HiPEAC info 59

JANUARY 2020

HiPEAC conference 2020 Bologna

The power: Delivering energy-efficient computing systems

Calista Redmond on open hardware

The wisdom of James Mickens

contents

14

RISC-V's Calista Redmond on the age of open hardware

16

James Mickens on security, architecture and more

17

Energy efficiency special feature

HiPEAC is the European network on high performance and 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. 779656.

Cover image: © Zelenenka | Dreamstime.com
Design: www.magelaan.be
Editor: Madeleine Gray
Email: communication@hipeac.net

25

NAND characterization with NanoCycler

26

Smart monitoring for electric motorcycles

34

Life-changing projects: The HiPEAC Student Challenge

First of all, I would like to wish you a healthy and prosperous 2020, personally as well as professionally. For HiPEAC, the year 2020 begins favourably with the start of HiPEAC6, a continuation of 15 years of HiPEAC.

I will remember 2019 as the year of anger. The 'yellow vests' in France have been demonstrating for more than a year against the French government, Greta Thunberg is known for her angry speeches to world leaders, the Catalans are angry at the Spanish government, President Trump is angry at the Chinese, while the Democrats in the United States are angry at President Trump.

In several European countries, the government no longer has a parliamentary majority. In some countries, political parties have had great difficulty forming a government, in some cases even resulting in fresh elections. In the UK, the previous parliament was unable to decide on Brexit. Political parties seem to have a hard time compromising and putting the country's interest above their own. This is not a good thing when facing major long-term societal challenges like the ageing population, migration, climate change and environmental degradation.

While the United States is preoccupied by Trump's presidency, and Europe is preoccupied with issues like migration and Brexit, China is preparing its future. It is heavily investing in its Belt and Road initiative to create a global network of trade routes, and it is gradually strengthening its influence in areas where the West is retreating. It is quickly becoming a leader in upcoming technologies like renewable energy, electric mobility and artificial intelligence.

I believe it is time for European countries to stop arguing about the present, and to start preparing for the challenges of the future. I therefore very much welcome Ursula von der Leyen's European Green Deal because it will not only reduce emissions and improve sustainability, but is also a call for action to work together on a common goal that will eventually improve the lives of millions of European citizens. I believe that the HiPEAC community has to contribute by proposing technical solutions to make society more sustainable, by urgently working on attitude changes, and by making the next generations enthusiastic about working on serious societal challenges. There are no good excuses for postponing action until tomorrow. Instead, we should all start working on it today.

This is my new year's resolution for 2020.

Koen De Bosschere, HiPEAC coordinator

Once the energy efficiency of computers was largely ignored; today it has become a key concern for computer scientists and engineers. Here, self-confessed geek and European Commission Programme Officer Sandro D'Elia (DG CONNECT) explains why energy efficiency in computing is a crucial endeavour for society as a whole.

Powering up: Energy and computing

I am a physicist by training, and I have always been interested in the physical aspects of computation. In my university years, I learned about Landauer's principle and I found it fascinating. Today, we know that it might be not be correct, but as a young geek in his 20s the idea that you could not go below a certain energy cost to execute any computation seemed to me the ideal connection between the 'real' world of physics and the 'virtual' world of computing. Then I discovered that, in digital computers, the actual energy consumption was enormously higher than Landauer's limit, and the topic suddenly lost interest for me, as just another theoretical subject which was irrelevant in the real world.

"Unbelievable as it may seem, pushing my 80 kilos requires less power than pushing electrons inside a microprocessor"

Fast forward 35 years (and more...). Today the energy cost of computation is a global issue; the 'electricity bill' of information and communication technology (ICT) on this planet is so big that it is difficult even to measure it. And things are not getting better: there are recent general-purpose processors with thermal design power of 400W. To make a comparison, the electric motor of an e-bike that could carry you comfortably around a city requires only 250W. Unbelievable as it may seem, moving my 80+ kilos requires less power than pushing electrons inside a microprocessor. There is clearly something wrong.

The energy consumption of ICT probably became so huge because nobody really cared: the advantages of ICT have been so impressive, in so many sectors, that the electricity bill was the least important problem. But now things are changing, because the total amount of energy wasted is enormous, and because our way of thinking has changed.

Ursula von der Leyen, president of the European Commission, has published a document named *Political guidelines for the next European Commission 2019-2024*. This makes for interesting reading; particularly interesting is the fact that the very first 'headline ambition' for the coming years is named 'A European Green Deal'. The ambition is to make Europe the world's first climate neutral continent by 2050, and to put climate neutrality into law. There is a budget proposal behind this, which means that we are getting serious.

European Commission President Ursula von der Leyen

Let's see how this policy will be implemented in practice. The future 'Digital Europe' programme, which focuses on digital capacity and infrastructure, will complement the 'Horizon Europe' research and innovation programme. Through this programme, the European Union and the Member States will jointly invest in the widespread deployment of digital technologies across society. In the orientation paper for Digital Europe is a section named 'Making ICT products and services sustainable, by prioritising their energy efficiency'. Simply put, this means that the European Commission is planning specific investments in this area over the next few years.

OK, but what's new? After all, for many years the European Commission's Horizon 2020 programme has supported the development of computing technologies which are more energy efficient. Thanks to research funded by the European Union, the power budget of the control electronics for self-driving vehicles has gone down from the kilowatt range to much more manageable numbers, and a new generation of servers is reducing electricity bills in a wide range of datacentre applications.

The problem is that these improvements are simply not enough. We need a lot of computing power, and a lot of artificial intelligence (AI), to address the challenges of sustainability, climate change and the circular economy. But this is a problem: some estimates suggest that training an AI application generates as much CO_2 as five cars over their entire lifetime. For the future, we need powerful AI that can be trained without warming the environment like a pizza oven. Because – if you think about it – even now the number of AI applications that you use every day is much bigger than the number of pizzas that you could possibly eat.

This is a very difficult problem, which probably requires a new generation of hardware beyond the current semiconductor technologies, and a new generation of software to run the hardware. I honestly don't know if we will be able to find a solution but I hope so, because I also have children.

"For the future, we need powerful AI that can be trained without warming the environment like a pizza oven"

Benvenuti a Bologna! – Welcome to Bologna!

Photo credit: Bologna Welcome

A city of exquisite architecture and the first European city to have a university, Bologna is an indisputable gastronomic capital, as well as being home to several major mechanical and electronic companies. We caught up with Luca Fanucci (University of Pisa), the general chair of HiPEAC 2020, and local organiser Andrea Bartolini (University of Bologna) to learn more about this enchanting city and the computing ecosystem in Italy.

What makes Bologna such a good location for the HiPEAC conference?

Bologna is the lively, historic capital of the Emilia Romagna region in northern Italy. One of most productive regions in Italy, Emilia Romagna is famous worldwide for its 'Motor, Food, Biomedical, and Packaging Valley'. Bologna is also the Italian 'city of high-performance computing', hosting the Italian Tier-0 data centre for Supercomputing (CINECA SCAI) and the Italian Tier-1 data centre for the high-energy physics experiments at the Large Hadron Collider in Geneva (CNAF INFN). It is also preparing to host the new European Centre for Medium-Range Weather Forecasts (ECMWF) data centre, as well as the EuroHPC pre-exascale data centre.

With beautiful squares, cafes, portici (arcades), and medieval and Renaissance structures, Bologna is a city of two intriguing halves. On one side is the hard-working, hi-tech city with suave operagoers who waltz out of the theatres and reconvene in some of the nation's finest restaurants. On the other side is the politically edgy city that hosts the world's oldest university and is famous for its graffiti-embellished piazzas filled with students swapping Gothic fashion tips.

Bologna is also an important Italian transport hub, and is very easy to reach. Thanks to its high-speed train connection, Bologna is just a few hours' distance from the major northern and southern Italian cities.

Bologna is a city which welcomes its visitors with a rich set of cultural activities, wonderful views, and tasty cuisine (although not spaghetti Bolognese!).

Can you tell us a bit about the computing systems community in Italy? Where are the industrial and academic centres, and what makes the ecosystem unique?

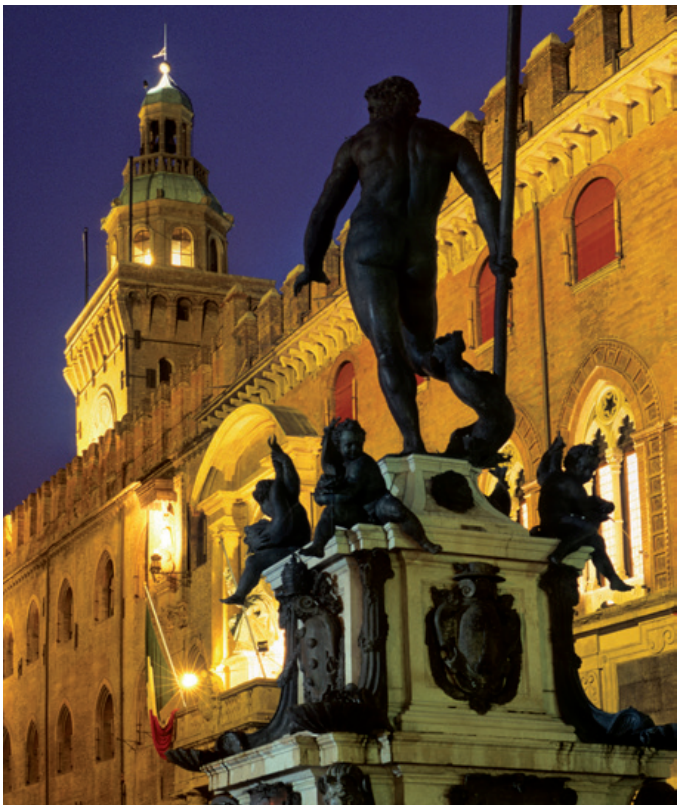
The Italian coomputing systems academic community is represented by the national computing consortium, the Consorzio Interuniversitario Nazionale per l'Informatica (CINI), which

consists of 47 public universities and over 1,300 teachers within the sectors of computer science, computer engineering, and information technology. In close cooperation with the national scientific communities, the consortium promotes and coordinates research and technology transfer activities, both basic and applied, in a number of computer science and computer engineering fields.

At the Alma Mater University of Bologna, the DISI (Computer Science and Engineering) and DEI (Electrical, Electronics and Information Engineering) departments are the main players in information technology development. At the University of Pisa, this role is taken by the DI (Computer Science) and DII (Information Engineering) departments.

Both Tuscany and Emilia Romagna have recently created two competence centres, namely Bi-Rex and Artes 4.0, which bring together representatives from universities, research centres and companies of excellence to help companies, especially small/medium enterprises (SMEs), adopt industry 4.0-enabling technologies.

In terms of industry, the Italian computing system community is represented by both large companies and SMEs. To name a just a few: STMicroelectronics, Engineering, Leonardo, Thales Alenia Space (large companies); EnginSoft, E4 Computer Engineering, CAEN, Sitael, Seco, Ingeniars (SMEs). *[Note: you can catch up with several of these companies at the industry exhibition at HiPEAC 2020.]*



Neptune´s Fountain

Piazza Santo Stefano

What are some of the most interesting projects taking place at the University of Pisa and the University of Bologna at the moment?

Both the University of Bologna and the University of Pisa are active members of the European Processor Initiative consortium currently implemented under the first stage of the Framework Partnership Agreement signed by the consortium with the European Commission (FPA: 800928). This aims to design and implement a roadmap for a new family of low-power European processors for extreme-scale computing, high performance, big data and a range of emerging applications. In addition, the University of Bologna is currently involved in the European Union (EU) H2020 ICT11 IoTwins project, with a focus on digital twin technologies for industrial SMEs, while the University of Pisa is involved in the H2020 innovation action TETRAMAX, which promotes the digitalization of European industries in the domain of customized and low-energy computing for cyber physical systems and the internet of things.

Another interesting activity in which the University of Pisa is involved is the CloudScout project in collaboration with the European Space Agency, due to launch in early 2020. This represents the first deep neural network onboard the satellite exploiting the Intel Movidius Myriad-2 VPU.

What we should do and see in Bologna while we're there?

Bologna is a great place to walk around: getting around the city is very simple with clear street signs and many distances that can be quickly be covered on foot. In the city centre the 'must-see' places are the main square (Piazza Maggiore), nearby Neptune's Fountain, the two towers, the San Francesco and Santo Stefano squares, and the Archiginnasio Public Library, as well as the Sala Borsa Public Library. Outside the city centre you'll find the FICO Eataly World Bologna Park. While staying in Bologna, make the most of the food: lasagna, tortellini and tagliatelle Bolognese are a must, but a cheese and cold meat board with local bread (*tagliere di formaggi e salumi con tigelle e crescentine*) must also be tasted.


Photo credit: Eneko Illarramendi

Powering digital transformation at Computing Systems Week Bilbao

Bilbao was an apt setting for a Computing Systems Week focused on digital transformation. With major research and technology centres and a strong industrial base, the capital of the Basque country has long been regarded as an engine for innovation.

Keynote talks from Onur Mutlu (ETH Zürich), Johan Stahre (Chalmers University of Technology) and Eduardo Quiñones (Barcelona Supercomputing Center) covered intelligent architectures for intelligent machines, industry in digital transition and the convergence of embedded and high-performance computing, respectively. The dedicated industry day focused on how business is being transformed by digital technologies, with examples from Digital Innovation Hubs and local companies, including how to integrate people with a learning disability into the workplace.

Meanwhile, technical sessions covered the computing spectrum from high performance to embedded systems, including a Eurolab4HPC workshop on programming models for future supercomputers, safety and modelling in embedded computing and the convergence of big-data analytics, cloud and high-performance computing.

Once again, the HiPEAC student track was a great success. Participants in the Student Challenge showcased a range of inspirational projects tackling societal issues from finding a train seat in Belgium to tracking missing people in Turkey or detecting wildfires in Spain. Thanks to our sponsors, Intel Movidius and Arm Education, students were awarded with the Intel Movidius Neural Compute Stick, STMicroelectronics embedded hardware and Arm internet of things / embedded systems course vouchers. Meanwhile, with presentations from local and international organizations, the Inspiring Futures session helped guide students in their career decisions.

The next HiPEAC Computing Systems Week is due to take place in Tampere, Finland on 27-29 April. Watch this space for further information.

Turn to p.34 for an interview with participants in the HiPEAC Student Challenge.

Photo credit: Miguel Martín

Photo credit: Evangelos Petrongonas

dividiti (dv/dt) accelerate omni-benchmarking for MLPerf Inference v0.5

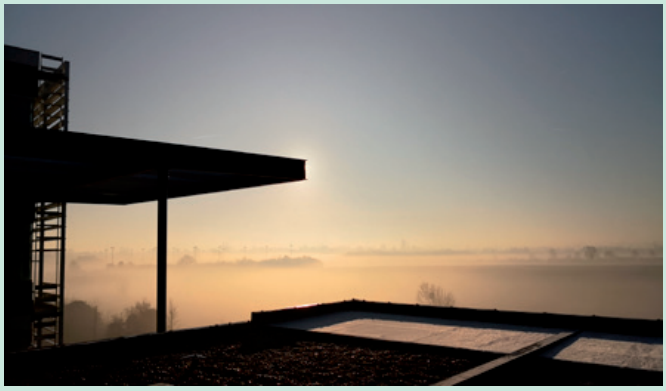
In November 2019, the MLPerf consortium released over 500 validated inference benchmarking results from 14 organizations measuring how fast and how well a pre-trained computer system can classify images, detect objects and translate sentences. Over 400 of these results were submitted by dividiti, a HiPEAC company based in Cambridge, UK.

