



Trinity Centre for  
High Performance  
Computing

## Trinity College Dublin Expands Advanced Research Capabilities Using DataDirect Networks Storage

www.datadirectnet.com

## Challenge

To provide a common storage pool with high levels of sustained throughput for a variety of users and applications that met cost targets while providing significant future scalability.

## Application

Trinity College Dublin operates the largest supercomputing and visualization facility in Ireland, supporting advanced chemistry, physics, computational science, and biomedical research. They needed a high-performance, low latency, scalable storage solution to enable the performance of their cluster computer.

## Solution

DataDirect Networks' S2A storage system provides a high-performance consolidated storage pool that sustains read/write performance for computationally intensive projects with simultaneous large data ingest operations.

## Trinity College Dublin Expands Advanced Research Capabilities Using DataDirect Networks Storage

Trinity College Dublin, founded in 1592, is the oldest university in Ireland. Today, it boasts the Trinity Centre for High Performance Computing (TCHPC), Ireland's premier high performance computing center providing large-scale supercomputing, data management and visualization facilities. The Centre lies at the heart of Trinity's computational science initiative with a primary mission to facilitate research, across a wide range of disciplines, through the use of advanced computing techniques. Its staff has expertise in numerical modeling, molecular dynamics, software engineering, supercomputing, visualization, systems administration, and Linux.

“The DataDirect Networks storage solution exceeded the price/performance targets we set for ourselves.”

— Dr. Geoff Bradley,  
TCHPC's Projects and Development Manager  
Trinity Centre for High Performance Computing

The Centre supports several leading edge research projects, with a focus on multidisciplinary efforts that link researchers in physical science with their counterparts in computer and biological science. It operates two cluster computers, each used to perform different computational tasks relating to computer science, including interaction, simulation, and graphics as they apply to nanomaterials and biomolecular systems; computational mathematics, physics, and chemistry; and biomolecular sciences including proteomics, structural biology, and biomolecular modeling (visualization, perceptually adaptive graphics and animation). In addition, the Centre just acquired a powerful 3-D visualization system that allows researchers to transform their work into 3-D images to further extend their research capabilities.

## Research Advances Require Leading Edge Storage

While the Centre's cluster computers provided the processing power needed by its constituent

researchers, its storage system was completely full and severely oversubscribed, resulting in significant delays, program failures, and restarts. Researchers had to delete files before they could start a new computational project.

“Our system was 90-95% full, so jobs would fail because the disks had no more space. Plus, our existing storage controllers were maxing out at 140-150 megabytes per second even with a typical workload on the cluster and adding more capacity to them didn't make sense because they were limited by bandwidth. We needed to look for a more scalable solution,” explained Dr. Geoff Bradley, TCHPC's projects and development manager.

TCHPC supports over 300 users, with 40-50 active users per month. Access to the compute resources is regulated by a queuing system that allocates time and resources among the users. Because the cluster supports such a wide range of research projects, processing and storage requirements can vary widely.

Two areas of research are particularly compute-intensive and have heavy storage demands: quantum chromodynamics (QCD) and lattice QCD, which are part of particle physics theory; and electronic structure calculations, which determine the optical properties of new organic molecules.

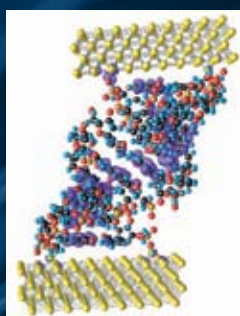
Lattice QCD simulations using the TCHPC facilities typically use 256 CPUs and are very disk intensive. They involve running Markov chain simulations that perform complex calculations, save data to disk, reload, and repeat the sequence again and again. It can take months to effectively carry out a single simulation and produce a result. There are many groups in Trinity working on electronic structure calculations. One group uses coupled cluster and complete active space methods to benchmark the optical properties of new organic molecules. These parallel simulations can read in and out terabyte size data files putting a considerable strain on the storage system.

Meanwhile, the cluster computer must simultaneously support other research such as quantum chemistry projects that typically require a large initial read-in of data, along with CPU-intensive physics projects. These types of jobs use an aver-

www.datadirectnet.com

## Physical Science Research

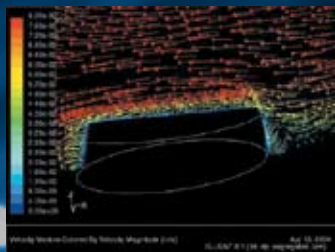
The college's physical sciences research encompasses computational mathematics, physics and chemistry. This research focuses on developing advanced algorithms and methodologies in molecular dynamics, ab-initio quantum chemistry and high performance computing to tackle problems in areas including lattice QCD, optical and electronic properties of materials, spin transport, surface reactivity and kinetics and medicinal chemistry.



Modeling Electronic Transport and the Atomic Level  
Credit: Professor Stefano Sanvito

## Biomolecular Science Research

Biomolecular science researchers in biochemistry, chemistry and pharmaceutical chemistry focus their research efforts on biomolecular modeling, rational drug design, structural biology and proteomics.



Investigation of Tablet Dissolution using  
Computational Fluid Dynamics  
Credit: Dr. Deirdre D'Arcy.

