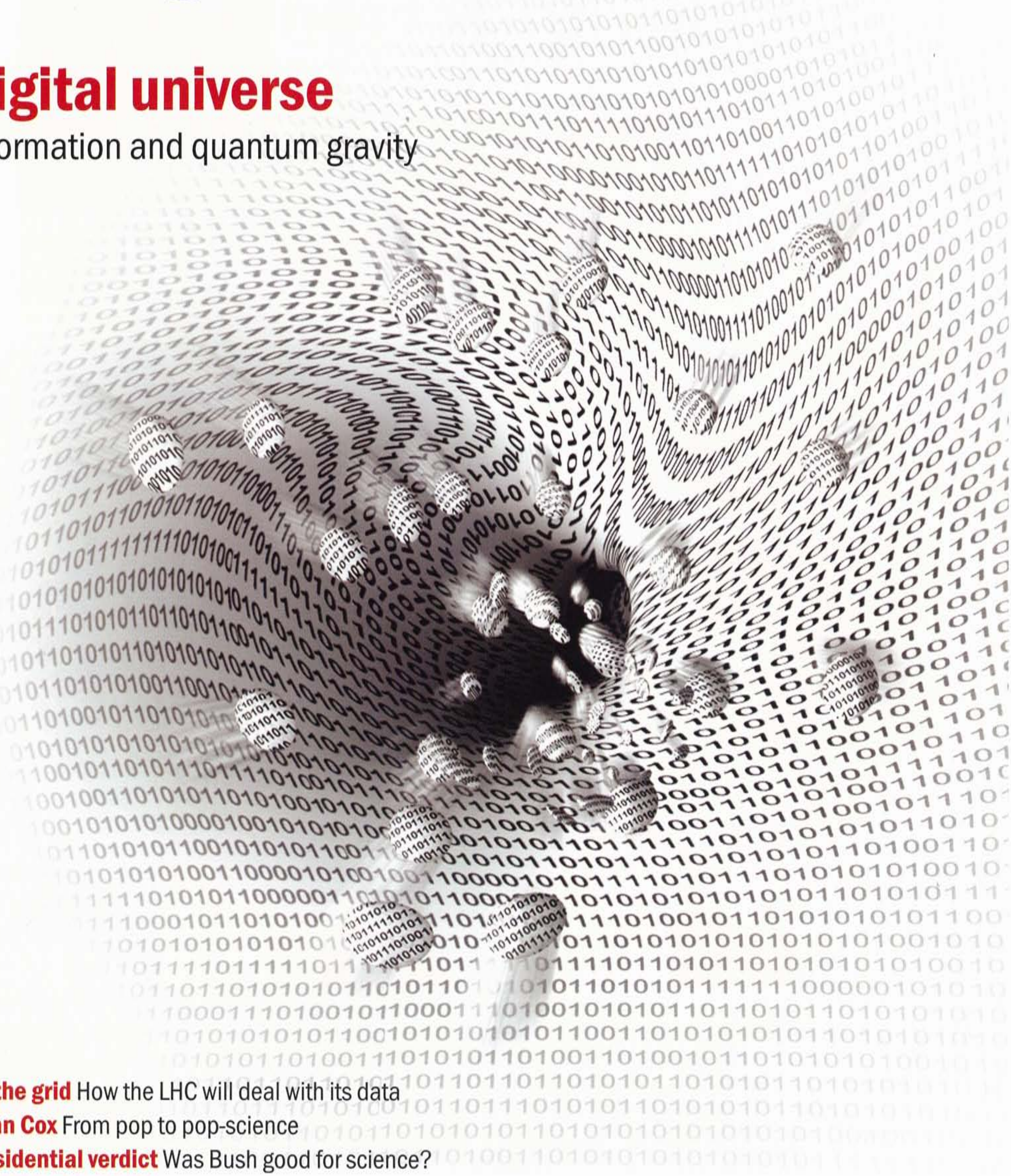
physicsworld
physicsworld.com
Volume 21 No 11 November 2008

