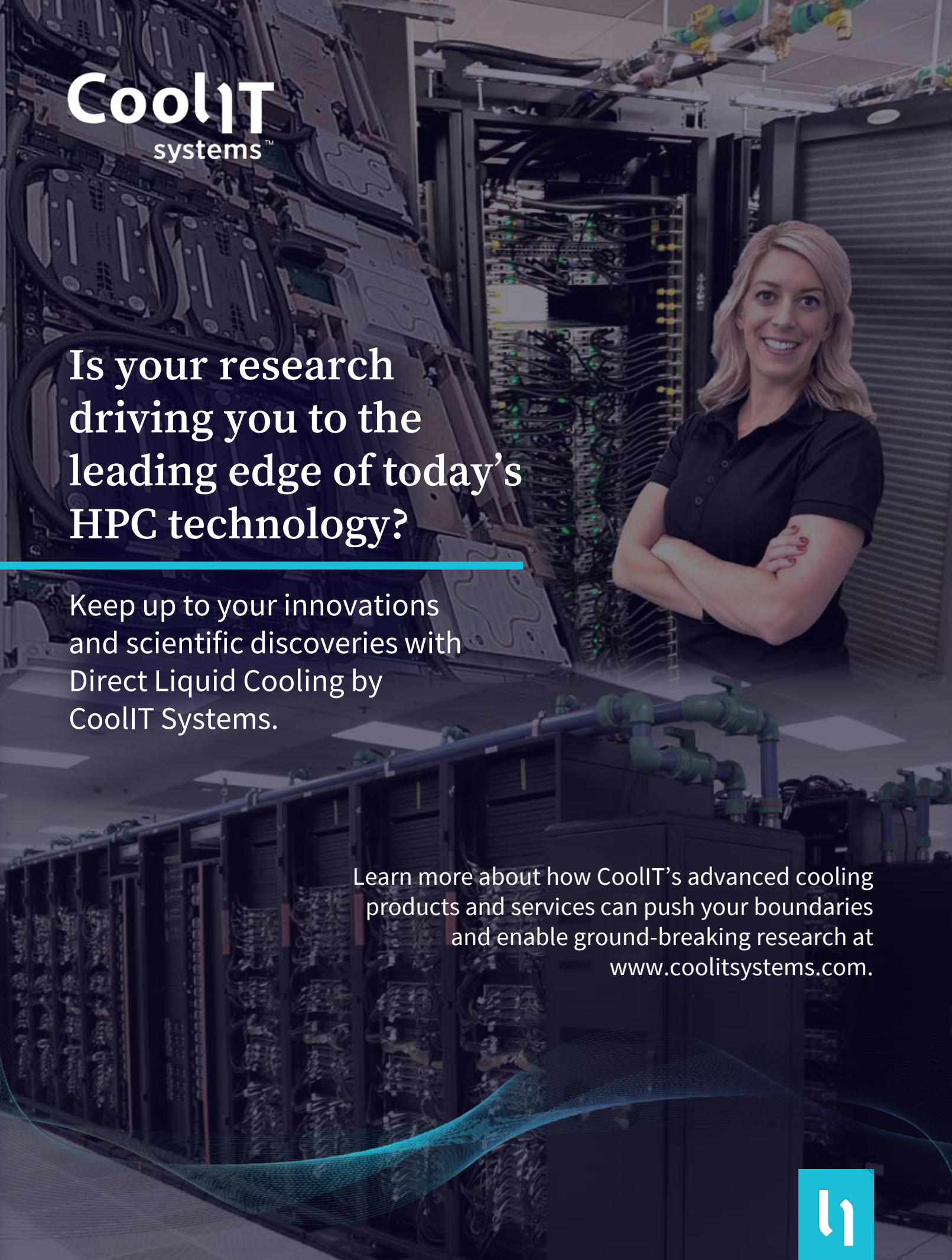
HPC Yearbook 2022/2023

Our annual review of high-performance computing for scientists and engineers

www.scientific-computing.com

From the publishers of
SCIENTIFIC COMPUTING WORLD


CoolIT
systems™
Is your research driving you to the leading edge of today's HPC technology?
Keep up to your innovations and scientific discoveries with Direct Liquid Cooling by CoolIT Systems.
Learn more about how CoolIT's advanced cooling products and services can push your boundaries and enable ground-breaking research at www.coolitsystems.com.

CONTENTS

Welcome to High-Performance Computing 2022/23




This edition of the HPC Yearbook aims to cover several key themes across the HPC ecosystem. Firstly, it would be remiss not to mention that the first exascale system was officially recognised this year. Frontier leads the charge of exascale systems coming from the US Department of Energy's National Laboratories over the next few years. On page 4, we will explore a small sample of the hardware developments that underpin exascale development and look at how future technologies, such as silicon photonics, might push the envelope even further in the coming years. There is the first of our expert interviews on page 8, with the European Processor Initiative's Ettienne Walter discussing the critical importance of chip security in establishing European HPC.

On page 10, we have a feature exploring the convergence of HPC and AI technology and the impact that is having on scientists and researchers. Following that is a feature looking at the explosion of quantum computing. While some way off from delivering real-world science, there is growing interest in the huge potential of quantum computing.

Exascale is the subject of discussion once again on page 18 as ECP's Project Director, Doug Kothe, discusses the lasting impact of exascale development. This interview also focuses on the software tools that will help to empower the next generation of HPC users. On page 22, we have another interview with DiRAC Deputy Director Clare Jenner and DiRAC Training Manager Richard Regan. They highlight the comprehensive training programme developed to support DiRAC users.

Last, but by no means least, on page 25, we have a contributed article from Pistoia's Anca Maria Ciobanu discussing the use of machine learning to help treat rare diseases.

Robert Roe
Editor

EDITORIAL AND ADMINISTRATIVE TEAM
Managing Editor:
Annabel Ola (editor.scw@europascience.com);
Editor:
Robert Roe (editor.scw@europascience.com)

ADVERTISING TEAM
Senior Account Manager:
Lexi Taylor (lexi.taylor@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:
Justin Zwierzanski (justin.zwierzanski@europascience.com)
Tel: +44 (0)1223 221 035

CORPORATE TEAM
Managing Director: Warren Clark
Head of Content: Mark Elliott

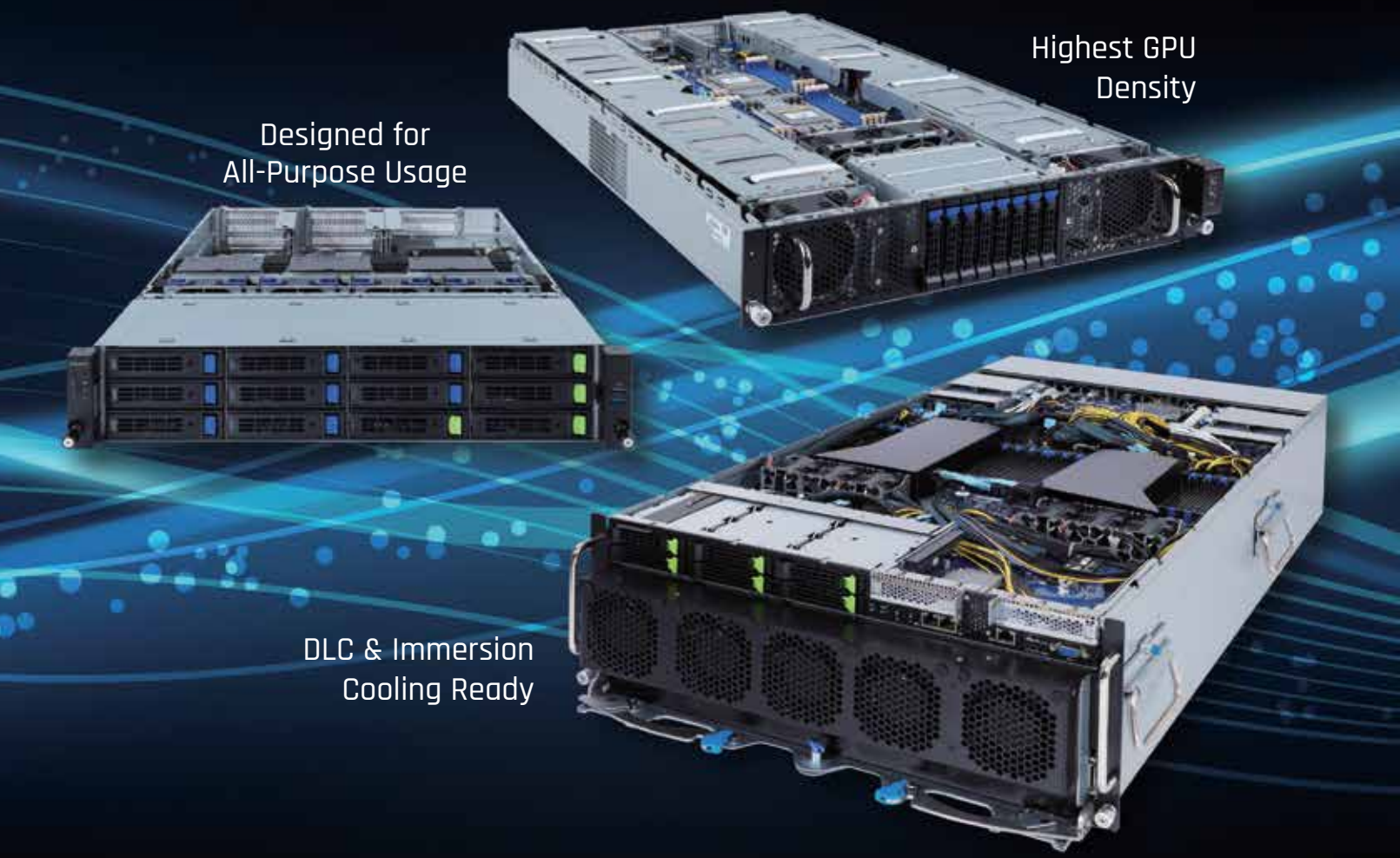
SUBSCRIPTIONS: HPC 2022-23 is published by Europa Science Ltd, which also publishes Scientific Computing World. Free registration is available to qualifying individuals (register online at www.scientific-computing.com). Subscriptions £180 a year for six issues to readers outside registration requirements; 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 image and all other images: Shutterstock.com

GIGABYTE™

NVIDIA



GIGABYTE NVIDIA GPU Servers

The Market's Largest Selection!

