HPC 2018-2019

A view into high-performance computing

From the publishers of

SCIENTIFIC COMPUTING WORLD

ALTAIR PBS WORKS™ 2018
ACCELERATING INNOVATION IN THE CLOUD

Cloud-friendly and Secure Workload Management Solution for the Entire Engineering Lifecycle

ENGINEERED TO MAKE HPC EASY, FAST, AND RELIABLE. THE ALTAIR PBS WORKS™ 2018 PRODUCT SUITE INCLUDES:

Performance, scalability and security advancements to **Altair PBS Professional™**, Altair's industry-leading workload manager and job scheduler.

Altair Access™, a new user portal environment providing intuitive, seamless access to HPC resources across the enterprise to run and manage HPC jobs, and remotely visualize results.

Altair Control™, a new HPC infrastructure management solution for administrators providing a single pane view environment to configure, deploy, monitor, and troubleshoot on-premises and cloud HPC infrastructures. A powerful what-if simulator is also included with PBS Control for advanced capacity planning along with seamless cloud-bursting capabilities.

Learn more at
pbsworks.com

Altair
PBS Works™

Contents

WELCOME TO HIGH-PERFORMANCE COMPUTING 2018-19

As the HPC community makes its final preparations for SC18, we have been taking a look at the challenges and technological developments that are shaping the industry in 2018.

On page 4 we look at potentially disruptive technologies that can take over from traditional scaling historically provided by Moore's Law. On page 8, we have a feature that explores the preparations for exascale in the US with a particular focus on the latest system at Oak Ridge National Laboratory, Summit. This system highlights the impact that AI is having on HPC development and also showcases the programming model and direction that the US is taking in the last system of the pre-exascale era.

Processor and accelerator technology are the focus on page 12, as we look back at some of the big announcements over the last 12 months. With Intel suffering setbacks in its fabrication processes, there is much stronger competition from companies such as AMD, IBM and Arm.

On page 16 we take a look at some of the Gordon Bell finalists. This year five of the six finalists for the coveted HPC award are focusing on the use of deep learning, AI or mixed precision workloads. Several are also running on Summit or Sierra showcasing the performance these new DOE systems can achieve.

On page 18 Intel's Cedric Andreolli, Jim Cownie and Kate Antakova highlight the importance of the Roofline model for performance optimisation. They also demonstrate how this model can be used to best effect when trying to optimise HPC applications.

On page 22 we have a piece from OCF's Mahesh Pancholi discussing the importance disaster recovery. This article has a focus on HPC storage and what can be done to remove or mitigate data loss in the face of disaster.

Robert Roe, editor

SUBSCRIPTIONS: *HPC 2018-19* 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. ©2016 Europa Science Ltd. Whilst every care has been taken in the compilation of this magazine, errors or omissions are not the responsibility of the publishers or of the editorial staff. Opinions expressed are not necessarily those of the publishers or editorial staff. All rights reserved. Unless specifically stated, goods or services mentioned are not formally endorsed by Europa Science Ltd, which does not guarantee or endorse or accept any liability for any goods and/or services featured in this publication.

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 image and all other images: Shutterstock.com

EDITORIAL AND ADMINISTRATIVE TEAM
Managing editor: Tim Gillett (editor.scw@europascience.com)
Editor: Robert Roe (editor.scw@europascience.com)

ADVERTISING TEAM
Advertising sales manager: Mike Nelson (mike.nelson@europascience.com) **Tel:** +44 ((0)1223 221039

DESIGN TEAM
Production manager: David Houghton (david.houghton@europascience.com) **Tel:** +44 (0)1223 221034
Senior graphic designer: Zoe Andrews (zoe.andrews@europascience.com) **Tel:** +44 (0)1223 221 035

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

Looking ahead

Robert Roe looks at technology that could disrupt the HPC ecosystem

As the HPC industry reaches the end of technology scaling based on Moore's Law, system designers and hardware manufacturers must look towards more complex technologies that can replace the gains in performance provided by transistor scaling.

There are many different potential solutions to the problems faced by in trying to extend computing performance. In the short term, HPC users have been able to look at parallelism and the increasing use of accelerators to drive more performance but this contributes to a need for memory bandwidth and more complicated coding. This is not a long-term solution as accelerators and parallel processing still rely on the same technologies and are also affected by the hard material limits placed on continued transistor scaling.

In the long-term there is some potential for multi-chip-module technology and there are rumours that Intel is working on a new dataflow computing system known as Configurable Spatial Accelerator (CSA) which could take the company away from the x-86 Von Neumann architecture.

3-D memory could also provide some much needed respite from the memory bottlenecks facing many HPC systems. Beyond the more conventional technologies there is also the promise of Quantum computing on the horizon which could also provide huge performance increases for certain types of applications.

Eric van Hensbergen, Arm fellow and director of HPC, doesn't think of Arm as a disruptive technology anymore as it has become well established in HPC. However, while the company gains momentum it must also look to a future that will require more effort to continue the performance gains seen in the past.

'Historically we were a disruptive technology in that we were in this market and coming at it from a different standpoint from others. But taking a step back, especially with the recent announcements from Intel on some of the approaches that they are taking with their new dataflow type of approach to things, and considering everybody else is GPU-based, I think we may be the only conventional architecture left in HPC.'

'As a researcher I find that profoundly disappointing, that we are no longer a disruptive technology, perhaps we are the boring one. Of course application developers and scientists like boring because it means things can work without a lot of trouble,' added Hensbergen.

Hensbergen noted the announcement of the Astra system at Sandia National Laboratories

as an example of the maturing Arm HPC ecosystem. This will be the first public petaflop system using the latest Arm processor from Cavium the ThuderX2.

'From an architectural perspective the tools and software environment have been maturing steadily, and we are quite happy with that but this is going to be larger scale and that where you find the other big challenges in HPC,' added Hensbergen.

Arm has been working with HPC partners for some time but the last 12 months has seen a number of system announcements. While most of these are test-bed systems the Isambard system at Bristol University, part of GW4 and the Post K computer developed by Fujitsu for the Riken research institute were the largest systems that had been announced. While the Post K computer will be larger than the Sandia system it is not scheduled to go into production until 2021.

Hensbergen explained that the Arm teams have been working towards large-scale systems for some time, but the Sandia Astra system will represent the first time they can see a system of that scale in action. 'Astra is one of the first at scale and I think that is going to help us mature the multi-node scaling aspect of the ecosystem a great deal,' said Hensbergen.

'Then we have Fujitsu coming further along, everything looks good on that end, and all of our other partners as well, so generally we feel that Arm is coming into its own now.'

Hensbergen also stated that the Arm is now staffing for business development in addition to its research activities as the company sees its role in the industry changing as the Arm ecosystem matures.

'We will be transitioning stewardship of the Arm HPC development roadmap from more of a research centric activity to more of a business-centric activity. Research will of course look for what is further down the road in HPC; we are not abandoning HPC or high-performance data analytics but we are staffing up more from a business perspective,' stated Hensbergen.

'A lot of our focus has been on the leadership class with a series of developments – especially the post K system. I think that in the leadership class we are feeling pretty good about the maturity of the ecosystem in being able to address that,' concluded Hensbergen.

Looking to the future

Quantum computing could be one answer to the problem of technology scaling but quantum computers remain largely untested for all but a small selection of applications. It is unlikely that quantum computers would ever replace general purpose computing systems but, for specific applications, they can deliver incredible performance. In May this year one of the biggest names in quantum computing, D-Wave, announced that it had opened a business unit that will focus on machine learning applications. Known as the Quadrant business unit, D-Wave aims to provide machine learning services that make deep learning accessible to companies across a wide range of industries and application areas. However, this new business currently uses GPU technology to give customers access to the Quadrant algorithms until the technology can be integrated into its quantum computing systems.

Quadrant's algorithms enable accurate discriminative learning (predicting outputs from inputs) using less data by constructing generative models which jointly model both inputs and outputs.

'D-Wave is committed to tackling real-world problems, today. Quadrant is a natural extension of the scientific and technological advances from D-Wave as we continue to explore new applications for our quantum systems,' said Vern Brownell, CEO at D-Wave.

D-Wave also announced a partnership with Siemens Healthineers, wa medical technology company.

Siemens Healthineers and D-Wave took first place in the CATARACTS medical imaging grand challenge, using Quadrant's generative machine learning algorithms to identify surgical instruments in videos. These algorithms are being researched as a way to improve patient outcomes through better augmented surgery and ultimately computer-assisted interventions (CAI).

'Machine learning has the potential to accelerate efficiency and innovation across virtually every industry. Quadrant's models are able to perform deep learning using smaller amounts of labelled data, and our experts can help to choose and implement the best models, enabling more companies to tap into this powerful technology,' said Handol Kim, senior director, Quadrant Machine Learning at D-Wave.

'Quadrant has the potential to unlock insights hidden within data and accelerate innovation for everything from banking and quantitative finance, to medical imaging, genomics, and drug discovery,' said Bill Macready, senior vice president of Machine Learning at D-Wave.

In addition to machine learning applications D-Wave has also been working to publish results of other application areas that can help to demonstrate the effectiveness of quantum computing technology.

The study published in the journal, *Nature*, explored the simulation of a topological phase transition using its 2048-qubit annealing quantum computer.