Digital universe

Information and quantum gravity



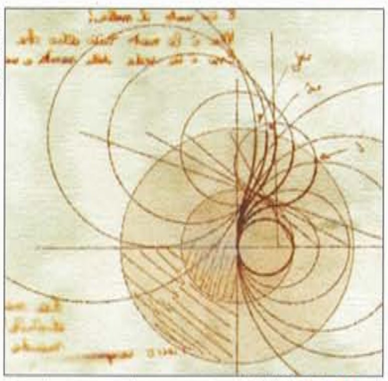
On the grid How the LHC will deal with its data

Brian Cox From pop to pop-science

Presidential verdict Was Bush good for science?

physicsworld

Data deluge – the other LHC challenge 24–27



Renaissance redraws – art and CERN 28–29

On the cover
Digital universe (*Jean-Francois Podevin/ Science Photo Library*) 30–36
How the LHC will deal with its data 24–27
Brian Cox 12–13
Was Bush good for science? 16–17

Physics World is published monthly as 12 issues per annual volume by IOP Publishing Ltd, Dirac House, Temple Back, Bristol BS1 6BE, UK

United States Postal Identification Statement
Physics World (ISSN 0953-8585) is published monthly by IOP Publishing Ltd, Dirac House, Temple Back, Bristol BS1 6BE, UK. Annual subscription price is US $520. Air freight and mailing in the USA by Publications Expediting, Inc., 200 Meacham Ave, Elmont NY 11003. Periodicals postage at Jamaica NY 11431. US Postmaster: send address changes to *Physics World*, American Institute of Physics, Suite 1NO1, 2 Huntington Quadrangle, Melville, NY 11747-4502

Features

The large hadron computer

Plans for dealing with the torrent of data from the Large Hadron Collider's detectors have made the CERN particle-physics lab, yet again, a pioneer in computing as well as physics. **Andreas Hirstius** describes the challenges of processing and storing data in the age of petabyte science

Andreas Hirstius is a physicist who switched into computing. He is now the technical manager of CERN Openlab and the CERN School of Computing, e-mail andreas.hirstius@cern.ch

In the mid-1990s, when CERN physicists made their first cautious estimates of the amount of data that experiments at the Large Hadron Collider (LHC) would produce, the microcomputer component manufacturer Intel had just released the Pentium Pro processor. Windows was the dominant operating system, although Linux was gaining momentum. CERN had recently made the World Wide Web public, but the system was still a long way from the all-encompassing network it is today. And a single gigabyte (10^9 bytes) of disk space cost several hundred dollars.

This computing environment posed some severe challenges for the computer scientists working on the LHC. Firstly, physicists' initial estimates called for the LHC to produce a few *million* gigabytes – a few petabytes (10^{15} bytes) – of data every year. In addition to the sheer cost of storing these data, the computing power needed to process them would have required close to a million 1990s-era PCs. True, computing capabilities were expected to improve by a factor of 100 by the time the LHC finally came online, thanks to Moore's law, which states that computing power will roughly double every two years. However, it was difficult to predict how much computing power the LHC experiments would need in the future and CERN computer scientists had to be aware that the computing requirements could grow faster than Moore's law. Farming out the number-crunching to other sites was clearly part of the solution, but data transmission rates were still comparatively slow – in 1994 CERN's total external connectivity was equivalent to just one of today's broadband connections, a mere 10 megabits per second.

The chief sources of the LHC data flood are the two larger detectors, ATLAS and CMS, which each have more than 100 million read-out channels. With 40 million beam crossings per second, constantly reading out the entire detector would generate more than a petabyte of data every second. Luckily, most collisions are uninteresting, and by filtering and discarding them electronically we reduce the flow of data without losing the interesting events. Nevertheless, ATLAS, CMS and the two other LHC experiments, ALICE and LHCb, will together produce 10–15 petabytes of data every year that have to be processed, permanently stored and also kept accessible at all times to researchers around the world. Dealing with such huge amounts of data was dubbed the "LHC challenge" by the IT departments at CERN and the other institutes that worked to solve the problem.

