



Sign up for our newsletter and get the latest HPC news and analysis.

Email Address

Why Storage Matters to Scientists

May 6, 2015 by [staff](#) [Leave a Comment](#)

As the Large Hadron Collider restarts at Cern, data storage has become as important to scientists as compute power. But, as Tom Wilkie from [Scientific Computing World](#) reports, the innovative technologies being developed have much wider applications.

[Cern](#), the European Laboratory for Particle Physics just outside Geneva, was the birthplace of the World Wide Web. Now, as Cern's Large Hadron Collider (LHC) starts a new round of experiments after the discovery of the Higgs boson in 2012, the laboratory is the crucible of yet more innovation that could change the face of computing in both science and the wider commercial world.



Tom Wilkie, Scientific Computing World

The LHC generated more than 100 petabytes of data in its earlier runs, and will continue generating data at the rate of two to three petabytes each month. This huge quantity of data has to be distributed to scientists around the world so they can use it to discover new physics. The way in which Cern has structured its data distribution, and the techniques it and its associated laboratories are using to store that tidal wave of data, will have ramifications for the growth and development of big data in science and in commerce.

According to Laura Shepard, director of HPC markets, for storage company [DDN](#): "Everyone knows that data is becoming more distributed and big data is key to large-scale endeavours not just in the scientific but also in the commercial community. It is about where the data is generated and where it is processed. The Cern project is an amazing example of this, and it is only going



[Education and Training](#)[Specialist](#)

Columbus, OH - Ohio
Supercomputer Center

[Senior Engineer](#)

Columbus, OH - Ohio
Supercomputer Center

[Data Applications Engineer](#)

Columbus, OH - Ohio
Supercomputer Center

Jobs by  SimplyHired

[See All Jobs](#)

to get more and more relevant as these large-scale scientific and commercial projects come on line.”

In March, [Seagate](#), the world’s leading hard-disk drive company, announced that it had concluded a three-year partnership with Cern Open Lab to collaborate on the development of the Seagate Kinetic Open Storage platform, in the context of the LHC program. Just as the scientists are having to learn about how best to store data in order that they can easily retrieve it for analysis, so the data storage companies are looking at the LHC project and cooperating with the laboratories in the distributed data network – DDN is working with the Canadian national laboratory, Triumf, in Vancouver, for example – to help them develop the next generation of storage technologies that will help not just science but more widely in the commercial world of big data as well.

It wasn’t always like this. “Everyone talks about compute, but storage was always the child that’s left behind. Everybody needs it, but not a lot of thought is given to it. But once users start running, they realize that they do need storage and need it fast. So we have to plan it out and think about it,” said Michelle Butler, technical program manager for storage enabling technologies, at the US National Center for Supercomputing Applications (NCSA) in Illinois.

Take the burden from the users

Scientists use computers as a means to an end – they want to focus on the science, she continued, “but you have to know what is under the covers to take advantage of what is there. You have to know how it’s laid out; what the storage is underneath it.” At the NCSA, she explained, “we’re working with users that have very large files. Right now the burden is on the users to move that data, and we have to take that burden off them. We also have to take the burden off them to find their data.”

Butler pointed out that distributed data is not confined to the LHC project. At the NCSA, the most data intensive sciences, storing up to 4PB, are weather and atmospheric sciences – including solar flares that would be included in weather so the solar and space environment. Nor is distributed data unknown: she cited the example of an earthquake engineering team that uses the Cray Blue Waters’ supercomputer at the NCSA but which stores its data in three different machines in geographically separate places. “How can they find the files that they have stored?” she asked. It is a complex problem, she went on, involving data management and the management of file locations but the location problem was not just within one computer it also involved location geographically. “I think there is a lot of different hardware and software that needs to be built here.” And, she stressed, it was imperative to work with the vendors: “They can’t just do their own thing with their hardware anymore; it is much more global than that. Everyone has to have more standards and more

flexibility and realize they do not own all the data at every level anymore.”