'Our success in MLPerf Inference v0.5 is due to our unique open workflow automation technology, Collective Knowledge (CK),' explains Dr Anton Lokhmotov, chief executive and co-founder of dividiti. 'We conducted hundreds of benchmarking experiments, followed by thousands of auditing experiments, with many combinations of machine learning models, libraries, frameworks and hardware platforms. Such experiments are notoriously hard to stage in an automated, portable and reproducible fashion, which explains why even well-resourced hardware vendors only submit a handful of results. In collaboration with Arm and the Politecnico de Milano, we staged experiments on systems ranging from Raspberry Pi class boards and Android phones to high-end workstations.'

'MLPerf is being contributed to by many organizations, from tiny startups to giant corporations with up to 50 contributors per organization. It is simply astonishing that a small organization with only three MLPerf contributors has submitted nearly three times more results than all other organizations combined,' stated Dr Vijay Janapa Reddi, Associate Professor, Harvard University, and MLPerf inference co-chair. 'Workflow automation will be critical not only for generating large volumes of high-quality results, but also for validating and finding the most optimal solutions in terms of performance, quality and cost.'

'Benchmarking modern-day platforms with multiple software branches, libraries, toolchains, datasets, and test and device configurations may deliver a set of inconsistent results,' said Colin Osborne, director of engineering and distinguished engineer, Machine Learning Group, Arm. 'Arm uses the Collective Knowledge (CK) framework to transform our multi-dimensional problem space into simplified building blocks and more manageable benchmark results.'

"MLPerf Inference v0.5 is a strong testimony for the HiPEAC internship programme"



View from dividiti's Cambridge headquarters

MLPerf Inference v0.5 is also a strong testimony for the HiPEAC internship programme, which enabled a three-month visit of Emanuele VItali, a PhD candidate at Politecnico di Milano, to dividiti's headquarters in Cambridge.

'Our collaboration focused on studying the performance and accuracy of convolutional object detection models across combinations of TensorFlow backends and run-time options. We initially published our results in a blog post, but with the encouragement of my mentors at dividiti we also submitted them to MLPerf Inference,' said Emanuele. 'Overall, I found my internship a deeply satisfying experience: an opportunity to try something different to my usual academic setting, take on an industry-scale challenge and also make PoliMi the first university ever to submit to MLPerf. On a personal level, my mentors were very friendly and approachable, happy to explain things at a high-level but also to get down to the nitty gritty details. I highly recommend that other PhD students try HiPEAC internships.'

'My own motivation for a career in computer engineering was propelled by an internship where I reported directly to the legendary ARM instruction set architect, Sophie Wilson, followed by PhD internships at exciting compiler and semiconductor companies,' added Dr Anton Lokhmotov. 'Over the last ten years, I have hosted many HiPEAC interns at Arm and dividiti, who have pursued their careers at leading research organizations and companies including DeepMind, Google, NVIDIA, Microsoft, Intel, Arm and TomTom. If you are a HiPEAC student, do yourself a favour and do an internship.'

FURTHER INFORMATION:

MLPerf consortium

mlperf.org

Blogpost: 'Omni-benchmarking Object Detection'

bit.ly/ob-od

BSC announces global collaboration facility to develop open computer architectures

BSC has announced the opening of the European Laboratory for Open Computer Architecture (LOCA). LOCA's mission is to design and develop energy-efficient and high-performance chips, based on open architectures like RISC-V, OpenPOWER, and MIPS, in Europe, for use within future exascale supercomputers and other high performance domains. Building on work done in the European Processor Initiative and MareNostrum Experimental Exascale Project, LOCA will be based in Barcelona and build open-source ecosystems, including the design and development of open-source hardware for high-performance computing chips and open-source software.

'LOCA will be a collaborative laboratory that welcomes companies, foundations and academic institutions that share the vision that it is necessary to create open architectures to guarantee transparency, competitiveness and technological sovereignty,' said BSC Director Mateo Valero. 'We are launching it with great conviction, because it is another step in our philosophy of paving the way for the creation of European high-performance computing (HPC) architectures. In the past, in the Mont-Blanc project, we created a cluster based on Arm processors, while in the EPI project we are developing the general software stack and a RISC-V accelerator in the Arm-based multicore chip.'

BSC's John D. Davis, who received his PhD from Stanford and has held multiple technical roles in startups and large companies, will direct the European Laboratory for Open Computer Architecture. 'We envision a future that is wide open, incorporating open-source software and hardware,' he says. 'LOCA is a mechanism to extend the success of open-source software like Linux to the hardware domain. To unlock the potential energy efficiency and performance of future systems, we must use hardware/software co-design, enabled by an open hardware and open software ecosystem. LOCA´s inaugural five-year plan focuses on developing and building key open European-made intellectual property as a basis for future exascale systems – and beyond.'

BSC is looking for interested parties from industry and academia to collaborate in LOCA. In addition, the centre is actively seeking top researchers to join the centre and contribute to LOCA.

FURTHER INFORMATION:
bit.ly/LOCA_announcement

Designing cloud-based solutions for global enterprises

In December, HiPEAC member Josip Pojatina gave a talk at .debug, the largest developers' conference in Southeast Europe, on the topic of designing cloud-based solutions for global enterprises.

As more and more enterprises start to migrate parts of their load to the cloud, many smaller companies wonder why their advanced solutions were rejected by global enterprises when competing for a tender with famous software vendors.

In his talk, Josip revealed the criteria used by global enterprises when developing or selecting cloud-based solutions and the most common pitfalls of even the largest software vendors.

The presentation covered onsite private / hybrid / public cloud architecture, IaaS / PaaS / SaaS / Infrastructure as a code, security, data privacy / handling and protection, regulations and compliance, cloudeconomy and performance tuning, integration, cloud sovereignty, support and other important, but very often overlooked, criteria.

Developers were given insights into how to build a solution that can fulfil all important criteria and how to find the right balance among opposite constraints such as performance, encryption, compliance and the cloud economy.

Solutions for the most common mistakes will be provided along with tips on how to manage ever-changing cloud components, pricing and data privacy regulations.

FURTHER INFORMATION:
debug.hr

Parallelware Analyzer: Quality assurance and development tools for parallel code

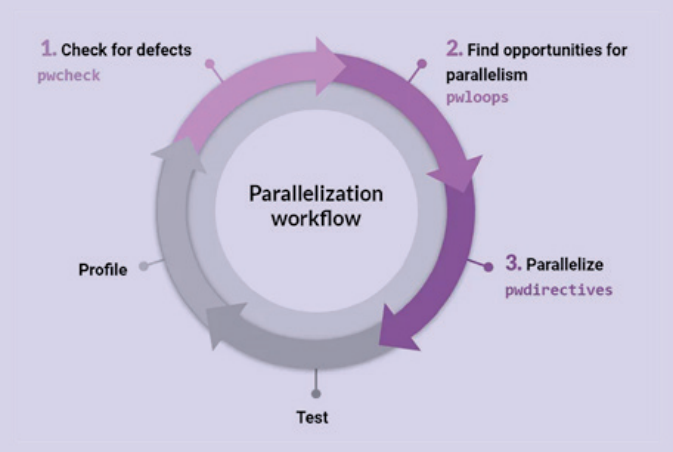
Manuel Arenaz, Appentra

Appentra is a Spanish deep tech software startup company that strives to minimize and eventually remove the parallel software development barrier, making parallel computing easier for everyone. Appentra's new product is Parallelware Analyzer, a suite of command-line tools aimed at helping software developers to build better quality parallel software in less time.

Thanks to the state-of-the-art static code analysis capabilities of the Parallelware technology, source code is analysed quickly, enabling real-time static code analysis for the development of correct parallel software.

While Parallelware Trainer provides an interactive learning environment where users can learn how to parallelize, Parallelware Analyzer provides the appropriate tools for the key stages of the parallel development workflow, aiding developers with code analysis that would otherwise be error prone, time consuming and completed manually. More specifically:

1. **Prepare code for parallelization.**
   How an algorithm is codified has a great influence on how it can be parallelized. Parallelware Analyzer provides users with recommendations on how to make the code follow best practice for parallelization.
2. **Detect and fix defects related to parallelism.**
   Data race conditions are very hard to detect and debug. Parallelware Analyzer finds data race conditions hidden in the code and obtains information about how to fix them.
3. **Identifying opportunities for parallelization.**
   The code most likely contains many loops which can be parallelized in many different ways. Users can discover which loops are most suitable for parallelization, how they can be classified into common parallel patterns and which is the best parallelization strategy for them.
4. **Parallelize opportunities, choosing from a range of technologies and heterogeneous computing platforms.**
   Different parallel hardware and technologies to exploit them are available, allowing the creation of new parallel versions of the code using technologies such as OpenMP and OpenACC and targeting different hardware, including multicore central processing units (CPUs) and graphics processing units (GPUs).



Parallelware Analyzer is available for Linux – on both x86-64 and Power architectures. Windows and Mac versions will be out soon.

Appentra is offering developers the opportunity to join the Early Access Program for Parallelware Analyzer, and we are keen to receive their feedback so that the tool fulfills their needs.

FURTHER INFORMATION:
appentra.com/products/parallelware-analyzer/

New decade, new phase for HiPEAC

HiPEAC 6, the latest phase of the HiPEAC project, began on 1 December 2019. In this new phase of the project, HiPEAC will be focusing increasingly on cyber-physical systems and the high-performance, embedded compute elements needed to power their increasingly sophisticated devices.

HiPEAC has been funded continuously by the European Commission since 2004, and has been giving the computing systems community in Europe an identity, a voice and a hub to share knowledge and experience ever since. We are excited about the challenges this new phase will bring and would like to thank the European Commission for their continued confidence in the project.

Keep up to date on the latest HiPEAC news, events and activities by visiting our website:
hipeac.net

Winners of the HiPEAC Tech Transfer Awards 2019 announced

The fifth edition of the HiPEAC Tech Transfer Awards were announced in December 2019. This year, eight winners have been selected, as follows:

- **Monica De Mier, Bytelab Solutions:** Bytelab Solutions
  The start-up company Bytelab Solutions, dedicated to the acceleration of materials research and development via digitalization, is developing and commercializing a software platform that integrates the open-source code BigDFT. This software, co-developed at Barcelona Supercomputing Center (BSC), can be used for the computational design and characterization of materials. Bytelab's mission is to democratize the use of computer simulations to enable faster (and thus more cost-efficient) development of materials.

- **Daniel Hofman, University of Zagreb:** Jobbee system
  Jobbee is a predictive, geo-location sensitive system for matching recruiters and candidates, tasks and skills, based on scientifically developed algorithms and inputs from human resources and human behaviour expertise. It was created thanks to the transfer of knowledge from the field of data management, data mining, recommendation systems, user tracking and parallel processing from the University of Zagreb to the private company VIDI-to.

- **Leonidas Kosmidis, Barcelona Supercomputing Center (BSC):**
  Brook SC
  Brook SC was developed to allow safety-critical applications to be programmed in a CUDA-like general-purpose graphics processing unit (GPU) language, Brook, which enables the certification of the code while enhancing programmer productivity. In this technology transfer, an Airbus prototype application performing general-purpose computations with a safety-critical graphics API was ported to use Brook SC.

- **Filippo Mantovani, Barcelona Supercomputing Center (BSC):**
  Data visualization for industry 4.0
  The BSC visualization tool Paraver, which analyses performance, was transferred to the company Aingura IIoT, which provides solutions to improve manufacturing productivity. This allows Aingura IIoT to represent time-stamped data series, combine data series and visualize the data in graphs, for example.

- **Farhad Merchant, RWTH Aachen University:**
  Parameterized posit arithmetic hardware generator and its integration to a RISC-V based platform
  Based on posit arithmetic, an alternative to the traditional floating point standard for computing, RWTH Aachen provided services to the Bosch Research and Technology Center in Bangalore, India, transferring a posit arithmetic hardware generator, named PAU generator, then integrating PAU-generated hardware onto a RISC-V platform. It is projected that Bosch will be able to source 10-12% more microprocessors for cameras enabled with posit arithmetic.

- **Tuan Nguyen, student of Akash Kumar, Technische Universität Dresden:** Automatic floorplanner for partially reconfigurable FPGA-based design
  One of the main benefits of field-programmable gate arrays (FPGAs) is the ability to reconfigure small areas of the FPGA without interrupting the whole system. However, this requires designers to floorplan the design on the FPGA, and automatic floorplanning for partial reconfiguration is not currently offered by commercial electronic design and automation (EDA) vendors. The floorplanning technology developed in this project, which was transferred from TU Dresden to Huawei, is able to automatically floorplan the design within minutes.

- **Horacio Pérez-Sánchez, José Pedro Cerón Carrasco, Jorge de la Peña García and Helena den Haan Alonso (UCAM):**
  Computational drug discovery of Zika virus inhibitors
  Members of the BIO-HPC group at UCAM created a marketable solution using advanced computational drug discovery technologies to predict compounds that could inhibit the Zika virus. Tests at the University of Hong Kong, showed one of the compounds, novobiocin, to potentially inhibit the virus. A patent was obtained and subsequently licensed to the pharmaceutical company Ennaid Therapeutics. Clinical trials will begin in 2020.

- **Hans Salomonsson, EmbeDL:** EmbeDL
  Developed by the artificial intelligence (AI) research and innovation company Machine Intelligence Sweden AB as part of the Horizon 2020 project LEGaTO, EmbeDL is a research spin-out that can help organizations bring deep learning (DL)-based AI into physical products (embedded systems). EmbeDL can make AI components in cars, drones and smart homes significantly faster while using less energy and memory.

Members of HiPEAC granted first place in the Xilinx Open Hardware Contest in two categories

Students Dimitris Danopoulos and George Tzanos, supervised by research associate and HiPEAC members Dr Christoforos Kachris and Professor Dimitrios Soudris, won first place in the Xilinx Open Hardware Contest with a novel platform for the hardware acceleration of machine learning applications. The platform was developed at the Microprocessors Lab at the National Technical University of Athens (NTUA).

Specifically, Dimitris Danopoulos received first prize in the hardware acceleration category for his work on the acceleration of FAISS, a widely used application for efficient similarity search and clustering of dense vectors on a Xilinx Alveo field-programmable gate array (FPGA) board and cloud FPGA. George Tzanos received the first prize in the PYNQ category (using the Zynq system-on-chip, or SoC) for the efficient acceleration of the naïve Bayes algorithm for training and classification.

This work was carried out as part of the 'CloudAccel: Hardware Acceleration of Machine Learning Applications in the Cloud', a project that has received funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT) under grant agreement no. 2212, as well as the Xilinx University Program.

As reported in *HiPEACinfo 52*, in 2017 students from the Microlab/NTUA also received first prize in this contest and later established the InAccel startup, now a world leader in application acceleration using FPGAs.