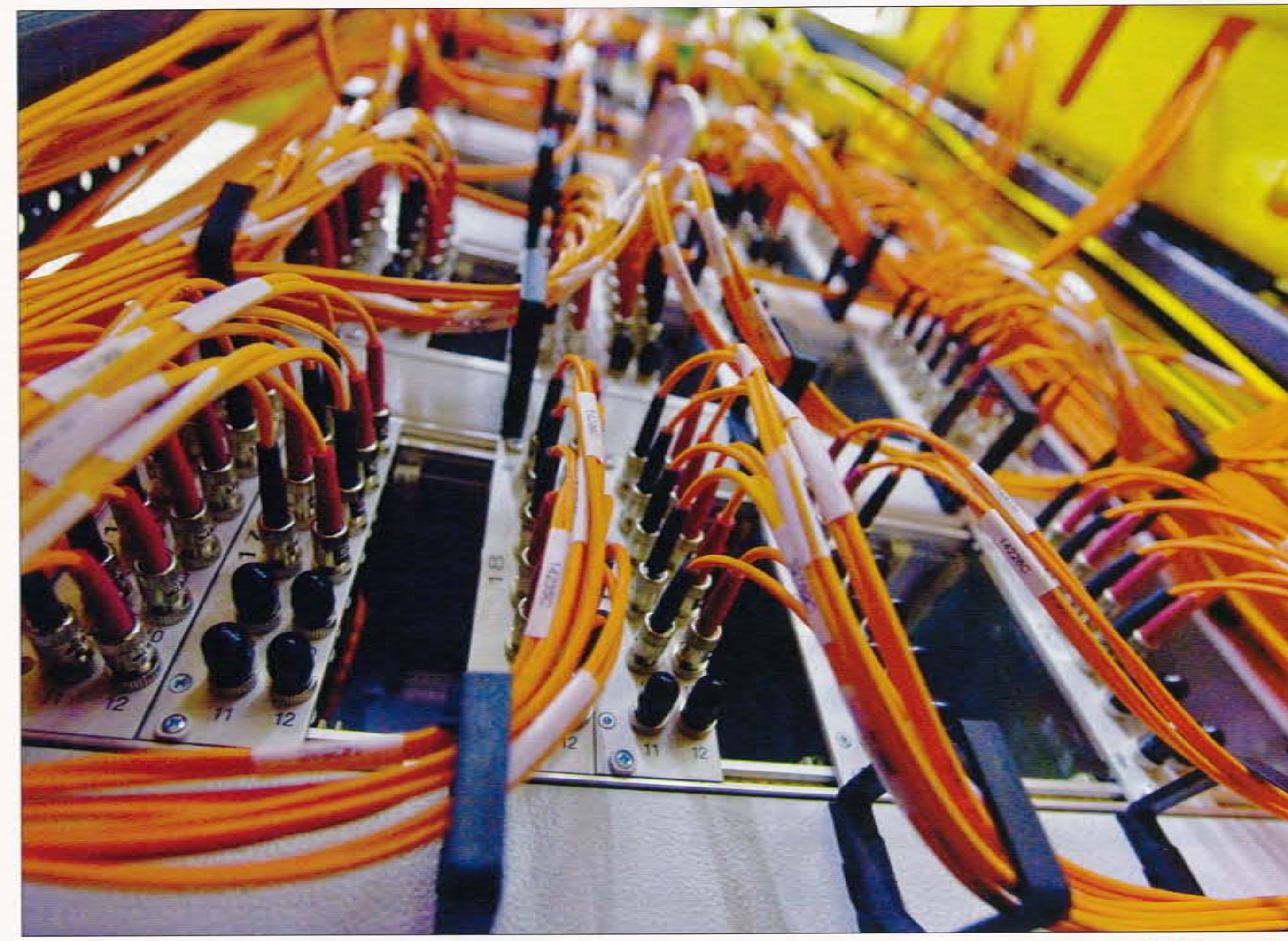
Building on past efforts

The estimated requirements of the LHC experiments were up to 10 000 times greater than the data volume and computing power of their predecessors on the Large Electron Positron (LEP) collider, which CERN closed in 2000. In the time between the end of LEP and the start-up of the LHC there have been a number of experiments on the Super Proton Synchrotron (SPS) that marked important steps towards computing for the LHC. For example, just before LEP was dismantled to make way for the LHC, the NA48 experiment on kaon physics produced data at peak rates of about 40 megabytes per second – only about a factor of five to eight less than what we expected from the LHC experiments during proton–proton collisions. Knowing that the available hardware could deal with such data rates was reassuring, because it meant that by the time that the LHC went online, the hardware would have improved enough to handle the higher rates.

Colliding heavy ions produces about two orders of magnitude more particles than proton–proton collisions, and so data rates for heavy-ion collisions are correspondingly higher. By late 2002 and early 2003, the specifications for the ALICE experiment, which will use both types of collisions to study the strong nuclear force, called for it to take data at a rate of about 1.2 gigabytes per second. Since the requirements for ALICE were so much greater than for the other experiments, it was obvious that if the computing infrastructure could handle ALICE, then it could handle almost anything – and certainly data coming from the other experiments would not be a problem.