Application developers and scientists like boring because it means things can work without a lot of trouble

"

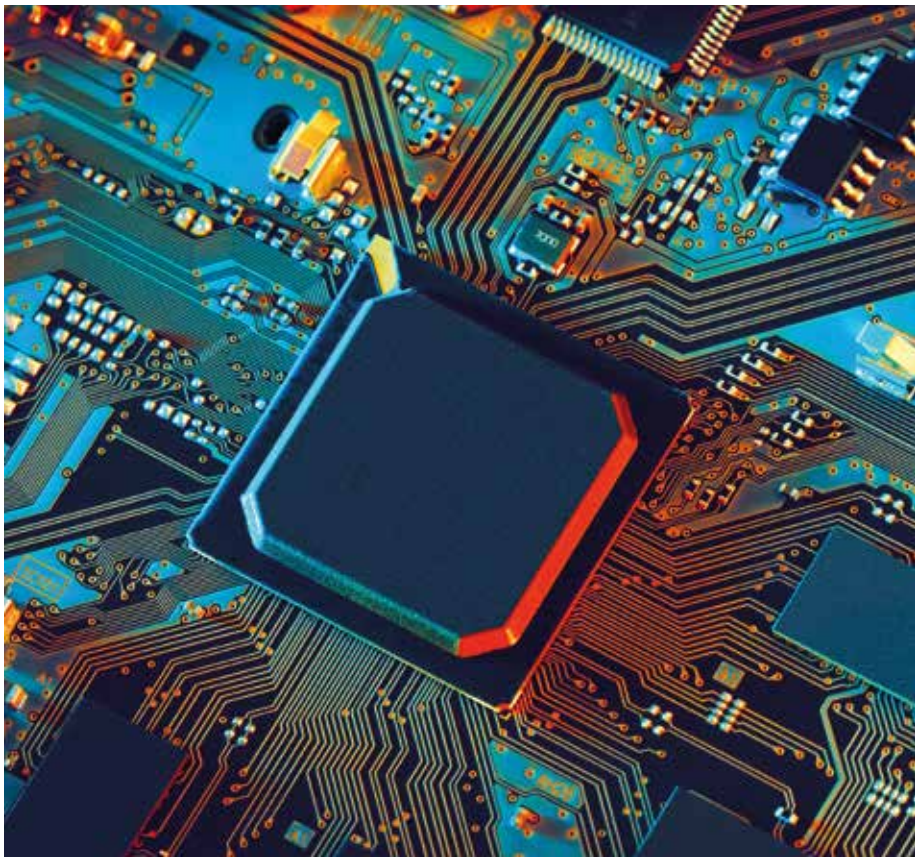
The study helps to demonstrate the fully programmable D-Wave quantum computer can be used as an accurate simulator of quantum systems at a large scale. The methods used in this work could have broad implications in the development of novel materials. This new research comes on the heels of D-Wave's recent *Science* magazine paper demonstrating a different type of phase transition in a quantum spin-glass simulation.

The two papers together signify the flexibility and versatility of the D-Wave quantum computer in quantum simulation of materials, in addition to other tasks such as optimisation and machine learning.

'This paper represents a breakthrough in the simulation of physical systems which are otherwise essentially impossible,' said 2016 Nobel laureate Dr Kosterlitz. 'The test reproduces most of the expected results, which is a remarkable achievement. This gives hope that future quantum simulators will be able to explore more complex and poorly understood systems so that one can trust the simulation results in quantitative detail as a model of a physical system. I look forward to seeing future applications of this simulation method.'

'The work described in the *Nature* paper represents a landmark in the field of quantum computation: for the first time, a theoretically predicted state of matter was realised in quantum simulation before being demonstrated in a real magnetic material,' said Dr Mohammad Amin, chief scientist at D-Wave.

'This is a significant step toward reaching the goal of quantum simulation, enabling the study of material properties before making them in the lab, a process that today can be very costly and time-consuming.'



Quadrant is a natural extension of the scientific and technological advances from D-Wave as we continue to explore new applications for our quantum systems

“

Finding a path beyond Moore’s Law

While not always the case, one of the primary drivers for innovation is necessity – to overcome grand challenges that require people to think outside of the box beyond what may seem possible today.

While we will still see continued improvements in transistor size and density the free ride of increasing performance is coming to an end. Intel has announced delays in its products based on its next generation 10nm fabrication process and while AMD appears to be preparing to announce processors based on 7nm transistor fabrication, the industry as a whole is running out of room for more improvements.

Whatever number the companies finally plateau at the main question then becomes what can be done to further increase performance in the future. It may no longer be a free lunch but that certainly does not mean that improvements will not be made.

‘If it was just us that were suffering the end of technology scaling then that would be problematic but it’s everybody. It forces all of us to think a bit harder about how we approach things. In some ways it levels the playing field instead of having one of the incumbents having a two technology node advantage over everybody else,’ said Hensbergen.

‘We have a bunch of different work going on through the research organisation on 3D integration, Multi-chip modules (MCM’s) and a whole slew of things, which is going to increase computational density, decrease latency and improve the memory situation irrespective of technology scaling,’ added Hensbergen.

As the options begin to stack up its clear that there may be no single answer to solving the computing performance challenges in the future. However as Hensbergen notes, it is important to be think carefully about which technologies the company should invest time and resources. Too much time and investment on a technology that ends up being abandoned could set a company back in the next performance race.

‘It’s not like we are coming to a halt. It’s more of a question of “wherever technology scaling ends up plateauing what do we build on top of that?” We have to be clever about how we do that especially in terms of the focus on exascale and the leadership class but they are already starting the discussions of post-exascale,’ stated Hensbergen.

‘I believe there are technologies across the spectrum that enables the continued march of computing advances. It’s not a free lunch as it has been with technology scaling; we do have to be clever about these things but we have to be careful as well,’ explained Hensbergen. ‘Things like accelerator technologies are going to be important as we move forward but it’s equally important that Arm helps to work to standardise how software interfaces are composed because that’s been such a key aspect of our value proposition.’

While Arm is exploring more technologies in the long term, in the short term the company aims to take its experience in defining and curating standards and see how this can be applied to accelerator technologies.

‘The way the accelerator world is unfolding at the moment it is a little bit like the Wild Wild West. Yes, you can abstract a lot of it behind library interfaces, but then people start pushing at the edges and if we are not careful we get into a situation where we are in a much more embedded environment where you are not going to have the same level of code portability or performance portability that you have in today’s general purpose systems,’ said Hensbergen.

‘How do we take the expertise that we have in standardising an instruction set and creating a stable software ecosystem and extend that into accelerator topics so we can keep the same level of portability within the ecosystem.

‘One of the technologies that we are looking at is as we go forward is how can incorporate accelerator technologies. But we are really trying to look at it from a standards based perspective. We really want to push it as far as we could with general purpose technology. Then as we build out acceleration technology we can do it in a way that is not going to throw the baby out with the bathwater,’ concluded Hensbergen. ■

"LEADING-EDGE DISRUPTIVE TECHNOLOGY INTEGRATORS"

vSCALER
PRIVATE CLOUD

FLASH-IO TALYN
BURST BUFFER SOLUTION

DGX-2
GPU ACCELERATED SOLUTION

SPEAK TO US ABOUT THE **NVIDIA® DEEP LEARNING INSTITUTE**

AMD Partner Program ELITE

NVIDIA ELITE PARTNER

intel Technology Provider Platinum 2018

WHY CHOOSE US?

TAILOR-MADE SOLUTIONS
Boston has the knowledge and expertise to tailor your ideal solution.

BOSTON HPC LABS
Remotely test and benchmark your technologies.

LEADING-EDGE TECHNOLOGY
Boston's R&D Labs facility offer the latest technology first.

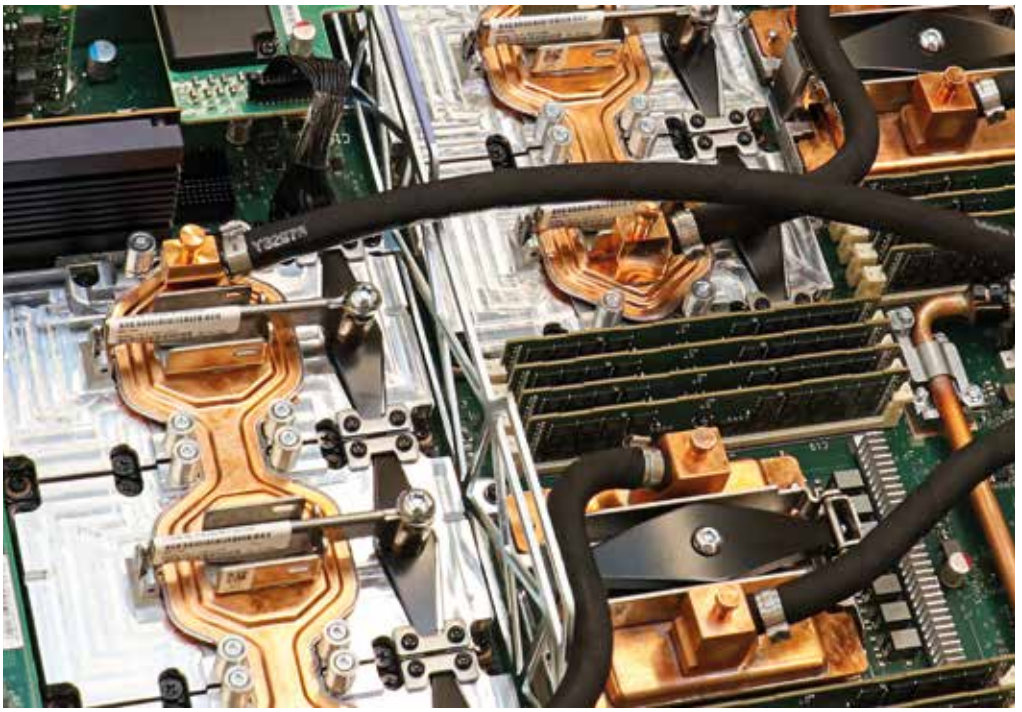
- STORAGE
- WORKSTATIONS
- SERVERS
- NETWORKING
- SOLUTIONS
- CLOUD SERVICES

BOSTON
Servers | Storage | Solutions

SUPERMICRO®

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

Approaching the Summit of exascale



Robert Roe takes a look at supercomputing development that is paving the way for the first generation of exascale systems

With exascale fast approaching, the HPC community is now looking to the last generation of pre-exascale supercomputers to see what architectures, tools and programming models can deliver before the next systems are installed in the next three to four years.

While we do not know how far developers will be able to push application performance it is possible to derive insight data from the latest generation of flagship HPC systems such as the Collaboration of Oak Ridge, Argonne, and Lawrence Livermore (CORAL). The first of these systems, Summit, installed at the Oak Ridge National Laboratory, is already breaking records for HPC and AI application performance.

The CORAL systems being installed in the US National Laboratories, similar-sized systems in China and the development plans for a European exascale system are already shedding light on what the architectures and programming models will look like on the first generation of exascale supercomputers.

