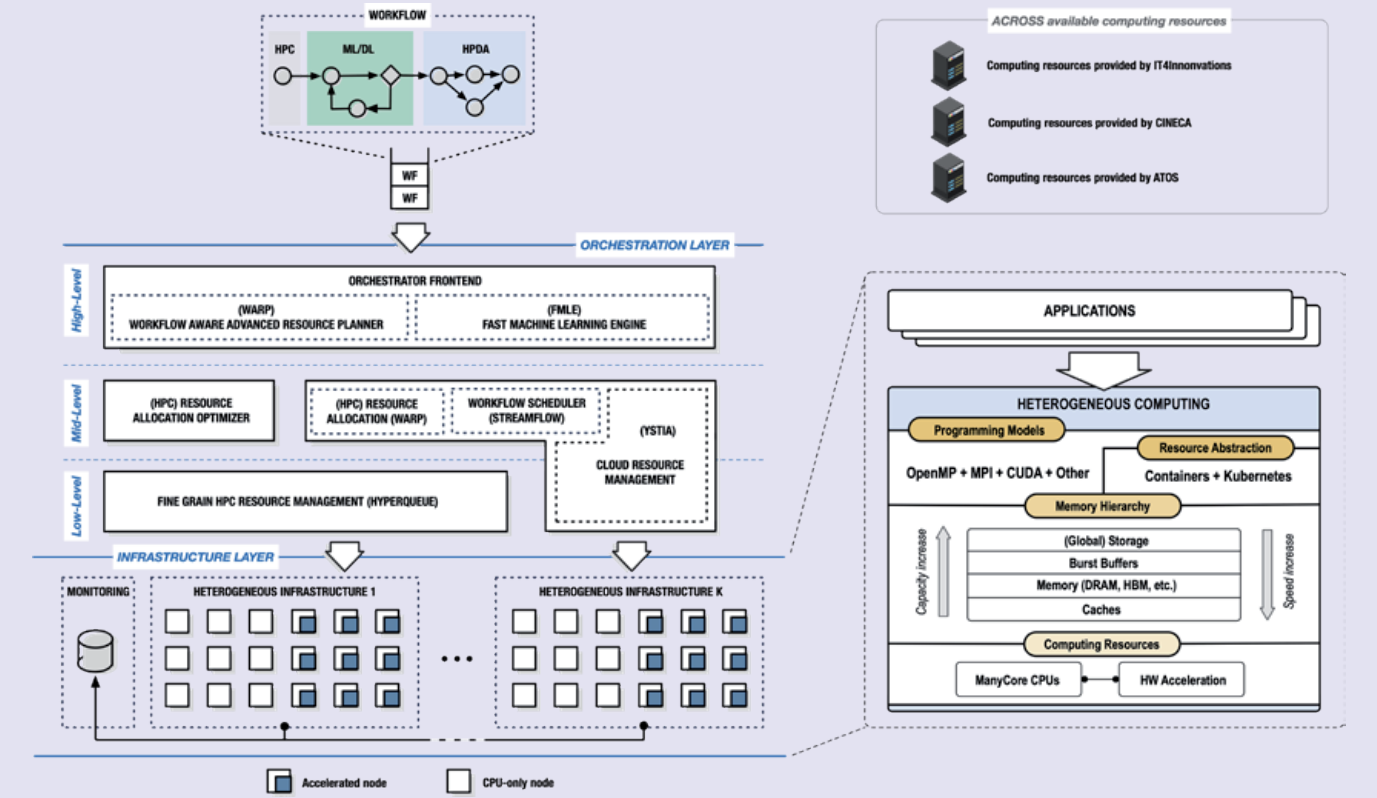
At the lower level of the stack, significant effort was devoted to exploiting libraries and frameworks targeting specific acceleration technologies (graphics processing units (GPUs), field-programmable gate arrays (FPGAs), etc.). This work, even at such a low level of the stack, enabled us to co-design part of the infrastructures we used to run the workflows, as well as to make optimal use of such heterogeneous resources.

Thanks to the appropriate coordination of all these activities, pilot applications saw significant improvements from the beginning of the project. For instance, the aeronautics pilot is now able to quickly run workflows mixing deep learning models training and inference with large computational fluid dynamics (CFD) simulations and a faster HPDA task. The weather and climate simulation pilot can run very large simulations (also by using finer-grained spatial meshes) including ocean and atmospheric complex models, thanks to the efficient management of the large data sets generated. The energy and carbon sequestration pilot can efficiently run simulations by exploiting GPU acceleration.

Last but not least, our software stack enables deterministic execution, thus overcoming one of the limitations of using batch scheduling systems.

What kinds of accelerators did ACROSS use?

