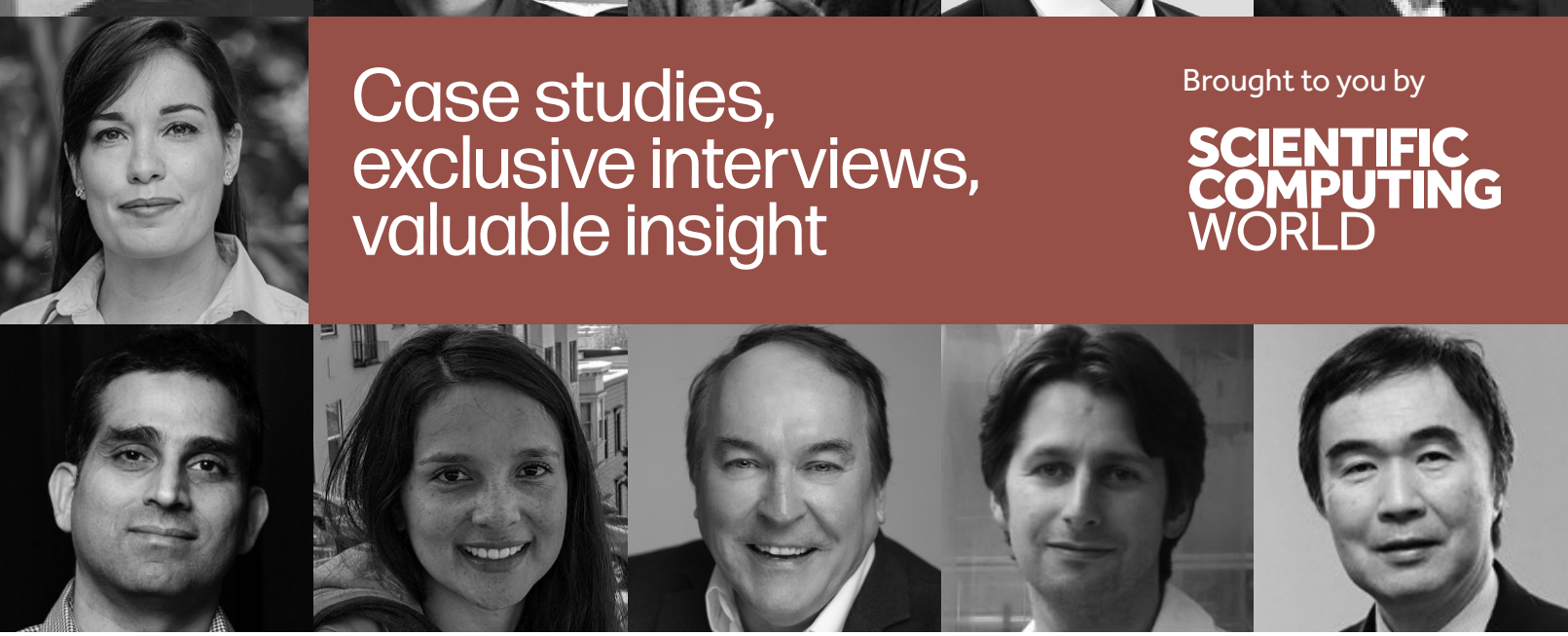
HPC Breakthroughs 2026

Unleashing scientific discovery with the power of computing

Case studies, exclusive interviews, valuable insight

Brought to you by

SCIENTIFIC COMPUTING WORLD



Including

Beyond Moore's Law: Pushing computing limits in 2026
Exascale Engineering: Solving the toughest engineering challenges
Brain-Inspired Computing: GPUs, CPUs, and neuromorphic tech converge
Causal AI vs. Malaria: HPC delivers life-saving insights
Quantum Meets Classical: Drug design and chemistry research

SEAGATE LYVE™ Cloud

SEAGATE

Smarter Storage for Scientific Research & Life Sciences

Designed for the **realities of modern research**: constrained **budgets**, expanding **AI workloads**, and sensitive **data**.

At Seagate, we collaborate with **research institutions** and **life sciences organisations** to address **real-world data challenges**. Our solutions support the **full lifecycle of scientific data**—from acquisition and analysis to long-term archiving.

What We Offer:

Infrastructure that fits your environment

- Compatible with open-source tools and HPC clusters
- Hardware-agnostic systems that integrate seamlessly with existing workflows
- Scalable from laboratory to national research facility

Budget-conscious design

- High-capacity Mozaic drives reduce cost per terabyte
- Transparent pricing with no hidden charges
- Subscription and financing options available

AI-ready storage architecture

- Tiered storage for training, inference, and data preparation
- High-throughput, low-latency pipelines for large datasets
- Edge-to-core-to-cloud data movement with Lyve Mobile

Secure, sovereign cloud storage

- Lyve Cloud offers GDPR-compliant, UK/EU data residency
- No egress charges, hybrid deployment options Enterprise-grade encryption and access controls

Proven support for regulated environments

- End-to-end ownership and global support network
- Trusted in genomics, imaging, and clinical research
- Lifecycle automation and policy-based tiering

Speak to **our team** to explore how we can **support your research goals**.

CONTENTS

3 Editor's Welcome
Making the most of advanced computing infrastructure

5 Innovation in the post-Moore's Law era
As transistors become ever more difficult and more expensive to produce, innovation in HPC and AI is driving advancements

8 Harnessing HPC – Cambridge's approach to computing at scale
Deepak Aggarwal explains how his team delivers services that support new and experienced users in HPC

12 Modelling truck explosions and fires using asynchronous tasking on exascale supercomputers
Martin Berzinson on developing large-scale, task-based software for solving complex engineering problems

16 Exact AI for aerospace
A start-up is developing a form of mathematically precise AI models for decision-making in mission-critical areas such as aerospace and defence

18 Merging GPU, CPU and neuromorphic processing
The SpiNNaker2 platform is a brain-inspired compute architecture combining elements of neuromorphic, symbolic and probabilistic AI

20 Preparing for agentic AI in healthcare
Harry Lykostratis discusses the development of a framework for the adoption of agentic AI, supporting more efficient and effective workflows

22 Causal AI breakthrough at exascale – modelling malaria
The combination of Causal AI and high-performance computing are unlocking faster, more transparent, and more actionable insights

24 Building better data networks
How can performance be improved to support the next generation of HPC and AI infrastructure?

26 Connecting qubits to photons
The development of transducers could help scientists interface superconducting quantum computing and photonic quantum networks

28 Optimising drug design with quantum computing
Combining classical and quantum computing can help to accelerate the development of new drugs and treatments

30 Quantum chemistry: Integrating quantum computing and HPC
Professor Peter Coveney explains why quantum's limitations mean classical computing still has a vital role in research

34 Advancing materials science with quantum computing
OTI's Scott Genin discusses how quantum computing can solve challenges and open up new possibilities for materials discovery

36 Understanding endometriosis using quantum computing
UQ's Mark Webber talks about tackling endometriosis with quantum computing and what is needed to make a useful quantum system

38 Developing quantum algorithms for the future
Andrew Childs discusses the development of algorithms for quantum computing and how this exciting field continues to evolve despite existing hardware limitations

42 Thank you to our sponsors
Making the most of advanced computing infrastructure

CONTENT AND MARKETING
Tel: +44 (0)1223 221030

Editor
Robert Roe
robert.roe@europascience.com
Audience development manager
Andrew Knight
andrew.knight@europascience.com
Senior designer
Zoe Wade
zoe.wade@europascience.com
Head of content
Finbarr O'Reilly
fin.oreilly@europascience.com

ADVERTISING & PRODUCTION
Senior account manager
Eleanor Waters
eleanor.waters@europascience.com
+44 (0)1223 221041
Head of business development
Stephen Russell
stephen.russell@europascience.com
+44 (0)1223 221044
Production manager
Nick Clark
nick.clark@europascience.com

CORPORATE TEAM
Chief operating officer
Mark Elliott
mark.elliott@europascience.com
Chief executive officer
Warren Clark
warren.clark@europascience.com

Scientific Computing World is published by Europa Science Ltd, St John's Innovation Centre | Cowley Road | Cambridge | CB4 OWS | ● ISSN 1744-8026
Tel: +44 (0) 1223 211170 ● Fax: +44 (0) 1223 213385
Web: www.scientific-computing.com

Receive our weekly newsletter: Go to www.scientific-computing.com/register

Where hype meets reality: why breakthrough technologies must prove their scientific worth

Over the past decade, scientific research has undergone a significant transformation driven by the rapid advancement of computational power. High-performance computing (HPC) and artificial intelligence (AI) are no longer confined to specialised research centres; they are becoming central to scientific progress. Quantum computing is the next technological horizon that promises to deliver significant input for scientific research. However, today, much of the impact is based on creating large-scale, fault-tolerant quantum computers rather than applying the technology to real-world research.

Together, these technologies promise breakthroughs across fields ranging from climate modelling to drug discovery. Yet, amid the excitement, a pressing question remains: how do we ensure that investments in cutting-edge technology translate into meaningful value and tangible scientific outcomes?

HPC systems remain the workhorses of computational science. They simulate physical systems at scales that were previously unimaginable, predicting the behaviour of galaxies, modelling protein folding, or designing advanced materials at the atomic level. Modern exascale systems can process quintillions of calculations per second, allowing researchers to probe scientific questions with unprecedented resolution.

AI has added a new dimension to scientific research. It excels at pattern recognition and inference, enabling scientists to extract meaning from massive and complex datasets. In astronomy, AI accelerates the discovery of new exoplanets by sifting through vast amounts of data from telescopes. In healthcare, machine learning models are revolutionising diagnostics and significantly reducing the time required to develop new drugs. In materials science, AI-driven generative models are being used to design molecules with properties tailored to specific industrial or medical needs.

AI models often require HPC-scale resources to train, while AI can, in turn, help optimise HPC simulations by reducing redundant computations or creating efficient surrogate models.

Quantum computing is still in its infancy compared with HPC and AI, but its potential impact should not be ignored. By leveraging quantum phenomena such as superposition and entanglement, quantum computers may eventually solve classes of problems that are intractable for classical machines.

One of the most promising areas is quantum chemistry, where accurate simulation of molecular interactions could enable breakthroughs in drug discovery and materials development. Quantum systems may also offer faster solutions to highly complex optimisation problems, for instance in energy grid management, and they could transform the field of cryptography by reshaping approaches to data security in an era of quantum-enabled decryption.

However, despite rapid progress, quantum computing is not yet a general-purpose tool for most scientists. Its promise is immense, but its current utility remains limited to highly specialised research.

Balancing investment with value creation

Private companies, governments and research institutions are investing billions in AI infrastructure. While this funding is essential to keep scientific capability at the frontier, investment alone is not enough. New technologies should be evaluated not only by their novelty but by the value they generate.

The key questions are straightforward. How many new insights, discoveries or validated models does a technology enable? Does the technology accelerate collaboration across disciplines such as biology, physics and engineering? And, perhaps most importantly, are the financial and energy costs of new computing infrastructures justified by the breakthroughs they enable?

For instance, building a next-generation exascale machine should not be considered an achievement in itself. Its value must be assessed by whether it accelerates scientific progress relative to its cost and carbon footprint. Current investment in AI, particularly LLMs, would require hundreds of billions of dollars worth of value generated over this current generation of hardware.

The convergence of HPC, AI and quantum computing represents one of the most exciting frontiers in human knowledge. Yet the pursuit of raw power or novel architectures must not overshadow the ultimate goal, which is to advance scientific understanding and push forward into new avenues of research. Responsible investment, guided by value generation and measurable scientific outcomes, will ensure that these transformative technologies fulfil their promise.

Robert Roe
Editor
Scientific Computing World

Evolve or stagnate: Innovation post-Moore's Law

As each new generation of transistors becomes more expensive and difficult to produce, innovation in HPC and AI technology is driving advancements

At the ISC High Performance 2025 event hosted in Hamburg, Germany, speakers discussed how advances in HPC (high-performance computing) and AI (artificial intelligence) technology is driving innovation that aims to keep pace with Moore's Law as each new generation of transistors becomes more expensive and difficult to produce.

As Moore's Law slows, the HPC community faces a pivotal moment: evolve or stagnate. At this year's ISC conference, leaders in computing from across the globe outlined a compelling vision of how AI and HPC are converging, not just out of convenience, but out of necessity.

As silicon scaling reaches physical limits, the industry is turning to architectural innovation, AI-centric designs and modular hardware, such as chiplets, to meet the ever-growing computational demands.

Isambard AI: a national leap forward

Professor Simon McIntosh-Smith, Director of the Bristol Centre for Supercomputing (BRiCS), discussed the development of Arm-based systems in the UK's Isambard-AI programme.

"We've established this organisation to deliver more than £300 million-worth of AI and HPC services over the next five years," he said. "We've already installed and got into production Isambard 3, which is an Nvidia Grace CPU system, and Isambard-AI phase one, based on the Grace Hopper architecture."

These HPE-built systems are not conventional supercomputers retrofitted for AI workloads; they're designed from the ground up to prioritise artificial intelligence.

"This isn't an HPC system that also does AI," McIntosh-Smith emphasised. "This is designed from the beginning for AI. The whole software stack in particular, is very AI-focused."

Phase one of Isambard-AI, a 168-node direct liquid-cooled system, was operational within just four months of arriving, transitioning from a university car park to a fully running production system. The second phase, a five megawatt system comprising nearly 5,500 Grace Hopper processors, recently debuted at number 11 on the Top500 list – the UK's best showing in 23 years.

"It's going to be the fastest supercomputer in the UK by far," he said. "The UK Government basically said, 'we think AI is important. We believe we're behind where we need to be. We want to address that quickly'."

"We've established this organisation to deliver more than £300m worth of AI and HPC services over the next five years"

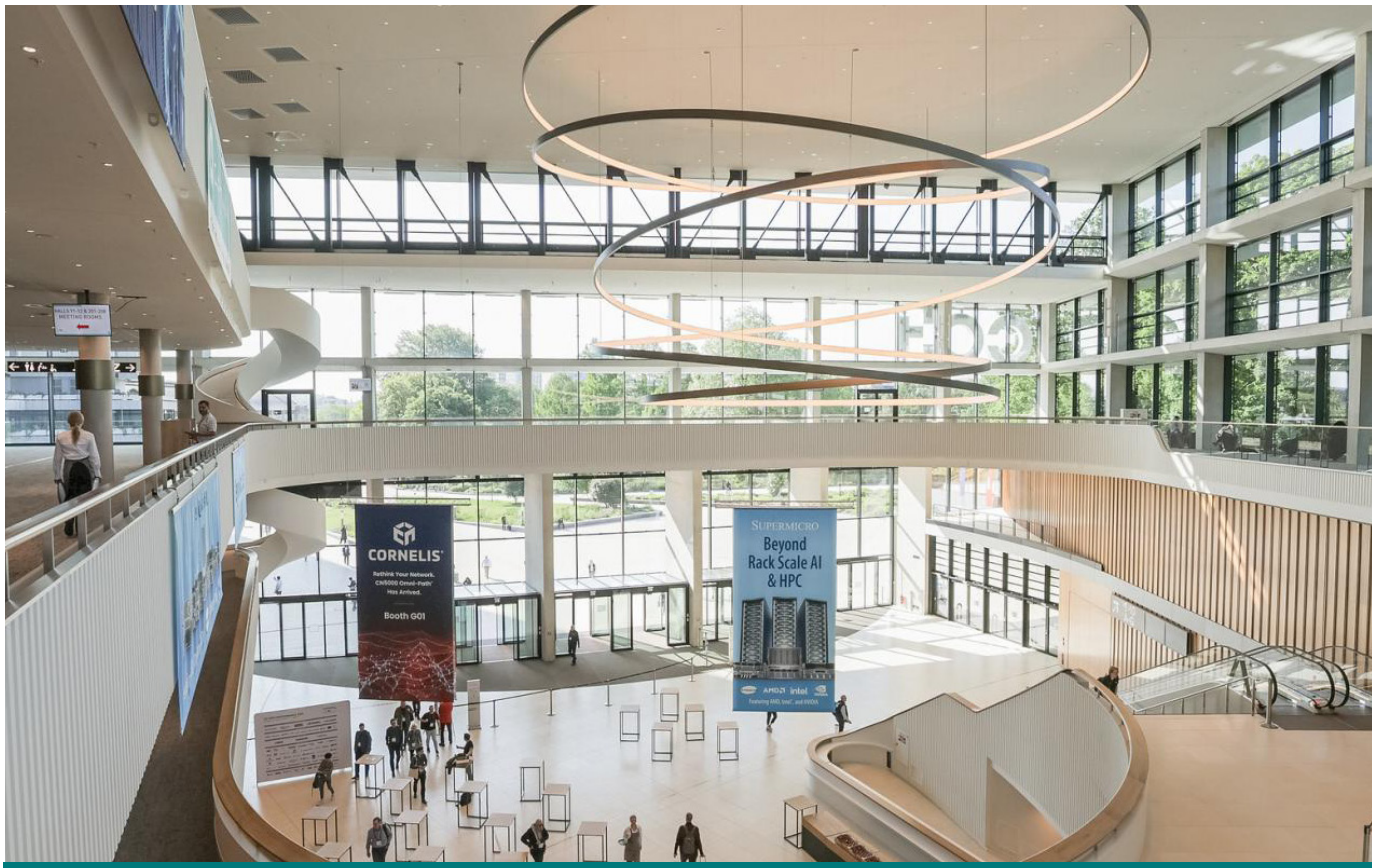
HPC + AI: not a zero-sum game

Dr Satoshi Matsuoka, Director of Riken's Centre for Computational Science in Japan, addressed a critical concern in the community: will the shift toward AI-centric design leave traditional HPC behind?