Many of the choices on hardware have already been put in place, with development well under way. It is likely exascale systems will take a similar path, because to start again from an architectural perspective would require a huge amount of work to prepare both the software and tools running on the next wave of leadership-class supercomputers.

In order to deliver real exascale application performance on these systems, it is important that the compilers and other software tools are already well understood by the user community. Also, the applications must be

tested on similar architectures. While this may be an order of magnitude, smaller scale developers can test their applications before they are given access to a full-scale system.

Jack Wells, director of science at Oak Ridge Leadership Computing Facility, explains that the team behind the Summit system have gained insight from past system upgrades that maintaining as much of the programming model as possible helps to accelerate the move from one system to another.

'When we jumped to Titan it was a pretty significant change. It was much newer then and when we made the decision in 2009 and got the machines all pulled together in 2012 it was harder to use the tools, the compilers and the programming approaches were much less mature. Even working just on Titan, things got easier over time, so you could accomplish these things in marginally less time.'

Some of this difficulty was based on the timeframe and the relatively new introduction of GPUs to HPC systems which meant that the tools and the user expertise had not matured sufficiently. Even so, the team behind the Summit system aimed to keep the programming model as similar as possible, to help accelerate uptake of the system by the user community.

'We have tried to keep the programming model as much the same as we could when moving from Titan to Summit. Of course, we don't have a Cray compiler on Summit, we have an IBM compiler but we brought PGI along with us and that was available on Titan. We have tried to keep the transition of the programming model as easy as it could be.'

Wells stressed that, even with the correct tools, HPC is not easy per se. The main aim is to try and reduce the difficulty as much as possible and to provide a familiar environment for the user community to help drive sustainable application development.

'Those are the kind of things that we need to concern ourselves with. Sustainability of applications software and programming models on our supercomputers,' concluded Wells.

We have tried to keep the programming model as much the same as we could when moving from Titan to Summit

"

Co-designing the next generation of supercomputers

The development of exascale systems has been focused on co-design for a number of years. European efforts such as the Mont Blanc project which explored the use of Arm-based clusters, have formed the foundations for the European Processor Initiative (EPI). In the US, projects such as FastFoward and FastForward2, have looked at different aspects of HPC hardware and programming such as processors, memory and node architecture for future supercomputing platforms. It is clear that no one company or government organisation can design a modern supercomputer. In order to continue to drive innovation, energy efficiency and performance, there needs to be considerable effort placed into partnerships between academia and research organisations, hardware companies and application developers funded through government-led projects that help subsidise the cost of exascale development.

The development of the CORAL systems, specifically Summit, has been no different. Geetika Gupta, product lead for HPC and AI, explains the importance of collaboration between Nvidia and Oak Ridge in designing Summit and influencing the development of GPU technology.

'Oak Ridge and some of the other partner labs started exploring the use of GPUs way back in 2008 or 2009. At the time they started with some cards they felt that they had the right level of parallelism. Some of the life sciences codes, such as Amber, can take advantage of the compute cores available as part of the GPUs,' said Gupta.

'In 2013 they installed the first GPU-based supercomputer called Titan. That was based on the Kepler GPU architecture. I was the product manager at that time for the Kepler series of products, so I was closely involved with the Titan deployment in the 2012/2013 timeframe,' added Gupta.

Since that time, Nvidia has been continually involved with Oak Ridge and the Oak Ridge Leadership Computing Facility (OLCF) working to understand the requirements of future systems and the applications used by the user community. Gupta stressed that this feedback on applications and workloads helps to inform what Nvidia should be building in future generations of GPU architecture.

'When the time came to come up with the follow up to Titan, Nvidia was closely involved. The way they had described the workloads, we saw that there was a need to increase GPU-GPU communication. In 2014 when we started thinking about the next system, we knew that we needed a fast GPU-GPU interconnect and that formed the basis for NV Link,' said Gupta.

'We continued to work with them and analyse the workloads to find the bottlenecks and see what new things were emerging, and that helped to influence the basis of the designs for the Volta GPU architecture,' stated Gupta.

However, not all trends and changes in

the development of computing can be neatly predicted years ahead of time. Early machine learning and deep learning methods have been around since the 1950's but the convergence of data availability, GPU acceleration and algorithmic improvements led to the explosion of deep learning across scientific and industry in the last few years.

Nvidia led much of the hardware for this with its GPUs, so it made sense for them to add in hardware specific to deep learning which would accelerate applications. This led to the development of tensor cores, which are now included in the latest Nvidia GPUs, including the Tesla V100 that is used in the Summit system.

Adding new hardware to the HPC toolbox

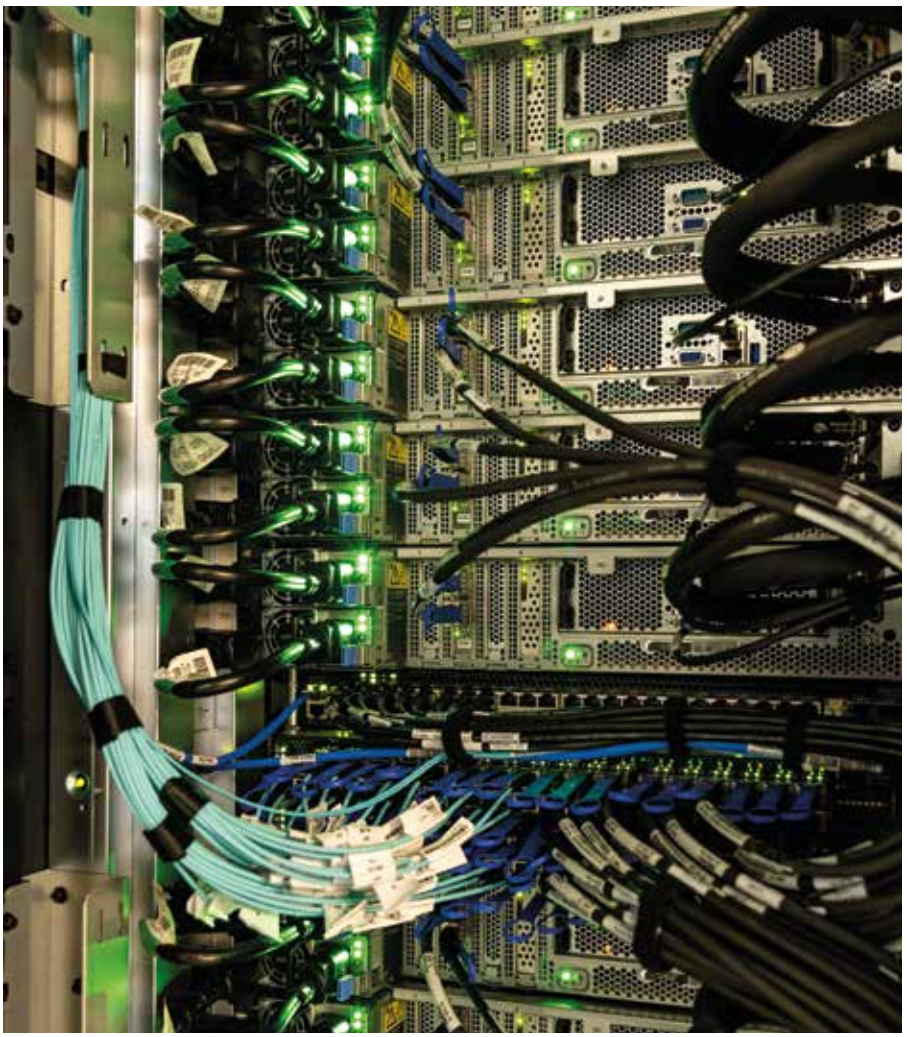
'While all this was happening we could see that AI and deep learning was also emerging as one of the primary tools to analyse large amounts of data, and that was the reason that we decided to include tensor cores in the Volta GPU architecture,' said Gupta. 'Now you can see how the V100 can be used to enable computational science and they have support for tensor cores, which can be used to assist scientific computation with AI and deep learning.'

'CUDA cores are great for doing FP64 based matrix multiplication but there is a lot of work that can be done at lower precision, so the tensor cores have been added to the Volta GPU architecture to optimise that lower precision computation,' added Gupta.

Gupta explained that the tensor cores take half-precision floating-point format (known as FP16) as an input. Tensor cores carry out operations such as matrix multiplications and accumulate that data into single precision floating-point format (FP32). 'The data pipeline for the tensor cores has been designed to do these operations much faster, just because of the way that the cores are fed data. They can do almost 12 times the operations that a CUDA core would be able to.

'It is quite interesting that a scientific workload on one single GPU architecture can get a mixed precision computation. They can decide where they want to use low precision for coarse grain analysis, and then in some of the later iterations,' stated Gupta.

'People are looking at AI and deep learning as a new tool in their toolbox. It's not something that is going to replace the existing way of doing scientific computation, but it is definitely a new tool that they can use to do certain types of workloads and speed up the process. The other reason is the amount of data that scientists need to analyse. It takes much longer if you are just relying on traditional methods,' Gupta concluded.



Getting applications up and running on Summit

As the Summit supercomputer has now been installed and is awaiting full acceptance, the OLCF has been inviting application users onto the system to test application performance and help the team to remove any bugs ahead of full production at the beginning of 2019.

The OLCF has several programmes to develop and test applications for new and existing users. The Innovative and Novel Computational Impact on Theory and Experiment (Incite) programme is one example of the US Department of Energy's efforts to allocate computing time to users across the globe.

The OLCF also run the Application Readiness Program and the Directors Discretionary Program (DD) among other projects to help new and existing users obtain access to the Summit system.

Oak Ridge also invited 13 projects to be part of its Center for Accelerated Application Readiness (CAAR). A collaborative effort of application development teams and staff from the OLCF Scientific Computing group, CAAR is focused on redesigning, porting and optimising application codes for Summit's hybrid CPU–GPU architecture. This gives users early access to explore the architecture but users can also receive technical support from the IBM/Nvidia Center of Excellence at Oak Ridge National Laboratory.

