HiPEAC
COMPILATION ARCHITECTURE
INFO 50
APPEARS QUARTERLY | APRIL 2017

April 2017: Computing Systems Week, Zagreb

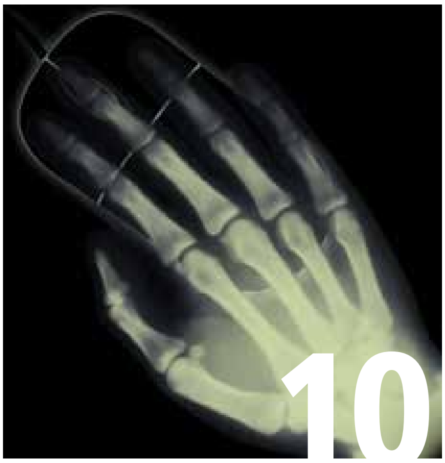
Advancing the digital healthcare revolution

A quantum computing breakthrough

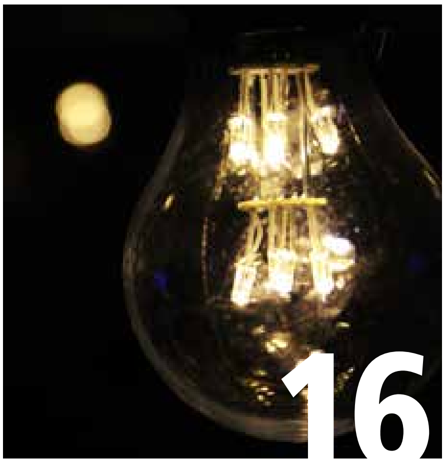
HiPEAC Technology Transfer Award winners

contents

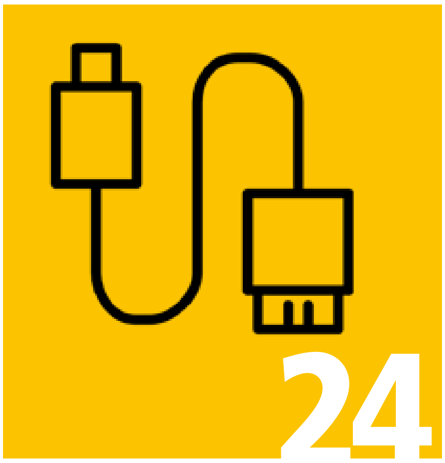
37 nations represented at HiPEAC17



Bringing the computing revolution to healthcare for a changing population



Innovation Europe



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.687698.

Design: www.magelaan.be
Editor: Catherine Roderick, Madeleine Gray
Email: communication@hipeac.net

The internet is disrupting everything… and fast. As more and more information, both recent and historical, becomes available, and as search engines become more powerful in interpreting unstructured information on the internet, our privacy is being invaded in unprecedented ways. Even if you do not disclose any information about yourself on social media, this will not stop others from sharing information about you. Denying that you know somebody is pointless if you appear in the background of a selfie taken by a tourist while talking to that person. High resolution pictures can reveal information that is not visible to the naked eye like messages on a smartwatch or a smartphone, or notes jotted down on a piece of paper. Even confidential documents get disclosed on WikiLeaks. Cover-up operations are often failing because it is very difficult to delete digital evidence on the internet.

The consequence is that candidates who run for highly competitive elective offices become very vulnerable. With millions of eyes zooming in on all available information, there are always things that can be used by an opponent to damage a candidate. On social media, anybody can create a storm based on real or fake news. Messages are copied, liked or retweeted at the speed of light. By the time facts have been checked and analysed, the damage to a reputation has long since been made. There are no places to hide from such a storm on the internet. Recently, there seems to have sprung up a new generation of politicians who have developed a strategy to deal with this situation. Instead of defending themselves, they just ignore the news, calling it a conspiracy, not relevant or fake, and continue their business as usual. The internet is known to cause disruption in many sectors. Could this be sign of the beginning of disruption in politics, a disruption that might make the political profession harsher and in which only the toughest men and women can survive and thrive? If this were to be the case, it is definitely not the disruption I was hoping for.

The theme of this HiPEAC magazine is health. Health is the second biggest market for embedded systems in Europe (after automotive and before military and aerospace). This means that developing IT-solutions for challenges in healthcare is a very good opportunity to generate impact. I wish you pleasant reading and I hope that the research and innovations presented in this magazine will inspire you.

Koen De Bosschere, HiPEAC coordinator

An update on European po

Sandro D'Elia of the Technologies and Systems for Digitising Industry unit at the European Commission updates us on progress in the various EU digital initiatives.

Most of the work of my office in the European Commission is centred on 'Digitising European Industry', the initiative aiming to 'ensure that any industry in Europe, big or small, wherever situated and in any sector can fully benefit from digital innovations to upgrade its products, improve its processes and adapt its business models to the digital change'. This started in 2015, and it's time to look back at what we have learned.

The most important feedback that we have got in the last few months is the enormous interest in the initiative across Europe. There are lots of meetings, events and workshops on this subject, across all industry sectors from manufacturing to health, transport or energy, and the message we get is invariably the same: this is something being taken very seriously and that is greatly needed for the future of Europe. Everybody is aware that there is no other option: European industry has to embrace digital technologies to stay innovative, and has to stay innovative to survive.

"EU-level cooperation is needed to achieve results"

A second message that we get is the widespread awareness of the possible negative impact of digitization on employment. We know that many jobs have already been replaced by computers, and that even more jobs will be replaced in the future by cyber-physical systems with varying degrees of autonomy, or by artificial intelligence. To create the new jobs that will replace the lost jobs, Europe needs digital skills across all sectors: not only programmers but also people capable of interacting with robots, training neural networks and generally using the technology of tomorrow in any aspect of life and work. This requires collaboration between the education system and industry; the HiPEAC community, which has one foot in the world of industry and the other in academia, can play a significant role.

The third message is the need for collaboration. The country most advanced in the digitization of industry is probably Germany, which invented the concept of 'Industrie 4.0', but even its government clearly says that this cannot be a national effort. EU-level cooperation is needed to achieve results, and the network of 'Digital Innovation Hubs' that we are trying to build will play an important role in spreading digital technology across all regions.

Of course, this requires adequate investment. The European Commission contributes directly through the Horizon 2020 programme, which dedicates several challenges to the digitization of industry. For example I4MS (Innovation for Manufacturing SMEs) aims to create innovation hubs and transfer technology to SMEs across Europe; other challenges aim to fund the development of digital industrial platforms.

It should be clear that the funding available from H2020 is too limited to achieve impact across all of Europe, and should be considered only as 'seed money': it will be useful to kick-start new initiatives and to guarantee coordination between local initiatives across Europe – in other words, to foster the European dimension which is needed to reach critical mass. However, Digital Innovation Hubs need long-term and stable funding, which is not

licy on digital technologies

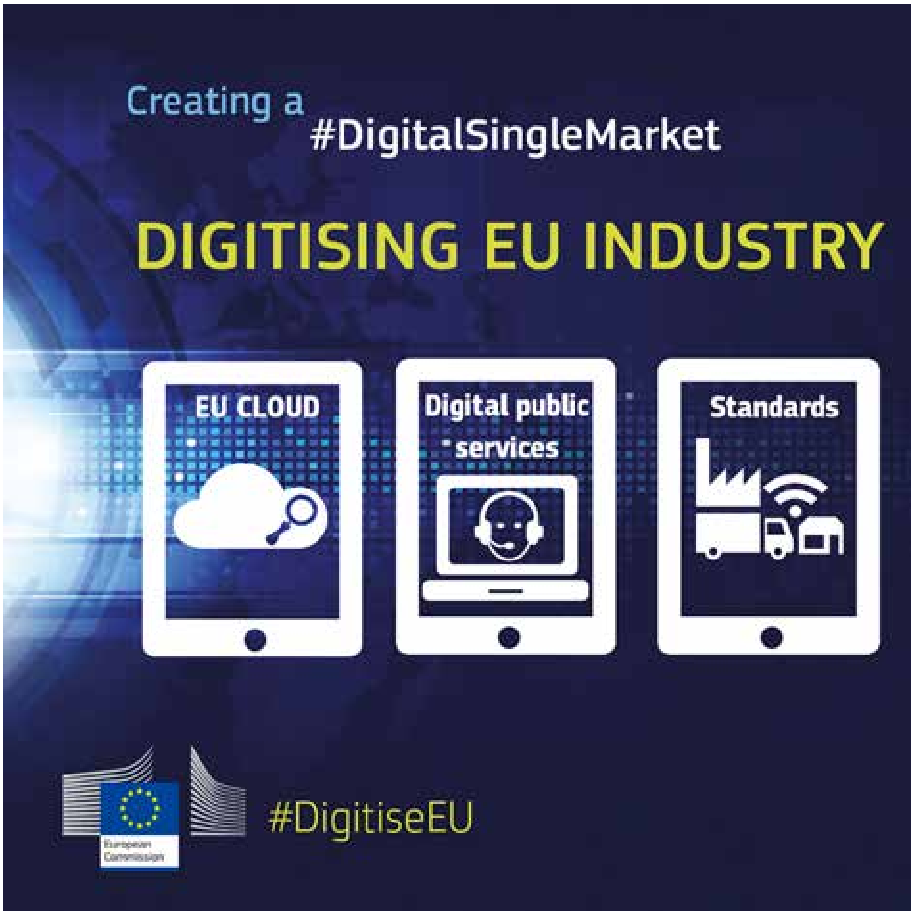
compatible with H2020 rules, so they will have to get their main financial support from other sources: local governments, national programmes, or European Regional Development Funds.

In this context, there is no 'one size fits all' solution: every innovation hub will have to find their best way to support its local industry. The European Commission will only have the role of supporting coordination and collaboration across Europe, namely through the Platform of National Initiatives which was launched at the end of March in Rome. In the same week another important event took place, which is also very relevant for the HiPEAC community: the launch of the European High-Performance Computing initiative, in which several Member States join forces to develop the next generation of 'exascale' computers, designed and built in Europe.

So, many things are shaping the digital policy of Europe in the coming months, and all these initiatives fall under the big umbrella of DSM, the 'Digital Single Market'. DSM has already delivered some spectacular results such as the reduction of data roaming costs across Europe. However, even more important are the ongoing activities in the areas of regulation for data ownership, free flow of data, liability and security, and autonomous systems. All these areas are prerequisites for our work in digital technologies: legal certainty is needed for investments in, e.g. big data or autonomous robots, and rules have to be coherent across Europe. If this does not happen, competitors in the US and China will outperform European industry thanks to the advantage they have in their home markets, which are true digital single markets.

This issue of the HiPEAC magazine has a special focus on healthcare, which is a very clear example of the need for a DSM in Europe: just think how data ownership and data privacy are important for the health profession. Who should own the data from your fitness sensors? Should the doctor that you see while on holiday be able to access your medical data from another hospital? Will you have the right to be informed in real time if your elderly grandmother becomes ill? Should you be free to bring your health insurance data with you when you move to another country? All these questions are very practical, but do not have a consistent solution across Europe. A DSM is needed to guarantee high quality of services and, of course, to also make the European healthcare sector efficient.

To summarize: what is happening now in the field of European policy will have a strong impact on the future development of digital technologies in all application areas. As a professional in the field, I advise you to stay tuned and to follow future developments closely, as they will be relevant not only for the overall market, but very likely also for your future career choices.



Welcome to Computing Systems Week Spring 2017 from Mario Kovač

HiPEAC: Mario, you were involved in the development of MP3 players and have a patent for JPEG compression. What new multimedia technologies are you excited about?

Photo: University of Zagreb

MK: There are several. For example, with IP video traffic reaching almost 90% of global consumer traffic by 2018 (as presented in the recent market analysis by Cisco), and given the plethora of devices on the market, the need to efficiently process and deliver video content will require enormous (exascale and beyond) HPC processing capabilities. Novel architectures and programming paradigms will need to be used to tackle this problem, but the results will enable companies in various market segments (including entertainment, health and security) to provide attractive and efficient new products and services. Our current research is strongly focused on this HPC/cloud architecture and application domain.

HiPEAC: You're also part of the EU's Expert Horizon2020 Leadership in Enabling & Industrial Technologies ICT Committee. What do you think Europe should be focusing on in terms of industrial ICT?

MK: An interesting new H2020 LEIT ICT work programme is currently in the definition process and hereby I encourage all of the HiPEAC community to participate in this process. We all know that ICT is both driver and enabler of industrial growth, so investments in technological development of ICT industry and integration of ICT in all segments of our industry is an important factor. Also, Europe has been dependent on non-EU processor technology for years. There are new EU initiatives that will try to change this, which I strongly support.

HiPEAC: What's the technology scene like in Zagreb? Also, where's the best place to grab a beer after a long day at CSW?

MK: Croatia is small country but the technology scene here is healthy and vibrant. The combination of good education and the possibility to provide ICT solutions/services globally makes this industry segment prosperous and competitive. As for a place to relax, with the centre of Zagreb being close to the CSW venue there are a number of places to have coffee, dinner and a few beers later. Some most popular spots in the centre are around Cvjetni trg (Flower Square) / Bogovićeva Street or Tkalčićeva Street.

Some useful Croatian for your time at CSW

Hi, I'm John and I'm great at computer science.

Bok, ja sam John i rasturam računarstvo.

I am lost. Please show me the way back to CSW.

Oprosti, izgubio sam se. Kako da se vratim na CSW?

Where's the nearest bar?

Gdje je najbliži kafić?

New impetus for Czech researchers in computing systems

Continuing the series of workshops in EU new member state countries, HiPEAC led a workshop at IT4Innovations in Ostrava, Czech Republic on 21 February. The aim of the workshops, which have been running since 2012, is to communicate to researchers in EU 'new member states' what HiPEAC is and what it does.

Representatives of five different technical universities as well as several companies came together in Ostrava for a very beneficial workshop hosted by IT4Innovations, the national supercomputing centre of the Czech Republic. Koen De Bosschere and Rainer Leupers presented the benefits of membership of the network for researchers from both academia and industry. Their presentations were followed by introductory talks by the attendees, which outlined the computing systems research ecosystem in the Czech Republic. Three blocks of presentations took place. The first was dedicated to speech and video processing. The second showcased Czech research related to low-power, high-performance computing. The final section was composed of talks on embedded systems and processors, networks and FPGAs. Prof. De Bosschere summarized his overall impression from the presented topics saying that: 'Had the presentations been anonymous, it would have been very difficult to tell whether they came from the Czech Republic, or from one of the "old member states". The research presented was of excellent quality. Several research outcomes were the result of European research projects, which shows that colleagues from the Czech Republic successfully compete for international research funding. HiPEAC membership can further expand their network, and get them involved in even more project proposals.'

The HiPEAC network hopes to welcome more members from the Czech Republic as result of the workshop.

37 nations represented at HiPEAC17

550 people from 37 countries came to a very sunny Stockholm 23-25 January for the annual HiPEAC conference. Over the years, it has developed into Europe's premier forum for experts in embedded and high performance systems architecture and compilation to network, forge new partnerships and find out about the latest developments in the field.

One of the reasons for the conference's popularity is the varied nature of the technical programme which is supported by exhibitions of university, project and industry-led research and innovation, and talks from companies. This year's company speakers came from both global giants like Intel and Ericsson and European SMEs including Silexica, Synective and INSYS. Keynote talks by Kathryn McKinley (Microsoft Research), Sarita Adve (University of Illinois at Urbana-Champaign) and Sandro Gaycken (ESMT Berlin) discussed data centre tail latency, memory coherence and consistency, and the immensity of the cybersecurity challenge.

The conference saw the launch of two startups: ZeroPoint Technologies (Gothenburg), which is in part a spinoff of the EC-funded EUROSERVER consortium, and Matryx Computers, the new business line of Embedded Computing Specialists. ZeroPoint is developing innovative compression technology with the potential to significantly compress the content of the cache and memory system while Matryx Computers specializes in FPGA-based embedded computers and operating systems for connected devices.

The Swedish capital, birthplace of Skype and Spotify and home to a vibrant tech startup scene, made an excellent host city, with the conference dinner taking place at the spectacular Stockholm City Hall. General Chairs Mats Brorsson and Zhonghai Lu of KTH Royal Institute of Technology in Stockholm noted the all-round positive ambience: 'We received a lot of positive feedback about the programme and the venue at the Waterfront Congress Centre. Three excellent keynote speeches led what has been a very interesting and diverse schedule of activities,' commented Mats Brorsson. 'It's been a very enjoyable experience to chair this edition of the HiPEAC conference and having been able to count upon the support of a very experienced conference committee! I'm now looking forward to attending HiPEAC 2018,' added Zhonghai Lu.

Werner Steinhögl of the EC's DG Communications Networks, Content & Technology, addressed a plenary session audience on the Digitising European Industry initiative, which aims to support and link up national initiatives for the digitization of industry and related services across all sectors and to boost investment through strategic partnerships and networks.

On the final day, Workshops and Tutorials co-Chair Diana Göhringer of Ruhr-University Bochum was awarded a HiPEAC Distinguished Service Award for her efforts in running this core element of the conference over the past three years.

The HiPEAC team would like to thank the conference sponsors, without whose generous support the event could not have been such a success.

See the keynotes speeches and other highlights at www.hipeac.net/youtube

Photo: Bagus Wibowo

Design for reliability in the era of the computing continuum

Concluded in Autumn 2016, the EU-funded CLERECO (Cross Layer Early Reliability Evaluation for the Computing cOntinuum) project proposed a scalable, cross-layer methodology and supporting suite of tools for accurate and fast estimations of computing systems' reliability.
As we enter the era of nanoscale devices, reliability is becoming a key challenge for the semiconductor industry. The now atomic dimensions of transistors result in a vulnerability to variations in the manufacturing process and can dramatically increase the effect of environmental stress on the correct circuit behaviour. Failures in early assessing computing systems' reliability may produce excessive redesign costs, which can have severe consequences for the success of a product.