Butler’s theme of taking the burden away from the users is very much in evidence also at Cern. Years of effort have gone into building a system where the “users do not know about the details of our infrastructure. They have the illusion of an infinite file system,” according to Massimo Lamanna, who is section leader, file systems and disk operations section, at Cern.

He explained that, since the 1990s, all the experiments have been directly connected to the computer centre. “They use the computer centre first of all as an archival system. In some senses, this may not seem very exciting, but it’s clearly super-important.”

With the beams running inside the LHC, the experiments register thousands of collisions between the fundamental particles, but local filters at the detectors down in the tunnels select those signals that might be interesting for the physics the experiment is trying to study. Once the event has met these criteria, the data is put on fibre cables directly connected to the data centre. “We receive these on disks and from the disks we stream to tape, because essentially it’s the quickest way to create a second copy. Which would typically stay for ever – it’s the archival component of what we are doing,” Lamanna continued.

Archiving is the first of three activities that the data centre supports, he explained. The second is to make the disk files immediately available for processing – starting the reconstructions of the sub-nuclear events in preparation for the final analysis across all the data. The important point to this initial analysis, he said, is to get an idea of the quality of the data, the calibration of the instruments, and the ability to gauge whether the data is okay or whether there are some subtle variations that the experimentalists need to correct immediately.

The third activity is export. From Cern’s disks, the data is distributed to the LHC computing grid. A network of “Tier 1” centers around the world take the data – depending on which experiment they are collaborating on – and the data is copied to tape there as well. This process ensures that, distributed across the 10 to 12 independent Tier 1 centers, there is a complete copy of the LHC data – a second copy in addition to the one held at Cern itself. Lamanna said: “At, for example, CMS [one of the LHC experiments] they are taking 1PB in a week. If you go to RAL [the UK’s Tier 1 centre at the Rutherford Appleton Laboratory] you get 10 per cent; go to Fermilab [the US particle physics laboratory] you get 20 per cent. So across Tier 1 you get a second copy – for business continuity/disaster recovery reasons. My team is in the middle. We are getting the data right from the pit, putting it on tape, giving the data for reconstruction – this is done on our batch machines – and then we run the services that can be interrogated by the experimenters to put us in connection

with the Tier 1.”

Innovative storage technologies

How the Tier 1 centers organize their storage is largely up to them, although there is a memorandum of understanding (MOU) in place “that has stringent requirements about up-times and what kind of services have to be provided,” as Reda Tafirout, group leader, Atlas Tier 1 at Triumf, the Canadian particle physics laboratory in Vancouver, pointed out. Atlas is one of the other LHC experiments. Not every country can provide a Tier 1 centre for Atlas, he continued, because it’s an expensive project “but we had to do it to help the science from the experiment.”

Every year Atlas reviews its computing model, based on a projection of how the LHC will perform, and how much data acquisition time will be happening. What happened in 2012 with the Higgs discovery, we knew there was a new particle so the LHC run for that year was extended by a couple of months – which has consequences for the computing resources. At that time, Canada was one of the very few Tier 1 centers that had spare capacity that was pledged to Atlas.”

In Tafirout’s view, the Tier 1 centers are such a critical component of the project that they have to have reliable hardware and infrastructure. But while Atlas makes projections of its requirements for up to three years in advance: “It is dynamic; there is always room for revisions or variations. So it’s a continuous process.” Triumf signed the MOU with CERN in 2006. “We needed really high-density systems at Triumf; we could not go with traditional servers that were being used at that time at Cern because they had such a large data centre that density was not so critical for them at that time. In our case, the real estate is really at a premium at Triumf, and in Vancouver in general, so it was critical to go with the highest density possible.” DDN offers some of the highest density storage, which was one of the reasons that Triumf opted for the company as its supplier.

All the Tier 1 centers have to be up and running when the experiment is collecting data because the buffer space at Cern is not infinite, it is only a few days. So it is important that that data is farmed out to the Tier 1 centers as soon as possible. Triumf has an end to end connection with Cern through the LHC’s own optical private network, a multi-Gigabit per second link distinct from the normal research network used by the universities. The providers are Géant in Europe and the Canadian Network for the Advancement of Research, Industry and Education (Canarie) in Canada.