'Our commitment is to be able to start the INCITE user programme in January. Our plan is to accept the full system soon and then have the fourth quarter of the calendar year for an early science period. This will allow us to get some additional hero users on there, to knock the cobwebs out for the system but we will start INCITE in January. Those proposals are being evaluated now.'

Wells stressed that it is important that many of the users come from different application areas but also from different institutions, from both the US and other nations, in addition to catering to US national interests.

Those are the kind of things that we need to concern ourselves with sustainability of applications software and programming models on our supercomputers

“

While INCITE looks at developing a broad community of users, other programmes focus more on US interests.

‘The big-user programmes are INCITE, ALCC programme. When the leadership computing facilities were established the resources should be allocated based on merit, and it should be available to US industry, universities, national laboratories and other federal agencies. The money comes from the DOE office of Science programme but the user base is much broader than that.’

‘The DOE implements this complicated combination of programmes through its “user facility model”, with the same business model that it would use for the light sources, nanoscience centres or the joint Genome Institute. It makes it available to the world so it is available to international competition and we do that through the INCITE programme,’ added Wells.

‘DOE Office of Science programmes have the need for capability computing too, and so they are in the best position to understand those programmatic priorities. They have a programme where they can support projects of interest to the DOE. The INCITE programme does not take the DOE’s interests per se, we don’t consider that it is based on scientific merit as determined by peer review.

These programmes provide many options for users to gain access to Summit or the other national laboratory computing systems but Wells explains that for new users, the DD programme is usually the first step.

‘You get started with the DD programme because the responsibility for that is given to me and the team that I lead, so that is part of my job. A lot of people do this role but I am involved in getting users started on the machine,’ stated Wells.

‘We have three goals for our DD programme, one is to allow people to get preliminary results for these other user programmes I have mentioned, because it is competitive and they need results to show and strengthen their proposal,’ added Wells. ‘Then we do outreach to new and non-traditional use cases. In the past we didn’t have so much data analytics and AI and, in order to get people started in these areas in 2016 and 2017, we gave several small allocations to test out their workflows and to try out the codes. This allows them to get some performance data and maybe write us a few introductory papers.’

Wells noted that the third use of the DD programme is to help support local teams that need help in order to get a further allocation, or to support other work done at the DOE or national laboratories. ‘We use a small amount of this time to help people get started at the lab. Maybe we have a new hire or an internally funded project that hasn’t had time to compete for a big allocation on a supercomputer. We will support a local team in that way.’

Application readiness

Many of the so-called ‘hero’ users and their applications were given access to the Summit system through the Application readiness programme. This is key to the OLCF, as it gives an indication of the kind of applications that will be capable of using the entire system.

‘The majority of those projects have so far been able to demonstrate that they can use around a thousand or two thousand nodes. Eventually, almost all of the projects will be able to use the whole machine but they have not demonstrated it yet, because they have not yet been given access to the whole machine,’ said Wells.

‘We did open up broader access to an early science call for proposals, where we had 64 teams ask for access. We tried to prioritise their early access also, as this enabled them to get some results.

‘Not everybody got equal access, so it’s not like it was a fair thing but we wanted to get as many people on, as early as we could, in the middle of all the development activities. From this early work, a set of them wanted to go for the Gordon Bell Prize. That visible competition is something that we want to support, so we really enabled those teams to have even more access,’ added Wells.

‘We had something like seven teams submit papers for the Gordon Bell Prize and, as we recently announced, five of them were finalists,’ noted Wells. He also stressed that this is a huge achievement, considering that the Summit system has been through full acceptance, yet still has a number of bugs. The teams were also using the test and development file system, as the full storage system has not yet been accepted. ■

A year in processor development

Robert Roe looks back at the processor news in 2018

With new and old companies releasing processors for the HPC market, there are now several options for high-performance server-based CPUs. This is being compounded by setbacks and delays at Intel opening up competition for the HPC CPU market.

AMD has begun to find success on its EPYC brand of server CPUs. While market penetration will take some time, the company is starting to deliver competitive performance figures.

IBM supplied the CPUs for the Summit system, which currently holds the top spot on the latest list of the Top500, a biannual list of the most powerful supercomputers. While a single deployment is not a particularly strong measure of success, the Summit system has generated a lot of interest, five of the six Gordon Bell Prize finalists are running their applications on this system, which highlights the potential for this CPU – particularly when it is coupled with Nvidia GPUs.

Arm is also gathering pace, as its technology partner's ramp up production of Arm-based CPU systems for use in HPC deployments. Cavium was an early leader in this market, delivering the ThunderX processor in 2015 and its follow up ThunderX2 was released for general availability in 2015.

There are a number of smaller test systems using the Cavium chips, but the largest is the Astra supercomputer being developed at Sandia National Laboratories by HPE. This system is expected to deliver 2.3 Pflops of peak performance from 5,184 Thunder X2 CPUs.

HPE, Bull and Penguin Computing have added the ThunderX2 CPU to its line-up of

We're putting that $1 billion into our 14nm manufacturing sites in Oregon, Arizona, Ireland and Israel. This capital, along with other efficiencies, is increasing our supply to respond to your increased demand

products available to HPC users. Coupled with the use of Allinea software tools, this is helping to give the impression of a viable ecosystem for HPC users.

With many chip companies failing or struggling to generate a foothold in the HPC market over the last 10 to 20 years, it is important to provide a sustainable technology with a viable ecosystem for both hardware and software development. Once this has been achieved, Arm can begin to drive the market share.

Fujitsu is another high-profile name committed to the development of Arm HPC technology. The company has been developing its own Arm-based processor for the Japanese Post K computer, in partnership with Riken, one of the largest Japanese research institutions.

The A64FX CPU, developed by Fujitsu, will be the first processor to feature the Scalable Vector Extension (SVE), an extension of Armv8-A instruction set designed specifically for supercomputing architectures.

It offers a number of features, including broad utility supporting a wide range of applications, massive parallelisation through the Tofu interconnect, low power consumption, and mainframe-class reliability.

Fujitsu reported in August that the processor would be capable of delivering a peak double precision (64 bit) floating point performance of over 2.7 Tflops, with a computational throughput twice that for single precision (32 bit), and four times that amount for half precision (16 bit).

Trouble at the top

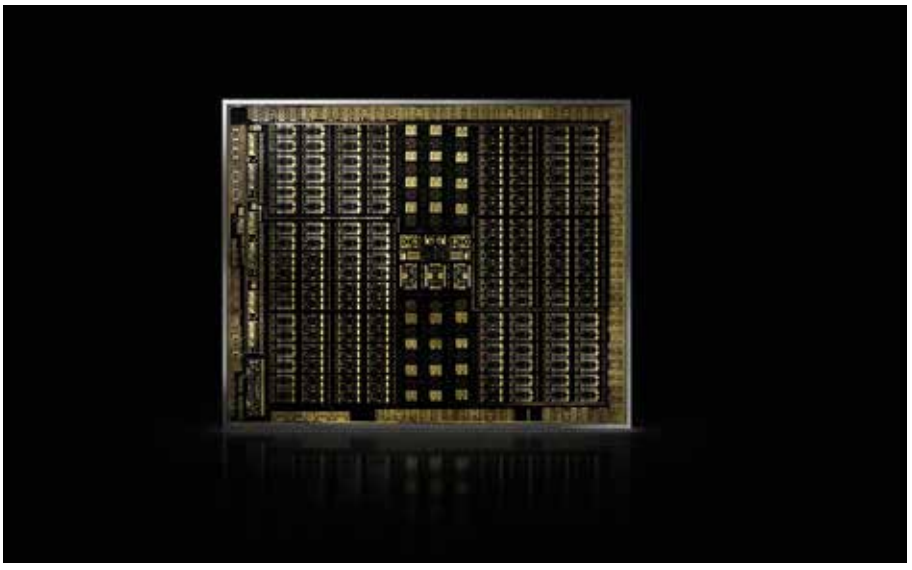
Intel has been seen to struggle somewhat in recent months, as it has been reported that the next generation of its processors has been delayed due to supply issues and difficulty in the 10nm fabrication processes.

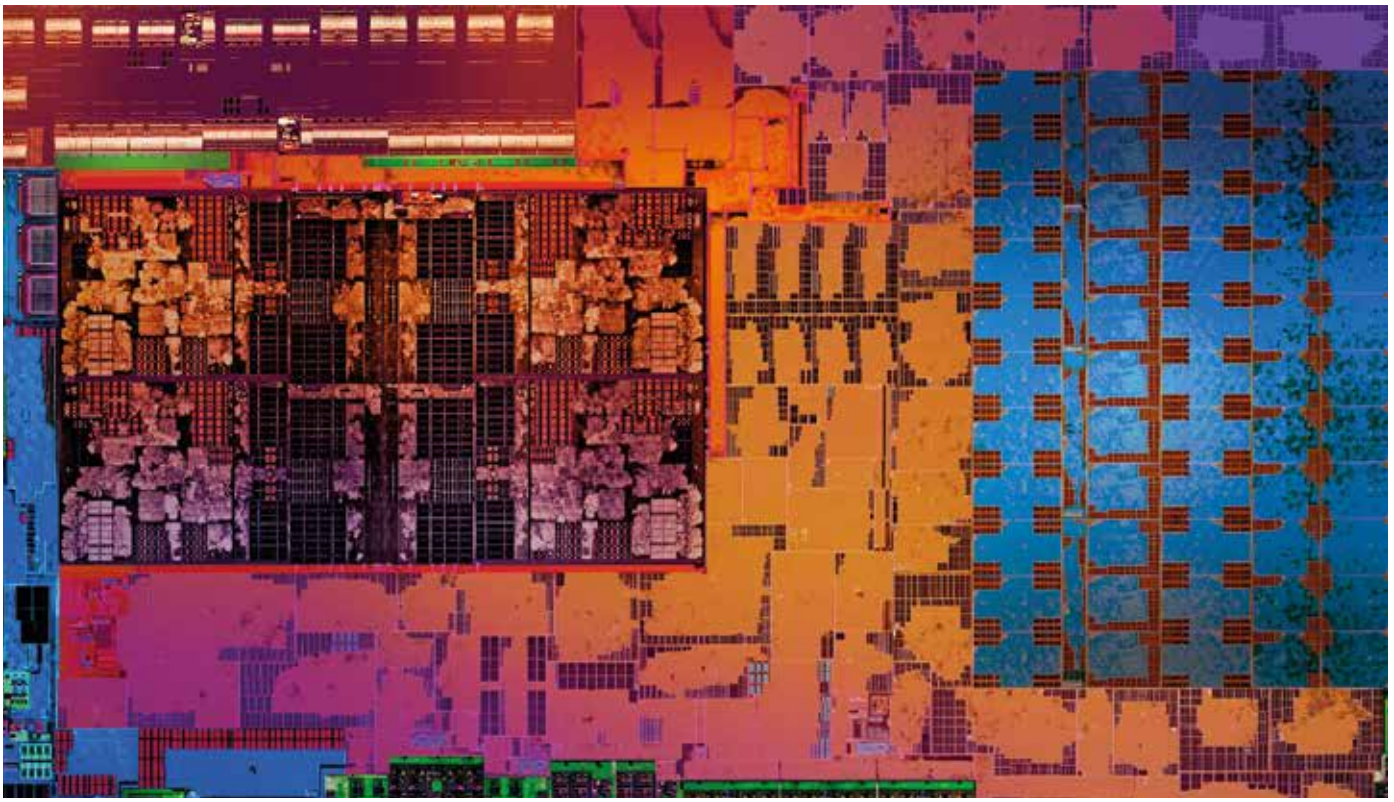
The topic was addressed in August by Intel's interim CEO Bob Swan, who reported healthy growth figures from the previous six months but also mentioned supply struggles and record investment processor development.

