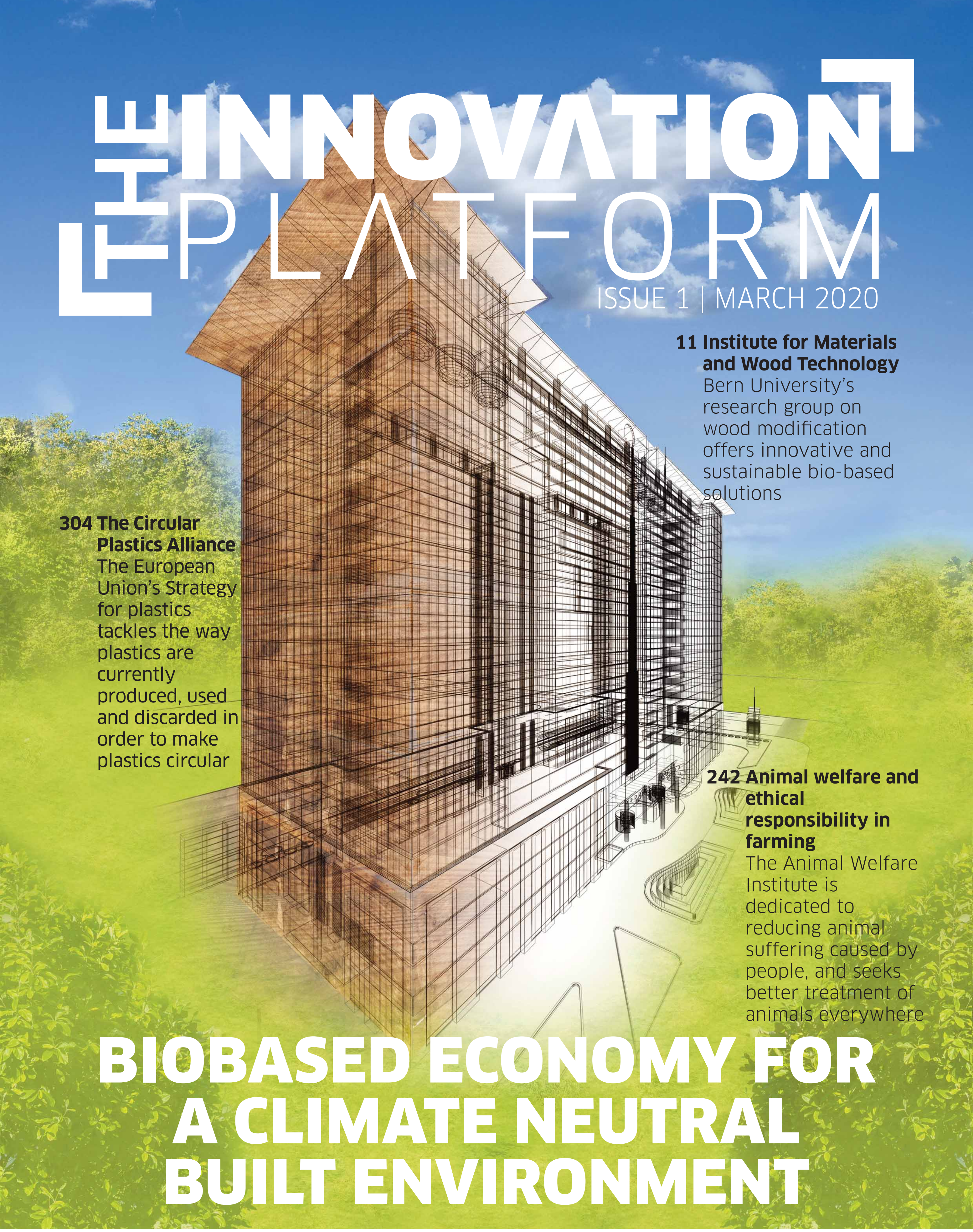


THE INNOVATION PLATFORM

ISSUE 1 | MARCH 2020

11 Institute for Materials and Wood Technology
Bern University's research group on wood modification offers innovative and sustainable bio-based solutions

