HiPEAC info 66

JUNE 2022

HiPEAC Conference Budapest 2022

Open source: A strategic move for Europe

Shifting towards the edge in European computing

Bianca Schroeder on smarter storage

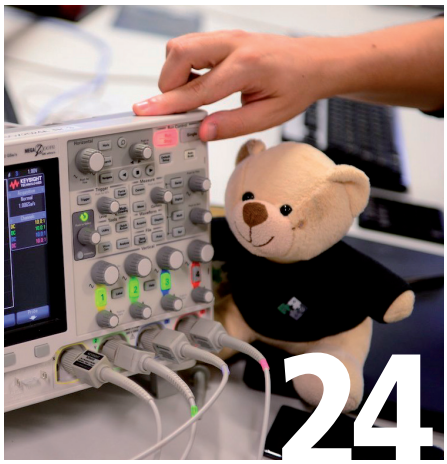
Hai Li on machine learning across the stack

contents

Max Lemke on the shift to the edge

Hai Li on machine learning

Bianca Schroeder's smart storage



A year ago, very few people could predict the situation we are finding ourselves in today. Most of us were hoping that 2022 would bring normality back to our lives. Little did we know that the news about the pandemic would be dwarfed by the news about a war near the border of the European Union. This war is a socially created disaster, with a potentially bigger impact than the pandemic on both society and the economy. The price of energy and of raw materials is increasing rapidly without a clear perspective for improvement in the short term. Many people will have to cut spending, and some will end up in poverty, especially if the job market crashes too. In some parts of the world, it might even lead to famine. This is definitely not the new post-COVID-19 normal we were hoping for.

Looking at the war itself, digital technology is playing a clear strategic role. The Ukrainian president is virtually present around the world. On the battlefield, smart weapons are making a difference; in the air, satellites are following the actions on the ground, and much of this information is also available to the public. Ukrainian citizens are active users of social media channels which are difficult to shut down. They are fairly well informed about the movements of the armies and can help each other. Ukrainian children who are temporarily living abroad can stay in touch with their family, friends and teachers through the internet. Every citizen with a smartphone can make videos and share them with the rest of the world, making this possibly the best documented war in the history of humanity. When Russian soldiers left the Kyiv region, within hours the world could witness what they left behind. Post-war tribunals will have access to plenty of undestroyable digital evidence, which is quite different from the wars of the past.

This is the conference issue of the HiPEAC magazine. The Budapest conference is the first in-person HiPEAC conference for over two years, and I am looking forward to the event with its rich programme. I also wholeheartedly hope that this will be the last event in this decade that we have to postpone in order to hold it in person. HiPEAC is a network, and in-person meetings are essential to create value for our members.

Koen De Bosschere, HiPEAC coordinator

HiPEAC is the European network on high performance embedded architecture and compilation.

hipeac.net

@hipeac hipeac.net/linkedin

HiPEAC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 871174.

Cover image: © stockshoppe
Design: www.magelaan.be
Editor: Madeleine Gray
Email: communication@hipeac.net

Along the compute continuum, we are witnessing a shift from centralized computing facilities to computing at the edge. In this policy article, Dr Max Lemke, head of the Internet of Things Unit at the European Commission, explains what this shift means for Europe.

From HPC/cloud to edge/IoT

A major paradigm shift for Europe

Today, 80% of the processing and analysis of data takes place in data centres and centralized computing facilities, and 20% in smart connected objects. This ratio is expected to reverse over the next five years, inevitably leading to a paradigm shift in where data is processed. Edge computing refers to processing which takes place closer to or even within an internet-of-things (IoT) device itself, that is at the 'edge' or periphery of the network.

At the European Commission's 2021 Fireside Chat on Next Generation IoT and Edge Computing, experts agreed that *'edge computing is the logical evolution of the cloud computing model, avoiding the transfer of mission-critical data in the cloud, supporting resilience, real-time operations, security, and privacy protection while at the same time reducing energy consumption and carbon footprint'*. It is expected that in the middle of this decade most of the data processing and analytics will take place where they are the most efficient, which is generally close to where the data is generated: at the edge of the network.

European companies have strong expertise and market shares in industrial and business applications, industrial IoT and 5G across many sectors, including mobility, energy, the home, agriculture, manufacturing, logistics, etc. By capitalizing on these unique, largely sectoral competences and by driving leadership in the field of edge computing and the IoT, Europe has a one-time opportunity to regain a significant role in the computing market by 2025 and, by doing so, make further progress towards digital autonomy. Leveraging a local, distributed computing infrastructure, this move to the edge and the IoT facilitates the creation of new services and business models rooted in verticals around the applications, rather than today's more general-purpose cloud business models.

This shift towards edge computing will lead to strong changes in the computing landscape and innovative use scenarios across our economy:

- **High performance at the edge and in devices:** Reinforcing current trends, high-performance processing will become a commodity at the edge and in the IoT. A good example comes from the mobility sector, which is increasingly electric, connected, autonomous and servitized. With the share of electronics in the total cost of a car rising, terms like 'software-defined vehicle' or 'smartphone on wheels' are emerging to define **next-generation intelligent vehicles**. At the core of such new vehicles will be a few central computer systems powered by high-performance chips rather than the 50 or more electronic control units – each with their own software and hardware – we see today.

- Strong computing capacity at device level in the emerging smart IoT will enable new concepts of **decentralized intelligence and swarm computing**. Whereas in the past we programmed each device in the IoT individually, in future, AI-enabled software engineering tools will support a functional approach or holistic orchestration of swarms of homogeneous or even heterogeneous devices. Such concepts have a high potential for example in the energy sector, powering the next generation of smart grids by **optimizing locally renewable energy supply and demand** by households, buildings and electric vehicle charging.

- AI-based **cognitive cloud** frameworks will integrate diverse computing and data environments seamlessly and securely across the **computing continuum**, spanning from HPC to core cloud to edge to device level. They will support automatic management of the computing continuum and cater for dynamic load balancing and optimized energy efficiency of computing resources and data traffic. An interesting use case is **high-precision farming**, with processing taking place on equipment (dispensing water, fertilizer and pesticides), at the edge (data from local farms) and in the cloud (high-level services) .

EU-level support

The European Union (EU) is supporting the evolution of computing with actions along several lines:

- **Chips Act:** In February 2022, the European Commission proposed a comprehensive set of measures to ensure the EU's security of supply, resilience and technological leadership in semiconductor technologies and applications. As the first of three pillars, the 'Chips for Europe Initiative' will pool resources through the enhanced 'Chips Joint Undertaking' resulting from the strategic reorientation of the existing Key Digital Technologies Joint Undertaking. A major target of the European Chips Act is the next generation of computing chips, including quantum chips.
- **Data legislation**: The European Commission is putting forward the legislative framework for a prospering data economy. For example, the Data Act, which was proposed by the European Commission to the Council and Parliament in March 2022, will make more data available for use and will set up rules on who can use and access what data for which purposes across all economic sectors in the EU.
- **Data and cloud strategy**: The European strategy for data aims at creating a single market for data that will ensure Europe's global competitiveness and data sovereignty. Common European data spaces will ensure that more data becomes available for use in the economy and in society, while keeping the companies and individuals who generate the data in control. The Commission also aims to enable access to secure, sustainable, and interoperable cloud infrastructures and services for European businesses. Through the DIGITAL programme it is co-investing in European dataspaces for sectors like mobility, manufacturing, energy, agriculture, health; and in cloud-to-edge services, cloud federation, and marketplaces.
- **Research and Innovation Programmes:** In Horizon Europe, Europe is supporting research and innovation on computing technologies 'at large'. Major initiatives include:
  - Under Cluster 4 ('Digital, Industry, Space') more than a quarter of a billion euro is planned to be spent between 2021 and 2024 on a targeted **research and innovation initiative on 'Cloud-Edge-IoT for European Data'** addressing the next generation of meta-operating systems for the IoT, environments for decentralized intelligence, and on cognitive cloud frameworks.
  - **Joint Undertaking on 'Key Digital Technologies' (KDT)**: The Strategic Research and Innovation Agenda for Electronics Components and Systems dedicates a chapter to 'Edge Computing and embedded Artificial Intelligence' addressing challenges like energy efficiency, system complexity, and sustainability. Calls in KDT are normally conducted once a year based on joint funding from the EU, participating states, and industry.
  - The **European High Performance Computing Joint Undertaking (EuroHPC JU)** is a joint initiative to develop a world-class supercomputing ecosystem in Europe. It aims at reaching exascale capabilities in Europe in the next two years. Among other things, it supports the development and uptake of innovative, energy-efficient and competitive supercomputing technologies and applications based on a supply chain that will reduce Europe's dependency on foreign computing technology
  - In many areas across the DIGITAL and Horizon Europe programmes, applications capitalizing on advances in computing are supported. These include applications in the manufacturing and energy-intensive industries sectors (Cluster 4), in the mobility and energy sectors (cluster 5), and in the agriculture and food sector (Cluster 6).

In developing the research programmes, the European Commission draws upon resources like the HiPEAC Vision, which provides an excellent complement to the strategic research agendas of European industrial associations.

Precision agriculture offers opportunities for automatic management of compute resources across the continuum

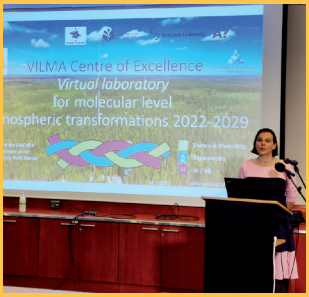
European Commissioner Thierry Breton announcing the Chips Act

CSW Tampere celebrates the best of the IoT

Attracting around 140 registered participants, HiPEAC finally managed to hold Computing Systems Week in Tampere, Finland after twice being postponed due to the COVID-19 pandemic. With a varied programme focusing on the internet of things (IoT), the event included keynote talks by Laura Ruotsalainen (University of Helsinki), on artificial intelligence (AI) and society, Teppo Hemiä (Wirepas) on non-cellular 5G, and Ari Kulmala (Tampere University) on industrial acceptance of open hardware.

Technical sessions included an exploration of system-on-chip technologies for the automotive sector, acceleration for the edge and IoT, machine learning across the compute continuum, and disruptive RISC-V-based architectures. Meanwhile, the European Digital Innovation Hub session on the opening day featured a roundtable of discussions on how to collaborate, while the HiPEAC Student Challenge and Inspiring Futures careers session offered activities for students.

Sessions from the event are available to view on HiPEAC's YouTube channel, HiPEAC TV. On behalf of HiPEAC, many thanks to Jari Nurmi and the local team for their efficient organization and warm welcome.



FURTHER INFORMATION:
CSW Tampere playlist on HiPEAC TV
bit.ly/CSWTampere_HiPEACTV_playlist

Intel oneAPI Center of Excellence established at TU Darmstadt

Andreas Koch (left) and Leonardo Solis-Vasquez

The Embedded Systems and Applications Group at the Technical University of Darmstadt (TU Darmstadt), led by HiPEAC member Andreas Koch, has been selected to establish an Intel oneAPI Center of Excellence. The centre's objective is to accelerate data parallel computing and simulation software used in medical and pharmaceutical research powered by oneAPI open cross-architecture programming.

Through the oneAPI centre, and in collaboration with Intel, the research group will port an accelerated version of the AutoDock application to create a single codebase that can be efficiently optimized and tuned for multiple hardware architecture targets.

AutoDock is widely used for simulating molecular interactions at close distances, aiming to predict the best 'fit' of two molecules to each other from a biophysical standpoint. These 'docking' results are an important initial step for the discovery of new drugs, as the computations can be performed much more quickly than experiments using traditional 'wet lab' chemistry. However, the simulation of these docking processes is computationally expensive.

To address this challenge, as described in *HiPEACinfo 66*, Leonardo Solis-Vasquez of the Embedded Systems and Applications Group is working on an accelerated version of AutoDock, named AutoDock-GPU. AutDock-GPU speeds up these simulations by parallel execution on different processors, including multicore central processing units (CPUs), graphics processing units (GPUs) and reconfigurable compute units.

Together with Intel experts, TU Darmstadt is now working on a next-generation parallel implementation of AutoDock-GPU, which will also leverage oneAPI. This enables easier integration of faster and higher-quality simulation algorithms, and also has the potential to be scaled up for execution on the upcoming Aurora supercomputer at Argonne National Laboratory in Illinois (United States), providing a performance of two exaFLOPS. The combination of an improved AutoDock-GPU with significant compute power will help address current and future challenges in medical and pharmaceutical research much faster than is possible today.

FURTHER INFORMATION:
bit.ly/TUDarmstadt_oneAPI

Üdvözöljük Budapesten – Welcome to Budapest!

Replete with iconic classical architecture set alongside the River Danube, it is easy to see how the Hungarian capital earned its UNESCO World Heritage Site title. In addition to its rich history, Budapest is also a scientific and technological hub – fitting for the city which produced mathematician John von Neumann. HiPEAC 2022 local host Péter Szántó (Budapest University of Technology and Economics) tells us more.

What makes Budapest a good location for the HiPEAC conference?

A vibrant city, Budapest is the capital of Hungary, and its political, cultural, commercial, industrial, and transport centre. It is easily accessible from almost every major European city. Budapest is the centre of research and development activities in Hungary, hosting the largest universities in the country, such as Budapest University of Technology and Economics (BME) and Eötvös Loránd University (ELTE), and offices of international development companies.

The two parts of the city, Buda and Pest, separated by the Danube, have different identities. In the narrow, winding streets of the Buda Castle district you'll find a wealth of architectural sights, cafés, pastry shops, and restaurants, while from the Citadel on Gellért Hill you can enjoy magnificent views of the city. The city centre on the Pest side offers a wealth of attractions within walking distance of each other, including the finest restaurants. At night, Pest's famous party district has a lot to offer, especially the Ruin Bars, a Budapest specialty. Budapest is also famous for its thermal baths.

What is the computing systems ecosystem like in Budapest, and in Hungary more generally?

The largest high-performance computing (HPC) infrastructure for academic research in Hungary is maintained by the HPC Competence Centre (HPC CC), in close cooperation with universities. Supercomputing clusters are located in four cities: at the HPC CC site in Budapest, and at the universities of Debrecen, Miskolc and Szeged. A new cluster will start operating in 2022; offering a performance of five petaflops, it will be ranked within the top 100 of the HPC TOP500 list.

The ELKH Cloud, a more heterogeneous and artificial intelligence (AI)-centric system, is operated by the Institute for Computer Science and Control (SZTAKI) and the Wigner Research Centre for Physics. There are also numerous other smaller clusters; for example, BME operates a local HPC system and a cloud infrastructure to support education.

In terms of industrial users, the purpose of the OMSZ supercomputer is meteorological forecasting, while the system developed in collaboration between OTP Bank and the Ministry of Innovation and Technology is used for AI-based language processing. On the industrial side, several SMEs and large companies are also represented, such as Arm, Ericsson, Nokia, Huawei, Flextronics, AI Motive and 4iG.

What are the 'must-dos' while we're in Budapest for HiPEAC 2022?

Walk in the Buda Castle area, and visit The Royal Palace of Buda, the Fisherman's Bastion, and Matthias Church. On Pest side, the Parliament is a good starting point; from there you can take a walk to St Stephen's Basilica and the Opera House. The Széchenyi Thermal Bath, close to Vajdahunyad Castle and Heroes' Square, is a great place for relaxation. Try Hungarian traditional dishes, such as goulash soup, halászlé (fisherman's soup), hurka and other traditional sausages, mangalica steak (made from a special breed of pig), Hortobágyi (savoury meat-stuffed pancakes), and desserts like Somlói galuska or Eszterházy cake. Be sure to taste quality Hungarian wine with your meal, for example from the Villány region, Tokaj, or another of the 20 official wine regions of Hungary.

Max innovation: TETRAMAX celebrates successes

Launched in 2017, the TETRAMAX Innovation Action sought to promote technology transfer in customized, low-energy computing solutions across Europe. Now that the project has come to an end, we caught up with TETRAMAX Coordinator Rainer Leupers to find out how TETRAMAX has helped bring products to the market, energizing the low-power electronics sector with the latest machine learning and edge computing innovations.

What prompted you to create TETRAMAX?

In my dual role as professor and entrepreneur, I get to observe the gap between academia and industry on an almost daily basis. When moving from shareholder meetings to project consortia events, it feels like switching between languages. The academic vocabulary emphasizes publications, h-index, and somewhat vague research buzzwords. In industry language, products, timelines, and financials are predominant.

TETRAMAX was conceived as a sort of interpreter that provides both sides with resources to better understand each other and that adds a missing link: focused and co-funded technology transfer experiments (TTX) as a low-risk testbed for bringing specific academic results into practice. We had a proof-of-concept in place with the TETRACOM project, strong encouragement from the European Commission (EC), a diverse and capable project team, and HiPEAC as a powerful networking infrastructure in the background. With this unique setup, the creation of TETRAMAX was simply an opportunity not to be missed.

How did TETRAMAX build on the experience of TETRACOM, its predecessor?

TETRACOM established the core team of the TETRAMAX consortium. It proved the existence of a European information and communication technology (ICT) tech-transfer market via 'bilateral TTX', where an academic institution would transfer a specific, ready-made hardware or software technology to an industry partner for productization. TETRAMAX inherited this instrument but took it much further via new kinds of TTX and a truly pan-European dimension. TETRAMAX also took advantage of the well-known TETRACOM brand name and network, while several TETRACOM TTX clients were elevated to full consortium partners for the sake of European Union (EU)-wide coverage. Last but not least, TETRAMAX included a dedicated task for measuring the long-term impact of all the former TETRACOM TTX. This was a rare opportunity within an EU project framework, and we obtained remarkable results that were documented in a public white paper.

What, for you, were the highlights of TETRAMAX? What are you most proud of?

I think the numbers speak for themselves. We were able to generate a huge multiple out of the EC investment of €7 million in terms of new jobs, enabling startups, better products and so forth. TETRAMAX was a highly experimental project with considerable management complexity. In addition, our team in the end made a tangible contribution to boosting pan-European tech transfer with all TTX being strictly 'cross border'. Close collaboration with HiPEAC on numerous channels and joint events was definitely among the highlights as well. We also helped pave the way for some follow-up projects, which inherited best practices and concrete materials from TETRAMAX. Personally, I'm proud that, supported by a fantastic core team, I had the chance to lead my second EU project in a row with an 'excellent' final assessment by the EC.

What kind of impact would you like to see from TETRAMAX? Why is it important to measure longer-term impact?

Right from the start, we tracked a comprehensive set of numerical key performance indicators (KPIs) to capture the impact of our TTX as well as TETRAMAX as a whole. Measuring impact by KPI numbers is more ambitious than the traditional qualitative approach. The advantages are higher precision and better monitoring capabilities over the project's duration.

The TETRAMAX team at the 2020 HiPEAC conference

Most of our KPI results are publicly available in our project reports. They largely revolve around economic benefits for our clients, such as revenue increases that may be attributed to a certain TTX. On a higher level, another key demonstration of impact is the mobilization of concrete academia-to-industry tech-transfer activities throughout the community. Many of the great TTX proposals that unfortunately could not be funded by TETRAMAX were eventually implemented by other means like EU projects with similar open-call, third-party funding programmes. Concerning long-term impact, as mentioned above, we had a unique opportunity for concrete measurements via the direct TETRACOM / TETRAMAX transition. In short, we typically observed a 3x to 5x increase in KPI values a few years after the conclusion of a TTX, which is a very encouraging result.

What developments would you like to see in European industry over the next few years?

The EC is currently funding a multitude of projects with a clear mission for tech-transfer stimulation, such as the Smart Anything Everywhere (SAE) initiative. This enables exciting opportunities for industrial segments where Europe plays a predominant role. A good example is the automotive sector, where we are witnessing a revolution in hardware and software architectures, driven by the need for higher efficiency with the advent of autonomous vehicles and more diversity and independence in the semiconductor space. Our community can contribute a lot in this field but would benefit from specific directions for research from industry. Currently, a lot of European research output is exclusively flowing to China. European industry should be more determined to embrace the novel technologies developed 'on their doorstep' by local universities. Likewise, academia must acknowledge that tech transfer means more effort than just sending a research paper to a potential industry partner. The good news is that the award of HiPEAC 7 by the EC guarantees the continuity of an important tech brokerage platform for European ICT over the coming years.

