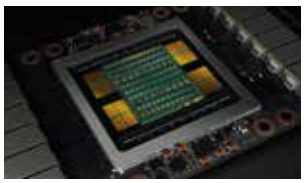
HPC 2017-18

A view into high-performance computing

From the publishers of
scientific computing world

Contents

SUBSCRIPTIONS: *HPC 2015-16* is published by Europa Science Ltd, which also publishes *Scientific Computing World*. Free registration is available to qualifying individuals (register online at www.scientific-computing.com). Subscriptions £180 a year for six issues to readers outside registration requirements; single issue £30. Orders to ESL, SCW Circulation, 4 Signet Court, Swann Road, Cambridge CB5 8LA, UK. Tel: +44 (0)1223 211170. Fax: +44 (0)1223 213385.

US copies: *Scientific Computing World* (ISSN 1356-7853/USPS No 018-753) is published bi-monthly for £100 per year by Europa Science Ltd, and distributed in the USA by DSW, 75 Aberdeen Rd, Emigsville PA 17318-0437. Periodicals postage paid at Emigsville PA. Postmaster: Send address corrections to *Scientific Computing World* PO Box 437, Emigsville, PA 17318-0437.

Cover: Stocklady/Shutterstock.com

A change of pace

While it is not yet time for the transition to exascale, much of the work is well underway and this is driving significant change in the HPC software and hardware ecosystem. There are other changes coming from the explosion of interest in AI and machine learning and also in the way that supercomputing performance is measured.

The impact of AI and machine learning research is the focus on page 10, as we explore the impact of this technology on the HPC market and the development of new processing technologies.

On Page 14, Robert Roe explores European and US plans for exascale. While both groups support approaches to co-design they have very different approaches to architecture and underlying hardware that will power the future exascale supercomputers.

The ways in which HPC performance can be measured are examined by Adrian Giordani on page 20. Giordani finds that the FLOPs centric methods used in the past are no longer suitable as memory and data access become more important to sustained performance on next generation HPC systems.

On page 26 Gemma Church explores application development and the challenges to scaling software to extreme levels on some of the world's largest supercomputers.

Also featured in the magazine we have an interview with Ansys on the use of Computer aided engineering (CAE) software in HPC that is featured on page 4. There is also an in-depth look at supercomputing in South Africa on page 6. This piece from Nox Moyake looks at the development of supercomputing infrastructure from its inception in 1994 to future projects such as assisting with the Square kilometre Array (SKA).

Tim Gillett
Managing editor

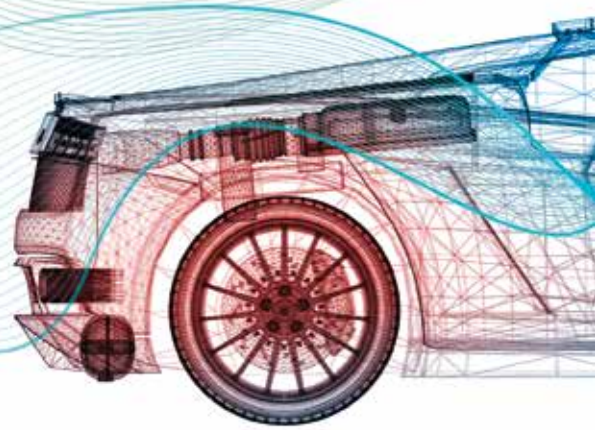
EDITORIAL AND ADMINISTRATIVE TEAM
Managing editor: Tim Gillett (editor.scw@europascience.com)
Technical writer: Robert Roe (editor.scw@europascience.com)
Circulation/readership enquiries: Pete Vine (subs@europascience.com)

ADVERTISING TEAM
Advertising sales manager: Mike Nelson (mike.nelson@europascience.com)
Tel: +44 (0)1223 275 465 Fax +44 (0)1223 213385
Advertising production: David Houghton (david.houghton@europascience.com)
Tel: +44 (0)1223 275 474 Fax +44 (0)1223 213385

CORPORATE TEAM
Managing Director: Warren Clark
Web: www.scientific-computing.com

CAE turns to HPC

Robert Roe interviews Ansys' Wim Slagter on the value of HPC technology within the CAE market



The CAE industry is no stranger to the use of HPC – but what was traditionally a tool employed by the largest engineering companies, such as the automotive and aerospace manufacturers, is now becoming much more ubiquitous as the barrier to HPC at entry level is lowered.

Although a large-scale on-premise cluster can still be out of the price range of many small or medium-sized engineering companies, they can still leverage the computational capabilities of HPC through cloud or appliance-based systems designed specifically for engineers or designers.

For large-scale users with access to their own HPC infrastructure, software companies are continually optimising their software to better fit today's highly parallel architectures through partnerships with HPC centres and hardware providers, such as Intel and Nvidia, to tune engineering software for the next generation of HPC architecture.

Wim Slagter, director, HPC and cloud alliances at ANSYS, commented on the importance of using HPC for engineering and design purposes. 'HPC is helping manufacturers cut costs and create new revenue streams because they can design completely new products they had not previously considered. Users can also produce more reliable products and reduce cost in the development cycle,' said Slagter.

He also stressed that many more engineers and designers have begun to see the potential value of using HPC in their workflows, as a survey published by Ansys found many users had issues that can be solved through the use of HPC technology. The 3,000 survey respondents were asked to name the biggest pressures on design activities, and about half of them stated reducing the time required to complete design cycles. Similarly around a quarter of respondents said producing more reliable products, which results in lower warranty-related costs, was a big concern. 'HPC is an enabler of innovation. HPC is a technology that addresses our customer's top challenges,' said Slagter.

Reducing the barrier to HPC

Today the benefits of using HPC are fairly well understood by the CAE community, but this does not make using and setting up HPC systems any cheaper. In the past, the cost of provisioning and managing an in-house cluster were prohibitively expensive for all but the largest companies, but now more engineers can gain access to HPC class computing capabilities through technologies such as the cloud.

HPC is an enabler of innovation. HPC is a technology that addresses our customer's top challenges

Particularly for smaller and medium-sized enterprises, Ansys recommend using cloud computing to reduce costs and facilitate access to HPC computation without the burden of managing a cluster.

'It is clear that hardware and software enhancements provide more value and have enabled HPC to deliver more value, but there are also challenges once companies start to deploy HPC. For example, in smaller enterprises specifying a cluster of provisioning, and managing a cluster, is not straightforward. They may lack even the basic IT staff in-house to set up and manage a cluster. These are real challenges, so we need to simplify HPC cluster deployments,' stated Slagter.

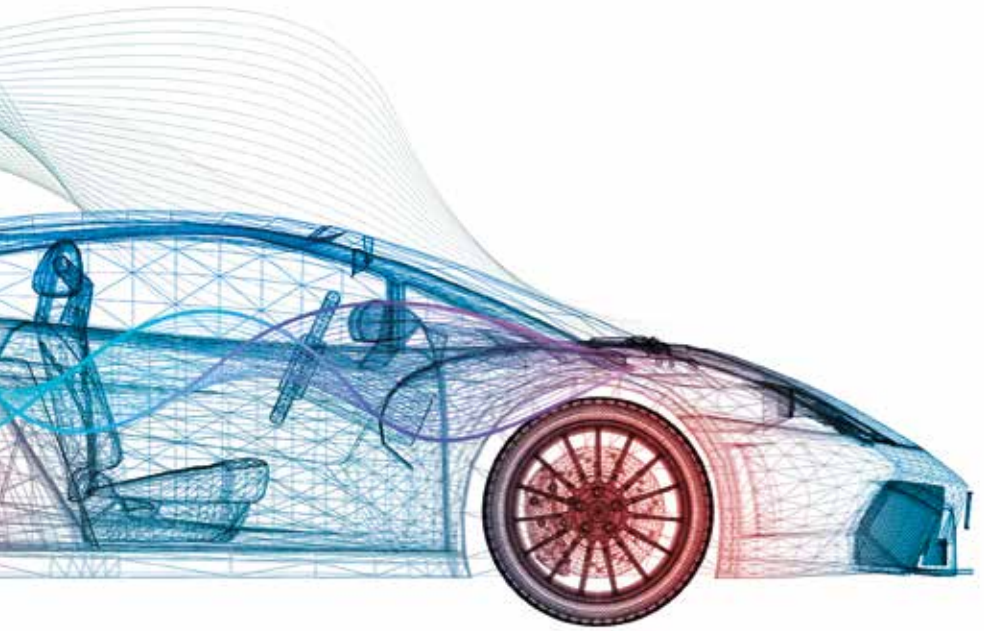
The main method that Ansys employs to democratise HPC to a wider audience is through partnerships with a number of companies such as cloud-hosting providers, HPC hardware manufacturers and supercomputing centres such as HLRS, in Stuttgart.

The cloud partners not only provide HPC services, but also the back-end infrastructure, to those customers who lack the in-house HPC or IT staff but still want the ability to increase computational resources quickly. In addition, Ansys also has partnerships with HPC partners that provide appliances, or pre-configured racks of computational hardware optimised and configured to run Ansys software.

'The theory of cloud is that those users can deploy these services only when they need it and ... they only have to pay for what they use. Smaller companies are interested in the flexibility, not only in terms of hardware deployment, but also software licences,' stated Slagter.

'Cloud is one aspect that is clearly helping to address these challenges, the other thing is that we have developed HPC appliances with specific partners – out-of-the-box, externally managed clusters. Again, this is aimed at those companies lacking infrastructure or IT staff, the right people to manage and provision a cluster,' Slagter said.

Through alliances and partnerships, Ansys is aiming to democratise the use of HPC for engineers in smaller firms by providing them with either cloud-based or appliance-based HPC solutions. For users without infrastructure or investment to support an in-house cluster, these options can provide HPC class computation at a



Italyan/Shutterstock.com

lower barrier to entry than traditional HPC.

'These partners work with system integrators to remotely manage HPC clusters for our customers, and those clusters are shipped to the customers but then everything is up and running, Ansys is pre-installed and the system is optimised to run Ansys workloads,' said Slagter.

'Gradually HPC resources are becoming more readily available to all engineers and hardware is becoming more affordable and also more powerful than ever before,' Slagter concluded.

Preparing for the future

Though the largest growth in the use of HPC comes from small and medium-sized companies, CAE software providers like Ansys are still putting a lot of time and effort into tuning their software for large-scale HPC simulations.

Ansys has developed partnerships with HPC centres such as HLRS in Stuttgart and King Abdullah University of Science & Technology (KAUST) in Saudi Arabia. These partnerships allow Ansys to scale its software to test the limits of engineering simulation on some of the largest supercomputers in the world.

For example, in July this year Ansys, Saudi Aramco and KAUST announced that they had set a new supercomputing milestone by scaling Ansys Fluent to nearly 200,000 processor cores. This supercomputing record represents a more than five-fold increase over the record set just three years ago when Fluent first reached the 36,000-core scaling milestone.

The calculations were run on the Shaheen II, a Cray XC40 supercomputer, hosted at the KAUST Supercomputing Core Lab (KSL).

'We need to deliver better HPC performance and capability. These customers are pushing the envelope; they come up with ever more challenging models that require more computational resources and better performance' said Slagter. 'We are constantly working on the optimisation of our software. We need to provide more parallelism throughout the entire process. We need to continue working with supercomputing partners.'

By partnering with these HPC centres, Ansys can test the limits of its software and implement optimisations to take advantage of the highly parallel nature of today's leadership class HPC systems. Without these partnerships the onus would be on Ansys alone to deliver these improvements, which are quickly becoming a task too complex for a single organisation.

'The computer industry has delivered enormous increases in computing speed at lower cost. As an ISV, we have also made significant improvements regarding parallel performance, robustness, and scalability' said Slagter. 'The same is true for GPUs as well. Any of our Ansys products, from structural mechanics to fluid dynamics to electromagnetics, are now taking advantage of GPUs,' concluded Slagter.