Discover Full Product Range

Web: gigabyte.com/Enterprise | Email: server.grp@gigabyte.com

Pete Hansen/shutterstock

Innovation at scale

Robert Roe explores advances in CPU, accelerator and networking hardware that is being designed to support exascale

Exascale has been a long sought-after goal for high-performance computing (HPC) because it represents the next order of magnitude in computing performance since the first petaflop systems were announced more than 10 years ago.

In 2022, the Frontier supercomputer, located at the Oak Ridge National Laboratory, was the first system to break the exaflop barrier. In the June 2022 edition of the *Top500 list*, Frontier recorded an HPL benchmark score of 1.102 exaflop/s. This makes Frontier not only the most powerful supercomputer ever to exist – it's also the first true exascale machine.

While there have been significant and wide-ranging advances in technology during the development of the first exascale machines, the elusive exaflop has been out of reach until 2022. However, in that period, the types of computation conducted on these systems also changed. There has been a shift towards Graphics Processing Unit (GPL)-accelerated systems to deliver exascale performance and a rise in Artificial Intelligence (AI) applications driving more mixed-precision performance, such as FP32 and FP16, rather than FP64 (double-precision floating-point format), which has been traditionally used in HPC.

Satoshi Matsuoka, Head of the Riken Center for Computational Science, location of the second-fastest supercomputer in the world, discusses the role of mixed-precision performance in modern HPC systems: "People sometimes contest us when we say that we offer the first exascale supercomputer. But what do we mean by exascale? Well, there are several definitions. If you think exascale is the FP64 performance, then an exaflop would be represented by the peak performance, or achieved LINPACK performance, and, of course, for Fugaku, this is not the case. Its RMAX (maximal LINPACK performance achieved) is 0.44 exaflops."

Matsuoka continues: "However, very few applications correlate with FP64 absolute dense matrix linear algebra performance in this context. So this may not be a valid definition when you really think about the capability of a supercomputer.

"The second possible definition is any floating-point precision performance that is bigger than an exaflop or a metric from some credible application. In that respect, Fugaku is an exaflop machine, because, for example, in HPL we achieved two exaflops. However, Oak Ridge National Lab's Summit machine has achieved two exaflops in the Gordon Bell-winning applications. So, although Fugaku is an exascale machine by this definition, it was not first," adds Matsuoka. "I think the most important definition when we started thinking about these exascale machines was to achieve almost two orders of magnitude speed-up, as >

> Very few applications correlate with FP64 absolute dense matrix linear algebra performance in this context

compared to the current state-of-the-art in 2011 and 2012, when we had 10- to 20-petaflop supercomputers.

"As I have demonstrated, Fugaku is about seven times faster across the applications than the K computer, an 11-petaflop RMAX machine. And because of the 'application-first' nature of the machine, we believe this is the most important metric. We have achieved two orders of magnitude speed-up over our last-generation machine, which you would call a 10- to 20-petaflop machine. Being application first, this was the most important – and, in this context, we have achieved what was expected out of the exascale machine," Matsuoka concludes.

In a presentation from the ISC high Performance Conference, Lori Diachin, Project Deputy Director for the US Exascale Computing Project (ECP), discussed the work going on in the US to prepare for the first wave of exascale-class machines. Diachin also touched on some of the challenges that the ECP faces in delivering three exascale systems for the US national labs.

"The ECP has three primary technical focus areas. The first is application development. In this area, we have selected 26 design centres, which are focused on the computational motifs that many applications can take advantage of – adaptive mesh refinement type kernels, high-order discretisation kernels, particle methods, and so on," said Diachin. "The second major technical area is software technologies. We are working very hard to develop an integrated software stack, so that applications can achieve their full potential on the exascale computing platforms. We have 71 different products that are focused on runtime systems, programming models, math libraries, data visualisation tools and the software ecosystem writ large.

"Finally, we have our hardware and integration area, which is focused on the delivery of our products on the DoE facilities. It's also been focused on partnerships with six of the US HPC vendors to look at the research and development that was needed for excess scale nodes and system design."

Specialised challenges

Whereas Fugaku was designed around Central Processing Unit (CPU)-only hardware to ensure the general purpose nature of its applications suite, the US exascale systems rely on large numbers of GPUs to deliver performance. This has thrown up additional challenges for the ECP – not only because they must cater for GPU technology, but also because there are several vendors and competing technologies being used in the various US exascale systems. The three systems will eventually use GPUs from Nvidia, AMD and Intel – and this means that the ECP has to optimise for all three technologies.

"One of the things that's very interesting in these systems, and that has been driving a lot of the work within the ECP, is the fact that the accelerator node architecture is changing from being an Nvidia-only ecosystem to an ecosystem that has a wider variety of GPUs, in particular, the exascale systems will have AMD, Intel GPUs… and that's driving a lot of our work for performance portability," explained Diachin. "In our application portfolio, we have some 24 applications that were chosen because they were of strategic importance to the Department of Energy," she continued.

"They range in topics from national security applications, such as next-generation stockpile stewardship codes, to energy and economic security scientific discovery, Earth System modelling and healthcare in partnership with the National Institute of Health (NIH). With these 24 applications in the six co-design projects, we have more than 50 separate codes that have more than 10 million lines of code collectively. When the project began, many of these codes were focused on MPI or MPI plus open MP – and were largely focused on CPU with a small number of them starting their work with GPU-accelerated computing. Since the beginning of the project in 2016, each code has had to develop a unique plan to make progress toward deployment on the excess scale systems, moving away from CPU-only implementations to performance portable GPU implementations. When we talk about preparing applications for the exascale systems, there is a hierarchy of adaptations that have needed to happen, so what do we need to do at the lowest levels to rewrite and optimise our codes," said Diachin. "There"s been a lot of data layout that has been done: loop reordering, kernel flattening, and so on. It's really focused on those lower-level operations that can really improve the performance in different kernels of the application."

Exascale networking

CPU and accelerator technologies are just one aspect of the exascale puzzle. As the parallelism in exascale systems will be vastly larger than anything seen today, it will be a significant challenge to deliver the I/O bandwidth needed to support application performance. This is not just a challenge for exascale HPC, but other markets including AI and traditional datacentre applications.

In 2022, Hewlett Packard Enterprise and Ayar Labs signed up for a multi-year strategic collaboration to really accelerate the networking performance of computing systems and data centres by developing a range of silicon photonics solutions based on optical I/O technology.

This was soon followed by the news that Ayar Labs had secured $130m in additional funding from Boardman Bay Capital Management, Hewlett Packard

> The accelerator node architecture is changing from being an Nvidia-only ecosystem to an ecosystem that has a wider variety of GPUs

Enterprise (HPE) and Nvidia, as well as multiple new and existing financial investors including GlobalFoundries and Intel Capital.

Silicon photonics will be used to enhance the networking capabilities and support future requirements for HPC, artificial intelligence and cloud computing architectures.

The technology also now has the climate-critical potential to reduce the amount of energy used in data centres and large computing systems.

Hugo Saleh, Ayar Labs Senior Vice-President of Commercial Operations, explains: "Within the press release, we talked about a few things. One is a future design of HPE Slingshot architecture, which has its genesis back at Cray. Today, it is their high-end, ethernet-like, networking solution that is targeted for HPC.

"We're also working with HPE on advanced architectures, where we're now talking about the composability of disaggregated resources with an intelligent software stack."

Solving a range of problems for extreme-scale HPC

The partnership between HPE and Ayar Labs is aimed at developing capabilities that leverage optical I/O, which is a silicon photonics-based technology that uses light instead of electricity to transmit data, to integrate with HPE Slingshot or other future networking products yet to be developed.

"Whether you're talking about HPC, or disaggregated computing, there is a real limiter on I/O," explains Saleh. "In HPC, it's usually referred to as a memory bottleneck. It's not a memory capacity issue; it's the ability to move the data out of memory DIMMs into the CPU and back.

> "Whether you're talking about HPC, or disaggregated computing, there is a real limiter on I/O"

"The other bottleneck that's been seen and talked about quite a bit, of course, is the bottleneck on the GPU: between the CPU and GPU transferring the data and then again, between the GPU itself and the memory."

These bottlenecks are a growing concern for any scientists and researchers that are using HPC and AI systems as they have the potential to limit application performance.

"What we are doing at Ayar Labs is an attempt to change the physical domain that data is transmitted in," notes Saleh. "Going from electricity, voltages and currents, to photons. And we do that coming straight out of the socket. So it's not a transceiver at the back of the server and it's not a mid-board optics. We design chiplets that sit inside of the package, that are nearly abutted to the CPU, memory, GPU or accelerator. We're agnostic to the host ASIC. Then we transmit photons and light outside of the package for your high-speed, low-power I/O."

Ayar Labs first demonstrated this technology at Supercomputing 2019, the conference and exhibition that is held annually in the United States. "We have a full test rig. We first demonstrated our technology to the HPC community at that Denver event in 2019," says Saleh.

"Since then, we have made two public announcements that are the projects we're doing with Intel. So Intel has demonstrated an FPGA with our photonics inside of it, transmitting massive amounts of data at much lower power."

This technology could massively increase the memory bandwidth for future HPC and AI systems. Each chiplet delivers the equivalent of 64 PCIe Gen 5 lanes, which provides up to two terabits per second of I/O performance. The system uses standard silicon fabrication techniques, along with disaggregated multi-wavelength lasers to achieve high-speed, high-density chip-to-chip communication with power consumption at a picojoule range.

Ayar Labs developed its technology alongside Global Foundries as part of its monolithic silicon photonics platform.

"We worked with the Global Foundries on developing a monolithic process; one that lets you put electronics and optics on the same chip," says Saleh. "A lot of traditional optics are separate; we have it all combined into one and that simplifies our customer's life when they're packaging all these components – it reduces power, it reduces costs and reduces latency."

GF Fotonix is Global Foundries' next-generation, monolithic platform, which is the first in the industry to combine its 300mm photonics features and 300GHz-class RF-CMOS on a silicon wafer. The process has been designed to deliver performance at scale and will be used to develop photonic compute and sensing applications

Ayar Labs also helped Global Foundries develop an advanced electro-optic PDK, which was released in Q2 2022, and will be integrated into electronic design automation vendor design tools.

Connect world/shutterstock

Securing chips for the EU

EPI's **Ettiene Walter** discusses the development of Europe's HPC processor and the importance of securing chip security

The European Processor Initiative (EPI) aims to design and implement a roadmap for a new family of low-power European processors for exascale and the wider HPC community, as well as big data and a range of emerging applications.

EPI focuses on two concurrent areas of development. The first is a general purpose processor (GPP) based on Arm IP, and the second is an accelerator technology based on the RISC-V instruction set architecture (ISA).