The platforms are available on GitHub:
github.com/dimdano/faiss-fpga
bit.ly/GitHub_AccelCloud_naive_Bayes

Video showing the main innovations of the platform:
youtu.be/a7iEHEheDxs

Dates for your diary

DAC 2020: Design Automation Conference
19-23 July 2020, San Francisco, CA, United States
Paper submssions until 22 January 2020
dac.com

ISC High Performance 2020
21-25 June 2020, Frankfurt, Germany
February 2020: deadlines for tutorials, project posters, and workshops.
isc-hpc.com

DATE20: Design, Automation and Test in Europe
9-13 March 2020, Grenoble, France
HiPEAC booth and jobs wall
date-conference.com

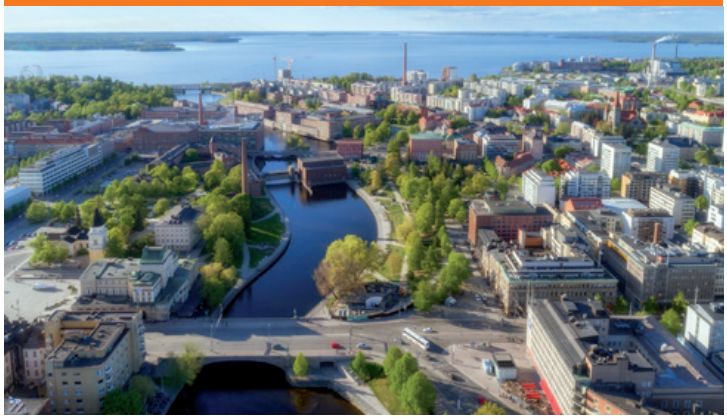
ASPLOS 2020: Conference on Architectural Support for Programming Languages and Operating Systems
16-20 March 2020, Lausanne, Switzerland
HiPEAC Paper Award conference
asplos-conference.org

EuroHPC Summit Week 2020
23-27 March 2020, Porto, Portugal
exdci.eu/events/eurohpc-summit-week-2020

Embedded Systems Week
11-16 October 2020, Shanghai, China
Abstract submissions: 3 April 2020
esweek.org

FMEC 2020: The Fifth International Conference on Fog and Mobile Edge Computing
20-23 April 2020, Paris, France
emergingtechnet.org/FMEC2020

Computing Systems Week Tampere
27-29 April 2020, Tampere, Finland
hipeac.net/csw/2020/tampere



'RISC-V is ushering in a new era of sili

With open hardware riding a wave at the moment, there has perhaps never been so much focus on the RISC-V instruction set architecture (ISA), which has been championed for several years by HiPEAC members. We caught up with HiPEAC 2020 keynote speaker Calista Redmond, chief executive of the RISC-V Foundation, to find out more.

Why is this an exciting time to be working on open hardware projects?

As computing demands continue to rise thanks to artificial intelligence, machine learning, the internet of things (IoT) and virtual/augmented reality (VR/AR), there is a growing demand for custom processors purpose-built to meet the power and performance requirements of specific applications. RISC-V is enabling the industry to optimize designs for today's computing requirements and innovate faster.

Throughout my career, I've been proud to be part of a number of open source initiatives to drive tech innovation forward and foster industry-wide collaboration. The RISC-V ecosystem is one of the most dynamic communities I've seen to date. Over the last few years, the membership has grown exceptionally fast and includes a broad mix of organizations in different industries.

As I've been travelling across the globe to promote the benefits of RISC-V at events and meet with our member companies, it's really stuck me how the level of commitment to drive the mainstream adoption of RISC-V is like nothing I've seen before. It's exhilarating to witness our community collaborate across industries and geographies with the shared goal of accelerating the RISC-V ecosystem.

What advantages does RISC-V offer as opposed to commercial ISAs? What opportunities does RISC-V open up, both for research and business?

Unlike legacy instruction set architectures (ISAs) which are decades old and are not designed to handle the latest workloads, RISC-V has a variety of advantages including its openness, simplicity, clean-slate design, modularity, extensibility and stability. Thanks to these benefits, RISC-V is ushering in a new era of silicon design and processor innovation.

We've seen that RISC-V:

1. Unlocks architecture and enables innovation. RISC-V is a layered and extensible ISA so companies can easily implement the minimal instruction set, well-defined extensions and custom extensions to create custom processors for cutting-edge workloads.
2. Reduces risk and investment by enabling companies to leverage established and common intellectual property (IP) building blocks with the development community's growing set of shared tools and development resources.
3. Provides the flexibility to create thousands of possible custom processors. Since implementation is not defined at the ISA level, but rather by the composition of the system-on-chip (SoC) and other design attributes, engineers can choose to go big, small, powerful or lightweight with their designs.
4. Accelerates time to market through collaboration and open-source IP reuse. RISC-V not only reduces development expenses, but also enables companies to get their designs to market faster.

In addition to the many benefits RISC-V offers companies, the simple fixed base ISA and modular fixed standard extensions also make it easy for researchers, teachers and students to leverage RISC-V to learn and push the boundaries of design.

Is open hardware really sustainable? How does the business case stack up?

With more than 420 organizations, individuals and universities that are members of the RISC-V Foundation, there is a really vibrant community collaborating together to drive the progression of ratified specs, compliance suites and other technical deliverables for the RISC-V ecosystem. While RISC-V has a BSD open-source licence, designers are welcome to develop proprietary implementations for commercial use as they see

icon design and processor innovation'

fit. RISC-V offers a variety of commercial benefits, enabling companies to accelerate development time while also reducing strategic risk and overall costs. Thanks to these design and cost benefits, I'm confident that members will continue to actively contribute to the RISC-V ecosystem to not only drive innovation forward, but also benefit their bottom line.

What's the role of the RISC-V Foundation in promoting the RISC-V ISA?

The RISC-V Foundation's role is to build an open, collaborative community of software and hardware innovators while directing the future development and adoption of the RISC-V ISA. The Foundation hosts over 20 work groups and committees that are hands-on developing the extensions, tools and strategy to accelerate implementation design and adoption. We have groups focused on security, ISA extensions, verification and compliance, emulation, software, and several on industry specific interests such as high-performance computing (HPC) and academia.

To gather the community together, each year the RISC-V Foundation hosts global events to discuss current and prospective RISC-V projects and implementations, commercial and open-source implementations, software and silicon, vectors and security, applications and accelerators, simulation infrastructure and much more. We also actively promote independently hosted Meetups and events centred around RISC-V.

We encourage organizations, individuals, and enthusiasts to join our ecosystem and together enable a new era of processor innovation through open standard collaboration.

Events such as the 2019 RISC-V Workshop in Zurich bring the community together. Photo credit: Andreas Kurth

Can you name some of your favourite projects using RISC-V?

I don't have a favourite project, but rather I love the amazing spectrum that RISC-V is engaged in – from a wearable health monitor to scaled-out cloud data centres, from universities in Pakistan to the University of Bologna in Italy or Barcelona Supercomputing Center in Spain, from design tools to foundries, from the most renowned global tech companies to entrepreneurs raising their first round of capital. Our community is broad, deep, growing and energized.

What are some of the plans for the future of RISC-V?

The Foundation's goal is to accelerate industry adoption of RISC-V for the shared benefit of the entire community of stakeholders. We'll continue to do that by driving progression and closure on standards and technical deliverables, making it easier for companies to implement RISC-V cores in their products. Another priority is growing the overall member community across stakeholder areas and deepening community engagement. We will do this by focusing on expanding the ecosystem across industries and geographies, along with offering more support and educational tools so operating systems, hardware implementations and development tools can scale faster.

The RISC-V ecosystem is poised to significantly grow over the next five years. Semico Research predicts that the market will consume a total of 62.4 billion RISC-V central processing unit (CPU) cores by 2025! By that time I look forward to seeing many new types of RISC-V implementations including innovative consumer devices, industrial applications, high performance computing applications and much more.

Don't miss the Eurolab4HPC session on open-source hardware on 21 January at HiPEAC 2020, featuring Calista Redmond, Rick O'Connor, Ted Marena and more!

bit.ly/HiPEAC20_Eurolab4HPC_Open_HW

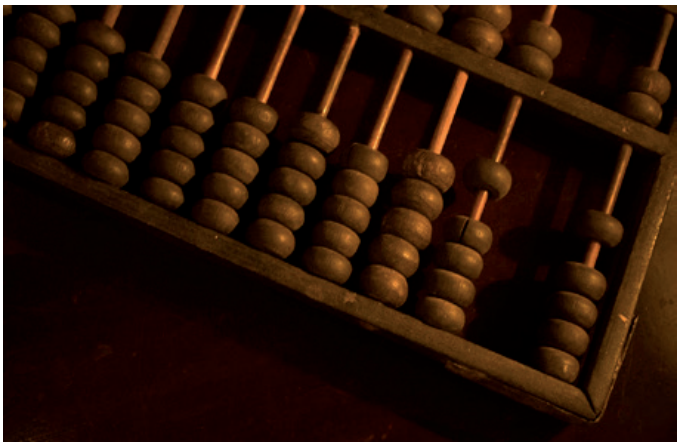
"It's exhilarating to see our community collaborate with the shared goal of accelerating the RISC-V ecosystem"

If you haven't heard HiPEAC 2020 keynote speaker James Mickens talk before, you've been missing out. From machine learning's similarity to a portal to a demon universe to the inherent unreliability of distributed systems, by way of disconcerting football mascots and potato hands, the Harvard computer science professor offers a fresh, funny and, above all, human take on computing systems.
[Warning: this interview contains irony.]

'Unlike SGX, an abacus isn't vulnerable to side channels'

Your talk at HiPEAC focuses on computer security – but everything about computer security is fine, right, so what will you actually have to say?
Indeed, I was shocked to learn that computer security became a solved problem in the past six months. The majority of my HiPEAC talk will now be devoted to lengthy rants about airport WiFi. I feel confident that the HiPEAC audience will enjoy my five-act, one-man play involving the relative merits of "Fr334irport_99GHZ" versus "nope_nope_A_WOLFMAN_APPEARS_waffles."

Aren't ethical considerations just something which liberal arts graduates invented to hold back the marvel that is machine learning?
Anything which limits the rate of initial public offerings is suspicious by default and likely related to communism. Unrestricted stock options are the best way to inspire the moral deviants whose imaginations will unlock tantalizing new markets like selling-blood-as-a-service and on-demand-bedbug-removal-because-earlier-today-I-put-bedbugs-in-your-bed. Ethical considerations are just tiny chains that we must break to enable a world in which gig workers create simulations of other gig workers, thereby creating a self-sustaining recursion of ambiguous social progress which becomes less ambiguous the more that you look at it. Science!



Look! No side channels. Photo credit: Thảo Vy Võ Phạm, Pixabay

It has sometimes been suggested that researchers might like to talk to people from other fields and look beyond their field of architectural or compiler brilliance. Why is it important to have a holistic perspective?
I was unaware that there are other types of brilliance besides architecture brilliance and compiler brilliance. Do you have citations to support this claim? Note that I only accept citations from ISCA, ASPLOS, or original napkin scribblings by Tomasulo.

Europe likes to hold itself up as a kind of moral authority on digital technology. Do you think that some of the approaches being applied, such as the General Data Protection Regulation (GDPR) for privacy, are in the right direction?
The GDPR assumes that Europeans have interesting data to steal. However, criminals don't care about the best way to wear a beret, or how you can build a manual-transmission car that seats one-and-a-half people uncomfortably. Americans have real data to steal, like the best way to make a turducken for Thanksgiving. Do you know what a turducken is? It's a chicken stuffed inside a duck stuffed inside a turkey. A single turducken contains 17,000 calories and is created via a process that resembles nuclear fusion. Turduckens also demand to be served with sauces that will add another 4,000 – 5,000 calories. If the American national anthem isn't playing in your inner monologue right now, STAND UP AND RESPECT MY FLAG.

Given the potential of digital technology for wreaking havoc, do we need to go back to executing computational tasks on an abacus or sticks?
Unlike SGX, an abacus isn't vulnerable to side channels. So, that's nice. The 0.5 Hz clock speed is somewhat unfortunate, but it does give you time to appreciate your inevitable mortality. A well-developed sense of your finite nature will encourage you to write code that compiles on the first try.

Want more? Visit The Wisdom of James Mickens:
mickens.seas.harvard.edu/wisdom-james-mickens

Energy efficiency is high on the agenda for computing systems of all kinds, from the behemoths of data centres and high-performance computing (HPC) installations to the myriad miniature devices populating the internet of things (IoT). In this special feature, we find out how the HiPEAC community is delivering low-power, energy-efficient solutions, from the chip to the network.

Bright sparks

Tackling the energy challenge in computing systems

A HOLISTIC SOLUTION FOR ENERGY-EFFICIENT COMPUTING, ACROSS THE STACK - THE LEGATO PROJECT

The LEGaTO project, funded by the European Commission, aims to address the issue of energy efficiency with a made-in-Europe toolset which also takes in security, reliability and programmability. We caught up with ***Osman Ünsal*** *(Barcelona Supercomputing Centre), longstanding HiPEAC member and coordinator of the LEGaTO project, to find out more.*

Why is reducing energy use so important for future computing systems?

Computing systems are responsible for increasing proportions of global electricity consumption. Take exascale computing systems, which currently would require similar amounts of energy to that of a small nuclear plant. Or data centres: with cloud computing becoming the dominant modus operandi, the enormous server farms needed to power it are making ever greater contributions to the global carbon footprint. The IoT also poses problems in terms of energy consumption; as noted in *HiPEACinfo 58* and elsewhere, while individual devices may consume very little energy, an explosion in connected devices could make the aggregate energy consumption a major issue.

What is LEGaTO doing to address this problem – across the computing stack?

The LEGaTO toolset has all levels of the computing stack covered. Our partners' hardware spans the gamut from the IoT to HPC, while on the software side we're using a task-based programming environment with many variants.

This programming environment is extremely versatile, from making many different machines appear as a single machine to the programmer, to making the most of energy-efficient field-programmable gate array (FPGA) accelerators, and it can be used for cloud or real-time computing as well as for high-performance applications. Task-based and dataflow focused, it is based upon: OmpSs, created and built upon at BSC; Xitao, the experimental execution model and runtime from Chalmers University; Maxcompiler, the Maxeler dataflow programming model; and Technion's DFiant high-level synthesis (HLS) language.

How do heterogeneous hardware platforms help enhance performance and reduce energy consumption?

Previously, chip manufacturers such as Intel could rely on producing faster and faster general-purpose processors,

negating the need for specialist processors. However, now that Moore's Law is becoming harder to sustain, heterogeneous platforms composed of different kinds of processors are being employed to ensure continued performance gains. At the same time, from the energy point of view it makes sense to build custom chips that are more energy efficient. The key is identifying the right chip for the right application.

