Image credit: Stewart Oak. Imperial College London.

ARCASTREAM USER CASE STUDY

Protecting & Managing Data in World-Leading Research

Solution Highlights:

Simultaneously serving
2,000 HPC nodes & 3,000+ users
seamlessly via the desktop

20 GB/s
throughput with no loss of interactive use performance

1 Billion files
replicated each night in less than 8 hours

Imperial College London efficiently stores, manages & protects large volumes of world-leading research data throughout its lifecycle with a high-performance & future-proof software-defined storage and data-management platform from ArcaStream.

Imperial College London is home to 17,000 students and 8,000 staff, attracting undergraduates from more than 125 countries and awarding over 6,700 degrees every year. The University focuses on the four main disciplines of science, engineering, medicine and business and is one of the world's leading university research centres - sharing ideas, expertise and technology to find answers to today's big scientific questions and tackle global challenges.

As a centre for high-impact research, the University's Research Computing Service (RCS) - part of the ICT department - plays a vital role in addressing the computing and storage needs of the research community. In 2018, the RCS team launched the Research Data Store (RDS) to provide new robust, reliable storage services to efficiently manage and protect large volumes of research data throughout its entire life-cycle. The innovative solution at the heart of these services was designed and delivered by ArcaStream.

ARCASTREAM

in partnership with Tectrade
Your data protection & recovery partner

Challenges

Fragmented Islands of Storage

For over a decade, Imperial College's academic research community had been served by a centralized compute service under the management of the Research Computing Service (RCS), with users charged based on reserved capacity. The system had been expanded over the years in a piecemeal manner to address incremental growth and any storage attached to it only served the purposes of transient data storage. The result was a very complicated and fragmented environment, with over 30 separate independently managed islands of storage, which were difficult to access, manage and use as well as expensive to fund and maintain. Faced with poor performance and the high-costs of data-centre space, academics - who should have been solely focused on the creation and use of data in their research projects - were forced to grapple with storage capacity issues and strategies.

Ensuring Regulatory Compliance

Many users treated the centralized compute service as a de facto storage repository for all their research data and many petabytes of data had been built-up over the years with neither structure or process. It was becoming difficult for the RCS team to gauge whether data was in active use, had been abandoned years ago, or whether it was hot or cold. By moving to a new single, centrally-managed and supported system, the University would ensure that it met the demands of data providers and funders who expect researchers to demonstrate responsible data management, as well as complying with increasingly stringent regulations for the responsible handling of personally identifiable information (PII).

Delivering a Continual Service

The Research Data Store (RDS) was defined to address these challenges and provide a longer-term strategy to manage and store research data. The ultimate aim was to enable researchers to access data with ease and speed, and store that data throughout its life-cycle whilst enabling the RCS to intelligently manage growing storage demands and efficiently recover costs. To deliver the RDS, a new robust infrastructure was required, fully-integrated with the University's legacy and future compute systems, and capable of delivering a continual service over decades.

Image credit: Stewart Oak. Imperial College London. The Toxicology Unit, Charing Cross Hospital campus

The Solution

The requirement for guaranteed performance, scalability and integration with current and future compute systems substantially increased the complexity of the project and a bespoke solution was sought. It was vital that the selected vendor be willing and capable of addressing Imperial College's future workflows. Following a competitive tender process, ArcaStream - supported by our integration partner Tectrade - was selected to provide a high-performance, scalable research storage solution to seamlessly integrate legacy infrastructure and support the University's future storage strategies.

A pioneer in high performance, data aware, software-defined storage and networking solutions, ArcaStream specifically designs solutions to accelerate the world's most challenging scientific data workflows and is trusted by leading organisations such as the University of Edinburgh, the University of Bristol, the University California, Irvine and CFMS Bristol.

PixStor™, ArcaStream's high-performance scalable storage platform based on IBM Spectrum Scale™ parallel file system, combines flash, disk, tape, and cloud storage into a single global name space. With a software-defined architecture, it uses open standard commodity hardware to avoid vendor lock-in coupled with powerful data management tools - including tiering, cloud integration, monitoring, search and analytics - to drive workflow efficiencies and reduce costs.