Current practice involves a worst-case design approach with large guard bands. Unfortunately, application of this approach is reaching its limit in terms of economic sustainability with regard to performance, size and energy costs. Coordinated by Dr Stefano Di Carlo of the Polytechnic of Turin, the CLERECO project aimed to address this challenge by focusing on reliability analysis in the early phases of the design. Early assessment within the design cycle provides the freedom for adaptive modification if the estimated reliability level does not meet the requirements. CLERECO methodology provides dedicated tools to separately analyse the technology, the hardware components (at the microarchitecture level) and the software modules of a complex system and to recombine the characteristics of single objects into a complex statistical Bayesian model. This can be used to perform statistical reasoning on the reliability of the system as a whole.

See the full version of this story at bit.ly/2mLHwn6

HiPEAC members win prestigious CGO Test of Time award

A big round of applause to HiPEAC members John Cavazos, Grigori Fursin, Mike O'Boyle, Olivier Temam, and their co-authors Felix Agakov and Edwin Bonilla for winning the Test of Time award for their CGO'07 research paper on 'rapidly selecting good compiler optimizations using performance counters' (dl.acm.org/citation.cfm?id=1252540). This annual award recognizes outstanding papers published at the International Symposium on Code Generation and Optimization (CGO) one decade earlier, whose influence is still strong today.

This paper set an early example of the benefits of applying machine learning to compiler optimization. Importantly, it also led to realizing the challenges of transferring this research into production: the need to perform and process a huge number of rigorously controlled experiments to train predictive models, all in the presence of the continuously evolving software and hardware stack.

These challenges motivated Dr. Grigori Fursin to continue this research as a community effort. He created an open-source framework to share research artifacts (workloads, data sets, tools, models, features, scripts) as reusable components with JSON API, crowdsource experimentation across diverse hardware and inputs provided by volunteers, continuously learn most effective optimizations, collaboratively discover important SW/HW features to improve predictive models via a public repository of knowledge at cKnowledge.org.

Ten years on, this collaborative approach to performance optimization is used and extended by dividiti, ARM, General Motors, IBM, Imperial College, University of Edinburgh, University of Cambridge and other leading universities and companies to develop faster, cheaper, more power-efficient, and more reliable computer systems. It also helped initiate the Artifact Evaluation initiative at the CGO, PPoPP, PACT and other premier conferences to encourage artifact sharing and reuse, as well as independent validation of experimental results: cTuning.org/ae .

Dr Fursin commented: 'We would like to thank the community for strong interest in our machine learning and community based optimization techniques over the past ten years. We also encourage you to join our community effort to accelerate computer systems research and thus enable efficient, reliable and cheap computing everywhere - from IoT devices to supercomputers!'

Award for TUDelft Team in International Big Data Apache Spark Competition: Ultra Fast and Low Cost Personalized DNA Analysis Using Big Data Approach

A team from the Computer Engineering Lab at Delft University of Technology team won the $25,000 2nd prize in the Big Data Apache Spark hackathon competition held in New York City.

This is an international competition in which contestants compete to create an innovative big data solution that addresses relevant societal challenges using publicly available datasets and big data techniques. The competition generated much interest, attracting more than 500 registered contestants, with 23 teams making it to the finals. The TUDelft team created a platform called DoctorSpark to enable high performance and low-cost computation of DNA analysis programs using the Apache Spark big data framework. This platform enables faster DNA diagnostics in hospitals and clinics for patients suffering from cancer or other genetic disease. The results were announced during the Data First Event in New York on 27 September 2016. More information about the winning project can be found at http://devpost.com/software/scalable-dna-analysis-pipelines-using-sparkz

Cristina Silvano named 2017 IEEE Fellow

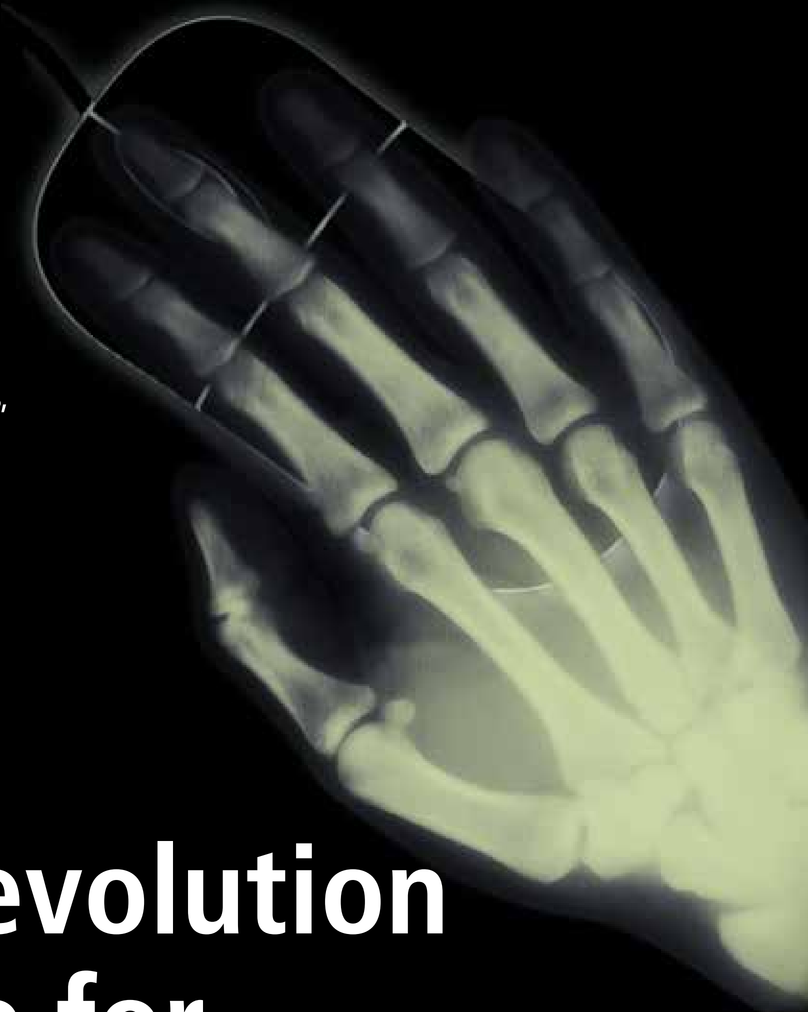
Professor Cristina Silvano of the Politecnico di Milano has been named an IEEE Fellow 'for contributions to energy-efficient computer architectures'. The IEEE grade of Fellow is conferred by the IEEE Board of Directors upon a person with an outstanding record of accomplishments in any of the IEEE fields of interest. The total number selected in any one year cannot exceed one-tenth of one percent of the total voting membership. IEEE Fellow is the highest grade of membership and is recognized by the technical community as a prestigious honour and an important career achievement.

At the early stages of her career, Cristina was part of the Bull-IBM Research team for the design of a family of scalable multiprocessor systems based on the PowerPC architecture, introduced in 1992 by Apple-IBM-Motorola. She then started investigating power optimization and estimation techniques for embedded architectures applied to the Lx/ST200 VLIW processors, designed in partnership between HP Labs and STMicroelectronics and widely used in a variety of embedded media processing products.

Her research interests are in the design of energy-efficient computer architectures with special emphasis on design space exploration and application autotuning for embedded manycore architectures. In these areas, she has coordinated several funded projects, including two EU-funded projects (MULTICUBE and 2PARMA). She is also active in the area of autotuning and adaptivity for energy-efficient HPC systems. On this topic, she is currently the Scientific Coordinator of the H2020 FET-HPC ANTAREX research project.

Prof. Silvano is an active member of the scientific community and served as General Chair and Program Chair of several conferences and workshops on computer architectures and design automation. She is Associate Editor of the ACM Transactions on Architecture and Code Optimization and served as independent expert reviewer for the European Commission and for several science foundations.

She has over 160 publications in peer-reviewed international journals and conferences, four books and has made several industrial patent applications.



Europe's national healthcare systems face huge challenges, including an aging population and the inevitable burden of chronic diseases and conditions, and limitations on economic resources. These have placed new demands on healthcare systems and so, to remain sustainable and meet populations' needs, a shift is required in the way that services are managed, delivered and funded.

Bringing the computing revolution to healthcare for a changing population

In terms of information and communication technology, a big data approach is needed to help address problems faced by traditional healthcare applications. As this article shows, these have access to a limited set of data, which is usually fragmented and stored in different and hard-to-access sites. As such, the introduction of increased automation into the healthcare sector has never been more appropriate.

Digital healthcare systems can offer a number of benefits, such as improved connectivity, information integration and data capture, increased of analytic and diagnostic speed and accuracy and long-term cost savings. They can also facilitate patient empowerment, enabling them to play a more active role in the management of their own health, and receive personalized medicines and health plans.

Reliance on such techniques is increasing, which means that the potential for growth in the digital health sector is huge.
However, the shift towards digital healthcare brings its own unique challenges, as reliability and security of the information captured by digital systems and devices is paramount. This has meant that development and evolution within the medical IoT sector has been slower than other digitized fields, due to the high levels of regulation and validation needed to bring products to market. Add to this the significant technical tasks of dealing with massive amounts of data, or the rigorous performance required by medical applications within minimal power or space constraints, and it is easy to see the complexity of bringing new health technologies to market.

In this special feature, we explore a few examples of how the HiPEAC community is at the heart of this revolution, developing cutting-edge biomedical technologies and enhancing the capacities and capabilities of existing ones. Whether it is helping to model the human brain, building a European ecosystem for large scale clinical data management, harnessing the power of high-performance systems for medical imaging, or adapting financial applications to intensive care, HiPEACers are laying the foundations for the healthcare of tomorrow, by trying to meet the demands of today.

You can read more on this topic in the 'Career talk' on page 35 with Darko Gvozdanović of Ericsson Nikola Tesla, which is leading the way in European eHealth systems.

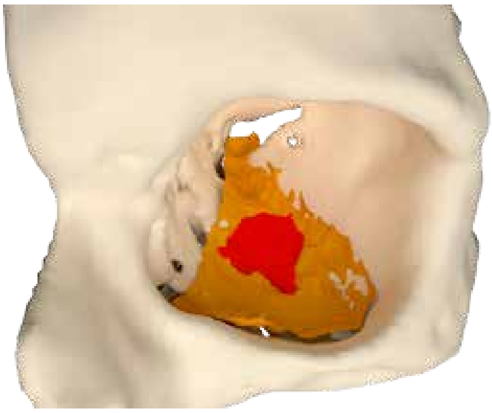
HIGH PERFORMANCE COMPUTING IN MEDICAL IMAGING

Researchers at IT4Innovations in the Czech Republic are constantly searching for new research directions and areas where high-performance computing (HPC) technology can be put to good use. One of our most important collaborations is with medical doctors from the University Hospital in Ostrava, Czech Republic, working on methodology for more precise measurement of orbital (eye socket) fracture size.

Very specific and precise information is required to assess the seriousness of orbital floor fractures – fractures at the base of the eye socket. Such assessments in turn determine whether patients should undergo surgery or whether less invasive treatment should be given instead. Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) scanners are currently used by doctors to create three-dimensional virtual models from two-dimensional CT and MRI images. The extent of an orbital floor fracture is determined directly from CT images using a simplified empirical approach.

> ***"We have developed parallel versions of all the tools for image processing, dramatically reducing analysis time and ensuring that patients receive the correct treatment sooner"***

Although the use of CT and MRI technology raises standards in diagnostic medicine, the process generates large amounts of data. It is not only very time-consuming and labour-intensive to analyse this data, but also inefficient because not all the required information can be extracted from such virtual models. Utilizing resources available at IT4Innovations, we have developed parallel versions of all the tools for image processing outlined below, dramatically reducing analysis time and therefore ensuring that patients receive the correct treatment sooner.



3D virtual model of eye socket (white), orbital floor (orange) and fracture (red)

To construct three-dimensional models from two-dimensional data sources, we start by using filters such as Gaussian smoothing, anisotropic diffusion or BM3D to reduce noise in the CT images. Secondly, k-means clustering is used for image segmentation. In this step, the image is simplified to allow us to localize objects and their boundaries. Finally, we use the Poisson method for surface reconstruction. After analysis of the 3D models, doctors carry out validation exercises, which helps us to improve existing algorithms, thus enhancing the accuracy of measurements of orbital floor fractures.

Overall, we expect this collaboration to lead to virtual models of the orbital floor with minimal user intervention, which would allow doctors to more precisely establish the size of orbital floor fractures and therefore make better decisions about the treatment of patients.

www.it4i.cz
Karina Pešatová, IT4Innovations National Supercomputing Center

IMPROVING RESPIRATORY VENTILATION WITH ADVANCED ICT ANALYTICS

We expect an intensive care unit (ICU) to be the safest possible place, yet patients routinely receive mechanically assisted ventilation, which leads to the possibility of ventilator induced lung injury. Inflation of the alveoli generates stress forces which in turn create strain on the cells, which may lead to damage. The stress forces created by the inflation process are proportional to the tidal volume, a parameter that is defined on the mechanical ventilator, but needs to be optimized for gender and ideal body weight.

Queen's University Belfast, co-ordinator of the FP7 NanoStreams project, developed a system to monitor tidal volume and other airway pressure parameters associated with the respiratory physiology of patients. The system is known as VILIAlert and it is deployed in an ICU where patients are monitored continuously. The system is programmed to monitor various threshold violations (e.g. pressure in the patient's airways becoming too high or too low) and to report such events to the attending physicians in real time via SMS and other electronic media.

This builds upon previous NanoStreams work that calculated prices of financial options from a real-time streaming feed of stock prices. Here, the kernels were driven by for loops and alternatively by navigation of a binomial tree, yet monitoring of physiological parameters involves more logic and many more parameters. In addition, a fundamental component of our ventilator monitoring systems is a database,

▶

whereas, in the market data application, data relating to prices is processed straight off the wire.

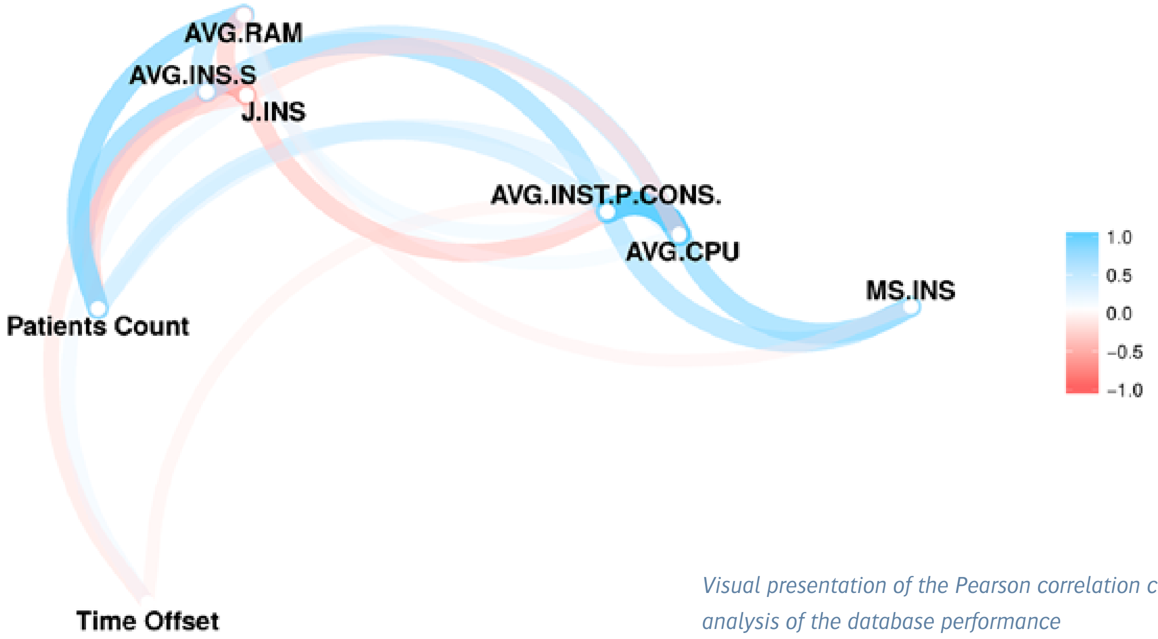
For the financial use case, we defined new metrics of 'seconds per option' and 'joules per option' leading to a quality of service metric. One has no control over the arrival time of the next price update although, on a typical trading day, arrival intervals can be modelled using Poisson distribution. In contrast, human physiology is a continuous process measured by sensors that can be set to take recordings at predefined time intervals before forwarding them to a central database. Data is even routinely filtered at source so that only every third or fewer reading might be transmitted to the database. This is as much a function of the network infrastructure as of the scalability of the compute infrastructure.

> *"Monitoring of physiological parameters involves more logic and many more parameters."*

We extended the analysis in NanoStreams to derive metrics for the database component in our VILIAlert system and applied this to four open source databases (MySQL, PostgreSQL, ScaleDB and MariaDB), all of which have similar interfaces. In order to provide rigorous coverage of test cases, but within a reasonable amount of time, we used the statistical method of non-parameter bootstrapping. This reduced our run-time from 16 days to 36 hours. The image below presents the Pearson correlation coefficients for our analysis. Each metric is a node in the graph and the proximity of the metrics to each other represents the overall magnitude of their correlations. Thus clustering of the metrics is easily seen. Each path represents the correlation between the two variables. Blue and red paths represent positive and negative correlations respectively, while the transparency and the width of the path represent the strength. Thinner and more transparent paths mean weaker correlation. We can see that Patients Count, average RAM used (AVG.RAM), average inserts per second (AVG.INS.S) and joules per insert (J.INS) form one cluster while AVG.INST.P.CONS (average instantaneous power), AVG.CPU (average CPU) and MS.INS (milliseconds per insert) form another distinct cluster. This means that increasing the number of patients has more impact on RAM usage than on CPU usage. This also means that databases that rely more on CPU than on RAM to handle an increased number of patients tend to have higher instantaneous power consumption than the databases that rely more on RAM. Apart from that, increased CPU usage implies an increment in the MS.INS metric. The best examples for this are ScaleDB and PostgreSQL, both of which had similar performance regarding the AVG.INS.S metric. ScaleDB handles an increased number of patients by using more CPU power and thus having the highest INST.P.CONS metric, while on the other hand PostgreSQL relies more on RAM and therefore has the lowest INST.P.CONS metric. Similarly, ScaleDB had the highest MS.INS metric, while the PostgreSQL had the lowest.