To address this challenge, CERN computer scientists and members of the ALICE team collaborated to design a system that could receive data at a rate of 1.2 gigabytes per second from an experiment and handle them correctly. The first large-scale prototype was built in 2003 and was supposed to be able to handle a data rate of 100 megabytes per second for a few hours. It crashed almost immediately. Later prototypes incorporated lessons learned from their predecessors and were able to handle ever higher data rates.

The Large Hadron Collider will produce about 10–15 petabytes of data every year that have to be permanently stored and kept accessible at all times



CERN

Getting connected
All the links in CERN's computing infrastructure must work together for the LHC to succeed.

Another project from the LEP era that helped computer scientists build the LHC computing environment was the scalable heterogeneous integrated facility, or SHIFT, which was developed in the early 1990s by members of CERN's computing division in collaboration with the OPAL experiment on LEP. At that time, computing at CERN was almost entirely done by large all-in-one mainframe computers. The principle behind SHIFT was to separate resources based on the tasks that they perform: computing; disk storage; or tape storage. All of these different resources are connected via a network. This system became the basis for what is now called high-throughput computing.

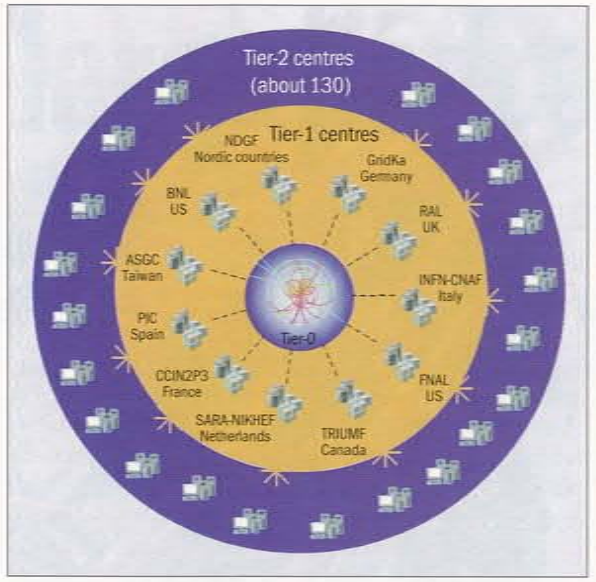
The difference between high-throughput computing and the more familiar high-performance computing can be understood by considering a motorway full of cars, where the cars represent different computing applications. In high-performance computing, the goal is to get from A to B as fast as possible – in a Ferrari, perhaps, on an empty road. When the car breaks down, the race is over until the car is fixed. In high-throughput computing, in contrast, the only thing that matters is to get as many cars as possible from point A to point B. Even if one car breaks down, it does not really matter, because traffic is still flowing and another car can get on the road.