Expanding the capabilities of design engineers

Another fast-growing area of CAE is the use of simulation tools by designers that can make quick changes to design concepts and see the changes in product performance without having to wait for verification from engineers. Slagter explained that Ansys has been working for a number of years on its design tool, Ansys Discovery Live. The aim of this product is to give simulation tools to designers who to accelerate prototyping of new products.

'We have recently launched Ansys Discovery Live a new product that is completely built from scratch on GPUs. It is empowered by the thousands of cores available in a GPU and it provides real-time simulation results that are primarily aimed at designers,' said Slagter. 'This was a multi-year development programme where people have drag and drop capabilities, but with a built-in simulation that solves in an instant, so people do not have to wait to get their results. For example, if a user changes the CAD, you can immediately see the influence on the flow or stresses of the structure.'

While this tool will not directly be used on HPC resources, it does feed into the same design cycle and could reduce the amount of simulation required, as small changes can be made early in the design process without requiring large simulations to verify their performance.

'It is not providing the full high-fidelity results that the analysts require, but that is not necessary for designers who want very quick results,' said Slagter.

This new tool can feed into HPC workflows, as designers may work quickly on a workstation and then pass data to engineering teams that can run a full analytical simulation on a HPC cluster. Ultimately, modern CAE requires ISVs to serve several user communities with different requirements from the large aerospace or automotive companies, to small engineering firms and design teams that have different uses for engineering software.

HPC is helping manufacturers both cut costs and create new revenue streams, because they can design completely new products they had not previously considered

'It may sound difficult if you looking for a single solution to meet the requirements of these different types of customers but that is not what you should do, or at least that is not our HPC strategy,' said Slagter. 'We are not working on a single solution that will meet all requirements, because that is impossible. For analysts, we will continue to improve parallelisation further, optimise software, extend the parallelism throughout the workflow, work with supercomputer centres to profile and benchmark the software at an extreme scale.

'It is impossible to provide just a single solution, you need an ecosystem. It is becoming more and more important to grow and come up with the right solutions by working with the right partners,' he concluded.

Firing up a continent

Nox Moyake describes the process of entrenching and developing HPC in South Africa

The year 1994 marked a much-anticipated turning point in South Africa's history, a time when all could have equal access to economic opportunities and when all could participate fully in the development of the country's economy.

From 1994 to 2014, the GDP annual growth rate averaged 3.08 per cent, reaching an all-time high of 7.1 per cent in the fourth quarter of 2006. In that year, the South African government, through the Department of Science and Technology, decided to invest significantly in research in order to drive the much-needed social and economic development, with a plan to increase expenditure on research to one per cent of GDP by 2010.

The idea was to increase the country's capacity to create and disperse knowledge, investing in new instruments required to ensure that the country not only had requisite capacity to generate new levels of knowledge, but also to cement South Africa's position as an attractive destination for science and technology projects.

In 2007, the Centre for High Performance Computing (CHPC) was launched, followed by the South African National Research Network and the Data Intensive Research Initiative of South Africa; together the three are referred to as the National Integrated Cyber Infrastructure System (NICIS).

Flagship projects (2007 to 2014)

Once the CHPC was established, the next question was how does the government develop and encourage use of HPC in South Africa. A few strategies were used to entice interest in the field.

First was the creation of flagship project grants. The centre advertised and granted two-year funding to 11 projects across a number of research domains. Each project aims to solve a major science or socio-economic challenge in South Africa and on the continent in general, using the centre's computational resources.

The CHPC team

The use of HPC resources provided training and exposure to parallel computing and set the ground for development of additional HPC capacity in South Africa. In addition, the flagship projects advanced research activities in these specific domains and accelerated the building of HPC capacity. University postgraduate students were central to these developments.

These 11 flagship projects covered the following topics:

- Computational space physics and astrophysics;
- Large-scale simulations of energy storage materials;
- Regional coupled ocean-atmosphere modelling;
- A South African high performance multi-physics computational fluid dynamics solver;
- Electromagnetic computer simulation for the Meerkat and SKA;

Tang Yan Song/Shutterstock.com

Once the CHPC was established, the next question was how does the government develop and encourage use of HPC in South Africa

- Modelling HIV-1 evolution;
- Modern South African astronomy and cosmology: confronting the simulates and the observed universe;
- Monte Carlo simulations of technological tools for quantum information processing and communication;
- Nuclear collision and data grid for the physics community;
- Biomechanics of myocardinal infarction and the development of novel therapies; and
- Computational research initiatives in imaging and remote sensing.

Human capital development

Second to the HPC role-out phase was the creation of internships, studentships and year-round training courses to CHPC users, to ensure that they make optimum use of HPC resources.

The CHPC Winter School, usually attended by 100 students, is a favourite among computer science and computer engineering students across South African universities. The school is an introduction to parallel programming and covers topics including: an introduction to HPC hardware, systems and techniques; multicore programming with OpenMP; using message passing (MPI) on cluster supercomputers; high performance co-processors; and rapid prototyping for HPC with Python.

Occasionally, the CHPC will invite international pioneers and innovative developers in the HPC field to present workshops and lectures in South Africa. These lectures are open to all researchers and HPC users to attend.

In 2012, the centre introduced the Student Cluster Competition as a component of its Winter School. Following a national call, students shortlisted are introduced to parallel programming and taken through various aspects of the competition. They are then divided into groups and given an assignment that determines which groups will proceed to the national competition, which takes place at the annual CHPC National Conference in December.

The winning team proceeds to the international competition at the International Supercomputing Conference in Germany. South African teams have performed very competitively in the international round, ➤

A million reasons to back SKA project

The Council for Scientific and Industrial Relations (CSIR) has, through its Centre for High Performance Computing (CHPC), donated a portion of its supercomputer to Ghana, where it will be used for human capital development purposes in preparation for the data processing requirements associated with the Square Kilometre Array's (SKA) project.

The donation of the supercomputer components is part of the launch of the first African SKA satellite outside of South Africa that recently took place in Accra, Ghana.

The single rack of compute nodes, with storage and network, was part of the CHPC's decommissioned Tsessebe supercomputer and is now dedicated to training and capacity development at the Ghana Earth Observatory and will form the initial part of the processing of the data emanating from the satellite.

The technology transfer forms part of the first phase of the CHPC's SKA Readiness project involving repurposing and transfer of HPC systems that are out of production to create HPC imprints in the eight SKA Africa partner countries: Namibia, Zambia, Botswana, Mauritius, Mozambique, Kenya, Madagascar and Ghana.

The Department of International Relations and Cooperation (DIRCO), through the African Renaissance Fund, funds the SKA Readiness project. SKA Africa partner countries, with the exception of Kenya and Mozambique, who are later in the year, have received HPC systems from three supercomputers: Ranger from the Texas Advanced Computing Center (TACC) in the United States, Cambridge supercomputer from the University of Cambridge in the UK and Tsessebe from the CSIR in South Africa.

The second phase of the SKA Readiness project will start in 2018 through the donation and distribution of Stampede, a next-generation supercomputer from TACC, which will expose the partner countries to different types of HPC technologies.

The conference serves as a showpiece of the work done by users of the centre in the previous year

The CHPC has around 1,000 users in academia and industry

➤ having won the overall prize in 2013, 2014 and 2016 and taken second place in 2015 and 2017.

The annual CHPC Conference

The third leg of entrenching an HPC footprint in South Africa involves the now established and growing CHPC National Conference. The conference serves as a showpiece of the work carried out by users of the centre in the previous year and brings together renowned international and domestic speakers in the field, academia and industry.

Delegates attended a number of specialised workshops and tutorials, birds of a feather and special interest group sessions. Over the last few years, the conference has held a special forum of the Southern African Development Community, that is devoted to the establishment of an HPC centre that will be focused on supporting research carrying a shared social and economic impact on the region.

The future

The CHPC currently has about 1,000 users; most are in academia and others in industry.

System engineers and research scientists provide regular support to users. The centre supports research from across a number of domains and participates in a number of grand international projects such as the Cern and Square Kilometre Array (SKA) projects.

The CHPC contributes as a tier-two facility to Cern. The SKA project is fast developing momentum and the centre has established a SKA Readiness Project that aims to source, and distribute recently retired HPC equipment as training machinery for the eight African partner (Zambia, Botswana, Mauritius, Ghana, Mozambique, Kenya, Namibia, Madagascar) countries that will be hosting SKA together with South Africa.

The first round of distribution of HPC machinery and introductory training is almost complete, with Kenya and Mozambique planned for later this year. ●

Noxolo Moyake is a research communications specialist at the CHPC

Students showing the way

Team South Africa took second prize in the prestigious International Student Cluster Competition held at the International Supercomputing Conference in Frankfurt, Germany. The results were announced following three days of hard work by 12 international teams.

The Centre for High Performance Computing (CHPC) trains Computer Science and Engineering students from across South African universities in its annual winter school held every July and selects some for the national Student Cluster Competition that takes place during the CHPC's National Conference every December. Winners of the national competition are entered into the annual International Student Cluster Competition that takes place in Germany.

South Africa is always represented by a new team of six undergraduate students who travel to Germany to build a small cluster of their own design on the International Supercomputing Conference exhibition floor, and race to demonstrate the greatest performance across a series of benchmarks and applications.

Students receive a unique opportunity to learn, experience and demonstrate how high performance computing influences our world and day-to-day learning.

Tsinghua University (China) took the overall prize based on their Linpack score and their performance in the 'mystery challenge'.

Team South Africa was represented by Mishka Mohamed, Kyle Jordaan, Tyrone de Ruiters and Liam Doult, all from University of Western Cape, as well as Phillip Goosen and Lydia de Lange from Stellenbosch University.

Team South Africa

The team members were honoured guests of Minister Naledi Pandor during the Department of Science and Technology's Budget Vote in parliament this year, where she wished them well for the international competition.

South Africa has won the international competition three times before and taken second position once. Asked on what the magic formula is, David Macleod, CHPC Engineer and the team's advisor, simply said: 'We have good sponsors and we come prepared.'

This year's team was sponsored by DellEMC for hardware equipment and Mellanox for network equipment.

AI drives new computing technologies

Robert Roe looks at advances in AI computing technology

As the market for artificial intelligence matures, it is helping to drive accelerated growth in computing technologies to support highly parallel workloads, artificial intelligence (AI) and machine learning (ML). This new paradigm in computing is opening up the benefits of GPU or accelerated computing to a broader audience – far beyond the traditional users of supercomputing.

The growth in AI and machine learning has been dramatic. In April this year, market research firm IDC predicted that western European revenues for cognitive and AI systems would reach $1.5 billion in 2017.

IDC predicts this rise will continue in the coming years as the company forecasts a growth rate of 42.5 per cent through 2020 when revenues will exceed $4.3 billion.

Much of this growth comes from comes from three key industries which were early adopters of AI and cognitive systems – banking, retail, and discrete manufacturing, although the IDC report does note that cross-industry applications have the largest share across all industries.

The report states that by 2020 these industries – including cross-industry applications – will account for almost half of all IT spending on cognitive and artificial intelligence systems.

'IDC is seeing huge interest in cognitive applications and AI across Europe right now, from different industry sectors, healthcare, and government,' said Philip Carnelley, research director for Enterprise Software at IDC Europe, and leader of IDC's European AI Practice.

'Although only a minority of European organisations have deployed AI solutions today, a large majority are either planning to deploy or evaluating its potential. They are looking at use cases with clear ROI, such as predictive maintenance, fraud prevention, customer service, and sales recommendation,' Carnelley added.

The report also notes that from a technology perspective, the most significant area of spending in western Europe in 2017 will be cognitive applications at approximately $516 million. This includes 'cognitively-enabled' process and industry applications that automatically learn, discover, and make recommendations or predictions.

'Cognitive Computing is coming, and we expect it to embed itself across all industries. However, early adopters are those tightly regulated industries that need robust decision support: finance, specifically banking and securities investment services, is one of these early adopters,' said Mike Glennon, associate vice president for customer insights and analysis at IDC.

'However, the cost savings to be found in automating decision support in a structured environment, together with the enhanced ability to identify previously hidden aspects of behaviour, ensure the distribution and services and public sectors embrace cognitive computing and artificial intelligence systems – where it can offer the dual benefits of lowering cost, and growing new business. We also expect strong growth in adoption in manufacturing in w estern Europe, at the core of industry across the region,' Glennon added.