FURTHER INFORMATION:
tetramax.eu

Open-source development kit enables plug-and-play FPGA acceleration in the cloud

IBM researchers in Switzerland have released the cloudFPGA development kit, named cFDK, which enables developers to deploy accelerated compute kernels as a network-attached function on field-programmable gate arrays (FPGAs) within minutes. Recently open sourced, it is the first development suite targeting standalone, network-attached FPGAs in the cloud, enabling scalable, FPGA-accelerated cloud-native applications.

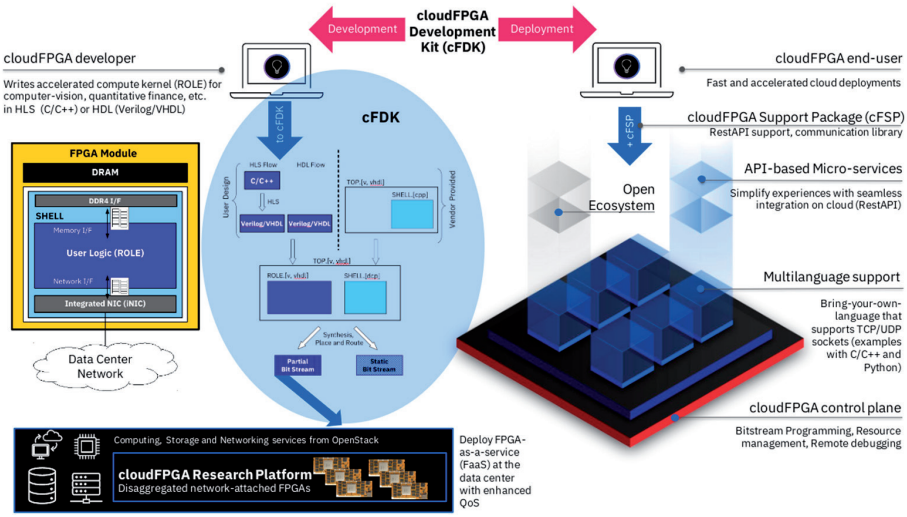
The researchers developed the cFDK with the goal of making FPGAs 'first-class citizens' in the cloud and consequently breaking them free of central processing units (CPUs). FPGAs are useful accelerators for various applications including video transcoding, high-performance computing, and data analytics.

As a consequence, there are many FPGA-accelerated compute kernels, but they have not been widely deployed in the cloud because there is a considerable barrier to adoption. Once the kernel has been coded, it usually takes months of work to interface this kernel with the host application, debug and finally deploy it in the cloud. This method is also at odds with the micro-service architecture widely used to set up scalable services in the cloud.

The cFDK addresses these pain points, and the IBM researchers have demonstrated how an accelerated quantitative finance kernel can be seamlessly integrated into scalable microservices, as well as how they can extend to efficient function-as-a-service (FaaS) offerings in the cloud (see references in 'Further information', below). In addition to the examples published as part of the cFDK, it is straightforward to integrate and deploy other accelerated kernels, such as those from whole custom kernels or the Xilinx Vitis Library.

In its initial release, the cFDK targets the cloudFPGA-platform also developed by IBM Research, which is available through a prototype deployment. The modular nature of the cFDK makes it easy to target different custom platforms or FPGAs offered by various cloud vendors, as long as the respective platform provides a network interface on the FPGA.

The cloudFPGA-platform is built on three main pillars, as shown in the diagram: 1) the use of standalone network-attached FPGA cards, 2) a hyper-scale infrastructure for deploying such FPGA cards at a large scale and in a cost-effective way, and 3) an accelerator service that integrates and manages the standalone network-attached FPGAs in the cloud. The main difference between cloudFPGA and other research projects and commercial products is the absence of a bus connection, e.g. via PCIe to a host CPU. The traditional communication channel and its associated card driver are replaced by a network stack (TCP/UDP) and its affiliated socket programming model. As a result of this paradigm change and an optimization of system integration, packaging, and cooling, the cloudFPGA-platform allows 1,000 FPGAs to fit into a single DC rack at 1/10 of the initial cost. The efficiency and ease of using the cloudFPGA-platform with the cFDK were the main reasons for selecting it as one of the target platforms in the Horizon 2020 project EVEREST.



FURTHER INFORMATION:

cloudFPGA on GitHub
github.com/cloudFPGA

EVEREST project
everest-h2020.eu

D. Diamantopoulos, R. Polig, B. Ringlein, M. Purandare, B. Weiss, C. Hagleitner, M. Lantz, and F. Abel, "Acceleration-as-a-µService: A Cloud-native Monte-Carlo Option Pricing Engine on CPUs, GPUs and Disaggregated FPGAs," 2021 IEEE International Conference on Cloud Computing (CLOUD), 2021, pp. 726-729, doi: 10.1109/CLOUD53861.2021.00096.

B. Ringlein, F. Abel, D. Diamantopoulos, B. Weiss, C. Hagleitner, M. Reichenbach, and D. Fey, "A Case for Function-as-a-Service with Disaggregated FPGAs," 2021 IEEE International Conference on Cloud Computing (CLOUD), 2021, pp. 333-344, doi: 10.1109/CLOUD53861.2021.00047.

ELES acquires Campera Electronic Systems

In January 2022, ELES Semiconductor Equipment acquired Campera Electronic Systems, which is specialized in maximizing productivity and optimizing performance for field-programmable gate arrays (FPGAs). We caught up with HiPEAC member Calliope-Louisa Sotiropoulou from Campera, to learn about the significance of this acquisition.

Congratulations on the acquisition! Why did ELES decide to acquire Campera?

A global leader in innovation and reliability, ELES' core business has so far focused on the semiconductor market. The acquisition of Campera Electronic Systems was prompted by their intention to strengthen their aerospace and defence division, established in 2000, and to contribute to their market diversification strategy. In addition to Campera's strong portfolio in aerospace and defence, we were specifically chosen for our experience in FPGA platforms, for the availability of our intellectual property (IP) and for our excellent reputation among our clients.

What does the buyout mean for Campera? How will it enable Campera to develop?

Campera Electronic Systems has extensive experience in the development of FPGA-based embedded systems. The partnership with ELES will allow us to broaden our offer to our clients. Together with our new owners, we can provide design, prototyping, development and production of electronic boards and systems, including mechanical design, reliability studies and complete system qualification, enhanced by the ELES reliability embedded test engineering (RETE) methodology, which guarantees the fastest and most competitive path towards zero defects and zero scraps. The ELES-Campera combination now offers a complete turnkey solution, a unique and distinctive offer for our clients' system needs.

We are proud to say that it was a strategic choice by our new owners to keep Campera as an independent company. This choice permits us to maintain all the flexibility and advantages of a small, innovative company with the security, solid base and experience that ELES, with its over 30 years in the market, can provide.

Any advice for HiPEAC members who are considering options for scaling up their commercial operations?

Our experience proves that, as long as your work is driven by quality and passion, you will be noticed and opportunities will follow. It requires patience and perseverance. ELES approached us because of our excellent reputation and because of the quality of our work. In the end, our best advertisement has always been our satisfied customers.

What are your plans for the future?

The ELES group keeps growing. Together with ELES, Campera Electronic Systems aims at creating a solid base in the international market and at strengthening our international profile. ELES already has commercial subsidiaries in the United States and Singapore that can open up whole new markets for us. In addition, we want to expand our research and development (R+D) activities, promoting cutting-edge solutions with partners in the European Union and across the world. New challenges have arisen in the new post-pandemic market and we need to be one step ahead.

Of course, we will continue to be active HiPEAC members and include ELES in this journey. We have already had five HiPEAC interns of the highest quality, and we currently have an open position for another. Through HiPEAC we are looking for new industrial and academic partnerships, employees and interns. Our new status has renewed our motivation and energy and we are looking forward to new ventures together with ELES.

CAMPERA ELECTRONIC SYSTEMS

ELES Your Partner in Semiconductor Reliability

The ELES headquarters in the scenic setting of Todi, Umbria

HEAppE website showcases the middleware path to HPC-as-a-service

Václav Svatoň, IT4Innovations

HPC-as-a-Service is a well-known term in high-performance computing (HPC). It enables users to access an HPC infrastructure without buying and managing their physical servers or data centre infrastructure. Through this service, small and medium enterprises (SMEs) can take advantage of the technology without an upfront investment in the hardware. This approach further lowers the entry barrier for users and SMEs who are interested in utilizing massive parallel computers but often do not have the necessary expertise in this area.

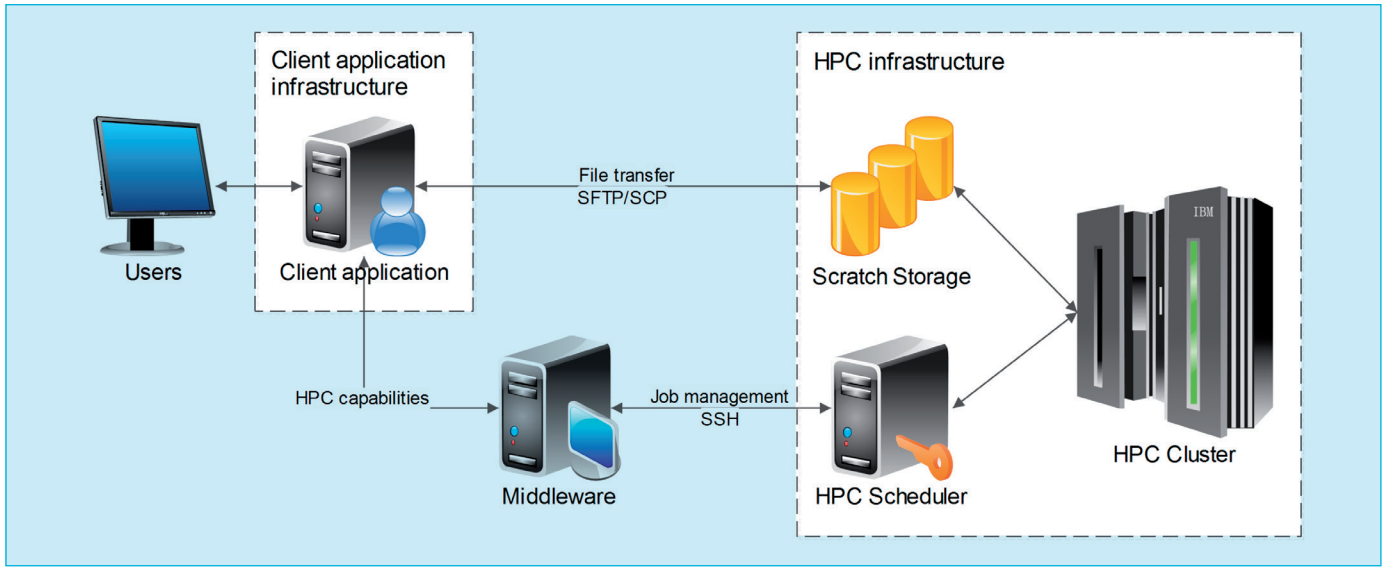
To provide simple and intuitive access to a supercomputing infrastructure via a representational state transfer application programming interface (REST API), an application framework called HEAppE has been developed. This framework uses a mid-layer principle known as middleware in software terminology. Middleware manages and provides information about submitted and running jobs and the data transferred by them between a client application and an HPC infrastructure. HEAppE allows the execution of required computations and simulations on HPC infrastructure, monitors progress, and notifies the user if the need arises. It provides necessary functions for job management, monitoring and reporting, user authentication and authorization, file transfer, encryption, and various notification mechanisms.

The first version of HEAppE, originally named HaaS Middleware, was developed as part of a joint project of IT4Innovations National Supercomputing Center and the transnational DHI company, a leader in hydrological software development. Primarily designed for application at IT4Innovations in hydrological modelling, HEAppE, currently available in version V3.0, has already been used in many different fields and operated by several supercomputing and data centres. HEAppE middleware is intended not only for one particular type of hardware for existing high-performance and future exascale computing systems but also for different systems in different supercomputing centres. By using HEAppE, all interested parties may benefit from HPC technologies.

The core of the HEAppE team is based at IT4Innovations at VSB – Technical University of Ostrava, which is a leading research, development, and innovation centre active in the field of HPC, data analysis (HPDA) and artificial intelligence (AI). IT4Innovations operates the most powerful supercomputing systems in the Czech Republic provided to Czech and foreign research teams from academia and industry.

To promote HEAppE, we have launched a new website, which offers essential information about this open-source solution. On this well-structured site, you'll find details of the know-how behind the middleware, a description of the HEAppE architecture and links to relevant publications. The HEAppE website helps visitors learn more about the technologies used or find out more information about supercomputers. In addition, it also offers references presenting all projects in which the HEAppE middleware solution has been used.

FURTHER INFORMATION:
HEAppE website
heappe.eu



LEXIS Platform lowers barrier to accessing HPC and cloud environments

Jan Martinovič, IT4Innovations

Today, many organizations aim to glean knowledge from data generated by industrial and business processes. However, combinations of high-performance computing (HPC), cloud and big-data technologies are required to process the ever-increasing quantities of data produced.

Traditionally, access to powerful computing platforms has been problematic for small and medium enterprises (SMEs), due to technical and financial considerations. This is now changing thanks to initiatives like the LEXIS project, financed by the European Union (EU), which built an advanced computing platform at the confluence of HPC, cloud and big data. The platform leverages large-scale, geographically distributed resources from existing HPC infrastructure, employs high-performance data analytics (HPDA) solutions and augments them with cloud services, allowing scientific, industrial and SME users to access these technologies across multiple HPC centres.

Driven by the requirements of several pilot use cases, the LEXIS Platform uses best-in-class data-management solutions provided by the EU-funded EUDAT initiative, and advanced, distributed orchestration solutions (Topology and Orchestration Specifications for Cloud Applications-TOSCA), augmenting them with new, efficient hardware and platform capabilities.

Common issues have been addressed, including distributed data storage, federated access, accounting and billing. This innovative platform aims to cater to the needs of European industry and create an ecosystem of organizations which could benefit from its HPC, HPDA and data management solutions.

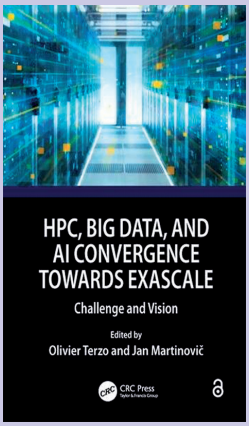
The LEXIS consortium has developed a demonstrator with a major open-source dimension, including validation, testing and documentation. It has been validated by large-scale pilots from industrial and scientific sectors, delivering significant improvements, including job execution time and solution accuracy.

LEXIS THE PLATFORM

Book: HPC, Big Data and AI Convergence Towards Exascale

Editors: Olivier Terzo and Jan Martinovič

HPC, Big Data and AI Convergence Towards Exascale: Challenges and Vision provides an update on the most advanced computing, storage and interconnection technologies at the convergence of high-performance computing (HPC), the cloud, big data and artificial intelligence.



Demonstrated by solutions devised within recent Horizon 2020 projects, the book gives an insight into challenges faced when integrating these technologies, as well as in achieving performance and energy-efficiency targets towards exascale. Emphasis is given to innovative ways of provisioning and managing resources, as well as monitoring their usage. In addition, industrial and scientific use cases provide practical examples of the need for cross-domain convergence.

Providing an overview of currently available technologies fitting the concept of unified cloud-HPC-big-data-AI applications, the book presents examples of their use in actual applications, including aeronautical, medical and agricultural case studies. The book is an example of how joined-up European research delivers real results: coordinated by the LEXIS and ACROSS projects, who cooperated closely on this work, it features contributions from CYBELE, DeepHealth, EVOLVE and the European Processor Initiative. Two chapters are available as open source.

FURTHER INFORMATION:

HPC, Big Data, and AI Convergence Towards Exascale

bit.ly/HPC_BigData_AI

lexis-project.eu

acrossproject.eu

LEXIS has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825532.

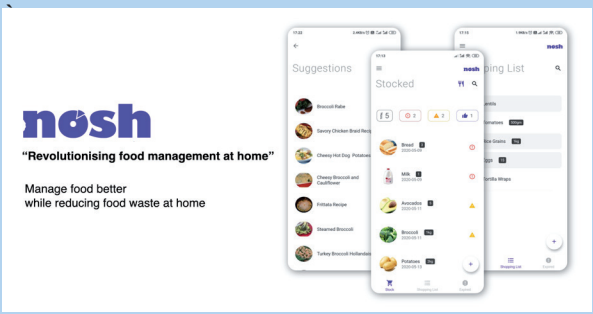
ACROSS has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement no 955648. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Italy, France, the Czech Republic, the United Kingdom, Greece, the Netherlands,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

Nosh Technologies develops multi-layered blockchain framework to reduce food waste

Nosh Technologies has announced that it has developed the world's first multi-layered blockchain framework using machine learning and cloud computing to reduce food waste. Named SmartNoshWaste, it was shown to reduce food waste by 9.46% in comparison to the originally collected food data.

'Instead of patenting the framework, we made it available as a peer-reviewed research article along with the code and data,' comments HiPEAC member Somdip Dey, one of the founders of Nosh Technologies. The paper was published in Smart Cities journal (see 'Further information', below).

Somdip grew up in India and developed the Nosh application after moving to the UK and discovering the extent of the food-waste problem there. He was recently named a Massachusetts Institute of Technology (MIT) Innovator Under 35 Europe, as well as a World IP Review Leader.



FURTHER INFORMATION:

Dey, S.; Saha, S.; Singh, A.K.; McDonald-Maier, K. 'SmartNoshWaste: Using Blockchain, Machine Learning, Cloud Computing and QR Code to Reduce Food Waste in Decentralized Web 3.0 Enabled Smart Cities'
Smart Cities 2022, 5(1), 162-176; doi.org/10.3390/smartcities5010011
Nosh Technologies article on Entrepreneur.com entrepreneur.com/article/421386
MIT Innovators Under 35 Europe 2022 emtecheurope.com/innovators-2022/

Dates for your diary

ACACES 2022: 18th International Summer School on Advanced Computer Architecture and Compilation for High-performance Embedded Systems
10-16 July 2022, Fiuggi, Italy
hipeac.net/acaces/2022

Euro-Par 2022: 28th International European Conference on Parallel and Distributed Computing
22-26 August 2022, Glasgow, UK
2022.euro-par.org

CPS Summer School
19-23 September, 2022, Pula, Sardinia (Italy)
cpsschool.eu

ITEM 2022: 3rd International Workshop on IoT, Edge, and Mobile for Embedded Machine Learning
19-23 September 2022, Grenoble, France
item-workshop.org

EFECS 2022: European Forum for Electronic Components and Systems
24-25 November 2022, Amsterdam, The Netherlands
efecs.eu

Lieven Eeckhout and David Kaeli named ACM Fellows

Lieven Eeckhout (left) and David. R Kaeli (right)

HiPEAC members Lieven Eeckhout (Ghent University) and David R. Kaeli (Northeastern University) have been named ACM Fellows in the most recent round of awards. Lieven is the first ACM Fellow in Flanders and only the second in Belgium.

The ACM Fellows programme recognizes the top 1% of ACM Members for their outstanding accomplishments in computing and information technology and / or outstanding service to ACM and the larger computing community. Fellows are nominated by their peers, with nominations reviewed by a distinguished selection committee.

Congratulations to both on this prestigious award!

Do you know a HiPEAC member who deserves an ACM award? Nominate them via the ACM website awards.acm.org/fellows/nominations

HiPEAC 2022 keynote speaker Hai 'Helen' Li is the Clare Boothe Luce Professor of Electrical and Computer Engineering at Duke University. We caught up with Hai in advance of the HiPEAC conference to delve into the fascinating – and now ubiquitous – topic of machine learning.

'Machine learning should consider vertical integration, from devices to circuits, architecture, systems, algorithms, and applications'

What attracted you to work in machine learning?

My background training was in very-large scale integration (VLSI) design and computer architecture. Fifteen years ago, I joined Seagate and started to work on emerging memory technologies. My interest quickly moved from high-density memory design to its use for neuromorphic computing. The goal of my approach was to renovate computer architecture and develop cognitive systems by emulating the structure and working mechanisms of the human brain.

The research has been a lot richer than I initially imagined – not only does it involve hardware implementation, but the methods are inevitably entangled with algorithm innovations. The huge potential of artificial intelligence / machine learning (AI / ML) attracted me to dedicate myself to this field – but I think the concept of ML is broader and must consider vertical integration, from devices to circuits, architecture, systems, algorithms, and applications.