NanoStreams' overall mission is to explore domain-specific software stacks for real-time data analytics. In our work on physiological monitoring, where data ingress and storage is the dominant workload in comparison to SQL queries, we have identified distinct energy and performance characteristics for different databases. We have found that ScaleDB is an optimum database technology when handling between 200 and 800 patients in this application, while PostgreSQL performs best outside of this range.

Charles J Gillan, Murali Shyamsundar, Aleksandar Novakovic and Dimitrios S Nikolopoulos, Queen's University Belfast



Visual presentation of the Pearson correlation coefficients from analysis of the database performance

WIDE-RANGING INNOVATIONS AT TU DELFT

Medicine and healthcare form one of the most notable achievements of all human endeavour, and resonate closely when our lives or those of our loved ones are affected by bad health. Traditionally, medicine has been a relatively conservative field in the way technology is used to support the activities of doctors or to facilitate new methods for diagnosis and treatment. However, as new technologies continue to prove their effectiveness and viability in clinical environments, more and more attention is being given to incorporating these technologies into common medical practices.

Our Computer Engineering Lab at the Delft University of Technology (NL) has taken notice of this trend, and has worked to establish a network of Dutch and European collaborators to investigate the potential impact of bringing the computer revolution to the medical world. The effort in our lab has two focal points: 1. investigating and enabling new technologies, and 2. facilitating and improving existing technologies.

Enabling new technologies

A good example is genetic research, which promises to become a game changer for the practice of medicine, by enabling personalized diagnostics and therapies to be developed for specific patient needs. Long and expensive compute times hinder the actual deployment of these techniques in patient care. Our lab has been collaborating with a number of institutes such as the German Cancer Research Center (DE) and Utrecht University Medical Center (NL) to accelerate their compute intensive algorithms, this enabling them to be used for patient diagnostics. Our lab has a high-tech startup called Bluebee that focuses on commercializing the genomics-related technologies that we develop.

Another standout example is research into the human brain, the so-called final frontier of science. This new and rather challenging field of research is expected to lead to a deep understanding of the root causes of mental illness and to help develop new effective therapies. The first step towards enabling this research involves simulating brain activity from the bottom up, by building brain models one cell at a time. Needless to say, such an activity is remarkably computationally intensive. Our lab is collaborating with partners such as the Erasmus Medical Center (NL) to accelerate and scale up these computations on high performance platforms, allowing the creation of bigger models that shed more insight into the functionality of the brain.

Improving existing technologies

Our lab is also working closely with a couple of organizations to improve the capabilities of existing medical procedures. One example is our collaboration with Leiden University Medical Center (NL) and Philips (NL) to manage the large size of medical imaging databases and to speed up image processing algorithms. This allows new modes of medical examination, where automated algorithms can support doctors to identify features in images or to combine and compare images for better or faster diagnosis. This also allows for new forms of intervention, such as minimally intrusive surgery, in which surgeons use imaging equipment and real-time processing to eliminate the need for direct visual inspection during surgery.

With our research, we aim to enable medical professionals to provide patients with better and more effective medical care, and give them a helping hand to integrate new technologies into this most valuable of human professions.

TU Delft

Zaid Al-Ars, TU Delft

AEGLE: HARNESSING BIG DATA TO FIND TOMORROW'S CURES

Currently, healthcare applications only have access to a limited set of data, as data are usually fragmented, stored in different sites and with no easy access from external locations. In order to unlock the value of these data, a big data approach is needed. Analytics will help us understand the nature of various scientific questions and will allow us to integrate different data sources to help answer them. In addition, the adoption of a big data approach will enable the discovery of new correlations that are currently not foreseen, due to the fragmentation of datasets.

Focusing on healthcare, this approach could have an impact on the fields of medical imaging, oncology, intensive care units and healthcare policy making, as well as on the movement towards personalized management of chronic disease. This impact is twofold: on the one hand, it will enable healthcare stakeholders to develop cost-effective interventions, simultaneously improving patients' quality of life; while on the other, it will boost the activities of businesses developing big data health solutions.

Big data analytics are, in fact, becoming increasingly common in human-centred sciences, and ever-increasing data volumes have led to the development of new parallel processing models. However, data volumes are increasing at a faster pace than the available processing power, making it increasingly difficult to keep up with processing requirements.

The AEGLE solution: Big data for healthcare

An EU-funded Horizon2020 initiative implemented by 13 partners across Europe, AEGLE provides a framework for the management of big bioclinical data. The project addresses a number of challenges which can be divided into four main categories: user, technical, business and ethical, which reveal both the complexity of the project and the potential for impact on healthcare. AEGLE tackles performance and scalability challenges by building on heterogeneous acceleration, cloud and big data computing technologies to deliver optimized analytics services. Issues regarding the acceptance of the platform, problems regarding data integration, the nature of the AEGLE use cases, the sustainability of its business model and the management of legal and regulatory issues have already been identified, and their solutions are being incorporated into the system design.

Rather than just providing another multipurpose big data analytics platform, AEGLE incorporates health into the core of its activities. In addition, to help overcome resistance to change, AEGLE is working on a regulatory framework needed for the adoption of new solutions, and has involved healthcare stakeholders in its activities from day one. Finally, AEGLE will provide a practical demonstration of the impact of big data on healthcare, by delivering three prototypes and by organizing awareness-raising activities to attract users and buyers. These activities are accompanied by a business model to enable the exploitation of results after the project ends.

Three use cases have been selected, covering a wide spectrum of healthcare:

- Type-2 diabetes, representing non-malignant chronic diseases. The AEGLE platform allows the interdependency of risk factors to be analysed so as to predict potential deterioration.
- Chronic lymphocytic leukaemia, an example of a malignant chronic disease. The AEGLE framework associates phenotypic data with personal genetic profiles and offers the possibility of identifying and evaluating treatment plans, with a view towards personalized medicine.
- Intensive care units, a typical paradigm of acute care. AEGLE aims to improve the management of clinical and laboratory data as well as physiologic waveforms. Its scalable data analytics will provide automated analysis of variables for the detection of unusual, unstable or deteriorating states in patients.

This approach will help AEGLE to include other cases within these categories, meaning the platform can be easily scaled up.
Overall, AEGLE aims to be the point of reference in big data applications for health that will create a multi-million euro business impact, enable thousands of researchers to exploit analytics and lead to increased acceptance of big data solutions in healthcare.

www.aegle-uhealth.eu

Andreas Raptopoulos, EXUS Innovation and Candela Bravo, LOBA

A TULIPP IN THE FIELD OF MEDICAL X-RAY IMAGING

Medical imaging is the visualization of body parts, organs, tissues or cells for clinical diagnosis and preoperative imaging. The global medical image processing market is about $15 billion a year. The imaging techniques used in medical devices include a variety of modern equipment in the fields of optical imaging, nuclear imaging, radiology and other image-guided intervention. The radiological method, or x-ray imaging, renders anatomical and physiological images of the human body at a very high spatial and temporal resolution.

Dedicated to x-ray instruments, the work of the Tulipp project is highly relevant to a significant part of the market share, in particular through its Mobile C-Arm use case, which is a perfect example of a medical system that improves surgical efficiency. In real time, during an operation, this device displays a view of the inside of a patient's body, allowing the surgeon to make small incisions rather than larger cuts and to target the region with greater accuracy. This leads to faster recovery times and lower risks of hospital-acquired infection. The drawback of this is the radiation dose: 30 times what we receive from our natural surroundings each day. This radiation is received not only by the patient but also by the medical staff, week in, week out.

While the x-ray sensor is very sensitive, lowering the emission dose increases the level of noise on the pictures, making it unreadable. This can be corrected with proper processing.

From a regulatory point of view, the radiation that the patient is exposed to must have a specific purpose. Thus, each photon that passes through the patient and is received by the sensor must be delivered to the practitioner; no frame should ever be lost. This brings about the need to manage side by side strong real-time constraints and high-performance computing.

We managed to lower the radiation dose by 75% and restore the original quality of the picture thanks to specific noise reduction algorithms running on high-end PCs. However, this is unfortunately not convenient when size and mobility matter, like in a confined environment such as an operating theatre, crowded with staff and equipment.
Yet by providing the computing power of a PC in a device the size of a smartphone, Tulipp makes it possible to lower the radiation dose while maintaining the picture quality. To achieve this, a holistic view of the system is required so as to achieve the best power efficiency from inevitably highly heterogeneous hardware.

With our power-aware tool chain, the application designer can see, for each mapping of the application tasks on the hardware resources, the impact on power consumption. He or she can thus schedule the processing chain to optimize both the performance and the required energy. The tool chain relies on a low-power real-time operating system. Specifically designed to fit in the small memory sizes of embedded devices, it comes with an optimized implementation of a necessary set of common image processing libraries and allows seamless scheduling of the application on the hardware chips.

Philippe Millet, Thales

A Tulipp Hardware Instance

This issue's round-up of news and results from EU-funded projects includes the final outcomes of major projects ASPIRE, EUROSERVER, ASAP and HARPA, as well as giving an update on work on aircraft design in the MIKELANGELO consortium.

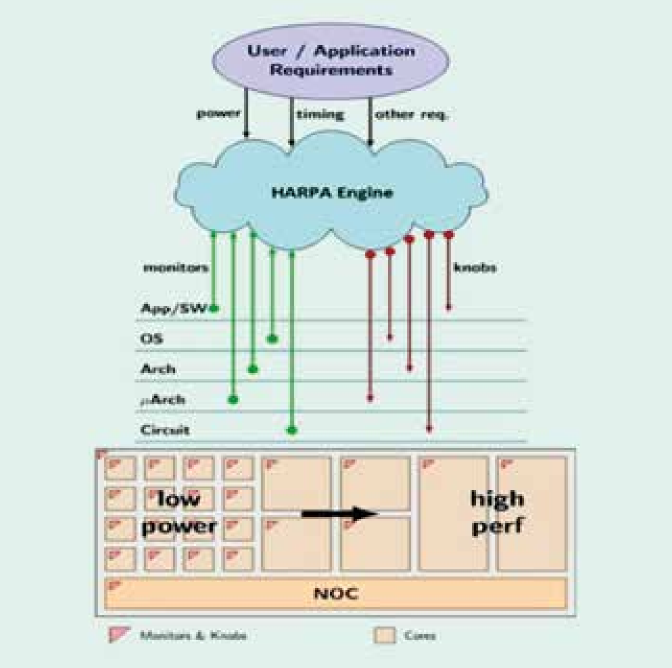
Innovation Europe

COST-EFFICIENT WAYS TO MANAGE PERFORMANCE VARIABILITY

Continuously increasing application demands on both high-performance computing (HPC) and embedded systems (ES) are driving the information and communications manufacturing industry to a never-ending scaling of silicon devices. Nevertheless, integration and miniaturization of transistors comes with an important and non-negligible trade-off: time-zero and time-dependent performance variability. The HARPA project, which ended in late 2016, aimed to enable next-generation embedded and high-performance heterogeneous many-cores to cost-effectively confront variations by providing 'dependable performance': correct functionality and timing guarantees throughout the expected lifetime of a platform within thermal, power and energy constraints. The HARPA novelty is in seeking synergies in techniques that have been considered virtually exclusively in the ES or HPC domains (worst-case guaranteed partly proactive techniques in embedded, and dynamic best-effort reactive techniques in high-performance).

The industry and academic partners of the pan-European HARPA team specialized in fields covering all abstraction layers, from hardware to application level. The project developed a set of monitors/knobs in hardware and software designs that observes performance unpredictability, triggering system reactions. The figure below provides an overview of the HARPA engine.

It is a middleware split between the Operating System (HARPA-OS) and the hardware actuators (HARPA-RTE) and provides run-time dependable performance guarantees. HARPA-OS applies resource allocation policies, arbitrating the OS calls with a by-second time granularity. HARPA-RT sits at a low level in the system stack, achieving a millisecond control on hardware resources. HARPA-OS and HARPA-RTE cooperate to ensure the performance dependability goals, keeping a prompt low-level control on hardware resources. Run-time reactive and proactive techniques have been deployed, ensuring that the combined monitor/scheduling/knob reaction latency never violates the application deadlines. These techniques were tested on industrial applications running on embedded platforms and a full-system evaluation framework simulating HPC setups.



A fundamental objective of the project was to provide solutions to mitigate reliability threats and ensure dependable system performance. To this end, the HARPA engine was developed, implementing various control frameworks across the system stack. The goal was to exploit different manifestations of platform slack (i.e. slack in performance, power, energy, temperature, lifetime or structures/components), in order to ascertain timing guarantees throughout the lifetime of the device. A component of the HARPA engine is the HARPA-OS, the system-wide resource manager developed by POLIMI. This component must include control policies capable of providing a response in a timeframe spanning from hundreds of milliseconds to a second. The HARPA-RTE sits at a low level in the system stack and is in direct contact with the various monitors and knobs. It has responsive control on hardware resources, enabling extremely fast adaptation to system behaviour in the scale of some milliseconds, which is ideal for providing guarantees for hard-deadline applications and complements the comparatively slower responsiveness of the HARPA-OS.

The concepts developed within the HARPA context addressed both the HPC and ES domains equally. Specifically, from the HPC domain we used disaster and flood management simulation, while, from the ES domain, a radio frequency spectrum sensing application, a face detection application, object recognition and the Beesper Landslide Multimodal Monitoring. In particular, HARPA use cases demonstrated in HPC platforms: (i) Intel Xeon, (ii) x86-64 multi-core plus a GPU and embedded platforms: (a) Freescale i.MX 6Quad, (b) ODROID XU-3 (Octa Core Linux Computer Samsung Exynos5422 Cortex-A15 2.0Ghz quad core and Cortex-A7 quad core).

NAME: Harnessing Performance Variability (HARPA)
START/END DATE: 01/09/2013 – 30/11/2016
KEYWORDS: many-core, high-performance architectures, thermal related reliability, dependability, adaptive systems, energy efficiency, performance and timing analysis, run-time resource management
PARTNERS: Politecnico di Milano (Italy), Interuniversitair Micro-Electronicacentrum Imec (Belgium), University of Cyprus (Cyprus), Vysoka Skola Banska - Technicka Univerzita Ostrava (Czech Republic), Thales Communications & Security (France), Institute of Communication and Computer Systems (Greece), HENESIS (Italy)
BUDGET: €3.9M
WEBSITE: www.harpa-project.eu

The HARPA project received funding from the European Union's FP7 Programme under grant agreement no. 612069.

THE MIKELANGELO APPROACH TO HPC SIMULATIONS AND AIRCRAFT DESIGN

When high performance of a computer infrastructure is needed, we usually choose to use HPC. However, when flexibility and adaptability are required, we tend to opt for the HPC cloud. In such cases, we amalgamate the best of both worlds: the performance of HPC and the flexibility of the cloud. However, the combination of these approaches presents us with challenges. When performance, flexibility and security of the virtualized infrastructure are required, software adaptations are necessary alongside the use of HPC. Enter MIKELANGELO, a Horizon 2020-funded HPC cloud research project. MIKELANGELO is boosting performance of VMs (Virtual Machines utilizing the hardware structure of a physical host) and I/O (input/output) operations by deploying their innovative technologies: I/O boosting updates to KVM (Kernel-based Virtual Machine), OSv unikernel for fast and secure workloads, OpenStack and Torque compatibility and deployment-ready OpenFOAM HPC cloud components.

Able to boot in less than a second, OSv (an open source operating system designed for the cloud) can execute applications on top of any hypervisor, resulting in superior performance, speed and effortless management. Many applications, including HPC and the big data business cases steering the MIKELANGELO project, directly benefit from those features.

Efficiency and speed of input/output operations is especially important in the light aircraft design process, running heavily parallelized numerical simulations to improve aerodynamic properties at an early stage. The Slovenian aircraft manufacturer Pipistrel uses computational fluid dynamics (CFD) simulations on a computer to simulate the flow of air around an aircraft and analyse aerodynamic features of their designs without time-consuming and expensive manufacturing.

OpenFOAM, the most widely used general-purpose open source software package for CFD is ideal when it comes to designing new aeroplanes or even just improving parts of existing aeroplanes. Pipistrel currently runs many consecutive cases either on a local machine or on a remote cluster. In either case, the target machines need to be specifically configured to run OpenFOAM requests.

The OpenFOAM cloud, developed within MIKELANGELO, along with highly optimized I/O components built directly into KVM can be deployed on top of any hardware (cluster, HPC hardware, cloud hardware). Its functionalities, flexibility, modality and ease of deployment are exposed through a lightweight OpenStack dashboard allowing users to focus on the simulation design rather than on cluster deployment, management and support.

The HPC cloud approach developed through MIKELANGELO brings together the best of both worlds: the raw performance of HPC infrastructure and the flexibility of clouds. The MIKELANGELO team are working tirelessly to maximize achievements in both of these areas of strength, using unikernels and optimized virtualization infrastructure (IO efficient KVM) to reduce the virtualization impact on one hand, and optimizing the actual software packages (e.g. openFOAM) to perform on such infrastructure on the other.

MIKELANGELO meeting – Pipistrel's headquarters, Ajdovščina, Slovenia

NAME: MIKELANGELO - MIcro KErneL virtualizAtioN for hiGh pErfOrmance cLOud and hpc systems
START/END DATE: 01/01/2015 – 31/12/2017
KEYWORDS: HPC, cloud, simulations, aircraft design, unikernels, OpenFOAM
PARTNERS: XLAB (Slovenia – coordinator), Huawei Technologies Düsseldorf (Germany), IBM Israel, Intel Research & Development Ireland, ScyllaDB (Israel), Universitaet Stuttgart (Germany), GWDG (Germany), Ben-Gurion University of the Negev (Israel), Pipistrel (Slovenia)
BUDGET: €5.99M
WEBSITE: www.mikelangelo-project.eu
MIKELANGELO is funded by the European Commission's Horizon 2020 Framework Programme under grant agreement no. 645402.