'The surprising return to PC TAM growth has put pressure on our factory network. We're prioritising the production of Intel Xeon and Intel Core processors so that collectively we can serve the high-performance segments of the market. That said, supply is undoubtedly tight, particularly at the entry-level of the PC market. We continue to believe we will have at least the supply to meet the full-year revenue outlook we announced in July, which was $4.5 billion higher than our January expectations,' said Swan.

Swan stated that the 10nm fabrication process was moving along with increased yields and volume production was planned for 2019: 'We are investing a record $15 billion in capital expenditures in 2018, up approximately $1 billion from the beginning of the year. We're putting that $1 billion into our 14nm manufacturing sites in Oregon, Arizona, Ireland and Israel. This capital, along with other efficiencies, is increasing our supply to respond to your increased demand.'

While Intel is undoubtedly the king of the hill when it comes to HPC processors – with more than 90 per cent of the Top500 using Intel-based technologies – the advances made by other companies, such as AMD, the re-introduction of IBM and the maturing Arm ecosystem are all factors that





mean that Intel faces stiffer competition than it has for a decade.

The Rise of AMD

The company had success in the headlines at the end of 2017 when the new range of server products was released but, as Greg Gibby, senior product manager of data centre products at AMD notes, he expects the company will begin to see some momentum as several 'significant wins' have already been completed.

Microsoft has announced several cloud services that make use of AMD CPUs and the two socket products are also being deployed by Chinese companies such as Tencent for cloud-based services and Baidu has adopted both CPUs and GPUs from AMD to drive its machine learning and cloud workloads.

AMD is generating huge revenue from its console partnerships with Sony and Microsoft.

While these custom CPUs do not directly impact HPC technology, the revenue provided valuable time for AMD to get its server products ready. In 2018 the server line-up has been successful and AMD is rumoured to announce 7nm products next year. If this comes to fruition AMD could further bolster its potential to compete in the HPC market.

Gibby also noted that as performance is a key factor for many HPC users, it is important to get these products in front of the HPC user community.

He said: 'I believe that as we get customers testing the EPYC platform on their workloads, they see the significant performance advantages that EPYC brings to the market. I think that will provide a natural follow-through of us gaining share in that space.'

One thing that could drive adoption of AMD products could be the memory bandwidth improvements which were a key feature of AMD when developing the EPYC CPUs. Memory bandwidth has long been a potential bottleneck for HPC applications, but this has become much more acute in recent years.

I believe that as we get customers testing the EPYC platform on their workloads they see the significant performance advantages that EPYC brings to the market

"

In a recent interview with *Scientific Computing World*, Jack Wells, director of Science at Oak Ridge National Laboratory noted it as the number one user requirement when surveying the Oak Ridge HPC users.

This was the first time that memory bandwidth had replaced peak node flops in the user requirements for this centre.

While AMD was designing the next generation of its server-based CPU line, it took clear steps to design a processor that could meet the demands of modern workloads.

Gibby noted that the CPU was not just designed to increase floating point performance, as there were key bottlenecks that the company identified, such as memory bandwidth that needed to be addressed.

'Memory bandwidth was one of the key topics we looked at, so we put in eight memory channels on each socket,' said Gibby. 'So in a dual socket system, you have 16 channels of memory, which gives really good memory bandwidth to keep the data moving in and out of the core.

'The other thing is on the I/O side. When you look at HPC specifically, you are looking at clusters with a lot of dependency on interconnects, whether it be InfiniBand or some other fabric.

'A lot of the time you have GPU acceleration in there as well, so we wanted to make sure that we had the I/O bandwidth to support this.' ■

Subscribe for free*

Do you compute?

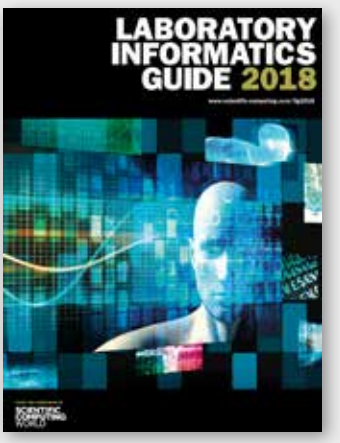
The only global publication for scientists and engineers using computing anwd software in their daily work

Do you subscribe? Register for free now!

scientific-computing.com/subscribe

*Registration required

Also published by Europa Science



Gordon Bell Prize finalists highlight the impact of AI

Robert Roe reports on the Gordon Bell Prize finalists for 2018, as the list is dominated by AI and deep leaning applications running on the Summit supercomputer

Each year the Association for Computing Machinery award the Gordon Bell Prize to recognise outstanding achievement in high performance computing (HPC). This year, five out of the six finalists are running on GPU-based system Summit, at Oak Ridge National Laboratory, or the Sierra system at Lawrence Livermore National Laboratory.

The finalist's research ranges from AI to mixed precision workloads, with some taking advantage of the Tensor Cores available in the latest generation of Nvidia GPUs. This highlights the impact of AI and GPU technologies, which are opening up not only new applications to HPC users but also the opportunity to accelerate mixed precision workloads on large scale HPC systems.

This is a very compute intensive task because they are basically comparing the DNA of all of the samples for a given population

"

Jack Wells, director of Science Oak Ridge Leadership Computing Facility (OLCF) comments: 'All five of those projects have already demonstrated that they can use the system at full scale. That is pretty remarkable, because we are still kicking the bugs out of the system software and they were only using our test and development file system, because we are still final acceptance of the full 250 petabyte file system.'

Deep learning and AI research has been adopted relatively quickly by the HPC and scientific communities and this is demonstrated by the ACM Gordon Bell Prize finalists. This relatively new opportunity in computing – HPC systems loaded with GPUs, high bandwidth interconnects, and Tensor Cores which can accelerate mixed precision or deep learning workloads – is already impacting the way research is carried out on large-scale HPC clusters.

In 2017 the Gordon Bell prize was awarded to a team from China that worked on the Sunway TaihuLight system to simulate the most devastating earthquake of the 20th century.

Using the Sunway TaihuLight, which at the time was ranked as the world's fastest supercomputer, the team developed software able to process 18.9 petaflops of data and create 3D visualisations of a devastating earthquake that occurred in Tangshan, China, in 1976.

The team was awarded the prize for their

application, which included innovations that helped the researchers achieve greater efficiency than had been previously possible running software on the predecessor to the Summit system, Titan, and the Chinese TaihuLight supercomputer.

The recipients of the prize noted that they had worked on several innovations. This included a customised parallelisation scheme that employs 10 million cores efficiently; a memory scheme that integrates on-chip halo exchange through register communication, optimised blocking configuration guided by an analytic model, and coalesced DMA access with array fusion; on-the-fly compression that doubles the maximum problem size and further improves the performance by 24 per cent.

The Tangshan earthquake was one of the most damaging natural disasters of the 20th century. The earthquake in Hebei province resulted in between 242,000 and 700,000 deaths.

Understanding the phenomena and how it might impact the local population is hugely important to preventing extensive damage and loss of life during future disasters. In order to create as accurate a model as possible, the team developing the simulations for the Tangshan earthquake used input data from the entire spatial area of the quake, a surface diameter of 320 km by 312 km, as well as 40 km below the earth's surface.

In 2018 several of the nominees for the award are employing deep learning, AI or mixed-precision computation that runs quickly on the Tensor Cores – a new hardware addition included with the Nvidia V100 GPUs to accelerate deep learning applications.

The finalists from 2018 include researchers working on weather and climate, earthquake simulation, genomics, electron microscopy and research that aims to quantify the lifespan of neutrons.

One of the projects that is a finalist for the Gordon Bell Prize, 'Development of genomics algorithm to attain exascale speeds', was developed by a team from Oak Ridge National Laboratory led by Dan Jacobson.

The team achieved a peak throughput of 2.31 exaops, the fastest science application ever reported. Their work compares genetic variations within a population to uncover hidden networks of genes that contribute to complex traits.

'This is a very compute intensive task, because they are basically comparing the DNA of all of the samples for a given population. In the past it was thought to be this heroic thing that no one had enough compute to run but it's intrinsically an integer code, it's not really even a floating point code because you are comparing DNA sequences,' added Wells.

With the end of Dennard scaling, getting performance out of computers is more difficult and people are going to start looking in the cracks and corners of the room for more opportunity to get performance
"

Wells noted that, when Jacobsen and his team converted the integer code to mixed precision FP16 code that could be run on the Tensor Cores, they were able to get a four-times speed boost on the integer code running on the Summit system.