What are some of the technical challenges involved in scaling up neural networks?

Scaling up neural networks can be achieved by increasing the depth and width of a network model. This is traditional thinking, but it is not an efficient way of improving learnability. We may reconsider other solutions, such as the neuro-symbolic AI approach, which uses deep-learning neural-network architectures and combines them with symbolic-reasoning techniques. In this way, many smaller neural networks can be composed into large ones to provide complicated and better functions.

The traditional design flow is composed of several abstract layers, which was mainly due to limited design capacity. The efficiency in hardware system design was obtained by sacrificing the optimization space. Holistic co-design tends to break the barrier across the vertical layers (devices, circuits, architecture and systems, algorithms, and applications) and therefore achieve global optimization. However, the optimization process could be more complicated, which will be a new challenge to address.

How can we help reduce the energy use of machine learning?

We must understand why training models consume so much power – essentially, we tend to retrain a model with large data samples, if not all the data samples. So novel training methods that inherit more information from an existing model would be the key. There is a lot of ongoing research in this direction, such as life-long learning, one-shot learning, etc. Another direction would be the renovation of neural-network model design with enriched data representation and processing capabilities, for example, the asynchronous, mixed-signal approach using advanced memory technologies.

Deep learning has blossomed since 2012, thanks to larger datasets and greater compute power. How do you see this field evolving over the next few years?

The advantages of deep learning have been demonstrated in diverse sets of applications and have attracted great interest. I believe that deep learning will be popularized in more application domains and present its great potential. Especially as data acquisition becomes easier, fast data processing becomes more critical. On the other hand, as not all the domain experts are machine learning experts, the automation of model generation (e.g. neural architecture search) and training will be important. This will further aggregate the requirement in computing power, which could be the bottleneck in preventing the AI/ML from progressing again.

Sweeping up best-paper awards and attracting widespread media coverage, Bianca Schroeder's research is shaking up the field of storage devices and system reliability. HiPEAC caught up with the University of Toronto professor in advance of her keynote talk at HiPEAC 2022.

'It is hard to imagine the capacity needs of the global datasphere being met entirely by SSDs'

Tell us about some of the main trends in storage over the last 10 years.

In terms of hardware, a major trend has of course been the proliferation of flash-based solid-state drives (SSDs). Also, more recently, after decades of research, phase-change memory has finally made its way from prototypes into products. While I think the jury is still out on what the main impacts of phase-change memory will be, SSDs have become ubiquitous everywhere ranging from personal devices to datacentre and high-performance computing (HPC) deployments.

In storage systems, disruptive innovation usually requires technical advances across the entire storage stack and the success story of SSDs is no exception. At the device level, the key for success has been to achieve higher density with minimal impact on cost, by successfully squeezing more bits into a cell (from originally one bit in single-level cell (SLC) drives to most recently four bits in quad-level cell (QLC) drives) and stacking cells vertically (3D NAND).

Going up the storage stack, we have seen a new interface (NVMe), flash-optimized systems software (e.g. flash-aware filesystems with TRIM support), and flash-optimized application software (e.g. key value stores based on flash-optimized data structures, such as log-structured merge trees), all of which have contributed to the success of SSDs.

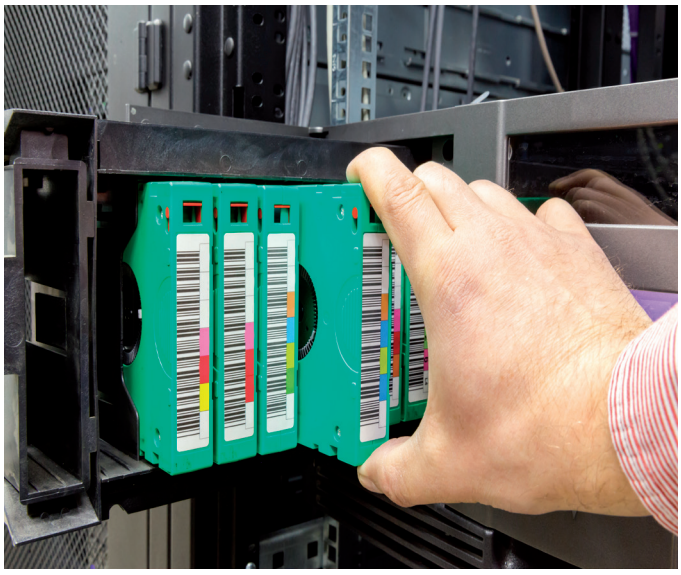
Seeing how SSDs have gone from use in only a few special performance critical systems 10-15 years ago to ubiquity today, it might be tempting to conclude that they will soon completely replace hard-disk drives (HDDs). And in fact, more SSDs are now being shipped annually than HDDs.

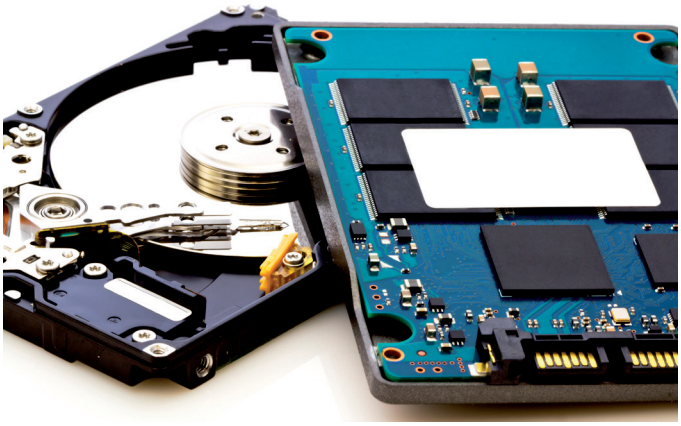
However, the story changes when looking at capacity: the total capacity of shipped HDDs exceeds that of SSDs by a factor of five times and one of the few things that have been a constant in the history of storage systems seems to be the ever growing user demand for more capacity. So essentially the future of HDDs will depend on their ability to continue to offer capacity at a significantly lower $/GB than SSDs. In the next few years this could be through new methods (such as heat-assisted or microwave-assisted magnetic recording, known as HAMR and MAMR) for increasing density.

What will probably change already in the near future is where we will see HDDs deployed. Especially with the increased use of cloud storage, SSDs might pretty soon replace HDDs in consumer devices. However, unless the $/GB for SSDs comes down significantly compared to HDDs it is hard to imagine that the huge capacity needs of the global datasphere can be met entirely by SSDs. Remember that even tape is not dead yet! It is still heavily used to meet the capacity needs of backups, archival applications and cold storage.

What techniques do you use for resource allocation and scheduling?

I have applied queueing theory to scheduling in web servers, database systems and high-performance computing systems in order to improve response times and I am currently exploring the





use of game theory for guiding garbage collection decisions in distributed storage systems. In much of my recent work I borrow techniques from machine learning (ML) to solve computing-systems problems. For example, I have applied ML to predict errors in storage devices and dynamic random-access memory (DRAM) and used those predictions to activate appropriate preventative measures; to predict failures of tasks in large-scale computing platforms and use those predictions to mitigate the effects of unsuccessful executions; or to predict resource consumption of virtual-machine workloads to achieve better resource utilization.

What are some of the most interesting real-life projects you've worked on?

What fascinates me about real-world data from production systems is that once you start digging you almost always find something interesting and unexpected. Besides personal curiosity it also seems that it is our duty as scientists to validate any assumptions we make against real data, even for things that might seem 'intuitive'.

Often even baseline statistics about device behaviour entail some surprises. For example, our work on hard-disk reliability showed that device replacement rates are significantly higher than what datasheet mean times to failure would suggest and that the rate of uncorrectable errors in SSDs is much higher than one would expect. It was also interesting to see that a higher price point does not always translate to better reliability, e.g. we didn't see evidence that the more expensive enterprise HDDs or SSDs are consistently and significantly more reliable than nearline devices (although they might have some other advantages of course).

An example of how field data analysis can lead to improvements in systems is error handling in DRAM. Traditionally, the assumption was that DRAM errors are caused by a random transient disturbance of the charge, e.g. by cosmic rays, and the best protection is through error-correction code (ECC). One of the things we found when analysing field data from Google's data centres is that DRAM errors often repeat on the same dual in-line memory module (DIMM) and even in the same cell or region of a DIMM. That observation points to hardware issues rather than transient errors caused by some external disturbance.

It also means that, rather than relying on ECC, a more effective mechanism to protect against future errors is by removing the affected memory page from further use ('page offlining'). We demonstrated the effectiveness of page offlining in simulation in a 2012 ASPLOS paper. Today page offlining is widely used in practice and, in addition, vendors have begun to provide different types of systems support for working around DRAM hardware problems, for example through post-package repair (PPR), adaptive double-device data correction (ADDDC) and partial cache-line sparing (PCLS).

What advances would you like to see in resource allocation over the next 10 years?

As a community we are good at continuing to build more and more powerful systems, but we could be using them more efficiently. When looking at field data I see so many different types of resource wastage. For example, when analysing traces of jobs in massively parallel computing platforms (e.g. traces published by Google or LANL), a third of the tasks fail or are being killed, resulting in wasted cycles. Similarly, analysing traces of virtual machine workloads shows that more than half of them use on average less than 10% of their resources. Meanwhile, when analysing field data from enterprise storage systems, we observe that even at the end of life (five years) more than half of the systems use less than 50% of their storage capacity. One problem is that users are poor at estimating their resource needs. So more tools that help in that direction would be useful.

For many years, championing free and open-source computing technologies was the preserve of a committed subset of computer programmers. Today, as demand for compute power grows against the backdrop of a slowing Moore's Law, open source is looking increasingly mainstream. For this article, we spoke to international experts to find out why open source matters, and what it means for Europe.

Breaking out of the black box

Open source gets serious in Europe

Commoditization to innovation

According to **Philippe Krief**, director of research relations at the Eclipse Foundation, the first wave of adoption of open source by companies was driven by cost reduction and commoditization: 'Linux, MySQL and PHP were all adopted to reduce the cost of operating systems and infrastructure software.' The second wave, he says, was embodied by projects like Eclipse and OpenStack that used open source to challenge the status quo in development tools and cloud infrastructure by sharing development resources.

Today, open source is an essential driver of innovation, says Philippe. 'By allowing common code to be shared, open source allows 80% of a development team's resources to focus on innovation and added value instead of maintaining code that does not differ from the competition,' he says. 'In addition, in the last 10 years open source has become an enabler of innovation by promoting the adoption of new technologies: the more your user community grows, the more valuable your innovation becomes. This has been shown in areas from big data and cloud, to artificial intelligence, to new development approaches like continuous integration / continuous deployment.'

Little wonder, then, that major commercial players have been snapping up companies in the open-source landscape, in a series of big-money acquisitions. 'Look at IBM acquiring RedHat for €34 billion, Microsoft buying GitHub for €7.5 billion, or Salesforce acquiring MuleSoft for €6.5 billion,' Philippe points out. This has been shown in areas from big data and cloud, to artificial intelligence, to new development approaches like continuous integration / continuous deployment,' he adds.

From open source to secret sauce

More than 40% of companies in the software industry participate in at least one open-source project, while, according to a recent study by Forrester, 96% of companies consider open source to be important in their digital transformation initiatives. For some companies, open source forms the centre of their entire value proposition.

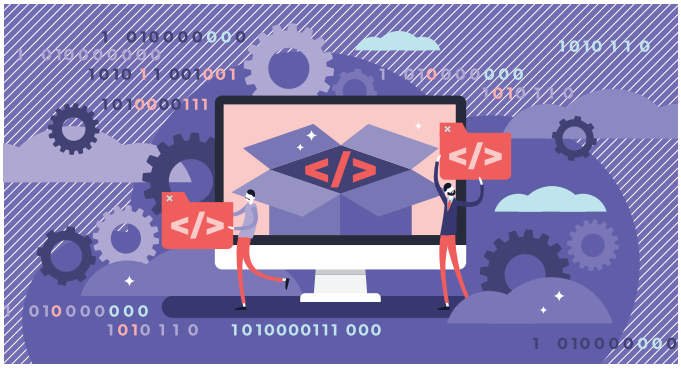
Philippe Krief

Gaël Blondelle

'Over the years, three main business paradigms have emerged: the product, the distribution and the platform,' explains **Gaël Blondelle**, vice president of ecosystem development at the Eclipse Foundation.. 'In the product approach, a single company uses open source to promote adoption of its technology and create a community. The business model relies on the ability to differentiate a "community" version of a product available under an open-source licence from a "professional" version under a classical end-user licence agreement. Usually, companies welcome external contributions but require copyright attribution to give them full control of the destiny of the project.'

'The distribution approach is used when a complex technology with a multitude of components requires both technical and legal expertise to package a distribution that brings value to enterprise customers,' says Gaël. 'Mostly associated with the Linux landscape, this business model is also used by RedHat, Suse and Canonical, to give other well-known examples.'

Finally, the platform business model enables collaboration between competitors on an open-source platform designed to be extended by companies within the ecosystem to create products and services, explains Gaël. 'This powers the Pareto principle in software development: collaborate with competitors on the platform that represents 80% of the effort to build a product, and focus your resources on the application(s) or service(s) representing 20% of your software, but 100% of your unique value proposition.' The platform model is supported by open-source foundations, which offer vendor-neutral governance, notes Gaël. Examples include the Eclipse Foundation (development tools and the internet of things), the Open Infrastructure Foundation with OpenStack, the Cloud Native Computing Foundation with Kubernetes and the Apache Foundation with big data.



Open source allows end users to easily start with the technology and to focus on their own application codes. However, for proprietary software companies, open source provides both an opportunity and a threat, explains Gaël. 'On the one hand, using open-source technologies shortens product development time. On the other, companies who sell proprietary software will have to evolve, refocusing what differentiates them in the context of a global open-source ecosystem.'

Opening the full stack

While the success of open-source software is undeniable – Linux, for example, is now the operating system of choice for the world's most powerful computing systems, running on all of the Top 500 supercomputers – open hardware has only recently become a mainstream concern. 'I would say that open hardware is currently at the stage where open software was around 20 years ago,' says **Luca Benini**, chair of digital circuits and systems at ETH Zürich and full professor at the University of Bologna, whose team has been pioneering open integrated circuit designs, as shown by the PULP platform (see pp. 24-25).

Luca Benini

The free and open-source software movement

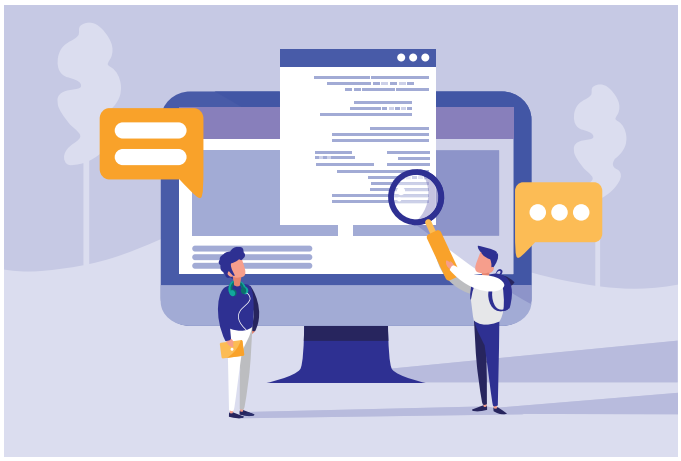
In the 1950s and 1960s, source code for operating systems and compilers was distributed freely as part of hardware packages, allowing programmers to fix bugs or add functionality. This changed in the 1970s in the face of soaring software production costs, with the introduction of software licences.

In response to this trend, in the early 1980s a community began solidifying around the concept of 'free' software which could be studied, one of whose pioneers was the GNU Project founder Richard Stallman. In 1991 the Linux kernel, created by Linus Torvalds, was released as freely modifiable source code.

The dot com movement in the late 1990s resulted in free software becoming popular in web development, including Apache's HTTP server, MySQL database engine and PHP programming language. In 1998, the term 'open source' was coined by Christine Peterson.

Linux and RISC-V are open-source successes

John D. Davis



For **John D. Davis**, a lead researcher at Barcelona Supercomputing Center (BSC), there are good reasons to open up the full computing stack, from applications to hardware. 'Trends like the slowing of Moore's Law, plus the need to increase performance while reducing power, require hardware-software co-design,' he explains. 'If all the layers of the computing stack are open, the layers providing the biggest impact can be changed in concert, reducing design complexity and enabling mutually beneficial design points in both hardware and software. Basically, an open computing stack enables optimization where optimization makes the most sense.'

There are additional benefits to open source, he adds. 'In terms of software, you can leverage a large ecosystem across implementations. Security is improved, as you have a fully auditable collection of intellectual property (IP), and safety benefits from the absence of black boxes. There is no vendor lock-in, and an ecosystem to enable custom development from small companies to large enterprises. Plus, the collaborative aspect – the existence of a large open-source community and existing open-source solutions – means you can achieve faster time to market.'

For Luca Benini, too, open hardware offers improved measures for security: 'In the past, security was achieved mostly through obscurity,' he explains. 'However, we've seen many times that hackers always manage to find a way in. But if you make your security open, you have a lot of "good guys" who can help you identify and fix vulnerabilities. Openness also leads to greater trust, as it allows others to inspect your security solution.'

> *"An open computing stack enables optimization where optimization makes the most sense"*

Open opportunities for European business

Thanks to its role in driving innovation, open source can allow researchers to challenge the status quo and become leaders in a specific field, says Philippe Krief, citing the RISC-V initiative and OpenHW Group as examples. This is particularly important for Europe, which lacks the major vertical companies dominant in other parts of the world but which has a strong research infrastructure, he notes. 'Open source enables digital autonomy because it replaces dependency on proprietary intellectual property mainly owned by non-European companies with the need to invest in people and skills. And the good news is that European countries, with their well-established universities and robust education systems, are a great source of skilled researchers and engineers.'

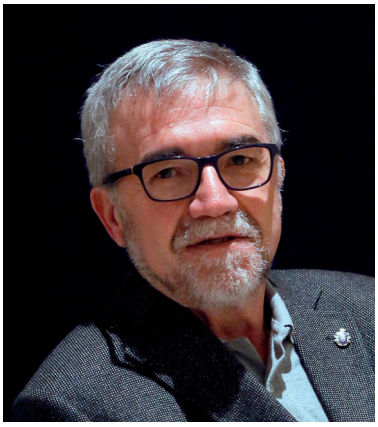
Gaël Blondelle agrees: 'European companies are not in the same league as those in North America or Asia. If they play alone in this category, at best they will be bought out, at worst put out of business. So if a European company wants a competitive edge, it needs to build a network of European companies to pool resources.'

One example of this strategy is Robert Bosch, notes Gaël. 'Executives at Bosch realized that if each platform promoted their technology to the detriment of others, they would eventually collapse, resulting in a dependence on technologies by big tech companies. Instead, they decided to put the IoT technologies developed internally into the open arena, in order to benefit from a snowball effect leading to wider take up of their technologies on one hand, and the continued improvement of these technologies on the other.'

Today, Bosch makes use of a stack of operating system (OS) components from its own research and development (R+D) offices, from the R+D offices of its competitors, and from other sources. 'Not only does Bosch have complete control over this stack; if one of the component owners proves to be fallible, the company could easily replace the component or development

team with one of its own,' says Gaël. 'What's more, as the platform gains in maturity and innovation, it is also adopted by its competitors, who, de facto, contribute to its sustainability and standardization as the AUTOSAR platform.'

Open source as government policy



In addition to growing enthusiasm for open source among companies, recently there has been a growing trend for governments to embrace open source. This can be witnessed by the emphasis on RISC-V in the EU's Chips Act, as well as a flurry of infrastructure investments by South Korea, China and the United States, for example. 'Open source provides a unique opportunity for Europe to regain technological sovereignty and superiority,' argues BSC director and HiPEAC co-founder and first coordinator **Mateo Valero**.

'For example, by expanding the RISC-V ecosystem, Europe can leapfrog into new technology areas and overtake the current leaders, taking advantage of the freedom of access and implementation from design to production offered by open source. It is gratifying to see this being taken seriously by states, such as the Spanish Government's recently announced €11 billion investment in semiconductor industries, which will create exciting new opportunities thanks to open-source technologies.'