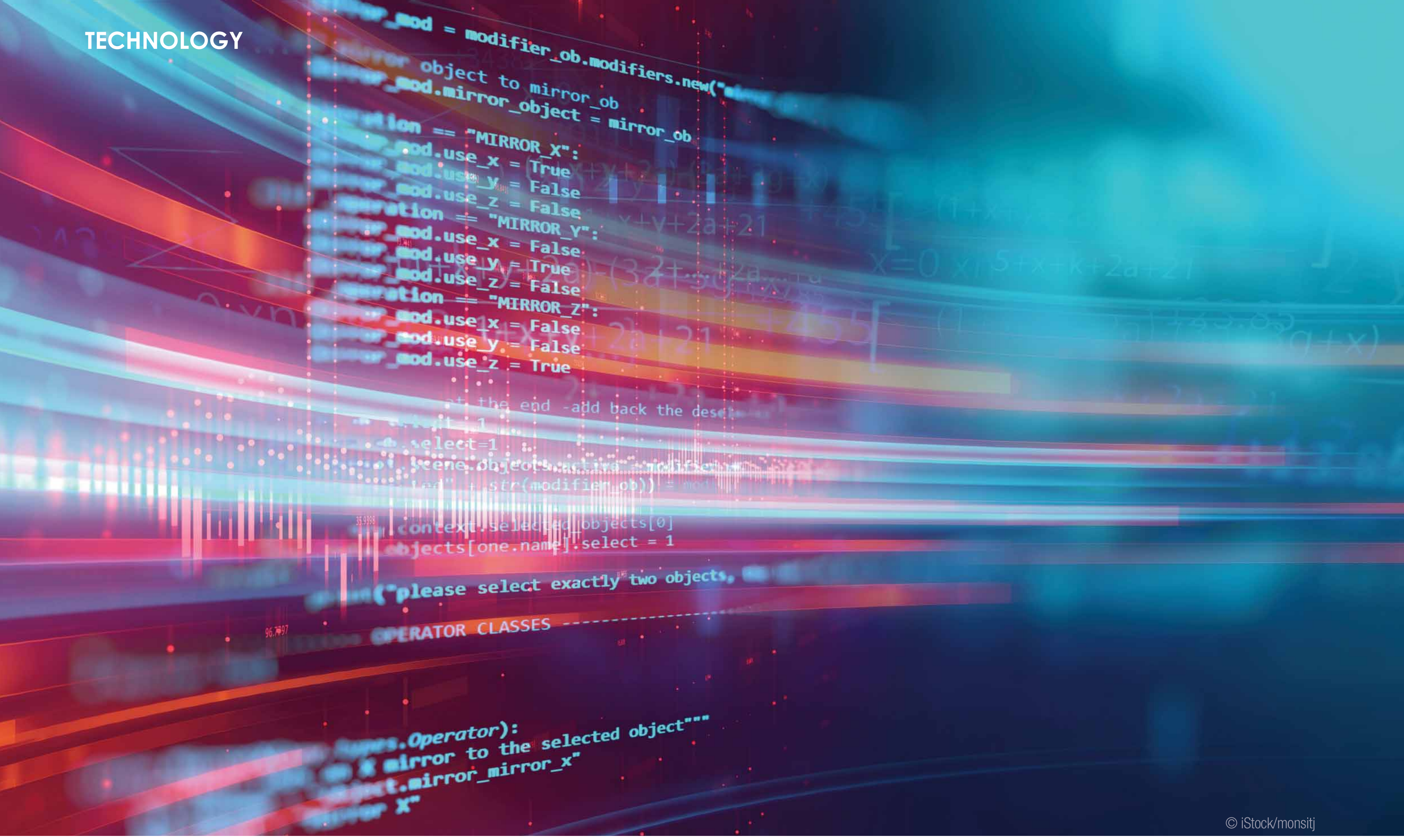
304 The Circular Plastics Alliance
The European Union's Strategy for plastics tackles the way plastics are currently produced, used and discarded in order to make plastics circular