"There's been lots of fear among the HPC community," he noted. "Some people feel that HPC will be compromised... compute, memory, network, storage, everything is shifting towards AI."

But Matsuoka is optimistic, seeing convergence rather than competition.

Left to right: Luisa Patricia Gonzalez Guerrer, Satoshi Matsuoka and Simon McIntosh Smith



PHILIP LOEPER/ISC High Performance 2025

ISC 2025 in Hamburg: speakers explained why they are turning to architectural innovation, AI-centric designs and modular hardware

> "What we believe is that advances and so much money being invested in the AI space... will help to accelerate HPC beyond what it would have been capable of."

The development of the processor that powers Japan's Fugaku supercomputer, known as the A64FX, was a pivotal moment in the evolution of HPC. The project was born from the Japanese Government's "Post-K" initiative, intended to build a successor to the K computer that could lead the world in both computational power and energy efficiency. At the heart of this vision were three main collaborators: Riken, Japan's top national research institute; Fujitsu, the prime industrial partner and system developer; and Arm, the British semiconductor and software design company that provided the instruction set architecture (ISA).

Historically, most supercomputers relied on x86-based processors from Intel or, increasingly, GPUs from Nvidia. However, Japan sought to break away from this dependency and design a homegrown solution tailored specifically for the diverse and demanding workloads of modern science, including simulation, AI and big data processing. Riken, in its role as the research lead, specified these needs in a performance-oriented blueprint. Fujitsu, drawing on its experience from developing the SPARC64 processors used in the K computer, took up the challenge of implementing this vision.

"Some people feel that HPC will be compromised... compute, memory, network, storage, everything is shifting towards AI"

A crucial innovation in this process was the decision to base the new processor on Arm architecture, specifically its 64-bit Armv8-A. While Arm cores had been traditionally used in mobile and embedded devices due to their power efficiency, they were not previously regarded as contenders in HPC. To make Arm viable for supercomputing, a radical expansion was needed. This is where the close collaboration between Fujitsu and Arm became critical.

Working together, they co-designed a new vector processing extension to the Arm ISA, called the Scalable Vector Extension (SVE). Unlike traditional SIMD vector units, which have fixed register sizes, SVE supports vector lengths from 128 bits to 2,048 bits in 128-bit increments, allowing future processors to adopt wider vectors without changing the software. This design made SVE extremely forward-compatible and well-suited to the kind of adaptable, high-throughput workloads seen in HPC. Fujitsu became the first company in the world to implement this extension in hardware through the A64FX.

Riken's contribution to the processor went beyond setting requirements; its Center for Computational Science worked closely with both Arm and Fujitsu to ensure that the resulting architecture would support a wide range of scientific applications. The software stack, compilers, libraries and simulation tools, had to be adapted to the new architecture, and Riken played a pivotal role in co-developing and validating these tools in tandem with the hardware.

This ensured that by the time the A64FX chips were ready for deployment, a robust ecosystem was already in place.

Japan is placing its bet on this belief with the successor to its Fugaku supercomputer.

"The speed-up from the K computer to Fugaku was about 70x," he said. "The question is: can we achieve similar speed-up going from Fugaku to FugakuNEXT in 2029 or 2030?"

This leap will require strategic exploitation of AI-driven advances. Even as double-precision performance stagnates in next-generation GPUs "for example, Nvidia GPUs going from Hopper to Blackwell, the FP64 performance actually decreased".

Matsuoka believes that circuit-level reconsiderations and cross-domain synergies can unlock new growth.

"Can we converge? Can we help each other to achieve yet another two orders of magnitude? Or even three?" Matsuoka asked.

Chiplets: a practical path to performance

While national programmes like Isambard show what's possible with strong investment, Dr Patricia Gonzalez-Guerrero from Lawrence Berkeley National Laboratory pointed to the physical realities underpinning the slowdown in performance scaling.

Beyond specialising monolithic processor designs, as seen in the Fugaku supercomputer. Chiplets take it a step further; instead of constructing a processor as a single, massive, monolithic chip, the chiplet approach breaks the system down into smaller, modular pieces, individual silicon dies, with specific functionality, integrated into a single package.

These chiplets might be compute cores, memory controllers, AI accelerators, or network interfaces, all brought together through high-bandwidth, low-latency interconnects. By decoupling components in this way, chiplet-based design enables engineers to mix and match technologies, optimise yields and tailor systems for specific performance goals or power envelopes.

As the pace of Moore's Law slows and the physical limitations of traditional monolithic chip design become more apparent, the world of high-performance computing finds itself at a crossroads.

For decades, increasing computational power relied on packing more transistors into a single silicon die. But in recent years, this strategy has begun to yield diminishing returns in terms of cost, power efficiency and flexibility, especially for the increasingly heterogeneous workloads that modern HPC systems must handle.

"We know that HPC performance does not increase any more," she said. "One of those reasons is that it is very hard to scale transistors beyond two nanometres, and that's a fact."

The solution, Gonzalez-Guerrero argued, is chiplets, modular components that allow for high-performance, scalable designs with improved manufacturing yields. "What we need to increase performance is more transistors. Then let's make the chips larger. But the problem is yield: with large monolithic dies, it drops dramatically."

Chiplet-based designs, particularly using 2.5D packaging, mitigate this by placing multiple smaller dies on a silicon interposer with high-bandwidth interconnects.

"Chiplets made the exascale possible," she explained. "The top three supercomputers are fuelled by this chiplet technology."

By separating design from yield constraints, chiplets enable both cost efficiency and flexibility, key qualities in a world where innovation must be decoupled from Moore's Law.

Lawrence Berkeley National Laboratory (LBNL) is engaged in advancing chiplet architectures specifically for HPC.

Under the leadership of John Shalf in LBNL's Computing Sciences division, the lab recognises that traditional monolithic chips are hitting physical and economic limits which results in yields falling dramatically as die sizes grow, and power efficiency improvements are stalling.

LBNL partnered with the Open Compute Project (OCP) in 2024 to launch the Open Chiplet Economy Experience Center on its Berkeley campus.

This centre showcases the latest technology demos, from UCIe link subsystems to modular test chips, offered by partners like Synopsys, Blue Cheetah Analog, and others, illustrating real-world progress in chiplet formation and integration.

A post-Moore's roadmap

The convergence of AI and HPC is not just a technological trend, it is a necessity in the face of physical and economic limits. As Gonzalez-Guerrero pointed out, "it's hard" to push beyond two-nanometre technology nodes. But rather than yielding to the slowdown, the community is pivoting.

From rapid deployments like Isambard-AI, to the chiplet revolution, to Japan's vision for Fugaku Next, the HPC world is embracing architectural ingenuity and AI-driven momentum. Rather than replacing traditional scientific computing, AI appears set to propel it forward.

In McIntosh-Smith's words: "It's incredible how fast you can go with this approach."

As Matsuoka concluded, perhaps AI's most significant contribution to HPC will be not in competition, but in convergence.

"Advances in AI hardware or hardware supporting AI will not only accelerate AI, but with smarts, with the right design, system design with innovative algorithm design, we can achieve much greater performance gains into the zettascale era," Matusoka stated. **B**

"We believe that advances and so much money being invested in the AI space... will help to accelerate HPC beyond what it would have been capable of"

Building bridges: How Cambridge makes HPC accessible to researchers

The University of Cambridge is reshaping its infrastructure to serve everyone from clinical researchers to humanities. **Deepak Aggarwal** explains how his team is balancing GPU diversity, simplifying user access, and ensuring the availability of advanced computational resources to all scholars

Deepak Aggarwal leads the storage portfolio for Research Computing Services at the University of Cambridge, overseeing the AI Research Resource (AIRR). With a background in physics and hands-on experience setting up HPC (high performance computing) clusters from scratch, Aggarwal brings both a user's perspective and technical expertise to his role. At Cambridge, he manages computing infrastructure that serves a diverse user base whose members are increasingly adopting AI-driven approaches to conduct research.

Aggarwal and his team deliberately maintain hardware diversity across Nvidia, Intel and AMD GPUs to support both AI-driven workloads and traditional HPC. This shift has implications for how infrastructure is managed: it demands new models of accessibility, user onboarding and workload optimisation, ensuring that researchers from all disciplines can take advantage of advanced resources without encountering steep technical barriers.

Here, Aggarwal discusses the challenges and opportunities of this transition, the role of community in advancing HPC, and how Cambridge is shaping the next generation of computational research infrastructure.

Could you tell me about your role at the University of Cambridge?

Deepak Aggarwal: "I'm currently working as a Principal HPC Systems Manager at the University of Cambridge, within Research Computing Services (RCS), part of University Information Services. I lead the storage portfolio for our research computing services and oversee the national AIRR service. As part of this national service, we manage a system called Dawn. I also provide infrastructure support to the broader research community, both within Cambridge and for external users.

Additionally, I serve as Secretary of the UK HPC SIG (Special Interest Group) community. This comprises individuals from academia who manage HPC services within their universities and we organise quarterly meetings. I assist with planning and delivery.

How does the HPC SIG UK community help its members?

DA: The main objective of the SIG community is to share knowledge. For example, if Cambridge develops a solution for our users, it may also apply to other universities, thereby avoiding the need to reinvent the wheel in isolation. At these quarterly meetings, a structured programme is in place, where people give talks on a wide range of topics, from cluster installation to new AI tools for managing support tickets. The aim is always to share knowledge and the community benefits greatly from it.

We also have an active Slack channel and email list. For instance, if someone is

Sashkin/Shutterstock.com

Deepak Aggarwal

Deepak Aggarwal, Principal HPC Systems Manager at the University of Cambridge

considering new storage hardware or cooling solutions, they can ask the community for feedback before approaching vendors, ensuring honest opinions. These meetings are deliberately informal; nothing is recorded, allowing people to speak openly about technical and professional challenges, including career pathways in professional services compared with academic roles.

What drew you to HPC and managing large-scale resources?

DA: I began my professional journey in 2011, after completing a master's degree in physics. I joined a nuclear fusion institute in India as a nuclear analyst, working on ITER, the International Thermonuclear Experimental Reactor in France. I ran my simulations on a small HPC system at the institute, which was my first exposure to clusters.

Later, there was a strong push to develop computational facilities. In 2016, I joined a team setting up a large one-petaflop machine. I learned how to set up an HPC cluster from procurement to commissioning, benchmarking, porting user applications and training researchers to run jobs. I worked closely with users, helping them understand HPC terminology and how to transition from workstations.

In my current role I do not work directly with users as much, as it is more about developing solutions, while the RCS community handles engagement. However, because I was a user before becoming an admin, I understand their pain points and how they communicate. That background has been invaluable.

At the institute, when we launched the cluster, we grew from 50 users to 200 in three years, out of a total staff of 500. Many researchers stopped buying personal workstations once they saw the benefits of shared resources. One big challenge was engaging isolated users who were hesitant to seek help. Some were very active, but others stayed quiet. To reach them, I started an HPC newsletter and published around 40 issues, one every month.

What does your role look like day to day?

DA: On a day-to-day basis, user requirements drive our services. I manage three types of storage: high-performance storage for HPC workloads, project storage for teams and tape facilities for long-term data archiving and back-ups. These services exist because of demand from various user categories, not because we designed them in isolation.

Our main HPC service is CSD3, a large cluster comprising a mix of compute nodes, including CPU-based and GPU-based nodes, and spanning multiple GPU vendors, such as Nvidia, Intel and AMD. This ensures users can run their workloads on the most suitable hardware. Dawn is entirely Intel GPU-based and supports national AI research projects.

What is the makeup of the user base at Cambridge?

DA: Our user base is very diverse, ranging from the clinical school, physics, >

"Supporting diversity in skills means meeting people where they are, not forcing them to learn everything we know"

computational chemistry and astronomy, to the humanities, which increasingly utilise AI tools that require GPUs.

We offer various access methods to accommodate individuals with different skill levels. Experienced users can connect via Linux terminals, while beginners can use Open OnDemand, a browser-based UI that makes submitting jobs as simple as running applications on a laptop.

Supporting diversity in skills means meeting users where they are. We do not require everyone to learn Linux before using HPC. Instead, we provide tools that allow researchers to run their workloads immediately, essentially treating HPC as an extension of their laptop. This approach is aligned with the Government's roadmap, which emphasises rapid access to computational resources rather than lengthy training requirements.

At Cambridge, gauging user requirements can be challenging due to the community's large and diverse nature. Sometimes solutions arise from administrative challenges rather than direct requests. For example, we had different onboarding processes for internal, industry and national users. With the launch of AIRR, which allows any UK researcher to access resources at Cambridge, Dawn, or Bristol, Isambard-AI, we needed a unified system.

We adopted a federation model using MyAccessID from GÉANT and the open-source project management tool Waldur. This allows users to authenticate with their home institution's credentials, while giving PIs control over project membership. It simplifies onboarding, ensures proper offboarding at project end, and removes the burden of managing external credentials. Both Cambridge and Bristol now use this model for AIRR.

Has the shift from CPU to GPU impacted your role?

DA: The shift from CPUs to GPUs has transformed HPC. When I started, everything was CPU. My first large cluster consisted of 80% CPU and 20% GPU. We struggled to utilise those GPUs effectively. Nvidia invested heavily in training and libraries, which made adoption easier. Over time, the balance flipped, and new systems are now GPU-centric.

AI has accelerated this trend. LLMs and AI workloads require GPUs, and researchers want quick access. While AI codes are relatively easy to port between GPU types, traditional scientific HPC codes, such as CFD, molecular dynamics, or MPI workloads, are more complex to migrate, especially legacy codes written decades ago. This creates challenges in avoiding vendor lock-in.

DC Studio/Shutterstock.com

We deliberately maintain hardware diversity with Nvidia, Intel and AMD GPUs to prevent dependence on a single vendor. We are also working on projects such as federated container services, which abstract away GPU differences and automatically optimise workloads for whatever hardware is available. This heterogeneity reduces complexity for users while promoting sustainability and flexibility for providers.

Ultimately, the HPC community is in transition. AI has brought new users from non-traditional fields, making accessibility and flexibility more important than ever. At the same time, we must continue to support traditional HPC users, many of whom depend on highly optimised, domain-specific codes that remain CPU- and MPI-heavy. Balancing both is the challenge ahead.

There are various kinds of communities. We are building infrastructures that act as a bridge between researchers and the computational power they need. Our biggest users come from the clinical school, but we also have users from physics, computational chemistry, astronomy and many other fields. We even see increasing use from the humanities, particularly with the rise of AI tools in their disciplines. They also need high-end GPUs for their workloads.

In principle, any user who requires computational power, regardless of their HPC experience, can utilise the services. For those experienced in Linux, they can SSH into the terminal and run jobs directly. For those without Linux knowledge or HPC terminology, we provide a browser-based service called Open OnDemand. It has a simple interface: you log in, fill in details, and submit a job much like you would on a Windows application on your laptop.

How does this impact services and training, particularly for new users?