Mateo recommends investment by the EU in RISC-V in all markets that are strategic to European digital sovereignty. 'Policies should promote the research, development and industrialization of RISC-V high-performance computing (HPC) solutions, a market that is currently non-existent in Europe. In other areas such as embedded computing and the internet of things (IoT), where Europe already has mature markets, policies should support the transition from the traditional, closed hardware IP to open IP. This includes supporting infrastructure IP for devices, input / output (I/O) and memory interfaces,' he says.

To help take open source forward in Europe, BSC has created the Laboratory for Open Computer Architecture (LOCA), a mechanism to bring together industrial partners and researchers from all over the world to develop the RISC-V ecosystem. 'LOCA provides the ideal sandbox for hardware / software co-design, matching industry veterans with researchers and students. The aim is to develop European know-how of high-performance chip design in the latest silicon technology nodes,' explains John.

'Stemming from our participation in the European Processor Initiative (EPI), LOCA was created due to the need to create a RISC-V roadmap for HPC at BSC and, more generally, in the rest of Europe. The exciting part is that RISC-V enables a new level of research based on hardware-software co-design, and LOCA acts as an instrument to facilitate education, training and research with a common goal across multiple projects.' Projects currently under the LOCA umbrella at BSC include MEEP, eProcessor and the EUPILOT; see pp.38-42 for more details.

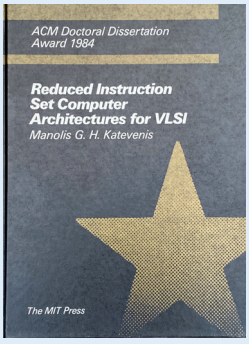
A brief history of RISC-V

It's hard to believe that the open-source RISC-V ISA – now the basis of a global foundation supported by thousands of organizations – was born out of a research project at the University of California, Berkeley as recently as 2010. The project aimed to create a 'clean-slate' ISA as an alternative to complex and costly proprietary ISAs.

As the Roman numeral 'V' ('five') in the name indicates, this work built on four preceding projects on the reduced instruction set computer – or RISC – concept, developed by renowned computer architects David Patterson, Carlos Séquin (both at UC Berkeley) and John Hennessy (Stanford University). Patterson and Hennessy were later recipients of the ACM Turing Award, often referred to as the 'Nobel Prize of computing'.

Did you know?

HiPEAC founding partner Manolis Katevenis (Foundation for Research and Technology Hellas – FORTH) designed the micro-architecture of the RISC-I and RISC-II chips, and was the chief implementor of the RISC-II single-chip microprocessor at UC Berkeley. For this thesis he received the 1983 Sakrison Memorial Prize and the 1984 ACM Doctoral Dissertation Award.



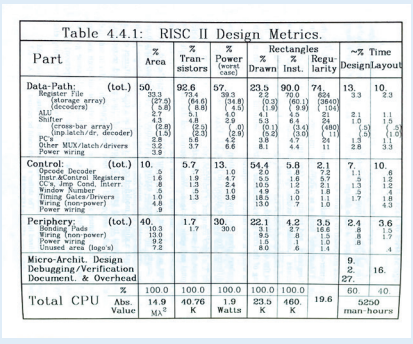


Table 4.4.1: RISC II Design Metrics.

Part	% Area	% Transistors	% Power (worst case)	Rectangles % Drawn	Rectangles % Inst.	Rectangles Regularity	~% Time Design	~% Time Layout
Data-Path: (tot.)	50.	92.6	57.	23.5	90.0	74.	13.	10.
Register File (storage array) (decoders)	33.3 (27.5) (5.8)	73.4 (54.6) (8.8)	39.3 (34.8) (4.5)	2.2 (0.3) (1.9)	70.0 (60.1) (9.9)	624 (360) (104)	3.3	2.3
ALU	2.7	5.1	4.0	4.1	4.5	21	2.1	1.1
Shifter (cross-bar array) (inp.latch/dr. decoder)	4.3 (2.8) (1.5)	4.8 (2.3) (2.3)	2.0 (.6) (1.4)	5.3 (0.1) (5.2)	6.4 (3.4) (3.0)	24 (480) (41)	1.0 (.5) (.5)	1.3 (.5) (1.0)
PCs	2.8	5.6	4.2	3.6	4.7	24	2.8	1.1
Other MUX/latch/drivers	3.2	3.7	6.6	8.1	4.4	11		3.3
Power wiring	3.9							
Control: (tot.)	10.	5.7	13.	54.4	5.8	2.1	7.	10.
Opcode Decoder	.5	.7	1.0	2.0	.8	7.2	1.1	.6
Instr.&Control Registers	1.6	1.9	4.7	5.5	1.8	5.7	.5	1.2
CC's, Jmp Cond. Interr.	.8	1.3	2.4	10.5	1.2	2.1	1.3	1.2
Window Number	.5	.5	1.0	4.9	.5	1.8	.7	.4
Timing Gates/Drivers	1.0	1.3	3.9	18.5	1.0	1.1	1.7	1.8
Wiring (non-power)	4.8			13.0	.7	1.0		4.3
Power wiring	9							
Periphery: (tot.)	40.	1.7	30.	22.1	4.2	3.5	2.4	3.6
Bonding Pads	10.3	1.7	30.0	3.1	2.7	16.8	.8	1.5
Wiring (non-power)	13.0			9.5	.8	1.5	.8	1.7
Power wiring	9.2			1.8	.1	1.0	.8	.4
Unused area (logo's)	7.2			8.0	.6	1.4		
Micro-Archit. Design							9.	
Debugging/Verification							2.	16.
Document. & Overhead							27.	
%	100.0	100.0	100.0	100.0	100.0		60.	40.
Total CPU Abs. Value	14.9 MA^2	40.76 K	1.9 Watts	23.5 K	460. K	19.6	5250 man-hours	

Photos courtesy of Jan Gray

How to make a success of open source

So you've got your project, you're convinced of the virtues of opening it to the world… what now? Achieving success in an open-source project requires serious commitment, Philippe Krief stresses. 'There are still a lot of people who think that pushing code to GitHub is enough to make the whole world adopt their technology. However, once public, it will be among millions of repositories.'

To gain support for an open-source project, Philippe recommends working on two levels in parallel. 'First, work on the quality of the code and the associated resource (documentation, licensing of the code and third parties, tutorials, automated tests, etc.). Second, ensure your project is well communicated and disseminated: inform the community, in all transparency, about progress made, problems encountered and future evolutions,' he says. 'To take advantage of cross-fertilization, meet up with communities related to the project. It is also vital to involve users to gather feedback or implement use cases showing the application potential of the platform.'

'It is important to bear in mind the four basic principles of open source: transparency, openness, meritocracy and vendor neutrality,' adds Philippe. 'Transparency in the code and project decisions allow newcomers to see that the project is in good health. Openness invites outsiders to contribute, unleashing the wisdom of crowds as opposed to individual expertise. Meritocracy means that developers have to earn their right to be able to change the code by demonstrating their skills and willingness to contribute. Finally, vendor neutrality is the cornerstone of open collaboration, with each participant contributing in a way that services their own goals as well as those of the community.'

Operation RISC-V reaches new heights

Thanks to a wealth of industry support, RISC-V International is becoming ever stronger: 'Many large companies have joined the foundation, meaning that no single company has the power to shut down or aggressively take over the open-source ISA,' explains Luca Benini.

'With over 2,700 members, the RISC-V community is growing rapidly,' adds **Calista Redmond**, chief executive of RISC-V International. This year, Intel was a high-profile addition to the foundation's premier membership category, joining the likes of Alibaba and Google, along with government bodies such as the Chinese Academy of Sciences, the RISC-V International Open Source Laboratory and Indian Ministry of Electronics and Information Technology.

This investment has translated to industry progress, according to Calista, citing as examples high-performance processors from Esperanto and MIPS, as well as a line of microprocessors from Alibaba. 'The traction has continued to build with Seagate and Western Digital in storage, and in multiple domains from SiFive and others, including automotive and other safety-critical applications. We've also seen the first proof points of RISC-V in mobile with Alibaba porting Android 12 to RISC-V.'

The RISC-V International headquarters recently moved to Switzerland, in a nod to the success of the ISA in Europe, Much investment and engagement has been driven by European Union initiatives, Calista notes. 'Recent examples include E4's Monte Cimone cluster, the ControlPULP HPC processor, post-quantum cryptography work from the Technical University of Munich, and the first European Processor Initiative RISC-V chips.'

Today, RISC-V is a viable option across all compute workloads, Calista says, including artificial intelligence, high-performance computing, and more. Beyond acceleration, it is also moving into the central processing arena, with preparations for general-purpose computing from laptops – as showcased by SiFive, the Institute of Software at the Chinese Academy of Sciences and ClockworkPi – to server-grade central processing units (CPUs), she adds.

'As an open architecture, RISC-V offers much more flexibility and scalability compared to proprietary instruction set architectures (ISAs),' says Calista. 'Companies can easily implement the minimal instruction set and add extensions to create custom processors for innovative workloads. Furthermore, RISC-V's shared tools and development resources help companies reduce risk and accelerate time to market.'

RISC-V is already found in nearly a quarter of application-specific integrated circuit (ASIC) and field-programmable gate array (FPGA) projects today, according to a Wilson Research Group Functional Verification study. RISC-V is also rapidly growing in the AI sector; by 2027, Semico Research projects that there will be 25 billion RISC-V AI system-on-chips. This rapid growth is bolstered by the expansion of the RISC-V ecosystem and the software, tools, and services it offers as building blocks. Sixty-six workgroups are currently working on extensions, with 15 new extensions ratified in 2021, while special interest groups are working to control for fragmentation, Calista adds. 'The more we work together, the more strategic and durable our collective future becomes.'

Hardware is… hard

So much for open-source software. Open hardware, however, is a little more complicated, as **Frank K. Gürkaynak**, a senior scientist in the digital circuits and systems group at ETH Zürich, explains (see 'Open source hardware is here to stay' in the HiPEAC Vision 2021). 'The manufacturing process for integrated circuits – which may have billions of components – is very complicated and takes several weeks. The factories which produce them, known as "fabs", represent billions of dollars in infrastructure investment.' Another complicating factor, he notes, is that 'there are far fewer integrated circuit designers that there are software developers, reducing the pool of people who can provide open-source solutions'.

Before the manufacturing stage, electronic design automation (EDA) software, which keeps pace with the increasing complexity in semiconductor blueprints, is used to design chips, Frank adds. 'Today three major companies – Cadence, Synopsys and Siemens EDA (formerly Mentor Graphics) – dominate the EDA tool market. Any serious IC design relies on these commercial tools, which come with significant licensing costs.'

In addition, pre-designed and validated subsystems, available through third-party providers, are also used to help manage the complexity. 'All this means that there are multiple entities with different commercial interests involved in the process, and these relationships have to be understood to develop more sustainable solutions for open hardware.'

Publishing descriptions of the lower levels of hardware designs, the physical blueprint and the mapping of the circuit onto components, is currently problematic. 'These levels require the use of EDA tools and third-party intellectual property (IP), and vendors are unsurprisingly unhappy to see information published which could affect their revenues,' explains Frank. 'However, register transfer level (RTL) descriptions, which do not require the direct involvement of EDA tools and where no technology-specific information is disclosed, have been behind the current success of open hardware.'

> ***"There are still a lot of people who think that pushing code to GitHub is enough to make the whole world adopt their technology"***

Supporting open-source hardware at lower abstraction levels and providing components that have already been implemented is an important next step, says Frank, and one that will involve a culture change at the respective companies. However, the shift to open hardware will be well worth it, he says: 'For small and medium companies, open-source hardware results in cost savings and a faster path to innovation. As for research, open source allows us to quickly get started on innovation, without spending a large amount of time dealing with legal overheads for what is essentially commodity infrastructure.'

With momentum growing, open hardware has a sunny outlook, according to Luca Benini. 'Europe is taking a strong position on RISC-V and open hardware, with major funding initiatives in both the EuroHPC and KDT (Chips) Joint Undertakings. It is a really exciting time to work in this area.'

FURTHER INFORMATION:

Eclipse Foundation eclipse.org

OpenInfra openinfra.dev

Cloud Native Computing Foundation cncf.io

Apache Foundation apache.org

RISC-V International riscv.org

AUTOSAR (AUTomotive Open System ARchitecture) platform autosar.org/

'Seize The Open Source Opportunity Through Comprehensive, Optimized Strategies' – Forrester report bit.ly/Forrester_OpenLogic_report

Wilson Research Group Functional Verification Study 2020 – report on Tech Design Forum bit.ly/3PPhrBx

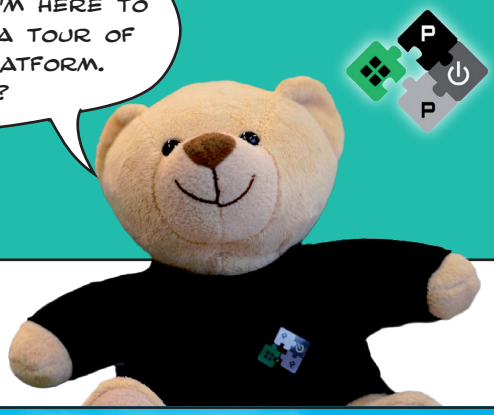
Analyzing the RISC-V CPU Market for SIP, SoCs, AI and Design Starts, Semico Research – report on RISC-V International website bit.ly/3PJiHpT

2020 TODO Group Open Source Program Survey Results bit.ly/TODO_OpenSource_Survey

'Open source hardware is here to stay', HiPEAC Vision 2021 doi.org/10.5281/zenodo.4719698

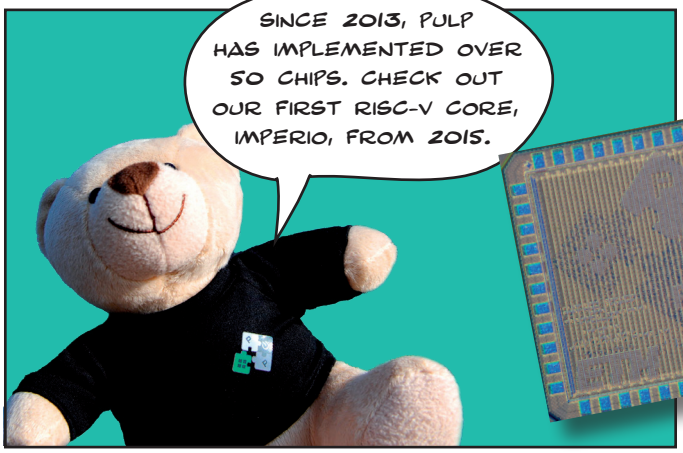
PULP FACTS

HI! I'M BIANCA, INTERNATIONAL BEAR OF MYSTERY, AND I'M HERE TO TAKE YOU ON A TOUR OF THE PULP PLATFORM. READY?



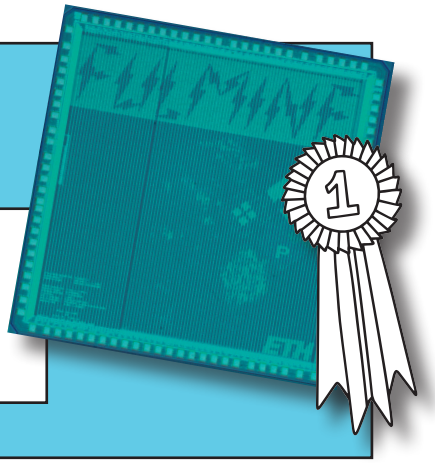
PULP STANDS FOR PARALLEL ULTRA LOW POWER. THE PULP PLATFORM STARTED AS A COLLABORATION BETWEEN THE INTEGRATED SYSTEMS LABORATORY AT ETH ZÜRICH AND THE ENERGY-EFFICIENT EMBEDDED SYSTEMS GROUP AT THE UNIVERSITY OF BOLOGNA.

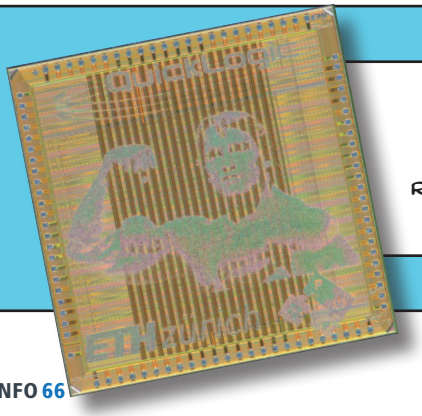
SINCE 2013, PULP HAS IMPLEMENTED OVER 50 CHIPS. CHECK OUT OUR FIRST RISC-V CORE, IMPERIO, FROM 2015.



(BACK THEN, MY PREDECESSOR, ALEX, WAS IN CHARGE. ALEX WENT MISSING BACK IN 2019. WE DON'T TALK ABOUT ALEX)

SOME OF OUR CHIPS HAVE WON ACCOLADES, LIKE FULMINE, WINNER OF AN IEEE CIRCUITS AND SYSTEMS BEST PAPER AWARD.

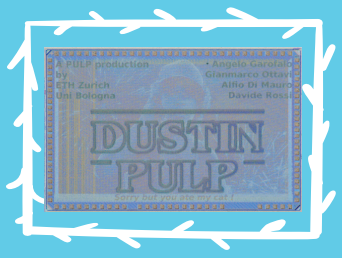




SOME PROVIDE EXTRA COMPUTATIONAL MUSCLE, LIKE ARNOLD, WHICH WAS PRODUCED IN RECORD TIME THANKS TO INDUSTRY COLLABORATION WITH QUICKLOGIC.

WHILE OTHERS ARE JUST PRETTY HOT CHIPS... EACH CONTRIBUTING SOMETHING UNIQUE TO THE OPEN HARDWARE ECOSYSTEM...

MR WOLF

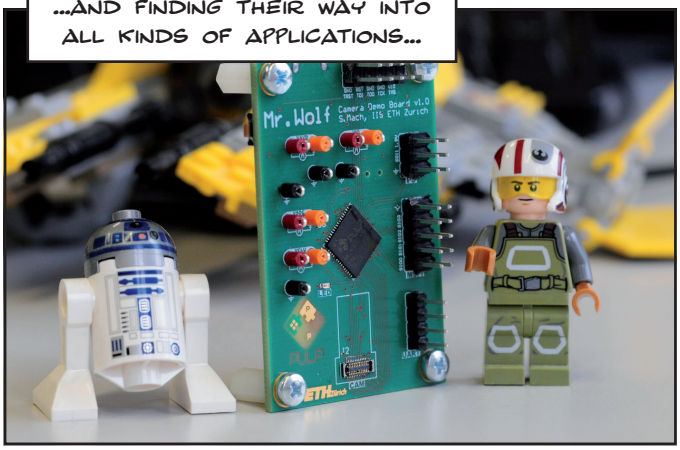


DUSTIN

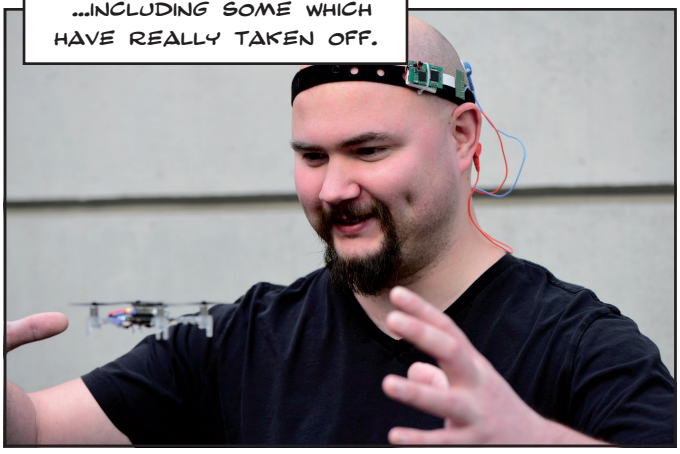
DARKSIDE

KRAKEN

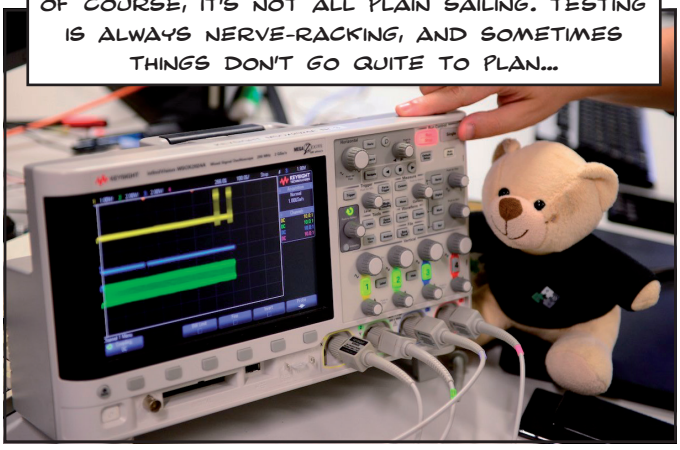
...AND FINDING THEIR WAY INTO ALL KINDS OF APPLICATIONS...