Overall the speedup from both algorithmic and hardware improvements has enabled this application not only to reach an unprecedented speed of 2.31 exaops, but also begin to compare genetic variations within a population to uncover hidden networks of genes that contribute to complex traits.

Geetika Gupta, product lead for HPC and AI at Nvidia, said: 'Accelerators enable multi-precision computing that fuses the highly precise calculations to tackle the challenges of high-performance computing with the efficient processing required for deep learning.

'There are many scientific applications that will be driven forward by the convergence of AI and deep learning. Something that started on the consumer side has been picked up by the scientific community, which is a great example of two different use cases coming together,' commented Gupta.

Another Summit application has already broken the exaflop barrier running a deep learning application using FP16 computation. The team from Lawrence Berkeley Laboratory, led by Prabhat, are working on the project entitled: 'Identification of extreme weather patterns from high-resolution climate simulations' which aims to analyse how extreme weather is likely to change in the future.

The team has already managed to achieve a performance of 1.13 exaflops, the fastest deep-learning algorithm reported.

'That is a more traditional deep learning neural network application of feature classification but they were able to use tensor cores as they were intended to be used for deep learning and we were able to scale up to an exaflop,' said Wells.

Wells also explained that mixed precision was not something that had been discussed much in the early development of the Summit system or more widely by the HPC community. There have been examples of mixed precision HPC workloads in the past but it had not been particularly widespread within the community.

Wells added: 'When we said that Summit was a 200 petaflop machine, it was left moot how much precision you have'.

Wells also stressed that these kinds of techniques were likely to become more widespread, as HPC users look for ways to increase performance: 'With the end of Dennard scaling, getting performance out of computers is more difficult and people are going to start looking in the cracks and corners of the room for more opportunity to get performance.'

'The earthquake research team from the University of Tokyo, which is also a Gordon Bell Prize finalist, described a trans-precision algorithm but they are actually not using the Tensor Cores yet because their problem is really more communication-bound', added Wells.

Their project – 'Use of AI and transprecision computing to accelerate earthquake simulation' – used Summit to expand on an existing algorithm. The result was a four times speedup, enabling the coupling of shaking ground and urban structures within an earthquake simulation. The team started their GPU work with OpenACC. They later introduced CUDA and AI algorithms to improve performance.

'They are doing an unstructured finite element analysis of earthquakes coupled to buildings, so they can develop a greater understanding of how Tokyo station, for example, might react in the event of a serious earthquake.

'They use it to reduce their communication burden. They will probably start to use the Tensor cores, but they have not got around to it yet.

'Right now, they are going from FP64 to FP32 actually to FP21, which they just do in software to FP16 in order to actually move the numbers around, in order to complete their communication and then iteratively refine towards the end of the calculation. For them, it's about reducing the communication burden,' concluded Wells. ■

Understanding and improving the performance of bandwidth bound code

Intel's **Cedric Andreolli**, **Jim Cownie** and **Kate Antakova** explain how the Roofline model can be used to improve HPC code

High performance computing is a field in which engineers often require knowledge of many different areas. An HPC developer must implement a solution to a problem which comes from 'real' science, but they also need to understand computer hardware and software. Understanding and improving the performance of an application is a difficult challenge and tools that can help are usually welcome.

In this article we will discuss how to understand the factors which frequently limit the performance of code on modern hardware, and, in particular the issue of data movement. When we are optimising our objective is to determine which hardware resource the code is exhausting (there must be one, otherwise it would run faster!), and then see how to modify the code to reduce its need for that resource.

The Roofline model provides us with a way to understand the trade-off between data-movement and computation so that we can understand which factor is limiting our code

"

It is therefore essential to understand the maximum theoretical performance of that aspect of the machine, since if we are already achieving the peak performance we should give up, or choose a different algorithm.

The Roofline model provides us with a way to understand the trade-off between data-movement and computation so that we can understand which factor is limiting our code, and how close to the limit we are. It has been heavily used recently to characterise HPC applications and understand their potential for performance improvement. Intel Advisor implemented a Cache Aware Roofline Model (CARM) a few years ago. This provides a very powerful way to characterise HPC applications.

More recently, the software implemented an additional version of the Roofline model called the Integrated Roofline, which allows it to detect specific bottlenecks related to threading and vectorisation, as well as inefficient data transfers between each level of memory (Level 1 cache, Level 2 cache, Level 3 cache, memory).

If you are tuning compute intensive code, you need to understand data movement, and the roofline model is a good way to do that.

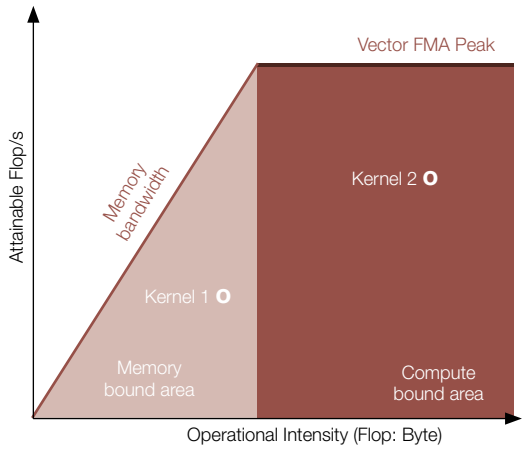
The main aim of the Roofline model is to incorporate performance, memory bandwidth, and memory locality metrics into a unified chart for the kernels of the analysed program. By comparing where each kernel is in the Roofline model, we can see how near

it is to the compute and memory roofs and we can get a measure of how much system performance is not exploited.

The original Roofline model developed by Berkeley focuses on the performance of specific kernels (in floating-point operations per second) mapped to the operational intensity which represents the ratio of floating-point operations to bytes of traffic transferred between the caches and DRAM. The computational performance roofs represent the maximum capabilities of the platform. For a modern processor the maximum (peak) performance for floating point operations is considered to be vector Fused-Multiply-Add (FMA) instructions. When other instructions are used, the peak performance may vary and multiple computational roofs can be included in the model based on the instructions used in the code, the data types processed and the machine architecture.

The memory bandwidth limit of the system is represented by the maximum main memory bandwidth in this roofline model. For any specific kernel plotted on such a chart the model gives a general answer about whether the kernel is memory- or compute-bound and suggests the general optimisation direction. However, the original roofline model does not give specific insight into which types of memory optimsations are profitable for the kernel and whether particular cache levels limit the performance.

FIG 1



Cache Aware Roofline Model Analysis Implementation in Intel Advisor

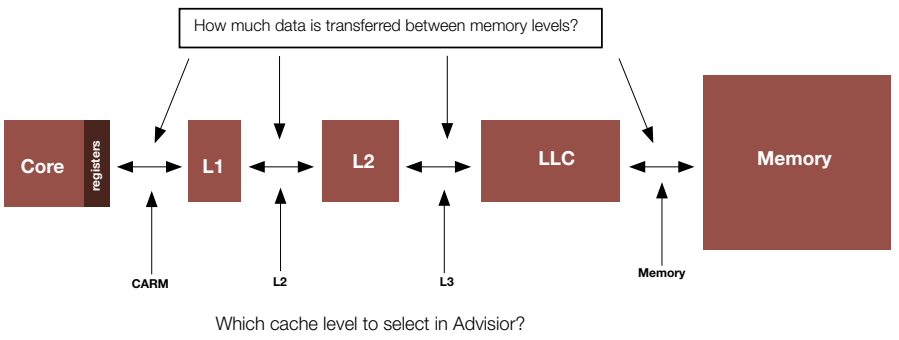
CARM was developed to provide additional information about application kernels' performance in a system with caches. This new model considers bytes of traffic transferred from CPU cores to each level in the memory hierarchy, making it easier to understand which cache-levels are important for optimisation.

Intel Advisor analyser implements the automatic collection of CARM for user applications. It does this by using instrumentation to understand your code. By running the instrumented application, it can count all the arithmetic operations and memory movements. To build the model it must acquire three types of performance data: the number of floating point operations, the amount of data that was transferred to execute those operations, and the execution time of the different sections of code. The instrumentation provides an accurate counting of the floating point operations based on what was really executed by the machine, allowing for masking I vector operations. This counting doesn't rely on performance counters and is not sensitive to any specific hardware. Counting the memory movement is also implemented by instrumentation, checking what ends up in the registers. This approach ensures that for a given platform and algorithm, the ratio of operations and memory transfers will stay the same, no matter what optimisations are implemented or the size of the dataset used, providing the CARM with strong characteristics to analyse an algorithm.

The implementation also has the advantage that the overhead (of around seven times slowdown) is bearable. To avoid impacting the performance of the application when measuring the execution time of each kernel, Intel Advisor split the analysis into two runs.

The first run uses sampling to time each uninstrumented kernel. The second run uses the instrumented kernel to count the flops, the memory transfers and the trip counts. Then, data from the two runs are merged to extract a set of metrics for each kernel such as the performance (flop/s) and the Arithmetic intensity (flop/byte). Even if detecting performance issues is possible with this model, identifying a bottleneck is challenging as every optimisation (memory, threading, vectorisation) will affect only the vertical position of a kernel plotted on the roofline model (flop/s). This is where the Integrated Roofline Model becomes interesting.

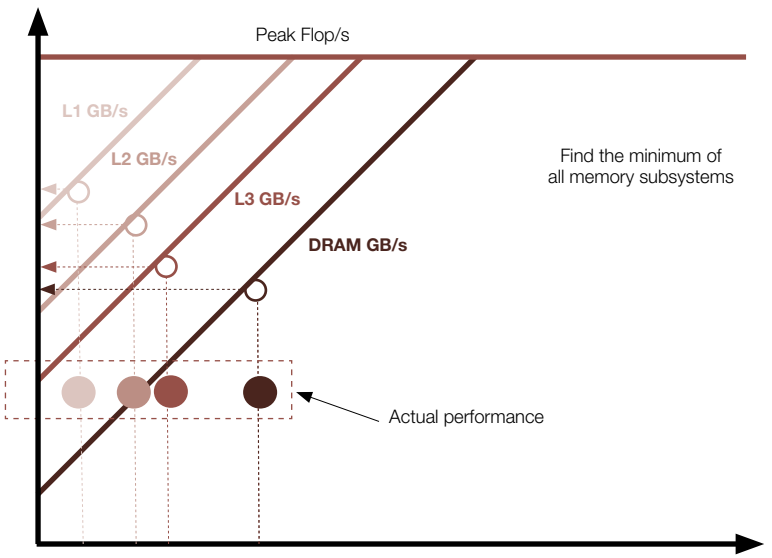
FIG 2



Integrated Roofline Model

The Integrated Roofline Model (IRM) implementation uses more than just code instrumentation. Intel Advisor also implements a cache simulator which can reproduce the behaviour of different cache levels. Using the cache simulator, it becomes possible to count the memory transfers between the cache levels. For a single loop, we

FIG 3



This new model considers bytes of traffic transferred from CPU cores to each level in the memory hierarchy, making it easier to understand which cache-levels are important for optimisation

“

can now generate several data points on the roofline plot for a single kernel, depending on whether we consider data transfers between registers and cache, L1 and L2, L2 and L3, L3 and memory.