The ACROSS project revolved around the idea that using hardware accelerators is the key factor in enabling energy-efficient execution. As such, the project explored the largest possible heterogeneity ranging from different central processing unit (CPU) architectures, to GPUs, to FPGAs. It was also a



The ACROSS software stack architecture

precursor and unique in exploring the neuromorphic paradigm of computation, by co-designing architectures supporting the execution of spiking neural networks (SNNs). In this regard, experimentation with open architectures has also been performed, to see how to scale them on high-end reconfigurable devices and, thus, better support the execution of large AI models. From this viewpoint, we needed to approach the challenge of training SNNs, whose underlying working mechanisms are more complex than those for traditional artificial neural networks (ANNs).

How has ACROSS teamed up with other European projects and initiatives to maximize results?

From the outset of the project, we sought to create synergies with other initiatives. We organized dedicated workshops, where many EU-funded projects were able to present their solutions and technologies, as well as sharing their outcomes. A major focus of the ACROSS project was a broader use of hardware acceleration technologies; as such, the outcomes of the European Processor Initiative (EPI) provided inputs to drive our co-design phase. In this regard, we are aware of the future deployment of EPI processors into supercomputers (and more generally of the growing adoption of RISC-V based architectures), and so we designed the software stack to be as modular and flexible as possible in order to accommodate any future libraries or frameworks designed to take advantage of the EPI processor architecture. We also worked in the direction of opening the use of our solutions to other applications, as explored as pilots in other European funded projects.

What lasting impact would you like to see from this project?

We are seeing a growing interest in mixing processing tasks belonging to very different domains, as this offers more opportunities for facilitating new discoveries or quickly improving complex engineering designs. During the project we became more aware of some limitations of the current systems and some of the acceleration technologies we explored. These limitations will be the target for future work; from this standpoint we think that the ACROSS outcomes will open the door to further investigation and different way of designing and managing exascale supercomputer resources. Besides the significant impact on the technological domain, ACROSS aims to generate a major impact on the pilots' application domains: to this end, we are strongly promoting the ACROSS solutions among the project stakeholders.

ACROSS has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement no. 955648. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Italy, France, Czech Republic, United Kingdom, Greece, Netherlands, Germany and Norway.


Partners in the EXTRACT project (EXTReme dAta Across the Compute conTinuum) are working on a data-driven, open-source platform integrating cloud, edge and HPC technologies for trustworthy, accurate, fair and green data mining workflows. In this article, Daniel Barcelona Pons and Enrique Molina Giménez (Universitat Rovira i Virgili) explain the data pipelines aspect of this project.

Taming a universe of data

How the EXTRACT project is parallelizing data-processing pipelines

The Cloud and Distributed Systems Lab (CLOUDLAB) research group from the Universitat Rovira i Virgili (URV) is a multi-disciplinary team that tackles key research lines of distributed systems. This research group has experience in scalable systems (cloud computing, serverless architectures, distributed storage, peer-to-peer) and web-based infrastructures.

As part of the EU-funded EXTRACT project, in which CLOUDLAB participates, project partners are working together to create enhanced workflows that will process extreme data reliably so that it can be used across a variety of scientific disciplines. Extreme data possesses a set of challenging properties such as high volume and speed, but also variability, that make it very hard to manage effectively.

The project's technology is being validated on two use cases:

- a personalized evaculation route (PER) system to guide citizens through a safe route in real time
- the TASKA (Transient Astrophysics with a Square Kilometre Array Pathfinder) use case, driven by l'Observatoire de Paris

This article focuses on the TASKA use case.

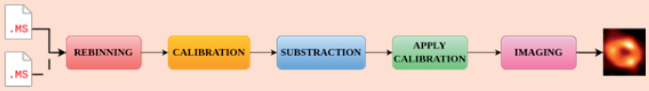
The TASKA data pipeline

Extreme volumes of data (from a few gigabytes to several terabytes) at variable speed are captured by many radio-telescope antennas. This data must be processed to generate high-resolution images of the cosmos that scientists can interpret. To help generate these images, the EXTRACT project is pursuing technical synergies that will help create and improve the workflows used in the TASKA use case by integrating the latest cloud technologies in data-processing parallelization.

The data processing necessary to obtain these images requires several composable steps to prepare and then analyse the data collected by antennas in the MeasurementSet (MS) format established by the Common Astronomy Software Application.

These steps are as follows:

1. rebinning
2. calibration
3. subtraction
4. applying calibration
5. imaging



Example of TASKA pipeline where antennas collect data and generate images

Currently, this pipeline yields poor performance since it is executed manually and monolithically onsite, meaning that it is difficult for scientists to explore these data effectively in a timely manner. Indeed, this process has the potential for many improvements. The CLOUDLAB research group is currently working on two notable improvements to increase the performance of data processing in the TASKA use case: inter-job parallelization and intra-job parallelization.