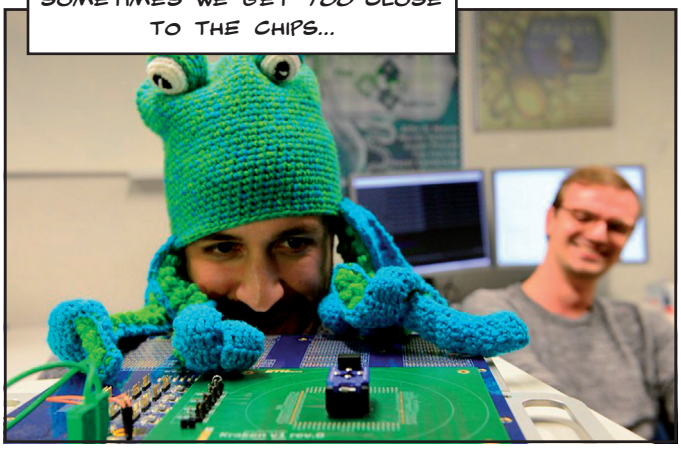
...INCLUDING SOME WHICH HAVE REALLY TAKEN OFF.



OF COURSE, IT'S NOT ALL PLAIN SAILING. TESTING IS ALWAYS NERVE-RACKING, AND SOMETIMES THINGS DON'T GO QUITE TO PLAN...



SOMETIMES WE GET TOO CLOSE TO THE CHIPS...



BUT WITH PARTICIPATION IN MAJOR EU-FUNDED PROGRAMMES LIKE THE EUROPEAN PROCESSOR INITIATIVE, THE FUTURE'S BRIGHT FOR PULP AND OPEN HARDWARE.

WANT TO FIND OUT MORE? VISIT PULP-PLATFORM.ORG AND CHECK OUT OUR CHIPS ASIC.ETHZ.CH AND FOLLOW US ON TWITTER @PULP_PLATFORM

Open source helps researchers share ideas and develop innovative technologies faster. In this article, we look at some of the range of HiPEAC technologies being developed as open source, from platforms for safety-critical systems, to internet-of-things (IoT) security solutions, to tools and libraries to exploit accelerators for high-performance and distributed computing.

Open for business

Open-source HiPEAC technologies across the compute continuum

OPEN-SOURCE SAFETY WITH SELENE

Carles Hernández (Universitat Politècnica de València), Sergi Alcaide (Barcelona Supercomputing Center), Konrad Schwarz (SIEMENS), Nicholas McGuire (OpenTech), Martin Rönnback (Cobham Gaisler), Charles-Alexis Lefebvre (Ikerlan) and Martin Matschnig (SIEMENS)

The SELENE platform is the first fully open-source RISC-V safety-relevant high-performance platform including a system-on-chip (SoC) and a hypervisor. The platform includes a set of safety features and acceleration support to meet the needs of high-integrity, performance-hungry applications in the space, automotive and railway domains, among others.

'Safety-related innovations in commercial hardware and software platforms are usually kept secret, as they are often regarded as a product-differentiating feature,' explains Carles Hernández, the coordinator of SELENE. 'However, making safety-related features open source allows anyone involved in a safety-related product, such as a car manufacturer, to audit and improve these platforms.'

For public research centres and universities, patenting or licensing can be counterproductive if there is no clear roadmap to exploit the results, says Carles. 'Meanwhile, the industrial partners involved in SELENE have been active in the open-source community for a long time, contributing to well-known open-source projects like real-time Linux or the LEON processor.'

Hardware architecture

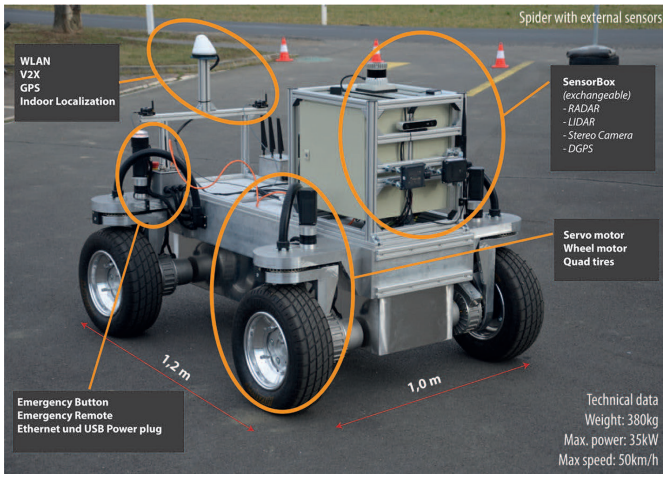
SELENE's hardware architecture comprises 6 NOEL-V 64-bit RISC-V cores featuring the hypervisor and compressed instruction set architecture (ISA) extensions. NOEL-V cores are dual issue, have L1 instruction and data private caches, and share a L2 cache. The SoC architecture also comprises several hardware acceleration modules that are interconnected to an AXI4 crossbar. The interconnect allows cores and accelerators to share the main memory and serves several general-purpose and domain specific input / output (I/O) devices.

Software architecture

The system software for the SELENE SoC utilizes Linux as the base operating system, well-known open-source frameworks such as ROS2 (depending on the needs of the different use cases defined in SELENE), and the Jailhouse hypervisor.

The ISAR build system is used to create the Debian-based Linux system image. This approach combines the advantages of classic embedded distributions and general-purpose distributions: it can tailor the kernel, drivers, and key software applications exactly to the hardware on the one hand, while reaping the benefits of centrally maintained, pre-compiled packages on the other.

Jailhouse is a simple partitioning hypervisor: it statically partitions multicore hardware into separate 'cells', each capable of hosting a guest operating system or bare-metal code. This allows the separation of critical and non-critical code, which simplifies the certification of safety-critical code and allows the consolidation of different software bases on a single multicore.



This autonomous vehicle, the Smart Physical Demonstration and Evaluation Robot (SPIDER) from Virtual Vehicles, is one of SELENE's use cases

Built-in safety support

The SELENE platform includes safety-related controllability and observability features, along with support to implement safety measures during operation. For validation purposes the DAVOS tool has been adapted to allow performing exhaustive fault-injection campaigns both at register transfer level (RTL) and for field-programmable gate arrays (FPGAs).

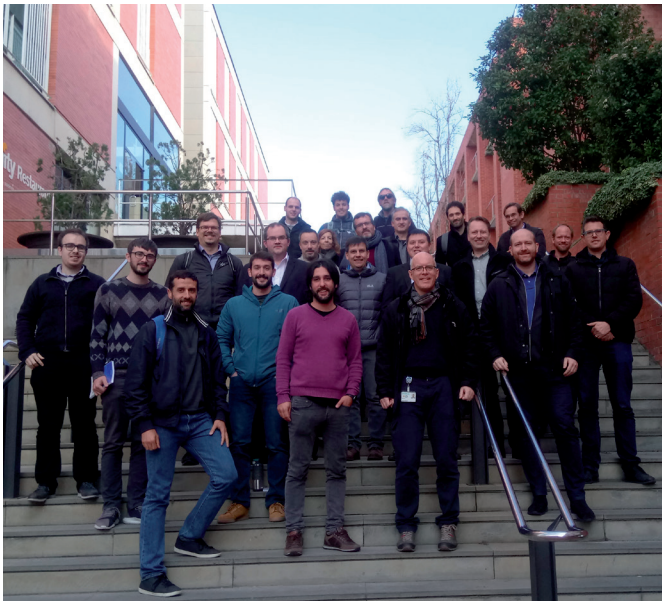
The most relevant features included in the SELENE platform are:

- SafeSU: a multicore, interference-aware statistics unit breaking down contention across contenders, supporting interference quotas, and measuring per-event-type maximum latencies.
- SafeDE: a module to enforce lightweight lockstepping across cores.
- SafeDM: a module to measure the diversity across cores running a task redundantly (See below for more information on these three safety-related hardware components).
- Rootvoter: a voting hardware module to provide N-modular redundancy support in the SELENE platform.

Artificial intelligence support

SELENE uses the open-source European Distributed Deep Learning (EDDL) library, created by the European Union-funded DeepHealth project, to deploy neural network (NN) models. Compatibility with other neural network frameworks is ensured using the Open Neural Network Exchange (ONNX) format. Accelerators are described in C and synthesized using commercial high-level synthesis tools capable of state-of-the art performance. Neural network accelerators are deployed in the platform for both application-specific integrated circuit (ASIC) and FPGA targets.

This work has received funding from the European Union's Horizon 2020 researchand innovation programme under grant agreement no. 871467.



The SELENE project consortium

Order out of CAOS: Safety-related open-source hardware modules

Jaume Abella, Barcelona Supercomputing Center (BSC)

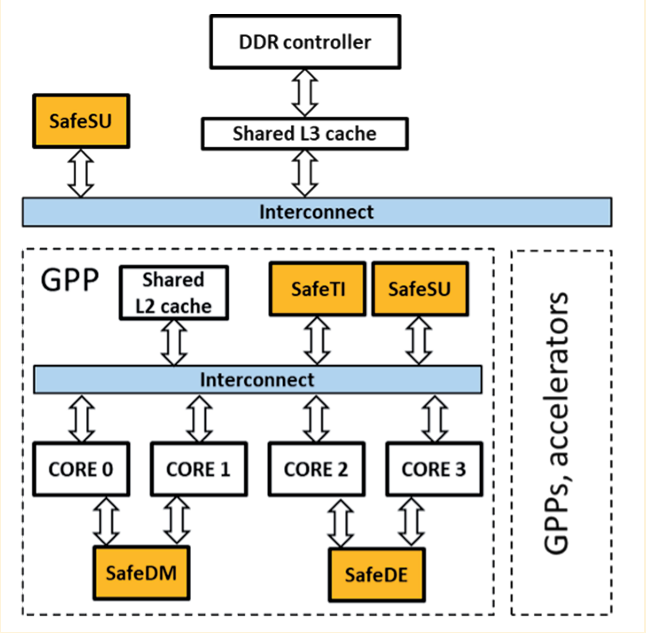
The increasing adoption of higher-performance processors in safety-related systems (such as those used in the space, avionics and automotive domains) has resulted in a need for appropriate hardware support to implement safety measures, as well as for system verification and validation (V&V). The advent of RISC-V in hardware design is ushering in a revolution in the field of open-source designs, and an opportunity to develop open-source processors for safety-related systems. However, this can only occur if those processors provide appropriate hardware support for safety.

In this context, the Computer Architecture and Operating System Interface (CAOS) group at BSC is developing a family of open-source hardware components intended for V&V of safety-related systems, as well as to implement safety measures. Those components, released with permissive open-source licences, include:

- SafeSU, a statistics unit to manage multicore timing interference
- SafeDE, a module enabling lightweight lockstepping across independent cores
- SafeDM, a monitor to measure the degree of diversity across independent cores

- SafeTI, a flexible and programmable traffic injector for platform validation

Some of these components can be downloaded via the GitHub link below.



Safety-focused hardware components developed by the CAOS group

Of these components, the SafeSU is being integrated and validated in a commercial system-on-chip (SoC) by Cobham Gaisler for the space domain reaching technology readiness level (TRL) 8 on FPGA by September 2022 as part of the Fast-Track-To-Innovation Horizon 2020 De-RISC project. All components (SafeSU, SafeDE, SafeDM, and SafeTI) are also being integrated as part of a prototype SoC for future product generations by Cobham Gaisler, reaching TRL 5 by November 2022 as part of the H2020 SELENE project. As part of the H2020 FRACTAL project, the SafeSU is being tailored to meet the needs of less-critical applications in the edge domain, and a software library providing the same functionality as that provided by SafeDE is also being developed and integrated (also to be offered as open source).

The goal of the CAOS group at BSC is to develop an even larger family of hardware and software modules enabling V&V of safety-related systems, as well as the implementation of safety measures, porting them to different protocols and interfaces (e.g. AMBA AHB, AXI4, etc.), and releasing them as open source with permissive licences. To promote their use and adoption, both in industry and academia, the CAOS group is open to a variety of collaborations including, among others, participation in EU project consortia, bilateral projects (e.g. for tailoring and integration), and providing support for integration by others. Similarly, the CAOS group is also open to retargeting their technology for other applications, such as providing security support, or testing high-performance computing (HPC) SoCs. Requests for collaboration and additional information can be directed to Jaume Abella who leads these activities at BSC.

jaume.abella@bsc.es

FURTHER INFORMATION:
BSC CAOS GitHub bsccaos.github.io

De-RISC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement EIC-FTI 869945. FRACTAL has received funding from the ECSEL Joint Undertaking (JU) under grant agreement no. 877056, and the MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR under project PCI2020-112010. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Italy, Austria, Germany, Finland, Switzerland.

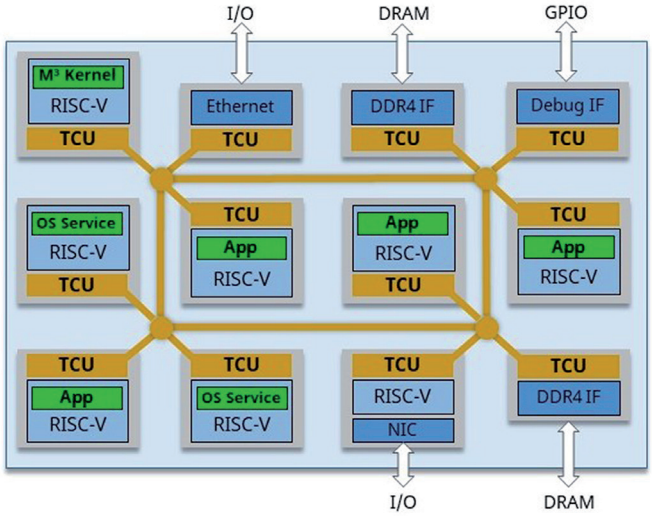
SECURING IOT DEVICES WITH OPEN-SOURCE HARDWARE AND SOFTWARE


Sebastian Haas and Nils Asmussen, Barkhausen Institut

We rely on smart systems and devices every day, most of which are connected via the internet. Their network forms the internet of things (IoT), which is becoming increasingly integrated into our everyday lives. These IoT devices must not only meet performance and energy-consumption requirements; at a time when attacks and data breaches are hitting the news almost daily, strong security and privacy properties are equally important. Providing these security and privacy properties requires not only addressing software problems, but also hardware vulnerabilities like Meltdown and Spectre, as recent years have shown.

Traditional multiprocessor architectures in IoT devices allow a modular system design and easy integration of different components into one system. However, hardware components are typically insufficiently isolated from each other. For example, an attacker with network access to the IoT device may exploit software and hardware bugs in a single component to get access to all other components in the system. This allows the attacker to manipulate the system or leak confidential data. In other words, the attacker can compromise the whole device.

The Barkhausen Institut is developing an open-source hardware / software co-design, called M³, that addresses security requirements at the hardware, operating-system, and application level. On the hardware side, M³ builds upon a tiled hardware architecture with physically separated tiles and a custom per-tile hardware component for cross-tile communication. This hardware component is called the trusted communication unit (TCU) which isolates tiles from each other and selectively allows communication.



The current hardware platform consists of multiple heterogeneous tiles which are connected by a network-on-chip. Tiles include open-source RISC-V Rocket cores or interfaces to input / output (I/O) peripherals and off-chip memory. The hardware platform is available as open source, which includes hardware register transfer level (RTL) code and a tool flow to compile and run the design on an field-programmable gate array (FPGA).

On the software side, M³ provides a microkernel-based operating system which is also available as open source. The microkernel runs on a dedicated 'kernel tile', whereas operating-system services (e.g. file systems and network stacks) and applications run on the remaining 'user tiles'. By default, all tiles are isolated from each other and only the kernel tile can establish communication channels between tiles. The microkernel is therefore programming the TCUs to enforce a desired security policy. For example, an application can access the file system only if permitted by the microkernel. This hardware / software co-design approach keeps malicious software and untrusted hardware components on the respective isolated tile and makes it much harder for an attacker to compromise the whole system.

FURTHER INFORMATION:

M³ open-source hardware available on GitHub
github.com/Barkhausen-Institut/M3-hardware

M³ open-source software available on GitHub
github.com/Barkhausen-Institut/M3

Talk by Sebastian and Nils at Computing Systems Week Lyon
bit.ly/CSWLyon_BI_video

COSSIM ADDRESSES SIMULATOR GAP FOR PARALLEL HETEROGENEOUS SYSTEMS


Nikolaos Tampouratzis and Yannis Papaefstathiou, EXAPSYS

Cyber-physical systems (CPS) and highly parallel and distributed computing systems – i.e. cloud and high-performance computing (HPC) systems – are growing in capability at an extraordinary rate, incorporating processing systems that vary from simple microcontrollers to high-performance units connected with each other through numerous networks. One of the main problems designers of such systems face is a lack of simulation tools offering realistic insights beyond simple functional testing, such as the actual performance of the nodes, accurate overall system timing, power / energy estimations, and network deployment issues.

In response, as part of the Horizon 2020 COSSIM project, we've developed the open-source COSSIM Simulation Framework. COSSIM efficiently integrates a series of sub-tools that model

the computing devices of the processing nodes as well as the network(s) utilized in the interconnections. It provides cycle-accurate results throughout the system by simulating together:

(a) the actual application and system software executed on each node, and

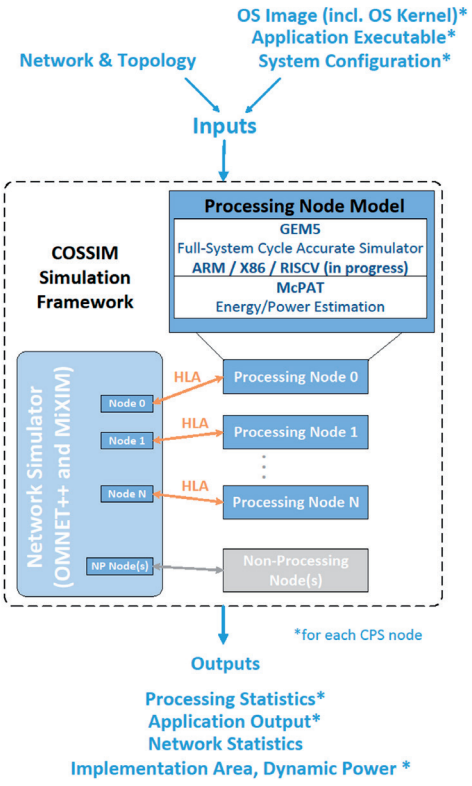
(b) the networks employed.

In so doing, COSSIM provides accurate performance / power / energy-consumption estimates for both the processing elements and the network(s) based on the actual dynamic usage scenarios. COSSIM also employs a standardized interconnection protocol between its sub-components, the IEEE Standard for Modeling and Simulation High Level Architecture – IEEE 1516.x HLA, meaning that it can be seamlessly connected to any other simulators (e.g. for simulating physical aspects).

The COSSIM Framework

COSSIM is built on top of several well-established simulators:

- **gem5**, a state-of-the-art full-system simulator, to simulate the digital components of each processing node in the simulated system
- **OMNeT++**, an established network simulator, to simulate the networking infrastructure
- **McPAT** to provide energy and power consumption estimations of the processing nodes and MiXIM (OMNET++ addon) to estimate the energy consumption of the network
- **CERTI HLA** architecture to bind the whole framework together



Framework features

- **Usability**: A unified Eclipse-based graphical user interface (GUI) has been developed to provide easy simulation setup, execution and visualization of results.
- **Support of multiple architectures**: The current version of COSSIM can support both ARM and X86 multicore central processing units (CPUs) in full-system mode (simulating a full Ubuntu-based operating system) using the latest gem5 version (v21.2.1). Over the next few months, the COSSIM framework will be extended in order to simulate the digital components of RISC-V systems, including their peripherals.
- **Extended component functionality**: Each component of the framework has been extended to provide advanced synchronization mechanisms and establish a common notion of time between all simulated systems.
- **Connectivity and expansion**: Through HLA and proper modifications to the basic components, the COSSIM framework can be connected to other tools to enable simulation of devices, events or physical processes (e.g. we are currently working on connecting COSSIM with the Ptolemy simulator)

Advancing the state of the art

To develop COSSIM, we took on the following research challenges:

(a) We developed a novel synchronization scheme supporting the notion of cycle accuracy throughout the sub-simulators.