Having several points on the roofline for the same loop brings many new possibilities in terms of interpretation. First of all, the way to interpret the theoretical maximum performance changes, as each dot must be projected on its appropriate roofline depending on the memory subsystem used.

Identifying vectorisation, threading and latency issues

The first problems which can easily be identified are inefficient vectorisation. Intel Advisor can display the IRM for a single thread.

In this situation, if we see that every dot is far from its dedicated roofline (as in the diagram), inefficient vectorisation is likely. In this case, it is important to check that loops were correctly vectorised.

In some cases, this problem can also happen on vectorised loops, meaning that vectorisation suffers from inefficient data accesses, due to latency.

Identifying threading issues can be done by running your analysis on a multithreaded system and displaying the roofline for all the threads used.

If all the dots of a single kernel are far from their respective rooflines and you had previously solved your vectorisation issues, threading might be the reason for your suboptimal performance.

Identifying cache overload

The IRM also allows us to identify too high a pressure on a given cache level. Observing the different arithmetic intensities of a single kernel at different memory subsystem levels provides meaningful information on how much pressure is inflicted on each cache/memory level.

A low arithmetic intensity for a given cache level will increase the chance that this cache is a performance bottleneck.

Usage of Non Temporal Stores (NTS) can also affect performance by releasing the pressure on caches. NTS allows the core to write data directly to memory. In some cases, when a kernel writes data that will not be used in the near future, there is no need to maintain this data in the cache (displacing useful data, and unnecessarily increasing the overhead of maintaining cache consistency).

Asking the compiler to generate NTS can therefore improve the performance by reducing pressure on caches. Based on the code, the software can recommend the use of NTS, instead of regular store operations.

Identifying inefficient memory access patterns

IRM helps us to identify inefficient memory access patterns by checking the dots for a given loop/kernel order from left to right. By default, we expect that, for a single kernel, the dots will be displayed in the following order (from left to right): CARM/L1, L2, L3, DRAM.

There are two cases where this order might not occur. Use of NTS is one. As the processor directly writes data from registers to memory, it is possible to see more data transfer on the DRAM side than at some other cache levels.

In that case, DRAM arithmetic intensity could be lower than L3 or L2 arithmetic intensities. This behaviour usually doesn't need to be fixed.

The other case is when the kernel is performing non-unit strides. If so, the CARM arithmetic intensity could be higher than L2 or L3 as every movement in any cache transfers a whole cache line, not a single value.

When this is observed, it is usually good practice to check how the data are laid out in memory and accessed. This can be done by running the Memory Access Pattern analysis (MAP).

Conclusion

Intel Advisor offers two different implementations of the roofline model that can answer different needs. The CARM is a very powerful tool for characterising an algorithm on a platform.

On the other hand, the IRM offers many possibilities to track down problems related to vectorisation, threading or memory, offering an important tool to help optimise an application. ■

Kate Antakova is a software developer at Intel Corporation, Cédric Andreolli is Intel's HPC application engineer, and Jim Cownie is a senior principal engineer at Intel Corporation UK

What is the holy grail of HPC disaster recovery?

Mahesh Pancholi, research computing specialist at high performance compute, storage and data analytics integrator OCF, explores the importance of disaster recovery solutions in HPC

Although disaster recovery (DR) has been around since the birth of computing, applying it to research computing saw its first spike in interest around 10 years ago. With the advent of the public cloud and its availability, interest in DR has been reignited and many more organisations involved in research computing are awakening to the possibilities.

The need for DR in high performance computing (HPC) has emerged alongside the increased role of CIOs in academia and research organisations, who have been recognising the value and the vital part HPC plays in keeping their organisations afloat. It is well known that the top five per cent of commercial organisations using HPC invest heavily in DR and are paying high prices for a cold cluster that sits in an offsite datacentre waiting to be used on that rare occasion that a disaster might occur.

Not only does this approach incur huge overheads, it also presents the challenge of how to make sure the cold cluster stays up to date and the data is always available.

With the availability of the public cloud, overheads can be severely reduced and organisations don't have to buy all their hardware upfront; so potentially you could have something that looks a lot like your cluster sitting in the public cloud.

In the past, research institutions have utilised software as a service offerings from major public cloud players which provide a whole range of applications running on their cloud. This allows users the facility to run the same applications in a disaster recovery scenario.

However, this approach is limited, as you may have the same applications, but you need the exact version of the application for your specific cluster.

For example, a CFD company may be assured that ANSYS or OpenFOAM is running on the cloud, but is it their actual version of ANSYS; does it have the exact libraries that it needs; and does it have the same environment? The last thing you want to introduce is an extra layer of variability, in what will be a difficult period of time if a disaster occurs.

The holy grail of HPC DR is having an exact mirror of your HPC installation, available if needed.

So the DR cluster sitting on the public cloud is exactly the same environment as your live cluster; with the same libraries and specific versions of your applications that your results are being derived from under normal circumstances.

Organisations need to carefully examine what they want from a DR service. They need to ask themselves is it a replacement service, is it an exact like for like, or is it to recover from that disaster instantly? These decisions will have a massive impact on the cost implications of putting together a disaster recovery plan.

Managing data storage in a disaster

A major challenge for HPC DR in the cloud is data storage. We've been working with a lot of our customers to shape their public cloud strategy and helping organisations to leverage public cloud storage as a tier of their storage infrastructure, together with a DR strategy.

If you already have a public cloud strategy, it isn't too hard to make sure you have the right datasets available in the right circumstances. However, it becomes more challenging if you don't already have a public cloud strategy.

It is important to consult with an experienced systems integrator who can establish that your data can move to the cloud in the event of an emergency, or more appropriately, ensure the right data is being periodically fed out to the public cloud, so when the emergency hits, you have data already in the public cloud.

Organisations need to carefully examine what they want from a DR service. They need to ask themselves is it a replacement service, is it an exact like for like, or is it to recover from that disaster instantly?

“

Benefits of mirroring

Replicating your existing HPC infrastructure in the public cloud goes beyond supporting DR, and can provide multiples of your HPC system on the public cloud for different requirements, whether DR, busting capacity during planned downtime, testing, development or expansion. For example, if a new technology or piece of software needs to be trialled, it could be tested out on the 'mirror' – HPC infrastructure on the public cloud – if it runs successfully, it could then be introduced to the on-premises HPC system.

Increased appetite for DR in HPC

Over the past five years, the costs for public cloud have become a lot more palatable, so it has become more accessible to more organisations. Over this time, the types of products being offered by public cloud platforms have also greatly expanded. Amazon Glacier is now a commonplace storage service and most know that it is the cheapest way to store your data in the public cloud, so people are now looking to take advantage of that.

HPC disaster recovery can be a reality now. Replicating your existing HPC infrastructure is something everyone who is running high performance or research computing should be looking into, because data and the ability to work with that data is what keeps an organisation running.

There are always concerns about cost implications, but it is important to consider what you really need in the event of a disaster. Identifying requirements, priority workloads and priority datasets are central to making sure that you have the appropriate DR in place, particularly for HPC. ■

Key developments of 2018

A round-up of some of the most important news stories of the last 12 months

Work starts on CERN milestone

Staff and researchers behind the Large Hadron Collider (LHC) have this year been celebrating the start of the civil-engineering work for the High-Luminosity LHC (HL-LHC): a new milestone in CERN's history.

By 2026 this major upgrade will have considerably improved the performance of the LHC, by increasing the number of collisions in the large experiments, to boost the probability of the discovery of new physics phenomena.

'The High-Luminosity LHC will extend the LHC's reach beyond its initial mission, bringing new opportunities for discovery, measuring the properties of particles such as the Higgs boson with greater precision, and exploring the fundamental constituents of the universe ever more profoundly,' said CERN Director-General Fabiola Gianotti.

The LHC started colliding particles in 2010. Inside the 27km LHC ring, bunches of protons travel at almost the speed of light and collide at four interaction points. These collisions generate new particles, which are measured by detectors surrounding the interaction points. By analysing these collisions, physicists are deepening our understanding of the laws of nature.

While the LHC is able to produce up to 1 billion proton-proton collisions per second, the HL-LHC will increase this number, referred to by physicists as 'luminosity', by a factor of between five and seven, allowing about 10 times more data to be accumulated between 2026 and 2036. This means that physicists will be able to investigate rare phenomena and make more accurate measurements. For example, the LHC allowed physicists to unearth the Higgs boson in 2012, thereby making great progress in understanding how particles acquire their mass.

The HL-LHC upgrade will allow the Higgs boson's properties to be defined more accurately, and to measure with increased precision how it is produced, how it decays and how it interacts with other particles. Supersymmetry (SUSY), theories about extra dimensions and quark substructure (compositeness) will also be investigated.

The HL-LHC project started as an international endeavour involving 29 institutes from 13 countries. It began in November 2011 and two years later was identified as one of the main priorities of the European Strategy for Particle Physics, before the project was formally approved by the CERN Council in June 2016. After successful prototyping, many new hardware elements will be constructed and installed in the years to come. Overall, more than 1.2km of the current machine will need to be replaced with many new high-technology components such as magnets, collimators and radio-frequency cavities.