FLEXIBLE & SCALABLE DATA ANALYTICS

Recently concluded, the ASAP FP7 project has developed a dynamic open-source execution framework for scalable data analytics. The driving idea was that no single execution model is suitable for all types of tasks, and no single data model (and store) is suitable for all types of data. Complex analytical tasks over multi-engine environments therefore require integrated profiling, modelling, planning and scheduling functions.

The ASAP project pursued four main goals:

1. A modelling framework that constantly evaluates the cost, quality and performance of available computational resources in order to decide on the most advantageous store, indexing and execution pattern.

2. A generic programming model in conjunction with a runtime system for execution in the cloud. The execution can target clusters using an extended and augmented version of Spark, or multiprocessors using the high-performance Swan task-parallel execution engine. State-of-the-art features include: irregular general-purpose computations, resource elasticity, synchronization, data transfer, locality and scheduling abstraction, ability to handle large sets of irregularly distributed data, and fault tolerance. To overcome Spark's limitations on irregular loads, the project has augmented the Spark runtime with full support for general-purpose, recursive computations.

3. A unique adaptation methodology that enables analytics experts to amend submitted tasks in later processing stages. In combination with visualization and monitoring of workflows, this enables data scientists and analytics engineers to fine-tune workflows and speed up development time as well as understand and adjust performance in production.

4. A real-time visualization engine to show the results of the initiated tasks and queries in an intuitive manner -- building on the dashboard of the Media Watch on Climate Change and the faceted search developed for the Climate Resilience Toolkit.

The ASAP consortium brought together partner expertise in data analytics, runtime systems, scheduling and cost estimation, programming models, optimization, data science and visualization. Towards the latter stages of the project, the consortium focused on integrating all of the ASAP modules into a single open-source framework. The ASAP platform is open and available for download and use, and incorporates research results that have advanced the state of the art in multiple fields and resulted in tens of publications.

The platform has already been deployed in production, on two industrial applications within the project, to manage complex workflows on web content analytics and telecommunication data analytics:

- The **Web Content Analytics** use case is centred on the services of Internet Memory Research. These services provide access to a very large collection of content extracted from the web, cleaned, annotated and indexed in a distributed infrastructure. Previously, this was mainly based on Hadoop components. ASAP extended the workflow interface used by IMR to make workflow editing easier and automatically produce optimal workflow materializations, by learning the performance of each component and automatically selecting optimal workflow components from all available implementations.
- The **Telecommunication Data Analytics** use case mines call data record data by WIND Telecomunicazioni, for user classification, prediction of network load and detection of unusual events from mobile phone calling patterns. The telecommunication data is combined with data mined from social media and visualized to help analysts gain better insights, detect special events that influence network traffic, and make overall better predictions and decisions. Technology developed within ASAP helped WIND engineers develop these applications, manage their execution, and scaled their analysis to many millions of mobile phone calls in a greatly reduced amount of time.

NAME: A Scalable Analytics Platform (ASAP)

START/END DATE: 01/03/2014 – 28/02/2017

KEYWORDS: big data analytics, heterogeneous platforms, workflow design, workflow scheduling

PARTNERS: Foundation for Research and Technology – Hellas (Greece), Université de Genève (Switzerland), Institute of Communication and Computer Systems (Greece), Queen's University Belfast (UK), Internet Memory Research (France), WIND Telecomunicazioni (Italy), webLyzard technology (Austria)

BUDGET: €3.6M

WEBSITE: www.asap-fp7.eu

The ASAP project received funding from the European Union's FP7 Programme under grant agreement no. 619706.

MAKING MOBILE DEVICES MORE SECURE WITH THE ASPIRE FRAMEWORK

In January 2017, the ASPIRE project was evaluated as 'excellent' at its final project review with the European Commission. The mission of ASPIRE was to integrate state-of-the-art software protections into an application reference architecture and into an easy-to-use compiler framework that automatically provides measurable software-based protection of the valuable assets in the persistently or occasionally connected client applications of mobile service, software and content providers.

For mobile devices like smartphones and tablets, security solutions based on custom hardware (as is traditionally done with smart cards, set-top boxes and dongles, for example) are not convenient. Software protection is therefore of utmost importance; it can be a maker or a breaker of a product or service, or even a business. Current software protection techniques are incredibly hard to deploy, cost too much and limit innovation. Stakeholders in mobile devices need more trustworthy, cheaper software security solutions and more value for the money they spend on software security. In this project, three market leaders in security ICT solutions and four academic institutions joined forces to protect the assets of one class of stakeholders: the service, software, and content providers. From their perspective, mobile devices and their users, which can engage in attacks on the software and credentials installed to access the services or content, are not trustworthy.

Final results and their potential impact and use

The software protection technology that has been developed consists of:

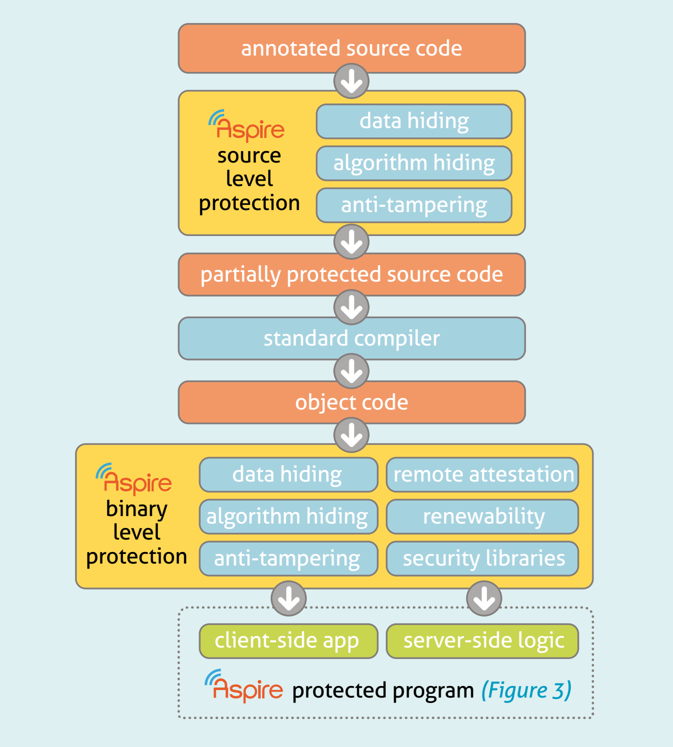
(i) the **ASPIRE reference architecture** for combining and composing multiple layers and types of software protections;

(ii) designs and implementations of a range of online and offline protections, some of which pre-existed, some which are new or significant improvements over the previous state of the art;

(iii) the robust **ASPIRE Compiler Tool Chain** that enables the automated, combined deployment of combinations of protections on real-world use cases;

(iv) the **ASPIRE Decision Support System** and its **ASPIRE Knowledge Base** to assist the user of the tool chain with the selection of the protections best suited to protect the software and the assets embedded in it; and

(v) the **ASPIRE software protection evaluation methodology** to assess the value of software protections vis-à-vis man at the end attacks.



The ASPIRE Compiler Tool Chain is based on plug-ins. Its overall flow is shown in the figure above. First, a sequence of source-to-source rewriting plug-ins are invoked. Each of them takes as input (pre-processed) C code and produces the same format. This facilitates the insertion of additional plug-ins. All the plug-in transformations are controlled by pragmas and attributes with which the assets to be protected have been annotated. Concrete annotations are available to specify concrete protections. Abstract requirement protections are supported as well, with which the developer can specify the security requirements on the assets (integrity, confidentiality, and so on). The ASPIRE Decision Support system then converts those requirements into specifications of protections to be deployed. The final source-level plug-in extracts the remaining annotations from the source code, which is then compiled with GCC or LLVM into standard object code, and linked with binutils (binary utilities). Plug-ins in the link-time binary code rewriting framework Diablo then apply further transformations to deploy additional protections and to finalize some of the protections of which the first analysis and transformation steps were initiated on the source code.

The prototype implementation available on GitHub supports the protection of Linux and Android ARMv7 binaries and dynamically linked libraries compiled from C and C++ code. Only the C code is protected, however. The tools have been extensively tested and validated on native Android libraries that are packed in Android packages (together with Java apps) and in plug-ins that provide vendor-specific crypto and DRM services in the Android DRM and mediaserver framework.

A large part of the developed software prototypes is available as open source with extensive documentation, and more than 30 demonstration videos have been published on the project's demonstration Youtube channel. A significant part of the research has already been peer reviewed and many additional papers are still in the pipeline. Through keynotes and tutorials, including in workshops organized by the consortium, the European software protection community has been revitalized and has been made well aware of the project and its results.

Exploitation and impact

Some of the project results are already ready for commercial exploitation. A spin-off is in the making at Fondazione Bruno Kessler, and a technology transfer from the University of Ghent to industry has already taken place. Some of the specific protections developed within the project are used in products in the pipeline in business units of the industrial partners. As such, the project strengthens the position of European companies, including, of course, the project partners, whose business models depend on securing the assets embedded in their software. Other results are not ready for immediate commercialization. But with the whole ASPIRE Framework encompassing the compiler tool chain, the decision support system, many protections, and tools that automate the application of the software protection evaluation methodology, the consortium has demonstrated that measureable, assisted deployment of software protection is feasible. The open source availability of the framework will help the European R&D community to bridge the gap to commercial deployment of the ASPIRE approach, not least by providing all the foundational infrastructure necessary for complementing and expanding the expert knowledge already amassed in the project from the researchers' expertise, from professional penetration tests, from a public challenge, and from external advice.

YouTube demo video channel: https://www.youtube.com/channel/UCntMGBjHr_oW5wEd5JgjD6g
Open source repository: https://github.com/aspire-fp7

NAME: Advanced Software Protection: Integration, Research and Exploitation (ASPIRE)
START/END DATE: 01/11/2013 – 31/10/2016
KEYWORDS: mobile software security, compiler, decision support, evaluation methodology
PARTNERS: Universiteit Gent (Belgium), Nagravision (Switzerland), SFNT Germany, Gemalto (France), Fondazione Bruno Kessler (Italy), Politecnico di Torino (Italy), University of East London (UK)
BUDGET: €4.6M
WEBSITE: www.aspire-fp7.eu
The ASPIRE project received funding from the European Union's FP7 Programme under grant agreement no. 609734.

LEADING DATA CENTRES INTO THE FUTURE: EUROSERVER

Tasked with developing an energy-efficient server design that could be used to meet the demands expected for exascale computing beyond 2020, the EUROSERVER team has concluded the project having produced solutions which could halve the cost of powering data centres and well as greatly increase performance through memory compression.

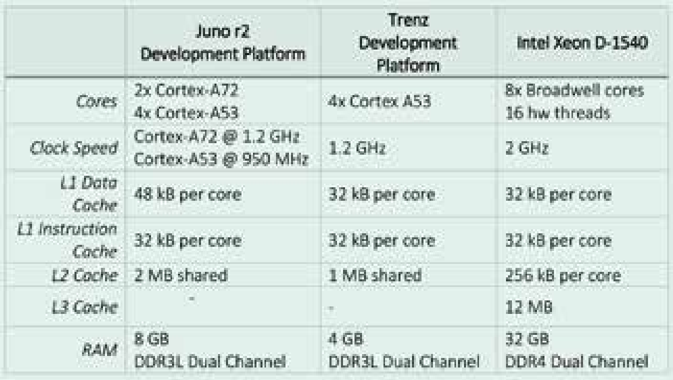
The project has also led to the development of two spin-off companies; KALEAO Ltd., headquartered in Cambridge, UK and ZeroPoint Technologies, a startup that has come out of Chalmers University of Technology in Gothenburg.

But what were the stages that took place behind these impressive outcomes and what new technical knowledge has been gained?

Getting ARM-based microserver designs into the data centre

Consortium partner ARM is a dominant force in the mobile device market where the energy-efficiency and popular instruction set of its processors has led to it being the instruction set of choice for mobile developers. Over the last few years, ARM-designed processors have looked to challenge the Intel-dominated data centre market.

The table below shows the experimental platforms that were investigated. They include a Juno ARM 64-bit development platform, a Trenz board with four energy-efficient Cortex-A53 ARM 64-bit processors and an Intel Xeon D-1540 that we believe is a realistic competitor to ARM in the energy-efficient compute domain.

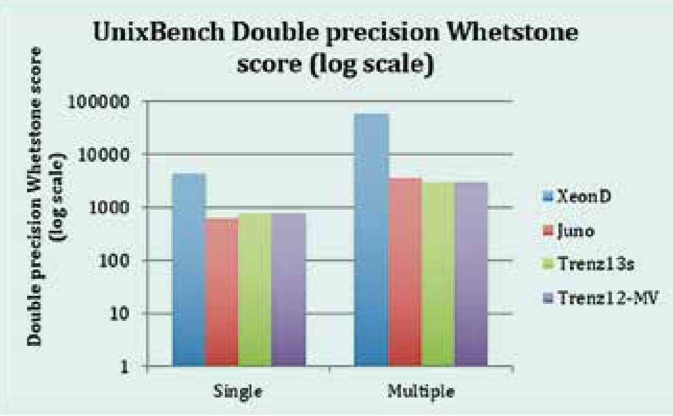


	Juno r2 Development Platform	Trenz Development Platform	Intel Xeon D-1540
Cores	2x Cortex-A72 4x Cortex-A53	4x Cortex A53	8x Broadwell cores 16 hw threads
Clock Speed	Cortex-A72 @ 1.2 GHz Cortex-A53 @ 950 MHz	1.2 GHz	2 GHz
L1 Data Cache	48 kB per core	32 kB per core	32 kB per core
L1 Instruction Cache	32 kB per core	32 kB per core	32 kB per core
L2 Cache	2 MB shared	1 MB shared	256 kB per core
L3 Cache	-	-	12 MB
RAM	8 GB DDR3L Dual Channel	4 GB DDR3L Dual Channel	32 GB DDR4 Dual Channel

The EUROSERVER platforms that were analysed

Some early adopters tried to integrate ARM processors into the data centre but used the ARM 32-bit architecture and hence the idea didn't gain traction. This has all changed with the advent of the 64-bit ARM architecture and, since then, many companies have investigated placing ARM-based micro-server designs into the data centre.

Yet ARM-based processors need to catch up with the large lead-time and massive inertia that Intel has established, the latter having control over the entire ecosystem from design through to fabrication. Intel-based processors make up 98% of the data centre market. The scores of typical benchmarks, such as UnixBench, suggest that Intel solutions are at least one order of magnitude more capable than the ARM-based solutions that are trying to compete with them, as shown in figure 1 below. Where EUROSERVER came in was to develop a server design that benefited from ARM's power efficiency and addresses some of its shortcomings so as to create a viable alternative to Intel-based solutions.



1: UnixBench, Whetstone test results for various devices under test (log scale)

> *"EUROSERVER's solutions could halve the cost of powering data centres and greatly increase performance"*

Hardware advances

Over the course of the project, a combination of hardware and software techniques were developed. On the hardware side, two prototype platform testbeds were created: a Juno R2 development board based system and a Trenz development platform. Both have energy-efficient, quad-core ARM 64-bit Cortex A53 processors, with the Juno differing in that it is also a big.LITTLE design and has a Cortex-A72.

The Trenz 0808-based, UltraScale+ system, seen in figure 2, combines a Trenz module with 4x A53 cores with a placeholder for a System-In-Package (SIP) 32-core A53. At the time of writing, the 32-core SIP is not ready but will be included in one of the follow-up projects that has resulted from EUROSERVER, including ExaNeSt, ExaNoDe and EcoScale.

2: The EUROSERVER designed, NEAT produced, prototype board. Not shown are a Trenz 0808 module and a SIP

Software breakthroughs

Processor manufacturers in recent years have been limited in how far the frequency envelope can be pushed due to power density, which has led to the rise of multicore chips. EUROSERVER has taken on board this change in design and has developed new scalable technologies, UNIMEM and the MicroVisor, that allow better scaling of compute and memory resources. These will be able to deal better with the exascale computing workloads that are expected in future data centres.

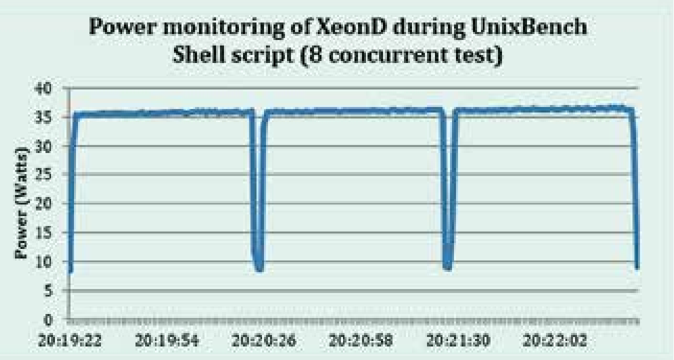
UNIMEM is a shared memory technology that allows multiple boards to share memory regions between them. This allows for better provisioning strategies and for greater in-memory workloads than are possible with current best-of-breed solutions. Memory from each board is divided into a local and a remotely addressable region. UNIMEM technology is a licensed IP technology and has been investigated by a number of companies and research organizations.

The MicroVisor is a new hypervisor technology derived from Xen. It is purpose made for low-power, energy-efficient platforms such as ARM that have many, albeit weaker cores. Traditional hypervisors are now quite 'bloated' and require a large amount of resources that are not available to ARM-based boards. Instead a lighter, more efficient platform has been developed that works natively with ARM and Intel architectures. The overhead for workloads running in virtual machines is near negligible, as seen in figure 1.