Not compute but storage capacity is critical

Raw data goes to tape directly. But the role of a Tier 1 centre is not just

archival. It has to ship the right data to the Tier 2 centers around the world. It is in these centers that the analysis and physics are done. Triumph holds the data for analysis on disk. This data, Tafirout pointed out, “is critical; it’s urgent. This is where DDN comes in. We need really high-performance disk storage to handle the derived data that users need access to, around the clock, from around the world. This performance is critical.”

Triumpf itself is not a Top500 site; indeed, none of the Tier 1 centers are in the Top500. “It is not the compute capacity, but the storage capacity with the highest performance that is critical,” Tafirout said. And, he continued, as a Tier 1 centre, it’s constant evolution. “We have to replace old technology, and keep up with what is available.” The technology dating from 2009, however, was a bulky storage system. Triumph bought it because it had a good price performance and the organization had enough space in the server room. “Now the data centre is larger, but still sitting in the same physical space. We approached DDN to integrate storage and servers in one solution,” he said.

According to DDN’s Laura Shepard, the company’s SFA products “are the most dense product in the industry so you have higher capacity for your footprint.” At Triumph, she said, they had been developing “in-storage processing – the ability to expose volumes on the storage controller itself and embed data-intensive applications on the storage controller.” This means that data-intensive applications can be co-resident with their data so there is less latency associated with the round trip: “This is very significant when you are having primarily a data transaction, as we are seeing here at Triumph. From a convergence standpoint, you are eliminating components such as switches, ports, and licensing – which reduces the acquisition cost as well as the management cost.” By eliminating external components, the system also achieves a smaller footprint, she said.

Innovating storage at Cern

At Cern too, change is the order of the day. Cern built most of the software infrastructure itself to support its data storage operations. “For the physics data, we use two systems: Castor and the other is much newer, called EOS,” Lamanna said. “So you inject files and they go to tape in Castor. The users see an infinite file system, in the sense that we have some disks, but it is just a cache and the users do not know about the details of our infrastructure. We catalogue the file and then it flows to tape and eventually is removed from the disk.” If it is still on the disk and the user wants it back, then there is essentially no latency. If the file is on tape – and in the long run, as Lamanna explained, the majority is on tape – then users have to wait a little bit for the tape to be mounted, the file to be copied to the disk, and they are again put in contact with the disk copy in a way that is transparent to them.

Castor was started in 1999/2000. It has given faithful service for a long time

and currently contains an archive of around 100PB of data. But three or four years ago, Cern started a new system, EOS, which is intended to be a disk-only system. The data will not overflow to tape but instead the system will handle an unprecedented quantity of disk space of around 70PB. This disk-only system has some unique features.

The main issue with any large disk-drive system is that the disks fail. Cern has around 2,000 machines each with a variable number of disks – typically between 24 and 48 – so at least 40,000 drives and therefore, statistically, it can expect a few disks to fail every night. Again, the answer is duplicate, geographically separate copies. “Every file is put onto two disks, so a single failure will not impact the user because a second copy of their data is somewhere else. And the system detects the imbalance. What is really amazing to me is that we have two computer centres: one is here in Meyrin; the other one with a more or less equivalent capacity is located in Hungary, at the Wigner Centre near Budapest,” Lamanna explained. If one of the two fails, the users are directed to the other until the missing files are recreated automatically by the system. He said: “To my knowledge this is absolutely unique. The distance between the two centers is around 1,000km; we have 22ms latency; but this works as a single system.”

Simplifying storage

Jim Hughes, chief technology officer at Seagate, believes that there are ways of simplifying the system even further and Seagate’s partnership with the Cern Open Lab project is intended to investigate how the company’s new Kinetic Open Storage system can be adapted to help with Cern’s data storage needs.