The EPI project builds on years of development in other European projects. The Mont Blanc projects, for example, which started back in 2011 and finished in 2020, provided the groundwork to develop software tools and prototype Arm-based systems for use in HPC.

The development of the Arm HPC ecosystem, specifically the development of the Scalable Vector Extension (SVE) and all of the work done by Arm and Fujitsu on the A64 processor used in RIKEN's Fugaku supercomputer, has further bolstered the use of Arm in HPC.

> "We need to secure our ability to source chips, and I think it's a global concern"

How has the progress of Arm in HPC impacted the development of European processors?

Ettiene Walter, EPI Phase Two General Manager: If you look at the entire picture, then it's more than 10 years in development. And, if you look at Japan, what they've done with Fugaku is the same – they also started on ARM. They have a sound system, but it's been 10 years of work. And so it's been a long story to develop this ecosystem that is stable and mature around the Arm ISA. Now we can say it's there – but this was not true five years ago.

Arm has come a long way in this time, from small testbed systems to large HPC systems where there is an installed user base. There is a lot of work being done on this, like SVE, to make this a platform that's suitable for HPC. And we can expect RISC-V to catch up, but it's still on the way; it's not at the same level yet.

Arm is known for delivering great energy efficiency, particularly with mobile or embedded processors. Is that why EPI chose this technology?

Yes – even if we are not at the same level. We are not providing chips for mobile systems. They have a long history and provide good benefits, but there is also a side-benefit of choosing

the Arm architecture. Arm provides an ecosystem of IP blocks; you can use these blocks in a similar way to Lego and bring things together. You can develop your own processor but the gap to deliver that solution can be more narrow. The customisable, configurable nature of the arm IP, and the system they've developed, make it easier to build from scratch – as opposed to traditional processor design that would be much more complex without the pre-existing IP blocks.

What can you tell us about manufacturing?

Now it's more of a side-aspect of technological development, rather than an architectural option. But to get good performance or energy efficiency, we need to deliver chips at a very small scale – at a nanometre scale. And to do that is complex. The smaller you go, the more complex it is. Added complexity comes due to the fact that we are not on the same level on the two different branches of development for EPI. The GPP will be based on the state-of-the-art process at TSMC. I prefer not to give any precise figure; it will be up to SiPearl to discuss when they will be able to.

> "The customisable, configurable nature of the Arm IP, and the system they've developed, make it easier to build from scratch"

We are not far from the state-of-the-art in the foundry for the accelerator. We are doing some test ships right now. This means that we are not at the industrial scale, but we foresee an improvement in the scale with the next generation of accelerators.

We need to secure our ability to source chips, and I think that's a global concern. We see that, for instance, with Covid-19 – to get access to Pfizer or other vaccines. We have the same potential issue in the automotive industry, with the chipsets that are lacking. Some cars couldn't be built due to the lack of processors, so this is very much a global issue.

We have to improve and be prepared for the future. And this is why I believe it's important for any country, but, for instance, also at the European level, to have some means to be sure that we can secure or source HPC chips. For me, there are two sides to this; one being that we need to be able to design and define our processors. The other point is the fact that we need to be able to make them, so we need the foundry; we need the factory. And we need to secure our suppliers – it's not one or the other; we need both.

Realising the potential of AI and HPC

Three experts in the field share their thoughts about the potential of an increased combination of AI and HPC. By **David Stuart**

The convergence of artificial intelligence (AI) and high-performance computing (HPC) promises to transform the scientific computing landscape with its potential to enable research groups to tackle challenges that would otherwise have been beyond their capabilities.

Over the past decade, we have begun to see AI penetrate nearly all industries and scientific disciplines, from the headline-grabbing integration in autonomous vehicles and the protein-folding predictions of AlphaFold, to the more quietly heralded work managing traffic flows, creating more efficient jet engines, and removing the noise from astronomical images. This is, undoubtedly, only the beginning – especially as AI is increasingly combined with the processing power possible from HPC. This is necessary for dealing with the very large data sets that are being made available, or may need to be simulated, and the complexities of deep-learning models.

There are, however, many challenges still to be overcome. These include challenges inherent to AI, challenges to the integration of AI in HPC, and challenges to successfully transfer knowledge to the people who need it.

The black box of AI

That AI is a black box technology, with the scientist unable to know what is going on inside, is one of the challenges that still faces the application. As Michèle Weiland, Senior Research Fellow at EPCC, the supercomputer centre at the University of Edinburgh, puts it: "If you are dealing with a black box, and you are getting the right answer, you don't necessarily know the reason why you are getting the right answer; you may just be lucky."

The problem of discerning why you are getting the answer you are getting can be exacerbated as the barriers to technologies are lowered and they start to be used by those who do not necessarily recognise the inherent limitations of AI.

Andreas Lintermann, of the Jülich Supercomputing Centre, and Coordinator of CoE RAISE (The European Center of Excellence in Exascale Computing – Research on AI – and Simulation-Based Engineering at Exascale), explains: "It's a new field for many domain scientists. They have to find out what models can be used, what architectures for neural networks make sense, and what kind and amount of data is suited. There's a danger you simply plug in a model that you believe works for your problem. It might generate some output that looks very promising but, in the end, it might not be as accurate as you think. These models hide very complex processes from the user and there are always questions – do the models always do what the user, or the scientist, expects them to do, and do they generate an output with a sufficient accuracy to reliably reason from?"

It is important that scientists don't blindly trust these models, rather there is a need for 'explainable AI', with a human who can explain and verify the AI outputs. Vikram Saletore, Principal Engineer, Super Computer Group at Intel Corporation, cites the example of interpreting the inference of a CT scan that might identify a tumour.

> Scientists are deservedly sceptical of AI replacing modelling and simulation

Explainable AI incorporates a radiologist who uses their expert interpretation to ensure that the identified region is indeed a tumour.

He says: "Scientists are deservedly sceptical of AI replacing modelling and simulation. To be useful, this technology must deliver demonstrably correct results to prevent the introduction of non-physical artifacts for a user's simulation, and provide benefits, such as faster inference performance."

Changing workflows

It's not just that AI is a black box, but the speed of development and workflows are also very different to what many scientists are used to. Weiland explains: "A lot of these machine learning tools are quite a long way removed from what computational scientists use in their day-to-day life. Scientists often use monolithic applications that they know inside out, and they are often quite old – 10, 20 or even 30 years old – and they have grown over time. These new applications have been very dynamic over recent years, so getting to grips with the different technology and the fast-changing landscape is really quite a challenge."

In the case of HPC, there are also significant differences in the way people are expected to interact with the systems, with command prompts to a computer in a data centre rather than graphical user interfaces on their own PC. There is also a need to rethink how the code is processed; it's not just the same code run faster. As the scientist moves from small data to big data, it becomes necessary to parallelise the code – so data can be analysed in parallel rather than sequentially, which could otherwise take years.

The challenge for the scientist is to break down the task into smaller parts without suffering from the overhead of the communication between the different parts, which inevitably can impact some use cases more than others, as Lintermann explains: "The problem is split into sub-problems, where each of the processes that live in a high-performance computer works on its own sub-problem.

"For example, if you want to simulate the fluid mechanics in a complete room, the problem is usually too big to be computed on a single processor and the room is split in such a way that each processor takes care of the computation in only a small fraction of the complete volume of the room. If you now, at one point in the room, initiate a pressure wave that travels through the room, it also needs to travel across these different sub-volumes with the consequence that the information from one processor needs to be transported to another processor. This communication is always a bottleneck when scaling from a small to a large number of processors.

"For a fixed problem size, you want to make sure that, if you use more resources, that you achieve a result faster. Usually, if you have the same size problem and use more processors, you would continually decrease the time to solution. However, as the volume sizes and the corresponding number of elements per sub-volume decrease, the communication share increases – and, at some point, using more processors does not make the computation faster anymore. This is usually when the scaling ends."

This task of optimisation is not one that the domain scientist is likely to have spent too much time thinking about previously – but, as Lintermann points out, it doesn't just relate to simulation but must also be explored for the optimisation of AI where, ideally, models can be trained faster by using more processors.

As Lintermann goes on to explain, simulations and AI can be part of combined full-loop implementations. A simulation can produce a lot of data, and these data are – maybe already at simulation run time – used to train artificial neural networks that may, as surrogates, directly be plugged into the original simulation. The same loop can iteratively be run to continuously optimise not only the simulation but also the surrogate model. This is something CoE RAISE is working on, for example, in the context of hydrogen combustion and its integration in aircraft engines.

Changing technology

At the same time, HPC hardware is also changing, and this is partly because of the influence of AI. As Weiland puts it: "The AI and HPC convergence is driving changes in hardware designs for HPC systems. Traditionally, HPC systems are designed for monolithic scientific >

Den Rise/shutterstock

> “This AI and HPC convergence is driving changes in hardware designs for HPC systems”

computing applications, numerical codes that don't do a vast amount of high-throughput IO; they predominantly read and write large files. HPC systems don't necessarily have the required flexibility with the workloads. AI has slightly different requirements on the hardware, and system designs will more and more reflect these changes, such as providing different types of file systems for different types of applications. Systems designs will have to vary a little bit, change and adapt to accommodate all the workloads equally. That's beginning to happen, but it'll actually happen more as these AI and HPC converge more."

Changes in hardware are increasingly possible because HPC systems are becoming more and more modular, consisting of different components. Some components are more suited for some tasks than others, and there are always new components coming along. Saletore speaks of how some of Intel's latest technologies are helping with both AI and HPC.

Saletore explains: "To meet the needs of our customers, the latest Intel technology is raising the bar in AI and HPC computing. XPUs provide an AI-enabled general-purpose hardware platform that incorporates high bandwidth memory; Xe-HPC accelerators provide a large number of cores and massive parallelism coupled with high bandwidth memory and interconnect – all of which are important to fast time-to-solution in a distributed HPC environment.

"General-purpose, next-generation Intel Xeon Scalable processors (code name Sapphire Rapids) are now architected with new instructions for AI including Advanced Matrix Extensions (AMX) and Tile matrix MULtiply (TMUL) ISA extensions. Some of these Intel processors incorporate High Bandwidth Memory (HBM). Intel Optane persistent memory with Distributed Asynchronous Object Storage (DAOS) have revolutionised storage performance to address the data handling issues."

As the acceleration of hardware capabilities continues, it becomes increasingly important that software doesn't become locked-in to the hardware, and Intel's oneAPI ecosystem is an important part of that, enabling scientists and organisations to quickly pivot to the fastest and most cost-effective hardware platforms. As Saletore puts it: 'OneAPI is an important initiative for HPC as it is open standards-based and is delivering performant portability."