> *"ARM-based designs will have a place in the data centre of the future"*

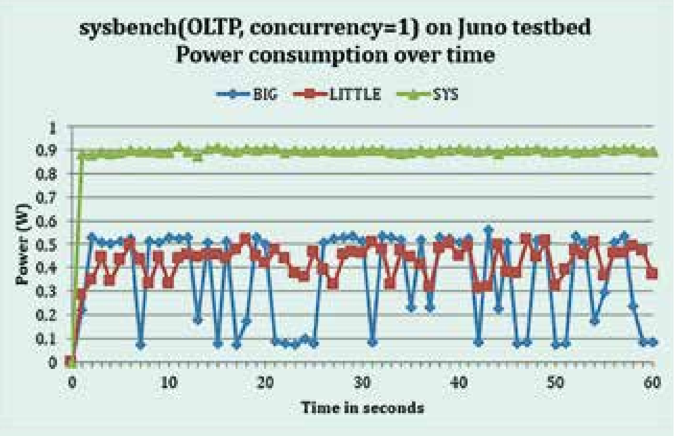
Energy-efficient platforms

Power monitoring techniques such as RAPL are used to expose the power utilized by the XeonD platform to be able to identify the power used by the processor during stages of a workload (see figure 3).



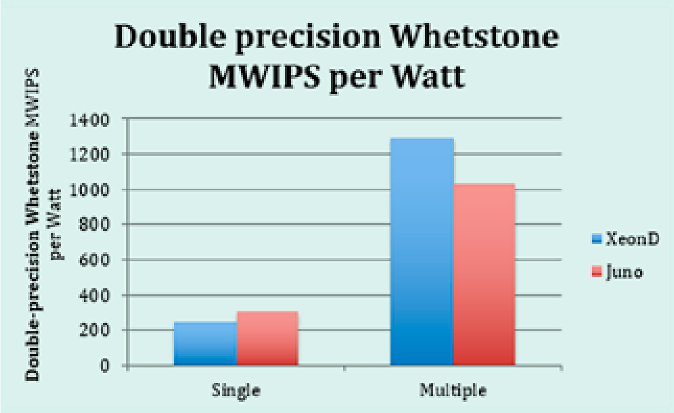
3: Power monitoring of the Intel XeonD while running a UnixBench Shell script test

The equivalent power monitoring has been exposed through kernel modules in the Juno platform to allow monitoring of the ARM system whilst running workloads (see figure 4).

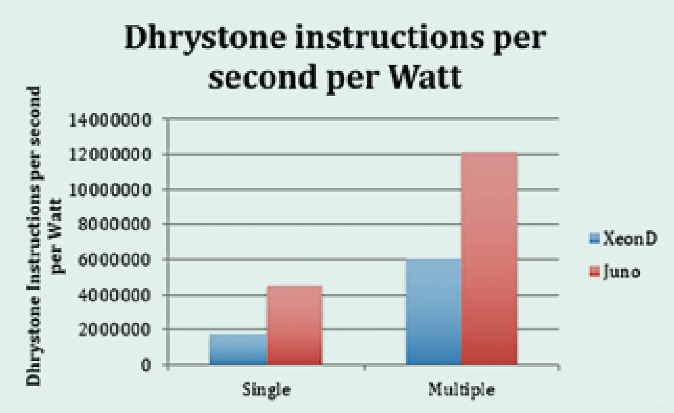


4: Power monitoring of the Juno R1 development board, whilst running SysBench OLTP workload

By looking at the power profile of the devices while investigating the workloads it is then possible to identify the power-efficiency of the platforms - as seen in figures 5 and 6. The power efficiency of the Juno platform shows that, although the ARM-based designs lag behind in raw performance values, they are more energy-efficient and will have a place in the data centre of the future.



5: These energy efficiencies were calculated by taking the Whetstone score and dividing by the average power usage recorded for the processor during this test

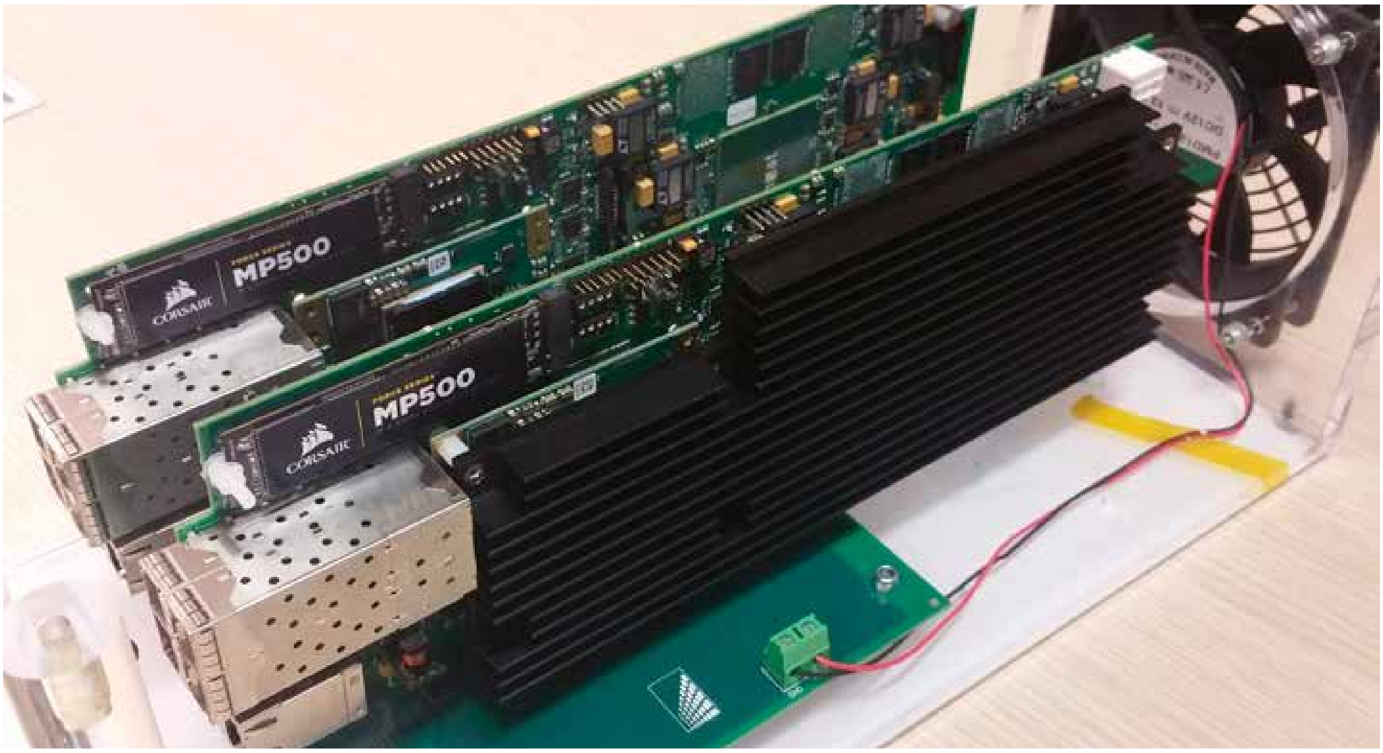


6: These energy efficiency values were calculated by taking the Dhrystone scores and then dividing by the average power usage during this test

The final EUROSERVER platform (see figure 7) combines a pair of UltraScale+ boards on a backplane that provides electrical and physical connectivity. These boards will be used in the several follow-up projects to form the basis of a 'European server', a server designed and built in the EU that will keep the continent competitive in the ever-changing global ICT market .

NAME: EUROSERVER: Green computing node for European micro-servers
START/END DATE: 01/09/2013 – 31/01/2017
KEYWORDS: microserver, energy-efficiency, memory compression, hypervisor, system integration, true convergence
PARTNERS: CEA-Leti (France), OnApp (UK), Foundation for Research and Technology Hellas (Greece), Barcelona Supercomputing Center (Spain), TU Delft (Netherlands), STMicroelectronics (France), NEAT (Italy), Chalmers University of Technology (Sweden) and ARM (UK)
BUDGET: €11.4M
WEBSITE: www.euroserver-project.eu

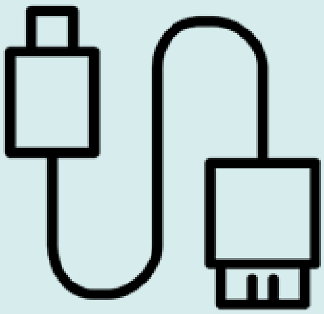
The EUROSERVER project received funding from the European Union's FP7 Programme under grant agreement no. 610456.



7: A pair of EUROSERVER boards, assembled onto a backplane with electrical connectivity, designed by EUROSERVER and produced by NEAT

Tech Transfer Award winners

2016 HiPEAC Technology Transfer Awards



In December 2016, we announced the winners of the latest round of Tech Transfer Awards. These annual awards recognize teams and individuals who have managed to turn research results into tangible services, products and enterprises. The winning technologies have had impacts spanning improvement of railway passenger safety, reduced cost of car insurance, and enhanced reliability and power efficiency from a wireless radio module for wide-ranging applications.

The 2016 winners were:

Daniel Hofman (University of Zagreb): S.W.A.T. – Sites of Web Assessment Tools

Silviu Folea (Technical University of Cluj-Napoca): Sub 1 GHz ISA100 technology for low cost and low power consumption embedded systems

Alastair Donaldson (Imperial College London): CLsmith in Collective Knowledge

William Fornaciari (Politecnico di Milano): Insurance telematics for reduced cost of ownership

Horacio Pérez-Sánchez (Universidad Católica de Murcia): Algorithmic developments in computational drug discovery, implemented on high-performance computing architectures

Jaume Abella (Barcelona Supercomputing Center): Increasing the real-time performance of the LEON family of processors

Martin Palkovic (IT4Innovations National Supercomputing Center): Improved passive safety and comfort of passengers in railway traffic

Bartosz Ziolko (Techmo): Sarmata speech recognition system

Per Stenström (Chalmers University of Technology): Blaze Memory: IP block for increasing the capacity of computer memory

Miguel Aguilar (RWTH Aachen): Automatic software parallelization and offloading technologies for heterogeneous embedded multicore systems

CLsmith IN COLLECTIVE KNOWLEDGE: Alastair Donaldson

The winning technology is CLsmith, a tool that automatically generates test cases to stress compilers for GPU programming languages. CLsmith originally targeted OpenCL, and was successful in finding a large number of defects in commercial OpenCL compilers (reported in a PLDI 2015 paper for which Alastair won a HiPEAC paper award).

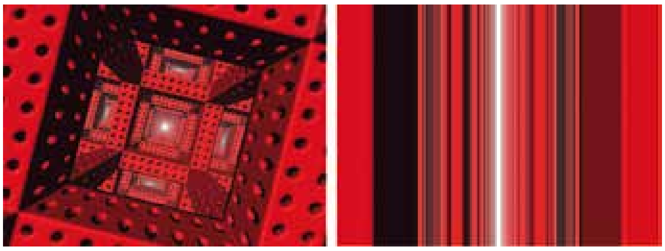
Since then, the Multicore Programming Group have developed a partner tool, GLFuzz, to generate tests for GLSL, the OpenGL shading language. Together, CLsmith and GLFuzz can be used to test a wide range of graphics compilers from vendors targeting both desktop and mobile graphics. A series of blog posts describes the GLFuzz technique and its application to industrial GPU drivers (bit.ly/2kRKgAR).

CLsmith and GLFuzz have enabled the discovery of a wide range of defects, including compiler crashes, compiler timeouts, cases where the compiler rejects valid code, cases where compiled code causes machine crashes when executed, and – arguably most seriously – cases where code that successfully compiles computes incorrect results with no other side-effects.

> ***"CLsmith and GLFuzz are being increasingly used by the many-core industry"***

Over the last year, and supported by technology transfer funding from the TETRACOM EU project, Imperial College London has worked with dividiti to integrate these tools with the company's Collective Knowledge (CK) framework. This enables seamless collection of data relating to compiler bug reports, querying of statistical properties of that data, reproduction of results across platforms, and comparisons between platforms.

CLsmith and GLFuzz are being increasingly used by the many-core industry; they are used routinely by some platform vendors to test their compilers. Their integration with Collective Knowledge will allow dividiti and Imperial to build on this early success, and move towards making CLsmith and GLFuzz the standard tools for assessing many-core reliability in industry.



On the left is a well rendered image; on the right is an image that has been badly rendered due to a bug. The framework detected the bug automatically.

ALGORITHMIC DEVELOPMENTS IN COMPUTATIONAL DRUG DISCOVERY, IMPLEMENTED ON HIGH PERFORMANCE COMPUTING ARCHITECTURES: Horacio Pérez-Sánchez

The Bioinformatics and High Performance Computing Research Group (BIO-HPC, http://bio-hpc.eu) at the Universidad Católica de Murcia works on the exploitation of HPC architectures for the development, acceleration and application of bioinformatics applications and its transfer to industry. The team's methodology can be applied to almost any bioactive compound discovery and design campaign and its main expertise resides in (but is not limited to) the discovery of drugs, biocides, pesticides, agrochemicals and nanomaterials. BIO-HPC created marketable solutions for implementation of computational drug discovery (CDD) technologies on HPC architectures in direct response to the needs of several specific companies. Projects include:

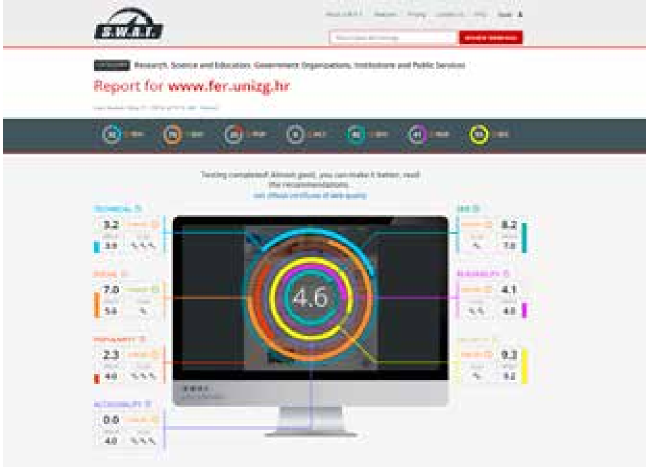
- Two international patents related with CDD and HPC were licensed to a multinational technological company in 2015. As a consequence, the Nanomatch company was created (https://www.nanomatch.de).
- BIO-HPC signed a technology transfer agreement with Artificial Intelligence Talentum SL (http://www.aitalentum.com/), so that the company would market the group's CDD developments on HPC architectures to other research groups and small pharma and biotech companies. This partnership took place as a result of funding from TETRACOM (http://www.tetracom.eu/).
- In the activity described above, BIO-HPC acquired relevant and practical knowledge about the interests of CDD on the HPC market. One particular idea of commercial interest was the commercialization (using the 'Software as a Service' or SaaS business model) of some concrete CDD on HPC technology developed by the group: Blind Docking Server (BDS). Funding was received from the Eurolab-4-HPC Business Prototyping fund. Three technological companies provide mentors; one is Angel Pineiro, founder of MD.USE (http://mduse.com/en/), a company specialized in offering scientific software to pharma companies. The company has confirmed that it is interested in the BDS system and wants to have commercialization rights.
- Alongside FX Talentum (http://www.fxtalentum.com/en/), a company working in this field, the group has been awarded technology transfer project funding from the Spanish government for research into the application of machine learning techniques, on HPC architectures, to CDD. Some algorithms that can be applied not only to CDD but also to other domains in scientific computing, such as algorithmic trading, have been developed.

S.W.A.T. – SITES OF WEB ASSESSMENT TOOLS: Daniel Hofman

To respond to the growing need for fast and reliable website quality assessment, knowledge in this domain has been transferred by the Faculty of Electrical Engineering, University of Zagreb to industry partner VIDI-to and turned into a valuable tool for website assessment: S.W.A.T. The tool was built in a modular and scalable manner so that it includes state-of-the-art programming models and is extendable to new methods in the future. It not only assesses quality of obvious components or those easily checked by technical specifications (adherence to standards) by simple pointing to non-adherences, but also benefits from much wider inputs and aspects.

> *"The impact of such a tool will have a profound influence on the creation, production and maintenance of websites, thus improving the web itself"*

Quality assessment results support decision-making on changes or improvements on web portals and sites. The impact of such a technological tool will have a profound influence on the creation, production and maintenance of websites, thus improving the web itself. The 'engine' of the innovation (source code, algorithms) could also be the basis for other sorts of services, therefore setting up a completely new technological field ready for further development and applications.



Report for www.fer.unizg.hr

The S.W.A.T. system operates on a highly virtualized infrastructure with data storage in a cloud. This allows flexible expansion of the system with the growth of the number of sites being tested. S.W.A.T. is a highly automated in-depth tool based on scientifically proven assessment algorithms, assessing and weighting elements and aspects of quality from various fields (technological, user, marketing, commerce). The algorithms, which are technological science-based innovation, are the most valuable component of the project.
http://swat.technology/

The DEWS (Design methodologies of Embedded controllers, Wireless interconnect and Systems-on-chip) Centre of Excellence at the Università degli Studi Dell'Aquila has been carrying out the testing and validation of a parallelization technique pioneered at ScienSYS, an SME based in France.

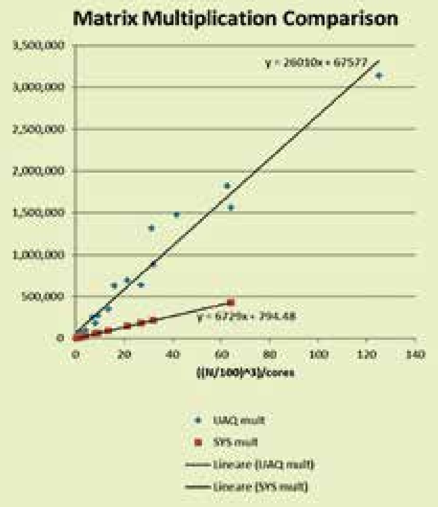
A runtime parallelization approach for shared memory architectures

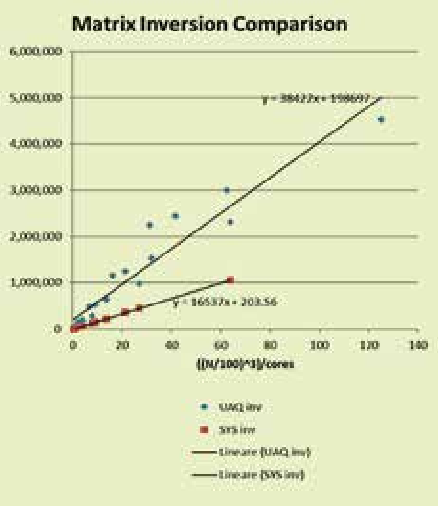
Multi-processor systems are becoming increasingly widespread in embedded systems thanks to the benefits of workload sharing, including faster computation time and decreased power dissipation. However, programming a multi-processor architecture generally requires effort from the programmer to exploit the platform to its true potential. ScienSYS has developed a new parallelization technique that targets multicore architectures with shared memory. Here at DEWS, we have evaluated it by running two computationally intensive algorithms and comparing the response times with the ones obtained using OpenMP-based parallelization.