About two and a half years ago, Seagate embarked on a path of what can be done to make disk drives simpler and faster and easier and cheaper,” he said. “One of the things we realized was that the interface that people use to talk to disk drives is the same interface that’s been used for the past 40 to 50 years – and it is trying to mimic a system that doesn’t exist anymore. Nowadays storage systems even inside a disk drive are, in essence, virtualized, so the question is: can we make things easier for us – for Seagate to build and therefore cheaper – and easier for customers to use by slicing away layer after layer of software so as to make everybody’s solution simpler?”

To build a scale-out storage system like Cern’s EOS, he continued, “you always have an architecture which is a server on the top, some distribution system, then a second server that holds the disk drives. Why can’t the machine that is making the EOS request – instead of making the request to another machine which then makes the request to the disk drives – why can’t the machine that is making the request talk directly to the disk drive? What we are attacking with Kinetic is elimination of the servers from the disk drives. Not cost reduction but elimination.”

By giving the disk drives an API, the applications can talk directly to the storage over the Ethernet. “Now we can have a very inexpensive switch in front of the disk drives, so that the disk drives are just on the network – peers to the compute nodes that are doing the work. When you do a project in Open Lab you have to have some pretty lofty goals, and that is the goal: to eliminate the servers in front of the disk drives,” he said. The value of Kinetic is cost savings, he continued, but for him personally, the goal was to create a better way of doing things.

Even Tier 2 centers – the ones that are tasked with analyzing rather than storing and distributing the data – can have storage problems. Michelle Butler, in charge of storage technologies at the US National Center for Supercomputing Applications, explained out that the NCSA is a Tier 2 centre. However, the LHC work does not use the main Cray Blue Waters supercomputer.

The University of Illinois has its own cluster infrastructure at the NCSA, on which individual departments can buy their own compute nodes, again using DDN disks. “That is the cluster that Atlas uses. One of their applications actually put so much stress on the data disk structure for the general campus cluster that we built our own storage infrastructure for them so they did not impact the campus cluster.” She said: “They had their own servers. They came in and said “we’re really going to knock that file system” and we didn’t believe them.”

There is a very large gap that I see between the HPC environment and the cluster environment. In future, they are going to marry a lot more,” she continued. “How do we build storage environments using commodity parts that make them just as reliable and fast as the large supercomputing environments? We’ve spent a lot of time learning more about smaller environments, and how to scale those.” Her team had been studying how to make disk environments smaller, cheaper, and use less power. She too sees a need for different software to make them more reliable, while at the same time cutting down on hardware.

At every aspect of the LHC data storage problem, the diagnosis is similar. The solutions that they advocate may differ, but for Laura Shepard from DDN, Seagate’s Jim Hughes, Cern’s Massimo Lamanna, Triumpf’s Reda Tafirout, and NCSA’s Michelle Butler, the outlook for the future is similar: simplify the system; reduce the hardware; make the architecture cleverer; and reduce the costs – and all the while, increase the capacity.



This story appears here as part of a [cross-publishing agreement](#) with [Scientific Computing World](#).

SHARE:



Related Content:

- [Europe Gets Why Big Data Needs Networks](#)
- [Why Storage Matters to HPC](#)
- [DDN Dominates TOP500 Storage](#)
- [High Throughput Data Acquisition at the CMS experiment at CERN](#)
- [Video: Genomic Storage - Wide Bands, Long Tails](#)

 Filed Under: [Government](#), [HPC Hardware](#), [HPC Software](#), [Industry Perspectives](#), [Industry Segments](#), [Research / Education](#), [Resources](#), [Storage](#)  Tagged With: [CERN](#), [DDN](#), [Seagate](#)

Leave a Comment

Name *

Email *

Website

Post Comment

- Notify me of follow-up comments by email.
- Notify me of new posts by email.

Resource Links:

Be ready for the challenges involved in ultra-fast computing of massive datasets. Discover Bull's exascale program

Join ISC High Performance in Frankfurt, Germany, July 12 -16

Simplify & speed HPC/cluster application development Try Intel Parallel Studio XE



[About insideHPC](#)

[Contact](#)

[Advertise with insideHPC](#)

Copyright © 2015