Serverless technologies create an ideal scenario for executing this pipeline. These technologies can support high scalability while abstracting the underlying infrastructure, which is key to adapting data processing to a variable volume of data and obtain results in real time. CLOUDLAB uses a function-as-a-service approach and its corresponding parallelization (specifically the Lithops cloud framework) to improve the current TASKA pipeline.

Inter-job parallelization and intra-job parallelization

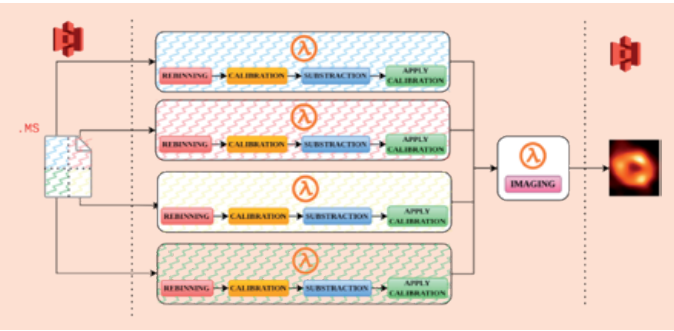
The first improvement that CLOUDLAB is implementing is inter-job parallelization. In this case, 'job' is understood as an instance of this process that takes a set of MS files and generates one of the

desired images. By running several jobs concurrently on different groups of functions, we already parallelize the execution of this process and generate a pipeline for the creation of images, which helps keep the process to a reasonable timeframe after the arrival of data.

The TASKA use case pipeline showing inter-job parallelization

By analysing the internals of the process, it is also possible to apply intra-job parallelization. This means parallelizing the different steps within the generation of the same image. It is possible to split each of the four first steps into multiple 'workers', with each taking a part of the original input dataset MS. Workers will consume partitions of the data to perform processing cooperatively. This parallelizes a large part of the workload within a job and yields further performance improvements. The last step (imaging), however, operates on the aggregate result of the previous steps and requires synchronization.

Performance improves even more when these two types of parallelism are combined. Combining both parallelization strategies results in the rapid processing of multiple datasets and the continuous generation of astronomical images, which can then be analysed by scientists in a timely manner.



The TASKA use case pipeline showing intra-job parallelization

The parallelization of the TASKA workflow will allow it to behave elastically when dealing with the variability of the extreme data that feeds into it. However, there are still challenges to solve in such a complex and demanding task, such as the correct and efficient partitioning of a complex data format like the MeasurementSet.

The EXTRACT consortium will continue towards further improvements on the TASKA use case. Specifically, they will examine a novel way to ingest data efficiently with smart partitioning. The success of this research will contribute to a better data-staging solution for the EXTRACT platform, which will help reduce data-processing latency thanks to more effective resource utilization.

More EU data projects

In addition to participating in EXTRACT, CLOUDLAB coordinates three EU-funded research projects:

- **NEARDATA**: Extreme Near-Data Processing Platform
  This project creates an extreme data infrastructure to mediate dataflows between object-storage and data-analytics platforms across the compute continuum. The NEARDATA platform is a novel technology for the mining of large and dispersed unstructured data sets that can be deployed in the cloud and in the edge (high-performance computing (HPC), internet-of-things (IoT) devices), that leverages advanced artificial intelligence (AI) technologies and offers a novel confidential cybersecurity layer for trusted data computation.
- **CloudSkin**: Adaptive virtualization for AI-enabled Cloud-edge Continuum
  This project aims to design a cognitive cloud continuum platform to fully exploit the available cloud-edge heterogeneous resources, finding the 'sweet spot' between the cloud and the edge, and smartly adapting to changes in application behaviour via AI.
- **CLOUDSTARS**: Cloud Open-Source Research Mobility Network
  CloudStars is a staff exchange programme that allows the mobility and exchange of researchers between academia and industrial institutions in the fields of cloud computing and AI technologies.

FURTHER READING

CLOUDLAB group cloudlab.urv.cat/web

EXTRACT project extract-project.eu

EXTRACT has received funding from the European Union's Horizon Europe programme under grant agreement number 101093110.

A team lead in cybersecurity research at Bosch, Dominik Sisejkovic's stellar computing career has taken him from Croatia to Germany, and from academia to industry. HiPEAC caught up with Dominik to learn what drives him and why Bosch is an excellent place to work.

Career talk: Dominik Sisejkovic

How did you first become interested in computer science and engineering?

Computer science has sparked my interest from the very beginning. Sometime at the age of 14, I first started exploring programming languages, specifically in the object-oriented domain, such as Java, C++, and C#. It was tremendously interesting to see how complex computing worlds could be simplified by modelling the problems in terms of objects – a human-relatable concept in code! Programming my first calculator with a graphical interface was – as funny as it sounds today – very exciting. To be able to create functioning logic and a graphical, interactive system with code still sounds like magic. This motivated me to enroll in the computing (computer engineering) track at the University of Zagreb, Croatia.