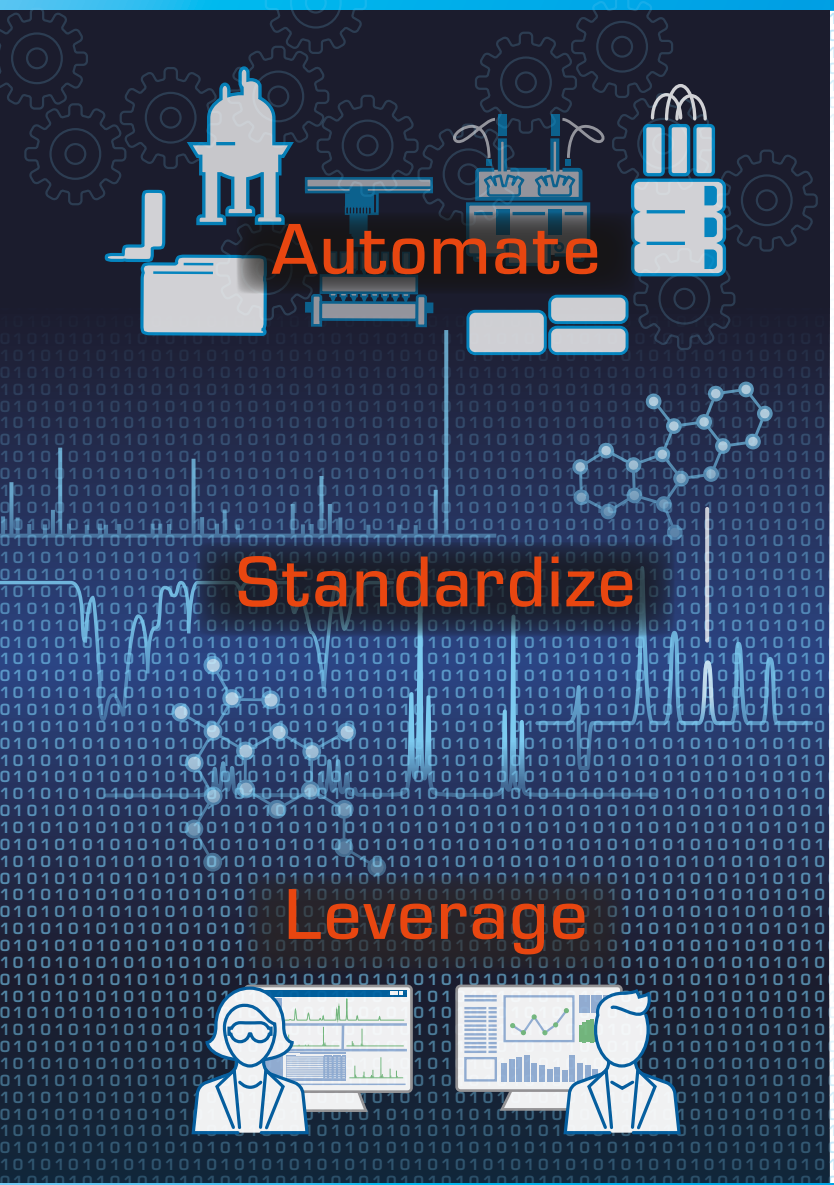
DA: Today, there is a strong push from new users running AI workloads on laptops, whether Linux or Windows. The challenge is to transition them to HPC without a steep learning curve. Supporting diversity in skills means meeting people where they are, not forcing them to learn everything we know. We want them to use their existing skills with the computational power we provide.

Some users may choose to learn more, but for those who just want to run workloads, we do not want the learning curve to delay them. The Government's current roadmap also emphasises providing resources as quickly as possible so that researchers can run workloads from day one. It should feel like a replacement for their laptop: take the code, move it to HPC, and run it.

This shift has expanded the community. We now have many users from non-computational backgrounds, including the humanities. Five or 10 years ago, nobody expected that. The AI boom has driven demand and brought in these new groups, which means we must be flexible in enabling access, just like public cloud providers. You log in, spin up resources and run jobs as if you were on your laptop, even though they are actually running on a cluster. At Cambridge, we have been doing this successfully. **B**

Deepak Aggarwal is the Principal HPC Systems Manager at the University of Cambridge

Enable Better Science



Create and manage automated dataflows with our no-code application for robust, reproducible experimentation.

Natively process and manage analytical data from >150 instrument vendor formats with chemical context.

Empower scientist decision-making with applications to review and interrogate data.
Deliver AI-ready data for data science.

Connect Heterogeneous Data Across Your Organization

Spectrus Platform

Learn more: www.acdlabs.com/Spectrus

ACD/Labs®

info@acdlabs.com ☎ 1-800-304-3988 (Toll free, USA & Canada)
+44 (0) 1344 668030 (UK)

Solving truck explosions and other complex physics using asynchronous tasking on exascale supercomputers

Martin Berzins, a researcher in computational science, discusses the development of large-scale, task-based software for solving complex engineering problems

Drawing on decades of experience spanning applied mathematics, computer science and engineering, Martin Berzins explains how the UINTAH framework and the Kokkos library enable high-performance, portable simulations across diverse architectures, including the latest exascale supercomputers.

Berzins is a multidisciplinary researcher in Computational Science whose work spans applied mathematics, computer science and engineering. His research focuses on the development of partial differential equation software for solving challenging engineering problems across a wide range of applications on extreme-scale computers.

He is a Professor of Computer Science in the School of Computing and the Scientific Computing and Imaging Institute at the University of Utah, and also a Visiting Professor at the University of Leeds. Since 2013, he has been the Computer Science lead for the DOE NNSA PSAAP2 Carbon Capture Multidisciplinary Simulation Center at the University of Utah.

Can you tell me about your role at the University of Utah and how you became involved in this HPC project?

Martin Berzins: I sit somewhere between applied mathematics and computer science, and I have always worked on mathematical problems in computer science. In 2003, I moved to the University of Utah after spending several years at the University of Leeds. It was simply time to do something different.

Utah was attractive due to the Scientific Computing and Imaging Institute (SCI), which was founded by Chris Johnson who invited me to join. Back in 1998, I had spent a sabbatical there, working with the group as they began tackling large-scale engineering problems.

One of the key ideas that emerged was pioneered by Stephen Parker, a graduate student of Chris's. Steve developed a framework called "UINTAH", a task-based code for solving large engineering simulation problems.

The word comes from the Uinta mountain range, which sits about 40 miles behind me as I speak. It is the only east-west mountain range in the continental United States. The word itself is a Native American term meaning "running water by trees," which captures the feel of that landscape.

UINTAH evolved from work on viewing programmes as collections of tasks, combined with a runtime that schedules them. Parker went on to Nvidia, where many people from Utah now work. After he left, I took on the responsibility of further developing UINTAH. My focus was to make the execution of tasks asynchronous. Originally, tasks were executed in a fixed order. By moving to asynchronous scheduling, we ensured there was always work ready to run, avoiding idle waiting for messages or data.

This idea is at the core of your question about portability and performance. Structuring a programme as a set of tasks, carefully specifying their dependencies, means you never stop in parallel computing. You always have something else to execute. That is the central principle behind UINTAH. It is not unique to UINTAH. Similar ideas have been incorporated into other codes, such as Charm++ from the University of Illinois, led by

petehansen/shutterstock

The PASC Conference

Martin Berzins is a Professor of Computer Science at the University of Utah's School of Computing and the SCI Institute

"The truck caught fire and, within minutes, exploded, leaving a crater 60 feet across and 20 feet deep. Fortunately, the area was evacuated, but the blast was massive"

Sanjay Kale, which has been in development since the late 1980s. Task-based programming has roots dating back to the 1960s. But it has never been mainstream. It is more difficult than more traditional models.

Why is this technique not widely adopted if it works so well?

MB: Because for many simpler engineering applications, conventional codes written with MPI or OpenMP scale well enough. If communication patterns are straightforward, or if you have enough manpower to brute-force the problem, you do not need a more complex system. It is only when applications become very complicated that manual control of communication breaks down. That is when task-based runtimes really shine.

The problems we work on are indeed very complex. For example, we often model coupled fluid and solid dynamics, as well as extreme events such as fires and explosions. Our colleagues in chemical engineering here in Utah have given us many challenging problems. One example is a real accident in Utah where a semi truck carrying 30,000 pounds of explosives overturned on a dangerous highway bend. The truck caught fire and, within minutes, exploded, leaving a crater 60 feet across and 20 feet deep. Fortunately, the area was evacuated, but the blast was massive. We modelled that event, simulating the truck, the explosives, the packaging, the fire and the detonation process. This problem required a task-based approach to handle all the complex physics at scale. The more complicated the physics, the more attractive task-based programming becomes. But it is more complex to implement. You have to manage how tasks access data carefully. Two tasks might safely read the same data, but if one writes while another reads, you have a hazard.

The UINTAH runtime manages this with a data warehouse, which tracks all variable accesses and ensures correctness.

How does asynchronous tasking differ from other models?

MB: The runtime also monitors which tasks are ready, which are waiting on communication, and whether tasks should run on CPUs or GPUs. This separation of concerns is key: the runtime system handles execution, while the application developers focus only on writing physics tasks and their dependencies.

By contrast, in MPI, you explicitly manage communication. You send messages and wait for responses. With asynchronous tasking, you relinquish control over when and where execution occurs. The runtime decides. For developers used to MPI, that can be difficult to accept.

Around 2010 to 2012, we hit a wall: our code was not scaling because too much time was spent in MPI waits. Tasks were sitting idle, waiting for data. I told my graduate students, Justin Luitjens and Ching-Yu Meng, that the runtime had to be asynchronous.

They redesigned it, and from that point forward, UINTAH scaled cleanly across >

PeachShutterStock/shutterstock.com

Department of Energy supercomputers. That work was carried forward by many students and collaborators, including Alan Humphrey, John Holmen (now at Oak Ridge), Mark Garcia (formerly at Argonne) and Alan Sanderson, with strong support from colleagues at Intel, Argonne and the Kokkos group at Sandia.

With that runtime, we were early adopters of new supercomputers, often running at full scale while other groups were still struggling to adapt. For example, we were among the first to run effectively on Aurora.

How can scientists adapt their applications to best suit exascale resources?

MB: It is less about the number of nodes and more about the power of each node. Machines such as Aurora and Frontier have fewer nodes than older CPU-only systems, but each node is far more powerful, with multiple CPUs and GPUs. The imbalance between compute and communication grows.

Communication speeds have improved, but not enough to keep up with GPU performance. GPUs can perform hundreds of floating-point operations in the time it takes to move a single word from main memory. That means GPUs are often starved for data. So you need careful runtime management to keep these powerful nodes busy without being bottlenecked by communication or memory transfers.

This is where Kokkos comes in. Around five-to-seven years ago, we began thinking systematically about portability to GPUs. Kokkos is a DOE-developed library that abstracts both execution and memory management. With Kokkos, you can write one code that runs on CPUs, Nvidia GPUs, AMD GPUs and Intel GPUs without modification. The library restructures loops and memory layouts to suit each backend.

For UINTAH, combining asynchronous tasking with Kokkos was crucial. One part provides resilience against delays and ensures there is always work ready. The other part makes that execution portable across architectures. Together, they allow our simulations to run efficiently on every DOE exascale machine we have tested.

We chose Kokkos because it was one of the first serious efforts in portability, developed at Sandia, while a parallel project called RAJA emerged from Lawrence Livermore. Both programming models have since been adopted widely. We knew the Kokkos developers and began collaborating with them approximately 10 years ago. They even added support for asynchronous tasking at our request, though relatively few people use it.

Over time, their implementation evolved, and we adapted our code accordingly; however, the collaboration has been excellent. Their long-term goal is to integrate Kokkos into the C++ standard library, which would greatly expand adoption.

Performance portability does mean you may not achieve the absolute best performance possible, because vendor-specific tuning is required for optimal performance. However, for million-line scientific codes, achieving efficient execution everywhere is far more valuable than squeezing out the last few per cent on a single architecture.

DOE's Exascale Computing Project recognised this and focused heavily on porting major codes to GPU-based systems using frameworks such as Kokkos. It was an outstanding project that brought together national labs and delivered real results.

There is also a social dimension. Scientific software represents billions of dollars of investment across DOE, NSF, DOD and other agencies. Rewriting those codes wholesale is not feasible. Most groups do the minimum necessary to keep things running, and inertia is a strong force. That is why new models take time to gain widespread adoption. Typically, it takes a decade for today's cutting-edge approaches to become mainstream in production codes. So, by running at exascale now, we gain a kind of 10-year head start.

How might this research evolve?

MB: Looking ahead, there is also interesting work at the intersection of task-based runtimes, AI and quantum computing. For example, researchers at RIKEN in Japan have proposed task-based approaches as a way to bridge classical and quantum computations, where latencies are unpredictable. Similarly, AI and engineering codes could both be expressed as tasks, allowing them to integrate naturally within the same runtime. This is an area of future opportunity.

Finally, self-adaptivity is vital. On any parallel machine, scalability collapses as soon as you wait for data. The asynchronous runtime eliminates that bottleneck. It makes UINTAH self-adaptive, automatically adjusting to communication delays, machine load and network topology. That adaptability has been the key to running effectively on all the largest DOE machines. **B**

Martin Berzins is a Professor of Computer Science at the University of Utah's School of Computing and the SCI Institute

Targeted MS Quantitation

Controlled and accessed from anywhere

Thermo Scientific™ Chromeleon™ 7.4 software

Consolidate your mass spectrometry (MS) laboratory with robust support for targeted quantitation and screening workflows in biopharma, environmental, and food safety sectors. Connect instruments, users, and data in a centralized scalable server-based system, providing remote data access and central management of methods, templates, and reports.

Full support for GMP compliance ensures data integrity, traceability, and verification throughout the data lifecycle, and native control of Thermo Scientific™ MS instruments, including single quadrupole, triple quadrupole and high-resolution accurate mass (HRAM).

Learn more at **thermofisher.com/chromeleonms**

‘Exact AI’ for aerospace and defence: Accelerating decision-making in mission-critical domains

PiLogic is pioneering a form of artificial intelligence that uses mathematically precise, expert-guided models for decision-making in sectors such as aerospace and defence. CEO **Johannes Waldstein** explains

Unlike generative artificial intelligence (AI), PiLogic’s models are grounded in logical and probabilistic inference, incorporating expert knowledge and engineering principles directly. The result is more accurate, efficient and lower-compute AI that can be deployed on remote devices such as satellites.

The company raised $4 million in seed funding led by Scout Ventures and Seraphim Space, with participation from Sovereign Capital, Flex Capital, FN Fund and angel investor Gokul Rajaram.

The technology combines logical inference, probabilistic reasoning, search and machine learning. It does not require large datasets or GPU infrastructure normally associated with LLM models.

The initial use cases PiLogic is focusing on are the radar model which tracks 3D objects such as drones and aeroplanes with high precision, adaptable to sonar, GPS and LIDAR and satellite power diagnostics which can be used to predict the health of onboard systems in real-time, validated with Nasa.

In the future, the technology could be expanded to include more use cases such as autonomous flight, threat detection and advanced diagnostics for space- and aircraft.

We spoke with Johannes Waldstein, PiLogic’s Founder and CEO...

Can you introduce PiLogic and its core mission?

Johannes Waldstein: Absolutely. PiLogic is an AI start-up pioneering what we call “Exact AI”, a mathematically precise, expert-guided approach to artificial intelligence. We’re focused on applying this to mission-critical environments such as aerospace, defence and space, where traditional AI models, especially generative ones, fall short.

Our platform combines logical inference, probabilistic reasoning, search and machine learning to produce deterministic, explainable outcomes. It doesn’t rely on large data sets or require heavy GPU infrastructure, which makes it perfect for deployment in constrained environments such as satellites, drones or field-deployed systems.

What exactly is “Exact AI,” and how does it differ from more common approaches such as GenAI?

JW: “Exact AI” is all about precision and accountability. It’s AI built from first principles, incorporating domain expertise, physics, logic, and known constraints directly into the model. This is a fundamentally different approach from generative AI, which uses massive datasets to learn statistical correlations.

Unlike LLMs, our models don’t hallucinate. Outputs are traceable, verifiable and deterministic. For example, if our diagnostic model says a satellite power subsystem will fail in 48 hours, it also tells you why and what corrective action to take.

That level of transparency and reliability is critical in sectors where lives and missions are at stake.

Why did you choose aerospace and defence as your entry point?

JW: These industries are full of legacy, rules-based systems that struggle to scale or adapt in real time, especially in edge environments where compute and bandwidth are limited.

Johannes Waldstein, CEO of PiLogic: ‘Unlike LLMs, our models don’t hallucinate’

At the same time, they're high-consequence domains where accuracy, explainability and trust are non-negotiable.

We saw an opportunity to bring truly intelligent, autonomous systems into these environments, not by replacing humans, but by augmenting their ability to make informed decisions at speed and scale. That's where Exact AI really shines.

We have two flagship models currently being deployed:

Radar tracking – this identifies and distinguishes flying objects such as drones vs planes in 3D space.

It remains accurate even when objects cross paths or cluster closely and it works across multiple sensing modalities such as GPS, sonar and LIDAR.

Electrical power diagnostics for satellites – built in partnership with Nasa, this model predicts failures in a satellite's power system using onboard sensors. It also provides real-time remediation strategies, such as rerouting power or activating back-up systems, all executed locally without needing to transmit data back to Earth.

Both models run offline on resource-constrained devices, which is essential for operations in space or in-theatre military use.

How does PiLogic's performance compare with LLMs or traditional ML models?

JW: The difference is dramatic. In targeted use cases, our models use up to 10,000× less compute than LLMs.