High-throughput computing is ideally suited to high-energy physics, because the "events" recorded by the experiments are completely independent of each other and can therefore be handled independently as well. This means that data analysis or simulation can be carried out on a large number of computers working independently on small chunks of data: the workload is said to be "embarrassingly parallel". In contrast, applications on a good, old-fashioned supercomputer are "highly" parallel: all available computing resources, perhaps tens of thousands of processors, are used for a single computing task.

Separating the different resources made SHIFT very scalable: each resource could grow independently in response to new demands. Also, the physical make up of each resource was largely irrelevant to the system as a whole. For example, more tape drives could be added to the system without the need to "automatically" add additional disk space as well, and old computer nodes could easily be retired and new nodes installed without disturbing the overall system. These aspects of SHIFT – an embarrassingly parallel network of computers, each of which could be updated to take advantage of Moore's Law, all working independently on different bits of data – proved to be the best possible basis for computing in the LHC era.

Distributed computing: the LHC grid

While researchers were developing and testing their respective software frameworks for data taking, data analysis and simulation, the computing environment at CERN and elsewhere continued to mature. Almost immediately after planning for the LHC began in the mid-1990s, it became clear to computer scientists that



Number-crunching network The LHC experiments will produce far too much data for CERN to handle alone. Hence, data analysis and storage tasks will be farmed out to many different sites around the world via a network known as the Worldwide LHC Computing Grid. From the CERN computer centre (Tier-0 of the grid), particle-collision data will be transferred to 11 Tier-1 centres along superfast 10 gigabits per second fibre links. These national sites will store about two-thirds of the raw data in huge tape libraries, while the rest will remain at CERN. Most of the data analysis will take place on computers at about 130 regional Tier-2 centres. Individual physicists can harness the processing power of the Tier-2 centres using links between the centres, university computing clusters (Tier-3), and their desktops and laptops (Tier-4).

the computing power at CERN itself would be significantly less than the computing power needed to analyse the LHC data and perform the required simulations. Therefore, computing power had to be made available elsewhere. The challenge was to build a system to allow physicists easy access to computing power distributed worldwide. That system is now known as the Worldwide LHC Computing Grid (WLCG).

The WLCG was built in a tier structure (see figure above). The CERN Computing Center is Tier-0, and all raw data are permanently stored here. There are 11 Tier-1 grid sites outside CERN, including the UK's Rutherford Appleton Laboratory, Fermilab in the US and the Academia Sinica Grid Computing Center in Taiwan. All of the Tier-1 sites have space for permanent tape storage, and the LHC experiments export their raw data from CERN to these Tier-1 sites. Most of the actual data analysis and simulation is done at about 130 Tier-2 sites.

In total, CERN will export 2–5 gigabytes of raw data to the Tier-1 sites every second. When planning for the LHC started, such rates did not seem feasible. However, by the turn of this century, fibre-optic technology had advanced far enough to make the first 10 gigabit transcontinental and (especially) transatlantic network links commercially viable. In response, CERN teamed up with other institutes and network providers to form the DataTAG project, which explored the potential of such fast links. The resulting collaboration set a number of speed records for transmitting data over long distances, starting with 5.44 gigabits per second between Geneva and Sunnyvale, California, in October 2003. (For historical reasons, network experts measure data in bits per second, while data-transfer specialists measure in bytes per second. A byte contains eight bits.)