Loihi – Intel's first self-learning neuromorphic computer chip

This interest is reflected in a large growth in the number of applications that are finding their way into use in both industry, academia. AI applications are being run on appliances and small servers all the way up to the largest supercomputers. This is true of even the leadership class HPC facilities which are pursuing research into machine learning and AI research applications.

Developing AI technology

Nvidia and Intel are the primary hardware providers for AI and ML applications, but several other companies have released processing technologies aimed at AI and ML applications.

Although only a minority of European organisations have deployed AI solutions today, a large majority are either planning to deploy or evaluating its potential

These range from Nvidia GPU's to Intel coprocessors or even FPGA technology.

While the DL and ML algorithms have been around for some time, it is the arrival of incredibly parallel computing technologies such as GPUs that have enabled the explosion in ML/DL applications. The falling price of these GPUs and other accelerators alongside increasing computational performance that is well suited to highly parallel applications have enabled accelerator technology to flourish in this new market.

This year has seen Intel and Nvidia release new hardware, tools and investment aimed at capturing the AI market.

Nvidia initially launched is DGX-1 appliance in 2016. This was updated in 2017 to deliver higher performance by adding the Volta GPUs to the DGX-1 system. The first the first examples of the Volta-based DGX-1 were delivered to supercomputing sites in September 2017.

But it is not just hardware that is the ground for stiff competition. The development of ➤

GIGABYTE™

INTENSIFY YOUR SCALE

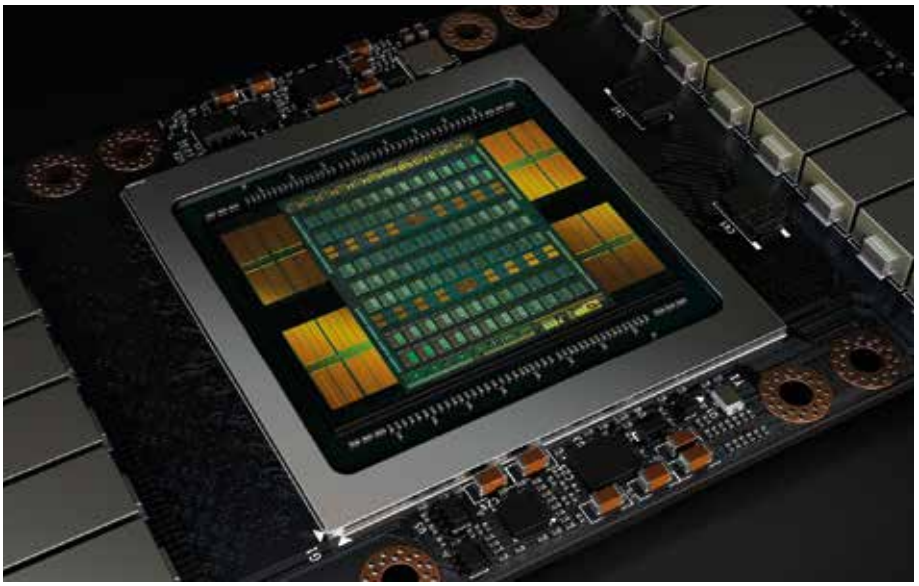
Optimized GPU Density and Cooling

G291 series

> Intel® Xeon® Processor Scalable Family
> Supports 8 x double slot GPU cards
> 6CH RDIMM/LRDIMM DDR4, 24 x DIMMs
> 2 x 10Gb/s BASE-T & 2 x 1Gb/s LAN ports
> 1 x Dedicated management port
> 8 x 2.5" hot-swappable HDD/SSD bays
> 80 PLUS Platinum 2200W redundant PSU

Support Intel® Xeon® processor
Intel Inside®. New Possibilities Outside.

Configure yours at:
> b2b.gigabyte.com



Nvidia Tesla V100 GPU

algorithms and faster application frameworks is another highly competitive area.

At Nvidia's GTC event in China, held at the end of September 2017, Nvidia CEO, Jensen Huang, delivered a keynote that focused more on software development and application performance than new GPU hardware.

There was one hardware tease of a new processor to aid in developing autonomous vehicles, known as 'Xavier', but this will not be unveiled until 2018.

Most of the keynote was focused on the development of new software NVIDIA AI computing platform, TensorRT 3, the AI Cities platform, and Nvidia DRIVE PX.

TensorRT 3 is Nvidia's AI inferencing software, which has been designed to boost the performance and slash the cost of inferencing from the cloud to edge devices. This new software could be used in applications from self-driving cars to robots.

During the keynote, Huang claimed that using TensorRT 3 and Nvidia's latest GPUs can process up to 5,700 images a second.

Nvidia's DRIVE platform has been designed as a platform for future autonomous vehicles and Nvidia has stated that users can adopt some or the entire platform. The systems is based on a scalable architecture that is available in a variety of configurations. These range from a single mobile processor, to a multi-chip configuration with two mobile processors and two discrete GPUs, or a combination of multiple DRIVE PX systems used in parallel.

The new processor Xavier will provide an added boost to this platform in 2018. Early access partners will be able to receive platform containing the Xavier SoC early next year, while general availability is expected at the end of 2018.

'Our vision is to enable every researcher everywhere to enable AI for the goodness of mankind,' said Nvidia CEO Jensen Huang. 'We believe we now have the fundamental pillars in place to invent the next era of artificial intelligence, the era of autonomous machines.'

Intel has also launched products aimed at deep learning such as its first neuromorphic, self-learning computer chip codenamed Loihi. Intel released information about the chip as part of a blog post at the end of September authored by Dr Michael Mayberry, corporate vice president and managing director of Intel Labs.

The blog post notes that the work on neuromorphic computing builds on decades of research that started with CalTech professor,

We believe AI is in its infancy and more architectures and methods – like Loihi – will continue emerging that raise the bar for AI

Carver Mead, who was known for his work in semiconductor design. The combination of chip expertise, physics and biology yielded some revolutionary ideas: comparing machines with the human brain.

As part of this research Intel developed a self-learning neuromorphic chip – codenamed Loihi – that mimics how the brain functions by learning to operate based on various modes of feedback from the environment. Intel claims that this is an extremely energy-efficient chip, which uses the data to learn and make inferences, gets smarter over time and does not need to be trained in the traditional way. 'We believe AI is in its infancy and more architectures and methods – like Loihi – will continue emerging that raise the bar for AI.

'Neuromorphic computing draws inspiration from our current understanding of the brain's architecture and its associated computations,' wrote Mayberry.

'The brain's neural networks relay information with pulses or spikes, modulate the synaptic strengths or weight of the interconnections based on timing of these spikes, and store these changes locally at the interconnections. Intelligent behaviours emerge from the cooperative and competitive interactions between multiple regions within the brain's neural networks and its environment,' Mayberry added.

The Blog post from Intel also notes that ML models such as deep learning have made tremendous recent advancements – particularly in areas such as training neural networks to perform pattern or image recognition. 'However, unless their training sets have specifically accounted for a particular element, situation or circumstance, these ML systems do not generalise well,' stated Mayberry.

The Intel blog post noted that the self-learning capabilities prototyped by this new chip from Intel have the potential to improve automotive and industrial applications as well as personal robotic. Any application that would benefit from autonomous operation and continuous learning – such as recognising the movement of a car or bike – is a potential application for the Loihi chip.

In the first half of 2018, the Loihi test chip will be shared with leading university and research institutions with a focus on advancing AI.

In addition to new chips, Intel has also committed $1 billion to AI research in recent months. Brian Matthew Krzanich, chief executive officer of Intel, stated: 'We are deeply committed to unlocking the promise of AI: conducting research on neuromorphic computing, exploring new architectures and learning paradigms. We have also invested in startups like Mighty AI, Data Robot and Lumiata through our Intel Capital portfolio and have invested more than $1 billion in companies that are helping to advance artificial intelligence.'

'AI solutions require a wide range of power and performance to meet application needs. To support the sheer breadth of future AI workloads, businesses will need unmatched flexibility and infrastructure optimisation so that both highly specialised and general purpose AI functions can run alongside other critical business workloads,' Krzanich concluded.

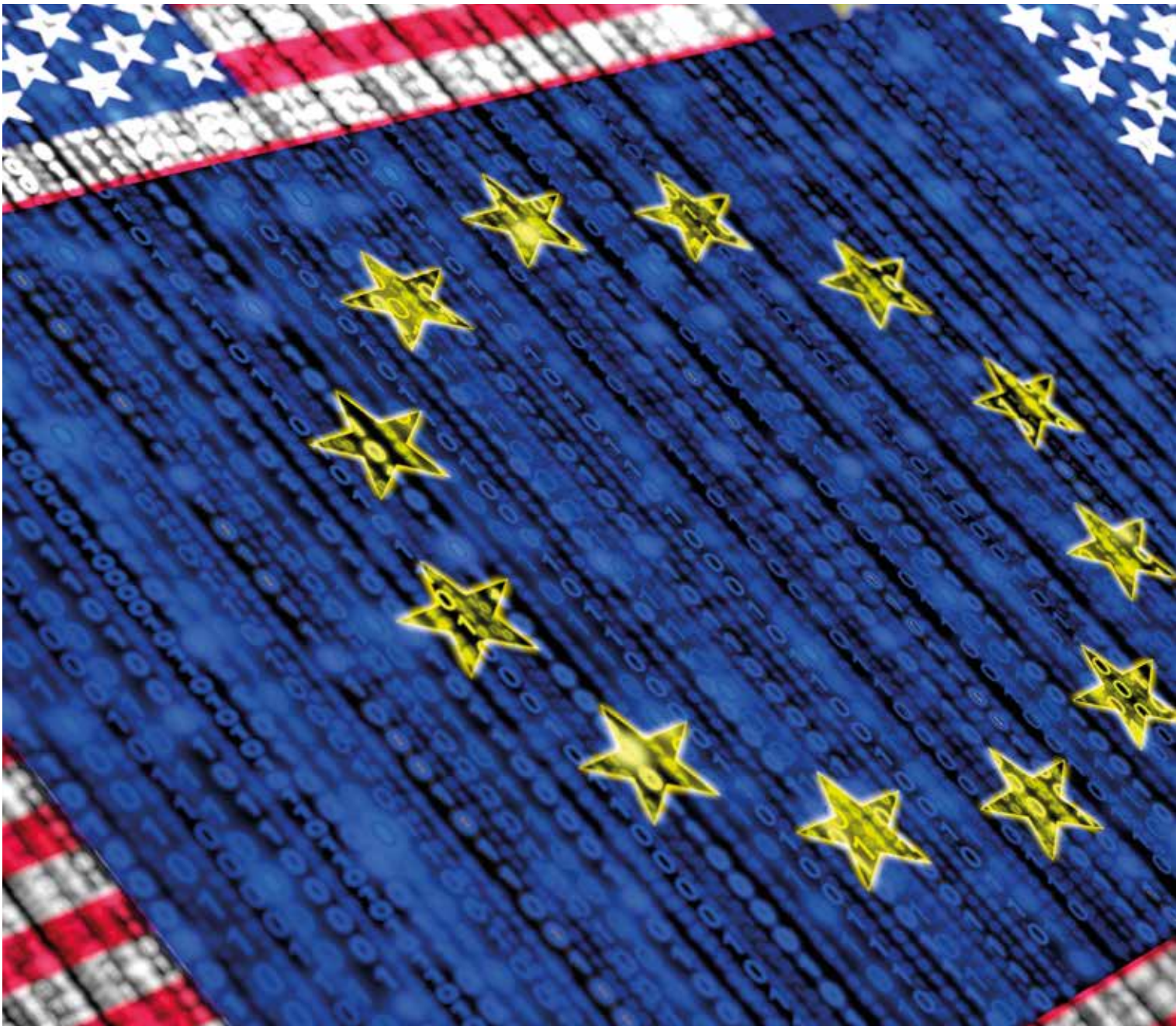
PBS Works is a Cloud-friendly Secure Workload Management Solution for the Entire Engineering Lifecycle.