Photo credit: David Ausserhofer / Körber Stiftung

Dominik was awarded a German Thesis Award by Körber Stiftung

Can you tell us about your education and work experience so far? How did you end up specializing in cybersecurity?

I started my university education at the Faculty of Electrical Engineering and Computing (known as FER) in Zagreb, Croatia. There, I had the chance to dive deep into all that the computing world can offer. During my studies, I was soon confronted with a very interesting concept: combining nature-inspired computing with optimization tasks. The very idea that, through algorithms, we can mimic natural processes such as genetics or immunity to solve hard optimization problems was the stepping stone for me to join the research world. In fact, I was very lucky; after only a few semesters, I was allowed to collaborate with international researchers on solving optimization problems in the domain of security – a first step that led me to become a researcher in (cyber) security in the years to come.

> ***"At Bosch Research, we are free to design the days around the tasks in a way that suits the work requirements and optimizes for success"***

After finalizing my master studies, I was out and looking for a PhD position in security. My wish was to live and work in Germany and find a location that will allow me to contribute to research with a tangible impact. And again, I was lucky! In 2016, I started as a research assistant at RWTH Aachen University, Germany, focusing on the research and development of a software ecosystem to design and evaluate trustworthy hardware designs. And with this, my research career was kickstarted; I had the opportunity to publish many papers and journals, travel the world, transfer solutions to industry, collaborate with researchers across the globe, work as a chief engineer for four years, write research proposals, and support the German Cyberagency on future research topics – among other things.

Finally, after finalizing my PhD and working as a post-doc in Aachen, I joined Bosch Research in Hildesheim as a team lead, focusing on bringing automation to various tasks in security engineering: a tremendous challenge and opportunity for all companies in the future.

What is a typical day like at Bosch Research? What do you most value about working there?

At Bosch Research, we are free to design the days around the tasks in a way that suits the work requirements and optimizes for success. In general, the task landscape is very broad. We do research and publish papers, develop proofs of concept, design

> *"I consider the HiPEAC community, and everything it has to offer, a crucial component of my success story"*

and submit patent applications, collaborate with university partners around the world, scout for new technologies, supervise students, and enjoy all steps along the way! For me, Bosch Research is a unique place as it combines high-class research under consideration of a business impact across a very large, heterogenous organization and beyond. This ensures that our research efforts always have the potential to be the next breakthrough – not just in science but also in industry. Having access to many business units that deal with real-world challenges is the right fuel for focused, impactful research and innovation.

What role has the HiPEAC community played in shaping you as a professional?

The HiPEAC community has been a great booster in my professional development from the first steps. In fact, it has provided value for my career in three acts (so far). First, the HiPEAC Jobs platform was where I found the PhD position in Aachen, which led me to where I am today. Second, during my time as research scientist, HiPEAC offered me a platform to substantially grow my network through various aspects: workshop organization, contributing to the HiPEAC magazine, and directly starting research collaborations – all components that have strengthened my research ecosystem. Finally, today, as part of industry, HiPEAC allowed me to shape the future by contributing to the HiPEAC Vision and having access to a wide range of experts. In that sense, I consider the HiPEAC community, and everything it has to offer, a crucial component of my success story. Thanks, HiPEAC!

Your doctoral thesis was recognized by a highly competitive German Thesis Award. What were the standout characteristics of your thesis? What advice would you give students who want to achieve excellence in their theses?

The German Thesis Award is a yearly award provided by the Körber-Stiftung to the best doctoral graduates in Germany from all disciplines. The main evaluation criterion, besides academic excellence, is the broader social relevance of a particular piece of research. Thus, the award encourages young scientists to highlight the value of their research to society.

Bosch Research is based in Hildesheim, Germany

With this context, the challenges in my doctoral thesis addressed the following question: how can we ensure trustworthiness in microelectronics? Over decades, both science and industry have developed countless security mechanisms to protect us from various attacks in the digital world; however, these have mostly been in software. Unfortunately, it has become clear that a tiny, delicate, malicious change in the hardware – during its design or fabrication – can break all the security mechanisms on top and provide back doors to controlled attacks. Having such a maliciously changed hardware component in the context of the telecommunication infrastructure, automotive industry, medical or military systems can lead to disastrous and far-reaching consequences.

Thus, my thesis focused on developing a holistic software framework with all necessary components to develop, evaluate, and apply protection schemes – concepts that were successfully transferred to industry. In that sense, I would urge students to derive research concepts for challenges followed by a simple question: how will my work benefit the society as a whole? If research ideas can lead to real innovation and have a tangible and beneficial impact on all of us, you are indeed on the right path!