Overcoming the challenges

The key to overcoming some of the challenges faced is bringing experts in the different domains together. Only then, by ensuring that science and industry get access to the facilities and technologies that they need and can use, do we have safeguards for the goal of scientific progress being achieved as quickly as possible.

There is an increasing focus at a European level on easing the process by which researchers and scientists can make use of HPC and AI. For example, EuroCC, a network of 33 international partners around Europe, is designed to identify knowledge gaps and identify HPC competencies to help with access to HPC for researchers. Similarly, the CoE RAISE project, funded by the European Commission under the Horizon 2020 Framework Project, is tasked with developing scalable AI technologies towards exascale with use cases from engineering and natural sciences. Examples of use-cases include the optimisation of the surface of aerofoils to reduce drag and increase lift, identifying potential porosities or weaknesses in metal additive manufacturing, and windfarm layout optimisation.

Lintermann explains the importance of getting this collaboration right, ensuring sufficient understanding between the different experts, and some of the steps taken in CoE RAISE: "There has to be a general understanding – what is the problem for a specific domain, and what do the domain scientists want to solve? We found out that the creation of so-called factsheets is very promising. The computer scientists, the AI experts, the HPC experts, and the domain experts gather to jointly draft the problem set-up in an understandable way and to carve out what would be necessary to solve the problem. This starts a discussion on how people can work together, how the AI people can contribute to solving the problem, and what HPC can do in reducing the time to solution.

"This all happens in an interaction room, which is an online live meeting room, where the people meet and have mural boards at their disposal to draw and add any information, images, and explanations. This has proven to be very helpful in equalising the language among the disciplines and helps people to communicate."

Conclusions

The potential of combining AI with HPC is phenomenal, although of course that doesn't mean it is suited to every task in scientific computing – and part of the challenge is to identify where it can be best put to work.

As Weiland explains, with reference to numerical computing: "In the most promising approaches, people are looking at those parts of the computation that can be replaced by machine learning, that don't necessarily influence the outcome, but can accelerate the getting to the solution by providing a better initial guess. Techniques where you effectively replace an entire numerical model or scientific computing model with AI, that's not ever going to be a solution, because AI is largely a black box. You can replace sections of scientific computing with machine learning, bits that matter for the performance, but don't matter for the solution."

Finding those places where AI and HPC can work – and ensuring that it works to the best extent possible – requires successful collaboration between AI experts, HPC experts and domain experts, and there are now many success stories in the field.

Saletore says: "Success stories and ready accessibility to advanced technology have raised awareness among scientists about how AI augmentation and even replacement can expand what computer models can do. This promotes new thinking and the freedom to really think big, as scientists pursue the creation of more accurate models and explainable AI models. Scientists are now realising that they can do research that just was not possible before."

Projects such as EuroCC and CoE RAISE can be seen to be playing an important part in this across Europe.

White Papers
now available online

VIEW FOR FREE*

Flexible, scalable and sustainable: the new power behind tomorrow's weather predictions
ATNORTH

The world's leading weather-forecasting companies, usually process weather data four times a day, which, depending on the speed of execution, can take around two hours at a time.

CoolIT Systems Direct Liquid Cooling Supports The Most Advanced HPC Research
COOLIT SYSTEMS

This whitepaper discusses the benefits of adopting liquid cooling for your HPC data centre and how the technology supports scientific research with high performance cooling solutions for the latest supercomputing systems.

Engineering reaches for the cloud
RESCALE

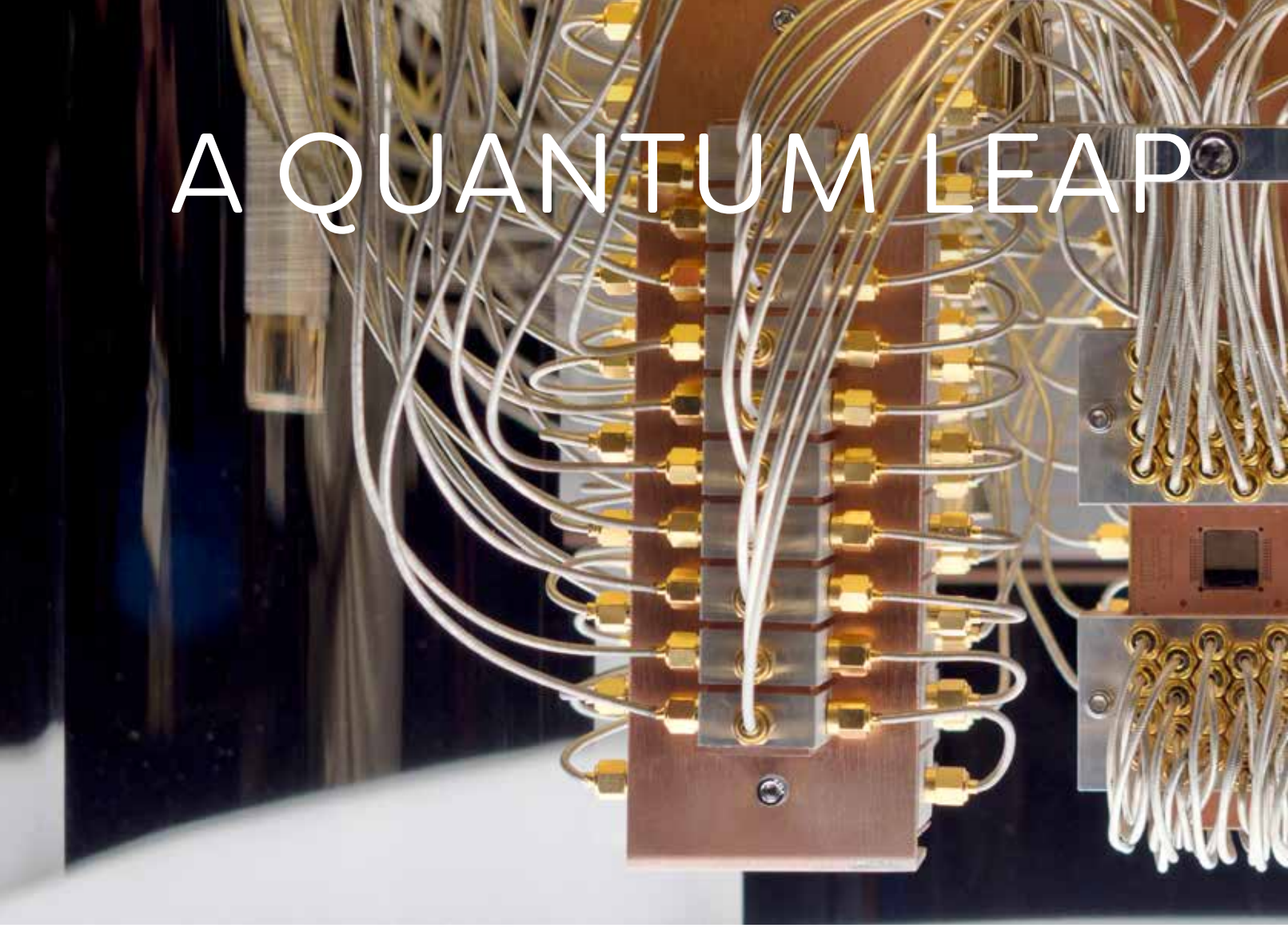
Computational engineering is set for strong growth – and the use of the cloud is burgeoning

*Registration required

www.scientific-computing.com/white-papers

SCIENTIFIC COMPUTING WORLD

A QUANTUM LEAP



Quantum technology is going through a period of rapid development, with several technologies driving the adoption of this emerging computing framework, finds **Robert Roe**

One of the biggest stumbling blocks in the development of quantum hardware is with the qubits themselves. This varies depending on the underlying technology used to create the qubits, but they are often error-prone and difficult to control, making quantum computers unstable and highly complex systems.

Advancing the technology requires larger quantum computers that can be scaled up and integrated with the cloud or existing classical computing systems. Scale is therefore of paramount importance in delivering real-world scientific insight.

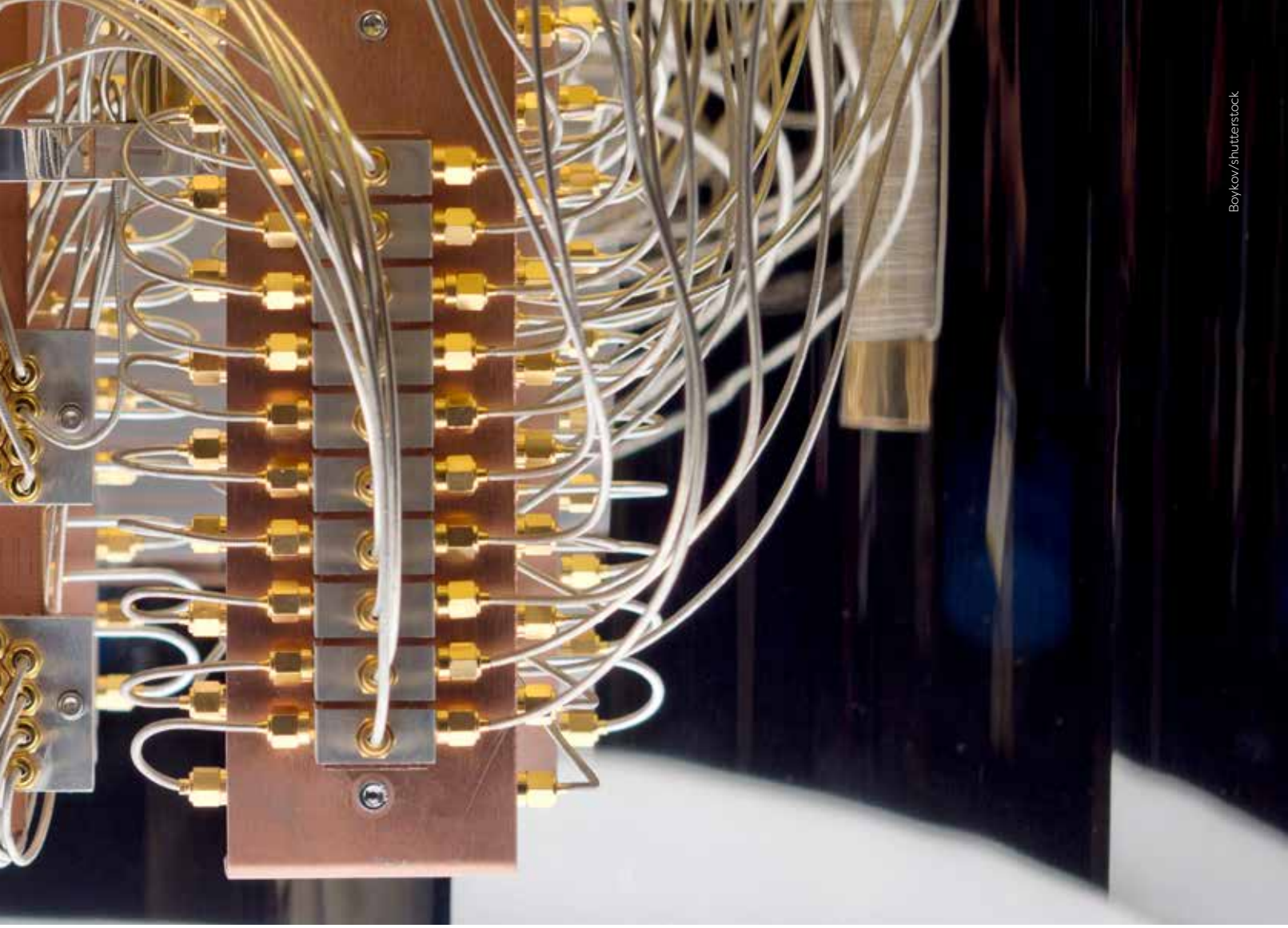
There are many ways to build these systems, depending on the type of technology used. Universal Quantum, for example, is trying to develop the world's first million qubit quantum computer using a technology called "trapped ion".