PBS WORKS 2018 — HPC LIKE NEVER BEFORE

- New intuitive user interface – desktop, web, mobile
- Create appliances anywhere – on premises, cloud, bare metal
- Access and manage HPC with uncompromised security
- Optimize consumption and control cost
- Automate cloud bursting to handle peak workload

pbsworks.com

Altair



BeeBright/Shutterstock.com

Continental ambitions

Robert Roe looks at European and North American efforts to establish exascale computing programmes

As the supercomputing industry approaches the exascale, Europe and the US have developed projects to deliver this huge jump in computing performance. Using novel technologies and building partnerships between hardware vendors, academic and research centres, the two groups are hoping to deliver exascale with very different computing architectures.

The US aims to continue with Intel CPUs with accelerators from Nvidia or Intel, while Europe aims to create an ARM-FPGA hybrid computing architecture. While they are very different technologies, both groups face similar hurdles in delivering energy-efficiency and parallelism on an unprecedented scale.

The exascale challenge

The pursuit of exascale is important for more than just the headlines and international bragging rights and accolades that come with hitting this milestone. Exascale promises to unlock the new level of computational performance that will provide the horsepower for future scientific discoveries not possible today.

Exascale class supercomputers will allow scientists and engineers to investigate problems at new levels of granularity and accuracy and will enable scientific breakthroughs.

However, reaching exascale is no easy task. While flops is no realistic measure of exascale computing performance, the most basic milestone for an exascale class supercomputer would be a billion billion calculations per second. To reach this level of performance requires systems at a scale that has never been seen before and that requires complex innovation of both hardware and software. All of this must be achieved within a given power limit of approximately 20 to 30 megawatts, which makes this all the more difficult as such a system would far exceed the efficiency of the most powerful supercomputers today.

There are several ways of approaching these challenges but with the scale of the problem increasing for each generation of computing development, most groups are choosing to tackle the challenge by creating their own computing ecosystem. This allows them to pool resources, investment and expertise with partners that can help to develop the launch pad for exascale computing.

Examples of this can be seen across several regions. Chinese HPC experts have opted for a somewhat home-grown approach but this still includes several component manufacturers, HPC centres, and organisations working together to create the necessary hardware and software to compete with other regions.

We hope to contribute to things beyond the delivery of just a couple of exascale systems in the United States

The American model of subsidising HPC development through investment in large supercomputing contracts from the DOE and other federal organisations is well understood but even this is taxed by the challenges of achieving exascale computing by the proposed date from the DOE of 2021.

The US has developed what it calls the Exascale Computing Project, a collaborative effort of two US Department of Energy organisations – the Office of Science (DOE-SC) and the National Nuclear Security Administration (NNSA).

'The Exascale Computing Project offers a rare opportunity to advance all elements of the HPC ecosystem in unison. Co-design and integration of hardware, software, and applications, a strategic imperative of the ECP, is essential to deploying exascale class systems that will meet the future requirements of the scientific communities these systems will serve,' said Paul Messina, former ECP director and senior strategic advisor at the Argonne Leadership Computing Facility (ALCF).

Messina commented that, alongside the two DOE organisations, there are also six vendor partners working on various aspects of the computing architecture as part of the PathForward exascale initiative: IBM, Nvidia, Intel, Cray, HP and AMD. He also noted that there are 22 institutions and 39 universities currently involved with some aspect of the research and development.

There are many moving parts for such a collaboration that uses co-design, but Messina stressed that there is more to this project than just reaching the exascale milestone. He hopes this work will leave a lasting legacy that provides benefits beyond just the use of the DOE exascale systems. 'I envision that some of the results might not pay off in 2020-2021, but it might in 2022 or 2024. Sometimes it takes a little bit longer but we hope to contribute to things beyond the delivery of just a couple of exascale systems in the United States,' said Messina.

He explained that the idea for meaningful collaboration extends across the entire project. One example off this is an industry council that has been setup with 18 commercial companies, of which 15 are end-user companies such as GM, GE, United technologies and FedEx.

'We established that industry council to make sure that we understand their needs. They all feel that they will eventually need exascale computing. We are working with them to understand their needs, what is realistic for them to do, and we will create a software stack that will meet their requirements.

This is a constant theme throughout the DOE ECP project. They do not just want to reach the milestone of exascale computing but develop the building blocks for other users and to develop meaningful application performance.

European plans

The Europeans have opted for a similar co-design approach banding several commercial entities and organisations together for a co-design approach. European plans for exascale are funded through the European Union, which has set out significant investment in IT infrastructure through the European Commission's FP7 programme – part of the Horizon 2020 initiative. Horizon 2020 is the biggest EU Research and Innovation programme ever, with nearly €80 billion of funding available over seven years (2014 to 2020).

The EuroEXA project is funded through this Horizon2020 programme and builds on previous European high-performance computing projects and partnerships, bringing together the focus of European industrial SMEs. Originally the informal name for a group of H2020 research projects ExaNeSt, EcoScale and ExaNoDe, EuroEXA hopes to coalesce all of these research projects into a single coherent exascale project.

The project is opting for co-design using a number of European developed technologies ➤

➤ and partners including HPC centres, research organisations and hardware manufacturers that can help to create a European exascale class supercomputer to rival competition in the US and Asia.

The €20m investment over a 42-month period is part of a total €50m investment made by the EC across the EuroEXA group of projects supporting research, innovation and action across applications, system software, hardware, networking, storage, liquid cooling and data centre technologies.

Funded under H2020-EU.1.2.2. FET Proactive (FETHPC-2016-01) the consortium partners provide a range of key applications from across climate/weather, physics/energy and life science/bioinformatics. The project aims to develop an ARM Cortex technology processing system with Xilinx Ultrascale+ FPGA acceleration at Pflop level by approximately 2020. This could then lead to exascale procurement in 2022/23 and commercialised versions of the technology available around the same time.

John Goodacre, professor of computer architectures at the University of Manchester, said: 'To deliver the demands of next generation computing and exascale HPC, it is not possible to simply optimise the components of the existing platform. In EuroEXA, we have taken a holistic approach to break-down the inefficiencies of the historic abstractions and bring significant innovation and co-design across the entire computing stack.'

Peter Hopton, founder of Iceotope and dissemination lead for EuroEXA, said: 'This is a world-class programme that aims to increase EU computing capabilities by 100 times, the EuroEXA project is truly an exceptional collection of EU engineering excellence in this field.'

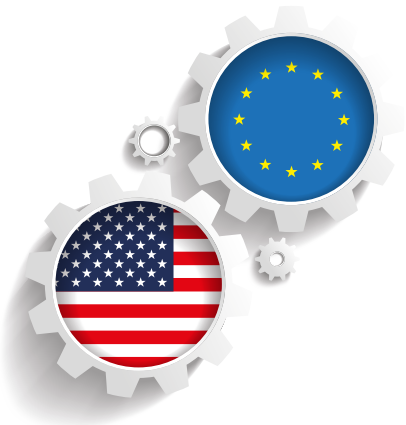
The EuroEXA project is certainly ambitious as it hopes to bring technologies from ARM and Xilinx together with Maxeler and memory technology from ZeroPoint Technologies to produce a new computing architecture for an exascale system.

Even if this process is successful, it will require a lot of application development to create the tools necessary to deliver sustained exascale performance. Alongside the scalable vector extensions (SVE) Arm has helped to provide Allinea debugging tools and the project has partnered with several research centres that bring their own large-scale application codes for development.

Arm is providing Allinea tools as a bridge between the hardware architecture and applications, evaluating application performance and pinpointing steps to maximise efficiency.

The tool selection for the EuroEXA Program enables Arm to collaborate with project partners and understand their challenges in application development and preparation on the novel EuroEXA platform. 'New capabilities are often a direct result of collaborating with leading research efforts such as EuroEXA. Arm is quickly applying learnings from EuroEXA, and similar efforts, into future state-of-the art Allinea tools designed to help reach the most efficient levels of exascale compute,' said David Lecomber, Arm's senior director of HPC tools.

Another project partner, Maxeler, hopes to port its Dataflow programming model to support the relevant components of the EuroEXA platform. Ultimately, this should allow the applications targeted to be brought on to the heterogeneous EuroEXA system platform.



Style-photography/Shutterstock.com

In EuroEXA, we have taken a holistic approach to break down the inefficiencies of the historic abstractions and bring significant innovation and co-design across the entire computing stack

'Joining EuroEXA is exciting for us because it allows us to bring our long-established Dataflow technology into Europe's latest effort towards achieving Exascale performance,' said Georgi Gaydadjiev, director of Maxeler Research Labs.

The Dataflow computing model will enable application developers to utilise the reconfigurable accelerators in a high-level environment. It also addresses the practical challenges of data movement when combined with other technologies, such as memory.

Memory specialist ZeroPoint Technologies uses novel compression approaches to store and transfer memory data more efficiently. The technology is based on more than 15 years of research from Chalmers University of Technology, Goteborg, Sweden. The firm has developed memory systems that use an IP-block that compresses and decompresses data in memory, so that typically three times more data can be stored in memory and transferred in each memory request. Their aim is to deliver added value and competitiveness concerning cost, power consumption and performance of exascale systems, plus added value in power consumption and memory performance by adapting its intellectual property blocks and integrating them in the computing chips. Additionally, the company will be responsible for the memory interface for the EuroEXA project.

Per Stenstrom, co-founder and chief scientist from ZeroPoint Technologies, said: 'We are very excited at having the opportunity to join the EuroEXA project and demonstrate the added values our unique technology can offer to Exa-Scale systems.'

Iceotope will also be aiming to provide a boost to power consumption and efficiency with its liquid cooling technology, which should allow denser computing racks and more efficient cooling technology. However, the company was not just selected for its capabilities in liquid cooling, but also for IP within power delivery, I/O connections, infrastructure management and data centre infrastructure.

Peter Hopton, founder of Iceotope said: 'It's a privilege to be selected as part of this programme, with this investment in development, our technology will now enable the biggest computers of the future, as well as the cloud computing environments and edge computing of today and the near future.'

Leaving a legacy

Creating an exascale supercomputer is a huge achievement, but if the accompanying software stack, programming models and even core design do not see widespread use by the wider HPC community, then all the effort that has been exerted will be lost in the transition to following generations.

While one measure of success is the system capable of a huge number of calculations, it is clear that there is an opportunity to develop better standards and approaches to computing that can provide benefits over the next five to 10 years.

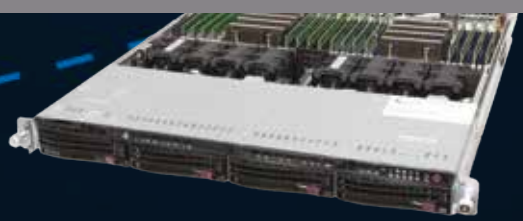
One example of the negative side effects that have come from several generations of computational evolution along the von-Neumann architecture is the memory bandwidth bottleneck that we see in today's most computationally intensive HPC systems.

The problem is not just confined to the performance from a lack of data transfer, as the energy ratio between control and arithmetic I/O ➤

"LEADING-EDGE DISRUPTIVE TECHNOLOGY INTEGRATORS"

NVIDIA® VOLTA™

AMD EPYC™



INTEL® XEON® SCALABLE FAMILY

SEE THESE SOLUTIONS AT:

SC17 - STAND 1775

ISC18 - STAND H-700

AMD

NVIDIA ELITE SOLUTION PROVIDER

intel Technology Provider Platinum 2017

TAILOR-MADE SOLUTIONS

From the specification to a unique design, Boston has the knowledge and expertise to tailor the ideal solution for you.

BOSTON HPC LABS

Remotely test and benchmark your technologies before investing to ensure our HPC solutions meet you exact requirements.

LEADING-EDGE TECHNOLOGY

With engineers dedicated to assessing new technologies in Boston's R&D Labs facility, Boston is able to offer new technlogy first.