The ScienSYS technique provides automatic parallelization at task level during runtime: any procedure can be automatically executed on any available executive unit of a multiprocessor system, as soon as required inputs are available. With this technique, the data availability alone drives the whole computing process. A private task stack is created for each executive unit that should be contained in a high-speed memory (e.g. a first cache level).

The tests we have carried out are represented by two algorithms involving N-dimensional square matrices. They perform respectively matrix multiplication and matrix inversion and both show a time cost equal to O(N3). They have been run on a platform composed of four Gaisler LEON3 processors, connected in SMP mode with shared memory, implemented on a Virtex 7 FPGA. Two platform versions were constructed: the first, V1, has on-chip memories for each core and no MMU. The second, V2, has no on-chip memories and does have MMU. V1 was selected to execute the tests with the ScienSYS approach to parallelization, which does not require an operating system (OS). V2 was chosen to exploit the OpenMP approach, implemented by using GCC implementation of OpenMP (i.e. gomp): this approach requires a Linux OS. We ran tests with N ranging from 100 to 400, and considering one, two, three and four processors. We collected response times using a hardware profiling system developed here at DEWS.

The performances achieved in terms of computing speed show that the ScienSYS approach works faster than when OpenMP is used, in both matrix multiplication and inversion. Figures 1 and 2 show the comparison for both cases respectively: the graphs represent the trend of the response time (y-axis) according to the theoretical factor applied to response time when varying inputs sizes N and number of cores, namely Qf. Qf is the cubed input dimension divided by the number of cores (in the ideal case of fully parallelizable code). The black linear trend lines show the mean growth rate in both cases, moving from left to right, namely increasing N and decreasing number of cores. Slope ratios indicate that the rate of growth in the case of the OpenMP approach is faster than in the case of the ScienSYS approach by a factor 3.9x for matrix multiplication, and 2.3x for matrix inversion.This indicates that the ScienSYS approach is better than the OpenMP one: for example, when inverting a 400 size matrix using three cores, computation time with the ScienSYS technique was 3.51x faster than with OpenMP. With this type of validation, the company is in a position to embark upon the process of bringing the product to market.





To commercialize research from a collaborative project requires not only an innovative and an in-demand product but also time, patience and funding. Chris Brown of St Andrews University explains the technology that he and colleagues are in the process of bringing to market.

ParaFormance™: Democratizing Multi-Core Software

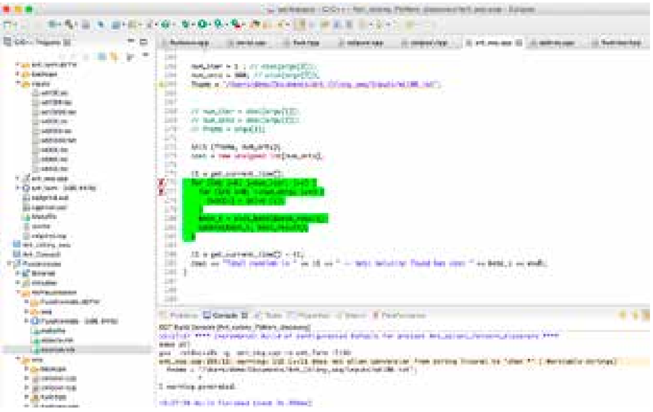
Multi-core computers have revolutionized the hardware landscape, providing high-performance, low-energy computing. However, as we are all painfully aware, programming highly-parallel systems remains complex, time-consuming and error-prone. Our research shows that fewer than 5% of programmers have the skills to deal successfully with the challenges that are posed by current multi-core systems, and this will become worse as we move towards heterogeneous many-core systems. ParaFormance™ takes a new approach. Building on the easily understood and widely accepted idea of programming patterns, and expanding on successful work from our FP7 and Horizon 2020 projects, we are developing a new toolset for building highly parallel software rapidly and safely. We aim to bring this to market quickly and effectively.

> *"ParaFormance™ delivers multi-core and many-core software on time, on budget, and without expensive errors."*

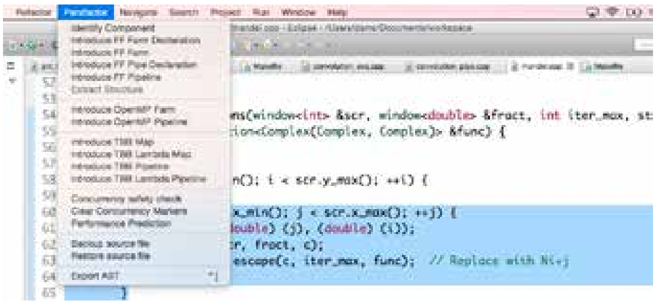
The ParaFormance™ Technology

ParaFormance™ comprises three core features:

- **Parallelism Discovery**: Our unique and sophisticated parallelism discovery feature finds the parts of the application that can be parallelized, automatically. With our own built-in intelligent heuristics and analytics, ParaFormance™ ensures that it reports only the parts of the application that will benefit from parallelization, removing false-positives. The results are displayed in an easy to read and clear way directly in the integrated development environment , and our sophisticated reporting system then allows them to be analysed at leisure.



- **Parallelism Insertion through Refactoring**: After discovering the sources of parallelism within the application, ParaFormance™ can then automatically refactor the code to prepare it for parallelization. Our advanced refactoring support is built on pattern-based technology that enables it to target many different parallelization libraries and platforms, e.g. Intel's Thread Building Blocks (TBB), OpenMP, pThreads, and more.



```
ns(window<int> &scr, window<double> &fract, int iter_max, st
tion<Complex(Complex, Complex)> &func) {

n(); i < scr.y_max(); ++i) {

x_min(); j < scr.x_max(); ++j) {
double) (j), (double) (i));
r, fract, c);
escape(c, iter_max, func);  // Replace with N(+j
}
```

- **Advanced Safety Checking**: Our advanced safety-checking features provide confidence that the parallel version of an application is correct and bug-free, both for parallelism that has been inserted via our refactorings or that is handwritten. This includes both static and dynamic checks, including race condition detection.

From EU research to spin-off

ParaFormance™ delivers a key technology that has been developed in a number of EU projects. ParaPhrase, a €4.5M EU FP7 project (2011-2015, http://paraphrase-fp7.eu), focused on new techniques and tools for improving the programmability of multi-core systems. The refactoring technology that now lies at the core of the ParaFormance technology was one of the tool prototypes that came out of ParaPhrase (2015-2018, http://rephrase-ict.eu), a €3.5M Horizon 2020 project that involves nine European partners: the University of St Andrews (UK, coordinator) IBM Research (Israel), EvoPro (Hungary), CiberSam (Spain), SCCH (Austria), PRQA (UK), the University of Pisa (Italy), the University of Turin (Italy) and University Carlos III Madrid (Spain).

Building on the success of these EU projects, the St Andrews team has successfully secured £450,000 of Scottish Enterprise funding (from the Scottish government) to take the technology to a commercial standard and to form an internationally recognized company.

User Trials

Initial user trials with two companies have shown very successful outcomes. In one trial, a complex 2.5M line legacy C++ application was analysed and parallelized using the ParaFormance™ technology, in about ten minutes. In a second trial, a 5000 line C++ application was analysed and parallelized by ParaFormance in a couple of hours (including installing the tool). Parallelizing either application would normally take a specialist developer weeks or months of manual effort. In both cases, we have been able to achieve significant and scalable speedups on the target architectures.

The ParaFormance Team

Philip Petersen – *Commercial Champion*

Philip brings significant commercial expertise as the former CEO of AdInfa, a successful high technology startup company which he has recently left, moving to Scotland from London. He has excellent connections with the UK business and investment communities. Prior to forming AdInfa, Philip established and ran successful sales and marketing teams at UK and international level.

Dr Chris Brown – *CTO Elect*

Chris brings key technical expertise to the project. His PhD work on refactoring, and subsequent research on three successful EU-funded projects, forms the basis for the ParaFormance technology. He will be responsible for developing the ParaFormance technology towards a successful commercial outcome and will transfer to the newly formed company as its Chief Technical Officer. He has previously worked as a software engineer for Technium CAST, a start-up software company in Wales.

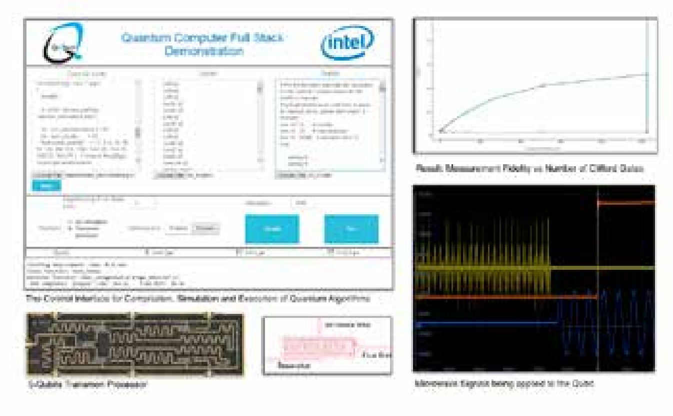
Professor Kevin Hammond – *Adviser*

Kevin has over 30 years of experience in parallel and multi-core computing. He is the author of over 100 research papers and books, and has been involved in the design and implementation of several programming languages. He has run over 20 national and international research projects, valued at over £14M in total, and involving up to 25 employees at thirteen sites.

For more information about ParaFormance™, contact Chris Brown (chris@paraformance.com) or visit www.paraformance.com.

QuTech and Intel demonstrate full stack implementation of programmable quantum computer prototype

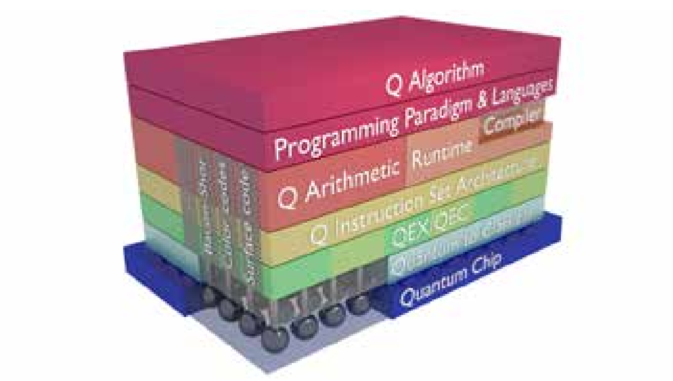
The potential for quantum computers to revolutionize computing systems is immense, but so far there have been few tangible results behind the hype. Now, researchers at the QuTech research centre, in collaboration with Intel, have made a significant step forward with their demonstration of a first full-stack implementation of a programmable quantum computer.


Quantum Computer Full Stack Demonstration
intel
The Control Interface for Compilation, Simulation and Execution of Quantum Algorithms
Result: Measurement Fidelity vs Number of Clifford Gates
5-Qubits Transmon Processor
Microwave Signals being applied to the Qubit

Quantum computing is evolving rapidly, in particular since the discovery of several efficient quantum algorithms, such as Shor's factoring algorithm, that can solve intractable classic problems. However, the realization of a large-scale physical quantum computer remains very challenging. To address this, researchers at QuTech, a quantum computing research centre founded by TU Delft and TNO, are collaborating with colleagues at Intel to investigate the different architectural components of a quantum computing system.

Thanks to their efforts, a first full stack implementation of a programmable quantum computer targeting two different superconducting quantum processors was recently demonstrated as a first proof of concept of an operational architecture. The proposed quantum computer system stack includes a quantum programming language to express quantum algorithms and a compiler that compiles these algorithms into quantum instructions. These instructions can then be executed on the quantum processor through the control electronics or can be simulated on the QX universal quantum computer simulator developed at QuTech. Although the two quantum processors are based on the superconducting qubit technology, the layers of the proposed system stack provide enough abstraction to offer high portability over different qubit technologies.

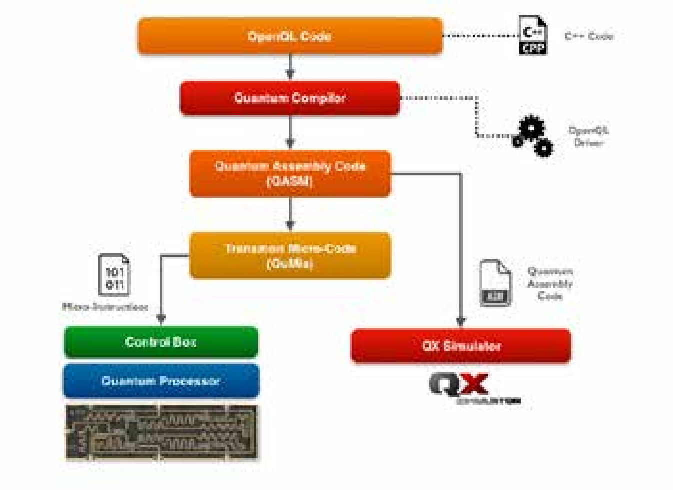
The quantum computer system stack



Overview of the quantum computer system stack

When defining and building an architecture for a quantum computer, it is necessary to understand how to address and control a larger numbers of qubits. As shown in Figure 2, building a quantum computer involves implementing different functional layers. At the highest level, algorithm designers formulate quantum algorithms such as Shor's factoring algorithm in a high-level language that is designed to represent not only quantum operations but also classical logic, which will always be necessary. A compiler will then translate those algorithms into the instruction set that can be executed on the quantum computer. Similarly to traditional computers, the code generated by the compiler is at assembly level, and the assembler we have extended for this purpose is called Quantum Assembler (QASM). A micro-architecture will provide the hardware-based control logic needed to execute the instructions on the target quantum chip. These instructions are translated into micro-instructions and, through the interface layer, sent into the qubit plane.

In our demonstration we implement a simplified version of the system stack while preserving its different layers.



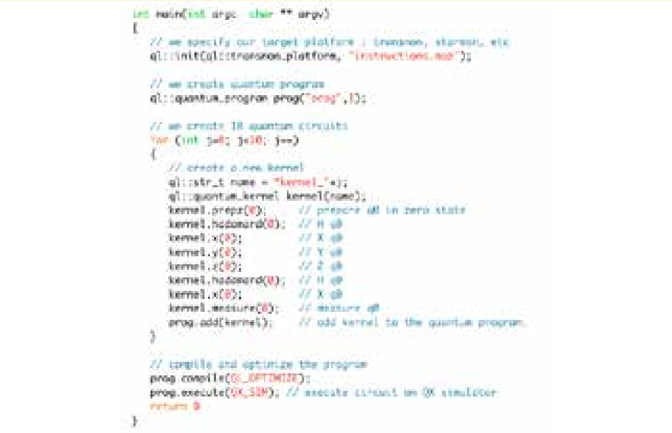
The functional flow: from quantum software to the quantum hardware

The full stack implementation: from software to hardware

The implementation of a simplified system stack is organized as follows: the quantum algorithms are expressed in OpenQL, which is a high-level quantum programming language. The OpenQL code is then compiled and optimized to produce an abstract (platform-independent) Quantum Assembly code (QASM) and a platform-specific Quantum Micro-code (QuMis). The QASM execution can be simulated using our QX universal quantum computer simulator [http://www.quantum-studio.com/] while the QuMis code can be executed by the Control Box (classical electronics) on the target quantum processor. We used two different quantum processors, the Transmon and the Starmon, to demonstrate the portability of the stack over different underlying hardware.

OpenQL : writing quantum algorithms

OpenQL framework is a high-level quantum programming framework that uses the standard C++ language as a host language and defines a quantum programming interface (QPI) to write quantum programs as a set of 'Quantum Kernels'. These kernels allows the programmer to write quantum algorithms while mixing quantum and traditional code. A quantum kernel is primarily composed of a set of quantum gates operating on different qubits. In the example shown in Figure 4, we create ten Quantum Kernels that apply an arbitrary sequence of quantum gates to one qubit. We add these kernels to our Quantum Program then we compile it while enabling optimizations to produce an efficient quantum code. After compilation, the code can be simulated using the QX simulator while the compiled micro-code can be executed on the physical platform.