For high-energy physics, permanent data storage does actually mean "permanent"

Within a year, transfer rates reached 7.4 gigabits per second, or about 9 DVDs per minute, for data transfers from the main memory of one server to the main memory of another server. This rate was limited not by the network but by the capabilities of the servers. Memory-to-memory transfers were just the beginning, since the actual data will be transferred from disks. This is a lot more demanding; nevertheless, in 2004, using servers connected to an experimental disk system, it was possible to transfer 700 megabytes, or one CD of data, every second, from Geneva to California with a single stream reading from disk – more than 10 times faster than a standard hard drive in a desktop computer today. This showed that the network connections would not be a problem and that the data could actually be transferred from CERN to the Tier-1 sites with the desired data rates. CERN is now connected to all the Tier-1 sites with at least one network connection capable of transferring data at a rate of 10 gigabits per second.

What to do with the data?

The challenges of LHC computing also included much more mundane issues, such as figuring out how to efficiently install and configure a large number of machines, monitor them, find faults and problems, and ultimately how to decommission thousands of machines. Data storage was another seemingly "ordinary" task that required serious consideration early in the planning phase. One major factor in the planning was that for CERN, and high-energy physics in general, permanent storage does actually mean "permanent". After LEP was switched off, physicists painstakingly re-analysed all 11 years worth of raw data that it had produced. The LHC might generate up to 300–400 petabytes of raw data over its estimated 15 year lifetime, and physicists expect all the data to remain accessible for several years after the collider has been switched off.

Data that are being used for computations are stored on disk, of course, but in the long term only data stored on tapes is considered "safe". No other technology has been proven to store huge amounts of data reliably for long periods of time and still have a reasonable price tag. These tapes are housed in libraries that can hold up to 10 000 tapes and up to 192 tape drives per library.

To ensure that the data stay accessible, all raw data are copied to a new generation of tape media as it becomes available. Historically, this has happened every three to four years, although the pace of change has accelerated recently. In addition to protecting precious raw data against routine wear and tear on individual tapes, such regular updates also reduce the number of tapes, because newer versions generally have higher capacities. Access to the data becomes faster with each successive upgrade, since the speed of new tape drives is faster. Old tapes are put onto pallets, wrapped in plastic and stored together with a few tape drives just to be sure that it is possible to access the original tapes again. Back-up tapes are also stored at multiple sites and in different buildings, to try to minimize the loss if

CERN

Access all areas Tape-reading robots navigate libraries full of data.

any disaster was to occur.

In addition to the actual data on particle collisions, the LHC experiments also have to store the "condition" of the detector (i.e. details about the detector itself, like calibration and alignment information) in order to be able to do proper analysis and simulation. This information is stored in the so-called conditions databases in the CERN computing centre and is later transferred to the Tier-1 sites as well. The LHC experiments needed to make changes to the database 200 000 times per second, but the first fully functional data-handling prototype could only handle 100 changes per second. After some intense efforts to solve this problem, Oracle, the relational database manufacturer that supplies CERN, actually changed its database software to allow the LHC experiments to meet the requirements.

One thing that industry partners say about CERN and high-energy physics is that the requirements are a few years ahead of virtually everything else. Pat Gelsinger, a senior official in the digital enterprise group at Intel (which has worked with the IT team at CERN in the past), has said that CERN plays the role of the "canary in a coal mine". By coming to CERN and collaborating with the physicists working on the experiments or with the IT department, industry is able to tackle, and solve, tomorrow's problems today.

The LHC challenge presented to CERN's computer scientists was as big as the challenges to its engineers and physicists. The engineers built the largest and most complicated machine and detectors on the planet, plus many other achievements that can only be described with superlatives. For their part, the computer scientists managed to develop a computing infrastructure that can handle huge amounts of data, thereby fulfilling all of the physicists' requirements and in some cases even going beyond them. This infrastructure includes the WLCG, which is the largest grid in existence and which will have many future applications. Now that the physicists have all the tools they have been wanting so long for, their quest to uncover a few more of nature's secrets can begin. ■