At Imperial College, a PixStor™ framework was deployed to deliver a protected, adaptable, scalable and collaborative central platform across the institution. It guarantees consistent high-performance with no degradation as the file system fills, and the platform has been designed to easily extend, upgrade and replace for the foreseeable future, with no limits on how large it can grow, or for how long it can operate a continuous service.

Highlights:

- A fully-scalable and adaptable ecosystem to deliver a protected, scalable collaborative central data storage platform across the University
- Eliminates silos and fragmented islands of storage
- A single global name-space with desktop access for a seamless user experience
- Guaranteed consistent performance, including interactive metadata
- Granular Scalability – Scale up or Scale Out
- Intelligent tiering to any source
- Enables efficient identification of costs and control of expenditure.
- A secure environment compliant with data regulations
- Software-defined future-proof architecture without vendor tie-in

The Software-Defined Advantage

When deploying petabytes of storage, costs can ramp up very quickly and using PixStor™ software-defined storage on commodity hardware offers a significant advantage. The Imperial College RDS system had to have a service lifetime of decades, with the user experience stable over that period. No hardware is going to last that long and still be viable.

With PixStor™, the University can scale to meet future requirements with confidence in guaranteed performance without degradation as the file system expands. New technology like NVMe, the latest object storage or cloud integration can be easily added from multiple vendors down the road.

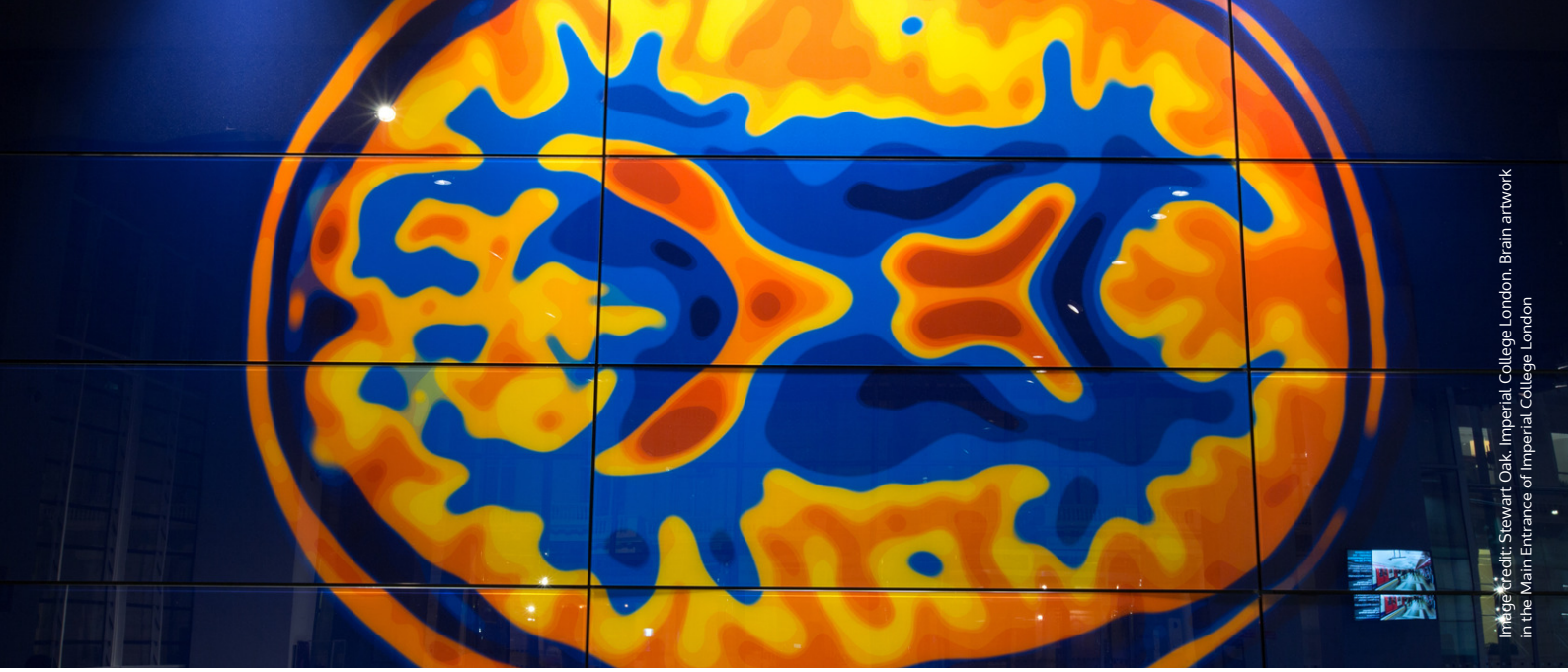


Image credit: Stewart Oak. Imperial College London. Brain artwork in the Main Entrance of Imperial College London

System Overview

Imperial College's Resource Data Store (RDS) infrastructure is data-centre based with geographically dispersed primary and secondary sites deploying PixStor™ with asynchronous replication and intelligent automated tiering to external storage targets.