FURTHER READING

German Thesis Award winners 2023
https://bit.ly/German_Thesis_Award_winners_2023

A PhD student at Babeș-Bolyai University in Cluj-Napoca, Romania, Laura Diana Cernău also has experience of working in an industry setting and is actively involved in helping other early career researchers in entrepreneurial activities. HiPEAC met Laura at womENcourage 2023 and asked her about academia, industry and incubating innovation.

'Innovation activities contribute significantly to students' personal growth'

Hi Laura! Tell us a bit about yourself.

Hello! I'm currently in my third year as a PhD student at Babeș-Bolyai University in Cluj-Napoca, Romania. My research focuses on software metrics, specifically their application in predicting the likelihood of code defects. This work has the potential to establish valuable connections between various software metrics and the overall quality of software systems.

In addition to my academic pursuits, I'm an active member of the Faculty of Mathematics and Informatics' start-up incubator, CS InnoHub. My work consists of scouting and mentoring student start-up teams as they embark on their entrepreneurial journeys. In addition to this, I've been navigating the world of software engineering for over six years now, mostly developing functionalities using backend technologies, particularly Laravel and Kotlin. I have pair programming sessions with my colleagues, work on functionality discovery when I am the technical owner of a subject, and address customer issues. My involvement also consists of mentoring a handful of my colleagues, helping them to pave their career paths and to work on self-improvement.

So what would you say are the main differences between academia and industry?

One of the notable distinctions lies in the approach to work styles. On one hand, much of a developer's time in the industry is spent translating requirements from diverse stakeholders into practical functionalities. The mindset in the corporate setting revolves around delivering value to the end users of the products you and your team created. On the other hand, the academic arena has its own unique work methodology. Here, emphasis is placed on the theoretical aspects, and theoretical concepts are demonstrated through experiments. Academic work demands a distinct level of rigour when it comes to work methods.

Laura at the EIT Digital Master School kickoff event at Babeș-Bolyai University in Cluj-Napoca

Laura is based in Cluj-Napoca, Romania

Why is it important for researchers – and particularly students – to get involved in innovation? What do you do to encourage this?

As previously mentioned, I play an active role in our university's Faculty of Mathematics and Informatics innovation incubator. My responsibilities within the incubator primarily revolve around identifying and guiding student teams with startup concepts. I'm engaged in a wide range of activities, from assisting in the coordination of events like hackathons and idea jams to managing the incubator's presence on social media platforms.

Participation in innovation initiatives is of essential importance for students, as it contributes significantly to their personal growth. It helps them nurture essential soft skills and provides invaluable insights into the journey of transforming an idea into a product that serves customers. Even if their ideas remain at the experimental stage, they will still carry with them the knowledge acquired and a strong work ethic that will benefit them regardless of their career trajectory.

OK, so we're ready to start our innovation adventure. What skills do people need if they are thinking of becoming an entrepreneur? What advice would you give students who are thinking of starting out on this path?

Effective communication skills are a must for students considering this path. In the context of entrepreneurship, building a solid network is a key factor. This network can only be cultivated through active participation in events and socializing with peers from similar backgrounds, potential investors, or prospective clients.

The ability to conduct thorough research is another critical skill. In terms of product development, market research plays a vital role in determining the success of a product. Apart from that, a risk-taking mindset and the ability to treat failures as valuable learning experiences are essential.

What's the technology / innovation ecosystem like in your local area? How does the environment promote entreprise?

I'll be talking mainly about Cluj-Napoca, a prominent city in Romania. The status of Cluj-Napoca as a thriving university town creates the perfect backdrop for innovation to flourish. The university environment promotes innovation initiatives, including our faculty's incubator, and numerous student organizations that host events, both locally and nationally.

Furthermore, the local government plays a pivotal role in promoting innovation and the information technology (IT) sector by endorsing a variety of events, innovation competitions, and the establishment of collaborative workspaces. It's safe to say that Cluj-Napoca boasts a robust innovation ecosystem bolstered by the synergy between universities, local government, and businesses.

While immensely satisfying, doing a PhD can be a daunting prospect, and maintaining motivation over several years of study is a challenge. HiPEAC caught up with Perry Gibson, now a postdoctoral researcher in the Glasgow Intelligent Computing Laboratory (gicLAB) at the University of Glasgow, to learn about his PhD experience.