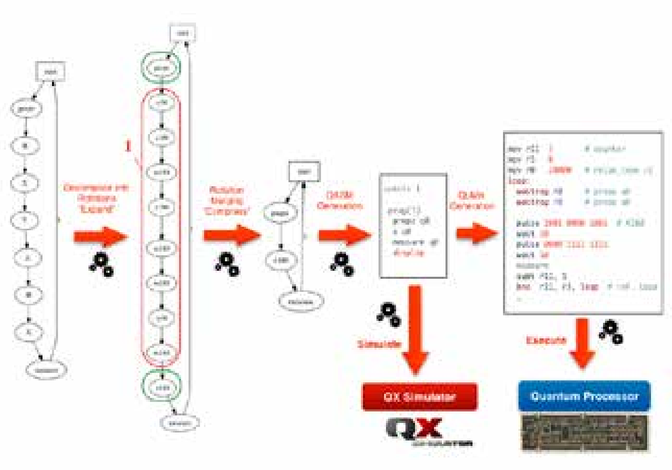
```
int main(int argc, char ** argv)
{
   // we specify our target platform : transmon, starmon, etc
   ql::init(ql::transmon_platform, "instructions.map");

   // we create quantum program
   ql::quantum_program prog("prog",1);

   // we create 10 quantum circuits
   for (int j=0; j<10; j++)
   {
      // create a new kernel
      ql::str_t name = "kernel_"+j;
      ql::quantum_kernel kernel(name);
      kernel.prepz(0);       // prepare q0 in zero state
      kernel.hadamard(0);    // H q0
      kernel.x(0);           // X q0
      kernel.y(0);           // Y q0
      kernel.z(0);           // Z q0
      kernel.hadamard(0);    // H q0
      kernel.x(0);           // X q0
      kernel.measure(0);     // measure q0
      prog.add(kernel);      // add kernel to the quantum program
   }

   // compile and optimize the program
   prog.compile(QL_OPTIMIZE);
   prog.execute(QX_SIM); // execute circuit on QX simulator
   return 0
}
```

Example of an OpenQL code which create ten arbitrary quantum kernels

Compilation and optimization of the quantum code

As we saw in the previous section, the quantum algorithms are composed of both traditional code and quantum code. The classical code is compiled by a standard C++ compiler while the quantum kernels are compiled using our OpenQL driver which converts the quantum kernels into quantum circuits, then optimizes and compiles these circuits to produce a QASM code and an executable QuMis code.

A simple overview of the main compilation phases is given in Figure 5, which depicts the compilation steps corresponding to the previous simple OpenQL code example. We can distinguish two main steps where the original quantum gate sequence is decomposed into elementary qubit rotations then optimized by merging them into shorter rotation sequences to perform the maximum number of operations within the limited coherence time of the qubit and achieve the highest possible fidelity.



Overview of the Quantum Kernels Compilation Phases

During the first stage of the compilation, the circuit gates are decomposed into a set of elementary qubit rotations which are supported by the target quantum processor. The rotations of the expanded circuit are then merged whenever possible to produce an efficient compact circuit. For instance, the first sequence of eight gates corresponds to an identity and can be cancelled out to leave only the meaningful rotation at the end of the circuit.

The compiler produces an intermediate quantum assembly code (QASM); the produced code is not platform-specific and can be simulated in QX. The next step is that a platform-specific micro-code is generated for the target physical platform.

Quantum Circuit Simulation using QX

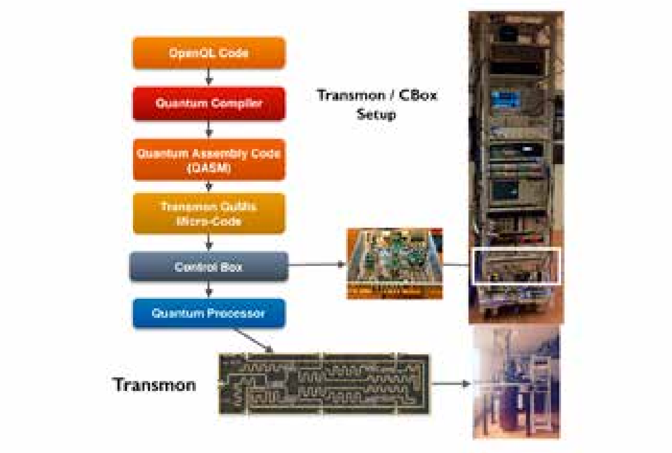
The QX Simulator is a high-performance universal quantum computer simulator that allows the simulation of quantum circuits under various quantum noise models corresponding to different quantum technologies. The QX simulator can simulate up to 34 qubits on a single node of our simulation server. The circuits are described by the input QASM code.

Besides keeping track of the quantum state during the circuit execution and displaying the qubit measurement outcomes, the QX simulator can emulate some control electronics units such as the measurement integration and averaging unit which averages the qubit measurement outcomes after multiple circuit execution iterations. For instance, this feature allows us to produce results that are similar to the real hardware.

Micro-code Execution

We defined the Quantum Micro-Instruction Set (QuMIS) which can be used to control quantum operations applied on the quantum processor with precise timing. We designed the QuTech Control Box (CBox) which implements a QuMA core. The QuMA core provides the execution support for the defined QuMIS to perform quantum computation in a programmable way by controlling the underlying electronic devices using QuMIS instructions. For now, the QuMIS contains five main instructions: pulse, wait, waitreg, measure and trigger.

The 'pulse' instruction triggers the arbitrary waveform generators to emit the specified RF signals, the 'wait' instruction control the timing, the 'measure' instruction triggers the measurement discrimination while the 'trigger' instruction generates digital outputs to control external hardware.



Hardware Setup for Operating the Transmon Quantum Processor

Hardware Setup

In order to demonstrate the high abstraction provided by the layers of our architecture, we used two different quantum processors which are the five qubit Transmon processor and the two qubit Starmon processor. Figure 6 shows the hardware setup driving the Transmon quantum processor.

Despite the two hardware setups being different, the exact same high-level OpenQL code can be executed on both platforms without any changes. The compiler adapts to the target hardware and produces a different micro-code for each platform. In future works, the hardware support will be extended to the spin qubit technology.

Just how powerful might quantum computing be?

The Shor's factoring algorithm is often seen as the 'killer' application that demonstrates the supremacy of quantum computing. It is designed to find the prime factors of a large integer number which can be used to break the widely used RSA asymmetric cryptography scheme. Based on this algorithm, a quantum computer can factor a large number of N bits in polynomial time (in Log(N)) using Shor's algorithm while a regular supercomputer requires exponential or sub-exponential time in the best cases (General Field Number Sieve (GNFS)) to solve the problem.

John Martinis (Google) made a very useful estimation of the required size and power of a traditional supercomputer to factor a 2048-bit number: it would require a supercomputer nearly as big as North America, which (assuming linear scaling) would:

- Cost $106 trillion
- Consume 106 terawatts of power (and would consume all of the earth's energy in one day)
- … and take 10 years to solve the problem!

In contrast, a quantum computer with 200 million qubits (admittedly, we are still far from that) would take only 24 hours to complete the task while consuming less than 10 MW of power, and would be just 10x10m in size.
See John Martinis' talk at bit.ly/2mHuRkc

The demonstration described was conducted by Nader Khammassi, Xiang Fu, Adriaan Rol, Leonardo Di Carlo and Koen Bertels with the contribution of a number of researchers from QuTech to the architecture design, the manufacturing of the quantum chips and the electronics design. This project is funded by Intel Corporation.

For more information about the quantum software and hardware used in this demonstration:
http://www.quantum-studio.com
http://www.qutech.nl

Jaume Abella, Francisco Cazorla and Carles Hernandez of Barcelona Supercomputing Center (BSC) explain Leopard, a new technology that will enable users to cope with the ever-increasing complexity of hardware in critical systems.

Leopard: a high-performance proce

The number and complexity of critical real-time functionalities in embedded systems is on the rise. This results in a relentless demand for increased levels of guaranteed computing performance that cannot be provided with simple single-core microcontrollers. Instead, multi-core processors with high-performance features such as cache hierarchies need to be used in those critical real-time embedded systems (CRTES). However, the intricate timing behaviour across complex hardware in multi-cores is a challenge for deriving worst-case execution time (WCET) estimates.

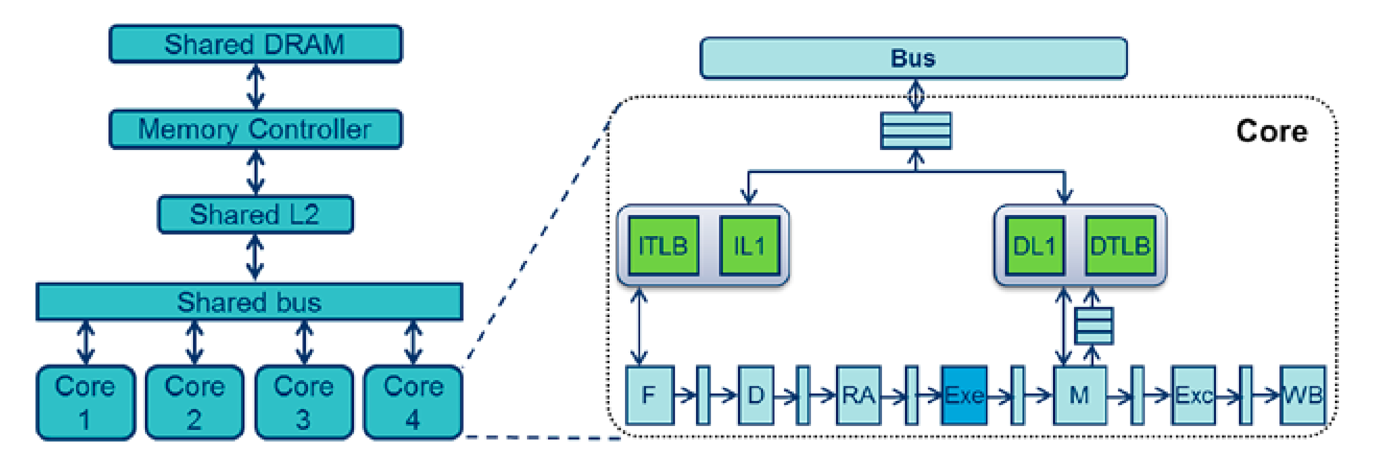
In response to this, BSC and Cobham Gaisler, as part of the EU-funded PROXIMA project, have jointly developed Leopard, a pipelined 4-core LEON-based processor with an advanced cache hierarchy. Leopard is especially suited for the space domain and provides higher levels of performance than average microcontrollers in CRTES. Indeed, Cobham Gaisler is already advertising it to customers. A key feature of Leopard is that, unlike common off-the-shelf multi-cores, it is well suited for measurement-based timing analysis.

Design principles

Leopard has been designed in such a way that the jitter (i.e. execution time variability) and worst-case behaviour of processor resources arise during the testing campaign. This helps reduce to quantifiable levels the uncertainty about unobserved timing behaviours. To that end, Leopard leverages time randomization and time upper-bounding techniques to naturally expose execution time jitter in the testing campaign while preserving high-performance features.

Time upper-bounding has been shown to be suitable for floating-point units with data-dependent latencies and for modelling the degree of contention in shared resources. Meanwhile, time randomization has been shown to fit several components such as cache placement and replacement, as well as arbitration to access shared resources (i.e. a shared bus or a shared memory controller).

- Random cache placement maps addresses to cache sets randomly and independently across different program runs so that whether two addresses are placed in the same cache set or not is a purely random event. This allows the dependence between memory location and cache set placement to be broken, thus releasing the end-user from having to control where objects are placed in memory, which is an arduous task due to the difficulty of controlling stack, code, libraries, operating system (OS) code and OS data location in memory and of preserving those locations upon integration of different functionalities.
- For the arbitration logic in a shared bus or network-on-chip, during system analysis Leopard deploys randomized arbitration across the maximum number of contenders. This allows WCET estimates that hold valid during operation to be obtained because the worst degree of contention has already been



ssor for critical real-time software

accounted for during the analysis phase, and the particular time when requests arrive at the shared resource is irrelevant because arbitration is random. Thus, the end user does not need to guess what other functionalities running in other cores will do in the shared resource or when. Instead, the end user can estimate the WCET of its application in isolation, still obtaining guaranteed high performance.

Timing analysis

By building upon time randomization, Leopard exposes time jitter in a probabilistic manner. Therefore, it matches perfectly the requirements of the measurement-based probabilistic timing analysis (MBPTA) techniques also developed as part of the PROXIMA project. MBPTA uses statistical techniques such as extreme value theory to predict the timing behaviour that can occur with arbitrarily low probabilities (e.g. 10-12 per run) based on small execution time samples (e.g. 1,000 execution time measurements).

Validation

Leopard implementation on an field-programmable gate array (FPGA) prototype has been successfully assessed with a number of use cases from the European Space Agency and Airbus Defence and Space, as well as with the central safety processing unit of the European Train Control System (ETCS) reference architecture provide by IK4-Ikerlan. Results show a moderate average performance degradation when compared with the original 4-core LEON-based processor: typically below 10%, and often close to just 1%. On the other hand, (probabilistic) WCET estimates are always above the observed execution time for the worst scenarios that could be produced manually. Yet they tightly upperbound observed execution times, therefore providing evidence on the reliability and tightness of provided WCET estimates, as needed for safety and resource efficiency.

In terms of the cost to hardware, all the modifications required to implement Leopard incurred an area increase as low as 2% in the FPGA and had no impact on the maximum operating frequency. Moreover, Leopard has been implemented with configurability in mind: time randomization and time upper-bounding can be disabled from the software level so that non-critical tasks can be run on the default setup. Also, worst-case conditions needed to estimate the WCET during the analysis phase can be enabled and disabled at will so that they can be accounted for during the analysis phase, but can be disabled during operation for better average performance and lower energy consumption.

High-speed tracing

Last but not least, different timing analyses require different degrees of tracing information from the applications under analysis. For instance, some timing analyses need to collect information about a subset of the instructions or even about all of them. The default tracing mechanism was unable to cope with the tracing speed needed for some timing analyses, so Leopard has been extended with a powerful Ethernet tracing feature able to collect abundant information at high speed. In particular, the debug interface is used to dump traces in a separate memory region with a dedicated memory controller so that those traces can be dumped to the host asynchronously through the Ethernet interface without interfering with the timing measurements.

What's next?

Leopard has already been acknowledged as a promising technology and received a HiPEAC Technology Transfer Award in December 2016. Cobham Gaisler, already advertising the technology on its website, has plans to include it in some of its future processors. Leopard is currently being enhanced at BSC in continued collaboration with Cobham Gaisler, within the scope of a project funded by the European Space Agency, to allow the WCET of critical tasks on a shared second level cache to be estimated for the first time in CRTES.

COBHAM
BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Magnus Peterson, Synective Labs AB
Technology opinion: FPGA acceleration goes mainstream

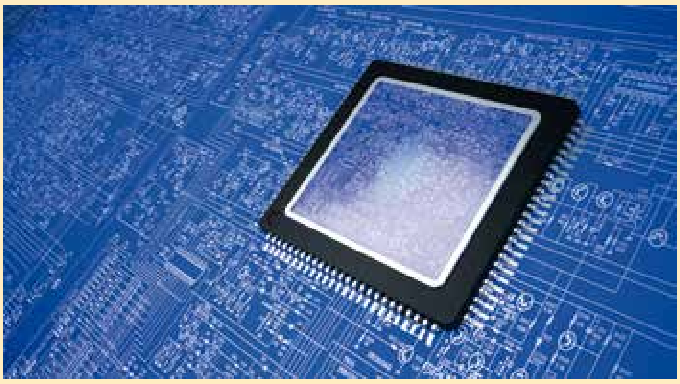
Field-programmable gate arrays (FPGAs) are those reprogrammable devices that for a long time have played an important role in very specific applications like mobile base stations and radars, but that have never really achieved a wider usage. With the ability to accelerate compute-intense tasks with an order of magnitude and with a fraction of the power consumption compared to competing devices, FPGAs are very appealing for embedded designs. Their flexibility to adapt to almost any interface standard and the potential cut in time to market they offer by being field re-programmable, makes the case even stronger. Unfortunately, FPGAs have been difficult and time-consuming to program, with only the low-level languages VHDL and Verilog at hand, and this has held back every attempt at wider acceptance.

But things finally now seem to be changing, thanks to several factors pointing in the same direction. Both Xilinx and Intel (Altera), the two big FPGA vendors, are finally offering tools for programming FPGAs using high-level languages like C/C++ and OpenCL. ARM cores have moved into FPGA chips forming SoC FPGAs, which have quickly become favourite system components for embedded designs. And with ARM cores on board, FPGAs have been discovered by software developers, who are now making use of the new high-level programming capabilities and realizing the potential these devices offer.

> *"Although FPGAs have been known to offer high performance, floating point operations have always been a weak spot. But that is no longer true."*

And probably even more important is that some of the big players have started to make their moves in the direction of FPGA-based server acceleration. Intel's acquisition of Altera is now resulting in the launch of a new Xeon processor with a tightly integrated Arria 10 FPGA, on the same chip. This will open the path to new, interesting possibilities. For their part, Microsoft has, after a successful project called Catapult that aimed to accelerate Bing searches with FPGA technology, launched the follow-up project Catapult v2. By integrating FPGAs into its Azure clusters, the company now offers FPGA-accelerated Deep Learning applications, completely seamless for the user, but with substantial savings in power and equipment for Microsoft. Amazon is also taking steps in the same direction by offering user programmable FPGA equipped nodes, 'F1 instances', as part of its BWS cloud services.

Although FPGAs have been known to offer high performance, floating point operations have always been a weak spot. But that is no longer true. By integrating hard floating point cores, the new Arria 10 FPGA family offers up to 10 TFLOPS of single precision floating point performance, making it a game changer.



On top of all this, FPGAs seem to be making their way into the automotive field, in systems for ADAS and autonomous driving – as image and signal processing at low power is one area where they really shine. This may ultimately lead to production volumes the FPGA vendors could so far only dream of.

High performance, low power, mature and easy to use tools, new high-volume markets and new, game changing FPGA devices – most things speak in favour of FPGAs right now. Will 2017 finally be the year that FPGAs have their ultimate breakthrough?

Career talk: Darko Gvozdanović, Manager Engagement Practice eHealth, Ericsson Nikola Tesla

With many years at Ericsson Nikola Tesla under your belt, you are a member of management of its Health Unit. Tell us a little about your career journey.

Since graduating with an MSc from the Faculty of Electrical Engineering & Computing in Zagreb in 2004, I have spent my whole career at Ericsson Nikola Tesla, the local Ericsson company in Croatia. In 2002, just as I completed two years in the research department, the Croatian government issued a tender for implementation of a national eHealth platform. From that moment onward, Croatia's eHealth system and my career have gone hand in hand. The initial years were dedicated to capturing and analysing requirements, remodelling eHealth processes and cooperating closely with different actors in the healthcare system to define the eHealth system architecture. Down the line, I become head of the eHealth department and responsible for our company's eHealth portfolio and overall solution architect for the Croatian national eHealth system. Indeed, one of the best moments of my career was when we launched the paperless national ePrescription functionality. Playing one of the lead roles in a solution which has transformed for the better the lives of everyone in the country, in an area as important as health, is magnificent. Not many jobs in the world offer such an opportunity. And this would not have been possible without my many great co-workers, the majority of whom eat, sleep and breathe eHealth just like me.