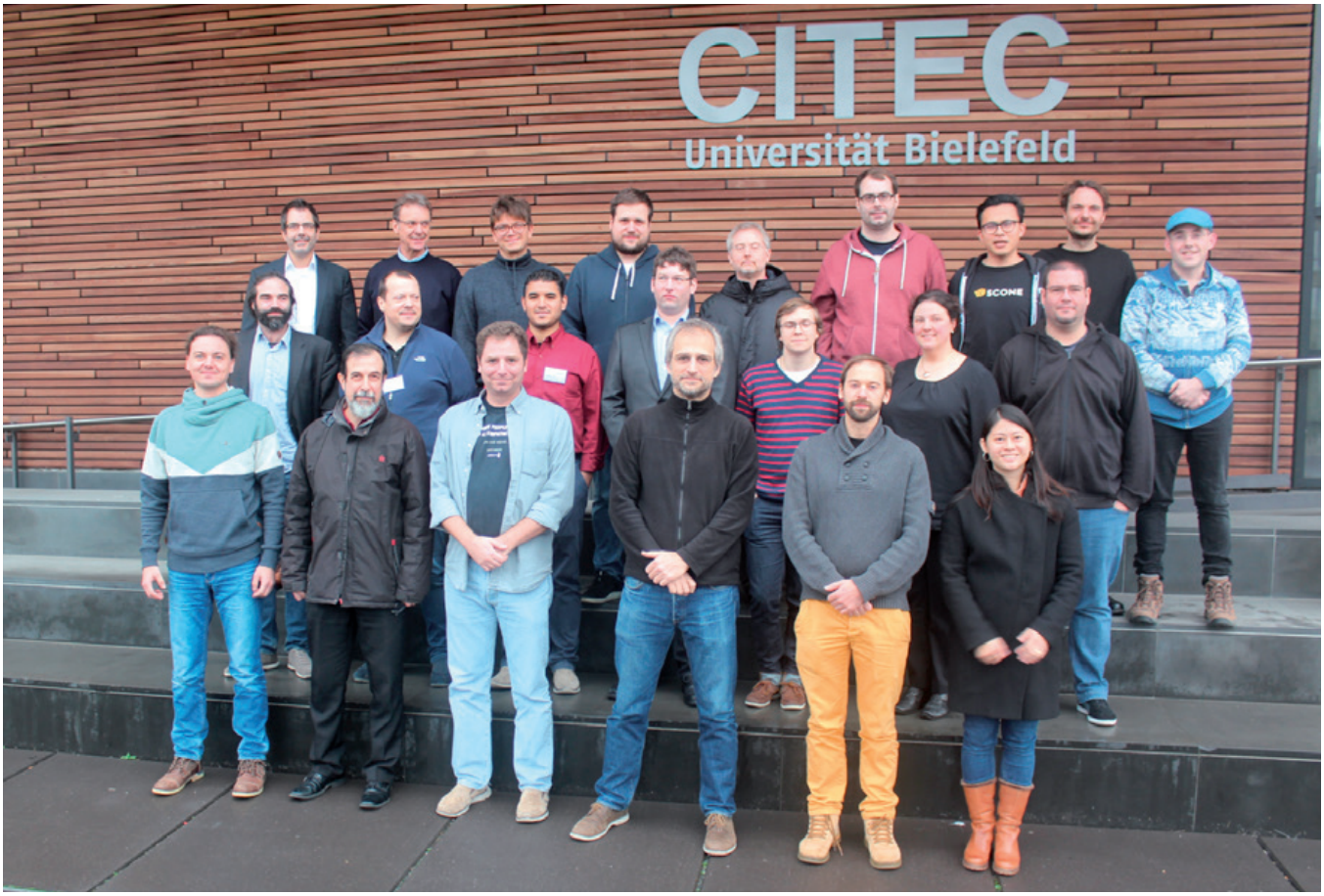
Think about how graphics processing units (GPUs) have transformed the HPC landscape; exascale levels of performance are currently dependent on these extremely powerful processors. While power hungry, with the topline around 300W, they are good for certain applications and offer good performance per watt.

If we consider processors on a continuum with general-purpose processors on one side, and application-specific integrated circuits (ASICs) on the other, ASICs offer far greater energy efficiency. However, ASICs are designed to run one specific application, such as speech recognition, and are unable to do anything else, as opposed their general-purpose cousins which can tackle a wide range of computing tasks while offering high performance.

FPGAs would come somewhere in the middle of this continuum: their energy efficiency is not as impressive as that of ASICs and their performance doesn't match that of GPUs, for example. However, these semi-custom hardware options allow you to tailor the application to the hardware and achieve relatively high levels of energy efficiency.

Working with FPGAs is like arranging different Lego bricks together. Different blocks deliver different functions – memory, compute, signal processing and so on – and you can assemble several of these to work together and work on different units of computation. This offers you the flexibility to match the compute offering to the application needs.

While FPGAs dissipate quite a lot of power, within LEGaTO we have been investigating techniques to reduce this. For example, a paper presented at MICRO 51 by BSC's Behzad Salami looked at the relationship between voltage supply and reliability. The paper found that, while reducing the voltage supply does result in reduced reliability, it is possible to run applications with similar levels of performance by ensuring that the less accurate results are the least important ones. This approach ▶



The LEGaTO consortium

works particularly well with neural network applications, for example, as neural networks are inherently resilient thanks to the built-in redundancy achieved by their many connections.

However, FPGAs are also very challenging to program; indeed, in the early days, programming an FPGA was akin to designing an integrated circuit. The emergence of high-level synthesis languages, such as the one developed by LEGaTO partner Technion, has helped this situation, while projects like LEGaTO are contributing to the software stack necessary to make FPGAs a viable option for a range of different applications.

Why is a toolset needed to fully exploit heterogeneous hardware resources while keeping energy consumption to a minimum?

For some time, people have been trying to take advantage of heterogeneous hardware and make it more energy efficient, including through the software used. While low-power solutions do exist, they are usually in system optimization and exist in isolation from other considerations such as security, reliability and programmability.

As computing systems increasingly populate the fabric of our everyday world – in cyber-physical systems such as self-driving cars, for example – issues such as security are gaining greater attention. Reliability is another area which has come under scrutiny as integrated circuits get smaller, with 3nm design also representing the limit of reliable operation. Meanwhile, as indicated above, programmability is becoming an increasing issue as heterogeneous hardware becomes more prevalent.

> *"LEGaTO aims to produce a toolset built around a single concept that draws upon various energy-efficient approaches"*

LEGaTO considers all these problems together and aims to provide a single software stack, taking a holistic approach through the programming model. The aim is to produce a toolset built around a single concept that draws upon various energy-efficient approaches.

Co-design is also a key plank of the LEGaTO approach, with our use cases guiding the technology research and allowing us to match the hardware to the application. For example, our small business partner, Machine Intelligence, specializes in energy-efficient neural networks. For the training phase of neural networks, which is compute-intensive, the kind of performance delivered by GPUs is required; however, the inference phase can be carried out on less powerful processors such as FPGAs and thereby save energy.

How does LEGaTO build on previous projects focusing on energy efficiency?

In earlier projects, such as ParaDIME, heterogeneous hardware was the new kid on the block, so we had to focus on modelling energy, predicting power and energy and consumption, and so forth. Given these energy profiles, we are now focusing on how a particular runtime mode could best exploit these profiles for a specific application – so moving to a higher level of the computing stack.

What are the main use cases in the LEGaTO project? Why were these chosen, and why are they so compelling?

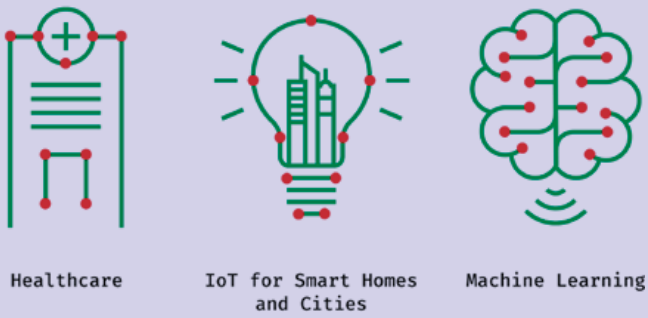
We wanted to focus on computational solutions for societal challenges which will become increasingly important in the future and for which we need to start improving energy efficiency now. For example, smart homes: with the ageing population, the LEGaTO smart mirror use case developed by Bielefeld is a great example of a solution to promote home independence. Or smart cities, as in our air quality use case in Barcelona, where air pollution data collected from sensors is processed using BSC's ALYA simulation software.

Machine Intelligence are working on self-driving technologies for the autonomous vehicles of the future, while at the Helmholtz Centre for Infection Research, researchers are discovering how using FPGAs can speed up their detection of biomarkers which identify certain diseases by up to 500 times.

LEGaTO

www.legato-project.eu

From smart mirrors to self-driving vehicles: The LEGaTO use cases



Machine learning

This use case develops a deep learning optimizer that uses OmpSs to reduce the energy consumption of deep learning inference significantly. The aim is to facilitate users to deploy deep-learning based AI on embedded systems in an energy- and cost-efficient manner. A startup company, EmbeDL, has been launched off the back of this work.

Hans Salomonsson, chief executive of EmbeDL, commented: 'The energy consumption of deep learning based artificial intelligence (AI) is an increasing environmental problem as more AI-based IoT systems are deployed in the real world. With the technology that we have developed in EmbeDL in the LEGaTO project, we have achieved multiples of energy improvements as well as execution speed.'

Biomarker discovery

This use case searches for a new method to discover biomarker candidates in datasets with a very low number of samples and up to thousands of features, which requires huge amounts of computational power. The LEGaTO technology with the usage of heterogeneous hardware provides the opporunity to enter new areas of research combined with a reduction of energy consumption in bioinformatics.

Sigrun May (Helmholtz Centre for Infection Research) said: 'With the LEGaTO technology, we can execute algorithms that otherwise could not be calculated in a reasonable time due to the calculation time of months or years.'

Smart city: air quality

The LEGaTO stack will leverage the processing capabilities and improve the energy-efficiency of an operational urban-scale air quality modelling system. This use case aims at demonstrating that monitoring of urban air quality through computational fluid dynamics simulations is feasible for nowcasting predictions in an operational workflow.

LEGaTO SmartMirror

Smart-home algorithms are often too complex and require too much computing power for energy-efficient usage. Within LEGaTO,

an energy-optimized processing platform is being developed to tackle this challenge. This is being used to power a smart mirror, which recognizes users and displays personalized information, and which could be used to support independent living for older people, for example.

Jens Hagemeyer, (University of Bielefeld) said: 'The tool flow and edge server platform developed within LEGaTO provide an order of magnitude improvement in energy efficiency for our SmartMirror system, all integrated in a compact, embedded processing module. In addition, it ensures local data processing, keeping your data private, without the need to upload it to a cloud environment.'

Secure IoT gateway

The secure IoT gateway will simplify the complexity of securing the connection of devices to a network, providing a network cockpit application for configuring and monitoring the system. This use case will help to help the other LEGaTO use cases by reducing the complexity of security.

Micha vor dem Berge (Christmann) commented: 'When thinking about security and privacy for IoT, the biggest fear is that your private data, such as a video of you walking in your house, could fall into the hands of somebody else. That's why Christmann is developing the easy to use secure IoT gateway within the LEGaTO project. It protects the whole data path from the IoT device to the destination by wrapping it into a virtual private network (VPN) tunnel. For an industrial context, advanced features like data filtering and mass-deployment are planned.'

SMART POWER MANAGEMENT: ADAPTING THE VOLTAGE AND FREQUENCY TO THE APPLICATION

Though it starts at the chip, the issue of power dissipation goes all the way up to full systems, says Kaijie Fan, a PhD student from Technische Universität Berlin (TU Berlin) who works with Professor Biagio Consenza (University of Salerno) and Professor Ben Juurlink (TU Berlin). 'In processor design, we're all aware of the power wall: the power consumption of a chip as a limiting factor for processor frequency increase. However, power is also a concern in large-scale computing systems, with next-generation computing systems needing to perform 10^{18} (a billion billion) calculations per second on a power budget of 20MW.'

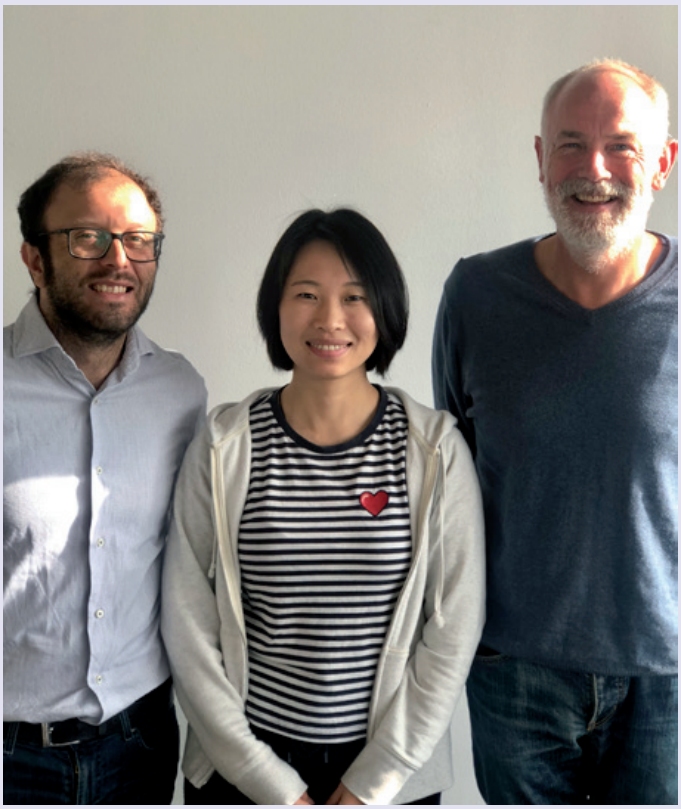
Meeting this target, she notes, will require major improvements in energy efficiency across the whole stack – from hardware, memory subsystem and interconnect up to the software level. 'Dynamic voltage and frequency scaling, where the voltage and frequency of the processor is scaled according to the requirements of the running application, is one of the most promising power management strategies,' says Kaijie.

The concept behind this technique is relatively simple, Kaijie explains: 'The energy consumed is proportional to the square of the operating voltage; therefore, by decreasing the power supply voltage, we can reduce energy consumption by the square of the voltage while only reducing performance by the voltage itself.' Measuring power and changing frequency can be done using libraries such as Running Average Power Limit (RAPL) on Intel processors and the NVIDIA Management Library (NVML) on modern graphics processing units (GPUs), she notes.

GPUs are a particularly interesting a case study for the performance/power relationship, says Kaijie. 'GPUs offer a significant improvement on peak performance and performance per watt in comparison to traditional multicore central processing units (CPUs). However, while power management for CPUs has been widely researched over a number of years, GPUs are a relatively new area of study in this field.' Hence further investigation could deliver rich results for high performance with lower power consumption.

Machine learning has the answer

Kaijie's research into dynamic voltage and frequency scaling uses machine learning to automate parts of this complex area. 'Different applications will have different "optimal" frequency configurations; for example, a compute-bound kernel will specifically benefit from lower memory frequency and higher core frequency,' she says. 'With the right problem formulation, machine learning is very good at understanding these code properties and can find a set of good (Pareto optimal) solutions.'