- STORAGE
- WORKSTATIONS
- SERVERS
- NETWORKING
- SOLUTIONS
- CLOUD SERVICES

BOSTON
Servers | Storage | Solutions

SUPERMICRO

WEB: WWW.BOSTON.CO.UK
EMAIL: SALES@BOSTON.CO.UK
PHONE: +44 (0) 1727 876 100

and the scalability through I/O communication are both concerns for future system designers.

This issue was described by John Goodacre, professor of Computer Architectures, Advanced Processor Technologies Group at the University of Manchester and in a presentation as part of a workshop at the Infrastructure for the European Network for Earth and System Modelling (IS-ENES) – another FP7 funded project. In the presentation, Goodacre noted that while the von-Neumann model was fundamental to many of today's systems, it did have certain limitations as we approach exascale computing.

While the memory bandwidth problem is significant, Goodacre lists several approaches to overcoming this challenge by increasing processor efficiency. These range from SIMD or vector machines, DSP, GPGPU, hardware accelerators or FPGAs as possible solutions.

The EuroEXA project chooses FPGA acceleration and lays down several steps to creating a new computing architecture leading to the creation of commercial systems in approximately 2023. The EuroEXA project, which builds on several earlier projects, Euroserver, ExaNODE, ExaNeSt, ECOSCALE aims to lay the building blocks for a European exascale system by creating a new computing architecture based on an ARM/FPGA hybrid processor, dubbed the 'ICT-42 EuroProcessor.'

This project will evolve through several phases, from designing the processor and node architecture, through to the interconnect technology and eventually OS, runtimes, programming models and application development.

However, designers must be careful not to create something which cannot be easily adopted by other users and this must be a clear thought from the outset, as the project is already considering potential commercialisation of these technologies once the initial exascale systems have been deployed.

Across the pond, the DOE is dealing with similar challenges. 'We want to build a software stack that will support a broad set of applications and that will have a life beyond the end of the exascale project,' said Messina.

'In other words, it would serve as the foundation, for some time after, for many applications to be able to take advantage of exascale. One of the broad goals is to come up with a software stack that can be used on medium class HPC systems, as well as exascale. People can adopt this software stack in order to make it easier to transition to higher and higher levels of performance,' Messina added.

While the EU and US exascale plans revolve around particular architectures or software stacks

Imageflow/Shutterstock.com

Portability is an important target or goal for applications, especially some performance portability, because there will continue to be more than one architecture in the lifetime of an application

that are being created specifically to bridge the hurdles needed to obtain exascale application performance, their efforts will be felt by HPC users across the globe as the technology, architecture and software design provide blueprints for other HPC users who wish to pursue their own exascale journey. Investment from US government funding US companies will surely benefit a worldwide community of HPC users, and this is also true of European efforts funded by European countries as part of the European Commission's efforts to accelerate computing efforts on the other side of the Atlantic.

'In recent years, if one is running HPC on a medium-sized cluster the software environment tends to be somewhat different from what you have for the leading-edge systems. That has been an obstacle that people have surmounted but we would like to lower the effort required to do that,' said Messina.

The hope from the US investment programme is to provide an ecosystem which can then support further development in the future. This is done through widespread adoption of tools and architectures that help to drive expertise and knowledge around a given technology, language or programming model. This development should provide a trickle-down effect to users who are not targeting exascale but can still take advantage of the same tools and methods for HPC.

Another aspect that Messina was keen to highlight was that another secondary goal would be to streamline some of the approaches to programming that are dealing with similar issues.

'Applications have, for very good reasons, adopted different approaches and algorithms and it is difficult for a community to converge to a single solution,' commented Messina. He gave one example of parallel I/O, explaining that different groups had their own approaches but it would be beneficial to converge if such a process could support all the required applications.

'I/O is just one of the examples, another would be common runtime API that could support tools like performance measurement tools, visualisation tools, or even compilers. 'In the long run, there would be substantial benefits for the high performance computing community if there were fewer choices, so long as they properly support the applications,' commented Messina.

Performance portability is another key aspect to the project, as it will help to ensure value in the work that has been done by the ECP partners. As Messina notes, 'portability is an important target or goal for applications, especially some performance portability, because there will continue to be more than one architecture in the lifetime of an application. There will probably be several different architectures that a given application will run on.'

Beyond the goals of exascale lie a new set of challenges and milestones that must be surmounted by future development. To lay the foundation for future innovation requires today's computer scientists to think carefully about the legacy they leave for the next generation. The industry needs real cooperation and teamwork to meet the requirements of exascale but also to lay the groundwork for future technologies that can successfully use the tools that are created today.

The path to an energy-efficient exascale supercomputer

Adrian Giordani explores the methods used to accurately measure the performance of supercomputers

The International Supercomputing Conference (ISC17) closed on 22 June 2017 in Frankfurt, Germany, ending an eventful week from a growing community energised on how to advance the sector. Highlights included the latest product announcements, updates on machine learning, GPU accelerators, the conclusion of the sixth iteration of the computing cluster competition and – of course – the Top500 supercomputer list.

China's Sunway TaihuLight climbed to the top of the supercomputing list, achieving 93 quadrillion operations per second (93 Pflops) on approximately 15 megawatts of power. Even with Sunway TaihuLight's achievement, it is clear that we may be living in a post-Moore's law world where processor performance growth is slowing overall.

The next question when you compare the Sunway TaihuLight's Top500 list achievement to the goal of exascale computing – a system that will run one billion billion calculations per second – is how will the first exascale supercomputer keep energy consumption low while sustaining far more computational power?

The Top500 list contestants all compete using the Linpack benchmark, which uses the IEEE Standard 754 floating point arithmetic, measured in floating operations per second (flops) that solves a dense system of linear equations, most of which are dense matrix–matrix multiplications. Currently, some mavericks within the supercomputing community are striving to shift the focus to methods that produce consistent reproducible results, while also looking at whole applications to give a better idea of real-world performance.

Diversifying benchmarks for real-world performance

While a flops-based approach keeps pushing managers of supercomputing centres onwards in one dimension of complexity, other lists have emerged to compliment the Top500. For example, the Green500 list looks at Linpack flops-per-watt for energy efficiency.

At this year's ISC17 another complementary benchmark to the High Performance Linpack (HPL), known as the High Performance Conjugate Gradients (HPCG) benchmark, entered its seventh year.

HPCG placed the Sunway TaihuLight system in fifth place on its list of 110 entries – and Japan's Riken/Fujitsu K Computer at number one.

To date HPCG, which measure performance that is more representative of how today's scientific calculations perform, has been run on many large-scale supercomputing systems in Europe, Japan and the US.

Last year, in a peer-reviewed paper published in the journal *International Journal of High Performance Computing Applications*, Jack Dongarra, director of the University of Tennessee's Innovative Computing Laboratory, US, who has been involved in Linpack's development since 1993, along with two other colleagues, analysed the performance of HPCG in comparison to HPL.

The team concluded that their preliminary tests show that HPCG exhibits performance levels that are far below the levels seen by HPL, one of the main reasons being the so-called memory wall. Still, HPCG scales equally well when compared with HPL.

'HPCG, in addition to HPL, is a good benchmark and should be run on every new system in addition to running HPL. HPCG shows a different characteristic of the system that is benchmarked and should be a good addition,' said Robert Henschel, chair of the Standard Performance Evaluation Corporation's High-Performance Group (SPEC/HPG).

'I disagree with the statement that HPCG measures "real application performance",' said Horst Simon, deputy laboratory director and chief research officer at Lawrence Berkeley National Laboratory, US, and co-editor of the biannual TOP500 list. 'Since HPCG is mostly determined by this fundamental speed of the machine, it will correlate with HPL in the foreseeable future.'

Then there is the High Performance

Kitdesign/Shutterstock.com

Computer Challenge (HPCC) benchmark, sponsored by the US's DOE, the National Science Foundation and DARPA. It comprises seven tests such as the HLP, Fast Fourier Transform, STREAM and communication bandwidth and latency. The HPCC benchmark looks at computational performance as well as memory-access patterns.

'Benchmarks like HPCC, HPL and HPCG are important and allow users to draw conclusions about the absolute best performance of a system, but they may not be representative of real-world workloads,' said Henschel.

SPEC/HPG benchmarks offer complementary metrics to HPCG that enable behaviour analysis of whole applications in a more in-depth view of real-world performance. SPEC/HPG benchmarks usually focus on parts of a large system or single nodes.

SPEC/HPG maintains three benchmarks called: SPEC MPI2007, SPEC OMP2012 and SPEC ACCEL. Each of the benchmarks addresses different ways that scientific applications can be parallelised. SPEC/HPG members include AMD, HPE, IBM, Intel, Nvidia and Oracle, as well as a host of associate universities.

Benchmarks like HPCC, HPL and HPCG are important and allow users to draw conclusions about the absolute best performance of a system, but they may not be representative of real-world workloads

SPEC ACCEL contains codes that make use of accelerators, such as GPUs or specific processors, to speed up performance of scientific applications in fields such as medicine, astrophysics, molecular dynamics, weather and fluid dynamics.

'All SPEC/HPG benchmarks are designed to measure the performance of real applications, not just a benchmark kernel or an algorithm. From our point of view, this gives users a more realistic picture of how applications are going to perform on one system compared to another, or how much of an advertised performance boost of a new processor is actually visible in application performance,' said Herschel.

In comparison, HPCC is a benchmark that uses very low-level benchmark kernels such as HPL and STREAM tests.

'Those benchmarks measure only small parts of what a scientific application would normally need to do during its runtime on a supercomputer. In contrast, SPEC ACCEL contains complete real-world applications, measuring the full execution cycle of a scientific application,' said Henschel.

With computational performance, one of the most important things to know is how much time is spent accessing each level of memory; including registers, cache, DRAM, ➤

➤ mass storage and all levels in between. With this level of detail you can forecast performance by understanding the scale of the problem.

'Yet, that is usually the first thing benchmarks discard: they fix the size of the problem!' said John Gustafson, currently a visiting scientist at the A*STAR (Agency for Science, Technology and Research) in Singapore. Gustafson is an accomplished expert on supercomputer systems and creator of Gustafson's law in computer engineering.

According to Gustafson, the original Linpack was a fixed size benchmark that did not scale. By persuading Jack Dongarra to switch to a 'weak scaling' model – for which Gustafson's law applies instead of Amdahl's law – this helped the TOP500 list to endure for 25 years. Since the 1980s, the benchmark's definition has become more goal-oriented, amenable to parallel methods and less subject to cheating.

From floats to a posit-based approach

This year a peer-reviewed research paper titled *Beating Floating Point at its Own Game: Posit Arithmetic* was published in the journal *Supercomputing Frontiers and Innovations*. The paper's authors believe this data type has the potential to revolutionise the supercomputing community's approach and attitudes to performance, both of applications and the systems they are run on.

'Benchmarks should always be goal-based, but usually they are activity-based,' said Gustafson. 'Which is where you get silly metrics like 'floating point operations per second' that do not correlate well with getting a useful answer in the smallest amount of time,' said Gustafson.

In the paper, Gustafson and his co-author, Isaac Yonemoto from the Interplanetary Robot and Electric Brain Company in California, US, conclude that the 'posit' data type can act as a direct drop-in replacement for IEEE Standard 754 floats, yet have higher accuracy, larger dynamic range and better closure – without any need to reduce transistor size and cost.

In short, posits could soon prove floats obsolete and further steer the community away from one-dimensional benchmarks altogether.

In another experiment by Gustafson, when comparing posit-based arithmetic with floats, the posit approach again came up on top. Gustafson ran random data through a standard Fast Fourier Transform (FFT) algorithm. Then he inverse transformed it and compared it with the original signal.

'For a 1024-point FFT and a 12-bit analogue-to-digital convertor data I was able to get back the original signal, exactly, every bit, using only a 21-bit posit representation,' says Gustafson. 'That's something even 32-bit floats cannot do. I can preserve 100 per cent of the measurement information with fewer bits.'

Maverick thinkers

In April of this year, Paul Messina, director of the US Exascale Computing Project (ECP), presented a wide-ranging review of ECP's evolving plans for the delivery of the first exascale machine – which has now moved its launch to 2021 – at the HPC User Forum in Santa Fe, US.