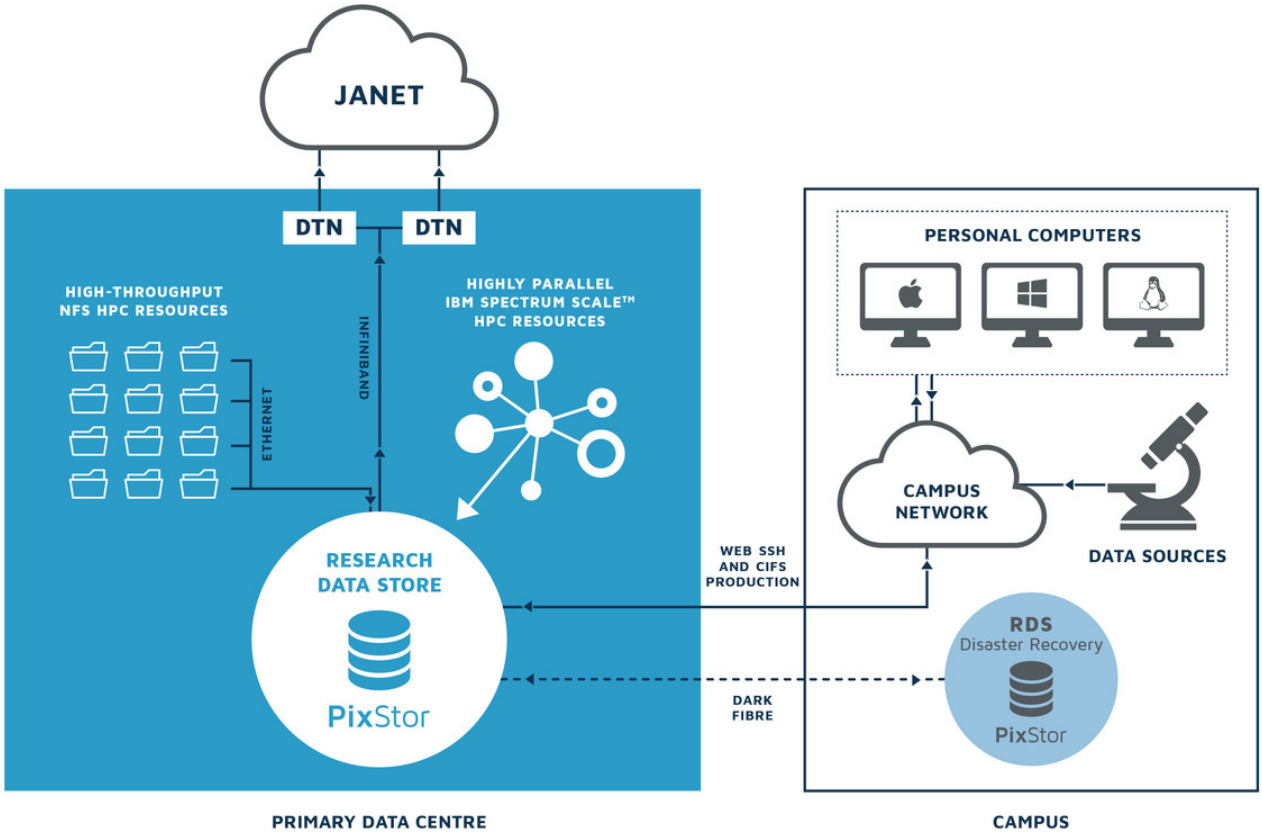
At the primary site, ArcaStream provided a 5PB research storage repository to the University using PixStor™. Capable of simultaneously serving their existing 2000 node high-performance computing estate as well as the desktop requirements of the wider research community, the combined solution delivered over 20GB/s of throughput with no loss in interactive usage performance – all consolidated into a single usable namespace.