From left to right: Biagio Cosenza, Kaijie Fan and Ben Juurlink

However, applying machine learning to this problem is not simple, says Kaijie. 'It requires good training data, in our case based on synthetic, code-generated benchmarks tailored to GPU architecture, accurate feature representation of the input source code, in our case OpenCL kernels, and good modelling methods, which should model performance and energy with different regression models.' The machine learning approach can be easily ported across different GPUs, says Kaijie: 'You just need to re-build the model on the microbenchmark results executed on the new hardware.'

The team's research showed that performance and energy consumption on GPUs behave differently with frequency scaling and that it is important to model them using different approaches to maximize performance and minimize energy consumption. 'Evaluating the approach on twelve benchmarks on the NVIDA GTX Titan X, we found frequency configurations that deliver a 12% increase in performance on the same energy budget, or the same performance while saving 20% of energy, for example.'

This approach can be applied to GPU compilers as well as high-level programming models and libraries to provide more accurate application-dependent energy frequency tuning, says ▶

Kaijie. 'Furthermore, it is very easy to port to very different target architectures, from high-performance Tesla GPUs to low-power embedded GPUs.'

FURTHER READING:

'Predictable GPUs Frequency Scaling for Energy and Performance' ICPP 2019: Proceedings of the 48th International Conference on Parallel Processing. Kyoto, Japan – 5-8 August 2019
bit.ly/GPU_frequency_scaling

Embedded Systems Architecture at TU Berlin
aes.tu-berlin.de/menue/home_aes

Biagio Cosenza's research page
cosenza.eu

A video of Ben Juurlink discussing performance and power in the LPGPU2 project can be found on the HiPEAC YouTube channel:
bit.ly/HiPEAC19_Ben_Juurlink_LPGPU2

THE POWER OF THE NETWORK: VIRTUAL NETWORK FUNCTIONS GET ENERGY EFFICIENT

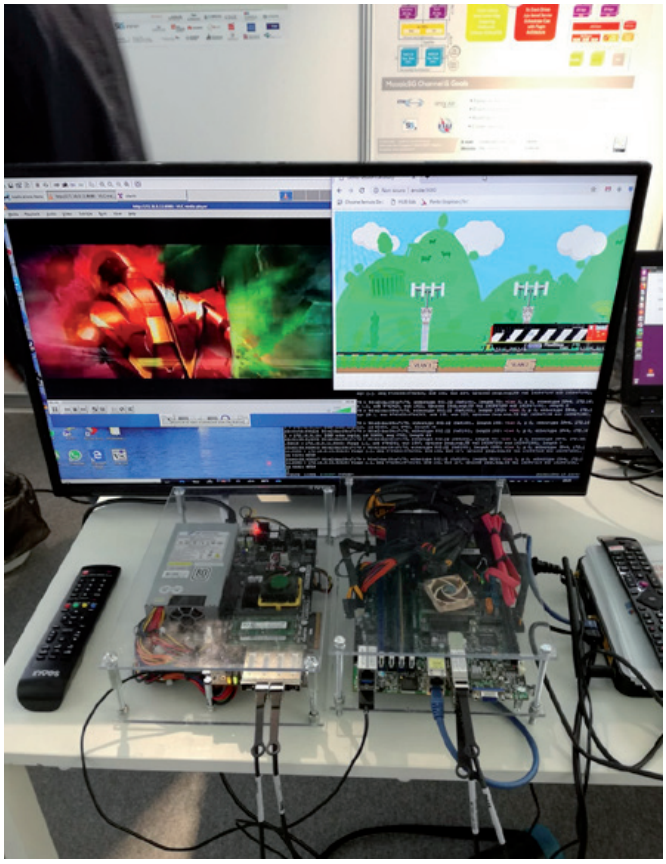
With all the focus in improving energy efficiency on the compute side, the network can sometimes get forgotten – but this is a mistake, argues Salvatore Pontarelli (CNIT – National Inter-University Consortium for Telecommunications, University of Rome 'Tor Vergata'). 'The amount of energy consumed by networking has been growing at a frightening pace over the last few decades, faster than 10% a year. This is despite significant improvements in the energy efficiency of equipment and operators' increased attention to energy costs,' he says.

There are a number of reasons for spiralling network energy consumption, says Salvatore. 'For example, we're seeing a growing number of users and terminals, increasingly data-hungry applications, increasing amounts of in-network computing and storage, more stringent performance requirements – especially in latency and security – and increased use of distributed software applications, for example based on microservices.' This makes it imperative to revisit network architectures and algorithms to improve user experience while at the same time taking the cost of energy into account, says Salvatore.

One way of addressing this problem is through network virtualization. As Salvatore notes, this approach allows a move away from a network composed of fixed functions and elements (firewalls, routers, load balancers and so on) towards a 'software-ized' infrastructure where all the network functionalities can be developed and deployed using commodity hardware. 'However, the price of this flexibility is a huge waste of resources, particularly energy, since commodity hardware is not optimized for the execution of network functions,' he warns.

Accelerating network functions while saving energy

Instead, Salvatore, along with colleagues at NEC and Axbryd, advocates offloading performance-critical network functions onto domain-specific programmable accelerators, which



lowBlaze demo at NSDI 19

'combine flexible programming with high performance and energy efficiency', he says. The team developed a framework for offloading virtual network functions (VNF) into FPGA-augmented network interface cards (SmartNICs). 'FPGAs are a natural choice, offering high throughput with low, predictable latency, as well as allowing their functions to be completely re-programmed as needed,' says Salvatore. 'This gives the infrastructure more flexibility: FPGA-based SmartNICs can be used as accelerators both for network functions and for other applications running on host systems as required.'

Some of the researchers who contributed to FlowBlaze: (left to right) Aniello Cammarano, Salvatore Pontarelli, Giacomo Belocchi

This framework, known as FlowBlaze, has been shown to deliver energy savings of 20% with respect to state-of-the-art VNF software offers. In addition, the FlowBlaze prototype, which provides programmability of network functionalities in the SmartNIC, was a great example of academia-industry collaboration. 'The academic environment [at the University of Rome "Tor Vegata"] allowed the exploration of disruptive ideas, providing the main abstractions and concepts. Axbryd, a spin-off of the university, was founded to commercialize the technology and to bring research prototypes to the level of maturity required by the market. NEC, as a Fortune500 company and large system integrator, contributed invaluable viewpoints on use cases and the market relevance of the technology,' explains Salvatore.

As for application areas, datacentre networking is an obvious choice, given the extremely low latency required, for example for high-performance computing tasks, says Salvatore. 'Other areas are emerging, such as cybersecurity, with so-called "zero-trust" deployments that need to enforce complex firewall policies at each infrastructure node. Responding to growing distributed denial of services (DDoS) cyberattacks, where the attacker interrupts internet services by flooding the target with superfluous requests, requires hardware that can scale to handle hundreds of gigabits of traffic, while applying complex network filtering algorithms. Finally, future 5G deployments, which are now leveraging datacentre-like infrastructure, are adding a number of use cases driven by the need to handle large volumes of network traffic.'

FURTHER READING:

Pontarelli, Salvatore, Roberto Bifulco, Marco Bonola et al. 'FlowBlaze: Stateful Packet Processing in Hardware.' NSDI 19: 16th USENIX Symposium on Networked Systems Design and Implementation, pp. 531-548. 2019.

bit.ly/NSDI19_FlowBlaze

ESTIMATING ENERGY IN EMBEDDED SYTEMS: THE POWOT SIMULATOR

Professor Ratcho Ivanov** and **Assistant Lubomir Bogdanov

Where embedded systems are concerned, parameters such as the pricing, size, marketing and distribution of embedded systems are usually taken care of by economists, analysts and mechanical engineers, notes Lubomir Bogdanov (Technical University, Sofia). 'For electronic engineers and software developers, two significant parameters are left: speed and power, with power becoming increasingly important in research since the early 2000s,' he adds. Since 2011, Lubomir and colleagues in the Laboratory for Embedded Microprocessor Systems at the Technical University, Sofia have been considering both hardware and software methods for reducing energy consumption.

'While the power requirements of an embedded system are negligible in comparison to a desktop computer, for example, the combined energy cost of billions of battery-equipped portable devices drastically changes the equation of consumed power,' says Lubomir. 'Software development is also affected, with developers aiming to reduce the power requirements of existing code.'

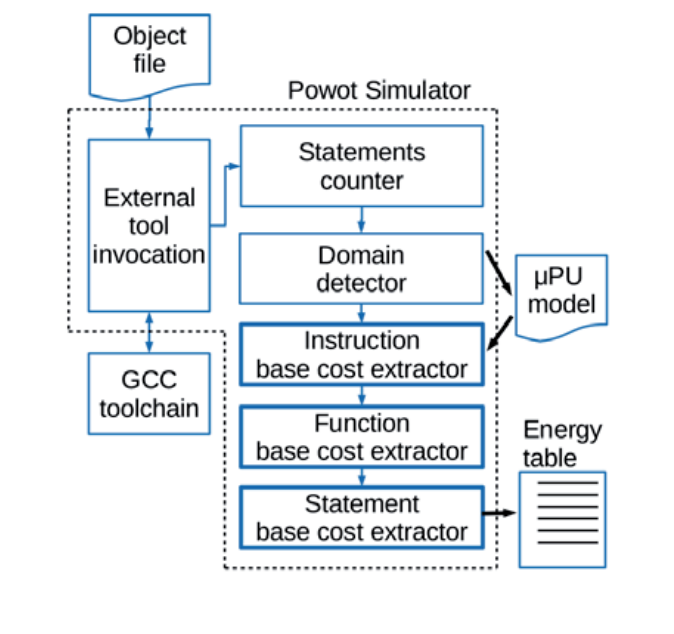
This is a problem, he says, because of a lack of tools helping programmers understand how much energy their code would consume. 'The only solution is to load the code in the target's memory and use measurement equipment to evaluate the consumption, and then try to predict battery life. This slows down the development process and forces the programmer to change the entire logic of the programme.'

Simulating power parameters

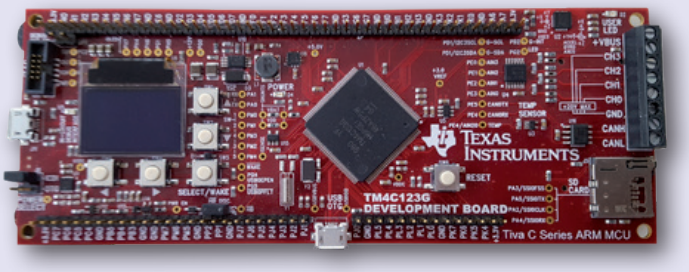
Hence the need for simulation, says Lubomir. 'Instruction set simulators (ISS) are used to estimate the energy consumption running on specific hardware by using previously measured parameters stored in files, or, as we call them, models. This allows the developer to run simulations and explore different versions of the code without requiring additional equipment.'

As Ludomir notes, many ISS capture the behaviour and parameters of a processing element on instruction level, meaning that no low-level information of the logic gates and their clock is present. 'However, creating such a model is not trivial and requires a lot of hardware insight that is not always available from the manufacturer,' he adds.

In response, the Laboratory for Embedded Microprocessor Systems has created the Powot Simulator, an ISS that hides the behaviour and operates at statement level of the C programming language. 'This means that an upward link between the assembler



Inner structure of the Powot Simulator



and the compiler has been created, allowing energy-related decisions to be made during the development phase. Given the corresponding instruction, or group of instructions, behind a C statement, we can use the model that contains the energy values behind each instruction. This allows us to find out the energy consumption of the statement in question.'

While the simulator currently uses the GCC toolchain, any other toolchain could be used, says Lubomir. 'The Powot Simulator is based on the Qt framework (C++) and the output file is a shared object (.so) that can be used in a bigger project, for example in an integrated developer environment (IDE). The resulting metrics are available to the user as an array containing the string of the C statement, its corresponding instructions, and energy values for each instruction.'

Future plans

In terms of future development, Ludomir says that the Powot Simulator could be extended to support cycle accurate simulations. It could also be upgraded by adding memory and peripheral energy models, which allow users to pinpoint power-hungry places on a chip or a whole system. 'The Powot Simulator originated as part of a bigger project – the POWer Optimiazatin Tool. The idea was to build a tool to make high-level, source-to-source optimizations in C programs for embedded systems, resulting in low power, energy-efficient code. Currently, firmware developers put a lot of effort into writing energy-efficient code; if we could automate this process, more and more embedded applications would use this kind of code,' he explains.

Finally, a further direction would be to refine the energy measurement of microprocessor-based devices. 'There are some existing solutions on the market, but building another one that could be interfaced by an ISS would complete the whole cycle of development, simulation and verification,' explains Lubomir.

Download the Powot Simulator from GitHub:
bit.ly/GitHub_Powot

In this article, Tamás Kerekes (NplusT) explains how his company partnered with PCB-Design – thanks to the TETRAMAX innovation action, funded by the European Union – to upgrade its NanoCycler technology for NAND characterization.

NAND characterization with NanoCycler

The use of solid-state storage devices is rocketing, driven by data centres, artificial intelligence applications, autonomous drive and other fields where large data volume, high computational power and low energy consumption are required. The core of the solid-state storage is NAND technology, which combines high density, fast access, low power and non-volatility.

Unfortunately, NAND memory devices are far from perfect. Their inherent failure modes include read errors, limited endurance and cell interference. Continuous evolution in pursuit of higher and higher density prevents the technology from reaching full maturity. In response, workaround techniques like error correction, wear levelling and read retry are used to overcome technology issues. Efficient workarounds require thorough knowledge of the NAND failure modes, which can be obtained through NAND characterization.

To this end, NplusT, a small Italian enterprise with many years' expertise in non-volatile memory testing, develops and markets the NanoCycler equipment, especially created for NAND characterization. NanoCycler has a 'tester-per-package' architecture, able to run independent experiments on each of the 48 simultaneously tested devices. Accurate, per-device temperature control and specific test intellectual property (IP) create an application-like environment for the devices under characterization.

The latest generation of NAND devices communicate with the host via a high-speed interface. Today's standard reaches 1.2 GT/sec which will double over the next two to three years. The test platform needs to keep pace with this evolution.

This is why NplusT partnered with PCB Design, a company based in Hungary and expert in all aspects of high-speed digital design, from the system level to the layout. This partnership was established with the support of BME MIT, a competence centre in the TETRAMAX consortium. After successfully applying for a TETRAMAX bilateral technology transfer call, the project started in June 2019 and finished in December 2019, with the successful characterization of the prototype and the transfer of the specific technology to the NplusT engineering team. During the execution, several critical technology issues were addressed, including signal integrity, picosecond-level fine tuning of the signal delays using field-programmable gate arrays (FPGAs), and algorithms to separate interface failures from real device failures.

The early feedback from the market demonstrates the validity of the project. None of the competing solutions is able to provide the required technical solution at a reasonable price. The upgraded NanoCycler technology is a leader in NAND characterization, where high-performance device management needs to be merged with accurate temperature control and scalability – at a reasonable price.

The technology which has been developed during this project will be the core of further solutions on NplusT's roadmap as a development environment for solid state device (SSD) firmware engineers and high-parallelism reliability test systems.