They are also several times faster and deliver consistently higher accuracy, without hallucinations or uncertain outputs.

Because our models are built from real-world knowledge and rules, they can outperform neural networks in situations where data is sparse or the stakes are too high for approximations.

That's a huge advantage in defence, aerospace and beyond.

How is your technology able to run on satellites and other edge devices without GPUs?

JW: That's a core strength of our architecture. We use advanced probabilistic inference optimised for high tree-width reasoning, something few systems can handle. Most inference engines max out at a tree-width of 36; ours has been benchmarked at 20-plus.

This means we are able to solve incredibly complex problems on lightweight processors. Therefore, no cloud infrastructure needed, no GPU dependency.

Who are your early customers, and how far along are you in deployment?

JW: We're already engaged with leading satellite operators and dual-use commercial-defence firms. One of our diagnostic models is scheduled to be included on a satellite launch later this year.

Our early traction is validating the need for AI that's built to operate in low-resource, high-trust environments, and we're seeing growing demand across both commercial and government sectors.

What are you planning to do with the recent $4 million seed round?

JW: We're using the funding to expand our core team, growing from five to around 15. A lot of the capital is going toward deepening our product suite, refining deployment infrastructure, and working closely with early adopters to tailor solutions to real-world challenges.

It's about building not just models, but a platform that can support an entire ecosystem of exact AI applications. **B**

Johannes Waldstein is the founder and CEO of PiLogic

Brain-inspired AI computing: ‘In drug discovery, we’ve seen up to a 100× speed-up compared with GPUs’

The SpiNNaker2 platform, a compute architecture combining elements of neuromorphic, symbolic and probabilistic AI with real-time, low-power performance, is opening new frontiers in AI and drug discovery

Artificial intelligence (AI) is playing a growing role in everyday life, but today’s AI hardware and algorithms still fall short of the brain’s efficiency and processing capabilities.

Here, Professor Christian Mayr discusses SpiNNaker2, a computing platform inspired by the brain that combines GPU, CPU, neuromorphic and probabilistic elements in a single system.

Unlike traditional neuromorphic chips that focus solely on mimicking individual neurons, SpiNNaker2 draws inspiration from biology across all levels of its architecture.

Mayr also explains how key principles from neuroscience, such as hierarchical structure, asynchronous communication, dynamic sparsity and distance-based connectivity, can be applied to reshape standard AI models. These approaches yield significant improvements in energy efficiency and processing speed for both training and inference.

What inspired the development of SpiNNaker2, and how does it differ from conventional AI hardware?

Christian Mayr: Traditional AI hardware, such as GPUs, is still far from matching the human brain’s incredible energy efficiency, low latency and large-scale parallelism. Our aim with SpiNNaker2 is to close that gap.

SpiNNaker2 is bio-inspired throughout its architecture, not just at the neuron level similar to most neuromorphic chips, but at all levels of design. It merges features of GPUs, CPUs and neuromorphic systems and integrates probabilistic computing. We’re not just mimicking how a neuron fires; we’re implementing principles such as dynamic sparsity, hierarchy and asynchronous communication, all things the brain does naturally and efficiently.

What are the key architectural features of SpiNNaker2?

CM: SpiNNaker2 consists of 4,848 individual chips, each with 152 parallel cores, essentially mini GPUs, and local memory. Each core has a CPU co-located, allowing us to run a deep learning model in the GPU section while running symbolic AI in the adjacent CPU, for example, in defence applications such as airspace anomaly detection.

What’s crucial is the distributed, local nature of memory and compute. Instead of funnelling everything through a single memory system, we activate only a fraction of resources at a time. For example, when running transformer models, only one chip might be active for a given token layer. This results in massive energy savings and bandwidth advantages.

The machine originated in the EU Human Brain Project. It began with brain simulation in mind, but today we use it for a wide range of tasks, from smart city AI and robotics to drug screening.

What kind of real-world performance gains are you seeing?

CM: SpiNNaker2 significantly outperforms conventional hardware in energy and speed for specific tasks. We see one-to-two orders of magnitude improvement in both performance and energy efficiency. In drug discovery, we’ve seen up to a 100× speed-up compared with GPUs.

That’s a game changer. If you’re tailoring a drug to a specific patient, and it currently takes a GPU-based data centre a week to simulate, that’s prohibitively expensive. But if SpiNNaker2 can do the same task in minutes or hours, suddenly personalised medicine becomes a real option.

How do you achieve low power consumption and high parallelism?

CM: The system borrows heavily from how the brain

Christian Mayr: ‘The closer compute and memory are ... the less energy is wasted on communication’

handles information: selectively activating resources only when needed, whether it's compute, communication, or memory. Everything is task-dependent. If a unit isn't needed, it doesn't consume energy.

That makes the whole machine extremely efficient.

Also, SpiNNaker2 is designed for strict real-time performance. Across its full 20-rack configuration, we guarantee sub-millisecond response times.

Push data in one side and you get results out the other – fast.

This architecture is inherently parallel, like the brain.

The brain has 80 billion neurons, each with thousands of synapses, all working in parallel at frequencies between 100 Hz and 1 kHz. It doesn't rely on centralised scheduling or fixed pipelines.

What brain-inspired algorithmic principles do you apply in SpiNNaker2?

Mayr: We look at neurobiological computing principles such as dynamic sparsity, hierarchy, distance-dependent topologies, and asynchronous updates. These principles help us reframe conventional AI algorithms in ways that drastically improve the energy-delay product, by up to an order of magnitude in both inference and training.

A key concept is information gating, the idea that you only activate memory, computation, or communication when necessary.

This selective activation mimics how the brain dynamically focuses attention and resources.

In large language models, for example, you want to activate weights only when a prompt demands it, not burn energy across the whole model.

Another crucial insight: the brain doesn't only run deep neural networks. The associative cortex operates more like symbolic AI. And it's highly probabilistic, another principle we've integrated into the hardware.

How important is locality in your system design?

CM: Hugely important. Energy is consumed when bits move, not just across chips, but across boards and racks. The closer compute and memory are to each other, the less energy is wasted on communication. Co-location isn't just at the die level; it matters at every scale. That's why we've built SpiNNaker2 around chiplets, racks, and memory systems that favour tight physical integration. But that also means you need to rethink your algorithms. You can't just assume centralised data access or synchronous execution. You need to recast your algorithms to account for locality and asynchronous processing.

What role does Saxony play in enabling this kind of innovation?

CM: Saxony is Europe's hidden semiconductor powerhouse. Approximately 36% of Europe's chips are produced in Saxony. However, when people think of semiconductors in Europe, they often think of France or Belgium, rather than Dresden.

There's a complete ecosystem here: from SMEs building fab tools, to design houses, to high-density processing and chiplet ecosystems. That's what enables high-performance hardware such as SpiNNaker2 to be constructed and refined locally.

We're working on expanding its application footprint, from AI accelerators to hybrid symbolic-probabilistic AI for autonomous systems and advanced robotics. And we're focusing more on sustainable AI. AI that's fast, accurate, and energy-efficient enough to be used in real-world, large-scale deployments. **B**

Christian Mayr is the Chair of Highly-Parallel VLSI-Systems and Neuromorphic Circuits at the Technische Universität Dresden

"The system borrows heavily from how the brain handles information: selectively activating resources only when needed"

Preparing the NHS for transparent, integrated AI

While the UK Government's 10 Year Health Plan promises an AI revolution in the NHS, the reality is more complex, says Open Medical's **Harry Lykostratis.** He explains how the Keychain project aims to solve healthcare challenges in integration, scaling, and transparency

The UK Government's 10-Year Health Plan aims to leverage the potential of AI software to support clinical decision-making, improve productivity and guide patients within the NHS. The optimism for the plan may be well-founded: AI models are being rapidly developed and improved.

However, applying AI models to real-world data can be difficult and currently many organisations are struggling with the applicability of AI tools.

Harry Lykostratis, Chief Executive of Open Medical, explains how the project his company has codenamed Keychain will address challenges and create a framework for the adoption of agentic AI, supporting more efficient and effective workflows.

Lykostratis notes that he has seen some success with ambient voice technology, but in many ways, that is an evolution of existing ways of working.

Scribes may capture more detail during a consultation and generate documents with less manual labour on the part of clinicians. Still, they are not going to unleash the revolution we have seen in other industries.

What is holding health back when it comes to other applications of AI?

Harry Lykostratis: From the conversations I have had with chief clinical information officers, it seems there is a lot of interest, but it is proving hard to experiment with these technologies. Healthcare is reluctant to feed models large amounts of personal data, so they get trained on very sterile data that may come from just a few hundred cases. Integration with NHS systems is hard and expensive, so for most suppliers it is just not worth the effort to work at that level.

Models trained on just a few hundred cases are hard to scale, which makes it difficult to incorporate them into organisational workflows. We need to resolve these issues and start thinking about the near future, because over the next few years, we are going to see a move from individual AIs to agentic AIs.

Today, we ask an AI a question, and it gives us an answer. Then we ask the AI another question or use another AI to refine that output. In the future, we'll use an agent that can handle multiple outputs to achieve a specified end with much less, or no, intervention from us.

In some industries, it will be acceptable for these agentic AIs to operate as black boxes. We won't really know how they achieve their goals, and they may evolve in ways that are hard to predict. I don't think that will be acceptable in healthcare. Each step taken by an agentic AI will need to be transparent, and we'll want to be able to interrogate it to ensure patient safety.

Can you explain the Keychain project?

HL: Open Medical is addressing these challenges through the Keychain project. Our cloud-based pathway management platform, Pathpoint, already includes many of the required features.

We are present in most trusts in England, so Pathpoint can already integrate with the major electronic patient record and enterprise IT systems in use in its hospitals. It can also handle the data cleansing and pseudonymisation needed to make large volumes of data available from these systems to AI models.

We want to create a single API, so that any AI tool integrated into the Keychain product can be securely and rapidly integrated with other systems, allowing trusts to both experiment with new models and rapidly scale those that work for them.

At the same time, we want to prepare for the coming era of agentic AI. Our vision for Keychain is that it should be able to accomplish four key objectives.

First, parallel processing. Healthcare generates a large amount of data in various formats, so we need a parallel model that allows us to utilise all to produce an output.

Second, sequential processing. We want to be able to take the production of one model and use it to inform the production of another.

Third, and critically, routing. We want to ensure that these outputs are generated in the correct order to act as agents that undertake specific tasks within a clinical workflow.

Fourth, feedback. We want Keychain to enable AI models to reuse their own outputs, allowing them to improve over time.

How would this system work in practice?

HL: To see how this might work in practice, consider referral management, which is something we are already being asked to support at scale.

In the traditional model of healthcare, a patient visits their GP, who then refers them to a consultant, who advises them, refers them to another service, or places them on a waiting list for treatment.

Referral management intervenes to triage the patient and decide whether a secondary care consultation is really the most appropriate next step. At the moment, this triaging can also be undertaken by clinicians, but in the future, it might be undertaken by an agent.

For this to happen, the agent would need to be able to handle the referral data.

This might include a large PDF from the GP. There might be an additional input from the summary care record.

There might be a handwritten note from the patient or a response from an online questionnaire. There might be an image from medical photography. So, we might send the electronic information to one AI model, and the questionnaire data to another.

Then we might feed those outputs into a

Harry Lykostratis, Chief Executive of Open Medical: 'In some industries, it will be acceptable for agentic AIs to operate as black boxes... I don't think that will be acceptable in healthcare'

"Healthcare is reluctant to feed models large amounts of personal data, so they get trained on very sterile data that may come from just a few hundred cases"

further model that can decide whether to send the patient back to their GP, or suggest they visit a nurse or physiotherapist, or put them on an urgent care pathway, and generate a summary to write back to the clinical record.

The outcome should be better use of resources and the removal of a significant bottleneck to timely care, but everything is done transparently, with guardrails so the patient can be picked up and referred back into the system, if necessary.

What steps have been taken to ensure this framework can be supported long-term?

HL: Open Medical is not looking to develop AI models. We are model-agnostic: we believe trusts and clinicians should be able to use any model they prefer. We are interested in agile models that take outputs from one stage to the next.

Our ambition is to create a framework that will enable this to happen.

Keychain will be able to integrate with the systems in use in an organisation, tap into and sanitise their data, route it to AI models, take their outputs, and present them back to the organisation in a useful way within the clinical workflow.

That will improve access for suppliers and enable trusts to move from having small pockets of AI within their organisations to using it to drive efficient, timely care.

The Government is right that there is huge potential in AI, but I think that now is the moment to address applicability.

We need to address the scale challenge and prepare for agentic AI. Otherwise, these potentially revolutionary technologies will have much less applicability in healthcare than ministers, trusts, clinicians and their patients hope. **B**

Harry Lykostratis is the Chief Executive of Open Medical

Modelling malaria: Causal AI breakthrough at exascale

In a scientific landscape increasingly defined by big data and AI computation, the next frontier isn't just about finding patterns, says **Neeraj Kumar,** it's about understanding the "why?" behind them

Dr Neeraj Kumar, of the Pacific Northwest National Laboratory (PNNL), is merging causal reasoning with exascale computing to transform how science is done. His work in Causal AI and high-performance computing is unlocking faster, more transparent, and more actionable insights across domains as diverse as pandemic preparedness, quantum materials design and drug discovery.

A causal inference engine for multiscale simulation (CIEMSS) enables rapid "what-if" and counterfactual analyses, compressing months of simulation into days and delivering reproducible, uncertainty-aware insights for complex challenges in biosecurity, health and materials science.

From malaria modelling to autonomous molecular discovery, Kumar's work illustrates how causal reasoning at scale can drive agile, trustworthy and cross-disciplinary scientific breakthroughs – charting a path toward autonomous discovery and quantum-informed causal exploration.

What is causal AI?

Neeraj Kumar: I would like to share how we are pushing the boundaries of what is possible when we combine causal reasoning with exascale computing. Causal AI is not just the next evolution, but a fundamental shift in how we conduct science at scale.

We have become very good at finding patterns. Traditional machine learning, deep learning and probabilistic programming have given us powerful tools, yet correlation is not causation. Recently, I read an article from the World Economic Forum stating that, while generative AI relies on patterns, causal AI uncovers the 'why?'. The 'what?' and the 'why?' are constants in scientific discovery. Understanding why things happen, distinguishing correlation from causation and recognising cause and effect are all critical. We have advanced in identifying what happened and finding correlations, but we are far less capable when it comes to explaining why.

Why is this AI being combined with HPC resources?

NK: The "why?" matters when making critical decisions that affect millions of lives, such as those related to pandemic response, climate change, drug development, or creating new materials for quantum technologies. The challenge is that causal inference and reasoning require enormous computational resources, which is where high-performance computing becomes essential. PNNL has developed an engine to address why? questions at scale. Rewriting complex simulations into probabilistic programming languages is often impractical. Running national-scale epidemic simulations can require thousands or even millions of core hours just for inference. There is also a lack of scalable causal inference and reasoning engines.

This allows us to integrate domain-specific simulators with causal inference on HPC systems, reducing months of analysis to days. This is not an incremental improvement but a transformative one. The name of this session is Transformational Impact on Scientific Discovery, and here is why this is the right moment for causal AI.

First, exascale computing now provides machines such as El Capitan, Frontier, Aurora and Lumi with extraordinary performance measured in exaflops.

Second, advances in causal discovery have shown that robust AI reasoning requires causal models, which allow us to ask both interventional and counterfactual questions.

With computational power and algorithmic sophistication now available, we can address significant challenges in pandemic preparedness, vaccine design, drug discovery, materials for fusion energy, quantum systems, and even mapping brain connectivity. This knowledge can also help us design hardware that operates more like the human brain.

This is the first time we have had both the computational resources and the algorithms to answer not only what happened but why it happened. Our approach is a CIEMSS funded by DARPA in collaboration with the University of Texas at Austin and the Basic Research Institute, with PNNL as the lead.

How does CIEMSS work?

NK: CIEMSS has four main components: a causal query language to make reasoning explainable because black boxes are not acceptable in science, multi-fidelity methods to balance accuracy with computational cost, continuous validation to test assumptions against reality, and integration across diverse scenarios so we are part of a larger movement in causal reasoning.

At the core of CIEMSS is Cairo, written in Julia, which uses effect handlers to transform any probabilistic programme into a causal one. It can answer interventional questions, such as what happens if we do X, counterfactual questions such as what would have happened if, mediation questions about the pathways or mechanisms involved, and discovery questions about what causes what.

This is not just theoretical. We have applied CIEMSS to real-world problems at scale. Our work follows Judea Pearl's causal hierarchy. The first level is association, which identifies patterns. The second level is intervention, which simulates what-if scenarios. The third level is counterfactual reasoning, which explores alternative histories and hypothetical cases. Traditional simulations explore one scenario at a time.