(b) The synchronization scheme was augmented with the ability to trade off the simulation time against the required timing accuracy.

(c) The sub-simulators were adapted so that they could efficiently handle real network packets.

(d) Both the sub-simulators and the synchronization scheme were adapted to implement a novel fully distributed system where no critical task and / or sub-simulator is centralized. This means it can efficiently handle the simulation of thousands of processing nodes interconnected with any network technology.

FURTHER INFORMATION:

Access COSSIM in GitHub

github.com/H2020-COSSIM

COSSIM official publication

dl.acm.org/doi/pdf/10.1145/3378934

COSSIM has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 644042.

ACCELERATING MADE SIMPLER WITH CELERITY


Faced with a lack of software tools to help exploit accelerators for high-performance computing, Biagio Cosenza (University of Salerno) and Peter Thoman (University of Innsbruck) created Celerity. In this article, they explain how this open-source software can help developers get the most performance out of different kinds of processor.

In recent years, the use of accelerators such as graphics processing units (GPUs) has become widespread in high-performance computing (HPC) systems. However, while such accelerators offer performance gains, the accompanying software has sometimes been slow to catch up.

'Several of those of us working on Celerity have worked with GPUs for many years,' explains Biagio. 'What we saw was that existing technologies made an already difficult task – writing and maintaining efficient software for distributed compute clusters – even more challenging: now you not only needed to manage the distribution of data and work across cluster nodes, but also to GPUs on each individual node, generally using a completely separate technology. An example would be an MPI + CUDA hybrid program, or MPI + OpenCL if you planned to support vendor-agnostic technologies.'

'However, we also had previous experience with academic projects seeking to automate this entire stack and saw how ultimately they fell short of their goals,' adds Peter. 'So when SYCL™ was released as a vendor-agnostic, high-level standard for writing single-node applications targeting heterogeneous hardware, we asked ourselves whether it would be possible to extend it to clusters of GPUs and accelerators with minimal code changes. We had one key idea – the concept of range mappers – and Celerity was born.'

Celerity is an open-source project which allows users to scale applications to a cluster of accelerators without having to be experts in distributed memory programming. 'We based Celerity on the Khronos open standard SYCL, which makes it particularly suitable for data-parallel algorithms which can be effectively parallelized on GPUs,' says Biagio. 'Depending on the complexity of the underlying data structures, extending a single-node SYCL GPU program to multiple distributed memory nodes, with one or more GPUs each, might take little more than a namespace replacement and the use of a built-in range mapper for each kernel,' he adds.

Celerity is currently deployed on the Marconi-100 supercomputer at CINECA and the JUWELS supercomputer at Jülich, as well as a number of smaller clusters. 'It is being used in high-performance applications such as molecular docking in drug discovery, for example, as in the EuroHPC project LIGATE. Celerity is also used in a range of other international and industry activities, such as optical room-response simulations in computer vision,' notes Peter.

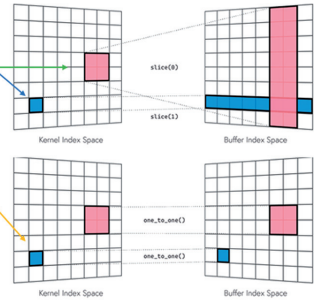
Open as standard

The researchers made Celerity open source to ensure it had staying power and practical impact. 'In our work with development tools and application programming interfaces (APIs) from academia, we noticed that, frequently, even well-designed, potentially useful tools can languish in obscurity or vanish without making much impact, due to a lack of support,' says Peter. 'To mitigate this issue, we are developing Celerity in the open, and run an extensive automated-testing and continuous-integration infrastructure that is publicly visible. With this approach, we hope to encourage more contributions and also provide a sense of security that the project is well supported, actively maintained and well tested on several architectures and backends.'

Participation in the SYCL working group is an important part of this work. 'By contributing to the working group, we can help shape the future direction of the standard and try to make sure that it remains a good target for Celerity requirements in particular and HPC use cases in general,' says Biagio.

LIGATE has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement no. 956137. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Italy, Sweden, Austria, Czech Republic, Switzerland.

```
queue.submit ([=](handler& cgh) {
  accessor a{mA,cgh,slice<2>(1),read_only };
  accessor b{mB,cgh,slice<2>(0),read_only };
  accessor c{mC,cgh,one_to_one{},write_only};
  cgh.parallel_for(range<2>(SIZE, SIZE),
    [=](item <2> itm) {
      float sum = 0.f;
      for(size_t k = 0; k<SIZE; k++)) {
        const float a_ik = a[{itm[0], k}];
        const float b_kj = b[{k, itm[1]}];
        sum += a_ik * b_kj;
      }
    c[item] = sum;
  });
});
```



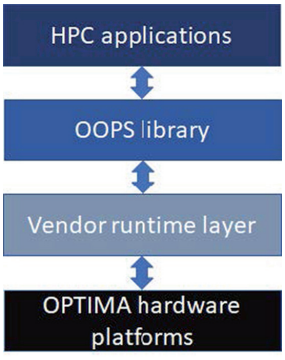
OOPS I DID IT AGAIN: PRODUCTIVE FPGA PROGRAMMING WITH OPTIMA'S OPEN-SOURCE LIBRARY

Panagiotis Miliadis, Chloe Alverti, Dimitris Theodoropoulos and Dionisios Pnevmatikatos, Institute of Communication and Computer Systems, National Technical University of Athens

Field-programmable gate arrays (FPGAs) can provide significantly higher performance than other processors when executing certain high-performance computing (HPC) applications. However, the difficulties in programming, interconnecting and handling them are well known.

To address this problem, the EuroHPC JU OPTIMA project is developing open-source libraries for FPGA-based HPC systems to deliver a significantly higher performance-to-energy ratio than that in existing HPC systems, including those made up of low-power central processing units (CPUs) like ARM and / or graphics processing units (GPUs). In addition, OPTIMA will provide guidelines and open-source reference designs, eventually allowing third parties to port applications to FPGA-based platforms in a similar way to porting to systems using GPUs and / or manycores today.

The OPTIMA OPen Source (OOPS) library will provide optimized software routines for a range of key operations in scientific and industrial applications, including vector operations, linear and differential equations, and matrix multiplications. OOPS will dramatically reduce the effort of mapping primitive computational kernels onto OPTIMA's reconfigurable logic, leading to faster execution time and improved energy efficiency. The library will be integrated into the OPTIMA toolflow (see figure below), enabling seamless utilization of the available hardware resources by software developers.



OOPS intercommunication with the OPTIMA toolflow

Leveraging the device vendor runtime layer, OOPS kernels will transfer data from the host processor to the hardware kernels, initiate and monitor data processing and send output results back to the application layer. The OOPS library set will implement a large subset of the basic linear algebra subprograms (BLAS), the building blocks for standard matrix and vector operations. It will also implement a sparse matrix-vector (SpMV) multiplication kernel and a subset of the Portable, Extensible Toolkit for Scientific Computation (PETSc) suite.

OOPS will expose a standard C-based application programming interface (API) in the form of function prototypes towards the application layer, meaning that developers will be able to integrate the library simply by including its API header in their software code. OOPS can easily be integrated or combined with existing frameworks, such as Parallelware, developed by Appentra, or GASPI, created by Fraunhofer. Other widely used programming languages like Python will also be evaluated.

Next steps

OPTIMA will shortly publish a repository containing:

- hardware implementation of the supported kernels
- host processor helper routines for easy deployment onto the FPGA
- example projects with instructions on how developers can use the OOPS library

OPTIMA will adopt a continuous integration / continuous deployment (CI / CD) approach for continuous updates and maintenance of the library, to ensure computer-aided design (CAD) tool compatibility and FPGA device support, as well as optimal performance with respect to the available resources.

FURTHER INFORMATION:

OPTIMA website optima-h2020-eu

BLAS netlib.org/blas

Balay S., Gropp W.D., McInnes L.C., Smith B.F. (1997) 'Efficient Management of Parallelism in Object-Oriented Numerical Software Libraries'. In: Arge E., Bruaset A.M., Langtangen H.P. (eds) Modern Software Tools for Scientific Computing. Birkhäuser, Boston, MA.

OPTIMA has received funding from the EuroHPC JU under grant agreement no. 955739. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Greece, Germany, Italy, the Netherlands, Spain and Switzerland.

In the latest in our series on red-hot deep-tech companies, Martin Croome, vice president of marketing, gives us an introduction to ultra-low-power GAP processors from GreenWaves Technologies.

GreenWaves

Enabling state-of-the-art machine learning and digital signal processing on energy-constrained devices

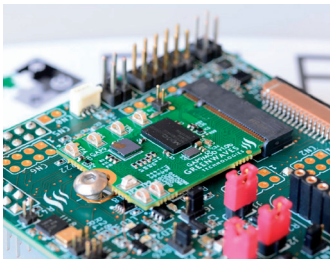
COMPANY: GreenWaves
MAIN BUSINESS: semiconductor design, systems-on-chips
LOCATION: Grenoble, France
WEBSITE: greenwaves-technologies.com

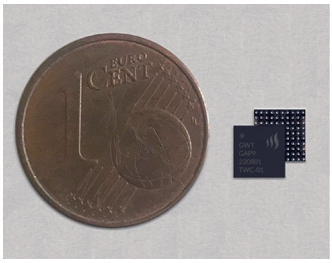
Located in the micro-and nano-electronics hotspot of Grenoble, GreenWaves is a fabless semiconductor startup that was founded in 2014. GreenWaves designs and brings to market advanced ultra-low-power artificial intelligence (AI) and digital signal processing (DSP) processors for energy-constrained applications. Energy-constrained devices generally have to last for years on one battery – as in the case of IoT sensors – or have a very small battery, as we see with hearable or wearable products.

GreenWaves' chief technology officer, Eric Flammand, was one of the founders of the Parallel Ultra Low Power (PULP) Platform that was started as a joint effort between the Integrated Systems Laboratory (IIS) of ETH Zürich and the Energy-efficient Embedded Systems (EEES) group of the University of Bologna in 2013 to explore new and efficient architectures for ultra-low-power processing. GreenWaves collaborates with PULP members on high-performance, low-power architectures – such as the PULP cluster – and transforms the results into real products. As reported in *HiPEACinfo* 58, the company's feedback has been invaluable in bringing PULP's technology to a higher level of maturity.

PULP itself uses processor cores that implement the open-source RISC-V instruction set architecture (ISA). GreenWaves is a key contributor to the PULP project and has staff on two committees of RISC-V International . It has been one of the first companies to focus on the 'very edge' and to leverage an open-source RISC-V core at large commercial scale.

GreenWaves' first product, the 55nm GAP8 processor, was one of the very first commercially available RISC-V processors and artificial intelligence (AI) microcontrollers. It allows massive deployment of low-cost, battery-operated intelligent devices that





capture, analyse, classify and act on a fusion of rich data sources such images, sounds, radar signatures and vibrations.

GreenWaves' second generation GAP9 processor enables a market-leading audio experience for hearable devices through ultra-low-power implementation of features such as neural-network-steered, ultra-low-latency, active noise cancellation, neural-network-based noise reduction and 3D sound in hearable devices such as true wireless stereo earbuds.

GreenWaves' GAP9 processor has been designed from the ground up to address a blend of classic DSP, ultra-low latency, sample-by-sample time-domain DSP and neural-network workloads while preserving a high degree of flexibility and programmability. While GAP9 shares some system components with traditional microcontroller units (MCUs), its architecture is quite unique, enabling a revolution in the performance of extremely energy-constrained devices.

We have had customers developing on simulations and field-programmable gate array (FPGA) versions of GAP9 for the past two years and on the real chip since February. Our customers have found the GAP9 hearable platform exceptionally power efficient for voice and music processing. It has given them headroom in both energy and processing power that they use to develop innovative new features in their products with no compromise in area, cost or energy. Our development tools enable software and hardware developers to productively harness the power of their processors using familiar neural-network and mathematical computing software. As one customer told us, 'we do in weeks with GAP9 what we do with other platforms in months'.

The European Commission actively promotes open science to facilitate the sharing of findings from publicly funded scientific research. In this article, HiPEAC's Xavier Salazar (Barcelona Supercomputing Center) explores the ramifications of open science in Europe, including its role in the digital transformation.

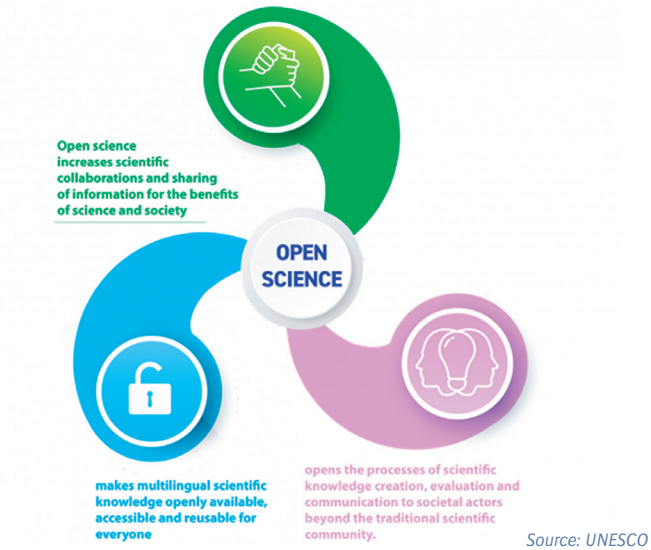
How open is open?

The term 'open' has become something of a buzzword, so it's worth taking a moment to consider what it really means. Here are some definitions:

UNESCO Recommendation on Open Science

The UNESCO Recommendation on Open Science was adopted by the 41st session of the UNESCO General Conference in November 2021. The recommendation, an international framework for open science policy, defines open science as follows:

> "*an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community. It comprises all scientific disciplines and aspects of scholarly practices, including basic and applied sciences, natural and social sciences and the humanities, and it builds on the following key pillars: open scientific knowledge, open science infrastructures, science communication, open engagement of societal actors and open dialogue with other knowledge systems*"



Source: UNESCO

The recommendation goes on:

> "***Open scientific knowledge*** *refers to open access to scientific publications, research data, metadata, open educational resources, software, and source code and hardware that are available in the public domain or under copyright and licensed under an open licence that allows access, re-use, repurpose, adaptation and distribution under specific conditions, provided to all actors immediately or as quickly as possible regardless of location, nationality, race, age, gender, income, socio-economic circumstances, career stage, discipline, language, religion, disability, ethnicity or migratory status or any other grounds, and free of charge.*
> - ***Scientific publications***
> - ***Open research data***
> - ***Open educational resources***
> - ***Open source software and source code***
> - ***Open hardware***"

Open science in the European Union (EU)

The European Commission has developed its own definition of open science, as follows:

> "*An approach to the scientific process that focuses on spreading knowledge as soon as it is available using digital and collaborative technology.*"

The European Commission has developed tools to monitor different aspects of open science, as can be seen on its 'Open science monitor' webpages (see 'Further reading', below). Metrics to monitor progress include **open access to publications** – defined by UNESCO as 'free access to information and unrestricted use of electronic resources for everyone' – and **open research data**, which the European Commission defines as 'the data underpinning scientific research results that has no restrictions on its access, enabling anyone to access it'.

In 2019, the European Commission's DG CONNECT commissioned a study to analyse the economic impact of **open-source software**

and hardware on the European economy. This found that 'open source software and hardware are key for the region's digital transformation and can be a major boost to the EU's GDP'. It estimated that open-source software contributes between €65 and €95 billion to the European Union's gross domestic product (GDP) and promises significant growth opportunities for the region's digital economy.

In addition, the study identified strengths, weaknesses, opportunities and challenges of open source in information and communication technology policies, including cybersecurity, artificial intelligence, 'Digitising European Industry', the connected car, high-performance computing, and distributed ledger technologies. In terms of open hardware, it discusses a number of initiatives and success stories, including the Open Compute Project, RISC-V and SiFive, and highlights the following European centres of academic excellence: the University of Bologna, Barcelona Supercomputing Center and ETH Zürich.

Noting the 'clear signals from investors on the huge value and potential of open source', the authors concluded that: 'The main breakthrough of the study is the identification of open source as a public good. This shows a change of paradigm from the previously irreconcilable difference between closed and open source, and points to a new era in which digital businesses are built using open source assets.' As this report makes clear, the use of open source has an overall impact on the economy, society, projects and individuals.

Building upon this analysis, the report offers a series of recommendations, including building capacity with open-source offices, providing more funding for research and development related to open source, and providing incentives to upload code to publicly accessible EU repositories.

As for the lower levels of the stack, open hardware (or open-source hardware) is a term for any physical artefact – machines, devices, or any other physical object – whose design is made publicly available through open licences. These licences specify the freedoms a licensee can exercise in studying, modifying, distributing, making and commercializing the hardware. Some basic concepts are discussed in the 'On licences for [Open] Hardware' paper (see 'Further reading', below).

The table below summarizes the main differences between open-source software and hardware.

Open Source Definition (v.1.9)	**Open Source Hardware Definition (v.1.0)**
Source code	**1. Documentation** **2. Scope** **3. Necessary software**
Derived works	4. Derived works
Free distribution	5. Free distribution
Integrity of the author's source code	**6. Attribution**
No discrimination against persons or groups	7. No discrimination against persons or groups
No discrimination against fields of endeavor	8. No discrimination against fields of endeavor
Distribution of license	9. Distribution of license
License must not be specific to a product	10. License must not be specific to a product
License must not restrict other software	11. License must not restrict other **hardware** or software
License must be technology-neutral	12. License must be technology-neutral

Differences between open-source software / hardware definitions

Source: Open hardware licences - Parallels and contrasts: open science monitor case study

With the value of open source becoming increasingly evident, along with its utility as a driver of new products and services, it will be interesting to see how this model continues to shape the technology arena in Europe and beyond.

FURTHER READING:

UNESCO Draft Recommendation on Open Science
bit.ly/UNESCO_open_science

Open Science | European Commission
bit.ly/European_Commission_open_science

Open science monitor | European Commission
bit.ly/EC_open_science_monitor

European Commission, Directorate-General for Communications Networks, Content and Technology, Blind, K., Pätsch, S., Muto, S., et al., The impact of open source software and hardware on technological independence, competitiveness and innovation in the EU economy: final study report, Publications Office, 2021
data.europa.eu/doi/10.2759/430161

Montón, Màrius & Salazar, Xavier. (2020). 'On licenses for [Open] Hardware.'
bit.ly/ResearchGate_concepts_open_hardware

European Commission, Directorate-General for Research and Innovation, Kauttu, P., Murillo, L., Pujol Priego, L., et al., Open hardware licences: parallels and contrasts: open science monitor case study, Publications Office, 2019
data.europa.eu/doi/10.2777/641658

High-performance computing (HPC) has the potential to revolutionize European industry and improve quality of life for citizens. In these case studies from the EuroCC project, Tomáš Karásek describes how the Czech National Competence Centre in HPC helps turbocharge local industry applications.

HPC to the rescue

Reducing pollution and optimizing exoskeletons: EuroCC case studies

SIMULATIONS TO REDUCE EMISSIONS

Czech energy company ORGREZ provides services such as electricity production and distribution, and fuel energy conversion. Today, it also focuses on greenhouse-gas issues, including air protection technologies and systems for monitoring, evaluating, regulating and reducing pollutant emissions, especially CO_2, NOx, and ash.

The Czech National Competence Centre worked with ORGREZ to determine whether computational fluid dynamics (CFD) simulations could describe the process of selective catalytic reduction (SCR), an emissions control technology, and whether such simulations could help design this technology. SCR is one of several DeNOx technologies used to reduce the concentration of nitrogen oxides in the exhaust gases of combustion plants to legal levels.

The SCR model was tested on the geometry of a coal combustion boiler and design data specified by the catalyst manufacturer, with a waste incineration boiler selected and municipal waste as the fuel. The CFD simulation resulted in a reduction in NOx concentration, additional oxidation of CO to CO_2, and conversion of SO_2 to SO_3.

Currently, SCR plants are designed using experience and simplified calculations. Using numerical modelling and simulations will make the process faster and more efficient, allowing subsequent design optimization. Thanks to HPC, these simulations can be completed relatively quickly.