A second site provides an offsite replica of the data for Disaster Recovery with dark fibre connection between the sites. This system uses ArcaStream's Ngenea™ to tier replicated data to a colder, external tier of storage, providing the University with a more cost-effective replication site without the need to expand the physical DR footprint as the production system grows. Using Ngenea™, the team has been able to redeploy existing Spectra Logic BlackPearl® object storage and a Spectra® T950 tape library for deep storage at the primary site, delivering a significant return on investment in their legacy hardware.

To facilitate the actual replication of data, ArcaStream's PixStor Sync™ utility was used. With over one billion files and directories in the RDS, PixStor Sync™ was able to meet Imperial College's requirement for a 24-hour recovery point objective. The complete replication process completes in less than 8 hours nightly, whilst also preserving all metadata (file ownership, access control lists, extended attributes, etc) such that the service is immediately and continually "live" at the DR site.

PixStor™ Capacity Analytics and Search were also deployed, providing the University with a far greater awareness of the contents of the repository, the age of the data and who is using the system. This has enabled RCS to make more informed decisions on expansion requirements, realise the true cost of existing data and become more aware of the behaviour and needs of the repository's users.

A smaller, isolated solution was also deployed, providing one petabyte of available capacity for secure requirements and sensitive data in compliance with data protection and GDPR regulations. The system provides the equivalent functionality but is subject to higher levels of auditing.



Solution delivered in partnership with

Tectrade

Your data protection & recovery partner

Tectrade specialises in protecting, recovering, managing, storing and securing business-essential data for organisations of all sizes around the world. With decades of experience, industry leading technologies and a proven track record of tackling data management challenges, Tectrade protects data through leading technologies ensuring this data can be recovered in case of IT outages of cyber-attacks in just minutes if necessary. Data is the DNA of any business, and at Tectrade we make it our mission to protect and recover it.

www.tectrade.com

Featured Technologies:

The overall solution leverages best-of-breed server and storage technology from Dell EMC, using a combination of Dell PowerEdge servers and Dell PowerVault storage to deliver exceptional reliability and performance at a commodity price point. R740 Servers run and serve the filesystem to the HPC Clients as well as to the general research community via the ArcaStream NAS stack.

Mellanox Spectrum™ and ConnectX® technologies were utilised to provide a hybrid network infrastructure that can deliver data via both Infiniband and Ethernet. The 100GbE networking solution backbone allows seamless scalability of the solution as the University's requirements increase.

Excelero's NVMesh® software provides a scalable NVMe tier for extreme metadata performance, running on Dell PowerEdge R740XD servers.

ArcaStream's PixStor™ combines these technologies into a single integrated platform.

Image credit: Stewart Oak, Imperial College London, Microbe Project in the Microbiology Laboratory, Department of Surgery and Cancer

Successful Outcomes

Powerful Usability

The University's RDS now supports over 3000 users, from senior academics securing funding for research projects to PhD students "working at the coal-face". Since system go-live, the speed of adoption has surpassed initial expectations and user numbers have increased substantially, with 10% of registered users now accessing the service solely for the purpose of accessing research data storage.

Research users can now access valuable data, whether a single geo copy or a secondary copy (replicated to the DR site), direct from the desktop rather connecting via the institution's HPC system. Interactive access has significantly improved by leveraging Excelero NVMesh® technology to accelerate metadata performance - ensuring that the system is always responsive even when it is really busy.

"The usability of the systems for interactive use has improved significantly," explains Matthew Harvey, RDS project-lead and RCS Manager at Imperial College. "Previously, there were frequent interruptions to interactive use because the file system load for some compute jobs effectively squeezed out interactive users. Users would log into the system, type in their search criteria but it could take more than 10 seconds to respond. Now that is a thing of the past."

Consumption v. Capacity

The move to the ArcaStream storage platform has enabled Imperial College to roll-out a new charging strategy that is more transparent and cost-effective for all, allowing researchers to cost storage, as well as compute, as services on their grants. As Harvey explains, this is a big win for researchers:

"In the past researchers would have had to cost compute or storage as physical hardware. For substantial requirements, they may have been obliged to actually buy a storage system or NAS to store their data on. Now the RDS service offers storage as a consumable, recharged based on consumption rather than reserved capacity. It's a far more simple and cost-effective approach for the user."