CIEMSS can simulate hundreds or thousands of counterfactual worlds simultaneously, enabling us to explore the entire interventional space, propagate uncertainty correctly, and scale to previously intractable problems.

Can you give an example of a real-world application for CIEMSS?

NK: One application is modelling malaria transmission. Malaria affects nearly 250 million people each year. Traditional models treat human and mosquito populations separately. Our approach models them as one interconnected system, run on both laptops

Neeraj Kumar is the Chief Data Scientist in the Advanced Computing, Mathematics, and Data Division at Pacific Northwest National Laboratory (PNNL)

"The model employs 11 coupled ordinary differential equations with stochastic parameters, which can take an extremely long time to run on a laptop, but can be solved in a matter of days on an HPC system"

and HPC clusters. Humans move through susceptible, exposed, intervention and recovered states. Mosquito populations have a different progression, but the systems are linked. Rather than point estimates, we express parameters such as transmission rate, death rate and insecticide effectiveness as decision variables with uncertainty.

The model employs 11 coupled ordinary differential equations with stochastic parameters, which can take an extremely long time to run on a laptop, but can be solved in a matter of days on an HPC system. We calibrate the model against real-world data, apply Bayesian inference to quantify uncertainty, and identify where more research is needed. In one scenario, CIEMSS determined that a targeted 50% transmission reduction would keep infections below 7,500, preventing approximately 202,000 cases.

This is actionable insight for intervention planning. CIEMSS has also been applied to other domains, including drug discovery through an AI chemistry agent called Cactus that generates small molecule candidates, as well as materials research for fusion and quantum applications. Looking ahead, we aim to scale exponentially by overcoming current computational limits.

The future is not simply faster computation, but autonomous scientific discovery, where we design optimal experiments, create new materials and build decision-making agents that understand cause and effect.

We are also exploring quantum computing for causal discovery, where superposition enables the exploration of multiple causal paths simultaneously, and entanglement reveals quantum causal relationships.

In summary, causal AI allows us to answer not only what?, but also why? Understanding why gives us the power to shape outcomes, design new materials, develop new drugs, and address global challenges from epidemics to climate change. **B**

Dr Neeraj Kumar is the Chief Data Scientist in the Advanced Computing, Mathematics, and Data Division at Pacific Northwest National Laboratory (PNNL)

‘Simply swapping the network can lead to a 30% application performance improvement’

Cutting-edge compute infrastructure is often underutilised – hamstrung not by processing power but by network inefficiencies. Cornelis Networks CEO **Lisa Spelman** explains how purpose-built interconnect technology is delivering dramatic performance gains

Cornelis Networks is a US-based technology company specialising in high-performance networking solutions designed to accelerate artificial intelligence (AI), high-performance computing (HPC), and data analytics workloads. Founded in 2020 as a spin-out from Intel, Cornelis continues to receive backing from Intel’s venture capital arm. The company focuses on developing purpose-built interconnect technologies that address the scalability and performance challenges of modern data centres and supercomputing environments.

Can you tell us about Cornelis Networks and the company’s goals?

Lisa Spelman: I’m CEO of Cornelis Networks. As a team, we are mission-driven to solve the challenges our customers face in the highest performance aspects of their data centre work, whether it’s AI workloads or traditional HPC.

We’re focused on improving efficiency, particularly in addressing the significant underutilisation of compute resources resulting from network inefficiencies.

Billions of dollars are being invested in compute infrastructure that can deliver next-generation scientific results and AI capabilities that seemed impossible just a couple of years ago.

However, much of that infrastructure operates with only about 30%-50% utilisation of GPUs or XPUs. A major contributor to this inefficiency is the network capability. While bandwidth is important, there’s more to a high-performance, purpose-built architecture than just raw speed.

Our mission is to deliver complete network solutions that help customers accelerate scientific discovery, reduce training times and increase inference throughput by solving these inefficiencies through the network.

Network bandwidth is always top of mind in AI and HPC. How does Cornelis differentiate?

LS: Network bandwidth is indeed critical, especially for workloads such as AI training and fine-tuning. But not all bandwidth is equal. In some architectures, a lot of bandwidth is wasted resending lost packets. So, while specs may match on paper, real-world performance varies significantly.

Take our 400G CN5000 network. Compared with a 400G InfiniBand solution, we deliver 2x higher message rates and 35% lower latency. Message rate is crucial for simulations and computational fluid dynamics, as well as for any application where problems are broken down into many small messages. Lower latency benefits both AI inference and HPC, ensuring predictable performance.

Our architecture, Omni-Path, was designed specifically for highly parallel workloads at massive scale.

InfiniBand and Ethernet, while reliable and widely adopted, were not originally built for these types of workloads and are being adapted to meet new demands. In contrast, our design is inherently suited to them.

What makes your architecture stand out?

LS: We’re a silicon-through-solution company. We design the SuperNIC ASIC, the switch ASIC, and the boards and systems they go into. These are manufactured at TSMC. We also developed an open-source fabric management layer, widely adopted and foundational to the Ultra Ethernet Consortium.

On top of that, we offer a performance layer. Between our SuperNIC and switch, we maintain feature consistency, support lossless data movement, and manage congestion at both hardware and software levels. Features like credit-based flow control, dynamic adaptive routing, and enhanced congestion control are built into the architecture from the start, not retrofitted.

Can you provide a specific example of performance impact?

LS: Absolutely. Consider our 400G product versus a competing 400G network in an HPC environment. With identical compute hardware, simply swapping the network can lead to a 30% application performance improvement. That’s performance typically only achieved by a new CPU generation or a new process node, gained entirely through the network. We call this network-led application performance acceleration.

For customers, this unlocks new possibilities. With the same infrastructure, they can handle larger datasets, expanding from just North America to include Europe in genomics research, for example.

Or, they can scale their AI models from X to Y billion parameters. Alternatively, they might reduce infrastructure costs by needing fewer GPUs or CPUs to achieve the same performance, thanks to more efficient compute utilisation.

Tail latency often causes bottlenecks in AI. How do you address that?

LS: Tail latency is a significant issue in AI and HPC. We use the analogy of a marching band, if one person falls behind, the entire group must wait. AI operations require all data to be present, so even small delays can slow everything down.

We’ve built acceleration features specifically to reduce tail latencies. These ensure that all parts of a workload move forward together. Our customers see noticeable improvements in collective communications and performance at scale with Cornelis products.

Lisa Spelman, CEO of Cornelis Networks: 'Lower latency benefits both AI inference and HPC'

What are the core performance metrics you focus on?

LS: We focus on what we call the "big five": message rate, bandwidth, latency, overlap (communication/computation) and scale. These metrics directly impact application-level performance.

With our 400G line, we're targeting enterprise AI and HPC in government, academia, automotive, manufacturing, energy discovery and healthcare/life sciences. While AI and HPC share some requirements, they also differ. For instance, AI training stresses bandwidth heavily, while some HPC applications are message-rate bound. We tailor our solutions to domain-specific needs.

What's next in your product roadmap?

LS: Our next-generation CN6000, the 800G product, is well into production development. It includes Ethernet capability and interoperability alongside our Omni-Path architecture. This is especially exciting for cloud customers who want standard Ethernet compatibility with advanced congestion and lossless features.

We've innovated to include Ethernet in the same chip, offering access to some Omni-Path features. It opens up new markets, especially in cloud infrastructure.

How do you see the market adopting 800G and beyond?

LS: The cloud market typically moves fastest toward new bandwidth levels, while HPC tends to move with application advancements. Much of the HPC market is still well served by 200G or 400G networks. We'll continue to support those versions while being ready for customers who want to push the edge with 800G.

We're not forcing customers forward. Our role is to meet their needs with the best price/performance ratio, helping them make smart infrastructure decisions at the right time.

We're fully committed to the Omni-Path architecture because it's designed correctly for these workloads. We're adding Ethernet and exposing features through it in a differentiated way that matters to customers.

Looking ahead to our 1.6T product, it's already under way and will align with Ultra Ethernet Consortium (UEC) specs.

But the truth is, our 400G product already meets most of the UEC's required features: no packet drops, optimal path selection, congestion prevention, fault tolerance and software support for major XPUs.

Our 800G product is "UEC-ready," even if the consortium spec isn't finalised. We've chosen not to delay product development while waiting for standardisation. Customers are aligned with that; let's keep innovating and moving forward together.

Our goal is simple: deliver the highest network performance and help our customers maximise their compute investment.

The payoff is immense, whether through larger workloads, faster training, or reduced infrastructure costs. With our silicon-to-software architecture and continued innovation, we're enabling the future of AI and HPC. **B**

Lisa Spelman is CEO of Cornelis Networks

"Billions of dollars are being invested in compute infrastructure... However, much of that infrastructure operates with only about 30-50% utilisation of GPUs or XPUs"

Connecting qubits to photons: breaking quantum computers free

Fermilab's **Dr Silvia Zorzetti** is working on quantum transducers that could enable hybrid quantum data centres, where superconducting qubits handle computation while photonic networks carry information across distances

Dr Silvia Zorzetti is the Principal Investigator on a multi-year DOE Early Career Award that will pave the way for long-distance communication and make it easier to integrate quantum computing with fibre optic data centre architecture. Focused on high-efficiency quantum transduction across microwave and optical domains, the project aims to interface superconducting quantum computing with photonic networks. Dr Zorzetti's research aims to help preserve quantum information by focusing on improvements to the methods of transporting it. Zorzetti earned an Early Career Award from the DOE for her research proposal that will provide $2.5 million over the next five years to support her work. This also includes overseeing the co-design of hardware and software stacks for 3D quantum processors, enabling initial use cases for quantum computing.

Zorzetti played a central role in establishing the first dedicated quantum research laboratory at Fermilab and is now a leading contributor to the Superconducting Quantum Materials and Systems (SQMS) Centre, one of five national centres funded by the US Department of Energy.

Silvia Zorzetti is a Principal Engineer and Department Head at Fermilab

Can you tell me about yourself and your role at Fermilab?

Silvia Zorzetti: I'm an engineer here at Fermilab, which is the US National Laboratory for particle accelerators. We design and build particle accelerators for projects at Fermilab and worldwide. A key component of particle accelerator technology is superconducting cavities. There have been decades of research on those materials. And so the research on quantum computing stems from the research on the superconducting materials for particle accelerators. Obviously, we are focusing on superconducting quantum computing.

In 2017, the lab began investing in quantum computing. I was there from the very beginning. Around 2020, we received a grant from the US Department of Energy, which funded the establishment of five quantum computing centres. The superconducting Quantum Materials and Systems Center (SQMS), led by Fermi National Accelerator Laboratory, is one of those quantum computing centres.

Why is coherence so crucial for developing quantum computing systems?

SZ: The basic concept is that greater coherence allows more operations to be performed on a quantum computer within a given time. Extending coherence is therefore essential to increase the computational

volume of a quantum computer. Superconducting cavities are sometimes even referred to as quantum memories, since their coherence times can be extremely long.

It is important to recognise that different physical platforms for quantum computing each have their strengths and weaknesses. Superconducting quantum computing is highly nonlinear, which allows precise control of individual quantum states and makes it well-suited for computation. Other platforms, such as ion traps or bosonic systems, can achieve longer coherence times, which can sometimes last several seconds.

This diversity of advantages suggests that the quantum computer of the future may be hybrid: superconducting qubits for computation, complemented by photonic components for state transfer and long-distance communication.

Transducers, which convert quantum information between different physical platforms, will play a vital role in such hybrid systems.

What are the limitations of the current technology used in superconducting quantum systems?

SZ: One of the main limitations of superconducting quantum computing is that it is confined within dilution refrigerators. Scaling up to the data centre level, which is part of the roadmap for many companies, requires overcoming this physical barrier. A promising solution is to transmit quantum information over optical fibres, which means converting microwave signals into photonic ones. However, this conversion process is very difficult to achieve efficiently. Currently, it is hindered by noise, making reliable single-photon transmission challenging.

Overcoming this limitation requires advances in materials, better engineering design and the development of protocols for quantum state transfer.

Some of these protocols allow for purification, correcting the effects of noise to recover useful quantum information.

The bottleneck is mainly in the development of materials, but also in the limited scale of current research efforts. Most work on transducers is carried out by small academic groups or small teams such as mine, while industry is watching closely to see which approaches prove most promising.

There is currently no single best method for quantum transduction.

For example, some approaches rely on acoustic waves, which are very noisy, but very fast, allowing conversion of a small fraction of photons at high rates.

Our approach is, instead, to attempt a single, high-quality conversion with much higher efficiency, aiming to reach 50% in one step. This is extremely challenging, as it requires much purer materials than are currently available, but it offers the potential for significantly lower noise.

The good news is that we already have a working transducer. Theory shows that efficiencies of up to 50% are possible, and this figure is widely accepted as the minimum threshold for a useful quantum transducer. In other words, at least half of the photons must be converted for practical applications. With the resources available, we are working to improve materials and designs to approach this limit, while collaborating with others to share knowledge and refine current demonstrations.

How important is collaboration when seeking to address these complex scientific challenges?

SZ: Collaboration is essential. Building a quantum computer is an enormous challenge that no single institution can tackle alone.

This is why the Department of Energy established the five quantum centres to build ecosystems of institutions working together towards shared goals.

SQMS focuses on superconducting materials, while the other centres address different aspects of the problem. Our collaborators include both academic and industrial partners, each bringing complementary expertise and facilities.

Currently, the primary focus of the superconducting qubit ecosystem is on enhancing coherence and gate fidelity.

The next stage is to achieve these improvements in multi-qubit systems and circuits, rather than just for individual qubits. Building large-scale quantum data centres will require thousands of logical qubits. This means not only improving single-qubit fidelity but also reducing noise in interconnections and developing software that can manage complex systems.

If quantum information can be reliably converted into optical fibre, large-scale quantum data centres will become both economically and practically feasible. Alongside this challenge of microwave-to-optical conversion, parallelisation will also be essential. Communication qubits, which link subsystems, will inevitably have the lowest coherence and fidelity, creating a bottleneck.

The solution is to design architectures where many subcircuits perform operations in parallel, with only a small part of the process relying on communication qubits. Optimising algorithms and architectures for parallel quantum computing will, therefore, be a key part of future development. **B**

Silvia Zorzetti is a Principal Engineer and Department Head at Fermilab

“It is important to recognise that different physical platforms for quantum computing each have their strengths and weaknesses”

Optimising drug design with quantum computing and photonic networking

Finding a molecule to target a disease without harming the patient remains one of the hardest problems in pharmaceutical science. **Gian-Luca Romano Anselmetti** explains how quantum algorithms, photonic architectures and hybrid computing approaches are bringing the dream of quantum-accelerated drug design to reality

Quantum computers have made significant progress over the past few decades, evolving from experimental novelties in scientific labs to large-scale industrial efforts aimed at developing a machine capable of tackling problems currently intractable by classical computers.

The selection of problem types where we can expect an advantage from transitioning from classical to quantum remains modest. However, drug design is usually among the selected few with industrial relevance, where quantum computers are attempting to make a difference. What problems can already be translated into a quantum algorithm that scales favourably over current classical methods, and what further research is needed to make the dream of quantum computers helping to develop a new treatment a reality?

Dr Gian-Luca Romano Anselmetti obtained his PhD in physics, developing quantum algorithms for chemistry and error mitigation from the University of Cologne. Before joining Boehringer Ingelheim, he had previous affiliations with Covestro and Microsoft Quantum.

How is Boehringer Ingelheim planning to use quantum computing?

Gian-Luca Romano Anselmetti: There's a team of five of us inside the company who work on a wide range of projects, ranging from sketching out the scale of a machine needed to tackle industrially relevant problems within our company to developing the algorithms to take it a step further and show scientists where our needs lie, what kind of projects would be interesting to us, and what information to extract from these calculations.

What's the main problem a pharmaceutical company has? I would argue it's drug design. All these processes begin with a target, which is typically associated with a specific molecular pocket related to a disease. Sometimes it's inside your body, sometimes it sits on a parasite or a virus, and so on. And your "only task" is to find a molecule that fits inside this pocket and binds to this inhibitor and doesn't kill you in the process. Sounds easy enough.

It turns out to be a very hard problem, and every little bit helps in finding the right molecule for each target. Typically, all the drugs we see here, which are similar to these small organic molecule drugs, are sold in the world today. So they are out of 10s of atoms. You usually don't have super interesting atoms around it, a couple of hydrogens, a couple of oxygens and some carbons.

How can quantum make a difference?

G-LRA: Using previous generations of hardware, you needed around four million physical qubits, which then corresponded to more than 1,000 logical qubits and had a runtime of three days, which was too slow for competitive modelling at the time. However, it was the first stick in the ground that provided an estimate of how expensive these calculations could become for problems of industrial relevance.