Messina stated that from the very start the exascale project has steered clear of flops and Linpack as the best measure of success. This trend has only grown stronger with attention focused on defining success as performance on useful applications and the ability to tackle problems that are intractable on today's petaflops machines.

For well over 25 years, leading voices in the supercomputing community have urged users to measure systems by their capabilities to solve real problems. A few of these individuals include Messina, Gustafson and Horst Simon.

I believe that, just as with the first petascale systems, the first exascale systems will run applications and problem sizes that would normally not run on smaller systems

Back in 2014, Simon said in an interview that calling a system exa-anything was a bad idea, because it becomes a bad brand, associated with buying big machines for a few national labs; therefore, if exaflops are not achieved, this will likely be seen as a failure, no matter how much great science can be done on the systems being developed.

'My views on naming anything 'exa' are still the same,' said Simon. 'However, what has changed is that we now have a well-defined exascale computing project in the US. This project includes a significant number of exascale applications – in the order of 24.'

These applications range from cosmology to genomics and materials science.

Europe is also working towards a supercomputing ecosystem effort known

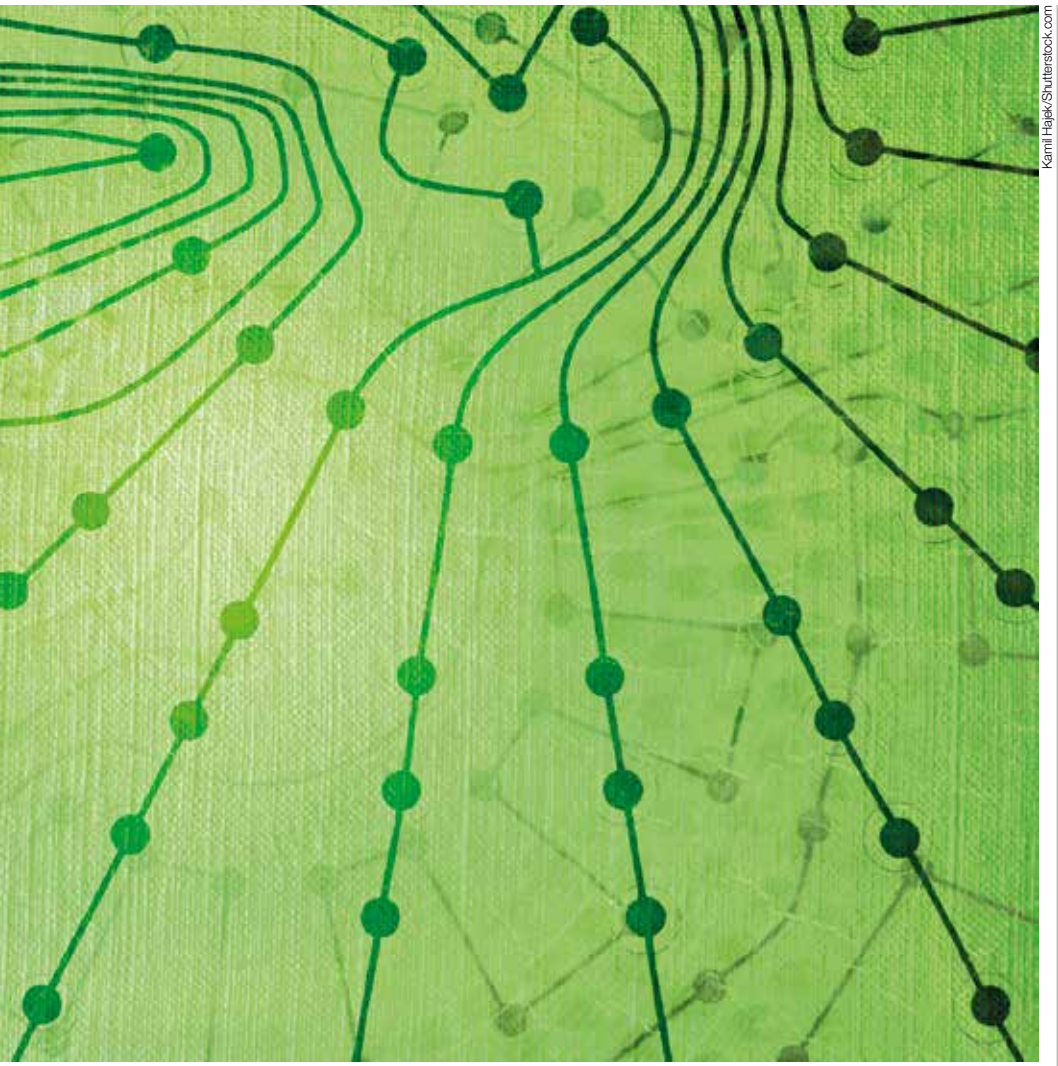
as EuroHPC. In July 2017, on a European Commission blog, an established voice in this field said the European supercomputing community faces a weak spot in technologies, such as the development and commercialisation of domestic computer or CMOS chips and processor technologies.

This view came from Wolfgang Marquardt, scientific director and chairman of the board of directors of Forschungszentrum Jülich (Jülich Research Centre), in Germany, home to a supercomputing centre and one of Europe's largest research centres.

'A single nation cannot make significant progress in this endeavour, and we need to work together to advance in this field,' said Marquardt.

For Europe's EuroHPC, or any exascale, effort the success factors that should take precedence have shifted from traditional metrics, according to Marquardt.

'In my opinion, there are more appropriate ways to discuss the power of supercomputers, rather than rigid benchmarking lists: energy-efficiency, scalability and adaptability for a


Kamil Hajak/Shutterstock.com

variety of different frontier applications have become more important parameters than the obvious efficiency metrics, such as the number of cores, or the peak performance on some standard test suite.'

Leading thinkers in the US tend to agree on this approach.

'What is most important is that they all have to measure progress on a metric that makes sense for the application. So there is no push to get some artificial results that may not make sense scientifically. Instead the application developers need to demonstrate a factor of a hundred improvement over 2017's state-of-the-art in their chosen metric,' said Simon.

But is there an approach or benchmark that will be able to meet the demands of future applications and their users, or will something new be needed for the 'exa-age' of supercomputing?

'There is no overall best benchmark. These benchmarks are not suited to actually make a purchase decision for a machine. You will need to first define what you want to accomplish with a supercomputer,' said Simon.

These questions could be whether the benchmark is for a single application or for a very diverse workload; or, is it for a small number of users or for many users – these are critical factors for managers of supercomputing systems. In many cases, one benchmark will be suitable for one problem, but not another.

'For an actual procurement, the Sustained Petascale Performance (SPP) is much more useful, but it needs to be tailored to the individual requirements,' said Simon.

The Sustained Petascale Performance metric tool is used on the Blue Waters system at the University of Illinois at Urbana-Champaign, US. It helps its users get a more detailed understanding of each application's performance, workload and the overall continual performance of the entire system.

To posit-operation-per-second processors and beyond

In an email interview, Dongarra emphasised that to get an idea of the best benchmark to work towards, the DOE's current exascale goal is a good guide: an application that can run 50 times better than on today's 20 Pflop systems, running under load at between 20 to 30 megawatts of power and with less than one fault per week.

According to Robert Henschel, the SPEC ACCEL single node benchmark is applicable to the future exascale system, but it would only evaluate a small section of a system – so a very comprehensive analysis, but not at scale. Posits could reset the approach of the community; first, it has to overcome natural scepticism and the current manufactured computer processors.

Simon said that in his opinion the main obstacle is that the global hardware computing industry is now a $350 billion enterprise. It will be very hard to move that market towards innovative concepts like posit-based architectures, even if they are probably better.

Benchmarks should always be goal-based, but usually they are activity-based

Marquardt also said, 'The concept of posits is interesting but – at least at this point in time – is not expected to be of relevance for the next generation of supercomputers.'

Perhaps a breakthrough innovation will not come from one of the larger players, such as big chip manufacturers Intel or Nvidia, but a smaller start-up on the bleeding edge of technology.

Start-up semiconductor company REX Computing, based in the US, is developing a novel low-power processor chip called 'Neo'. This 28 nanometre-sized technology is touted by its creators to have up to 25 times energy efficiency improvement for supercomputers and digital signal processing over conventional CMOS chips.

REX Computing's initial test chip was produced last year and uses a custom-designed IEEE compliant floating point unit that is being sampled by early customers.

The team at REX are also experimenting with posits and see great potential in them. A processor variant using posits is in production under contract with A*STAR.

'We are a very small team, but are punching outside our weight class,' said Thomas Sohmers, CEO of REX Computing. 'For a start-up like REX, we want to cater to early adopters and customers that have the absolute highest requirements for their systems, which ➤

➤ is a niche too small to base major product decisions on for the big guys.'

With $2 million in funding they have already developed a new processor architecture, created silicon chip units and the initial software. In comparison, the typical cost for a 28-nanometre node process for traditional semiconductor companies ranges from between $30 to $250 million.

'A number of start-ups in the past decade raised tens of millions of dollars without ever producing a working chip,' said Sohmers. 'While those big companies may look at posits as a risky proposition, we see opportunity in being the first to offer innovative solutions to those early adopters.'

To date, the Neo general-purpose float-based processor is achieving 128 single precision and 64 double precision gigaflops per watt in tests. 'In comparison, that's more than double what you see at the top of the latest energy-efficient Green500 supercomputer list,' said Gustafson.

According to Sohmers, the latest Intel 'Knights Landing' Xeon Phi chip, made on a 14 nanometre process, has a theoretical peak performance of about 10 double precision Gflops per watt.

The current theoretical peak performance on their Neo chip is better. On an older 28 nanometre process, the Neo performs 32 double precision Gflops per watt, 26 Gflops per watt for a DGEMM benchmark (designed to measure the sustained floating-point computational rates of a single node) and 25 Gflops per watt for a FFT – a very communication intensive function. This will widen advantages over x86-type processors for 64-bit operations.

'We are showing three to 25 times better energy efficiency while we have a huge (us on 28 nanometres versus Intel's 14 nanometres) process technology disadvantage. For our production chip, which is roughly on a par with Intel's 14 nanometre, our numbers would just about double,' said Sohmers.

Based on these conservative estimates, a 32-bit REX-type design processor, based on posits, instead of a 64-bit processor based on floats, could achieve 60 billion real-world operations per second per watt.

Scaled up, this is the energy-efficient exascale computer that Sohmers and Gustafson envision. 'With a 20-megawatt power budget, yes, you're definitely beyond exascale at that point,' said Gustafson.

However, as mentioned, some in the community have their doubts, which is based on the current dominance of the larger players in the semiconductor industry. 'That may have been possible in the late 1980s when the industry moved to the IEEE floating point standard, but at that time, the market was much smaller and floating point arithmetic was indeed faulty and counterproductive for software development,' said Simon.

But, according to Gustafson, IEEE 754 floats are obsolete: it's just that the world doesn't know it yet. 'The small companies have early-mover advantage and the big companies have amazing resources to apply, but are always conservative. That's where the revolutionary fun is – and always has been. Very much like the disruption of parallel computing in the 1980s,' said Gustafson.

Smarnok/Shutterstock.com

The established companies won't lift a finger until they see their market share threatened by an upstart; and sometimes, not even then

Large and established chip manufacturers are still squeezing as much out of CMOS technology by investing in Fin FET (fin field-effect) and seven-nanometre scale transistors.

'The established companies won't lift a finger until they see their market share threatened by an upstart; and sometimes, not even then. With the belief that these initially risky ideas will gain more mainstream adoption once they are proven out as being viable... it would only be at that time that the rest of the industry would be practically forced to change.'

The innovation that REX Computing is making is by taking a lot of unnecessarily and complex logic out of their hardware design for their processor through the use of 'scratchpads'. They have written unique code that gives exact latency guarantees for all operations and memory access, allowing a compiler to be able to handle all of the memory management just within software, not hardware.

'While it sounds simple and obvious, the actual algorithms and compilation techniques we are using are very unique, and up until us doing it, many said it would be impossible,' said Sohmers.