Dr Luuk Earl, Quantum Engineer at Universal Quantum, highlighted the path the company has taken since its inception in 2018. "It's a company that spun out from a research group at the University of Sussex. Two senior scientists, Professor Winfried Hensinger and Dr Sebastian Weidt, decided that the research they were doing was promising for quantum computing. They formed this company with a bit of venture capital funding.

"The real aim is to make quantum computers that can solve real-world interesting problems," Earl continues. "That is the main difference between other quantum computing companies and us at the moment. We're just focusing on that big-scale stuff. We're not interested in making toy models that can do some interesting science but aren't going to impact humanity. We're focused on that point where quantum computing is useful to everyone, which is a big challenge."

That is why a million qubit system is such an important goal, as this is the point at which some scientists and researchers believe that quantum computing systems will start to impact science and engineering. Earl notes scientists at Universal Quantum had carried out modelling of the resources required to solve particular problems.

"One of them is synthesising a particular chemical for fertiliser. They've done some simulations of how many resources you need to do the quantum chemistry for that to make an efficient process. And that kind of



Boykov/shutterstock

> The real aim is to make quantum computers that can solve real-world, interesting problems

comes out around a million qubits.

"Of course, there's always more to the story," stresses Earl. "We trade off three things in a real-world application. One of them is the number of qubits, which is, of course, always important.

"There are also the error rates and coherence time. So if you have really good error rates, you can probably do fewer qubits; if you have long coherence times, you can do with a few less qubits. So it's a bit of a rough number. But that's the kind of order of magnitude where quantum computers become really interesting."

Trapped ion quantum systems

Universal Quantum is developing its quantum computers based on trapped ion technology. A simple explanation would be that ions are trapped and precisely controlled using electromagnetic fields. Each ion levitates above the surface of a silicon microchip. The idea behind trapped ion systems is that the ions are relatively easy to control, as they are all precisely the same shape and size.

"A trapped ion system is what you'd initially think of if you think of a quantum computer," says Earl. "The qubits you use are naturally quantum systems; an atom is the closest thing we can come to a single quantum bit. Whereas a superconducting qubit is kind of an analogy to a qubit.

"The big benefit is that every qubit is identical because every atom is identical to every other atom; the energy levels are very well defined – and, especially with trapped ions, we can control the position and environment of those atoms very precisely," Earl continues. "By tuning electrode voltages, you can move the qubits around the surface of a chip very precisely; this means we have ultimate control over what these qubits are doing and how they're interacting."

Another important consideration that goes into developing all of these systems is the availability and price of the components that are used to control the qubits. For example, in the Universal Quantum system, the company uses lasers and microwaves to control the qubits.

Earl notes it is important to focus on developing systems with technology that is available today. "From the work – at Sussex and the preliminary work – we've done at Universal Quantum already, we think we can build million-qubit machines using technology that already exists. We're all very >

> “A lot of the algorithmic work is in imagining this future where you don’t have to worry about qubits and how they interact”

passionate about making real-world impactful systems as soon as possible. The quickest way to do that is using technology that already exists.”

Quantum accelerators

Founded in 2019, Quantum Brilliance has a very different take on the development of quantum computing. The venture-backed company develops quantum computers using a diamond substrate to help boost the reliability of the qubits and increase coherence time. The goal of Quantum Brilliance is to enable mass deployment of quantum technology to propel industries to harness edge and supercomputing applications.

The first generation of the company’s technology has already been installed in Paswey Supercomputing Centre, which is exploring how this technology might be used alongside high-performance computing (HPC) systems in the future.

Mark Mattingley-Scott, Managing Director, EMEA for Quantum Brilliance, explains why the company opted for this radically different quantum technology: “What diamond does is give you coherence for free,” he says. “So you get qubits if you make them in diamond the right way. They maintain quantum coherence, even at room temperature. What it means is that all the stuff you have to do with other quantum computing technologies, such as keep it cold, or ensure it’s under a really high vacuum, or use precise lasers to get photons aligned – all those things fall away.”

This is an important distinction from other quantum systems, as it means the quantum brilliance prototype systems can be smaller and more easily integrated with existing computing systems. They operate at room temperate and do not require compex systems or advanced cooling.

“I was actually at the Pawsey Supercomputing Centre this afternoon and I saw our quantum computer,” Mattingley-Scott says. “It’s a 6u device, so it’s a little bit higher than the standard 19-inch rack unit. That’s the first generation of machines. Inside this device is a small piece of diamond with the qubits on it, and some simple optoelectronics to interact with that – stuff you would find in a 5g antenna mast. We’re working on miniaturising that, and we’re pretty sure we can get down to a graphics-card-size accelerator within the next few years.

“Once you’ve got something like a graphics accelerator, then you’re in the same world as a normal graphics processing unit (GPU) or tensor processing unit (TPU),” Mattingley-Scott adds. “You can start to put these things in large quantities in a standard compute environment.” Another benefit to diamond-based quantum systems is they can be used for edge computing systems, such as in robotics and autonomous vehicles.

This is relevant to high-performance computing and research centres in general, because it allows quantum computing to be more easily integrated with classical computing architectures. This could help drive the adoption of the technology and allow more scientists and researchers to get access to quantum technology.

But scaling these systems to the point of mass adoption is still some way off. There are significant challenges facing today’s quantum computing developers. One significant challenge is getting to the point where there are enough qubits to provide a measurable performance improvement in some way. The second challenge is how to integrate a quantum computer with classical computing. While Quantum Brilliance wants to connect quantum processing units (QPUs) to classical systems directly, some other organisations want to connect these systems via the cloud.

However, Mattingley-Scott thinks this is a mistake: “Most companies are using the cloud, so they’re looking at a cloud hybrid execution model. We believe – and I think history bears this out – the QPU needs to be physically as close to the other classical compute devices, like CPUs and GPUs, as possible.

“We envisage a future in which you’ll go into your computing centre and pull a blade out – maybe it’ll be a CPU blade, maybe it’ll be a GPU blade, or maybe it’ll be a hybrid, and it’ll hopefully be a Quantum Brilliance QPU sat next to AMD or Nvidia or Intel CPUs and GPUs.

“Quantum computing must operate in a quantum-classical hybrid – it has to be the case,” Mattingley-Scott continues. “If you talk to almost all the hardware vendors, there will not be isolated quantum computers churning away doing stuff, and then delivering their results, at least for the foreseeable future – the next few decades. It is all going to be hybrid.

“In which case, just bite the bullet and put your QPU actually in an accelerator card... next to the GPU, next to the CPU. And then you’re not worried about data throughput, latency times and interaction times,” Mattingley-Scott concludes.

Quantum cloud

Quantinuum, on the other hand, is a company that has embraced the use of the cloud to help facilitate access to its prototype quantum systems. Quantinuum’s H1 generation of quantum computers is already commercially available. The Quantinuum H1 generation, currently consisting of two computers, the H1-1 and the H1-2, are fully accessible over the cloud and compatible with a variety of software frameworks.

Tony Uttley, President and Chief Operating Officer at Quantinuum, highlights the company’s growth from both hardware and software provider: “Quantinuum is the combination of Cambridge Quantum with Honeywell Quantum Solutions,” he says.

“Honeywell Quantum Solutions did a lot of work directly with the products that Cambridge Quantum Computing was making.

“What we found, as we were working together as separate companies, was that most people who are developing hardware will extrapolate away from the metal layer,” Uttley explains. “They will make a separation to protect IP, and you can’t get the full integrated benefit if you have that separation layer. We realised we could make fully integrated solutions based upon both the application layer on top of the middleware on top of our hardware.”

However, although the platform is based on the integration of two distinct companies, they also choose to make the software platform inclusive. “The

applications, the operating system that we develop, is designed to work on everybody's hardware," Uttley says. "And, as a practical example, we are one of the biggest users of IBM's quantum computers in the world. IBM is also an investor in Quantinuum."

While the hardware stack continues to mature, scientists and researchers are now getting access to software-development tools to create quantum algorithms and quantum simulators, or emulators, that allow them to simulate how a quantum computer might work in a classical system. This allows researchers to develop expertise and test out how applications might benefit them in the future.

"A lot of the algorithmic work is in imagining this future where you don't have to worry about qubits and how they interact, because all of that has been 'taken care of' by universal fault tolerance," says Uttley. "That's a decade away. So the key is: what do you do in the intervening time? How do you make progress? Can you do things with some of these earlier systems? And the answer is: yes you can. However, this requires users to begin to think about their problems differently," stresses Uttley. "What I mean by that is: don't think about what the problem is, and how you abstract that into the system. You need to think about what these systems are good at. And how do I use that for these kinds of problems?"

Not all qubits are created equally

One critical aspect of the development of these quantum systems, particularly in the early days where coherence and error rates make these systems relatively unstable, is that different hardware architectures are more suited to different problems. A simple example of this would be a large number of qubits with a low coherence versus a much smaller number of qubits with a high coherence time. Another factor is the amount of communication required between qubits.

"If you're trying to simulate a molecule, then ostensibly it depends on how that molecule is shaped, believe it or not," Uttley says. "This is because what happens in a superconducting quantum computer and semiconducting ones are similar; where they have an architectural property that's called 'nearest neighbour'. This means the qubits are physically manufactured on a piece of silicon."