The main issue is that it is quite a lengthy process. The ground set energy calculation that this constitutes is not an application, per se. Typically, on the classical side, this calculation occurs millions of times to enable dynamics and modelling of how things work. To know if these two systems bind or not, you have to compute thermodynamic properties, which is much more expensive.

What we did was to take this first estimate and optimise it, because we knew there were different knobs you could turn. And one of the main knobs you can turn is using in classical computing, again, the problem is modelled by the Hamiltonian matrix. Then the task becomes one of massaging this matrix into another one that, hopefully, or if done carefully, still encodes the same problem you had before, but at a lower cost to your quantum computer. Then there's something you can do that's called double factorisation, where you use your tricks in linear algebra and find a matrix that still encodes the same problem.

Gian-Luca Romano Anselmetti is a Quantum Computing Scientist at Boehringer Ingelheim

The "Active Volume" approach developed by PSI Quantum leverages photonic architecture to enhance the scalability of quantum algorithms. It is essentially a different approach to compiling algorithms for fault-tolerant quantum computers using photons that leverages the long-range connection between qubits enabled by the photonic architecture, resulting in a more compact circuit and, consequently, reduced cost.

By exploiting how their systems are interconnected more heavily than others, they can structure their problems in different ways. This is quantum addition as a circuit, which would typically scale radically because it requires back-and-forth operations.

However, in their approach to compiling this algorithm, the scaling is more linear, resulting in a significant saving.

How does the Hamiltonian matrix represent problems in quantum chemistry, and how can techniques like double factorisation reduce computational costs?

G-LRA: After setting the chemical problem up from first principles (ab initio), e.g. just assuming quantum mechanics and adding in a description of the molecule in terms of which atom sits where in 3D space and a list of the corresponding orbitals associated with each type of atom, you arrive at the Hamiltonian representing/encoding the problem in question. Picture this as a large matrix, and now the task becomes finding a solution to this matrix that minimises the energy and gives you the electronic configuration that nature would have likely prepared at room temperature.

To find this solution, classical and quantum methods exist that attempt to tackle the problem. Focussing on the quantum side of things, the current best quantum algorithm scales in runtime what is called the 1-norm of the Hamiltonian (essentially just a sum over all the entries, as each entry must be loaded on the quantum computer).

Therefore, to decrease the cost, one tries to find a similar, but different, Hamiltonian to the first one that still very closely encodes the original problem of interest but has a lower associated cost.

Techniques such as double factorisation can find alternative representations with these reduced costs.

In what ways can hybrid approaches shorten computation times and improve practicality?

G-LRA: Hybrid approaches integrate classical pre-processing with quantum acceleration to optimise the overall computation process. As different operations are cheaper or more expensive on classical or quantum hardware, for example, addition is incredibly cheap on classical computers but relatively expensive on quantum computers.

The main difficulty in this approach is that quantum computers are currently believed to be I/O limited; it is hard to load and read out data, so switching between classical and quantum computers frequently increases cost significantly.

Currently, it remains an open research question which part of the calculation should be performed quantumly.

What barriers remain before quantum computers can routinely support pharmaceutical drug discovery pipelines?

G-LRA: Several barriers need to be addressed before quantum computers can be routinely used in pharmaceutical drug discovery. On the hardware side, these include improving qubit fidelity and developing more effective error correction methods in both software and hardware to build a large enough machine to host calculations that are too large for classical computation to overcome its current limitations. Additionally, regarding the software, we still need significant development on the algorithmic side to flesh out further problems that can be solved efficiently on a quantum computer, providing an advantage over classical computing.

How might advances in quantum hardware, such as better interconnects or improved qubit fidelity, change the role of quantum computing in the industry?

G-LRA: Advances in quantum hardware, such as better interconnects and improved qubit fidelity, will significantly enhance the capabilities of quantum computers. These improvements will lead to more accurate and reliable computations, enabling quantum computing to play a more substantial role in various industries, including pharmaceuticals, by solving problems that are currently intractable with classical computers. The cheaper a computation becomes, e.g. the error correction overhead decreases, the earlier one can expect to have a quantum advantage.

Why is international collaboration between academia, start-ups, and industry essential for scaling quantum applications in drug design?

G-LRA: International collaboration is crucial for scaling quantum applications in drug design because it brings together diverse expertise and resources. Collaborations between academia, start-ups, and industry foster innovation, accelerate the development of new technologies and ensure that advancements are effectively translated into practical applications. There are still many challenges on both the hardware and software sides to be overcome to bring this to industrial applications. Some of the work is better suited for large industrial efforts (e.g., scaling hardware), while others, such as the development of principal algorithms, are more suited for academic research. Therefore, this collaboration is necessary to overcome the challenges and fully realise the potential of quantum computing in drug design. **B**

Gian-Luca Romano Anselmetti is a Quantum Computing Scientist at Boehringer Ingelheim

“They utilise a photonic architecture, which allows them to make more connections between their individual elements”

Cracking chemistry's quantum code (and why it needs classical HPC)

Quantum computers promise to solve intractable chemistry problems, but the qubits required are hampered by noise and coherence limits. **Prof Peter Coveney** explains why the future lies in tightly integrating quantum processing units with high-performance classical computing

At University College London (UCL), Professor Peter Coveney holds a chair in Physical Chemistry, is Director of the Centre for Computational Science (CCS), is an Associate Director in the Centre of Advanced Research Computing and is an Honorary Professor in Computer Science.

He is a Professor in Applied High-Performance Computing (HPC) at the University of Amsterdam (UvA) and an Adjunct Professor at Yale University School of Medicine (USA).

Coveney's research covers a broad area of interdisciplinary research, including condensed matter physics and chemistry, materials science, as well as life and medical sciences. In all of these areas high-performance computing plays a significant role. He has led numerous large-scale projects, including the EPSRC RealityGrid e-Science Pilot Project (2001-2005), its extension as a Platform Grant (2005-2009), and the EU FP7 Virtual Physiological Human (VPH) Network of Excellence (2008-2013).

What are the current limitations of quantum computing?

Peter Coveney: The scale of electronic structure calculations feasible on current or near-term quantum hardware is constrained by several inherent limitations, including coherence time, qubit count and connectivity, and device noise.

All these limitations, taken together, severely impact the number of qubits that may be put to work constructively for chemical applications. While we have access to quantum computing devices up to and exceeding 100 qubits, only a fraction of these can be utilised effectively. However, if these resources target a key subcomponent of a molecular system, they can still be of significant value. This fosters the integration of conventional HPC resources with quantum processing units to address problems of scientific interest in which only a key sub-component of the system is analysed on the quantum device.

My talk at ISC focused on integrating quantum computing and quantum processing units with pipelines and computing systems, not for their own sake, but to enable different types of research that would not be possible otherwise. I want to convey to you why that's actually important today, and, inevitably, it's quite a challenging topic because it involves people with interesting scientific applications. It involves a quantum computing company. Most of the work I'm going to discuss actually is being done with iqm, which is this Finnish/German company, collaborating directly with both us and the Leibniz Rechenzentrum which is in Garching near Munich.

What is the purpose of this collaboration?

PC: They've a quite advanced approach to integrating quantum processing units onto one of their supercomputers. The collaborators include participants from different institutions in the UK, the US, and Munich.

This particular application scenario exploits a natural affinity between the problem and the

ArtemisDiana/shutterstock.com

Professor Peter Coveney is the Director of the Centre for Computational Science UCL

"While we have access to quantum computing devices up to and exceeding 100 qubits, only a fraction of these can be utilised effectively"

required hardware, specifically the quantum computing infrastructure, as it's fundamentally focused on advancing capabilities in quantum electronic structure calculations.

Quantum computing for quantum chemistry has been a long-standing area of research and is a significant application. Why is that? Conventional computers, when attempting to solve electronic structure problems for molecular systems, encounter a wall of intractability very quickly. They're solving a Schrodinger equation for a many-body system. And the concept of the parameter of interest here being n, you could think of it as the number of electrons in the problem, or it's, more accurately, the number of spin orbitals that are just used to describe the electronic configuration.

The problem is that conventional algorithms scale in a nasty way with that number n. And, the most accurate calculations that you could do, which are called full configuration, are worse than exponential in n. It's factorial in n. This means that you can't obtain exact solutions to electronic structure problems unless you're working with very small molecular systems.

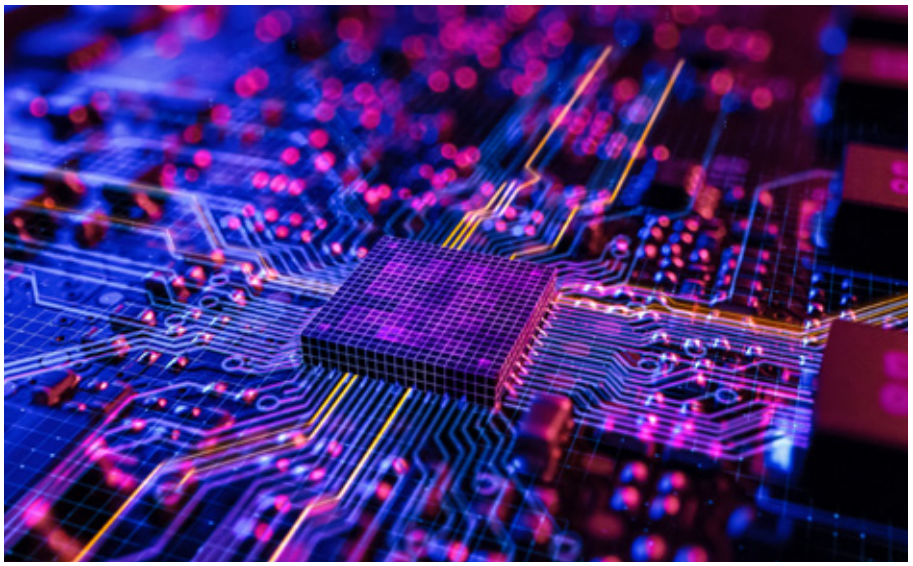
There's a whole hierarchy of approximations and so-called levels of theory that guide the field in its attempt to examine systems of interest. And what kind of systems might we be concerned with? Most quantum computing applications for quantum chemistry have been limited to examining gas-phase molecules and small molecules. But how small are we talking about? Well, if it's a gas-phase molecule, it's not usually very large. The type of molecule that might interest a pharmaceutical company is a small molecule, which is a small organic compound consisting of tens of atoms. The number of electrons becomes extremely large, rendering a full configuration interaction calculation impractical. We have to use approximate methods.

How can quantum computing solve this problem?

PC: Quantum computing is designed to solve larger problems in contexts where they are intractable on conventional computers. Our primary challenge in electronic structure theory is to understand electron correlation, which refers to the interaction of electrons with one another and the resulting structures within molecules, thereby determining their behaviour.

You can think of cases where you have a single covalent bond between two nuclei. That would typically be two electrons in a bond. Still, there are cases where you may have multiple electrons, so you could have a double bond with four electrons, or even, let's take a nitrogen molecule, it's got a triple bond, so there's six electrons crammed between the nuclei, and the correlation effects there really become very important to capture.

If you don't perform the calculations accurately, you won't obtain meaningful >


Gorodenkoff/shutterstock.com

results. But the problem here is, as is said in the last line, the accuracy comes with high computational cost, and the hope would be that you can deal with this on a quantum computer more effectively than you can on a conventional machine.

The reason for this is that, if the molecular system is quantum mechanical, you represent its state by a wave function. That wave function has to be represented on a conventional, or in the jargon, classical computer, in terms of two to the power n bits, where n, as I mentioned earlier, is the number of spin orbitals. So it becomes exponential in n, and you can't handle the size of the molecules very effectively, whereas on a quantum device, in the Hilbert space, it can be represented linearly, so you're suddenly going from an exponentially large problem to something that can be described linearly.

What are the limitations of today's quantum systems?

PC: That sounds good, and that's what all the happiness about quantum computing revolves around here. But the reality is that these computers today, in this NISQ (Noisy Intermediate-scale quantum) era, with a very high degree of noise and unreliability in the machines, rapidly prevent you from exploiting that potential benefit. So you have a huge amount of noise. You need to control the noise. You need to mitigate all the errors that take place on these qubit devices. And there's also a problem with the coherence time, which is the time you have available when the wave function is in its sort of superposition state, which is the state when you can exploit the quantum parallelism, which can often be in the sort of microsecond time domain or less.

So your calculations can never last very long. You have to do a large number of them. Additionally, designers of quantum devices are creating topologies because they have determined what would be beneficial for them. Still, it's not optimal for the computations you're interested in.

As a result, the design of the topology can, on its own, hinder the application.

And this is what often happens. So, you can be as we are, in the situation where we have access to the entire IBM Cloud, which has now got up to 156 qubit chips, the Heron chips, with a plan for IBM to take the quantum to 1,000s of qubits by using quantum interconnects between those in the sort of sense we deal with conventional computers.

But the reality is, at the moment, we can't make more than a small handful of those qubits on a single chip play a tune.

So, it doesn't matter if you have a 127-qubit device; you can't utilise most of those qubits together. The fidelity is so low; the struggle is to reduce the size of the calculation in terms of the qubits needed to get onto the device, while constantly controlling the noise and dealing with error mitigation.

So there's this problem, which is paradoxical and ironic, that you've got this quantum device, but what you want to do is touch it as it were, as infrequently as you possibly can, and make sure that when you do use it, you're using it without being overwhelmed by the noise, and that means the quantum capabilities are rather diminished.

However, suppose you attempt to apply those quantum capabilities to an interesting problem that involves more than a single gas-phase molecule.

In that case, you will need to interact with a conventional computer, which can handle that scale of problems much more effectively.

Quantum hardware on its own in a cloud is unable to support larger and more interesting scientific problems. You have to shift data. You have all these problems, including the difficulty of moving data from the classical compute to the remote site, as well as latency issues and so on. Many people who start using a quantum computer simply have a laptop and attempt to run qubit jobs on, for example, an IBM Cloud or a public version of it, which restricts their capabilities.

How can HPC help to accelerate quantum computing?

PC: What we're after is where the two things come together closely. I would say that LRZ is one of the world leaders in trying to make this sort of thing work with IQM devices. The work I'm referring to here involves 20-qubit devices, but IQM has now developed a 54-qubit series of devices called Emerald.

The interesting aspect of the architecture is that it features a square topology. In contrast, IBM's heavy hex topology restricts, as I mentioned earlier, the capability to perform longer-range computations with more qubits. However, we have the IQM chip, which we can link to the superMUC-NG machine located at the LRZ facility.

One of the problems we've been interested in is understanding how protons hop between water molecules in bulk water. Note, I'm not talking about a single water molecule any more. I'm talking about a group of these molecules and atoms within them.

This scale of calculation is amenable to a quantum computer, as I described. Still, it's in the context of the bulk fluid, because water molecules, when they club together, form water, not just individual water molecules through the interactions they exhibit, and some of those we will connect to through the multiscale element. Suppose you're into multiscale modelling and simulation. In that case, you'll be familiar with the idea that you can connect different levels of as it were, physical representation of a system. You need to do some parts of the problem in more detail, others in less, and you couple the two things together, and that's the spirit in which we're now adding a quantum computational.

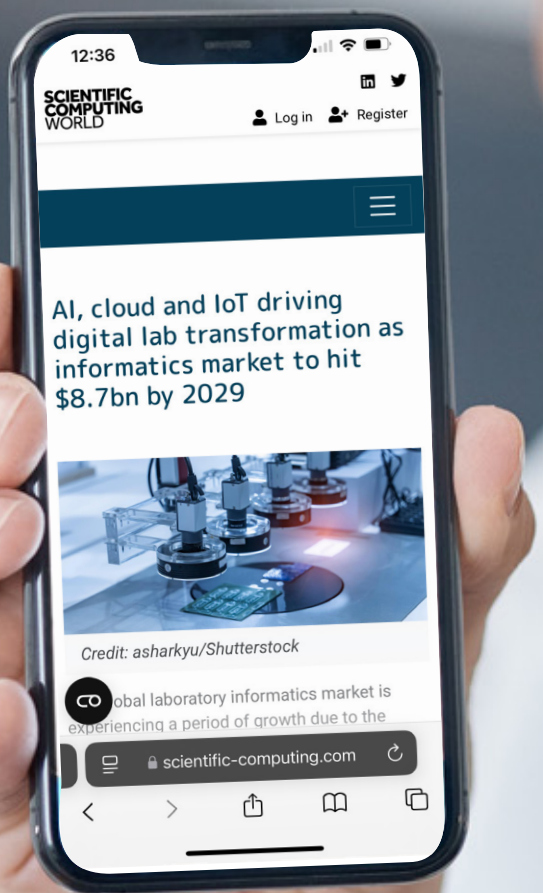
In the situation I described, I've a proton that's jumping between two water molecules. We can take that out of the broader picture of the quantum mechanical representation that's performed on a classical machine using density functional theory and a wave function calculation of what that proton is doing as it moves across.

