



STORAGE

When protons collide

Smashing protons together is very hard to do and, when it is done, 15 petabytes of data will be generated annually and stored on tape

CERN, the European Organization for Nuclear Research, is the world's largest high-energy physics research establishment and approximately half of the world's particle physicists use its facilities. It has embarked on a multi-year effort to find and observe some of the most elusive particles in sub-atomic physics. To find them, CERN is building the largest and highest-energy particle accelerator in the world at its Geneva headquarters. This is the LHC, the Large Hadron Collider (protons belong to a class of subatomic particles called hadrons).

The aim of the LHC is to smash protons together in head-on collisions at high enough energies to create new particles not stable in our Universe today. One such particle is the Higgs boson, which has been described as the Holy Grail of particle physics.

Observation of the Higgs boson could confirm a fundamental part of modern physics theory, and provide a deep explanation to the nature of mass, and why particles such as electrons and protons, the building blocks of the world around us, have such different masses.

The LHC is scheduled to start operating in 2008. It is a world-wide collaboration of over 8000 thousand physicists of over 55 countries. They will all have access to LHC experimental data via a global multi-tiered computing grid. The data generated during the multi-year LHC project is expected to dwarf every other scientific experiment in history, amounting to an incredible 15 petabytes a year.

A petabyte is one million gigabytes, equivalent to some 200,000 DVDs. To put that in perspective CERN, established in 1954 and therefore active for 53 years, currently holds 8 petabytes of historic physics data. This is going to almost double in the first year of the LHC's operation, and carry on swelling year by year because of the flood of LHC data.

The data flood will involve I/O rates of over 1GB/sec. That's equivalent to storing a movie on DVD every four seconds.

None of this Niagara of data can be lost, not one byte of it. The primary storage medium will be tape, in T10000 and 3592 formats, stored in IBM 3584 and Sun StorageTek StreamLine tape libraries. Nothing else is remotely feasible as only tape has the capacity and cost that allows the data to be kept semi-online inside a

library. It will require tens of thousands of tape cartridges and CERN must have confidence that the data, once written, can be read.

Duty of care

Tape error rates, though very low, will be a major consideration for CERN. Doctor Charles Curran is the Storage Consultant and physicist at CERN responsible for its tape storage operations. He said: "We are probably one of the biggest data holders in the world with our 8 petabytes of physics data. Because of the LHC project the number of tape cartridges we need will increase markedly. We face a whole series of problems - it could run away from us. We can't throw the data away and we are planning to keep it entirely in automated tape libraries.

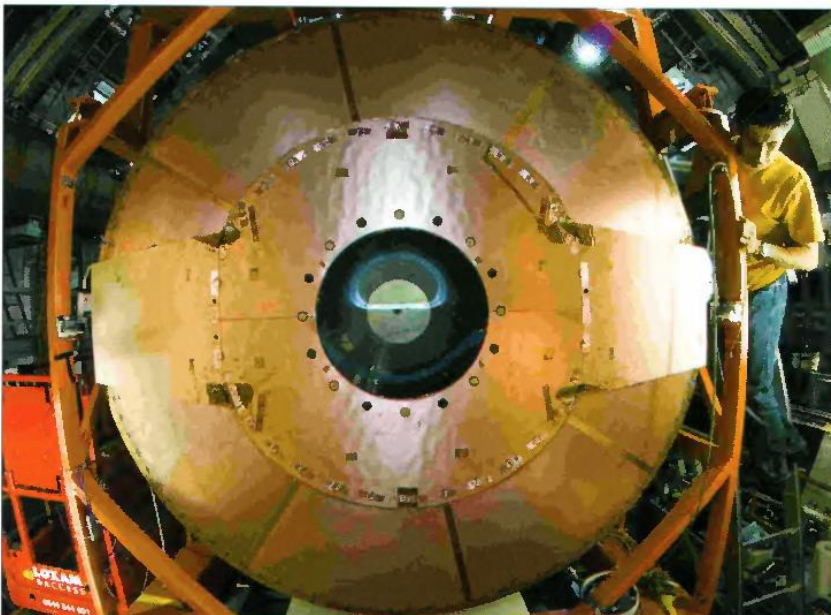
"LHC experiments will generate data instantly and we are the bit bucket at the end of this. There will be deep unhappiness if we lose even one percent of this data. We have a duty of care and can't afford to lose data. About one percent of media proves problematical with errors and file mis-reads."

Dr Curran explained: "We try not to abandon data. If it is critical we can ask our suppliers to have a go at recovering it. It is very labour-intensive for them but, generally they do succeed. There could be 5,000 data files on a tape, possibly up to 100,000. If some are irretrievable you can spend an enormous amount of time trying to recover them.

With an annual need for between 15,000 and 19,000 cartridges, even this 1 percent cartridge error rate means 150 - 190 reject cartridges a year. The problem is more acute than this as each time there is a tape generation change CERN will migrate data to the new generation, taking advantage of its capacity and I/O speed increases. Dr Curran is also expecting a StorageTek Powderhorn library to reach its end-of-life in 2008, meaning a migration of thousands of tapes to a StreamLine replacement.

Such a migration involves a lot of manual tape handling, not a good idea as that too prejudices tape quality.

Dr Curran does not want to find that existing data is unreadable when he tries to migrate it. That is worse than finding a new tape has too many errors to be usable,





much worse, as precious data will be lost.

CERN uses its own software to write thousands, perhaps many tens of thousands of files of CERN's experimental data to any individual tape. This use of CERN's own software created its own difficulty when dealing with tape media problems, as Dr Curran explains: "We use Linux everywhere and there is an area of grey fog between us and the tape automation and media manufacturers." In other words, it is harder than it would otherwise be to persuade suppliers that there is a problem with media or drives.