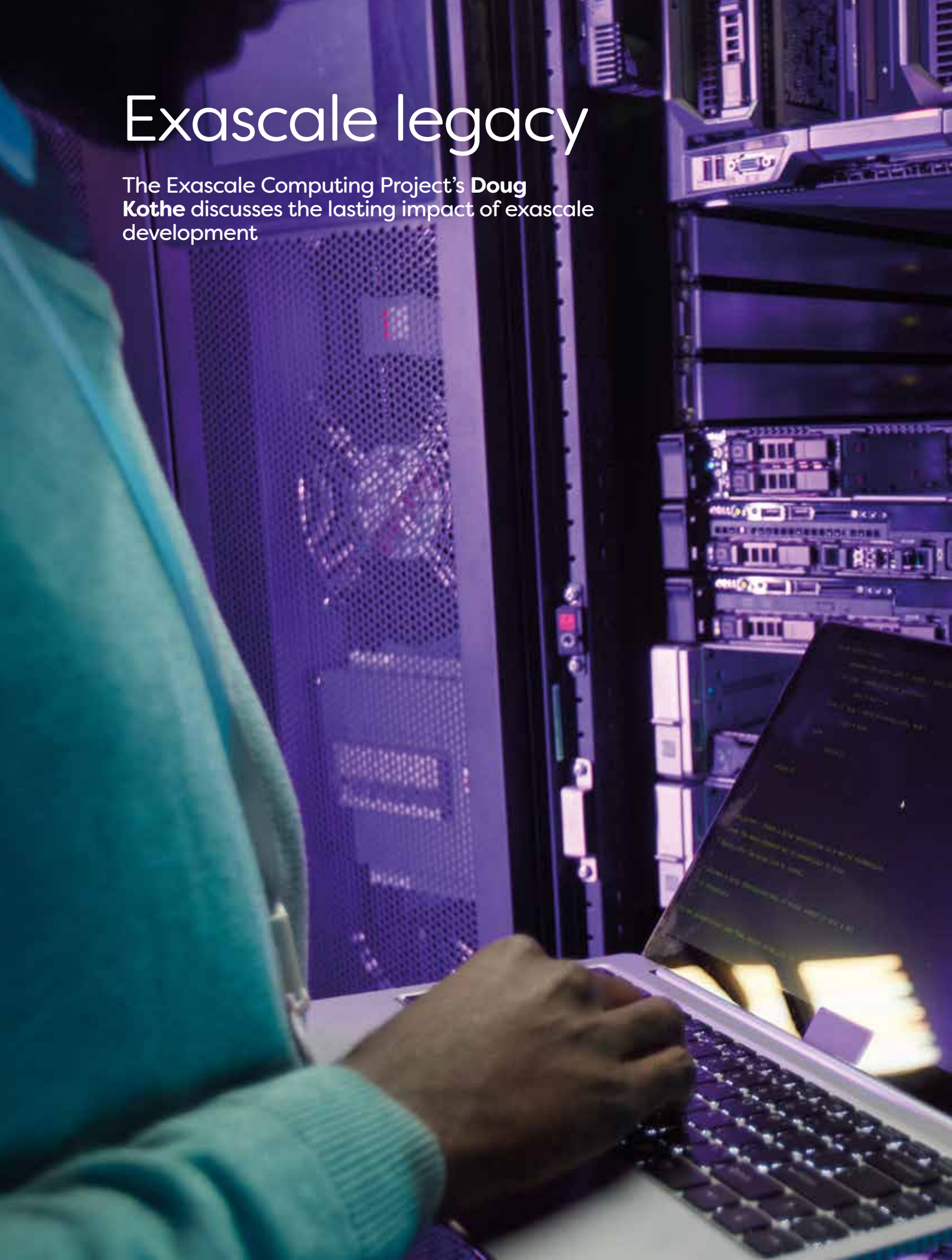
Uttley gives an example of a grid, or several rows of qubits, where the qubits can easily communicate with their nearest neighbour, but where communications from one side of the grid to the other take much longer and adds additional error to the system.

"There are molecules where the shape of the molecule itself is kind of a nearest neighbour interaction," Uttley continues. "A nearest neighbour molecule running on a nearest neighbour quantum computer actually can work pretty effectively. But if you have complex molecules, where you need these qubits to talk to every other qubit arbitrarily, then our trapped ion hardware works well. This is because we can physically transport our qubits so that any one can talk to any other one without introducing any additional error.

"It's that kind of deep knowledge about both the problem, and the way these systems work, that allows us to know which hardware or platform will be most suitable for a given problem," Uttley concludes.

Exascale legacy

The Exascale Computing Project's **Doug Kothe** discusses the lasting impact of exascale development



SeventyFour/shutterstock

Can you tell our readers about yourself and your role at the Exascale Computing Project (ECP)?

In terms of my current role, yes; I'm the Director of the Exascale Computing Project. It is a seven-year, almost $2 billion project that started in 2016. I joined the project at the beginning, actually a year before in 2015, after I had led a five-year effort to build what we call a virtual nuclear reactor – essentially a new, high-fidelity simulator of operating reactors. That experience was something that positioned me well to run ECP.

AT ECP, I initially stood up all the applications in the project, and there are 24 of them, plus some other efforts. For the first two years, Paul Messina at Argonne National Lab was the inaugural Director of ECP. I actually took over in the autumn of 2017, so I've been on point for five years. I'm also dual-hatted here at Oak Ridge National Lab. I'm now the Associate Laboratory Director for our Computing and Computational Sciences Directorate (CCSD). My passion is in building applications to address problems of worldwide interest.

How will exascale development impact the wider HPC community?

The Advanced Scientific Computing Research (ASCR) Office in the Office of Science and the Advanced Simulation and Computing Office, part of the National Nuclear Security Administration (NNSA), both realised that they needed to make a concerted investment in software. As application and software developers, we just couldn't agree more. Often, these sorts of investments are driven by the passion and commitment of the scientists. In other words: in the past, we had not developed this area as much as we should.

At ECP, our sponsors realised that they needed to make a heavy investment in software, many years before the arrival of a new system. This huge investment in software – that's not just boutique software for one system, but is essentially the software tools and technology – will develop the scientific and engineering tools for our nation. And – in many cases – the world, for decades to come.

A good application can live for decades, through many systems. It's essentially like constructing a large scientific instrument. So, in this case, I like to think of our applications as the beginning of a new app store for the nation. And our software stack is a new dynamic OS. This stuff is going to be around long after I retire. In a similar way, the DOE will build a large neutron source or large light source to be a nationwide scientific instrument.

> "At ECP, our sponsors realised that they needed to make a heavy investment in software, many years before the arrival of a new system"

You say that investment has changed? What is different about the ECP?

At the Department of Energy, there have always been investments in software development – but generally, at least in my career, never in such a concerted and integrated way. By concerted I mean with ample investment for innovation, agility, and trial and error.

We're all about agile software development – which means, frankly, sometimes you fail early and often. The other thing is to bring all the activities together under one roof. There were some growing pains and culture clashes initially, but that has given us a huge return on investment.The fascinating thing with ECP is that we put together this huge project, and we've got around 85 different teams working with each other. We developed this inherent co-dependency upon each other that I've never seen before. We are working with our sponsors right now to ensure that this co-dependent ecosystem is sustained well beyond ECP, and I'm very confident that's going to happen.

Can you give an example of this codependency?

At ECP, we were afforded the opportunity to build integrated teams – and, in some cases, we forced it >

> “A good application can live for decades through many systems”

because certain domain scientists work in their own bubble. That doesn't mean they are not very successful, but the whole thing about ECP is bringing people together. This has led to substantial dividends. As an example, we were building some abstraction layers, one out of Sandia called Kokkos, and one out of Lawrence Livermore called RAJA, that helps to demystify and, to some extent, hide the complexity of heterogeneous hardware.

Many of our application teams didn't know about that development, because they were at other labs or other institutions or, in some cases, didn't care because they didn't think it was going to help them.

You highlighted heterogeneous hardware. How has this trend affected the development of exascale?

Accelerators are here to stay. And I'm going to call it an accelerator, not a GPU, simply because what we're seeing now is hardware designed to accelerate certain operations. GPUs are probably mis-named because they are designed for graphics and to accelerate integer operations. But they do darned well with integer, logical and floating point operations, which is where scientific computing comes in.

ECP recognised that accelerated node computing is here to stay – whether it's your laptop, your desktop, a cluster down the hall, the cloud or Frontier, a node is going to be an eclectic mix of hardware.

If you don't have software that recognises how to lay out data, and how to utilise that hardware to exploit all these floating point operations, you're going to be in trouble.

I don't want to imply that this is easy going from one piece of hardware to another. The whole porting exercise is a contact sport. But the point is, we've designed our software to compartmentalise the pieces that we know can be accelerated. they have been separated with certain data structures that are more amenable to these accelerators. It doesn't mean that we might have to make some mods or changes.

But, in many cases, we have re-architected the software to be much more agile and flexible. To give you an example, the Kokkos abstraction library from Sandia National Lab basically handles all your data structures for you. You might say: “I would like a three-dimensional, 64-bit, floating point array”. I hand that to Kokkos and it looks at the hardware and says: “Okay, I know how to do this; I know how to lay it out”. To some extent, application developers aren't used to this. The mindset can often be: “I can do this, I know how to do this, I don't need you”. This has become more of a co-dependent situation where you need to take advantage of the effort put into certain libraries. There's not enough time and resources for you to do it well, or at all. We're making a strong push to show folks what we've done. Not necessarily that we are finished, but here's a really good start. You can see what we've done, you can add to the stack, and there are capabilities to add components to develop it further. It's a real thrill, a really exciting development.

What do you think will be the lasting impact of the ECP?

Our Department of Energy sponsors empowered leadership, myself included, to make tough decisions about funding and integration. That doesn't mean: “here's a bunch of taxpayer funds, go forth and do great things”. We are reviewed, scrutinised and advised all the time, and we should be because we're empowered with US taxpayer funds, and we have to do the best we can. The point is, being part of a large project, we were given the flexibility to make tough decisions.

That experience has been indispensable and, as a result of that experience, we've not only developed a current next-generation group of fantastic software and application developers, but also a group of new scientific leaders. I'm excited as I watch these young kids, as I call them, go out and lead other big endeavours. I'll probably do it from the golf course, frankly, but it's been fantastic!