Destination PhD

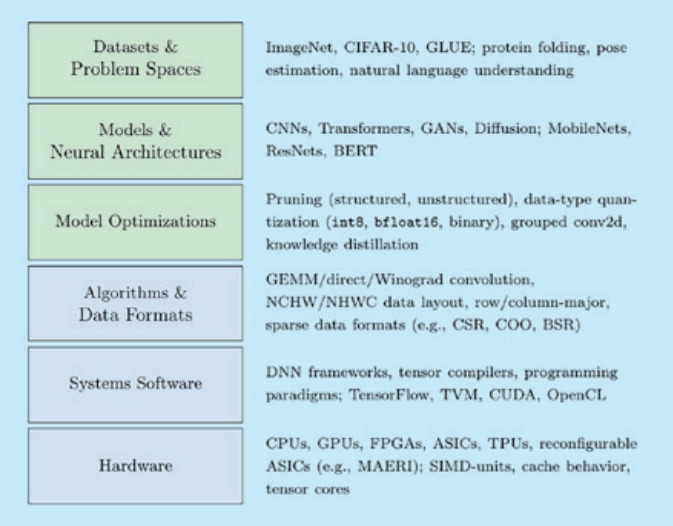
One researcher's doctoral journey

'When you're at high school, it feels like there's this pressure to decide what you want to do for the rest of your life. I didn't know what I wanted to do forever, so I decided to go down a route where I wouldn't get boxed into a corner,' says Perry Gibson. 'The informatics course at the University of Edinburgh attracted me as it offered a lot of scope: you could end up being as close or as far away to working with people as you wanted, and it offered courses ranging from theoretical and "mathy" to more flexible and artistic.'

After completing his integrated master's at Edinburgh, Perry moved back to his native Glasgow where he began studying a PhD under José Cano, focusing on optimizing machine learning. This choice was guided by two main considerations, he recalls. 'First, I liked the fact that it was driven by metrics, that you could see clearly what was improving, such as making something faster or using fewer resources. The goals are clear and achievable, and you can tackle the problem in lots of different ways. Second, I had read into the history of artificial intelligence (AI) and knew there had been "AI winters" in the past where research funding had dried up and companies went bust. So I wanted to get involved in the "hot topic" of AI, but with one foot in systems optimization to keep my options open.'

Perry's thesis focused on characterizing the systems stack – from hardware all the way up to machine-learning models – from a machine-learning perspective, in order to identify cross-stack interactions opportunities for deep neural networks (DNNs). 'A given acceleration technique could have unexpected effects in another layer of the stack, or require additional techniques from other layers to fully realize its potential. Studying these interactions for DNNs has not been sufficiently explored before,' he says. This observation was used to shape the three main challenges for his thesis:

- identifying unrealized gains (i.e. pinpointing the source of a performance regression in a deployed solution)
- identifying and exploiting cross-stack interactions
- efficient design-space exploration



Overview of DLAS, spilt between machine learning and systems techniques, with examples

His research led him to conclude that, to help manage complexity, the compiler should be the central element to address these challenges and that infrastructure needs to be composable. The platforms tested spanned the spectrum from small edge to server class, from Raspberry Pis and NVIDIA Jetson Nanos to Arm-based processors to larger-scale x86 machines, as well as more powerful NVIDIA graphics processing units (GPUs) and field-programmable gate arrays (FPGAs).

The resulting conceptual framework of deep learning across the stack, says Perry, really comes into play for the holistic viewpoint it provides. 'While people are experts in their own area, sometimes there are missed opportunities for optimization overall. For example, the systems community may be evaluating outdated machine-learning models,' says Perry. 'So there is a definite need for two-way collaboration.' As well as publishing results in peer-reviewed conferences, Perry's work has also been included in an open-source code database, meaning that it is used in actual deployments.

Writing the thesis

Before embarking on writing his thesis, Perry spoke to several people who had already obtained their PhD, and found that it pays to be selective about the advice you take. 'I found it helpful to focus on the actionable things – for example, that the thesis needs to be structured. So I researched winners of the Scottish award for the best PhD (SICSA), then I analysed how they had structured their theses and made a template based on this,' says Perry. 'At the time of starting my PhD I had already spent a few years writing, so I did an audit of all the text I had already written to see what I could reuse. Once I had the template, I could slot in text I'd already written, before filling in the gaps. My strategy was to go wide first before going deep.'

Tools, such as Perry's own thesis-o-meter, helped the writing process. 'Sometimes you get this frustrating feeling of getting stuck. The thesis-o-meter is a line graph showing what you've achieved in terms of word count. In the early days, this could be very motivating, as it showed the work starting to take shape.' Another tool he recommends is one to automatically check the bibliography, while he also used a large language model to proofread the text.

PhD defence

Once the PhD thesis was submitted, it was time for the viva, where the thesis would be defended. 'In the UK, this is presented to a panel of two people – one internal from your university, one external from another university. The PhD candidate gives a 15-minute presentation, before the panel goes through the thesis almost page by page, asking clarifying questions.' While this may sound nerve-wracking to the uninitiated, the advantage, says Perry, is that 'it's almost like an exam where you get to set the questions', In any case, simply doing the PhD was its own reward, he says: 'The feeling I had when I submitted my thesis was that, even if I didn't pass, I was very proud of how far I'd come.'