"The upgraded NanoCycler technology is a leader in NAND characterization, where high-performance device management needs to be merged with accurate temperature control and scalability – at a reasonable price"

Electric vehicles are riding a wave, and supersports motorcycles are no exception. Here, Isabelle Dor (CEA), Ana Gheorghe and Ramona Marfievici (both Digital Catapult) explain how Italian motorcycle manufacturer Energica is enhancing its machines with help from the innovation action FED4SAE.

The e-revolution starts here: Smart monitoring for electric motorcycles

With 2019 considered the 'base year' for the electric revolution in the global high-performance motorcycle market, it is no surprise that this year marks the inaugural season of the FIM Enel MotoETM World Cup. Energica Motor Company SpA, the first Italian manufacturer of supersport electric motorcycles, which combines innovation with the long tradition of the Italian Motor Valley, has been chosen as single manufacturer for the competition.

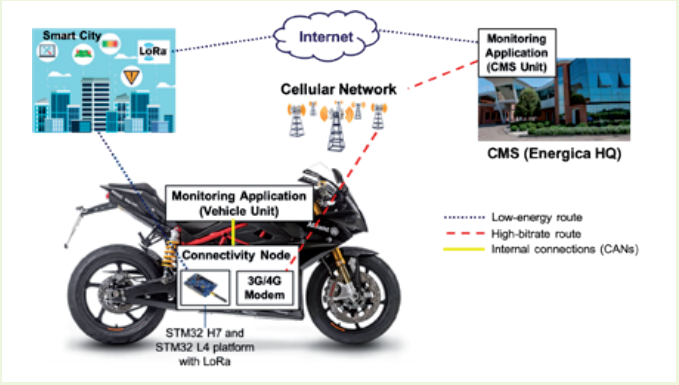
One recent major challenge – and opportunity – for Energica was to find a smart connectivity solution for electric two-wheelers which would enable a whole range of services. The company turned to the European Union-funded innovation action FED4SAE to support its application experiment MAMMUT, or 'Monitoring Applications exploiting on-board Motorbike's Multiple Transceivers', to design, implement and test a novel, smart connectivity architecture.

With this architecture, the company aimed to achieve the following goals:

- enable remote battery and vehicle monitoring
- support component and vehicle tracking during key-off periods
- provide geolocation services for motorbike and charging station localization
- allow real-time communication on the status of the vehicle and over-the-air updates during active riding periods

These led to four key requirements for the connectivity architecture. First, power consumption must be minimized when in key-off mode. Second, monitoring range must be maximized. Third, status and parameter updates during active and racing periods must be delivered in both a reliable and timely manner. Fourth, connectivity must be supported in multiple types of environments: dense urban, rural and indoor (i.e. basements and underground where vehicles are parked).

Energica approached these requirements by using both LoRa (long range) and 3G/4G communication technologies, as depicted in figure 1. This gives MAMMUT the low-power consumption, long-range and deep indoor penetration of LoRa, while building on the GPS-free geolocation functionality of LoRaWAN (LoRa wide-area network). In addition, it supports high throughput connectivity for high end services and vehicle firmware over-the-air (FOTA) updates through the cellular modem.



The core of MAMMUT is a novel platform with a STM32L552 MCU from STMicroelectronics, which leverages a LoRaWAN geolocation testbed and a time difference of arrival (TDOA)-based solver. The testbed with ten Kerlink iBTS gateways is deployed and managed by Digital Catapult in London. The localization solver for computing the final position is provided by the Swiss Center for Electronics and Microtechnology (CSEM) and uses probabilistic techniques to improve final precision.

Applications and services have already been implemented on a STM32 evaluation board, NucleoH7 with a LoRa-Discovery module, and will be easily ported to the final board.

In November 2019, a prototype was completed and tested during the preliminary trials of the MotoE World Cup in Valencia, Spain. In addition, power consumption and communication performance assessment 'in vivo' along with geolocation experimental campaigns in London are already on the 2020 agenda for MAMMUT.

FED4SAE is a three-year innovation action with a budget of €7.6 million, which started in 2017. Its partners are the French Alternative Energies and Atomic Energy Commission (CEA), Intel, STMicroelectonics, Thales, AVL, Digital Catapult, Fraunhofer, fortiss, the Swiss Center for Electronics and Microtechnology (CSEM), Stockholm Royal Institute of Technology (KTH), Budapest University of Technology and Economics (BME), University of Cantabria and Blumorpho.

fed4sae.eu

In the latest in our series on innovative HiPEAC companies, we find out how Worldsensing is powering operational intelligence and internet of things (IoT) monitoring solutions for risk management in critical infrastructure.

Worldsensing
Making sense of operational intelligence

COMPANY: Worldsensing S.L.
MAIN BUSINESS: Operational intelligence to traditional industries and cities, IoT monitoring system Loadsensing
LOCATION: Barcelona, London, Los Angeles and Singapore
WEBSITE: worldsensing.com
worldsensing.com/product/loadsensing
CONTACT: Dr Denis Guilhot, senior EU project manager
dguilhot@worldsensing.com

Worldsensing is a widely recognized global IoT pioneer. Founded in 2008, the technology provider delivers operational intelligence to traditional industries and cities. It provides IoT-based monitoring solutions for risk management in critical infrastructure such as buildings, bridges, tunnels, ports, wells, construction works and mining operations. It also offers seismic monitoring capabilities for engineering, oil / gas / water acquisition and CO_2 sequestration purposes.

Loadsensing is the leading wireless monitoring system allowing industrial companies to connect and wirelessly monitor infrastructure in remote locations. The data acquisition system is the industry reference for wireless geotechnical monitoring, as it is currently used to monitor over 50,000 sensors worldwide. Loadsensing is compatible with most geotechnical instrumentation and monitoring sensors and may be integrated with various kinds of data visualization software.

With over 100 employees and offices in Barcelona, London, Los Angeles, and Singapore, Worldsensing is active across the world, with 600 deployments worldwide and clients in over 60 countries across all continents. Worldsensing's investors include Cisco Investments, Mitsui & Co, McRock Capital and ETF Partners, among others.

The nature of Worldsensing's business makes it compulsory for the company to invest heavily in innovation, which represents an annual budget of almost €1 million. Worldsensing also ranks number 6 among Spanish small / medium enterprises in terms of participation and budget from the European Union's Horizon 2020 (H2020) research and innovation programme.

Within H2020, Worldsensing is leading the development and implementation of IoT technologies in energy harvesting and storage, integration of IoT with machine learning, secure maintenance of railway infrastructures, industry 4.0, smart cities and intelligent mobility, and integration of cloud computing and edge computing, to name a few. It also participates as a technology provider in several H2020 projects specifically aimed at improving cybersecurity frameworks in companies and in different kinds of critical infrastructure, such as water, to bring cybersecurity solutions to society.

WORLD SENSING

In this article, Daniele Cesarini (CINECA / University of Bologna) , Andrea Bartolini (University of Bologna), Carlo Cavazzoni (CINECA) and Luca Benini (University of Bologna / ETH Zurich) present their COUNTDOWN Slack runtime library for power management.

COUNTDOWN Slack

Reducing energy consumption at runtime while retaining high performance

Left to right: Daniele Cesarini, Andrea Bartolini, Carlo Cavazzoni and Luca Benini

High-performance computing (HPC) applications waste a significant amount of power in communication and synchronization-related idle times. By default, when processes are waiting in a synchronization primitive, MPI libraries use a busy-waiting mechanism. However, during MPI primitives the workload is primarily composed of wait times and input / output (I/O) or memory accesses, for which running a processor in low power mode may result in lower power consumption with limited or no impact on the execution time. MPI libraries implement idle-waiting mechanisms, but these are not used in practice to avoid performance penalties caused by the transition times into and out of low-power states.

A number of works have focused on strategies to maximize energy savings at the expense of performance. The main drawback is that they lead to a systematic increase in time-to-solution (TtS), which is not acceptable to users or facility managers. While methodologies have been proposed to reduce HPC energy with negligible or low impact on the TtS of running applications, these usually require complex toolchains, often based on prediction algorithms, which can lead to costly misprediction in irregular applications.

This is why, at the Energy-Efficient Embedded-System laboratory at the University of Bologna, in collaboration with CINECA, we developed COUNTDOWN Slack: a runtime library for profiling and fine-grain power management. An open library, COUNTDOWN Slack is based on a simple but effective strategy to reduce energy consumption in production HPC systems without performance penalties. COUNTDOWN Slack is able to scale down the core's frequency in slack regions of the application leaving unaltered the performance for both computation and data copy regions to avoid possible overhead.

COUNTDOWN Slack could benefit large-scale HPC systems like the Marconi supercomputer at CINECA

To extract the slack from MPI primitives, we propose a novel approach based on the insertion of artificial / instrumental barriers. This mechanism is agnostic to the MPI implementation, and it is built on top of standard MPI primitives.

We tested COUNTDOWN Slack in a real HPC system using a large set of HPC benchmarks extracted from the NAS parallel benchmark suite, and production runs of OMEN, a quantum-transport application which has twice been an ACM Gordon Bell finalist. We compared the proposed approach with state-of-the-art power management libraries, showing that COUNTDOWN Slack can preserve application execution time even in worst cases, while reducing the energy consumed by the compute units on average by 9.96% and up to 22.11%. According to our experimental results, COUNTDOWN Slack always leads to an energy saving (proportional to the communication slacks) with negligible execution time overheads (<3%).

FURTHER READING:

COUNTDOWN Slack on GitHub
github.com/EEESlab/countdown
Borghesi, Andrea, et al. 'Pricing schemes for energy-efficient HPC systems: Design and exploration.' The International Journal of High Performance Computing Applications 33.4 (2019): 716-734
arxiv.org/abs/1806.05135
Kerbyson, Darren J., Abhinav Vishnu, and Kevin J. Barker. 'Energy templates: Exploiting application information to save energy.' 2011 IEEE International Conference on Cluster Computing. IEEE, 2011
ieeexplore.ieee.org/document/6061058
Cesarini, Daniele, et al. 'COUNTDOWN Slack: a Run-time Library to Reduce Energy Footprint in Large-scale MPI Applications.'
arxiv.org/abs/1909.12684

The latest in our series on cutting-edge research in Europe finds out how De-RISC is commercializing European, RISC-V technology for space, how Maestro is delivering a solution to data movement for data-intensive applications, and how EPEEC is getting the most out of heterogeneous hardware for exascale.

Innovation Europe

LIFT OFF: DE-RISC TO CREATE FIRST RISC-V, FULLY EUROPEAN PLATFORM FOR SPACE

De-RISC
Dependable Real-time Infrastructure for Safety-critical Computer

The RISC-V instruction set architecture (ISA) has been attracting increasing interest thanks to its open-source nature. RISC-V allows the development of hardware platforms without needing to buy expensive licences and also circumvents any export restriction. This is particularly important in non-consumer markets where the sale volume is low, such as safety-critical systems in space or avionics.

By allowing Europe to develop its own hardware, RISC-V can also help reduce dependence on technology from the United States, which leaves Europe vulnerable to changes in policy or trade restrictions. As an example of this dependence, the GAIA (Global Astrometric Interferometer for Astrophysics) mission by the European Space Agency (ESA) was forced to use United States proprietary technology due to the lack of European alternatives.

Step forward De-RISC, which aims at producing one of the first fully market-ready RISC-V based platforms for the space domain and of European development. The goal of the project is to productize a multi-core RISC-V system-on-chip design already owned by Gaisler and port the XtratuM hypervisor owned by fentISS to that design to create a full platform consisting of hardware and software for future European developments within space and aeronautical applications.

By the end of the project, in 2022, the partners will have created a De-RISC board, radiation hardened and compatible with the European Space Agency's CoRA platform. On the software side, the project will deliver the XtratuM hypervisor to fully support the platform and new ISA. This will advance platform's readiness to technology readiness level (TRL) 8. In the avionics domain, the platform will be ready reach level B DO-178C certification.

'With the first RISC-V based, fully European platform for space, De-RISC will guarantee access to made-in-Europe technology for aerospace applications thus contributing to the "Technologies for European non-dependence and competitiveness" programme in these strategic markets,' said Paco Gomez Molinero, chief executive officer of fentISS and coordinator of the De-RISC project.

The project involves four partners. Cobham Gaisler will develop the hardware platform based around their previous designs of LEON processors, particularly the GR740 board. fentISS will port their existing hypervisor currently available for other ISAs to RISC-V. Barcelona Supercomputing Center (BSC) will provide safety and security features to the hardware design to

overcome the validation and certification challenges that new critical multicore embedded systems are facing. Finally, Thales will validate the design and software providing use cases for the verification of the complete platform (i.e. hardware, software stack including hypervisor and real-time operating system, and application). To that end, the main application used to validate platform will be telemetry and telecommand, which offer a wide range of processing characteristics from encryption to data compression and low-latency requirements.

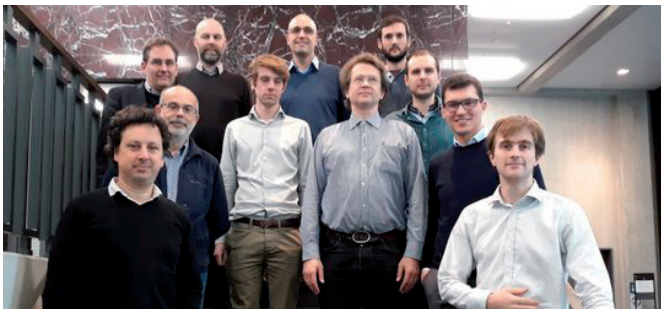
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 869945.

NAME: De-RISC: Dependable Real-time Infrastructure for Safety-critical Computer
START DATE: 01/10/2019
END DATE: 31/03/2022
KEY THEMES: RISC-V, system-on-chip, space, avionics
PARTNERS: fentISS, BSC, Thales Research and Technology, Cobham Gaisler
BUDGET: € 3,444,625
derisc-project.eu
@DeRISC_H2020_EU

SOLVING DATA MOVEMENT: THE MAESTRO PROJECT

Formed by Appentra, CEA, Cray, ETHZ/CSCS, ECMWF, Forschungszentrum/JSC and Seagate, the Maestro consortium addresses the ubiquitous problems of data movement in data-intensive applications and workflows. The consortium is supported by a three-year grant from the European Commission's H2020 Future Enabling Technologies for HPC (FETHPC) programme to build a data- and memory-aware middleware framework.

The Maestro project has been set up to tackle one of the most important and difficult problems in high-performance computing (HPC), namely the orchestration of data across multiple tiers of memory and storage hardware. Although data movement is now recognized as the primary obstacle to performance efficiency, much of the software stack is not well suited to optimizing data movement, and was instead designed in an age where optimizing arithmetic operations was the priority. The Maestro project aims to capture the data- and memory-aware aspects of applications and the software stack into a new middleware layer which will perform basic data movement and optimization on behalf of the application. Such capabilities will be crucial to facilitate efficient use of deeper memory hierarchies and tiered storage architectures.



The consortium includes a diverse set of partners that bring together strong expertise in HPC technologies and architectures as well as expertise in applications. The project endorses a co-design approach whereby the underlying technologies take the needs of applications into account; to this end it has selected a diverse set of relevant applications, such as numerical weather forecasting. For such applications, performance is increasingly constrained by the lack of fast data transport and efficient data orchestration capabilities.