Doug Kothe is the Director for the Exascale Computing Project and Associate Laboratory Director, Computing and Computational Sciences Directorate, at Oak Ridge National Laboratory. ●

Timofeev Vladimir/shutterstock

The science of appliance

Discover Scientific Computing World Online

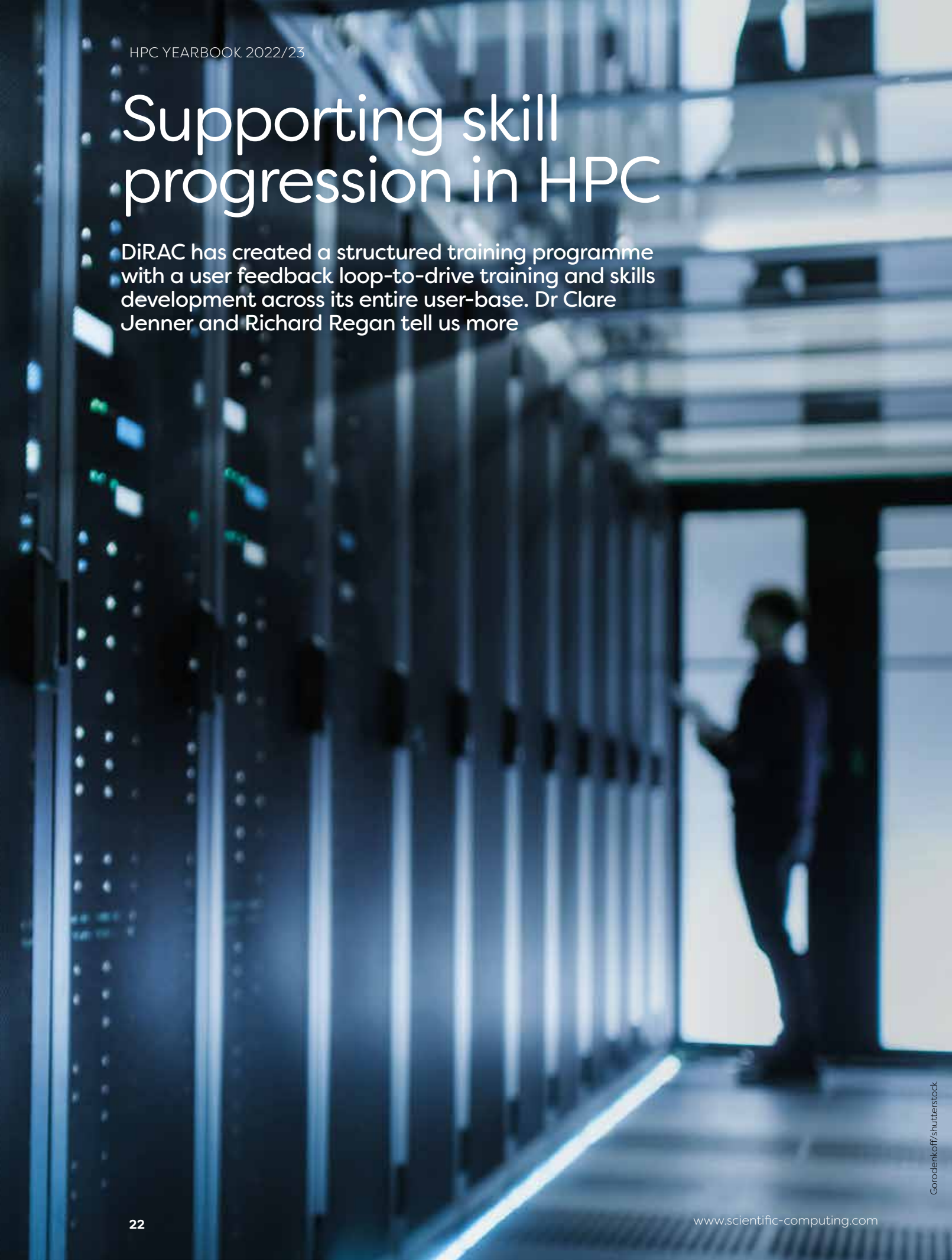
- Read news as it is published
- See industry press releases
- Refer to archived feature content
- View webcasts
- Study white papers
- Find relevant suppliers for your business
- Subscribe to the magazine, find our Twitter feed, or connect to us on LinkedIn

www.scientific-computing.com

SCIENTIFIC COMPUTING WORLD

Supporting skill progression in HPC

DiRAC has created a structured training programme with a user feedback loop-to-drive training and skills development across its entire user-base. Dr Clare Jenner and Richard Regan tell us more



Gorodenkoff/shutterstock

What can you tell our readers about the DiRAC facility and its resources?

Dr Clare Jenner, Deputy Director, STFC DiRAC High Performance Computing Facility: DiRAC is supported by the Science and Technology Facilities Council (STFC), funded through UK Research and Innovation (UKRI), and it is free at the point of use for academic researchers. DiRAC provides distributed HPC services to the particle physics, nuclear physics, cosmology, astrophysics and planetary science theory communities across all of the UK.

Due to our mixed research requirements across that very broad field, we run three different compute services hosted at four university sites across the country. Each of our resources has an architecture that is specially tailored to the different algorithmic problems our communities need to solve.

All of our users belong to an academic research project, which is run by a principal investigator based at a university in the UK. These groups will usually have codes in place that are being developed and added to by PhD students, early-career researchers and the post-docs and academics working within the team. People may come into a group with their own codes or write codes from scratch.

In terms of requesting resources, we have an annual call for time on our services, which includes a scientific justification and a technical application demonstrating that the DiRAC resource you've requested is right for the problem to be solved. You can also request research software engineering (RSE) time from our in-house RSE team if codes need to be ported or optimised for our technology.

Richard Regan, Training Manager, DiRAC: We service a broad spectrum of users; some researchers are really experienced and know what they're doing, but new PhD students, for example, might not have an awful lot of experience, having only just come from their undergraduate programme, and these users need a lot of training to make the best use of our services. Some users may have already worked on codes on local systems or their local clusters and have limited experience of HPC or have GPU skills. Still, it is up to us to build on those skill sets so they can be translated quickly into solid scientific output.

> “Some new users need a lot of training to make the best use of our services”

The first step for new users is the Essential HPC-Skills training programme. Can you explain how that was set up? What has changed over time?

Dr Clare Jenner: The training programme started back in 2011 with the creation of one of the UK's first HPC-Skills tests – the driving licence. DiRAC was a much smaller service then and the objective of the test was to make sure everybody had the skills they needed to use the machines.

As the programme has evolved, we have created a set of supportive online materials to go with the test by trawling the web to find appropriate existing training resources and putting together a set of selected links on our website. We start off with a basic set of essential skills by covering the Unix environment, command scripts, version control, software engineering and testing and code scaling, for example. These are the very basic skills that aim to expose users to what they need to have confidence and a good understanding of HPC when they start to work on their code on our systems.

The skills that we cover were pinpointed by asking the project investigators what skills they thought their new users should have when they joined their teams and what they would like them to have when they left, and this helped us to create a progression path through the skill set.

The training programme is continually evolving to meet the user's needs and now we're moving away from linking to other people's training materials and developing a new suite of material covering the same basic curriculum – but developed specifically for DiRAC users using DiRAC resources. This is a combination of bespoke, instructor-led classes and self-taught materials, and we're going to launch that course to our users early next year.

What opportunities are there for skills development after the initial programme?

Richard Regan: As I mentioned, we've got a lot of users with different skill sets and the Essentials programme is just the first step where we give them the key skills to get onto our systems and to get their research started.

Once they've got those skills, we then support them in their growth by giving them opportunities for other courses that are specific to our systems. We run CPU essentials, GPU taster sessions and fundamental and advanced CUDA courses. For example, in a few weeks' time, we've got a CPU course: 'AMD Induction Training', which will help users to get the best out of our AMD systems. We're looking at compilers, optimisation and profiling. But you can only get the best benefit from that training if you already understand the basics of how to use our systems. Then, these more advanced courses help researchers to develop more specific skills that allow them to take direct advantage of DiRAC's particular resources.

Dr Clare Jenner: As Richard says, we have a broad spectrum of users in DiRAC and some of them are extremely proficient at coding. They are very highly experienced, and this allows them to explore the advantages of our resources in a very in-depth way. For these users, we have our hackathon programme.

Here, we team up with our vendors, for example, AMD, Intel, or Nvidia, for a three-day event where we invite users from our core code sets to come along and sit in a classroom with vendor technology specialists and in-house DiRAC experts from our RSE and technical teams and work on their specific codes, optimising them and testing new hardware or new software. Hackathons are for our most experienced researchers – it's cutting-edge; it's an opportunity for knowledge transfer directly from the people who make the hardware to the researchers in a science project team.

More recently, we have branched into providing training on advanced applications for science for all of our users in parallel to the HPC skills. >

> Science is undergoing a data explosion and artificial intelligence (AI) and machine learning (ML) techniques are revolutionising the way that scientists tackle their research.

We now run a 'Machine Learning Techniques for Science' course, where we've collaborated with SciML, STFC's Scientific Machine Learning Group, to put on a hands-on practical course showing how to use techniques such as decision trees and neural and deep neural networks within codes.

These courses are extremely popular and are often over-subscribed with 24 hours of advertising, so we are now developing a more advanced AI/ML course tailored to our research communities as a follow-up to this more introductory course.

How do you know when to update or make changes to training opportunities?

Richard Regan: We get regular feedback from our users who have taken our training, to tell us how useful the skills training is and what they would like to see in the training programme going forward, and we survey our entire user-base annually. We also get feedback from our vendors on what technologies are coming up and if our users want to use it. Then we get together and get them trained.

Dr Clare Jenner: DiRAC also has a long track record of collaborating with industry and we also run an industrial placement programme, where we partner with companies in the industry and the public sector to give our users the opportunity to take a six-month sabbatical from their research and work on an industrial project that is of interest to both the company and the user.

> "These students come back and spread the knowledge and experience of the placement through our community"

For example, we've run these with Transport for London (TfL), where our users applied natural language processing to classify work orders, and with Guy's and St Thomas' NHS Foundation Trust, where students used ML methods to measure the effect of deprivation indicators on asthma in young adults.

These students come back and spread the knowledge and experience of the placements through our community, promoting the programme to other users. We've also recently arranged some quantum placements where our users have worked with Atos' quantum simulator at the Hartree Centre on projects looking at quantum field theory and the classification of pulsars.

As part of those placements, the students also designed and delivered a workshop to introduce other DiRAC users to the basics of quantum computing and the practical application of quantum ML skills in research. This allows us to feed direct knowledge from the placements back into the DiRAC community. This kind of feedback is important because in four or five years, when DiRAC comes to design a new set of services, we will have users who are familiar with the new technologies and can really help inform our procurement decisions.