OPTIMIZING EXOSKELETONS

Bringing together an international team of engineers, doctors and physiotherapists in the Czech Republic, MEBSTER is a research and development company that develops innovative, cost-effective assistive devices for clients with mobility disorders. Undertaking patient-centred research, the team designs their devices with users, rigorously testing products to ensure they are as simple and comfortable as possible.

MEBSTER and the Czech National Competence Centre joined forces to demonstrate numerical modelling and simulation in the design of a UNILEXA exoskeleton for gait assistance, created for people with partial or complete loss of lower limb function. The computational model is based on the finite element method (FEM). Since exoskeleton assembly is a mathematically complex, nonlinear problem with a wide variety of boundary conditions, HPC is necessary to solve it.

A simplified model of the UNILEXA exoskeleton assembly was developed to estimate the required HPC resources. Next, the Barbora supercomputer was used to solve a complex numerical model, including nonlinearities such as contact interfaces with friction, large displacements, etc.

Thanks to the simulation, researchers were able to verify the safety and effectiveness of the project, and increase user comfort. This resulted in a more competitive product on the market.

EuroCC has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement no 951732. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Germany, Bulgaria, Austria, Croatia, Cyprus, the Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, the United Kingdom, France, the Netherlands, Belgium, Luxembourg, Slovakia, Norway, Switzerland, Turkey, Republic of North Macedonia, Iceland, and Montenegro. This project has also received funding from the Ministry of Education, Youth and Sports of the Czech Republic (ID: MC2101).

Most unstructured mesh databases are not designed for direct usage with highly parallel numerical solvers. Here, Ondřej Meca, Lubomír Říha, Branislav Jansík, and Tomáš Brzobohatý (IT4Innovations National Supercomputing Center, VSB – Technical University of Ostrava) describe how they created a tool named MESIO, which provides a scalable interface for effortless unstructured mesh processing by massive parallel libraries.

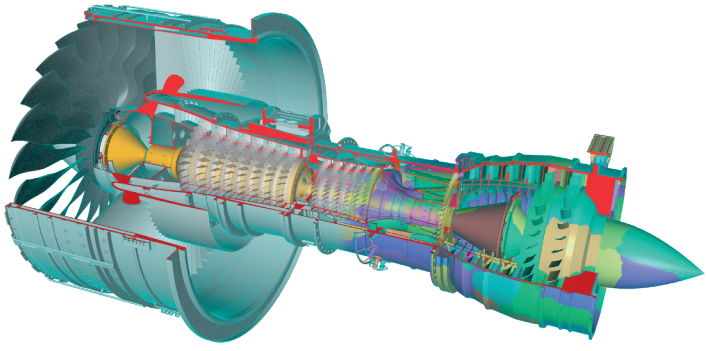
MESIO: Highly parallel loading of unstructured meshes

Ondřej Meca, Lubomír Říha, Branislav Jansík, and Tomáš Brzobohatý

Unstructured meshes are commonly used as a model description in numerical methods such as the finite element method, finite volume method, etc. Decades of development have established a rich set of database formats used by theoretical researchers and mainstream engineers. Formats were created in response to the requirements of a given tool, and many of them were made before the extensive development and adoption of HPC technologies.

This means that many are subobtimal for use in high-performance computing (HPC), as formats were optimized mainly for sequential tools, and they are not prepared for reading by massively parallel libraries. However, adoption of more HPC-centric formats for mesh databases is rare, as many engineering professionals understandably prefer to use well-known, tried-and-tested modelling tools for the most critical part of the design pipeline: the creation of high-quality numerical models.

Technically, parallel libraries usually require a particular data organization within a database that allows simple parallel reading with minimal communication and synchronization steps. If needed, it is relatively easy to build a sequential converter that transforms a sequential input database into a selected parallel database. Unfortunately, no tool can convert a general database in parallel, as it dramatically slows down and sometimes even inhibits (e.g. due to memory limits) the connection between tools that generate sequential mesh databases and parallel solvers. This is one of the key factors contributing to the low interest in HPC resources from the mainstream engineering community.



Unstructured meshes are used for HPC simulations

To bridge professional tools for the creation of complex engineering models and massively parallel libraries developed for the HPC environment, we created the MESIO tool. This open-source tool contains an algorithm that avoids all sequential parts of the conversion process and is general enough to be able to load an unstructured mesh from an arbitrary database format.

Using MESIO allows users to obtain the fully parallel loader with a provable linear speedup for large databases independently of any particular data arrangement and data duplication within a database. Since the tool can reconstruct a mesh with hundreds of millions of elements in seconds, it is possible to completely skip conversion to a suitable parallel database, as differences between loading the original (sequential) database and a parallel one are negligible. Hence, users can use the same database file regardless of whether they want to use it on a workstation or a cluster with thousands of computational nodes, even if the format was not designed for HPC technologies. This significantly simplifies and accelerates design pipelines that combine high-quality sequential tools to create numerical models and parallel libraries, allowing fast prototyping due to their ability to utilize modern HPC clusters fully.

MESIO was published in the prestigious journal *Advances in Engineering Software* and is available under the terms of the Berkeley Source Distribution (BSD) License.

FURTHER INFORMATION:

Access MESIO via Github github.com/It4innovations/mesio

Ondrej Meca, Lubomír Říha, Branislav Jansík, and Tomáš Brzobohatý: Toward highly parallel loading of unstructured meshes. Advances in Engineering Software, Volume 166, 2022.
doi.org/10.1016/j.advengsoft.2022.103100

Innovation Europe

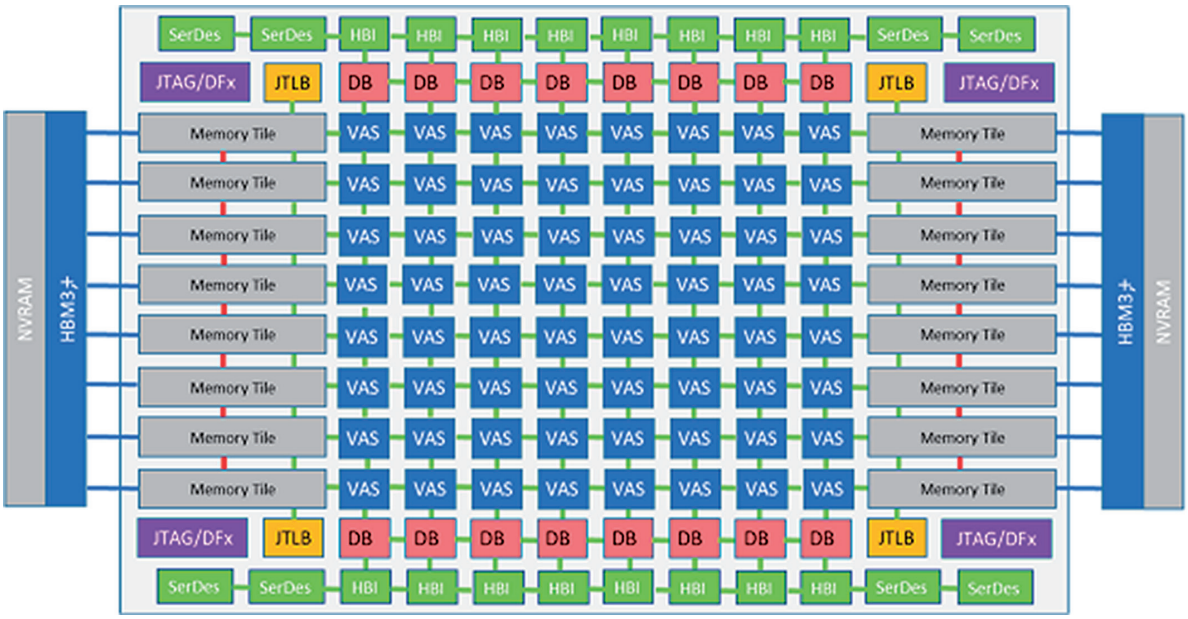
RISC-V® special

MEEP MEEP: SPEEDING UP EXASCALE ARCHITECTURE DEVELOPMENT



Launched in 2020, the MareNostrum Experimental Exascale Platform (MEEP) is a EuroHPC Joint Undertaking project exploring architectural options for future-generation high-performance computers, such as the MareNostrum supercomputer hosted by MEEP coordinator Barcelona Supercomputing Center (BSC). Open source is pivotal to the project, which makes use of the open-source RISC-V instruction set architecture (ISA). We caught up with MEEP coordinator John D. Davis and research engineers Teresa Cervero and Xavier Teruel, all based at BSC, to find out more.

'To meet the power and performance demands of today's high-performance computing (HPC) systems, we need hardware and software co-design,' explains John. 'It is difficult to develop the software and test the hardware before the chips have been made. This leads to a serial design process, and, consequently, very high risk on the hardware and software development paths.'



The ACME architecture

But what if you could play around with hardware and software combinations before committing to a physical product? 'MEEP was conceived as a digital laboratory to enable pre-silicon validation of hardware architectures and a software development vehicle. This means that MEEP can both validate hardware architectures, reducing the complexity and cost of post-silicon validation, and provide a hardware environment for the development of the associated software,' explains John.

The idea is to provide a foundation for developing European-based chips within the diverse European HPC research landscape, notes John. MEEP builds on work undertaken in other projects, such as the vector processor developed in the European Processor Initiative, while the aim is for the platform to be used by projects like the eProcessor and the EUPilot (see overleaf).

Innovative experimental features

There are three main components to this 'laboratory'. 'First, MEEP is composed of field-programmable gate arrays (FPGAs), which can be used to emulate other hardware devices. This allows hardware designs to be mapped onto the FPGA and utilize standard memory and input / output (I/O) interfaces, as required by many central processing units (CPUs) and accelerators,' says John. 'Second, we propose a research accelerator hardware architecture – the Accelerated Compute and Memory Engine (ACME) – as a demonstrator for pre-silicon validation. Finally, we are using HPC and high-performance data analytics (HPDA) workloads in the co-design process.'

Thanks to these innovations, MEEP facilitates groundbreaking new approaches to exascale supercomputers, allowing experiments at all levels of the computing stack. 'ACME's accelerators are characterized by their tight integration with the core, their support for vector processing, their support for systolic arrays and by being self-hosting. Putting this all together dramatically shifts the balance between traditional host processors and accelerators in HPC systems. With self-hosting accelerators, the role of host processors is greatly diminished, and more resources (i.e. silicon) can be devoted to the highly energy-efficient accelerators, resulting in improved energy efficiency,' explains John.

MEEP also provides a communication wrapper for the FPGAs, the MEEP shell, which helps the developer to avoid having to deal with the complexities of the host-accelerator communication, and a novel open-source simulator for multi-core RISC-V architectures, called Coyote.

Openness as standard

'While open source has established itself as a crucial factor in the ecosystem, most of the contributions so far have related to software, from applications to operating systems,' says Teresa. 'Collaboration is key to the success of open-source software initiatives, with the community playing an active and important role. Hardware is not yet at that stage, due to factors such as maturity, robustness, reliability and vendor lock-in.'

To help overcome this, MEEP wants to help bridge the gap between the hardware layers of the ecosystem, opening the door to contributions from the community and providing a mechanism for future developments, according to Teresa. MEEP's contributions include enabling a whole infrastructure to develop, test and validate hardware designs; developing tools to make this possible; and offering open-source intellectual property (IP) for academic development.

The project is working with RISC-V architectures, in some cases using them as they are, and in others contributing to the ecosystem with new proposals. 'This gives us the opportunity to leverage open-source resources at the architecture and system software level, while at the same time developing aspects of our vision that will allow us to specialize and differentiate,' says Teresa.

As for software, MEEP is assembling a Linux distribution, based on Fedora, which includes several frameworks targeting specific project objectives, including Podman, TensorFlow Lite, Apache Spark and COMPSs, says Xavier. 'The project is also contributing forked versions of the LLVM compiler, implementing several aspects of the compiler and runtime. These include a custom in-house systolic array ISA extension, a prototype implementation of the OpenMP spread construct, which has been submitted to the OpenMP standard, and improved code-generation algorithms based on automated loop transformations.' In addition, the project will contribute to new versions of the RISC-V benchmark suite developed at BSC, providing BLIS porting, OpenMP and bare-metal versions of a subset of its benchmarks (SpMV, GEMM, AXPY, Somier and FFT), adds Xavi.

FURTHER INFORMATION:
meep-project.eu

MEEP has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement no. 946002. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Croatia, Turkey.

MEEP
MareNostrum Experimental Exascale Platform

ACCELERATING EUROPEAN EXASCALE: THE EUROPEAN PILOT PROJECT

As part of the European Union's investments in the RISC-V ecosystem, The European PILOT project (EUPILOT) is creating accelerators –designed, implemented, manufactured, and deployed in Europe – to power pre-exascale systems. HiPEAC caught up with EUPILOT coordinator Carlos Puchol to get the lowdown on the project.

THE EUPILOT

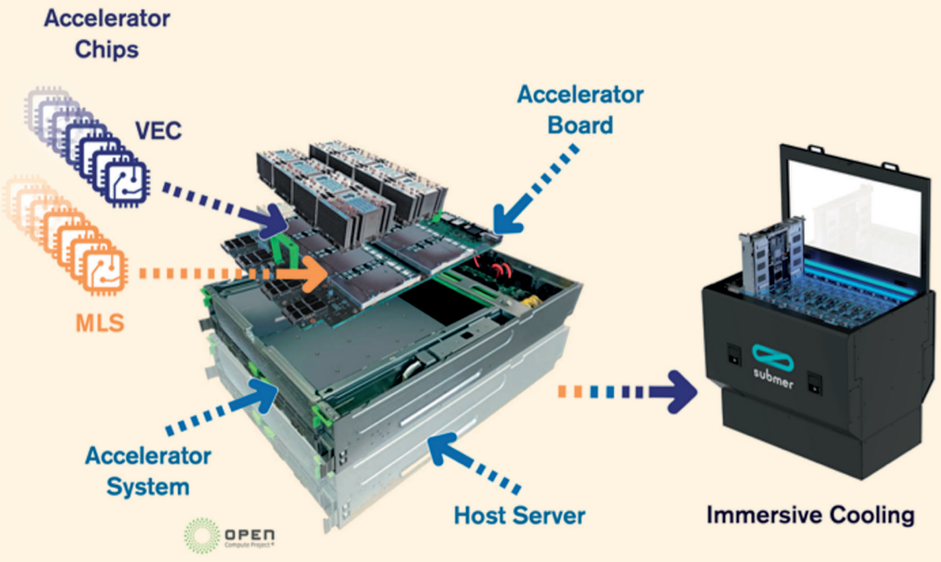
Thanks to Rotary International and Fulbright graduate scholarships, Carlos Puchol studied computer science at the University of Texas at Austin, before carrying out research at Bell Labs and later Transmeta, whose employees included Linux founder Linus Torvalds. After working in semiconductors for some years he founded the startup Amahi, an open-source multimedia server Linux distribution.

Recently, Carlos returned to his native Spain, where he is now coordinating the European PILOT project at Barcelona Supercomputing Center (BSC). 'The EUPILOT project aims to build an end-to-end demonstrator of accelerators that could be used in a pre-exascale system,' says Carlos. 'Making full use of European and open-source technologies and standards, the project will produce three chip tapeouts. The first will be a test chip to validate the use of the 12nm technology node. The second and third, developed concurrently, will contain a vector accelerator with up to 16 cores and a machine learning and stencil accelerator with up to eight cores, respectively,' he adds. Mounted in groups of up to four dies with low-power double data rate (LPDDR) memory, the chip modules will be installed into accelerator boards going into systems and, when paired with host servers, deployed into liquid immersion tanks supporting ultra-efficient power densities.

To carry out this work, EUPILOT has assembled a consortium of 19 partners across Europe, 'from well-known research institutions to smaller companies pushing the boundaries of tech', says Carlos. BSC leads the software effort, with contributions from Tubitak, ETH Zürich, FORTH and Chalmers. Semidynamics leads the hardware, architecture and verification aspects, with contributions from Extoll, FORTH and Fraunhofer. When the chips come back from the fab, the hardware /systems team will build the circuit boards (EXAPSYS) and systems (2CRSi), integrate them, and immerse them in liquid immersion cooling tanks (Submer). Other partners contributing to the project are Leonardo, the University of Bologna and others associated with the Italian information technology consortium CINI, KTH Royal Institute of Technology, CEA, Jülich, TU Kaislerslautern and the University of Zagreb.

Open path to EU tech sovereignty

Open-source technologies are pivotal to the success of EUPILOT, according to Carlos. 'This project aims to help Europe achieve technological independence. By enabling faster integration and deployment, open-source technologies and standards help us to reach that goal faster.' Indeed, EUPILOT uses open-source elements in everything from software (operating systems, libraries, applications) to hardware (intellectual property (IP), tools, modules, boards, systems) as far as possible, in every area of the system.



The project also represents a step forward in terms of European technological sovereignty, according to Carlos. 'EUPILOT contributes to a sustainable exascale HPC ecosystem in Europe, helping lay the groundwork for long-term technical independence by delivering an end-to-end proof of concept, from chips to advanced datacentre deployments,' he says. 'These European IP accelerators and the customized software ecosystem will demonstrate a path to exascale levels of performance at an unparalleled scale of integration. The know-how to build these supercomputers, the boost in industrial competitiveness and closer cooperation will all help establish European digital autonomy.'

To do so, the project will build on the work of other EU-funded initiatives, explains Carlos. 'Hardware-wise, EUPILOT leverages and significantly scales up advancements made within the European Processor Initiative (EPI), such as the EPI Accelerator (EPAC), in the form of the massively parallel arrangement of the HPC vector, machine learning and stencil accelerators.' The project also has close ties to MEEP (see previous pages), which is providing infrastructure and tools to simulate and emulate the accelerators and provide a software development vehicle for new hardware features in the cores, as well as multicore and system-level environments for the EUPILOT accelerator chips.

FURTHER INFORMATION:
eupilot.eu

The European PILOT project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement no. 101034126. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Italy, Switzerland, Germany, France, Greece, Sweden, Croatia and Turkey.

EPROCESSOR RISCS A MADE-IN-EUROPE CPU

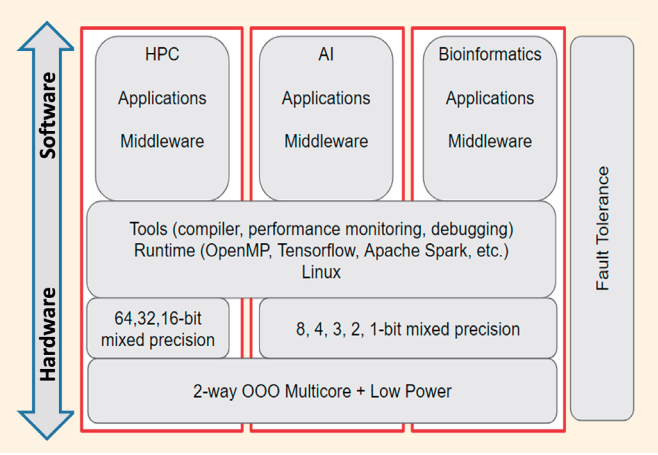
While a number of EU-funded projects focus on RISC-V accelerator development, the eProcessor project aims to create a fully fledged central processing unit (CPU) based on RISC-V. This open-source out-of-order processor core, plus accompanying accelerator and software stack, will be made in Europe. HiPEAC caught up with eProcessor coordinator Nehir Sönmez (Barcelona Supercomputing Center) to find out how the project will contribute to the RISC-V ecosystem and provide a free alternative to Intel, AMD and ARM-based designs.

For Nehir Sönmez, a senior researcher at Barcelona Supercomputing Center (BSC), the transparency of open source is what makes it so important for research. 'The code is visible for everyone: you can test it, alter it, and make it visible again,' he says. 'This is essential for research on modern computer architectures: you need clearly understandable baseline implementations on which you can build and experiment.'

Performing accurate research is currently hindered by the hermetic approach of proprietary hardware companies, explains Nehir. This means researchers are limited to unsatisfactory workarounds, such as using complex software-based architectural simulators like gem5, or making educated guesses to describe the hardware using schematics, hardware-description languages (HDL) or high-level synthesis (HLS) tools. 'However, even if you know how to implement a certain part, it may be illegal to do so because there is a patent on it,' adds Nehir, citing MIPS' response to the Yellow Star core and the inability of the Plasma MIPS core to fully comply with specifications until the patent expired as examples. Beyond research, Nehir believes that the use of open-source hardware can have a positive impact on society by shedding light on the inner workings of computers that people use every day.