More effective management of storage capacity also allows Harvey's team to avoid costly additions. "The ArcaStream platform provides us with the tools and insight needed to understand the access patterns of data on the file system for each project allocation. We can now use management commands to tell when files were last accessed and then using Ngenea™, we can set a policy that says, for example, that anything that hasn't been accessed for six months is pushed to cold storage. This information governance is enabling us to store valuable data more intelligently and economically."

in partnership with Tectrade
Your data protection & recovery partner

A Collaborative Future

Since the ArcaStream PixStor™ platform was commissioned and the RDS service was launched in late 2018, Harvey's team has supervised the complex multi-petabyte migration of several research groups from self-administered storage. "The feedback of users has been nothing but positive and we are now meeting the needs of new groups of users and researchers who we have previously not supported."

With a continuous focus on improving the working environment and delivering competitive advantage, Imperial College has recently upgraded its Internet connection to 200 gigabits per second enabling high bandwidth interconnection to JANET, the UK's research and educational network. Supported by this, the University is now developing new services for greater research collaboration between other research institutions. The RDS, as a centralized storage platform providing high-bandwidth and guaranteed consistent performance, will make it very easy to move data into and out of the new system when it comes online.

" The usability of the systems for interactive use has improved significantly. Previously, there were frequent interruptions to interactive use because the file system load for some compute jobs effectively squeezed out interactive users... Now that is a thing of the past."

Matthew Harvey, RDS project-lead &
RCS Manager, Imperial College London

The ArcaStream Approach

Whilst the technical solution delivered has exceeded expectations, it is ArcaStream's consultative approach and support that has impressed the RCS project team. From the outset, ArcaStream invested time to understand the specific requirements, providing an in-depth analysis of the existing environment to identify issues and bottlenecks.

Workflow-focused, ArcaStream takes a holistic view of systems and uniquely provides a single-point of contact for multi-vendor incident management. All hardware and software escalations are performed by dedicated support engineers, who ensure SLAs and top-quality support experience is delivered on the University's behalf.

Commissioning any complex system naturally exposes issues, particularly when stressing new systems in new ways. ArcaStream has the technical resources and depth of knowledge to meet these challenges head-on and resolve them quickly and with minimum disruption.

End Results

- Greater agility to manage capacity and performance.
- Reduced complexity and silos.
- Improved security to meet stringent data regulations.
- Efficient control of expenditure and growth strategies.
- Delivers significant ROI, with integration of legacy and new systems.
- Future-proof scalability and flexibility without hardware lock-in.
- Peace of mind and an improved user experience.

Researchers at Imperial College London are now using ArcaStream's PixStor™ platform with absolute confidence in its ability to support their research data storage needs. High performance, with enterprise-level reliability and integrity, along with robust continuous service through expansion, replacement, upgrades and refreshes, mean that the University can confidently plan for the long term - addressing continual multi-petabyte per year growth of their data holdings, safe in the knowledge that both the primary and DR solution can keep up.

The PixStor™ platform has already been expanded with additional capacity, and further expansion into tape storage using Ngenea™ is underway to provide a massive capacity boost without compromising on the service provided. This ongoing investment in the service speaks volumes about Imperial College London's confidence in the PixStor™ platform and ArcaStream's ability to deliver the level of support and service they require to strategically meet their always-evolving research data requirements.

in partnership with

DELL

Excelero

Mellanox TECHNOLOGIES

Tectrade
Your data protection & recovery partner

ARCASTREAM

ArcaStream designs solutions to accelerate the world's most challenging scientific data workflows. A pioneer in high performance, data aware, software-defined storage and networking solutions - with offices in the USA, UK and Germany and strategic partnerships with best-of-breed technology providers and integrators - ArcaStream is trusted by industry leaders and institutions worldwide to store, manage and protect vital assets throughout their lifecycle and beyond.

t: +44 (0)845 052 3721
e: sales @arcastream.com
www.arcastream.com

© 2019 Arcastream. All rights reserved. Arcastream, PixStor and Ngenea are trademarks of Arcapix Holdngs. All other marks are the property of their respective owners.

CS01-IMP-ARCA-EN-0619-01