Professor Dirk Pleiter, coordinator of the Maestro project, said: 'The Maestro project will provide a unique opportunity to challenge traditional approaches for handling data objects and data movements in complex HPC applications and workflows, which will be key for efficient exploitation of future exascale level supercomputers.'

PROJECT NAME: Maestro (Middleware for memory and data-awareness in workflows)
START/END DATE: 01/09/2018 - 31/08/August 2021
Key themes: computing systems, parallel/distributed systems, high performance computing, system software, data awareness, memory awareness
PARTNERS: Germany: Forschungszentrum Jülich; France: CEA; Spain: Appentra; Switzerland: ETH Zurich (CSCS), Cray; United Kingdom: ECMWF, Seagate
BUDGET: € 3,989,491.25
maestro-data.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme through grant agreement no. 801101.

ENABLING ENERGY-EFFICIENT COMPUTING FOR EXASCALE WITH EPEEC

Heterogeneity is enabling the performance and energy gains necessary for the exascale era, but programming heterogeneous hardware is notoriously difficult. Here, EPEEC Coordinator ***Antonio J Peña*** *(Barcelona Supercomputing Center) explains how the project will promote greater performance, productivity and energy efficiency.*

Why do we need an EPEEC programming environment for the exascale era? What's wrong with the programming models already in existence?

To reach exascale, we need to become much more energy efficient. A big part of this energy efficiency is expected to come from resource heterogeneity – by using specialized hardware components, such as processors or memories, we can be more efficient in certain tasks. However, heterogeneity is not easy to exploit efficiently, and requires considerable coding efforts with current programming environments. We are evolving European programming models, runtime systems, compilers, and tools so that they are aware of the upcoming resource heterogeneity and assist application developers to use it.

What are EPEEC's three main objectives? How has the project worked towards these so far?

First, programming productivity; second, performance; and third, energy efficiency. When targeting the former, it is expected that there will be some some performance loss in comparison to 'ninja programming', so we are also working on minimizing that loss. And we couldn't target exascale without being energy aware.

So far we have started targeting the first two main objectives. We have included plenty of features that will facilitate the use of a range of accelerators and memories. This is done by high-level programming based on directives, jump-start assisted by a tool suggesting the most appropriate directives, an optimized version of the automatically-distributed OmpSs flavour, and cool features for extreme scale in the GPI GASPI implementation.

Why is energy efficiency a particular problem for high-performance computing, and what specifically is EPEEC doing to address this issue?

Since energy consumption is the main concern for exascale, EPEEC is greatly concerned about it. We have planned to tackle this issue from the resource scheduling perspective. This task will be started in about a year's time, once our productivity and performance targeted developments are more mature.

What results have you had so far and what applications have you helped make 'exascale ready'? How will these applications improve quality of life, further scientific goals, etc.?

It is a little too early to see an impact on production runs, since we've just finished our first year, and we have only released early prototypes with limited functionality. EPEEC covers five big European applications representing different scientific domains, including fluid dynamics, combustion, nanophotonics, nanoplasmonics, plasma physics, material sciences and life sciences.

For now, these applications have been used in co-design efforts, meaning that they have helped those developing the programming environment understand their needs and design features accordingly. We are planning to release intermediate software prototypes next March, and those will be leveraged by our five applications to make them ready for upcoming supercomputers and hence enable science at a higher scale.

EPEEC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 801051.

EPEEC
PRODUCTIVITY AT EXASCALE

In this special edition of our regular career talk feature, José L. Abellán (Universidad Católica San Antonio de Murcia-UCAM) explains how a collaboration with colleagues in the United States both helped deliver an innovative new simulator for the community and enhanced his personal career.

Career talk special: Creating MGPUS

From September 2012 to August 2014, I enjoyed a post-doctoral position in the Integrated Circuits and Systems Group (ICSG), which is led by Professor Ajay Joshi, in the Electrical and Computer Engineering department at Boston University (USA). During that time, I was working on a two-year DARPA project titled 'Electro-photonic Network-on-chip Architectures in 1000+ Core systems', with Professor Joshi as the principal investigator.

The expertise we gained with this project in large many-core systems helped Professor Joshi initiate a research collaboration with Professor David Kaeli, who is an associate member of HiPEAC and leads the Northeastern University Computer Architecture Research (NUCAR) laboratory at Northeastern University (NEU). He is a well-recognized expert in high-performance computing (HPC) and graphics processing unit (GPU) architectures. In September 2014, I took up an assistant professor position with the Computer Science and Engineering Department at UCAM university and continued the research collaboration.

In October 2015, Professor Joshi and Professor Kaeli received a three-year National Science Foundation (NSF) collaborative research grant for the project 'Leveraging Intra-chip/Inter-chip Silicon-Photonic Networks for Designing Next-Generation Accelerators'. In this project, we combined our experience with silicon-photonic link technology and GPU architecture to optimize on-chip communications, and memory-hierarchy architecture in the context of both single- and multi- GPU architectures.

Our current team includes researchers from the United States (NUCAR and ICSG, led by Professor Kaeli and Professor Joshi, respectively), from South Korea (Professor John Kim at the

José L. Abellán has been collaborating with Ajay Joshi (left) and David Kaeli (right)

Sim through transatlantic collaboration

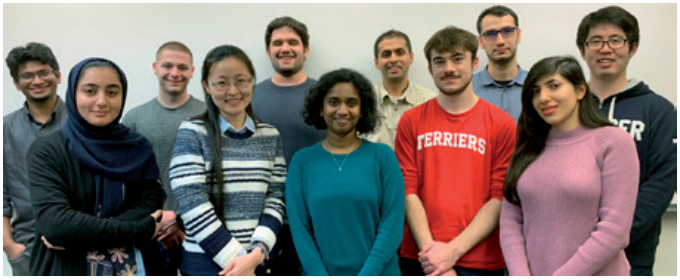
Korea Advanced Institute of Science and Technology-KAIST), AMD Corporation, and myself from Murcia, Spain (UCAM). Thanks to this joint effort, last year our team built a novel multi-GPU simulator called MGPUSim that was recently accepted for publication at ISCA 2019, and we are preparing a tutorial on this simulator for HPCA 2020 that will be held in San Diego, CA in February 2020.

I am very fortunate as I maintain an active collaboration with the team via regular weekly meetings over Slack. Undoubtedly, this is providing me with new opportunities in my career and has enhanced my skills as a professor and computer architect. In March 2017 I became a full member of HiPEAC, and at the beginning of this year I was promoted to associate professor.

Creating MGPUSim, a multi-GPU simulator

A multi-GPU system is a promising energy-efficient computing platform that can provide the required computing power demanded by today's large-scale data-driven applications. As a result, both industry and academia are looking for better multi-GPU solutions. For example, NVIDIA ships DGX-1 and DGX-2 systems, integrating up to 16 GPUs in each node. The systems are targeted primarily at deep neural network workloads. Similarly, AMD integrates four MI25 GPUs in its TS4 servers to accelerate deep learning applications.

However, the computer architecture research community does not have an open source, flexible, high-performance and reliable multi-GPU simulator. Current publicly-available GPU simulators were originally developed for single GPU platforms and cannot support simulation of state-of-the-art multi-GPU platforms. This is mainly because: 1) existing GPU simulators simulate dated GPU architectures and cannot easily model multi-GPU communication features; 2) existing simulators lack modularity and extensibility, making modelling and configuring a multi-GPU platform a tedious task; and 3) existing simulators are not efficient in terms of simulation speed. So researchers are disadvantaged when studying multi-GPU systems.



Ajay Joshi's research group

David Kaeli's research group

To address this, we created MGPUSim, as described in *HiPEACinfo 58*. We validated MGPUSim against real hardware with an error as low as 5.5% when compared with real GPU execution. The value of MGPUSim is not limited to only multi-GPU system simulation, but can be used to drive studies on state-of-the-art single-GPU performance. We have released MGPUSim as an open-source tool under the MIT License, so researchers from both academia and industry can take advantage of a flexible, high-performance and reliable simulator for their research.

Our plan is to continue our fruitful collaboration and apply for funding under the umbrella of designing optimized multi-GPU systems. Professor Kaeli and Professor Joshi have recently submitted an NSF Computer and Network Systems four-year project proposal where, if granted, I will be participating as an external collaborator. We will also look for other calls in Europe.

Now in its fifth year, the HiPEAC Student Challenge is a constant source of inspiration, with students coming up with solutions for societal problems from finding a seat on a train to tracking missing people. We caught up with Gianluca Giuffrida, Silvia Panicacci, Gabriele Meoni and Marco Marini from the University of Pisa to find out why they got involved, what projects they are working on and what their future plans are.

'It's beautiful to create something

Why did you decide to participate in the HiPEAC Student Challenge?

Gianluca: After hearing about the Student Challenge at the ACACES summer school, we decided to get involved in the next edition, which was at Computing Systems Week Bilbao.

Silvia: We also took the opportunity to display a poster on work for our PhDs, so that we could get feedback from more senior researchers.

Marco: When I started my master's thesis, I wanted to focus on artificial intelligence (AI) and helping people. My supervisor, Luca Fanucci, suggested developing assistive technology for disabled people, so I studied the problem and the literature before going ahead. The Student Challenge seemed like the perfect opportunity to share it.

Gabriele: The format of the challenge is great because you can feedback from others and discuss each other's projects.

Silvia: It's nice to have an intimate group where you can discuss the projects. At ACACES, for example, you get a chance to talk about your work in the poster session, but during the classes it's more about listening to presentations.

Left to right: Gianluca Giuffrida, Silvia Panicacci, Gabriele Meoni and Marco Marini

Gabriele: Yes, ACACES is great for hearing about professors' new ideas and finding out about new research directions.

Tell us a bit about your projects.

Gianluca: In our project, Parloma, Silvia and I are developing an AI-based robotic arm which allows deafblind people talking in sign language to communicate remotely. This is an extremely rewarding project; it is really beautiful to create something which has the power to change someone's life.

The system has two cameras, which produce a combined image which is sent to the cloud. We used AI to extract 18 key points on the body and on the hands, so that we can represent the whole body of the signer. The system then translates the Cartesian coordinates of the key points to joint angles to control the robot arm.

Silvia: Our demo works like a walkie-talkie. When the signer stops sending images – that is, messages – the robotic arm starts moving. We've achieved good results in terms of time, but we'd like to aim for real-time communication in the future.

In terms of the technical challenges involved, the filtering of the depth image was tricky. As the cameras are low cost, the image isn't very accurate – it's very noisy and needs to be filtered and RGB aligned.

Gabriele: The project I'm working on with Marco is about automatic speech recognition for people with dysarthria. For a range of reasons, such as cerebral palsy or a brain injury, for example, this condition causes people to have difficulties articulating, meaning that their speech may not be recognized by commercial speech-recognition systems.

Marco: With deep learning systems, you always need a lot of data, but in the case of detecting the speech of people with dysarthria, you need even more. We don't have enough data at the moment for continuous speech, but we are collaborating with hospitals and medical facilities which work with dysarthric people to find people who can record their voices.

which can change someone's life'

Gabriele: In addition to working with specialist centres, our colleague Davide Mulfari has been developing an app called CapisiciAMe, which allows dysarthric speakers to send voice recordings. The first demos are now available, with about 30 voices per word. The more voices we get, the more we can improve the speech recognition system. By the way, donations towards the development of the app are very welcome!

We're currently focusing on controlling household appliances with the app: voice-activated devices will be increasingly important in the future, and they are often particularly useful for disabled people.

What about plans for the future?

Silvia: In addition to enabling real-time communication, we'd like to integrate all the functions at the edge. At the moment, we use a completely cloud-based architecture, which can create latency issues. Integrating the inference stage onto an NVIDIA Jetson Nano, or the Intel Movidius Neural Compute Stick, for example, we could get a faster response.

Gianluca: Another advantage to edge computing is, of course, that it provides a response to privacy issues, as the images wouldn't be uploaded to third-party servers. We foresee a situation where we would just use the cloud for multiclient connection, similar to Skype. With two robotic arms and four cameras, it would be possible to communicate between two deafblind people, for example.

Silvia: In terms of commercializing the system, we've tried to keep costs down; the robotic arm itself is 3D printed, while, as mentioned above, we've used commercial cameras, so the whole system could be priced at around €2,000. To help us with the engineering aspects we've been working with biomedical engineers at the Scuola Superiore di Sant'Anna.

Gabriele: As for our speech-recognition system, we're planning to enhance it so it works for continuous speech, not just single words, although this will be challenging.

Marco: With regard to our speech-recognition system, we're planning to develop a software classification tool for speech therapists and doctors, called Recordia. Currently, there is no objective procedure for classifying levels of dysarthria, and we think this would be very useful. We have weekly meetings with speech therapists who make suggestions for improvements.

In addition, the system could be adapted to detect other kinds of speech which vary from stereotypically 'standard' forms, such as marked regional accents. It all depends on how you collect the data, manage the data and train the AI system.

Where do you see yourselves in a few years' time?

Silvia: I'd like to be doing research, although possibly in an industrial rather than academic setting.

Marco: I'd like to be working in the research and development section of a company. In terms of the near future, I plan to do a research stay at the Centre for Assistive Technology and Connected Healthcare at the University of Sheffield.

Gabriele: I would like to continue my research – I'd also like to do a post-doc abroad, to help improve both my research and soft skills.

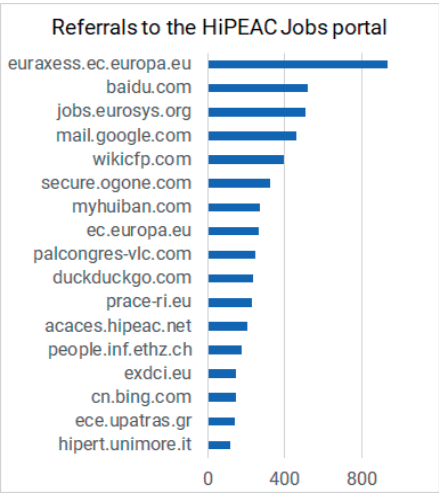
Huge thanks to Intel Movidius and Arm, who sponsored the HiPEAC Student Challenge with hardware gifts

We're often asked how jobs perform on the HiPEAC Jobs portal, so HiPEAC Jobs' Xavier Salazar (Barcelona Supercomputing Center) decided to share some basic insights from the portal, along with some suggestions to help your opportunities get the highest traffic. Statistics are from January to October 2019.

Maximize the performance o

With over 120,000 visits a year (or around 10,000 visits per month), the portal has been optimized so that it receives a significant amount of organic traffic from search engines. Among these are a decent number of direct visits, which means that we have loyal users who come to our portal over and over again. In addition, we get a relatively high number of referrals, along with redirects from social media and email, which demonstrates the impact of the campaigns, especially for attracting new users to the portal.

Taking a closer look on the referrals side shows how the connection with the Euraxess jobs portal, whereby jobs posted on HiPEAC Jobs automatically appear on the Euraxess website, has been very successful – it's already the website bringing most traffic to the HiPEAC website. Meanwhile, links from the Eurosys jobs portal, PRACE and EXDCI show how we are getting the right specialized traffic from advanced computing systems and high-performance computing (HPC) domains. Other sources of traffic are HiPEAC members' institutional websites, while we also have a notable number of visitors from the European Commission's website.