In regards to REX Computing's IEEE-float-compliant Neo processor, they have had evaluation units in use by early customers since May 2017. And they are planning on sampling 16 nanometre-scale chip units in spring of 2018, with larger volume availability in the last quarter of 2018.

Sohmers said, 'Depending on our results with the posit project, we expect to have evaluation units available for a variant of our processor replacing the IEEE float unit available in spring 2018.'

Based on their current posit-based simulations, they are very confident that they will they exceed 60 Gflops per watt with their first production chip next year, which has one potential 'peta-scale' supercomputer installation in the pipeline for 2019. This shows the potential for a reasonably priced exascale supercomputer by 2020 using Neo chips.

Back in the late 1990s, quips Gustafson, the goal was a 'tera-ops' machine, staying clear of flops and Linpack. But it wasn't long before the supercomputing community said, 'Yeah, yeah, sure. So does it get a Tflop on Linpack?'

This cycle repeated itself in the 2000s with the peta-scale computing goal: Pflops became the flavour of the decade. Exascale will probably reach the same fate with the first questions being about Eflops. 'It's just too much fun to plot trend lines for a benchmark that is older than dirt,' said Gustafson, who is still undaunted of the potential for posits.

'I did a quick scan of my email and found 40 entities working on making posit arithmetic real at the hardware level. Most are start-up companies, but also national laboratories, universities and companies like IBM, Intel, Qualcomm, Samsung, Google, Microsoft and Nvidia.

'Mostly, the feedback I've gotten: When can I have it? I want it now! Frankly, I'd be surprised if people are still using IEEE 754 floating point in 2027.'

In the supercomputing chip race, perhaps the surprise will come from a smaller country or start-up that will develop paradigm-shifting solutions first, and drag the race to a new path. Nonetheless, the past has shown that for any big idea, it takes time for change; the clock is ticking.

Subscribe for free*

scientific computing world

Do you compute?

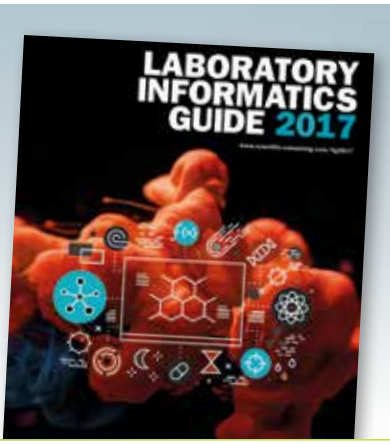
The only global publication for scientists and engineers using computing and software in their daily work

Do you subscribe? Register for free now!

scientific-computing.com/subscribe

*Registration required

Also published by Europa Science



The Best HPC Applications of 2017

Gemma Church explores applications at the forefront of HPC research

The term 'best' could be used to describe the biggest HPC application, the boldest, or the most innovative. This article will highlight those HPC applications where real progress has been made in the past 12 months and where challenges in this sector have been addressed.

These challenges are not necessarily application specific, according to Pak Lui, co-chair of the HPC|Works Special Interest Group at the HPC Advisory Council and Principal Architect at Huawei Technologies, who said: 'In my opinion, the HPC community faces generic challenges that are shared by different applications. The performance of the typical groups of HPC applications are all bounded by the compute, network communication, and storage I/O infrastructure or hardware. The performance of the HPC applications depends on the hardware development cycles. But the HPC community is always coming up with tools and libraries to make such systems accessible for all applications to use.'

In other words, cross-application collaboration is vital for the HPC sector to progress. Such collaboration is prevalent at the Gauss Centre for Supercomputing, which combines three major German supercomputing centres, the High Performance Computing Center Stuttgart (HLRS), Jülich Supercomputing Centre, and Leibniz Supercomputing Centre (LRZ).

All three centres work with researchers across the science and engineering spectrum, however each one does have some specialisations. Jülich is renowned for its fundamental research, physics work, and neuroscience work, and environmental sciences; LRZ strongly supports projects in geoscience, life sciences, and astrophysics; and the HLRS has a very strong focus on scientific engineering and industrial applications.

Eric Gedenk, a science writer for the GCS, said: 'Currently, we have about 300 research projects that use over 100 different applications. Many of them use community codes, but many of our research projects have in-house codes to suit their particular research needs.'

One of LRZ's highlights this year is the earthquake/tsunami research done by Dr Michael Bader and his team, which is a finalist for a best paper award at SC17. It presents a high-resolution simulation of the 2004 Sumatra-Andaman

earthquake, including non-linear frictional failure on a megathrust-splay fault system.

Last fall, the HLRS also helped Ansys scale the Ansys Fluent CFD tool to more than 172,000 computer cores on the HLRS supercomputer Hazel Hen, a Cray XC40 system, making it one of the fastest industrial applications ever run.

Dr Wim Slagter, director of HPC and Cloud Alliances at Ansys, explained: 'To overcome large-scale simulation challenges, we established a multi-year program with HLRS and worked on improving the performance and scalability of our CFD solvers. Apart from improving the transient LES (Large Eddy Simulation) solver, we focused on a variety of aspects, including the enhancement of our advanced multiphysics load balancing method, the optimisation of file read/write operations and the improvement of fault tolerance for large-scale runs.

'We started by further optimising the linear CFD solver, predominantly the AMG (Algebraic MultiGrid) portion of it. We optimised the partition and load balance capabilities to enable the good balancing at very high core count. To enhance the simulation throughput, we developed better reordering algorithms for improved memory usage, and we enhanced the transient combustion convergence speed. We also improved the parallel IO capability and developed better data compression strategies. Because of these very high core counts – up to 172,000 CPU cores – parallel solver robustness was obviously crucial here; we wanted to have a robust solver that can be "fired and forgotten", Slagter added.

Another major research finding out of Jülich in the last year, which came from the Center for Theoretical Chemistry at Ruhr University Bochum, uncovered the previously unknown complexities in the relationship between sulphur atoms' bindings. These bindings link long molecules together to form proteins and rubber. For example, if you stretch a rubber band again and again, the sulphur bridges will break and the rubber becomes brittle.

This rubber band example is familiar to most people, but a correct interpretation of the experimental data was lacking. However, this research found that the splitting of these bonds between two sulphur atoms in a solvent is more complicated than first assumed.

'Depending on how hard one pulls, the sulphur bridge splits with different reaction mechanisms,' Dr Dominik Marx, professor and director of the Center for Theoretical Chemistry at the Ruhr University Bochum, explained. In essence, the simulations revealed that more force cannot be translated one to one into a faster reaction. Up to a certain force, the reaction rate increases in proportion to the force. If this threshold is exceeded, greater mechanical forces speed up the reaction to a much lesser extent.

The High Performance Computing Center in Stuttgart, part of the Gauss Centre for Supercomputing

Previous simulation and modelling methods drastically simplified the effects of the surrounding solvent in order to reduce the processing power required. Work done at the Cluster of Excellence RESOLV in Bochum had already uncovered the key role the solvent plays in chemical reactions.

But correctly incorporating the role of the surrounding solvent requires immense computing effort. This computational power was made available to Marx and his international team by a special 'Large Scale Project' granted by the GCS on the Jülich Blue Gene/Q platform Juqueen. Without which, the detailed simulations to interpret the experimental data on sulphur atom bindings would not have been possible.

To overcome large-scale simulation challenges, we established a multi-year program with HLRS and worked on improving the performance and scalability of our CFD solvers

There has also been some fascinating extreme scaling work on the Juqueen supercomputer (based at the Jülich Supercomputing Centre) in the last 12 months at the High-Q Club. Dr Dirk Brömmel, senior scientist from High-Q Club, explained: 'The club was set up to showcase codes that scale to the 450,000+ cores or (ideally) 1.8 million threads. This helps to identify where possible bottlenecks are in future systems and if a solution is found on Juqueen, many of the codes have also found that their scalability has been transferable to other machines.'

For example, the Model for Prediction Across Scales (MPAS) is a collaborative project that develops atmosphere, ocean and other earth-system simulation components for use in climate, regional climate and weather studies. The MPAS-Atmosphere (or MPAS-A) component is a member of the High-Q Club. The primary applications of MPAS-A are in global numerical weather prediction, with a special interest in tropical cyclone prediction and convection-permitting hazardous weather forecasting and regional climate modeling.

The extreme-scale performance of MPAS-A has improved greatly in the last 12 months, as Dominikus Heinzeller, a senior scientist at the Karlsruhe Institute of Technology, explained: 'Our alternative I/O layer is based on the SIONlib library, developed by Dr Wolfgang Frings and colleagues at Research Centre Jülich, and integrated in the MPAS model framework in a completely transparent way: users can choose at runtime whether to write data in SIONlib format or in the netCDF format that has been traditionally used in MPAS.'

'In numbers, we diagnosed a speedup of a factor of 10 to 60 when reading data in the SIONlib format, and a factor of four to 10 when writing data. Combined with a simplified bootstrapping phase, the model initialisation time, as one example, could be reduced by 90 per cent in large-scale applications on the FZJ Juqueen and LRZ SuperMUC supercomputers. This strategically important development work was supported by Michael Duda, a software engineer at NCAR (National Center for Atmospheric Research), and funded by the Bavarian Competence Network for Technical and Scientific High Performance Computin', Heinzeller added.

The Hartree Centre

The Hartree Centre in Daresbury is home to some of the most technically advanced high performance computing, data analytics, machine learning technologies and experts in the UK. Alison Kennedy, director at the Hartree Centre, said: 'Our goal is to make the transition from pursuits in research that are interesting, ➤

➤ to meaningful solutions for UK industry.'

A recent success is its cognitive hospital project the Alder Hey Children's Hospital, which was awarded 'Most Innovative Collaboration' at the North West Coast Research and Innovation Awards 2017. Hospitals produce a huge amount of data, yet it is very difficult for clinicians and patients to use that data to improve their hospital experience. Using the power of the IBM Watson, an app has been designed so that children can engage and ask questions about the procedure that they're about to undertake at the hospital. Answers are then given through a friendly avatar.

The IBM Watson cognitive computing system can process the huge amounts of data it receives quickly, extracting the most relevant and important parts. It can then transform this mountain of information into useful and personal insights that can be used to improve services or treatments at Alder Hey.

This information also helps the patient to understand and prepare for a procedure at home. 'This could help to reduce the number of no shows and will help hospital staff address any questions the patient has before they get to the hospital, which frees their time too,' Kennedy added.

Such large-scale data analysis has wider implications across the HPC sector, as Daniel Reed, vice president for research and economic development at the University of Iowa and fellow of the Association for Computing Machinery, said: 'Traditional HPC applications usually start with a question and want an answer. Now, we are starting with a set of equations and we want to compute the implications. Such large scale data analysis has turned that concept on its head.'

Machine learning

The work at Alder Hey is the tip of the iceberg as the advancement of scalable artificial intelligence (AI) and machine learning (ML) applications has really stood out for HPC in the 12 months.

While the concepts of AI are not new, progress has been facilitated by HPC in this area as the increased volumes of data we now collect allow us to train computers to make decisions based on past examples. Faster network speeds enable us to move greater amounts of data around, and better compute elements for parallel execution of data.

Gilad Shainer, chairman of the HPC Advisory Council, said: 'As these three conditions are now met, we can actually leverage AI. AI will impact nearly all aspects of our lives – from making better financial decisions, improving our security, developing self-driving vehicles, forecasting health issues and many other areas.'

Lui said: 'HPC has helped AI/ML by reducing runtime to process a workload on a single workstation which may take months to complete; the people involved in AI/ML development learned to use HPC to deploy a scale-out approach which makes use of the hardware accelerators like Nvidia Volta GPUs, and EDR InfiniBand for high-speed, low-latency network in a HPC cluster environment to reduce the runtime of training a deep neural network.

'The applications that use AI/ML that have really exploded in the last year include algorithms for detecting objects and lane in self-driving cars, object classification in image recognition, fraud detection in financial transactions and speech recognition in videos,' Lui added.