In this sense, RISC-V has provided a welcome alternative to black-box solutions, says Nehir. 'RISC-V stemmed from an academic instruction set architecture definition, straight out of our computer architecture textbooks. Simple and efficient, it quickly attracted a large community of supporters, with many of the early efforts in Europe led by ETH Zürich through their PULP platform,' he adds (see pp.24-25). 'Actually, RISC-V builds on a tradition of dedicated open hardware efforts,' says Nehir, 'such as the OpenCores community and OpenRISC (c. 2000), and their heroic efforts such as the Zet X86 open implementation'.

Using open-source technologies can help Europe make inroads into the technological lead established by the world's technology superpowers by attracting more people to pitch in, says Nehir. 'Of course, the industry has a huge competitive advantage and decades of experience producing hardware. When it comes to actually producing open-source hardware, there are also the prohibitive costs associated with professional vendor tools and with taping out chips. That said, there is significant effort involved in providing open-source tools for hardware design and synthesis,



while there are also efforts such as the Google-Skywater process design kit to help bring down tapeout costs,' he notes.

Direct memory access

To contribute to the open-hardware technology ecosystem, eProcessor will deliver an innovative architecture, including an out-of-order central processing unit (CPU) – the execution paradigm commonly used for high-performance computing (HPC) – along with a single, unified vector, bioinformatics and artificial intelligence (AI) accelerator that can access the memory hierarchy directly. The architecture also includes an off-chip link, allowing the incorporation of additional CPUs and / or accelerators in a cache-coherent way.

As for applications, the project is focusing on three main domains: HPC, artificial intelligence (AI) and bioinformatics. 'In HPC, we are targeting the NAS parallel benchmarks, European Processor Initiative (EPI) RISC-V benchmarks and the vector benchmark suite to measure performance,' explains Nehir. The project's AI use cases include the 'smart mirror' developed at the University of Bielefeld and enhanced in EU-funded projects including LEGaTO, while Thales will use the technology to power a border surveillance application.

Nehir presenting eProcessor at HiPEAC's Computing Systems Week in Tampere

In bioinformatics, the focus will be on performing secondary analysis of genomic data, using the FM-index, Smith-Waterman, Smith-Waterman-Gotoh and Wavefront Alignment algorithms. For AI applied to bioinformatics, the project will draw upon the EU-funded DeepHealth toolkit: the European Distributed Deep Learning Library (EDDLL) and the European Computer Vision Library (ECVL). In all cases, the project follows 'a software / hardware codesign methodology, while producing two ASIC chips (single-core and multicore), where we configure the processor and accelerators according to the best metrics, such as performance or power, while running these applications in a feedback loop,' says Nehir.

As for the future of open-source hardware in Europe, Nehir points out that RISC-V is still strongest in the embedded domain: 'We need to properly cover the mid-range before delving into high-end systems.' The software layer is crucial in this aspect. 'More applications, runtimes and compilers need to support RISC-V. With Hi-Five Unleashed/Unmatched we are slowly starting to see an off-the-shelf RISC-V system with full Linux support, and all kinds of common user applications, a stable graphic user interface (GUI), web browsers, word-processing, spreadsheets, games, and so on. Only once this has become commonplace can we healthily push forward to providing high-end systems,' he cautions.

Beyond simply processing data fast, HPC also poses more challenges in terms of network interconnection and storage capabilities. 'All of these need to keep pace with one another, in order for it all to make sense,' says Nehir. 'If we are providing heterogeneity, at what level will that be done: node or cluster? How are we balancing scaling up and scaling out? What do we do with the smarts: the smart network interface card (NIC), smart storages, processing in memory, etc? How will we deal with the upcoming novel technologies, such as photonics?' However, while there is still a long way to go, eProcessor aims to be another major step in the right direction, he adds.

FURTHER INFORMATION:

eprocessor.eu

Yellow Star core brej.org/yellow_star

Plasma MIPS core github.com/adrianj/Plasma

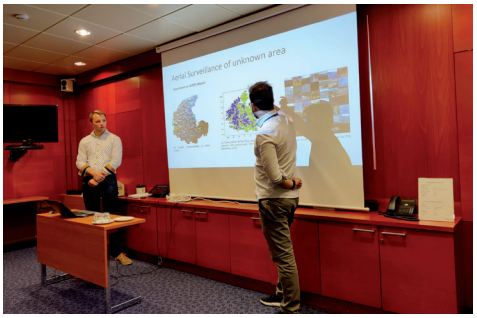
eProcessor has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement no. 956702. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Sweden, Greece, Italy, France, Germany.

The HiPEAC Student Challenge showcases projects by computer science and engineering students, whether undergraduate, master's or PhD. With a theme of the internet of things (IoT) for everyone, the latest edition, at Computing Systems Week Tampere, featured projects ranging from disaster responses to traffic-sign management, as shown in this article.

Students storm the IoT in Tampere

A real-time UAV surveillance system for natural disaster management

TEAM: Daniel Hernández, José D. Padrón, Kiyoshy Nakamura and Jamie Wubben, advised by José M. Cecilia and Carlos T. Calafate, Universitat Politècnica de València



Natural disasters cause devastating loss of life and massive economic losses, and, according to the Red Cross and Red Crescent societies, have increased by 35% since the 1990s. In response, a team from the Universitat Politènica de València came up with an unmanned aerial vehicle (UAV) solution allowing real-time monitoring of natural disasters using artificial intelligence (AI) techniques. The aim was to automate the process of image and video interpretation that will reduce or avoid possible losses, provide the necessary attention to the victims, or speed up decision making during this type of event.

The team proposed a framework comprising a digital twin and a set of UAVs, equipped with a powerful but lightweight computer, such as the NVIDIA Jetson Nano. The UAVs provide the digital twins with real-time data of the natural disaster. In addition, edge machine-learning techniques help analyse images provided by the UAVs and hence optimize decision making under pressure. As such, the project integrates the coordination of drone swarms for visual information gathering and computer vision at the edge into a proof of concept.

Traffic-sign recognition IoT-based application

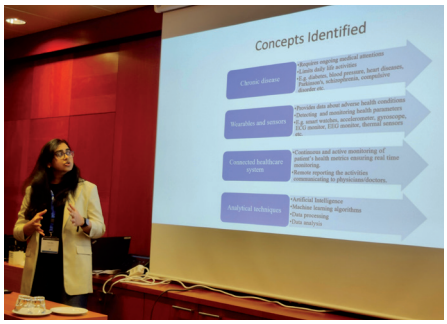
TEAM: Narges Mehran, Dragi Kimovski, Zahra Najafabadi Samani, Radu Prodan, Alpen-Adria-Universitaet Klagenfurt

This project involved the design of a traffic-management application to respond to road-safety concerns. The application receives a raw video, encodes it with a video CoDec in high resolution, divides it into frames and detects the traffic sign shown in the video using a multiclass machine-learning model. The end user – either a vehicle driver or passenger – then receives a warning related to the traffic sign.

The application, along with its source code, can be downloaded from the team's GitHub repository:
github.com/SiNa88/Traffic-sign-recognition-application

Analysis and processing of digital health data using the IoT and multiple analytical techniques

RESEARCHER: Adite Site, supervised by Jari Nurmi and Elena Simona Lohan, Tampere University



New technologies such as mobile apps, novel sensors and wearables are creating a major new source of healthcare data. In addition, they allow patients and healthcare professionals to store, process and share medical data for the coordination of care. However, continuous sensor data from wearables is voluminous, varied and unstructured, and needs to be analysed to offer meaningful insights. Lack of uniformity, consistent standards and interoperability also pose a problem for generating and adopting a common analytical platform for better management and delivery of healthcare services.

Using digital health data, machine learning algorithms and IoT, this research project aims to determine how digital health data can be used for connected healthcare systems to enable remote monitoring and report critical events. It also aims to explore how heterogeneous healthcare data can be combined, processed and analysed using machine learning algorithms.

Finding your ideal career path with HiPEAC Jobs

Deciding on a career path is not always easy, but HiPEAC Jobs is here to help. With careers sessions from roundtables to STEM student days, explore your options and hear from all different kinds of professional. Who knows, your dream job might be one you haven't even heard of yet?

As a community of over 2,000 computing systems specialists, with a well-established calendar of networking events, HiPEAC offers the ideal environment for researchers to build fulfilling careers, whether in academia or industry. Since HiPEAC was established in 2004, numerous cohorts of researchers have progressed from the beginning of their PhDs to postdoc and senior positions. Along the way, they may have participated in the HiPEAC Student Challenge, attended the ACACES summer school, done an internship or received a collaboration grant for a research visit, presented a poster or discussed a project proposal at the HiPEAC conference and subsequently worked on a project during a Computing Systems Week.

For some time, HiPEAC has been leveraging the robust experience base of our community to organize dedicated careers sessions during HiPEAC events. Many HiPEAC members are generously willing to share what they have learned with the next generation of computer scientists and engineers, while for others participating in a careers event is a welcome opportunity to scout for talent.

Careers sessions may include presentations, a roundtable discussion, and / or a tour of company booths, for example. The main objectives of the sessions are as follows:

- to provide or receive inspirational insights into different career paths
- to give or get advice on career development
- to share or learn the main skills employers look for
- to present or hear about potential vacancies or internships, as well as related projects, training networks and programmes
- to discuss and discover the best strategies for talent attraction, on the employer side, or for impressing your interviewer, on the employee side

There's a special chemistry at HiPEAC events, thanks to the blend of technical specialists, students and business developers, and this is obvious in the careers sessions. Topics that have been explored include the value of studying for a PhD, as a door to other opportunities, or the differences between working at a research centre, a university, a major multinational company, a small business, or even your own startup. Students are invited to think out of the box and have their preconceptions challenged, while in general the sessions seek to help them answer the question: 'Why am I doing what I am doing?'

FURTHER INFORMATION:

Find out about forthcoming careers activities on the HiPEAC Jobs career centre webpage bit.ly/HiPEACJobs_career_centre

Check out presentations from former events, interviews with experts and more on HiPEAC TV https://bit.ly/HiPEACJobs_YT_playlist

HiPEAC careers events take different formats

Available to PhD students and junior postdocs, HiPEAC collaboration grants support a three-month research stay at another institution within the HiPEAC network. In this article, Iván Fernández Vega (University of Malaga) explains how his collaboration with Onur Mutlu's group at ETH Zürich led to a paper submission to a top computer architecture venue.

Breaking new boundaries through international exchange: HiPEAC collaboration grants

Accelerating time-series analysis with a memory-based accelerator

NAME: Iván Fernández Vega
ROLE: PhD Student
UNIVERSITY: University of Malaga
HOST INSTITUTION: ETH Zürich
HOST SUPERVISOR: Onur Mutlu
COLLABORATION DATES:
30.08.2021-29.11.2021

Having previously spent several months as a visitor at ETH Zürich, I was keen to return to the team and carry out further work. Thanks to my HiPEAC collaboration grant, I was able to spend three months in Onur Mutlu's research group, focusing on memory-centric computing.

As Professor Mutlu has outlined (see for example *HiPEACinfo* 55), data movement within computing systems is a huge drain on energy use. During my research stay, my main objective was to study how dynamic time warping (DTW), a time-series analysis algorithm, could be sped up using an accelerator based on non-volatile memories. These TSA algorithms are usually memory bound, so reducing the impact of data movement is crucial.

Using the processing-using-memory (PUM) approach, we aimed to reduce the impact of data movement in terms of energy consumption, while simultaneously improving performance by carrying out the computation where the data resides.

Implementing the algorithm using UPMEM

First, we started to characterize the algorithm. To do so, I developed baselines on commodity processors – central and graphics processing units (CPUs and GPUs) – along with a field-programmable gate array (FPGA) implementation based on high-level synthesis. Next, I developed an implementation of the algorithm using UPMEM, a commercially available processing-in-memory architecture where small general-purpose processors are placed very close to the memory array.

Did you know?
As reported in *HIPEACinfo* 63, in 2021 UPMEM announced that it was the first company to mass produce a general-purpose accelerator based on processing in memory, a breakthrough awaited by the computing industry for nearly two decades. UPMEM, a startup incorporated in 2015, is based in Grenoble, France.

Developing a memristor accelerator

Second, we started to develop the idea of building the accelerator. We wanted to evaluate how memristor-based crossbars can perform in-situ computation, enabling a faster, more energy-efficent computation of state-of-the-art TSA algorithms. MATSA, which stands for MRAM-based Acclerator for Time Series Analaysis, uses SOT-MRAM crossbars to perform the computation in situ. I developed the analytical model to evaluate performance and energy use, and we are currently working on getting the results and comparing them to the baselines.

Did you know?
MRAM refers to magnetoresistive random access memory, a type of non-voltatile memory which stores data in magnetic domains, while SOT refers to spin-orbit torque, a technology to manipulate magnetization through electric currents.

As a side project, I contributed to the evaluation of the SpMV kernel on the UPMEM architecture.

Professor Onur Mutlu commented: 'The collaboration with Iván was extremely productive and helped us advance the state of the art in processing-in-memory and algorithm-architecture co-design. As a result of this work, we submitted a strong technical paper to a top venue in computer architecture, and will continue collaborating on non-volatile memories and processing in memory.'

HiPEAC internships are a great way to get work experience at a deep tech company. Beyond that, they can even be life changing, sometimes leading to a job offer, as Maria Oikonomidou found out.

HiPEAC internships: Your career starts here

Data-driven decision making at inbestMe

NAME: Maria Oikonomidou
RESEARCH CENTRE: Foundation for Research and Technology Hellas (FORTH) and University of Crete
HOST COMPANY: inbestMe
DATE OF INTERNSHIP:
15.06.2021 - 15.09.2021

As a graduate assistant at the Institute of Computer Science, FORTH, my studies centred on parallel and distributed systems, algorithms and systems analysis. I was particularly drawn to data science, graphs and applied machine learning for large complex networks, as well as parallel and distributed computing. I'd had experience of addressing problems on social media, particularly on Twitter and YouTube, and found that I was really passionate about finding motifs in individuals' digital footprints and using them to understand behaviours.

The internship opportunity at inbestMe was therefore an ideal fit for me. A roboadvisor which offers tailored investment plans for its clients, inbestMe uses technology to automate as much as possible, which helps keep costs down and consequently offer better returns. My role was to gain as many insights as possible from data to help improve the customer experience.

During my internship, I analysed inbestMe's website traffic, seeking answers to questions about the source of the traffic – including which marketing campaigns had brought visitors to the website – as well as users' behaviour on the site, such as how long they spent on each page and how they navigated the site. I also helped develop a tool to help predict potential future clients: using machine learning techniques and by analysing existing customer behaviour, this tool is able to distinguish which of those users who sign up to the company are more likely to become clients.

Thanks to the HiPEAC internship, I got to know the team at inbestMe and show them what I can do. I'm happy to say that, once it was complete, inbestMe offered me a full-time role as their data scientist, allowing me to concentrate fully on providing enhanced data-driven decision making for the business by building out scalable data infrastructure.

Ferad Zyulkyarov, inbestMe's chief technology officer, commented: 'When I was a PhD student, HiPEAC helped me to develop in my career by offering range of opportunities to collaborate with industry and academia. Now, it still helps us by allowing us to reach to some of the best talent in Europe. For example, through the HiPEAC internship programme, we were able to find, train and then hire Maria. Besides benefiting from HiPEAC, I am very happy that we can also contribute and give back to the HiPEAC community by mentoring the students in applying their skills and expertise in solving real-life business problems.

FURTHER INFORMATION:
inbestMe website inbestme.com

The HiPEAC network includes almost 1,000 PhD students, representing the next generation of computing systems experts. Last year, Fabian Schuiki was one of those students; today, he is a senior compiler engineer at SiFive. In this article, we find out about Fabian's thesis on energy-efficient architectures for high-performance computing.

Three-minute thesis

Featured research: Breaking the von Neumann bottleneck for energy-efficient HPC

NAME: Fabian Schuiki
RESEARCH CENTRE: ETH Zürich
SUPERVISOR: Luca Benini
THESIS TITLE: Streaming Architectures for Extreme Energy Efficiency in High-Performance Computing

As the HiPEAC community is well aware, the end of Moore's Law and the breakdown of Dennard scaling in silicon manufacturing has prompted a paradigm shift in the way we approach computer architecture design. Now that increased performance cannot be achieved through advanced manufacturing processes alone, research is increasingly focused on developing technology-aware computer architectures.

Energy efficiency is a critical focus, as a significant proportion of the physical limitations we see on today's high-performance computing (HPC) systems stem from heat dissipation. Hence performance at low power is the key to achieving high utilization of the available hardware, in order to mitigate the effects of limited frequency and overcome dark silicon.

The von Neumann bottleneck, in which instruction fetches compete with data accesses for memory bandwidth, is one of the key challenges in this area. This bottleneck also applies to the instruction pipeline of a processor, where load-store and control instructions compete with compute instructions for issue slots.

Specialist vs generalist

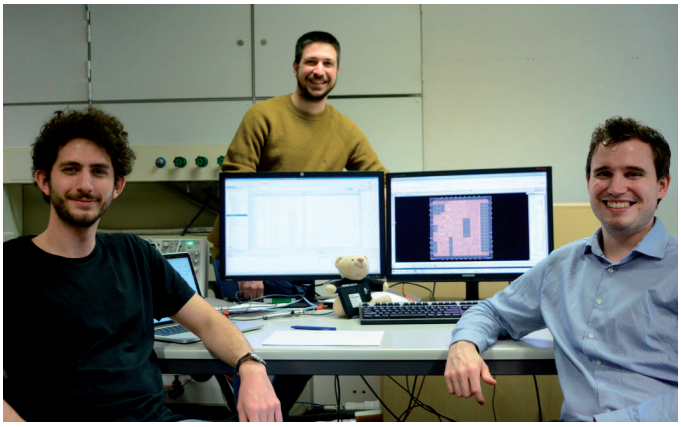
A popular way to overcome this bottleneck is to implement dedicated accelerators for a specific problem, an approach which has gained popularity with the rise of machine learning. This is based on the observation that, all other things being equal, specialization in hardware always wins. However, the lack of general programmability limits the accelerator's use to a specific problem. At a time of fast-moving algorithms, today's hardware accelerators will be unable to compute tomorrow's algorithms.

In parallel, general-purpose processors have also evolved to mitigate the von Neumann bottleneck, for example in the complex to reduced instruction set (CISC-to-RISC) translation in modern processors, which can act as an instruction compression scheme. Similarly, single-instruction, multiple-data (SIMD) and single-instruction, multiple-thread (SIMT) paradigms offer a fixed increase in computations per second, while Cray-style vectorization offers a more dynamic and potentially higher increase.

Data-oblivious algorithms lend themselves particularly well to these kinds of acceleration. Including many algorithms from the fields of linear algebra, machine learning and scientific computing, in the case of these algorithms control flow decisions do not depend on the data it processes.

Hardware address generation and direct memory streaming

This thesis develops the concept of hardware address generation and direct memory streaming as a method to mitigate the von Neumann bottleneck. It then applies the concept to in-order, single-issue processors, allowing them to achieve full utilization of compute resources. Next, it introduces pseudo-dual-issue execution with dedicated compute hardware loops, and distils these extensions into an architectural template for high-performance computers capable of concentrating a significant part of its energy footprint in the arithmetic units.



Fabian (right) and colleagues in the lab

HiPEAC

Thanks to all our sponsors for supporting #HiPEAC22!

arm
CISCO
DeepMind
HUAWEI
OPENHW GROUP
ASML
Collins Aerospace
Google
Hewlett Packard Enterprise
monetdb
nexter A COMPANY OF KNDS
Think Silicon AN APPLIED MATERIALS COMPANY
ZeroPoint TECHNOLOGIES
SUNDANCE
CAES PIONEERING ADVANCED ELECTRONICS
codeplay
ECMWF
evopro
MAXELER a groq company
Qualcomm
Reservoir Labs
semidynamics
Springer
STREAM HPC
THALES

Sponsors correct at time of going to print. For the full list, see hipeac.net/2022/budapest

Join the community

@hipeac
hipeac.net/linkedin
hipeac.net

This project has received funding from the European Union's Horizon2020 research and innovation programme under grant agreement no. 871174