## Scientific Visualization

TCHPC manages Ireland's only 3D immersive visualization facility complete with tracking. Researchers at the Centre are developing visualization software for groups in biochemistry, physics, chemistry and Neuroscience. Biochemists use the facility to study ligand protein docking and conformational changes, Chemists and Physicists use the facility to view electronic density and Neuroscientists use the facility to map and model functional activity in the brain.

These facilities were provided by the IITAC research programme <http://iitac.tchpc.tcd.ie> and funded by the Irish HEA PRTL Cycle 3 program.

age of 32-64 CPUs and run for an average of one to four days.

"In addition to these projects, many jobs perform check-pointing, which puts a severe strain on the file-system and storage system. Every four hours, each cluster node writes its state to disk to allow restarting in the event of a termination or crash. Each CPU node has four Gigabytes of RAM, which means that more than a terabyte of data must be quickly flushed to disk. Reducing the time of the checkpoint operations is a key consideration so that jobs will process faster, improving the ROI on the cluster," said Dr. Bradley. He added, "Plus, while checkpointing is going on, another job may be doing a read-in, so the storage system has to be able to gracefully handle multiple reads and writes at the same time."

## DataDirect Networks S2A Storage Solution

TCHPC recently received funding to expand their research capabilities. First on the shopping list was a new storage system. After evaluating several options, Bradley's team selected the DataDirect Networks S2A (Silicon Storage Architecture) storage solution because it met their needs for high performance, reliability, and scalability within their budget.

The S2A system's ability to sustain high-throughput reads and writes as multiple CPUs simultaneously access data while new content is being ingested at the same time has been an important advantage in maintaining performance.

Dr. Bradley described his success with the DataDirect Networks solution. "We now have the system installed and it's doing exactly what we expected it to do with zero issues. We are seeing nearly a Gigabyte per second of sustained throughput on the S2A, which is significantly greater performance than we had before. As we expand the system, we expect this to improve further."

The S2A solution is paired with IBM's GPFS parallel file system, and uses large capacity SATA disk drives.

He continued, "Now our researchers don't have to delete data and resubmit jobs to recreate it. They can keep data and do statistical analysis

over multiple data sets – and get their jobs done faster with the better performance afforded by the DataDirect Networks storage array."

DataDirect Networks' SATAssure intelligent SATA drive management system, which is included in the S2A appliance, makes large pools of SATA drives reliable, increases their uptime, and ensures their data integrity. Because the Centre's budget relies on precious research funding, SATAssure was a key factor because it allowed the use of lower cost, larger capacity drives without sacrificing reliability, performance, or data availability. In addition, the system's fault-tolerant architecture with inherent zero-time failover further assures reliability.

DataDirect's service and support were also positive factors during the system's installation. "The installation went remarkably smoothly, with no real hiccups along the way. We were well-briefed by our DataDirect representative before we began and the documentation was very complete. It was a very positive experience and painless compared to past upgrades," according to Dr. Bradley.

As the Centre's projects receive more funding, there will be additional storage requirements. The scalability of the S2A system was a decisive factor in TCHPC's selection.

"We can just drop in more drives using our existing S2A appliance and enclosures, and then continue to expand beyond that by adding more disk enclosures," explained the TCHPC projects and development manager. "We liked the S2A because we can add more disks without a performance hit – in fact, we expect to get even more performance as we expand the system."

The Centre anticipates extending its reach beyond Trinity College Dublin as it rolls out a service to users throughout Ireland next year. They are targeting their existing DataDirect Networks infrastructure to increase their storage capabilities to accommodate this growth.

"The future is more compute resources, more users and more specialized research projects, all of which will drive the need for more storage. We see the S2A as an important enabling technology for the Centre's continued expansion," concluded Dr. Bradley.

DataDirect Networks is the leading provider of open, scalable storage systems for performance and capacity driven applications. DataDirect's S2A (Silicon Storage Architecture) appliance enables modern applications such as video streaming, content delivery, modeling and simulation, backup and archiving, cluster and supercomputing, and real-time collaborative workflows, that are driving the explosive demand for storage performance and capacity. DataDirect's S2A technology and solutions solve today's most challenging storage requirements, including providing shared, high-speed access to a common pool of data, minimizing data center footprints and storage costs for massive archives, reducing simulation computational times, and capturing and serving massive amounts of digital content.

Major corporations, supercomputing centers and rich media organizations, including AOL, Ascent Media, Boeing, CINECA, CGGVeritas, CNN, Disney, Federal Reserve Board, Ford, Hess, Kodak Gallery, Lawrence Livermore National Laboratories, NASA Ames, RIOT, Sandia National Laboratories, Sony, Technical University Dresden, Technicolor, Time Warner, Thomson, Trinity College Dublin and Universal, utilize DataDirect high performance, high capacity solutions.

**DataDirect**  
N E T W O R K S  
Performance. Capacity. Innovation.

9351 Deering Avenue . Chatsworth . California 91311  
phone +1.800.TERABYTE (837.2298) . fax +1.818.700.7601  
sales@datadirectnet.com  
www.datadirectnet.com