The Hartree Centre is home to some of the most advanced high performance computing in the UK

High performance computing and artificial intelligence share similar hardware requirements. Important to both is the ability to move data, exchange messages and computed results from thousands of parallel processes fast enough to keep the compute resources running at peak efficiency

The accelerated use of HPC for AI has been, in part, facilitated by better hardware resources, as Scot Schultz, senior director of HPC/Artificial Intelligence and Technical Computing at Mellanox Technologies, explained: 'High performance computing and artificial intelligence share similar hardware requirements and important to both is the ability to move data, exchange messages and computed results from thousands of parallel processes fast enough to keep the compute resources running at peak efficiency.'

For example, IBM Research just announced unprecedented performance and close to ideal scaling with its new distributed deep learning software, which achieved a record communication overhead and 95 per cent scaling efficiency on the Caffe deep learning framework with Mellanox InfiniBand and over 256 Nvidia GPUs in 64 IBM Power systems. Schultz said: 'With the IBM DDL (Distributed Deep Learning) library, it took just seven hours to train ImageNet-22K using ResNet-101. From 16 days down to just seven hours not only changes the workflow of data scientists, this changes the game entirely.'

In May 2017, Nvidia also introduced Volta, the world's most powerful GPU computing architecture, created to drive the next wave of advancement in artificial intelligence. The world's first Nvidia DGX systems with Volta AI were recently shipped to the Center for Clinical Data Science (CCDS). Paresh Kharya, group product marketing manager at Nvidia, said: 'More specifically, with this technology, CCDS data scientists can develop a host of new training algorithms to help them see medical abnormalities and patterns within medical images.'

The Tesla V100 GPU broke through the 100 Tflops barrier of deep learning performance. Kharya added: 'Demand for accelerating AI has never been greater across every industry, including healthcare, pharma, financial services, auto, retail, and telecommunications. Developers, data scientists, and researchers increasingly rely on neural networks to power their next advances in fighting cancer, making transportation safer with self-driving vehicles, providing new intelligent customer experiences, and more.'

AI-based applications have certainly dominated in the last 12 months, and this trend shows no signs of stopping. We can expect AI to become increasingly integrated in the HPC landscape, as Shainer explained: 'We will see continuous development in this area, from hardware elements to software elements and it will just keep progressing. Several years from now, we will probably be talking less about AI as it becomes mainstream and tightly integrated into more solutions.' ●

Directory of suppliers

A list leading of suppliers, consultants and integrators

Altair

1820 East Big Beaver Rd
Troy, MI 48083, USA
Tel: +1 (248) 614-2400
Fax: +1 (248) 614-2411
info@altair.com
www.altair.com

Altair's PBS Works is the trusted leader in HPC workload management, with powerful products that simplify job submission, scheduling, analytics and remote visualization. For over 30 years Altar had delivered HPC and engineering software/services to 5000+ customers in multiple industries. Altair's 2600+ employees serve clients over 68 offices in 24 countries.

ANSYS UK Ltd

Sheffield Business Park,
6 Europa View,
Sheffield, S9 1XH
UK
Tel: +44 (0)114 281 8888
www.ansys.com

We help the world's most innovative companies deliver radically better products to their customers. By offering the best and broadest portfolio of engineering simulation software, we help them solve the most complex design challenges and engineer products limited only by imagination.

Boston Ltd

Unit 5, Curo park,
Frogmore,
St. Albans,
Hertfordshire,
AL2 2DD, UK
Tel: +44 (0) 1727 876 100
sales@boston.co.uk
www.boston.co.uk

Boston Limited has been providing cutting edge technology since 1992 using Supermicro® building blocks. Our high performance, mission-critical server and storage solutions can be tailored for each specific client, helping you to create your ideal solution.

CoolIT Systems

10 – 2928 Sunridge Way NE,
Calgary, Alberta, T1Y 7H9, CANADA
Tel: +1 866 621 COOL
Fax: +1 403 770 8306
sales@coolitsystems.com
www.coolitsystems.com

With over two million units deployed worldwide, CoolIT Systems is the world leader in energy efficient liquid cooling solutions for the HPC, Cloud and Enterprise markets. CoolIT's solutions liquid cool racks of high-density servers to save energy, increase performance and provide a quiet, reliable cooling environment for customers.

GIGABYTE Technology Co., Ltd.

Bao Chiang Road No.6, 231
New Taipei City, Taiwan
Tel: +886-2-89124000
server.grp@gigabyte.com
b2b.gigabyte.com

At GIGABYTE, carried by our many years of know-how in motherboard design, it is our passion to create products for the server industry. Using the most reliable components, the best features, and the highest quality standards only, we make reliable and ultra durable server hardware for the most demanding professionals.

Mellanox Technologies, Inc.

350 Oakmead Parkway, Suite 100
Sunnyvale, CA 94085
Tel: +1 (408) 970-3400
Fax: +1 (408) 970-3403
info@mellanox.com
www.mellanox.com

Mellanox offers a choice of high performance solutions: network and multicore processors, network adapters, switches, cables, software and silicon that accelerate application runtime and maximize business results for a wide range of markets including high performance computing, enterprise data centers, Web 2.0, cloud, storage, network security, telecom, and financial services.

PGI Compilers & Tools

20400 NW Amberwood Drive
Beaverton, OR, 97006, USA
Tel: +1- 503-682-2806
Fax: +1- 503-682-2637
sales@pgroup.com
www.pgicompilers.com

PGI is a premier supplier of software compilers and development tools for parallel computing. PGI high-performance parallel Fortran, C and C++ compilers and tools are supported on 64-bit x86 and OpenPOWER processor-based systems with or without NVIDIA Tesla GPU accelerators running under the Linux, macOS and Windows operating systems. PGI offers both for-fee and no cost license options.

Samsung

Samsung Semiconductor Europe
Kölner Straße 12,
D-65760 Eschborn, Germany
Tel: +49-(0)6196-66-3300
semi.eu@samsung.com
www.samsung.com/semiconductor

Samsung is the undisputable, long-term leader for DRAM and a leading manufacturer of NAND Flash solutions. We offer DRAM solutions including 3DS TSV DRAM, and all our advanced SSD products adopt Samsung's unique 3D V-NAND technology. With Samsung Memory, you're in for industry-leading performance, density, energy efficiency, and reliability.

Super Micro Computer, Inc.

980 Rock Avenue
San Jose, CA 95131, USA
Tel: +1-408-503-8000
Fax: +1-408-503-8008
marketing@supermicro.com
www.supermicro.com

Supermicro® (NASDAQ: SMCI), the leading innovator in high-performance, high-efficiency server technology is a premier provider of advanced server Building Block Solutions® for Data Center, Cloud Computing, Enterprise IT, Hadoop/Big Data, HPC and Embedded Systems worldwide. Supermicro is committed to protecting the environment through its "We Keep IT Green®" initiative.

For news updates direct from the editorial team
@scwmagazine

SUPERMICRO®

GPU

Maximum Acceleration Highest Efficiency Application-Optimized

8 GPU with NVLink in 4U
SYS-4029GP-TXRT
Volta Ready

4 GPU with NVLink in 1U
SYS-1029GQ-TXRT
Volta Ready

10 GPU in 4U
SYS-4029GP-TRT2
Volta Ready

Designed for Highly Parallel Applications such as
AI, Deep Learning, Autonomous Vehicles, Energy and Engineering/Science

- Supports Intel® Xeon® Scalable Processors
- Up to 10 GPUs in 1U/2U/4U
- Supports Active or Passive GPUs
- 3TB DDR4-2666MHz in 24 DIMMs
- Supports Up To 100Gb/s InfiniBand or Omni-Path fabric
- Tower, Rack, and Deep Learning Optimized Models
- Unique Thermal Designs to Support High Performance GPUs and CPUs
- 3 UPI Links at 10. 4 GT/s Supported

intel XEON PLATINUM inside™

We Power the Cloud

SUPERMICRO X11 SERVER PLATFORMS

TITANIUM

nvm EXPRESS

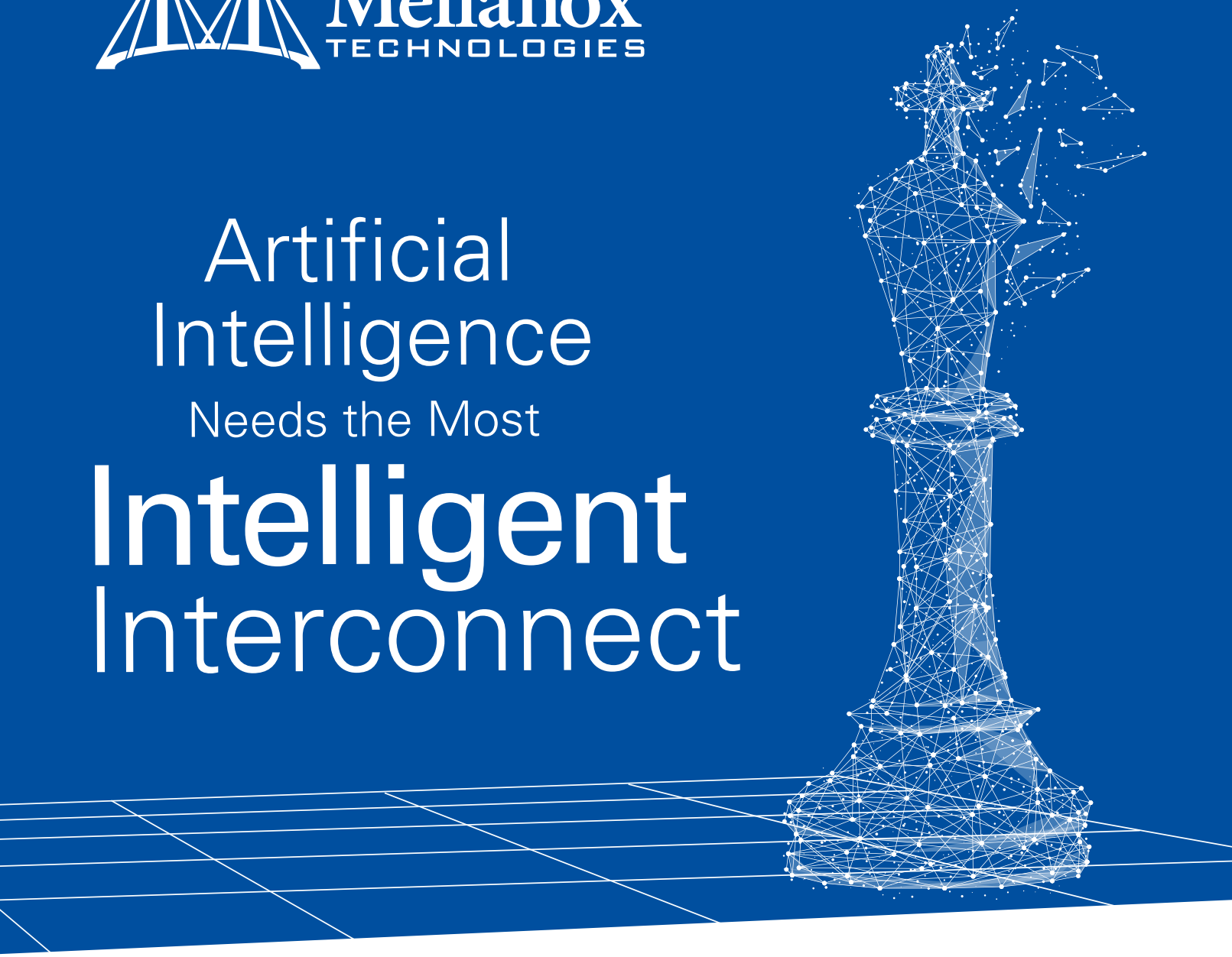
SUPERMICRO®

Intel Inside®. Powerful Productivity Outside.

Learn more at supermicro.com/X11

Mellanox TECHNOLOGIES

Artificial Intelligence Needs the Most Intelligent Interconnect



The World's Highest Performing 10/25/40/50/100 and 200Gb/s End-to-End Solutions

ConnectX-6

LinkX

Quantum

Spectrum™

Learn more: www.mellanox.com/deep-learning