Success Story

# Durham University: Pushing the Boundaries of Cosmology with VAST Data

In the north of England, Durham University's Department of Physics is pushing the boundaries of scientific discovery, particularly in cosmology led by the Institute for Computational Cosmology (ICC). As one of the world's top 100 universities, Durham University hosts a national supercomputing service for the Distributed Research utilizing Advanced Computing (DiRAC) HPC Facility funded by the UK Science and Technology Facilities Council (STFC). This HPC system, the COSmology MAchine (COSMA), is specifically designed for the scientific workloads its users need. The DiRAC facility comprises four HPC sites across the UK, with Durham being one of them, each focusing on different types of workloads.

Dr. Alastair Basden, the HPC Technical Manager at Durham University, oversees the daily operation of COSMA, which was established in 2001. COSMA is primarily used for extensive cosmological simulations to unravel the mysteries of the universe and requires significant memory, storage, and compute resources. Dr. Basden explains, "Cosmology is the study of the universe's origin, development, structure, history, and future. We do large simulations starting shortly after the Big Bang, introducing all the known physics such as gravitational interaction and fluid dynamics, propagating that forward in time, watching how different galaxies form, collide and merge."

## The Critical Need for Efficient Data Management in HPC

Cosmologists worldwide use the insights gained from these simulations. The simulated data is made available for researchers to analyze and compare with real world data obtained from telescopes. This process helps scientists understand the fundamental gaps in our knowledge, such as the nature of dark matter and dark energy. "What makes us unique is the focus that we have on bespoke system design and designing for workloads, which allows us to be incredibly cost effective for doing a particular science. So our cost per science output is very low across DiRAC as a whole," notes Dr. Basden.

However, managing and storing the tremendous amounts of data generated by these simulations can be a significant challenge. As well as bulk data stored on a parallel file system, there is also a requirement for application space, configuration files, code, git repositories and visualisation outputs. Durham University's previous solution for these mixed file types was based on an aging NFS server, which was struggling to meet the requirements. Dr. Basden notes, "The previous way of doing things didn't scale well. It was a single NFS server with a single namespace, which didn't scale to meet our needs. Whereas with the VAST Data Platform, we can add drive nodes to give more data or add compute nodes to give more parallelism, and it scales very well."

### Why the VAST Data Platform?

The VAST Data Platform is a highly scalable and affordable all-flash data system that allows you to run AI-scale analytics at less than half of the cost of traditional all-flash solutions.

**DASE Architecture**
With its unique Disagreggated and Shared-Everything architecture, VAST breaks tradeoffs deliver significant savings and make flash affordable for all of your data.

**VAST DataStore**
The VAST DataStore is an enterprise all-flash NAS platform built to meet the needs of today's powerful AI computing architectures and beyond.

**VAST DataBase**
The VAST DataBase is a key strategic piece of the Data Platform targeting structured data. It provides transactional and analytical data warehouse functionality on top of the scalable, performant, all-flash base layer to ingest and process data at best-in-class speeds.

**VAST Data Catalog**
The VAST Data Catalog leverages the Database and SQL to provide a built-in metadata index that allows you to search and find data easily – structured or unstructured – at scale and in a fraction of the time of traditional methods.

# A Transformative Partnership and Data Platform

Faced with the need for a next-generation solution, the ICC at Durham University chose the VAST Data Platform. The choice wasn't just about finding a replacement, it was about making a transformation. The decision to adopt VAST Data was driven by several factors, including the platform's; ability to handle small file operations efficiently, cost savings from advanced data reduction techniques, excellent analytics, and the global namespace capability, which should allow seamless collaboration with other UK sites.

Dr. Basden highlights the benefits of VAST Data: "It provides much higher performance, so users waste less of their compute time reading the data onto the system. Its ability to handle large numbers of small files efficiently also allows us to encourage users to use virtual environments more, particularly within Python and the SPACK package management tool, which traditionally haven't been great - we have had to limit the total number of files per user previously to avoid significant performance reductions. With the VAST Data Platform, this is no longer the case. We can now encourage users to make the most of the system to help their workflows and set up their bespoke environments."

Since implementing VAST Data, the COSMA system at Durham University has achieved a 3.4-to-1 data reduction ratio on their user data. Thanks to the massively increased network bandwidth and the VAST Disaggregated Shared Everything Architecture (DASE), they've seen significant performance improvements compared to the previous solution.

# Enhancing Research Capabilities and Collaboration Across the UK

Looking ahead, Durham University plans to leverage the spare capacity of its VAST Data Platform to support other research clusters and explore new opportunities for researchers. The potential for shared innovation is immense. The global namespace feature will enable seamless collaboration with other UK sites, which may include trials at Cambridge and Bristol Universities. It allows users to efficiently migrate their work between different systems and access a broader range of resources and expertise.

Moreover, Durham University is actively working with VAST Data's engineering team to explore ways to reduce the data platform's energy consumption, aligning with their commitment to sustainable and efficient research practices.

As Durham University and DiRAC continue to push the boundaries of cosmological research, its partnership with VAST Data empowers them to manage and analyze ever-growing amounts of data more efficiently, accelerate scientific discoveries, and foster collaboration across the UK research community. With VAST Data's scalable, high-performance data platform, Durham University is well-positioned to tackle future challenges and remain at the forefront of scientific innovation.

*"Cosmology is the study of the universe's origin, development, structure, history, and future. We do large simulations starting shortly after the Big Bang, introducing all the known physics such as gravitational interaction and fluid dynamics, propagating that forward in time, watching how different galaxies form, collide and merge."*

Durham University

## About VAST Data

VAST Data is the data platform company built for the AI era. As the new standard for enterprise AI infrastructure, organizations trust the VAST Data Platform to serve their most data-intensive computing needs. VAST Data empowers enterprises to unlock their data's full potential by providing simple, scalable, and architected AI infrastructure to power deep learning and GPU-accelerated data centers and clouds. Launched in 2019, VAST Data is the fastest-growing data infrastructure company in history.

For more information, visit vastdata.com