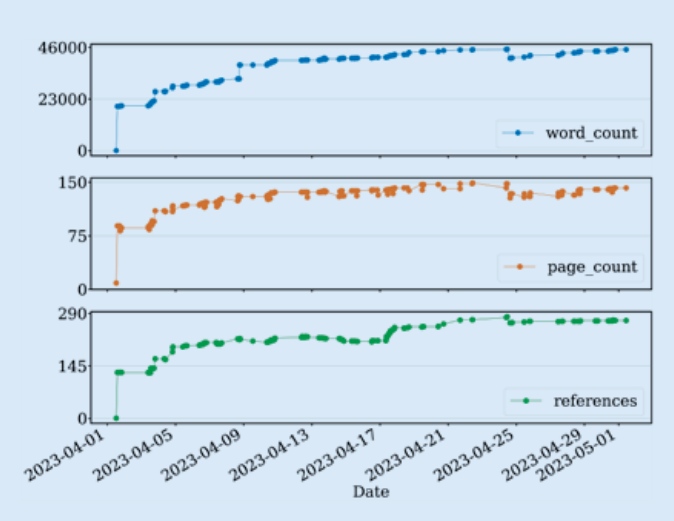
As for next steps, Perry says he's interested in taking the work done on AI compilers further, although whether in an academic or industry setting remains to be seen.

Thinking about a PhD? Read this first

So what advice would Perry give prospective PhD students? 'I think a lot of it is about getting the right balance between explore and exploit: figuring out what's going on in an area and where this can be exploited,' he says. 'Your PhD is one of the few points in your career where you actually have control over this explore / exploit ratio.'

Another important piece of advice, he says, is to choose your supervisor wisely. 'I was fortunate enough to have a highly dedicated supervisor,' he says. 'Others I knew weren't so lucky: some supervisors were very unsupportive, and some people even quit as a result. You don't need a good supervisor to succeed, but it's important that you can communicate well and find a way to deal with disagreements. Talking to their previous students can help you decide – and bear in mind that an early career academic might have more motivation to invest in you than a well-established name.'

As a final thought, Perry highly recommends getting involved in events where you can speak to your peers. 'Events like the HiPEAC conference are great for speaking to people working on related topics. These events help you realize that you're not a fraud because you can speak to experts, but you also meet a lot of people that know more than you, so it helps you to "calibrate" yourself,' he concludes.



Perry's 'thesis-o-meter' helped motivate him

Perry with his supervisor José Cano at HiPEAC 2023

Thanks to its enduring popularity, the Young People Programme is back for the 2024 edition of the DATE (Design, Automation and Test in Europe), and HiPEAC is once again participating. Young People Programme coordinator Marina Saryan (Synopsys) tells us more.

The DATE Young People Programme is back

The largest electronic design automation (EDA) conference in Europe, DATE will be held on 25-27 March 2024 in Valencia, Spain. It includes the Young People Programme (YPP), which supports PhD and master's students in their career development, with several initiatives.

Sponsorship of attendance

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

Industry Careers Fair

Participating companies will post open positions on the HiPEAC Jobs portal for students to apply. At the conference, companies and students will participate in a speed-dating recruitment event. Registered students will also have the opportunity to ask recruiters questions before the conference in a 'Meet the recruiter' call.

A keynote by Matt Venn, a science and technology communicator, will discuss job opportunities created by the open hardware movement. The event will also feature a panel of young industry experts, who will talk about career paths in microelectronics and related industries.

GPT Design Contest

This new initiative is a contest in which student teams work for 24 hours to create a design using GPT Tools. The contest is organized by Cadence, Arm, TU Munich, the University of Southampton and UNSW (Australia), who will provide the specifications of the design, plus a verification intellectual property (IP) to check the protocol compliance of the generated design. Training material will be provided to students prior to the conference.

Student Teams Fair

The Student Teams Fair brings together university student teams with EDA and microelectronics companies. Student teams will present their activities, success stories and challenges, while companies can provide support (such as free tool licences, personalized webinars and financial support) for future activities. Before the conference, student teams will be updated on what to expect from DATE 2024 and how to identify relevant events, to get the most from this experience.

PhD Forum

The DATE PhD Forum is a great opportunity for PhD students to present their work to a broad audience in the system design and design automation community, as well as to make contacts. For their part, representatives from industry and academia get an insight into state-of-the-art research in the system design and design automation space.

Academic Careers Fair

It's not just industry representatives who are looking for top-talent graduates; academic institutions also need high-potential candidates. Academic programme leaders will promote their research projects and open positions to YPP attendees, and this year's fair will also feature a workshop on how to write proposals, a skill researchers need throughout their academic career.

University Fair

The DATE University Fair (previously known as the 'University Booth') provides a platform to disseminate mature projects, ideally with a live demonstration. In addition to academics, this is also of interest to industry, as the outcomes of fundamental research can be potentially applied to commercial products. As there are often follow-up projects, which need a new generation of researchers, this closes the loop with the Academic Careers Fair.