The secret to increasing the collision rate is to squeeze the particle beam at the interaction points so that the probability of proton-proton collisions increases. To achieve this, the HL-LHC requires about 130 new magnets, in particular 24 new superconducting focussing quadrupoles to focus the beam and four superconducting dipoles. Both the quadrupoles and dipoles reach a field of about 11.5 tesla, as compared to the 8.3 tesla dipoles currently in use in the LHC. Sixteen new 'crab cavities' will also be installed to maximise the overlap of the proton bunches at the collision points. Their function is to tilt the bunches so that they appear to move sideways – just like a crab.

Another key ingredient in increasing the overall luminosity is to enhance the machine's availability and efficiency. For this, the HL-LHC project includes the relocation of some equipment to make it more accessible for maintenance. The magnet power converters will be moved into separate galleries, connected by new innovative superconducting cables capable of carrying up to 100 kA with almost zero energy dissipation.

'Audacity underpins the history of CERN and the High-Luminosity LHC writes a new chapter, building a bridge to the future,' said CERN's director for accelerators and technology, Frédérick Bordry. 'With its innovative technologies, it is a window to the accelerators of the future and new applications for society.'

To allow these improvements, major civil-engineering work at two sites, in Switzerland and in France, will include construction of buildings, shafts, caverns and underground galleries.

During the work, the LHC will continue to operate, with two technical stop periods to allow preparations and installations to be made for high luminosity alongside yearly maintenance activities. After completion, the LHC is expected to produce data in high-luminosity mode from 2026 onwards. By pushing the frontiers of accelerator and detector technology, it will also pave the way for future higher-energy accelerators.

US reclaims first place in TOP500

The TOP500 list of the fastest supercomputers, published in June, saw significant changes as the US took the top spot for the first time since November 2012.

Summit, an IBM supercomputer running at the Department of Energy's (DOE) Oak Ridge National Laboratory (ORNL), took top spot with a performance of 122.3 petaflops on High Performance Linpack (HPL) benchmark used to rank the list. It has 4,356 nodes, each equipped with two 22-core Power9 CPUs, and six NVIDIA Tesla V100 GPUs. The nodes are linked with a Mellanox dual-rail EDR InfiniBand network.

Sunway TaihuLight, a system developed by China's National Research Center of Parallel Computer Engineering & Technology and installed at the National Supercomputing Center in Wuxi, dropped to second after leading the list for the past two years. Its HPL mark of 93 Pflops has remained unchanged since it came online in June 2016.

Sierra, a new system at the DOE's Lawrence Livermore National Laboratory was third, with 71.6 petaflops on HPL. Built by IBM, Sierra's architecture is similar to that of Summit but uses less GPUs per node and slightly different clock speeds on the Power 9 CPU. Each of its 4,320 nodes are powered by two Power9 CPUs plus four NVIDIA Tesla V100 GPUs and using the same Mellanox EDR InfiniBand as the system interconnect.

Tianhe-2A, also known as Milky Way-2A, moved down two notches to fourth, despite receiving a major upgrade that replaced its five-year-old Xeon Phi accelerators with custom-built Matrix-2000 coprocessors. The new hardware increased the system's HPL performance from 33.9 petaflops to 61.4 petaflops, while bumping up its power consumption by less than four per cent. Tianhe-2A was developed by China's National University of Defense Technology and is installed at the National Supercomputer Center in Guangzhou, China.

The new AI Bridging Cloud Infrastructure (ABCI) is fifth, with an HPL mark of 19.9 petaflops. The Fujitsu-built supercomputer is powered by 20-core Xeon Gold processors along with NVIDIA Tesla V100 GPUs. It's installed at the National Institute of Advanced Industrial Science and Technology in Japan.

Piz Daint (19.6 petaflops), Titan (17.6 petaflops), Sequoia (17.2 petaflops), Trinity (14.1 petaflops), and Cori (14.0 petaflops) move down to number six through to 10 on the list, respectively.

While the US has retaken the top spot, overall the country's total number of systems featured in the list has fallen to 124 – the lowest since the TOP500 began. In the previous list published six months previously, the US had 145 systems.

However the US has increased its share in the performance category. Systems installed in the US now contribute 38.2 per cent of the aggregate installed performance, with China in second place with 29.1 per cent. Much of this change in aggregate performance comes from the introduction of Summit and Sierra at the top of the rankings.

China improved its representation to 206 total systems, compared to 202 on the last list. The next most prominent countries are Japan, with 36 systems, the United Kingdom, with 22 systems, Germany with 21 systems, and France, with 18 systems. These numbers are nearly the same as they were on the previous list.

For the first time, total performance of all 500 systems exceeds one exaflop, with the total systems delivering 1.22 exaflops. This has increased from 845 Pflops in the November 2017 list. While this is an impressive milestone, the increase in installed performance is well below the previous long-term trend shown until 2013.

The overall increase in installed capacity is also reflected in the fact that there are now 273 systems with HPL performance greater than one petaflops, up from 181 systems on the previous list. The entry level to the list is now 716 teraflops, an increase of 168 teraflops.

Big, bigger, biggest

Also in June, Hewlett Packard Enterprise (HPE) announced its collaboration with Sandia National Laboratories and the US Department of Energy (DOE) to deliver the world's largest Arm supercomputer.

As part of the Vanguard project, Astra, the new Arm-based system, will be used by the National Nuclear Security Administration (NNSA) to run advanced modelling and simulation workloads for addressing areas such as national security, energy and science.

In today's data-intensive environment, there is an increasing demand for higher compute performance as organisations conduct research-intensive tasks that require processing and analysing large data sets to address challenges across medicine, climate change, space, and oil and gas exploration.

'By introducing Arm processors with the HPE Apollo 70, a purpose-built HPC architecture, we are bringing powerful elements, like optimal memory performance and greater density, to supercomputers that existing technologies in the market cannot match,' said Mike Vildibill, vice president, Advanced Technology Group, HPE. 'Sandia National Laboratories has been an active partner in leveraging our Arm-based platform since its early design, and featuring it in the deployment of the world's largest Arm-based supercomputer, is a strategic investment for the DOE and the industry as a whole, as we race toward achieving exascale computing.'

Introducing new processors like Arm to the HPC ecosystem, which has been historically dominated by x86-based technologies, HPE is building a diverse network to offer more competitive options to power next-generation supercomputers, while accelerating the path to exascale.

HPE is also delivering memory-centric designs to support rapidly growing data-intensive HPC workloads, while enabling greater density with more performance-packed servers by bringing robust, Arm-based HPC technologies to power the Astra supercomputer and future systems.

With Astra, a major stepping stone in HPE's path to exascale, HPE is delivering over 2.3 theoretical peak Pflops of performance, 33 per cent better memory performance than traditional market offerings, and greater system density. The NNSA, an agency within the DOE that is responsible for the management and security of the nation's nuclear weapons, nuclear non-proliferation, and naval reactor programs, will advance modelling and simulation tools to improve analysis on data-intensive science experiments.

'We are committed to fielding the most advanced technologies as part of the Vanguard program,' said James Laros, Vanguard project lead at Sandia National Laboratories. 'We are collaborating with HPE to advance the Arm ecosystem and prove the viability of this architecture to support our national security mission.' ■

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 PBS Works™ is the trusted leader in HPC workload management, with powerful products that simplify job submission, scheduling, analytics, remote visualization and cloud computing capabilities. For over 30 years, Altar has delivered HPC and engineering software/services to 5000+ customers in multiple industries. Altair's 2600+ employees serve clients over 71 offices in 24 countries.

CoolIT Systems

10 – 2928 Sunridge Way NE,
Calgary, Alberta, T1Y 7H9, CANADA
Tel: +1 403 235 4895 Fax: +1 403 770 8306
sales@coolitsystems.com
www.coolitsystems.com

CoolIT specializes in scalable liquid cooling for the world's most demanding data centers. Through its modular, rack-based Direct Liquid Cooling technology, Rack DCLC™, CoolIT enables dramatic increases in rack densities, component performance & power efficiencies. Offering Coldplate Loops for the latest high TDP processors & high-capacity Coolant Distribution Units, CoolIT's reliable technology installs into any server or rack, ensuring ease of adoption & maintenance.

Panasas, Inc.

969 W. Maude Ave.
Sunnyvale, CA 94085, USA
Tel: + (408) 215-6800
info@panasas.com
www.panasas.com

Panasas is the premier provider of high-performance storage solutions that deliver the raw speed and enterprise-grade reliability required for compute-intensive and complex workflow environments. The fully integrated Panasas ActiveStor® scale-out NAS appliance features a plug-and-play parallel file system, client protocols, and flexible components that easily adapt to dynamic business needs.

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.

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

GIGABYTE Server supplies highly competitive, precision engineered servers. With almost 20 years' track record in the industry, GIGABYTE has built extensive experience in developing leading-edge technology products based on a focus on design and materials.

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.

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.

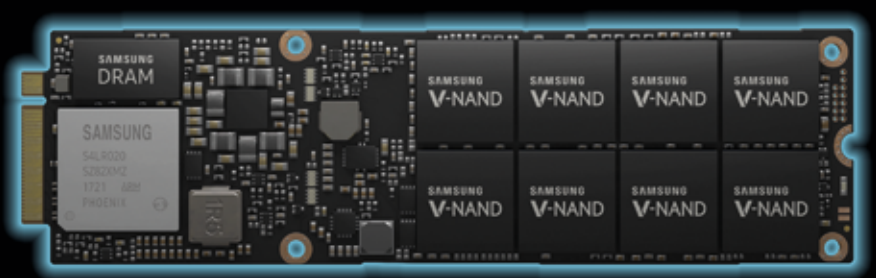
Configure yours at:
> b2b.gigabyte.com

SAMSUNG

Rethink Storage

Small shape, huge capacity

Despite its tiny size, our new PM983 NVMe SSD allows you to pack more storage capacity in your servers than ever before.



Available in capacities ranging from **4TB to 16TB** for up to 0.5PB in slim 1U server designs.
PM983 is depicted in its real NF1 form factor size (110mm x 30.5mm).

www.samsung.com/semiconductor

ISC EVENTS
GOLD SPONSOR
ISC 2018

PM983 NVMe SSD