That's where the quantum computer excels in effectively representing the data. We can sample the wave function on the quantum computer, retrieve the sample results and perform the remaining calculations on the classical device. **B**

Professor Peter Coveney is the Director of the Centre for Computational Science UCL

Stay informed.
Stay connected.
Stay ahead.

Get expert insights, in-depth analysis & industry updates delivered your way.



Visit
scientific-computing.com

Follow
linkedin.com/showcase/scientific-computing-world/

SCIENTIFIC COMPUTING WORLD

Why quantum chemistry needs better theory and algorithms, not just more qubits

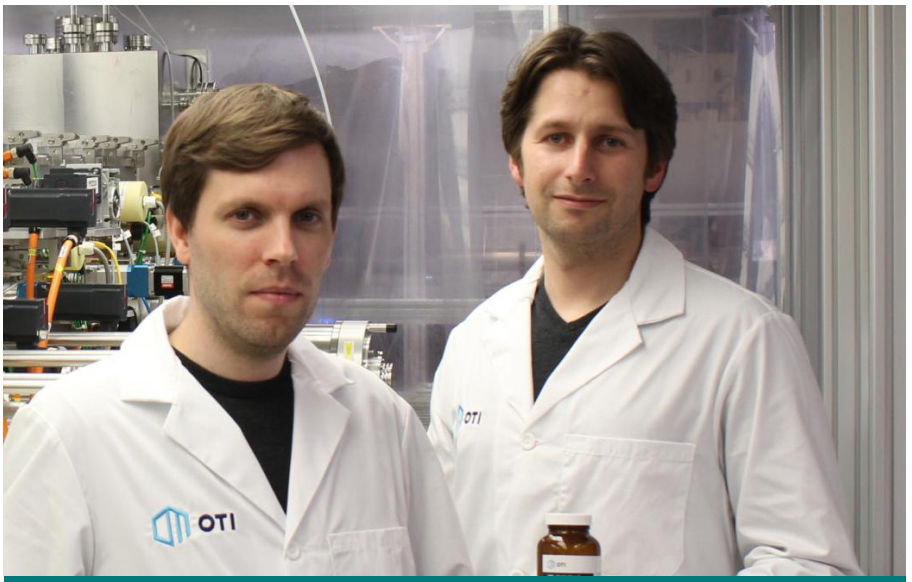
Scott Genin explains how OTI's qubit-centred algorithms achieve high-efficiency simulations on emulators, outperforming traditional materials discovery methods

OTI's team of theoreticians, quantum chemists, computer scientists and software engineers are building powerful quantum and quantum-inspired software to solve the most complex problems in materials science and chemistry.

This involves teams of theoreticians and quantum chemists developing key foundational theory to enable quantum computers to solve the most difficult and high-value problems. A team of quantum chemists and computer scientists builds upon the foundational work done by the quantum theory team to develop quantum and quantum-inspired algorithms for computational chemistry.

Software engineers can then begin the process of translating quantum algorithms into optimised code that can be run on any quantum computing hardware.

OTI Vice-President of Materials Discovery Scott Genin discusses the work done to harness quantum computing to solve challenges and how it could open up new possibilities for materials discovery.



Scott Genin (right), Vice-President of Materials Discovery at OTI Lumionics

Can you provide an overview of your work applying quantum computing to materials discovery?

Scott Genin: For the past seven years, OTI has developed and implemented efficient materials discovery methods to create novel materials in the semiconductor and display industry. This is demonstrated by OTI's cathode patterning material product line.

Developing a novel material requires a deep understanding of the underlying physics that enables a class of materials to have a structure-property relationship. In general, this means that machine learning (ML) and artificial intelligence (AI) models are not suitable for such a design problem, as they require multiple examples of materials that work. In contrast, we often start with minimal information.

High-accuracy quantum chemistry simulations enable us to achieve two goals. First, we need to understand the underlying physics and, second, build reliable databases for ML/AI models.

There are far too many ML/AI models for simulating density functional theory (DFT) energies; however, these models are neither useful nor relevant, as small changes in the basis set, DFT functional, or geometry of the material can significantly alter the output.

There is another problem with DFT, in that it makes too many simplifications of the electron-electron interactions within a molecule, which can lead to issues when designing OLED materials. A small change in either phosphorescence emission spectra or optical absorption can result in a material failing to meet application requirements.

Quantum computing presents a unique opportunity, as it holds the potential to enable high-accuracy simulation methods that surpass the accuracy and reliability of DFT.

OTI has investigated the fundamentals of modelling electron-electron interactions in qubit space. A slight misconception is that a Universal Quantum Computer will exactly simulate electron-electron interactions, but this is not quite the case – most qubits are considered para-fermions instead of just fermions.

It's not an issue since we already know how to transform fermionic Hamiltonians into Qubit Hamiltonians. However, it is a gross oversimplification to assume that all fermionic logic will equally apply in qubit space.

When we founded OTI, the assumption was that unitary fermionic logic would easily apply to unitary qubit logic, but this wasn't the case. Fast forward to today, and OTI has demonstrated through multiple research articles that developing qubit-centred logic for electronic structure algorithms is far more efficient.

During the development of theory, OTI has consistently focused on evaluating whether quantum chemistry simulations executed on a quantum computer would outperform current methods that run on conventional computers.

The current focus among academics and other industry researchers has been on simulating small, insignificant systems, such as a chain of hydrogen atoms or a few orbitals of a Diels-Alder reaction.

It's been fairly disappointing to see how some of these groups manipulate the system to force them onto a quantum computer by making simplifications that just can't be done in a real materials discovery process. OTI had to develop its own quantum emulator that uses the qubit coupled cluster (QCC) method to simulate materials up to 300 qubits, currently, which we have applied to simulate large-scale systems

as demonstrated in our academic publications (Estimating Phosphorescent Emission Energies in IrIII Complexes Using Large-Scale Quantum Computing Simulations). To date, no physical quantum computer has come close to matching the scale at which OTI can efficiently simulate in qubit logic.

What are the main challenges in materials science and chemistry that quantum computing has the potential to solve?

SG: There are two main applications: electronic structure problems and the vibronic spectra problem. These are two core calculations that lead to applied quantum chemistry.

How does quantum computing provide performance advantages over classical computing for materials discovery?

SG: The physical quantum computers currently do not provide an advantage, but we know from our research that solving quantum chemistry problems using qubit logic does provide better results than standard methods.

What specific role does quantum computing play in mapping excited states or modelling chemical reactions?

SG: Using quantum computing methods such as the QCC, we know that the resulting calculations are variational and properly incorporate electron interactions.

Modelling excited states accurately is difficult using low-accuracy methods such as Time-Dependent DFT and is known to cause prediction issues when simulating certain types of OLED materials. Since the methods on quantum computers are known to result in high-accuracy calculations, these methods will be able to accurately model excited states and chemical reactions.

Could you discuss the quantum computing-based algorithms your team developed, and how they're utilised for solving material science problems?

SG: Our team developed the QCC theory, which is based on the principle of developing an Ansatz or instruction set in qubit space, as opposed to fermionic space. A universal quantum computer does not actually understand what fermions are, since a universal quantum computer runs on the principle of Pauli operator logic, opposed to anti-symmetric fermionic excitation operators. This is not necessarily a problem since a fermionic excitation operator set when transformed into qubit space is, effectively, identical with respect to their eigenvalues; a direct translation of a fermionic excitation operator into qubit space can be inefficient.

OTI, with the University of Toronto, developed the first method to evaluate the sections of an Ansatz to determine which components actually contribute to energy minimisation. This theory was developed before other derivative methods, such as ADAPT-VQE. The QCC Ansatz does not suffer from the barren plateau problem and is still the most efficient construction for an electronic structure Ansatz to date.

Can you provide examples of how your quantum algorithms address strong electron correlation and molecular geometry?

SG: Strong electron correlation is addressed by the algorithm through similar principles as how Configuration Interaction addresses strong correlation. This is achieved by determining the combination of Pauli operators that creates the combination of determinants that minimises the energy to the ground state.

For the second part of the question, this is a misconception. Molecular geometry is addressed through a geometry optimisation calculation. This calculation computes the forces that are acting upon the atoms, which iQCC can contribute to by accurately calculating the electronic structure of the molecule. This will lead to a better understanding of the geometry for strongly correlated systems.

Can you explain the difference between quantum algorithms and quantum-inspired algorithms, and how both are utilised in your platform?

SG: Quantum algorithms are algorithms that use quantum circuit logic or wavefunction propagation via an Ansatz to solve a problem using logical qubits that are composed from multiple physical qubits. Quantum-inspired algorithms perform the same logical operations, except they use classical compute hardware. Both algorithms should produce the same result, it's just is a matter of which one takes more time.

How does your quantum software engineering team translate algorithms into optimised code for quantum computing hardware?

SG: Very carefully.

Are there any specific quantum computing hardware systems you are focusing on, or is your software able to work across multiple platforms?

SG: Our software is compatible with multiple platforms. We are currently focusing on working with Nord Quantique since their hardware is natively error-suppressing. Electronic structure problems will require zero error to run, not just error correction, so it makes sense to work with hardware that is already robust and has high-speed gate operations. **B**

Scott Genin is the Vice-President of Materials Discovery at OTI Lumionics

"It's been fairly disappointing to see how some groups manipulate the system to force them onto a quantum computer by making simplifications that just can't be done in a real materials discovery process"

From trapped ions to treatment: how million-qubit architectures could help to tackle endometriosis

Dr Mark Webber explains how high-fidelity, long-range connectivity and error correction are paving the way for quantum drug discovery in underserved regions

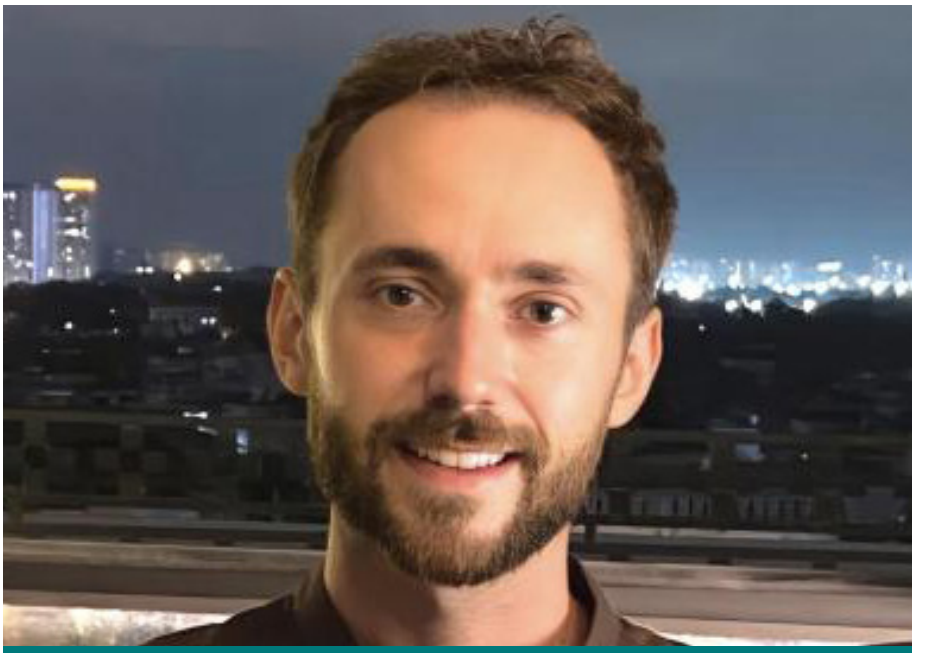
Universal Quantum has joined the Open Quantum Institute (OQI), an initiative hosted by CERN in partnership with the Geneva Science and Diplomacy Anticipator (GESDA), to launch a pioneering effort applying quantum simulation to endometriosis, a chronic condition impacting approximately 10% of women globally.

The collaboration will unite experts in biomedicine and computational chemistry to explore non-hormonal, shelf-stable therapeutics, embedding quantum insights into drug-discovery workflows to benefit underserved regions.

UQ plans to advance the field of quantum computing by engineering modular trapped-ion quantum systems designed to scale to millions of qubits, all within an industrially manufacturable framework. Their core innovation lies in the iQPU (integrated Quantum Processing Unit), a trapped-ion chip built on standard 200mm silicon foundry technology, housing hundreds of qubits driven by global microwave control instead of complex laser optics.

To connect these modules, UQ developed UQConnect, a record-breaking interconnect that enables 2,424 links per second at an ultra-high fidelity of 99.999993%, paving the way for scalable inter-module communication efficient enough to support million-qubit architectures.

However, Universal Quantum's Dr Mark Webber stresses that there has always been a focus on solving real-world problems, not just building the largest quantum system.



Dr Mark Webber is the Head of Quantum Algorithms at Universal Quantum

Can you tell us about yourself and your role at UQ?

Mark Webber: I'm the Head of Quantum Algorithms at Universal Quantum. Before that, I completed my PhD with the Ion Quantum Technology Group at the University of Sussex, the research group from which Universal Quantum spun out several years ago.

My PhD was in theory, while my colleagues focused on building trapped-ion hardware. I worked on how to use a quantum computer and optimise its performance. I joined UQ in its early days when there were about eight employees. Today, we have more than 100. Over the past four-to-five years, it's been rewarding to see the company and my own team grow.

My team has two main areas of responsibility. First, we work with end users to understand applications, mapping classical problems to quantum algorithms and then onto our hardware. Second, we focus on quantum error correction, which is crucial because today's qubits are not reliable enough for large-scale applications. Isolating a quantum system from its environment while maintaining precise control is highly challenging, resulting in imperfect qubits. For example, an error might occur roughly once in every thousand operations. To run meaningful applications, we need far better quality qubits, and error correction provides that through redundancy, similar in spirit to classical error correction in communication systems.

How many qubits are required for practical applications? Are all qubits equally useful?

MW: In practice, we need 100-1,000 physical qubits for each logical qubit. In our case, each qubit is a trapped ion atom, and scaling to practical applications requires hundreds of thousands or millions of qubits, many orders of magnitude beyond today's devices, which are in the hundreds.

UQ's mission has always been to tackle this scalability challenge, focusing on how to reach a million qubits as efficiently as possible.

Different hardware platforms have their own strengths and weaknesses. Superconducting qubits (IBM, Google), photonics and trapped ions all vary in terms of fidelity, operational speed and connectivity. Trapped ions are known for the highest fidelities, while superconducting qubits are faster.

Connectivity is another key differentiator: superconducting systems typically allow only nearest-neighbour interactions, requiring chains of gates to connect distant qubits, which introduces errors.

By contrast, trapped ions can be physically shuttled, enabling high-fidelity long-range

2018 Group photo from the Sussex Centre for Quantum Technologies

connectivity. This has significant benefits both for near-term algorithm design and, especially, for error correction, where long-range connectivity allows for more efficient codes, thereby reducing the number of qubits and runtime required.

How do you choose which applications are likely to provide a good candidate for quantum computing?

MW: When choosing suitable quantum algorithms, the key idea is the complexity advantage. Quantum computers are not faster versions of classical computers; they solve specific problems in fundamentally different ways. Problems that scale exponentially on classical machines, such as factoring large numbers or simulating molecules, are well-suited to quantum algorithms.

Chemistry simulation, in particular, is one of the most exciting applications.

Classical computers quickly become intractable when modelling molecules of realistic size and complexity, while quantum computers can exploit quantum properties directly.

In drug discovery, for example, companies already rely heavily on computational methods, but classical simulations are limited in accuracy and scale.

Approximations and machine learning help, but still fall short. Quantum computers could one day provide exact simulations of complex biological environments, unlocking insights into drug binding and molecular interactions.

In the near term, hybrid approaches, combining quantum computers with classical techniques, may provide earlier benefits.

This is particularly relevant to our collaboration through the OQI project, where we focus on applying quantum chemistry to women's health and endometriosis. Universal Quantum's vision is to build quantum computers for the benefit of humanity, so working on impactful real-world applications such as this is a natural fit.

What are the main challenges remaining for fault-tolerant quantum systems?

MW: The main hurdles are both technical and collaborative. On the technical side, we need much larger, error-corrected hardware.

On the collaborative side, we must work closely with domain experts in areas such as drug discovery, who understand the bottlenecks of classical methods. Our role is to identify subroutines where quantum computing can provide an advantage and integrate them into existing workflows. We've followed this model in projects with partners such as Rolls-Royce on fluid dynamics, and we're taking a similar approach in life sciences.

How does the collaboration with OQI help accelerate quantum development?