242 Animal welfare and ethical responsibility in farming
The Animal Welfare Institute is dedicated to reducing animal suffering caused by people, and seeks better treatment of animals everywhere

BIOBASED ECONOMY FOR A CLIMATE NEUTRAL BUILT ENVIRONMENT

FRONT COVER INSPIRED BY: FH B Bern University of Applied Sciences

SPONSORED BY: KIT Karlsruhe Institute of Technology | Card Lab Innovation | cabmm center for applied biotechnology and molecular medicine | Laboratory for Muscle Plasticity



© iStock/monsitj

Large data sets are arguably the biggest driver towards High-Performance Computing

Solving the storage conundrum to accelerate life science innovation

AN inability to process large and complex datasets is hindering innovation in life sciences, but change is on the horizon as the sector increasingly makes use of High-Performance Computing (HPC) technology in a more productive way. The biggest driver towards HPC is the progressively large data sets that researchers are working with that have grown almost exponentially over the last few decades.

In 1990, the Human Genome Project started as an international research effort to determine the sequence of the human genome. Co-ordinated by the National Institutes of Health and the U.S. Department of Energy, contributors grew to include universities across the United States and international partners in the United Kingdom, France, Germany, Japan, and China.

The work of the Human Genome Project has allowed researchers to begin to understand the blueprint for building a person and has resulted in a major impact in the fields of medicine, biotechnology, and the life sciences. The project lasted 13 years, requiring millions of hours of compute and hundreds of terabytes of data flows – the resulting 2.9 billion base pairs of the haploid human genome correspond to a maximum of about 725 megabytes of data, since every base pair can be coded by two bits. Since individual genomes vary by less than 1% from each other, they can be lossless compressed to roughly four megabytes. One of the most significant pieces of scientific research of the last three decades can fit, conformably on just three circa 1990 floppy disks.

Bigger data

The current generation of projects in areas such as microbiology generate data flows that are an order of magnitude greater. Cryo-EM, a Noble prize-winning breakthrough technology that enables 3D models of the structure of biological molecules, in near-atomic detail, is leading to advances in research across a range of genetic domains. However, in the course of a day a single cryo-EM microscope generates a huge volume of data (typically one to two terabytes), and if organisations purchase multiple microscopes, the data growth multiplies.

Another example of this data overload is Next-Generation Sequencing (NGS), a high-throughput method of DNA and RNA sequencing. Although first appearing in the early 2000s, the rapid and continuous increase in data generation paired with an accompanied reduction in sequencer costs means that today, NGS devices are responsible for generating the most data produced, analysed, and stored by life sciences organisations. These sequencing devices, ranging from desktop units to clustered sequencers, can generate several hundred gigabytes to several terabytes per day. Many organisations maintain multiple sequencers, further accelerating the data generation problem.