Standard backup products such as IBM's TSM (Tivoli Storage Manager) are used to write only a few percent of the data in CERN's IBM and StreamLine 8500 libraries. Because of this in-house tape software there has been no way for CERN to check the quality and reliability of the 30,000 or so tape cartridges currently in the libraries. Unlike the situation with disk drives there is no such tape equivalent for disk bad block checking. Until now that is.

Tapewise

In January of 2007 Dr Curran was introduced to Tapewise, software produced by Data Product Services from Farnborough, England. This software writes data to or reads data from tape, tracking any errors, soft recoverable ones or unrecoverable ones, that occur. It streams a whole tape through a drive at its maximum rate in this way and, with its Tape Error Map (TEM) technology, produces a 3D graph showing errors encountered along the length of a tape when data was being read or written.

The user can decide what an acceptable error rate is and that boundary will be shown on the graph with any error rates above the user-defined norm instantly visible. Any media problems can be immediately identified. The software supports a large number of tape formats: 3480; 3490; DLT; SDLT; 3570; 3590; 9840; 9940; T10000; LTOs 1, 2 and 3 and 3592.

Of course, tape errors can be caused by a faulty drive as well as by faulty tapes. Known good tape "TEMS" can be used to compare one tape drive with another and, once again, errors on the tape are instantly visible and will indicate drive problems such as shoe-shining; constant stop-start cycles because of inadequately writing data or reading data by the drive.

Now Dr Curran would have a way of checking tape quality with an independent software product, one that can be, and is, used by tape library and media manufacturers. It means that tape media problems and drive problems can be identified quickly and their scope found out before data is put at risk. The cost-effectiveness of Tapewise

was instantly visible to CERN and it is now a Tapewise customer.

Gaining credibility

Dr Curran described another benefit of Tapewise: "By using Tapewise it helps convince our management internally that there are problems. We can then talk to our supplier of automation and media. Without it we would just be relying upon our own judgment. We need proper evidence to talk effectively to our managers and to our suppliers. It can stop us just giving up on the media. Tapewise will help us identify that a problem exists and define its scope very much more quickly.

How does Tapewise help when talking to tape library and tape media suppliers about tape problems?

"They listen. We've complained for many years to suppliers about tape problems. Generally they have improved the next generation of a tape format to solve problems with an existing format.

"For example, there was a problem 15 years ago with a batch of 3480 format tapes and a particular type of glue holding a coating in place. The glue was susceptible to melting if the tape was left parked on the read/write head. We replaced 1,000 cartridges because of this. It was a huge problem with a solution that involved migrating the data on affected tapes to replacement media. Every so often you had to wet-clean the drive head to ensure reliable head performance.

"Other media had a leader block problem that involved particular batches. It was tedious to fix as the leader blocks had to be replaced. It took some time to convince the supplier that there was a problem.

"Thankfully these kind of serious problems seem to have disappeared." Nonetheless the one percent tape media error rate seems pretty constant.

Tapewise use will enhance CERN's ability to engage in positive discussions with suppliers: "Because of our 'home-grown' software we can start to lose the ability to convince our suppliers that we see a problem. Tapewise can produce independent evidence and help us gain credibility." When asked if there are any equivalent products to Tapewise, Dr Curran replied: "I haven't detected one. There are companies that will attempt to recover data from damaged media. But they require a fee up front and it is expensive. They have a good reputation though. Not having to use one of these agencies will justify the cost of Tapewise software right away.

Tapewise and Clareti EDT

Dr Curran is going to use Tapewise Online, DPS' newest edition of Tapewise

that leverages Gresham Enterprise Storage's Clareti EDT technology to provide ACSLS library control. Tapewise applies quality checks to tape cartridges and drives. Clareti EDT is tape library operation management software. Dr Curran explained how he will use the two products: "By embedding Clareti EDT technology in the Tapewise product I don't have to actually go to the tape library and start a Tapewise checking process manually. I will be able to set up Tapewise and have tape checking carried out automatically." The Clareti EDT technology within Tapewise will automate the mounting of tapes in Sun tape libraries. "Hopefully we will be able to automate the sampling of the thousands of tapes that we have in the libraries. We will be able to check our media much more manageably."

For a library with a Sun T10000 drive, an IBM 3592 J or B drive and a legacy StorageTek 9940B drive, Dr Curran can look at media on any one or all three drives at once with Tapewise thanks to the inclusion of Clareti EDT in the product.

He described the type of tape quality checking task that will now be very much more manageable: "We have had 3 petabytes of media delivered recently. It weighed two and one half tonnes. All of it will be inserted into the libraries. As it gets loaded we will run a full write test on one percent of it. We will do the same with other media when it arrives. We have to get this ready for LHC start-up in May 2008.

Could Tapewise be useful for other tape library users?

Dr Curran's opinion is straightforward: "I think it is important that they have to have a plan and the tools to ensure their data is stored in a reasonable condition. Every five years they will have to migrate it and that is not the time to find out that you have a problem."

The elusive Higgs boson

Tapewise, enhanced by Clareti EDT, will provide CERN with a tool that has been needed for many years; a way to simply and reliably check tape media and drive quality. The tape media management task is much improved and CERN's ability to talk credibly with suppliers about quality problems will be much enhanced. Its ability to store the expected 15 plus petabytes of LHC data reliably and dependably for its global user base of thousands of physicists will also be enhanced.

Hopefully, somewhere in the avalanche of data pouring to Dr Curran's tape libraries will be the evidence the physicists are looking for, that the elusive Higgs boson exists. With the help of Tapewise that data won't itself be elusive.