FURTHER INFORMATION:

DATE Conference Young People Programme
date-conference.com/young-people-programme
For further details about individual events, please contact:
Careers Fair - Industry and Student Teams Fair:
ypp-industry@date-conference.com
PhD Forum: ypp-phd@date-conference.com
Career Fairs - Academic and University Forum:
ypp-academia@date-conference.com
HiPEAC Jobs portal for DATE conference
hipeac.net/jobfairs/date24
HiPEAC Jobs portal hipeac.net/jobs

Pablo Antonio Martínez tells us about his doctoral studies tackle performance, portability and productivity issues for accelerators.

Three-minute thesis

NAME: Pablo Antonio Martínez
RESEARCH CENTRE: University of Murcia
SUPERVISORS: José Manuel García and Gregorio Bernabé
THESIS TITLE: Improving the performance, portability, and productivity of hardware accelerators

Improving microprocessor performance is becoming increasingly complex, while applications like artificial intelligence (AI) are demanding more and more computational power. In recent years, we have witnessed a paradigm change: rather than using the central processing unit (CPU) for everything, computers are evolving into more heterogeneous organizations, in which multiple specialized chips compute specific workloads. These specialized chips are usually called accelerators. Thanks to their specialization, accelerators are significantly more efficient than CPUs in terms of performance and / or energy consumption.

However, accelerators also come with notable challenges to the programming workflow. In environments with multiple accelerators, writing code for each is very inefficient, since each accelerator is programmed with different languages. Performance is also concerning because programming languages often struggle to exploit hardware to exploit its full potential. Lastly, portability is also complicated, since when a program is designed for a specific accelerator, it cannot run in a different one. Achieving programming languages that provide productivity, performance, and portability is known as the P^3 problem.

Enhancing performance and programmability

The first contribution of this thesis is a thorough study of existing languages for programming multiple accelerators with a single-source code, such as PHAST and oneAPI. We found that they are lacking in some desirable areas, especially productivity and performance. Motivated by this fact, this thesis proposes HDNN, a new domain-specific language based on the MLIR infrastructure for programming accelerators.

Another issue is how to run code that is already written on accelerators because rewriting the code for each accelerator is extremely costly. Thanks to a collaboration with the University of Edinburgh, funded by a HiPEAC collaboration grant, we propose ATC: a compiler solution that allows you to run the compute-intensive parts of a program written in plain C/C++ on an accelerator automatically. ATC uses program synthesis and novel compilation techniques to map regions of code to vendor libraries backed up by accelerators.

Furthermore, a notable challenge is how to use multiple accelerators at the same time in accelerator-rich environments like system-on-chips (SoCs). The idea of using multiple accelerators concurrently is often referred to as accelerator-level parallelism (ALP). In this thesis, we show a new proposal for exploiting ALP in heterogeneous environments. We present a framework capable of orchestrating multiple accelerators to run a single task jointly, significantly improving performance.

Practical applications

We expect that the proposal described in this thesis will help improve the usability and the performance of accelerators, which will establish the standard for future-generation computing systems. In particular, HDNN can significantly reduce programming effort while improving accelerator performance. On the other hand, a compiler that can replace handwritten code with accelerator code can enable the execution on these devices, drastically improving the efficiency of millions of programs. Lastly, exploiting multiple accelerators simultaneously can enable next-generation computing capabilities in terms of performance and energy efficiency.

Pablo's supervisor **José Manuel García** commented: 'Modern SoC designs incorporate multiple accelerators within the chip. Pablo has taken on the challenge of tackling the performance, portability, and productivity issues at different levels. This includes using programming languages (single-source and a new MLIR dialect), replacing existing code with application programming interface (API) calls, and exploiting multiple accelerators concurrently. We are optimistic that these developments will enhance the usability and performance of heterogeneous computing, and will establish it as the standard for future computing systems.'

HiPEAC

Thanks to all our sponsors for making #HiPEAC24 such a success!

AMD
HUAWEI
FRONT GRADE Gaisler
arm
ASML
AXELERA ARTIFICIAL INTELLIGENCE
CLOUDIAN
Collins Aerospace
doitnow
E4 COMPUTER ENGINEERING
Google
DeepMind
Hewlett Packard Enterprise
MAXELER a groq company
nexter A COMPANY OF KNDS
redpitaya
BOSCH Invented for life
SCALETORCH
ZeroPoint
SEA Projects
ECLIPSE FOUNDATION
ECSPEC
hoCOMPUTER compiler & tools
Qualcomm
Springer
STREAM HPC
TASKING
THALES
REGALE
VEDLIoT Very Efficient Deep Learning in IoT
EUPEX European Pilot for Exascale

Sponsors correct at time of going to print. For the full list, see hipeac.net/2024/munich

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

hipeac.net

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.