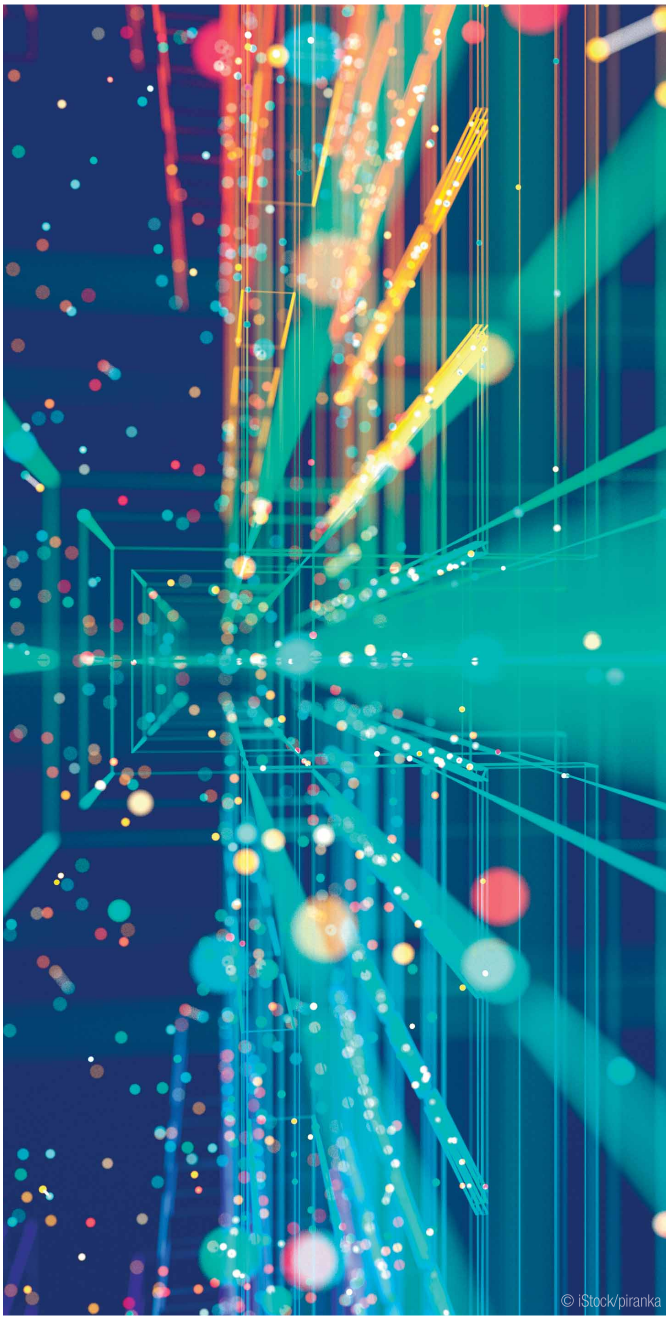
Large research projects using both cryo-EM and NGS workflow can generate hundreds of terabytes of data, which is increasingly being analysed by compute-based systems. However, according to Dale Brantly, Director of Systems Engineering at Panasas, many research organisations have started to experience a mismatch between the creation of digital research data and the compute and storage needed to analyse it. US-based Panasas supports industry and research innovation around the world with HPC storage solutions that deliver the PanFS parallel file system on the ActiveStor Ultra turnkey storage appliance.

“Probably the biggest thing that I see is as organisations move to HPC they will typically increase the compute nodes, but overall throughput does not increase by the required scale, or worse still, everything slows down due to a bottleneck at the network – and more so, within the storage.”

Brantly, who has helped deliver technology solutions for the likes of Oxford University’s Advanced Research Computing (ARC) facility and Garvan Institute of Medical Research, stresses that it’s important to maintain a balanced infrastructure even as the storage environments grow.

Moving in parallel

Although there is a proliferation of vendors offering storage, networking and compute within HPC, the research community has in the past favoured Unix and Linux-based file systems such as ZFS. While a venerable file system, the high growth in research data plus the need to accelerate processing is promoting a change towards more efficient and easily managed storage platforms.



© iStock/piranka

One of the most significant advances within this field is the use of parallel file systems, which not only provide higher performance data throughput, but also create a highly scalable data storage environment to support future computing needs. The life sciences sector has taken note of how other big data intensive research projects such as climate change modelling are dealing with the need for more performant storage.

For example, Rutherford Appleton Laboratory (RAL), one of the United Kingdom’s principal government labs and part of the Science and Technology Facilities Council (STFC) that supports research in such varied areas as astronomy, astrophysics, biology and climate modelling. RAL’s climate research has invested in nearly eight petabytes of high-performance storage in order to expand its highly data-intensive climate modelling efforts. The parallel file system is critical to obtain the scalability and affordable performance required for its rapidly expanding climate modelling workloads.