What are your department's main current priorities? And what's the best part of your job?

In the meantime, we have successfully implemented a national eHealth system in the Republic of Armenia and we are in the process of implementing a 'Health Information Systems Informatization and Interoperability Platform' in the Republic of Kazakhstan. I would say that the main priorities of the eHealth department are constant improvements in our eHealth portfolio and building capabilities to support multiple projects in Croatia and abroad.

Supported by the innovative atmosphere of my company and surrounded by such smart and passionate colleagues, I often catch myself spending several hours discussing different ways of supporting improvement in healthcare systems in different countries and in general. Knowing that you can transform these ideas into a concrete portfolio and, even more, witnessing real life implementation is very rewarding. Although interactions with many professionals with completely different backgrounds (doctors, pharmacists, public health specialists, and so on) might be challenging, it is the spice that makes my working days so interesting.

Caption: Darko and company President Mrs Kovačević welcome Albert II, Prince of Monaco

You are doing your PhD at a later stage of your career than many researchers. What are the main advantages (and disadvantages) of doing this?

I am currently a PhD student in the area of eHealth systems interoperability. Interoperability in healthcare is still a long way from being mastered, at least in national electronic health records and other similar programmes. This is a topic that has been with me throughout my career, and that I am very familiar with. Being involved in the actual implementation of new systems and services and having in-depth knowledge of real life issues is both an advantage and a disadvantage for PhD research. It is of course beneficial when you are very familiar with the domain, but the number and diversity of tangible issues to solve could be overwhelming. The key is to focus, to select a subset of issues to solve and to make a contribution in that area before moving on to the next one.

WHERE WILL YOUR CAREER TAKE YOU NEXT?

Check out the numerous job opportunities on the HiPEAC jobs portal: www.hipeac.net/jobs

If you're passionate about your career and would like to share it with the HiPEAC community, we'd love to hear from you. Email communication@hipeac.net with your story

Collaboration grants allow PhD students and junior post-doctoral researchers in the HiPEAC network to work jointly with a new research group. For further information, visit www.hipeac.net/mobility/collaborations.

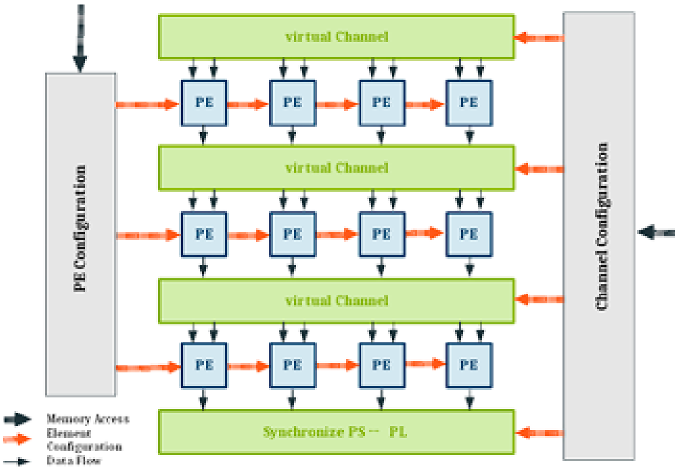
Creating the future through international exchange: HiPEAC collaboration grants

NAME: Amit Kulkarni
INSTITUTION: Ghent University - Belgium.
HOST INSTITUTION:
Ruhr University Bochum - Germany.
DATE OF COLLABORATION:
14/06/2016 - 30/06/2016
and 25/09/2016 - 07/12/2016

The research I did during my time at Ruhr University Bochum led to a paper being published at the 3rd International Workshop on Overlay Architectures for FPGAs at the FPGA 2017 conference. In the era of dark silicon, efficient computation with low power consumption is a must for any heterogeneous computing platform. HPC systems need ultra-efficient heterogeneous compute nodes. To reduce power and increase performance, such compute nodes will require reconfiguration as an intrinsic feature, so that specific HPC application features can be optimally accelerated at all times, even if they regularly change over time. Although modern embedded SoCs have CPUs and GPUs on the same die that can handle stringent performance requirements, they consume undesirable amounts of power, resulting in heat dissipation.

To tackle such problems, integrating a programmable logic with the SoC has resulted in efficient computation with low power consumption. This is because a CPU can leverage its complex computation to the custom hardware loaded onto the programmable logic. However, this comes at a price: the development costs incurred to generate suitable bistreams to configure the programmable logic.

Virtual Coarse Grained-Reconfigurable Arrays (VCGRA) come to the rescue in such situations. These arrays enable ease of programmability and result in low development costs. They specifically enable the ease of use in reconfigurable computing applications. The smaller cost of compilation and reduced reconfiguration overhead enables them to be attractive platforms for accelerating HPC applications such as image processing. The CGRAs are application-specific integrated circuits (ASIC) and therefore expensive to produce. Field Programmable Gate Arrays (FPGA) are comparatively cheap for low volume products but are not so easily programmable. We combine the best of both worlds by implementing a VCGRA on FPGA. VCGRAs are a tradeoff between FPGA with large routing overheads and ASICs. The paper presents a novel heterogeneous VCGRA called "Pixie" which is suitable for implementing high-performance image processing applications. The proposed VCGRA contains generic processing elements and virtual channels that are described using the hardware description language VHDL. Both elements have been optimized by using the parameterized configuration tool flow and result in a resource reduction of 24% for each processing element and 82% for each virtual channel respectively.



Spending time at another institution and working with new people broadened my research horizons and helped me make long-lasting contacts. I really recommend applying for a collaboration grant!

A. Kulkarni, A. Werner, F. Fricke, D. Stroobandt and M. Huebner: *Pixie: A heterogeneous Virtual Coarse-Grained Reconfigurable Array for high performance image processing applications* in 3rd International Workshop on Overlay Architectures for FPGAs (OLAF2017), Monterey, USA, 22/02/2017

The HiPEAC industrial mobility programme aims to give PhD students access to leading research teams in industry and to give such teams access to bright young minds. For more information, see www.hipeac.net/mobility/internships

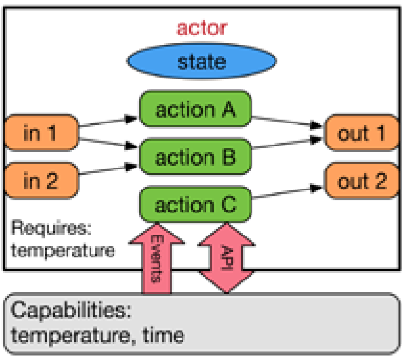
Training the next generation of experts: HiPEAC internships

NAME: Amardeep Mehta
RESEARCH CENTRE: Umeå University
HOST COMPANY: Ericsson Research, Sweden
DATES OF INTERNSHIP: September-December 2016

I am a PhD student at Umeå University, Sweden and, thanks to a HiPEAC internship, spent three months at Ericsson Research in Lund. My area of interest is resource management for mobile edge clouds and IoTs.

We are seeing a dramatic increase in small wireless devices connected to cloud services and expect there to be over 50 billion connected devices in the near future. Programming and managing them will be a major challenge. During the internship, I worked on development of a framework for IoT applications that can run in heterogeneous environments such as clouds, regional data centres, or servers at radio base stations, or inside embedded devices. A wide range of IoT applications, for example traffic safety applications for automated vehicles, could benefit from them. We worked on a development environment and management platform for IoT+cloud applications, Calvin, which is available as open source (https://github.com/EricssonResearch/calvin-base).

Calvin is a framework for application development, deployment and execution in heterogeneous environments, such as cloud, edge, and embedded or constrained devices. Inside Calvin, all the distributed resources would be viewed as one environment for an application. The framework provides multitenancy and simplifies development of IoT applications, which are represented using a dataflow of application components, Actors (internal structure of an actor is shown in figure 1), and their communication.

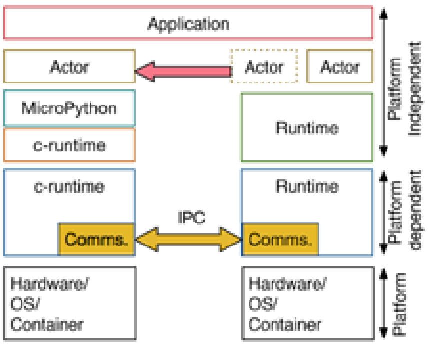


Anatomy of an actor. Tokens arriving at input ports or events can fire an action on the actor.

The Calvin distributed execution environment provides a distributed runtime, supporting an actor/data flow based programming paradigm, aimed at simplifying the development of IoT and cloud applications; in particular applications combining the two. Actor instances can be migrated between runtimes according to application specified conditions, allowing dynamic application distribution over runtimes.

The application's actors are implemented in Python for the Python-runtime and in C for the C-runtime. This work aims to support Python actors on the smaller C-runtime. The main task is to port a python virtual machine, e.g., MicroPython to an mbed platform and develop support libraries to interact with the mbed platform.

In this work, we implement Calvin Constrained (https://github.com/EricssonResearch/calvin-constrained), an extension to the Calvin framework to cover resource-constrained devices. The Calvin-Base and Constrained runtime stacks are shown in figure 2. Due to the limited memory and processing power of embedded devices, the constrained side of the framework can only support a limited subset of the Calvin features. The current implementation of Calvin Constrained supports actors implemented in C as well as Python, where the support for Python actors is enabled by using MicroPython as a statically allocated library. We thus enable the automatic management of state variables and enhance code re-usability.

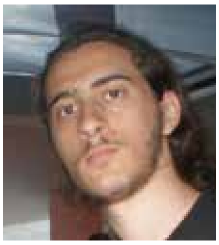


The Calvin runtime stacks. An actor being migrated from calvin-base to calvin-constrained runtime.

As would be expected, Python-coded actors demand more resources over C-coded ones. We show that the extra resources needed are manageable on current of-the-shelve micro-controller-equipped devices when using the Calvin framework.

Being one of the 800+ HiPEAC affiliated PhD students gives access to a vibrant research community spanning academia, large industry and smaller enterprises. It also provides the opportunity to take part in the mobility programme and to take part in networking and training events.

Three-minute thesis



TITLE: Java on Scalable Memory Architectures
AUTHOR: Foivos Zakkak
AFFILIATION: University of Crete and FORTH-ICS
COUNTRY: Greece
ADVISORS: Dr. Polyvios Pratikakis and
Prof. Angelos Bilas

As servers become more and more compact, it is expected that, within the near future, a single rack unit (1U) will feature hundreds of cores. These cores are expected to be grouped in coherent islands; groups of cores that will share a coherent memory. Coherent islands are also expected to communicate through efficient global interconnects but without hardware coherence.

In this thesis I study how high productivity languages can be run efficiently on such architectures. High productivity languages, like Java, are designed to abstract away the hardware details and allow developers to focus on the implementation of their algorithm, thus reducing the time to market of new products. At the same time, they offer increased security by automatically managing memory, and provide consistent behaviour across different platforms. To achieve these, high productivity languages rely on process virtual machines, like the Java virtual machine (JVM). Porting process virtual machines to the emerging architectures enables us to utilize the latter with legacy code, while allowing developers to exploit the scalability of them without the need to worry about the complexity of keeping data consistent across non-coherent memories. In this thesis I focus my work on the JVM since it is one of the most popular and widely studied process virtual machines on which tens of languages are being implemented, the most well-recognized being Java and Scala. JVM implementations need to adhere to the Java language specifications and the Java memory model (JMM). In this thesis I study JMM and present an extension of it that exposes explicit memory transfers between caches. This extension, called Java Distributed Memory Model (JDMM), aims to demystify the implementation of JMM on non-cache coherent architectures and, therefore, ease the process of showing that a JVM targeting a non-cache coherent architecture adheres to JMM. JDMM achieves this by providing explicit rules regarding the ordering of memory transfers in respect to other operations in a Java execution. I also argue that JDMM complies with the original JMM and allows the same optimizations.

I present a Java virtual machine design targeting non-cache coherent and partially coherent architectures. My design aims to minimize the number of memory transfers and messages exchanged while still adhering to the Java memory model. My design also takes advantage of partial coherence by sharing some structures across different cores on the same coherence island. Based on my design I implement a Java virtual machine and evaluate it on an emulator of a non-cache coherent architecture. The results show that my implementation scales up to 500 cores and its scalability is comparable to that of the HotSpotVM – the state-of-the-art Java virtual machine – running on a cache-coherent architecture.

Last but not least, I model my implementation in the operational semantics of a Java core calculus that I define for this purpose. I show that these operational semantics produce only well-formed executions according to the Java memory model. Since the operational semantics model my implementation, I argue that the latter also produces only well-formed executions, thus it adheres to the Java memory model.

European Research Council funding is one of the EU's tools to help top researchers carry out high-risk/high-reward research. Recently awarded an ERC Starting Grant, David Black-Schaffer, Associate Professor in the Department of Information Technology at Uppsala University tells us about his exciting new work.

Funding focus: ERC Starting Grants

I recently had the pleasure of chatting with HiPEAC Coordinator Koen De Bosschere at this year's conference in Stockholm. His energy and enthusiasm, combined with that of the HiPEAC staff team and Steering Committee, once again reminded me of how much the network has contributed to computer systems research in Europe, and, in particular, how much of a difference it has made for my own career.

My interactions with HiPEAC began seven years ago when I left Silicon Valley and moved to Sweden as a postdoc in computer architecture. In moving to Europe, I left behind my existing networks and found myself in a very different research environment. I volunteered to help write the 2011 HiPEAC Vision roadmap. This opportunity put me, a young researcher, in the same room as some of the world's leading experts in their field. Through these interactions, I learned the basics of the European funding and lobbying system and developed a better understanding of Europe's strengths (and weaknesses) in computer system research.

Over the years, at each conference and Computing Systems Week, I have been impressed by the smorgasbord (to use the Swedish term) of different activities, and by the levels of industrial participation. The strong academic and industrial connections that I have made through HiPEAC have been key in building multiple EU grant consortia and helping me to win an ERC Starting Grant late last year.

The grant for the project *Coordination and Composability: The Keys to Efficient Memory System Design* will fund PhD students and postdocs to work with me to build on breakthroughs in tracking and accessing data already acheived with colleagues at Uppsala. In all computing systems, whether small mobile devices or huge data centres, increases in performance must come from more power-efficient designs so that the benefits of enhanced performance are not outweighed by the negative impact of increased power consumption. My focus is on optimizing data movement energy, as the energy used to move data inside a computer processor is greater than that used to actually compute answers. Today's systems search through vast memory systems to find and retrieve data. If we can avoid searching by keeping track of where specific data is located, we can access it more quickly and more efficiently. However, while knowing where the data is allows us to access it more efficiently, the greater challenge is learning where to put it in the first place. The core of the ERC grant is to investigate how to integrate information from both the hardware and the software to enable smarter data placement and movement.

Photo: Knut and Alice Wallenberg Foundation

As computing power has become indispensable for everything from weather forecasting to medical monitoring, it is essential that we develop techniques to enable even faster computers in the future. If we can dramatically improve data movement efficiency, this ERC project will have a profound impact on a huge range of things that affect people's lives. It's going to be an exciting five years!

Read more about ERC funding at
https://erc.europa.eu/funding-and-grants/funding-schemes/starting-grants

Dates for your diary

European HPC Summit Week 2017
15-19 May 2017, Barcelona, Spain
https://exdci.eu/events/european-hpc-summit-week-2017

ISC High Performance 2017
18-22 June 2017, Frankfurt, Germany
www.isc-hpc.com

MEMSYS EU 2017: MEMSYS Europe International Symposium on Memory Systems
21-23 June 2017, Frankfurt, Germany
https://memsys.io/

13th International Summer School on Advanced Computer Architecture and Compilation for High-Performance and Embedded Systems (ACACES)
9-15 July 2017, Fiuggi, Italy
www.hipeac.net/acaces

10th International Symposium on High-Level Parallel Programming and Applications (HLPP 2017)
10-11 July 2017, Valladolid, Spain
https://hlpp2017.infor.uva.es

27th International Conference on Field-Programmable Logic and Applications (FPL 2017)
4-8 September 2017, Ghent, Belgium
www.fpl2017.org

26th International Conference on Parallel Architectures and Compilation Techniques (PACT)
9-13 September 2017, Portland, Oregon, USA
https://parasol.tamu.edu/pact17/

2017 ARM Research Summit
11-13 September 2017, Cambridge, UK
https://developer.arm.com/research/summit

International Conference Micro Energy 2017, Gubbio, Italy, 3-7 July 2017

http://www.microenergy2017.org
Registration open until 15 May 2017.

The ambition of this international conference is to bring together international scientists from academia, research centres and industry to discuss recent developments in the topic of micro energy and its use for powering sensing and communicating devices. We expect to welcome representatives from funding agencies including the European Commission's FET unit and the ONRG. Proceedings will be published as regular articles in a major science journal.

Conference topics include:

Session I - Micro energy harvesting
Energy transformation processes at micro and nano scales, mathematical models, harvesting efficiency, thermoelectric, photovoltaic, electrostatic, electrodynamic, piezoelectric, harvesting in biological systems, novel concepts in energy harvesting.

Session II - Micro energy dissipation
Noise and friction phenomena, fundamental limits in energy dissipation, Landauer bound, heat dissipation, thermodynamics of non-equilibrium systems, stochastic resonance and noise induced phenomena.

Session III - Micro energy storage
High performance batteries, super capacitors, micro-fuel cells, non-conventional storage systems.

Session IV - Micro energy use
Autonomous wireless sensors, zero-power computing, zero-power sensing, IoT, approximate computing, energy aware software, transient computing.

Co-located with the conference will be the NiPS Summer School 2017 – Energy Harvesting: models and applications, 30 June - 3 July
http://www.nipslab.org/summerschool