In terms of social media, it's no surprise that LinkedIn and Twitter are the most common sources of traffic, since they are platforms actively supported by HiPEAC. More surprising is the number of visitors coming from Facebook, which shows how our content is shared across other platforms.

Finally, when looking at the countries where the visits are coming from, we see that European countries, mostly countries of HiPEAC members (the UK, Germany, Spain, France, Italy, Greece and the Netherlands), feature prominently in the top 10. However, surprisingly, the number one country is the United States, while other non-European countries such as China and India are also in the top ten. This shows the international dimension and major impact of the portal not only within Europe, but also overseas.

Job posts which took off

In addition to evaluating the overall picture, we analysed the best performing job posts during the month of October 2019, which can help highlight strategies to maximize the impact of your vacancies.

We found that the job post which has attracted the most visitors to the portal over time is a perennial post that has been on the portal for some time, with the deadline being frequently renewed so that it doesn't expire, while ensuring it stays among the top positions on the portal and benefits from regular, automatic announcements on HiPEAC's social media channels. It is indexed on Google and has received many organic visits. Peaks in traffic can be observed when the poster of the job shared the vacancy via LinkedIn.

Another job post which saw a peak in traffic during October was boosted by being sent to mailing lists, such as those at universities and student associations, as well as being shared on social media. Indeed, the selected candidate learned about the job due to this promotional effort.

In the case of another vacancy, it was viralized at the beginning, meaning that it received a huge peak in traffic, while over the longer term it was likely indexed by Google. The job was well described, making it more attractive to potential applicants.

Another important point to ensure the job post performs well is local support, such as a university careers centre. Some universities actively promote HiPEAC, including sending regular jobs posts to internal mailing lists and recommended websites, printing and posting HIPEAC Jobs posters and displaying jobs on panels at the university. This results in a huge payback in the good performance of their postings.

FOLLOW US

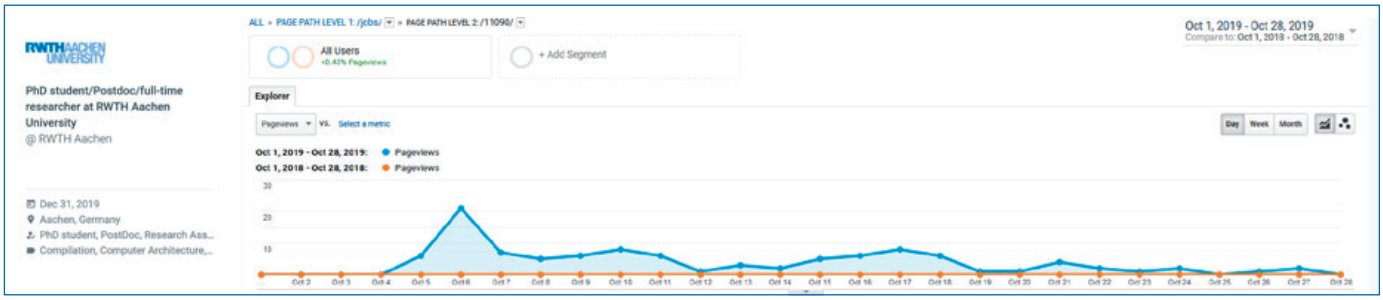
hipeac.net/linkedin

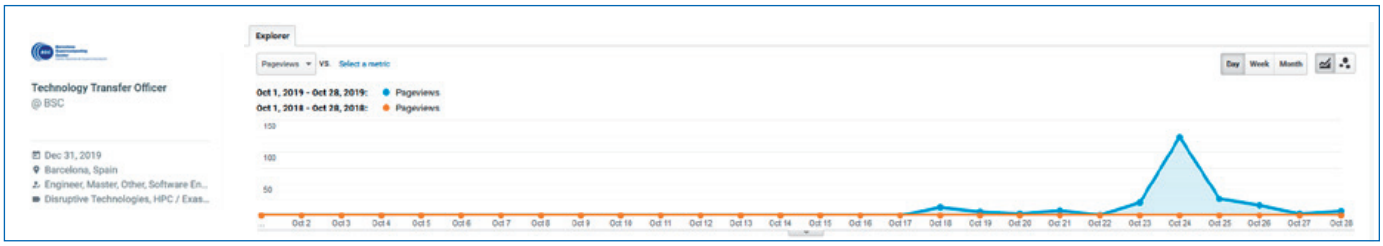
@hipeacjobs /@hipeac

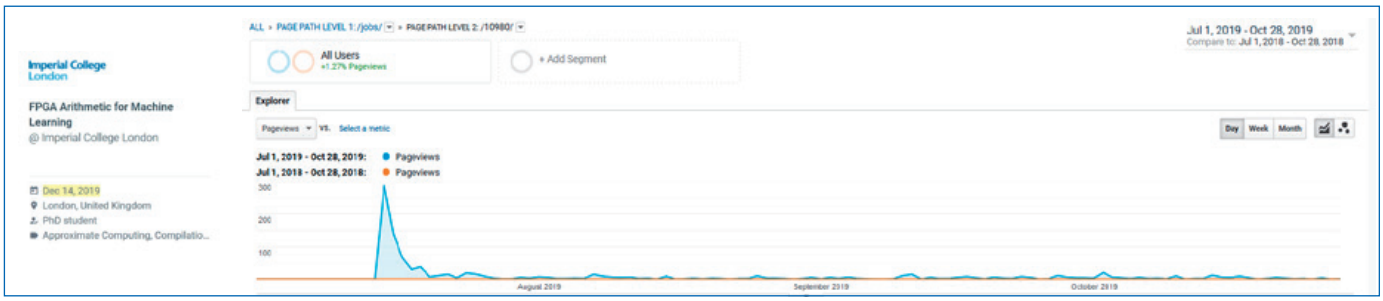
hipeac.net/jobs

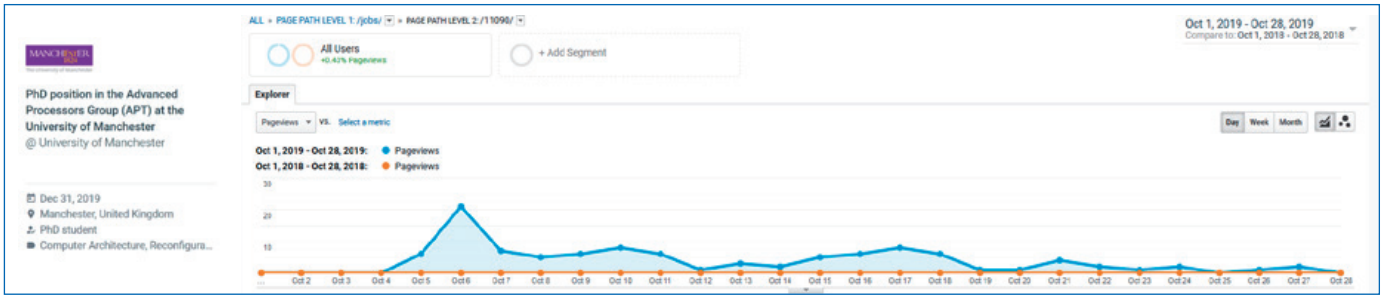
recruitment@hipeac.net

f your posts on HiPEAC Jobs









Top tips to maximize the performance of your job posts

1. Take ownership of the job opportunity. If you are a researcher you are best placed to understand your own needs, as well as the skills and profile you are looking for. The HiPEAC Jobs platform helps you to get into contact directly with the specialist people you need, without intermediaries. A good description of the opportunity is essential.

2. Be proactive. The portal has many tools to help you to share and spread the word. You can share it to social media automatically, as well as forwarding the vacancy to your preferred mailing lists. It pays off to do it, and HiPEAC makes it easy

3. Involve your institution. Tell your university career centre, your human resources department or talent manager about the HIPEAC jobs portal – it can be very useful for them.

Want to expand your horizons while gaining new skills? A HiPEAC internship is a great way to do so – check out the full list of opportunities on the HiPEAC Jobs portal: hipeac.net/internships

In this issue, Konstantinos Blantos explains how he helped develop surgical magnifying glasses during his internship at Campera Electronic Systems Srl, an Italian company which specializes in the development of high-performance, field-programmable gate array (FPGA)-based embedded systems.
Check out *HiPEACinfo 57* for our SME snapshot profile of Campera.

HiPEAC internships: your career starts here

NAME: Konstantinos Blantos
RESEARCH CENTRE: Aristotle University of Thessaloniki
HOST COMPANY: Campera Electronic Systems
DATE OF INTERNSHIP:
01/07/2019 - 01/10/2019

As a Master Student in Electronics in Aristotle University of Thessaloniki, I was looking for a challenging experience in an industrial working environment. At Campera Electronic Systems I had the privilege to work on a cutting-edge project and, at the same time, use the findings and the knowledge acquired for my MSc thesis.

In particular, I worked on the development of a set of surgical magnifying glasses with real-time video stabilization and the addition of an augmented reality overlay. The system consists of a set of magnification glasses that utilize an optical shutter in order to project the processed real-time video stream. The system also includes a belt on which the main processing unit, a Zynq Ultrascale+ FPGA, is placed. On this device the video stabilization and the augmented reality algorithms will be implemented.

In my role as an FPGA developer intern, I was involved in the study and design of the main processing unit implementation on the Zynq Ultrascale+ FPGA. Specifically, I worked on the implementation of the top-level frame buffer and its sub-modules that are an indispensable part of the optical acquisition system. This frame buffer allows the system to write and read more than one frame at a time in parallel. This module is a critical part of the implementation as its performance is crucial for achieving real-time video stabilization and processing.

A deeper approach

Following the Campera Electronic Systems HDL coding standards, I developed a generic and vendor independent module. The system itself can operate in various screen resolutions, different clock timings, different number of frames stored in the memory and other different parameters. Furthermore, within the frame buffer structure I developed two sub-modules: one to collect and process the incoming pixel data from the visor video-cameras and one to transfer the data to an external memory by using the Xilinx AXI bus. In addition, a control unit was developed to arbitrate processing and data flow.

Photo credit: © Wavebreakmedia Ltd | Dreamstime.com

I also developed all the necessary test benches in VHDL to test, debug and verify the functioning of the various components and the system as a whole. The project was developed using Xilinx Vivado and Aldec Active-HDL, while the coding standards and code quality was verified using Aldec Alint-PRO.

Calliope-Louisa Sotiropoulou, Campera Electronic Systems' research and development manager, said: 'HiPEAC has provided us with a great way to find highly qualified candidates with suitable skills in embedded systems and FPGA development. Kostas' work was of great value to our research project and we are looking forward to extending this opportunity to other interns in the future.'

CAMPERA ELECTRONIC SYSTEMS

The HiPEAC network includes almost 1,000 PhD students who are researching the key topics of tomorrow's computing systems. In this issue, we find out how Eva García-Martín's thesis tackles the issue of energy efficiency in machine learning, which is ever more prevalent and energy hungry.

Three-minute thesis

Featured research: Machine learning gets energy efficient

NAME: Eva García-Martín
RESEARCH CENTRE: Blekinge Institute of Technology
SUPERVISORS: Niklas Lavesson, Håkan Grahn, Veselka Boeva and Emiliano Casalicchio
DATE DEFENDED: 06/2018
THESIS TITLE: Thesis title: Energy efficiency in machine learning: Approaches to sustainable data stream mining

Energy efficiency in machine learning explores how to build machine learning algorithms and models with low computational and power requirements. Although energy consumption is starting to garner interest in the field of machine learning, the majority of solutions still focus on obtaining the highest predictive accuracy, without a clear focus on sustainability.

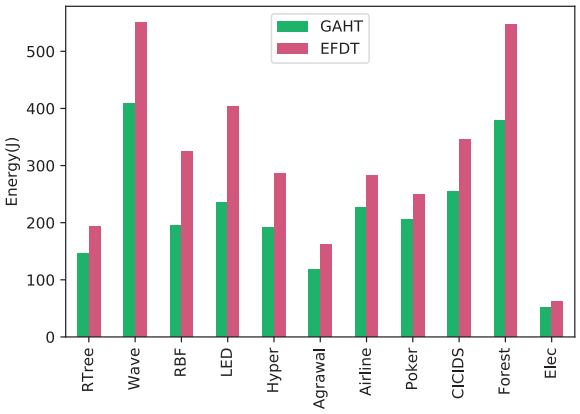
This thesis explores green machine learning, which builds on green computing and computer architecture to design sustainable and energy-efficient machine learning algorithms. In particular, we investigate how to design machine learning algorithms that automatically learn from streaming data in an energy-efficient manner.

Energy as a key metric for algorithm design

Energy consumption is a measure that is challenging to obtain, especially in the context of computer programs. Although software energy consumption had widely been studied in the context of computer architecture, this has not been the case for machine learning.

To address this issue, this thesis first illustrates how energy can be measured in the context of machine learning, in the form of a literature review and a procedure to create theoretical energy models. We then use this knowledge to analyse the energy footprint of one class of streaming algorithms, presenting an energy model that maps the number of computations and memory accesses to the main functionalities of the algorithm.

The uniqueness of this thesis lays on setting energy efficiency as the key metric when designing machine learning algorithms.



Energy comparison between the GAHT (Green Hoeffding Adaptive Tree, our algorithm) and the EFDT (Extremely Fast Decision Tree, the one we compare against), per dataset and per number of instances

This is showcased by two novel extensions of online decision tree algorithms. These solutions are able to reduce the energy consumption of the original algorithms by twenty to thirty percent, with minimal impact on accuracy. This is achieved by setting an individual splitting criteria for each branch of the decision tree, spending more energy on the fast growing branches and saving energy on the rest.

Energy efficient machine learning solutions open new directions towards moving machine learning to the edge. Edge machine learning focuses on building tiny machine learning algorithms with low computational, power, and energy requirements that can run on embedded devices. This improves privacy, since the data does not leave the device, and enables on-device machine learning in those devices with no internet connection.

We believe that this thesis provides the necessary tools to inspire and encourage researchers in the field of machine learning into a greener future.

Eva's supervisor Håkan Grahn commented: 'Energy consumption is of paramount importance in today's computer systems, and this thesis shows how energy consumption can be significantly reduced in machine learning algorithms while still maintaining the same or similar accuracy.'

Thanks to a record number of sponsors for making #HiPEAC20 a great success!

arm
CINECA
DeepMind
ECMWF
GREENWAVES TECHNOLOGIES
HUAWEI
hoCOMPUTER compiler & tools
monetdb
INFN
SYNOPSYS
United Technologies Research Center
XILINX
ZeroPoint TECHNOLOGIES
E4 COMPUTER ENGINEERING
NVIDIA
TEAMPLAY
epi European Processor Initiative
MONT BLANC
CAEN Tools for Discovery
COBHAM
codeplay
ECSPEC
emmtrix Technologies
MediCon
OPENHW GROUP
Springer
life.augmented
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA
THALES
UNIVERSITÀ DI PISA
evolve

Sponsors correct at time of going to print. For the full list, see hipeac.net/2020/bologna

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. 779656