There is much more information about our past projects and the future industrial and academic placements on our website – dirac.ac.uk.

Dr Clare Jenner is Deputy Director of the STFC DiRAC High Performance Computing Facility. Clare also directs the facility's training strategy, supporting the DiRAC user community's educational, research and innovation activities.

Richard Regan is the Training Manager for DiRAC. Richard is responsible for coordinating training events throughout the DiRAC community. He is the Principal Content Developer for the DiRAC training programme and is involved in the industrial engagement between DiRAC and its partners.

Gorodenkoff/shutterstock

Machine learning offers hope for patients living with rare diseases

Anca Maria Ciobanu discusses the use of machine learning to help treat rare diseases

As Strategic Theme Lead for Improving Effectiveness of R&D at the Pistoia Alliance, I constantly talk with different leaders from the R&D ecosystem to identify the common main challenges and create cross-company projects and programs that can positively impact the whole industry. One important topic of discussion is how to speed up the drug discovery process through artificial intelligence and machine learning.

During our recent London conference, thought-leaders gathered to discuss the important opportunity that we, as an industry, have to help patients with rare diseases find new and better treatments. By harnessing the power of AI and ML, we can speed up the discovery process and make a difference in the quality of life for families of children with rare genetic diseases – HAE or lupus, for example.

In this article, I will highlight ways in which my pharma colleagues are putting the spotlight on rare diseases by applying machine learning principles and techniques. At the same time, I would like to call on pharma and computing professionals to continue developing best practices in this area, and case studies that can move the needle on this important aspect of our industry. At Pistoia Alliance, our mission is "collaborate to innovate"; this is one of my biggest priorities in this space.

So how can machine learning point to novel treatments for this important patient population? Several pharma companies have already developed algorithms that can be used in medical applications. According to one study, published in *Orphanet Journal of Rare Diseases*, the majority of machine learning projects focus on images (32.2%), demographic data (27.0%) and 'omics' data (26.5%). Most studies used machine learning for diagnosis (40.8%) or prognosis (38.4%), whereas studies aiming to improve treatment are still relatively scarce (4.7%).

An increased number of organisations are already looking to have a more systematic approach to using AI and ML, aiming especially at improving the quality of life and giving more support to rare and orphan-disease populations.

As we heard from my colleague Bryn Roberts, of Roche, during that Pistoia Alliance conference in London, when a child is born with a potentially-fatal genetic disease, it becomes very personal to all of us, and we must collaborate for the greater good.

With the use of omics, we are able to study disease processes, down to the minute cell and genetic level, in great detail. Wearable digital devices allow clinical trial participants and HCPs to continuously monitor data. These digital tools are improving clinical trial participation because patients with rare diseases can now perform assessments at home, without the hassle of having to travel and arrange a hospital stay around their work or school schedules.

All of these advancements in pharma R&D are paving the way for a future explosion of machine learning applications, which hopefully will lead to novel drug discoveries. We have come a long way since I first focused on a rare disease in my own work.

In the past, I worked with a rare-disease association for the pediatric population with Hereditary angioedema (HAE Junior z.s.) and also with children having multiple sclerosis. In my volunteering activities during the past 10 years, I've been working with different rare-disease associations, and I saw the need for a long-term strategy for helping chronically ill people to live a normal life, not just to survive the disease or to ease down some symptoms. I strongly believe that the industry should work more closely with the patient associations even if, such as in any rare disease, there are a relatively small number of beneficiaries.

> "Several pharma companies have already developed algorithms that can be used in medical applications"

Change is coming, but it's slower than we would like due to the different barriers related with the long-term implementation strategies and use of new technologies. Most of these barriers can be overcome through collaboration and the non-competitive environment fostered by the Pistoia Alliance and its members.

For example, making data FAIR across the industry can speed up the discovery process by making the data easily shared across pharma, clinical and computing platforms.

Anca Maria Ciobanu is a strategic theme lead for Improving the Efficiency and Effectiveness of R&D at the Pistoia Alliance.

Directory of suppliers

A list of leading suppliers, consultants and integrators

whiteMocca/shutterstock

CoolIT Systems, Inc

2928 Sunridge Way NE #10,
Calgary, AB T1Y 7H9
Tel: (403) 235-4895
sales@coolitsystems.com
www.coolitsystems.com

CoolIT Systems specializes in scalable liquid cooling solutions for the world's most demanding computing environments. In the enterprise data center and high-performance computing markets, CoolIT partners with global leaders in OEM server design to develop the most efficient and reliable liquid cooling solutions for their own leading-edge products. Through its modular Rack DLC™ technology, CoolIT enables dramatic increases in rack densities, component performance and power efficiencies.

GIGABYTE

server.grp@gigabyte.com
www.gigabyte.com/Enterprise

GIGABYTE offers over 30 years of engineering expertise and success stories in the development of server and workstation solutions that cover a myriad of use cases. From high-density CPU computing to HPC, AI, Virtualization and exclusive liquid-cooled servers, GIGABYTE is poised to offer top performance, product quality, and engineering resilience, all built on GIGABYTE's constant attention to customer needs for delivering best-fit solutions.

atnorth
Steinhella 10. 221
Hafnarfjörður. Iceland
+354 539 3282
https://atnorth.com

Boston
Unit 5, Curo Park,
Frogmore, St.Albans,
Hertfordshire, AL2 2DD
United Kingdom
+44 (0)1727 876 100
sales@boston.co.uk
www.boston.co.uk

DUG Technology
3rd Floor, 192–198 Vauxhall Bridge Road
London SW1V 1DX,
United Kingdom
Tel: +44 20 7290 5380
sales@dug.com
www.dug.com

Motivair
5900 Genesee St.
Lancaster, NY 14086
USA
+001 (716) 691-9222
info@motivaircorp.com
www.motivaircorp.com

Nortek Data Center Cooling
8000 Phoenix Pkwy, O Fallon,
Missouri, 63368, United States
+001 (405) 263-7286
www.nortekdatacenter.com

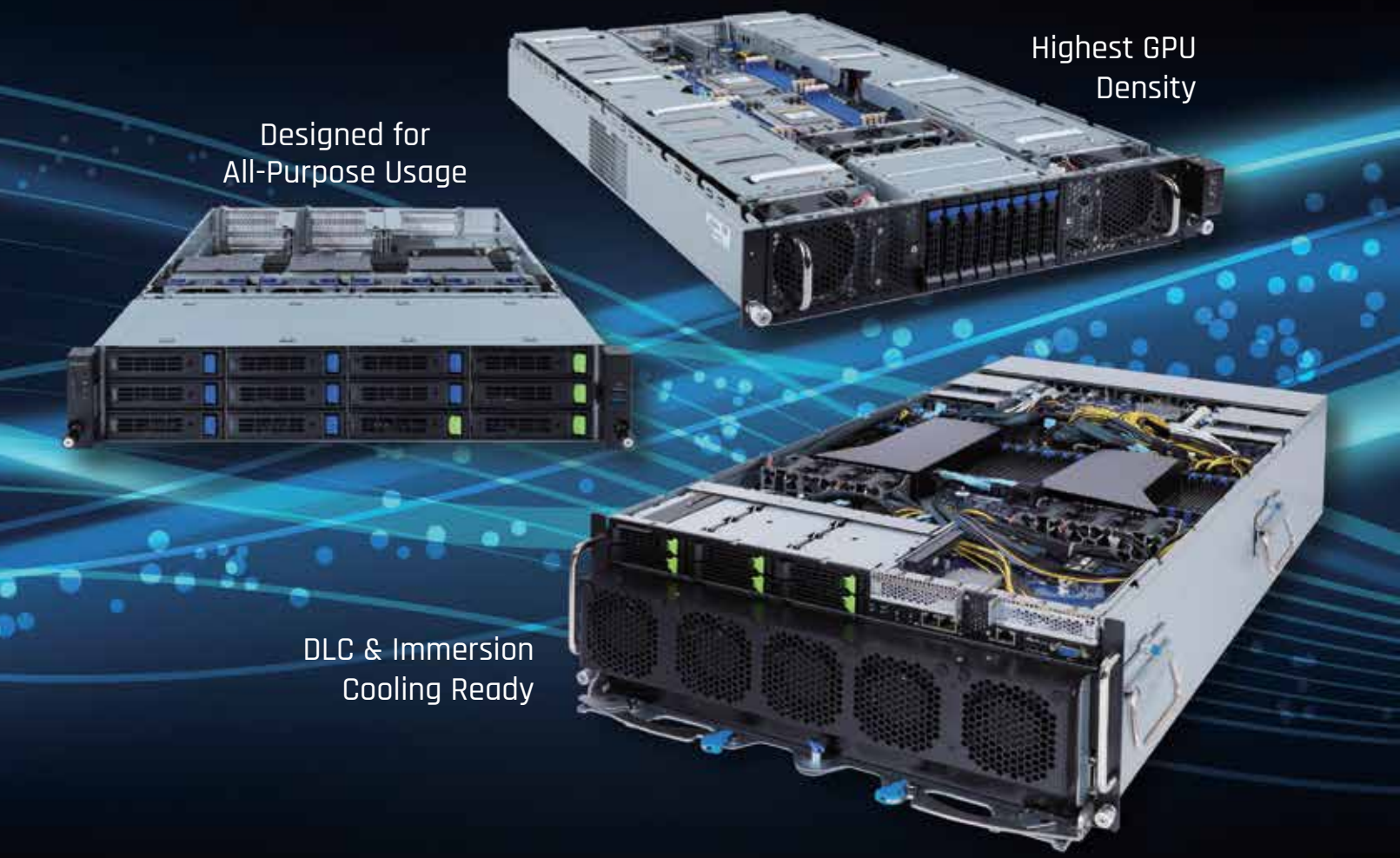
OCF
Unit 5 Rotunda Business Centre,
Thorncliffe Park, Chapeltown,
Sheffield, S35 2PG
+44 (0) 114 257 2200
support@ocf.co.uk
www.ocf.co.uk

SoftIron
Level 1, Devonshire House
One Mayfair Place
London W1J 8AJ
United Kingdom
+44 (800) 368 8646
info@softiron.com
https://softiron.com

Rescale
33 New Montgomery St.,
Suite 950
San Francisco, CA 94105
+001-855-737-2253
https://rescale.com

GIGABYTE™

NVIDIA



GIGABYTE NVIDIA GPU Servers

The Market's Largest Selection!

Discover Full Product Range

Web: gigabyte.com/Enterprise | Email: server.grp@gigabyte.com