MW: Being part of OQI is valuable because it connects us with a broad ecosystem of researchers and industry partners, helping shape how we prioritise algorithm development. We're organising workshops to bring communities together and align on realistic goals: what machine specifications are needed, what timelines are feasible and how quantum computing will complement classical methods.

Finally, we design algorithms with scalability in mind. While we explore demonstrations on near-term hardware, our focus is on algorithms suited for large-scale, fault-tolerant quantum computers. That's where we expect the most significant, long-term advantages to be realised. **B**

Dr Mark Webber is the Head of Quantum Algorithms at Universal Quantum

"Isolating a quantum system from its environment while maintaining precise control is highly challenging, resulting in imperfect qubits"

‘A simple idea could lead to a dramatic new quantum speed-up’

Professor Andrew Childs explains how new simulation algorithms achieve exponentially better error scaling, and why identifying problems where quantum beats classical remains one of the field’s greatest challenges

Andrew Childs is a professor in the Department of Computer Science and the Institute for Advanced Computer Studies (UMIACS) at the University of Maryland, College Park. He serves as interim director of UMIACS and was co-director of the Joint Center for Quantum Information and Computer Science (QuICS) from 2014 to 2024. Currently, he directs the NSF Quantum Leap Challenge Institute for Robust Quantum Simulation, a prestigious national initiative advancing quantum simulation technologies.

Childs has been recognised for his contributions to quantum simulation and Hamiltonian dynamics. His work has focused on developing fast quantum algorithms for simulating Hamiltonian dynamics, which form the theoretical foundation for understanding how quantum systems evolve. His research includes the development of efficient methods for simulating Hamiltonian dynamics using truncated Taylor series approximations and time-dependent Hamiltonian simulation with optimal scaling properties.

Before joining the University of Maryland, Childs was a DuBridge Postdoctoral Scholar at Caltech from 2004-2007 and held faculty positions in Combinatorics & Optimization and the Institute for Quantum Computing at the University of Waterloo from 2007-2014. He received his doctorate in physics from MIT in 2004.

In 2024, Childs was awarded the Kirwan Faculty Research and Scholarship Prize, recognising his contributions to quantum computing research. His most recent work includes developing optimal routing protocols for quantum atom arrays and advancing quantum symmetrisation techniques.

Andrew Childs is a professor in the Department of Computer Science and the Institute for Advanced Computer Studies (UMIACS) at the University of Maryland

How did you first get involved with quantum computing and algorithm development?

Andrew Childs: I am based in the Computer Science Department at the University of Maryland (UMD), where we have a quantum computing centre. I came to UMD about 10 or 11 years ago, when the centre was just being established, to help set it up. It is a joint centre with National Institute of Standards and Technology (NIST). Interests include frequency standards, time standards and cryptography, all of which are connected to quantum computing. That was the origin of the centre.

I became interested in quantum computing as an undergraduate student, more than 25 years ago. This was not long after the discovery of Shor’s factoring algorithm in the mid-1990s. I was studying at Caltech, which had an active group working on quantum computing. It was one of the most exciting places to work in the field at the time. I became involved there and have been working in quantum computing ever since.

My interest in the subject comes largely from my background. I was originally a physics student as an undergraduate, and my PhD is also in physics; however, I have always been interested in mathematics and related fields. At Caltech, researchers such as Jeff Kimball and others were already exploring experimental approaches to quantum information processing, and that had a strong influence on me.

As technology advances, will the use cases increase, or are there specific types of mathematical problems that are suited to quantum computing?

AC: Looking ahead, I hope to see more applications of quantum computing. It is difficult to predict exactly, but one important point I emphasise is that quantum computers require structure to deliver significant speed-

ups. Results show that for certain types of problems, such as unstructured search, the best possible improvement is only quadratic in nature.

This is what Grover's algorithm achieves: a square root speed-up, which is helpful but relatively modest. For such an advantage to be worthwhile in practice, a quantum computer would need to be extremely fast, reliable and able to handle very large problems.

Much of the field is therefore focused on problems where it may be possible to obtain exponential or, at the very least, super-polynomial speed-ups. These require a special structure, and while we are aware of a few examples where that structure exists and can be exploited, we lack a general understanding of what kinds of structure lead to a significant quantum advantage.

This makes it a broad and difficult question. As a result, quantum computing will likely remain useful mainly for specialised purposes, although it is too early to say definitively.

Another point I emphasise is that our understanding of the power of quantum computers primarily comes from what can be proven mathematically. Unlike in classical computing, we cannot simply create an algorithm, test it on large-scale instances, and observe whether it performs better in practice, because current quantum devices are still small and limited.

They are improving, and we may soon enter a regime where they surpass classical capabilities, but we are not yet at the stage where large-scale experimentation is possible.

In classical computing, many techniques are adopted because they work in practice, not because they are backed by strong theoretical guarantees. Machine learning is a clear example: large models produce highly useful results even though there are no theorems proving their performance.

If we had access to much larger quantum computers, we might well discover new use cases and behaviours that theory alone has not predicted. Whether this will be the case remains uncertain, but without the ability to test large systems directly, our knowledge is necessarily limited.

What I find most interesting about quantum computing is the abundance of mathematical challenges. It connects with many different areas of mathematics and provides a rich setting in which to explore a wide variety of questions. That is what I enjoy most.

It is not that I am indifferent to the prospect of building a quantum computer. On the contrary, that is an extraordinary and important challenge, and achieving it would have many fascinating consequences. However, I am equally satisfied working on the mathematical problems that arise independently of whether such a machine exists. Quantum computing provides a compelling framework for thinking about information processing, naturally generating a wealth of problems to explore.

"For such an advantage to be worthwhile in practice, a quantum computer would need to be extremely fast, reliable and able to handle very large problems"

What are your current areas of research?

AC: I have several research threads. One area I am particularly interested in is simulating quantum mechanics with quantum computers. This was one of the earliest proposed applications of quantum computing, and perhaps remains the most natural: using these devices to simulate physics in a highly controlled way.

There are many fascinating questions here, such as how to design algorithms that run as quickly as possible and what happens when they are applied to specific systems? This is a major focus of my work.

I am also interested in more fundamental questions about the power of quantum computers, particularly within the model of quantum query complexity. This model is appealing because it is very precise: it allows one to analyse problems in a setting where tight characterisations are possible. In the standard model, where an algorithm is given input data and asked to compute something, proving lower bounds is notoriously difficult. By contrast, in the query model one can determine how many queries are required to solve a problem, giving a clear measure of computational speed. I am pursuing a number of projects in this area.

Another line of research concerns the movement of quantum information under locality constraints. Consider a quantum processor in which some qubits are directly connected while others are far apart on the chip. If one wants to perform a gate between distant qubits, it is possible in principle to move the information, perform the gate, and move it back. The challenge is to determine the most efficient way to do this, given the geometry of the system and the allowed interactions. This raises a host of interesting questions about how best to transport and manipulate quantum information under such physical constraints.

If we think broadly about what can be computed efficiently, that is, within polynomial time, then the cost of moving qubits around only introduces relatively modest overheads. In this view, it does not make much difference whether the qubits are arranged on a line with only nearest neighbour interactions, or whether all qubits can interact directly. These variations only affect the running time by polynomial factors. However, if the goal is to optimise algorithms and make them as fast as possible, then such details become crucial. I tend to consider problems from both perspectives.

What research milestones excite you most?

AC: One particularly exciting development has been the creation of algorithms for simulating Hamiltonian dynamics with much >

Ar_TH/shutterstock.com

better scaling in terms of error. The earliest methods had overheads that grew polynomially with $1/\varepsilon$, where ε is the error. In contrast, newer approaches achieve running times that scale polynomially with $\log(1/\varepsilon)$, which is exponentially faster. This breakthrough led to an entire class of simulation algorithms that have since been generalised and developed into a new way of thinking about Hamiltonian simulation and related problems.

Is it an exciting time to be in quantum computing? It seems like there is a lot of new ground to cover?

AC: Quantum computing is still a relatively young field. In some sense there is more low-hanging fruit to be found, although much of it has already been picked as the field matures. Over time, the research has become more technical, with papers growing longer and more detailed. Yet alongside this trend, there remain broad and open questions. It is still possible that a simple idea could lead to a dramatic new quantum speed-up and an entirely new class of algorithms. That possibility continues to make the field very exciting.

What are the biggest algorithmic bottlenecks when simulating complex quantum systems?

AC: One of the most challenging aspects of Hamiltonian simulation, in terms of limiting the potential advantage from quantum computers, is the existence of highly effective classical algorithms for many problems of interest. In quantum chemistry, for example, although I am not a computational chemist, my understanding is that many of the problems practitioners care about can already be addressed successfully with classical methods. Density functional theory and other heuristic approaches may not work efficiently in every case, but they are remarkably effective for a wide range of problems of practical importance.

This makes it essential to identify simulation problems that are both scientifically valuable and resistant to classical techniques. The challenge lies in combining an understanding of what quantum computers can offer with expertise in the target domain, whether chemistry, materials science, nuclear physics, or another field.

Only by integrating both perspectives can we pinpoint problems where quantum simulation might provide a genuine advantage.

At the same time, we are still trying to understand how quantum computers perform on specific problems. For any given system to be simulated, there are many possible methods: a variety of algorithms, numerous variants and different approaches to simulating the dynamics. When fermionic systems are involved, for example, they must be mapped onto qubits, and the choice of encoding interacts with the choice of simulation algorithm. This, in turn, affects how efficiently the algorithm can be compiled into gates.

These interdependencies mean that even for a fixed system, determining the most efficient simulation strategy is far from straightforward. It is not uncommon to see papers estimating the gate counts required for a particular simulation, followed by later work that reduces the estimate by several orders of magnitude. This illustrates both the early stage of the field and the enormous scope for progress. It also shows that we still have significant work to do before achieving implementations robust and efficient enough for widespread use.

At present, much of this progress is still being made in a pencil-and-paper fashion, which highlights the limitations of our current tools, but also underlines the potential for major advances once larger devices become available.

Could you explain how quantum signal processing can help reduce the resource overhead for simulators?

AC: Quantum signal processing is closely connected with the development of Hamiltonian simulation algorithms that achieve better scaling with error. The earliest approaches were based on product formulas, which approximate time evolution by breaking it into a sequence of evolutions according to elementary terms. The error can be reduced by simulating those terms for shorter and shorter intervals, but the scaling with error is polynomial in $1/\varepsilon$.

Later, more advanced methods were developed that achieve scaling polynomial in $\log(1/\varepsilon)$. These methods work by directly implementing the Taylor series of the exponential function, providing a mathematically very different way of simulating dynamics. Building on this, the idea of quantum signal processing was introduced as a powerful and general framework. Roughly speaking, it enables transformations of operators that represent quantum dynamics by encoding spectral information into a qubit and then manipulating that spectrum to perform the desired evolution.

Quantum signal processing has proved to be a remarkably versatile technique, not only for Hamiltonian simulation but also for a variety of other quantum algorithms. However, its practical value for simulation is still uncertain. Although it offers asymptotic improvements in error scaling compared with simpler methods, it introduces additional overhead. Whether this trade-off is beneficial in practice depends on the scale at which one aims to apply the method. This is part of what makes it difficult to determine the best approach to Hamiltonian simulation. One must weigh the advantages of product formulas, quantum signal processing, and their various variants to identify which methods yield the most efficient circuits in practice. The result is a complex and still-evolving landscape. **B**

Andrew Childs is a professor in the Department of Computer Science and the Institute for Advanced Computer Studies (UMIACS) at the University of Maryland

Thank you to our partners

HPC Breakthroughs would not have been possible without the support of its sponsors. *Scientific Computing World* is proud to partner with these inspiring organisations and to help showcase their work in this project

ACD/Labs

8 King Street East, Suite 107, Toronto, Ontario, M5C 1B5, Canada
Tel: +1 (416) 368 3435
Email: info@acdlabs.com

More than 4,600 labs worldwide choose ACD/Labs software because it is designed with scientific tasks and workflows in mind. Serving the scientific community since 1994 via offices in Asia, Europe and North America, it employs a team of more than 200, including PhD-level scientists, and specialists in informatics and chemical disciplines. It has an international client base in the chemical, pharmaceutical, food and beverage, environmental, agrochemical, flavours, fragrances, government, and educational sectors. Its product mix includes enterprise applications, processing software and databases, and expert systems for analysis, predictions and simulation.
www.acdlabs.com

CDD (Collaborative Drug Discovery)

Nine Hills Road, Cambridge, CB2 1GE, UK
Tel: +44 (0)1223 803830

Collaborative Drug Discovery (CDD) provides an intuitive software suite extensively used by creative biologists and chemists working in academic, biotechnology and pharmaceutical settings. Its flagship product, CDD Vault, enables researchers to intuitively organise and analyse both biological and chemical data, and to collaborate with partners through a straightforward web interface. CDD helps scientists register entities, track inventory, manage assay data, capture experiments, calculate Structure-Activity Relationships (SAR), and mine their data for drug candidates. CDD was founded in 2004 and presently serves thousands of researchers doing drug discovery all around the world.
Learn more at www.collaborativedrug.com

Thermo Fisher Scientific

168 Third Avenue, Waltham, Massachusetts 02451, USA
Tel: (800) 556 2323

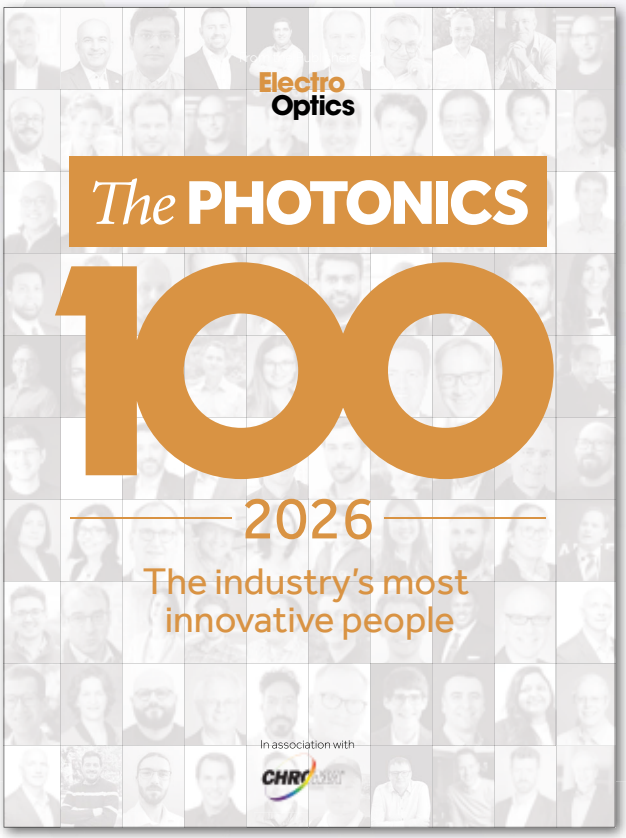
Thermo Fisher Scientific Inc. is the world leader in serving science, with annual revenues of more than $40bn. Its mission is to enable its customers to make the world healthier, cleaner and safer. Whether its customers are accelerating life sciences research, solving complex analytical challenges, increasing productivity in their laboratories, improving patient health through diagnostics or the development and manufacture of life-changing therapies, it is here to support them. Its global team delivers a combination of innovative technologies, purchasing convenience and pharmaceutical services through its industry-leading brands, including Thermo Scientific, Applied Biosystems, Invitrogen, Fisher Scientific, Unity Lab Services, Patheon and PPD.
www.thermofisher.com

Seagate Technology

Seagate Technology (Netherlands) BV, Tupolevlaan 105, 1119 PA Schiphol-Rijk, The Netherlands
Tel: +31 (6) 385 34387

Seagate Technology is a global leader in data storage and management, empowering scientific and research organisations to unlock the full potential of their data. Founded in 1979, Seagate has pushed the boundaries of data technology, delivering high-performance solutions that support the most demanding computational environments. Today, Seagate's portfolio includes industry-leading hard disk drives and enterprise-grade storage systems, as well as Lyve Cloud, a scalable and secure object storage platform designed for multi-cloud and hybrid workflows. Seagate understands the unique challenges faced by researchers: the need to store petabytes of raw data, access it quickly for analysis, and preserve it securely for long-term use. Whether supporting high-throughput sequencing in life sciences, modelling complex systems in physics, or training machine learning models in computational research, Seagate's solutions are engineered to deliver speed, reliability, and scalability.
www.seagate.com

THE 2026 LIST IS AVAILABLE NOW



Introducing The Photonics100: The Movers and Shakers Transforming Photonics

Now in its fourth year, The Photonics100 shines a spotlight on 100 trailblazers nominated by the global photonics community - those who are accelerating innovation and shaping the future of photonics technology.

Brought to you by Electro Optics

The PHOTONICS 100 2026
The industry's most innovative people

In association with CHROMA

www.electrooptics.com/thephotonics100