RAL deployed parallel storage system, linearly scaling capacity and performance to 150 gigabytes per second – which is one of the world's fastest implementations of a single file system throughput per terabyte of enterprise SATA storage. RAL can simply add individual blade chassis or entire racks to non-disruptively scale capacity and performance as its storage requirements grow.

As Dr Bryan Lawrence, Professor of Weather and Climate Computing at the University of Reading, and Director of Models and Data at the National Centre for Atmospheric Science (NCAS) explains, "… [the] parallel file system remains resilient even at scale, and the direct and parallel access to the storage pool means that we can work on our most complex simulations, unaffected by the system bottlenecks of our previous equipment."

When the Garvan Institute of Medical Research, one of Australia's premier medical research institutes, decided to make changes to their existing storage infrastructure to run their Illumina sequencer system at full capacity, they opted for a parallel file system. Garvan chose the Panasas PanFS parallel file system on ActiveStor to provide its researchers with a storage solution able to deliver the fast data access needed to support the rapid prototyping and evaluation of specific analyses required for genomic sequencing. After combining Panasas storage with Illumina sequencers, Garvan increased its sequencing capacity to 50 genomes per day on average – a fiftyfold improvement.

"Fundamental to our work is maintaining an extraordinary infrastructure that makes it all possible and Panasas is a key part of that," said Dr Warren Kaplan, Chief of Informatics at Garvan. "Panasas lives up to its promise of terrific performance with negligible maintenance and administration time. In addition, our sequencing data stays in the central repository throughout the analysis which makes for a more streamlined workflow, saving time and bandwidth."

Boosting performance

The increase in performance of even a few percent over legacy ZFS-based storage can have a significant impact on the adjoining compute system. However, gaining an understanding of the potential benefits is a complex task. To test this hypothesis, BioTeam, a high-performance consulting practice staffed by scientists and IT specialists, was engaged in 2018 to test the benefits offered by a modern parallel file system over legacy ZFS style equivalents. BioTeam created three real-world test scenarios including a Burrows-Wheeler Aligner (BWA) genomic indexing, BWA genomic alignment, and cryo-EM 3D classification.

© iStock/MF3d

‘Technology enablement needs to consider the human operator effect.’

The test used RELION, the industry-standard open source software product for performing cryo-EM reconstruction. The tests were performed using Panasas ActiveStor to provide a network filesystem for reads and writes as well as a reference ZFS storage array configured to provide a network filesystem for reads and writes. Conducted within a rigorous test methodology, the Panasas ActiveStor configuration with Panasas DirectFlow clients delivered a performance advantage of around 20 percent over the reference ZFS configuration during BWA indexing.

For the cryo-EM test, the researchers found approximately a 20 percent improvement in run time per iteration for ActiveStor vs the ZFS reference array for the cryo-EM 3D reconstruction without GPU acceleration. There is approximately a 10 percent improvement in run time vs the ZFS array with dual clients running the same job. The full whitepaper outlining the test methodology and results is available from Panasas.

The human factor

Technology enablement needs to consider the human operator effect. “Everyone is challenged for the battle for talent,” says Brantly. “We hear that everywhere and it’s very hard to find a top-notch storage or system administrator to run these systems – a key question that researchers and IT staff have to ask is the system simple to run and scale?”

Lastly, Brantly urges that joined-up thinking is vital, especially when you go into an HPC infrastructure with parallel storage and key infrastructure. “You need to implement sound IT principles such as backup and replication to ensure that the data is always available – this is potentially a lifetime’s work with potential to help society. Performance with protection is paramount.”

Dale Brantly
Director
Worldwide Storage Systems Engineering
PANASAS

Tweet @Panasas
www.panasas.com