exploring the same physics: as the lifetime of products in information technology is short, hardware and software obsolescence may have to be faced.

Data re-use is undoubtedly valuable for its discovery potential. However, it has an intrinsic limitation: the "raw data" kept for permanent storage are the result of the trigger process (taking place before the initial data recording) designed for specific physics reactions. The bulk of the data produced by the collisions are irreversibly lost and may carry with them undiscovered information.

Over the years, experiments have become larger and more complex, and the size of the collaborations reflects the size and ambitions of the experiments. Collaborating groups are located on all continents. This mode of operation is made possible by fast long distance networking [Highlight 9.4] and by software tools that ensure data coherence. It goes without saying that the Web [Highlight 9.7] and other means of digital communications have been essential to the success of these distributed collaborations.

9.2 Computing Clusters and Data Storage: The New Factory and Warehouse

Les Robertson

Infrastructure for innovation

Since the earliest days of CERN the availability of computational and data storage capacity has been one of the most important factors in enabling the extraction of physics from the data collected by the experiments. It has evolved from the automated analysis of bubble chamber photographs in the 1960s to the worldwide grid used today to distribute the mass of data from the Large Hadron Collider for processing in 170 sites in 42 countries where high energy physicists have access to computing resources. Funding has always been limited, driving a continuous search for new technologies to provide more cost-effective solutions — in data storage, networking and processing power.

In the 1960s and early 1970s the clear choice for computational capacity was the super-computer, designed for fast floating point calculations, though not particularly good for handling large volumes of data. CERN's innovative role in that era included working with the manufacturer to optimize mathematical libraries to exploit best the detailed architecture of the machine. The beginnings of the hierarchical mass storage management architecture were developed, the evolution of which is still continuing today. In the 1970s CERN already had a managed two-level data buffer ("cache") for its vast magnetic tape library.

During these early years the rather visionary idea emerged that physicists needed more than just access to a large computer to run their programs. They

needed to acquire, store and analyse data from experiments and simulations, and do this from all parts of the CERN site. Experimental data files needed to be transported and processed on central batch systems, with control of the job flow and delivery of the results. With the first accelerators, data was transported on magnetic tapes to the computer centre "bicycle online" and physicists carried large trays of cards and reams of paper to and from their offices. But in 1968 the first of a series of CERN-developed site networks was implemented culminating ten years later in the high speed general purpose packet switched CERNET [Highlight 9.3].

By the mid-70s supercomputer architecture had evolved towards support for vectorizable codes. Sterling efforts to exploit this hardware for High Energy Physics (HEP) algorithms were unsuccessful and commercial mainframes with their simpler architecture became the cost-effective solution, ushering in a period of heterogeneity as the competitive acquisition criteria led over the years to the installation of systems from different manufacturers with their proprietary architectures, operating systems and local network media and protocols. The challenge was to interconnect these systems to the site network and, later, to the growing number of incompatible wide area networks that were appearing in support of science in different countries and regions.

By the end of the 1980s and the start-up of the LEP collider, CERN had acquired in-depth experience in storage management, networking and distributed processing. This was a key factor in three major developments that proved essential for HEP data processing over the following decades: cluster computing, extending this to the wide area as a data grid, and of course the Worldwide Web.

Hierarchical mass storage service for the CDC 7600 supercomputer

In March 1972 CERN installed serial number 3 of the fastest supercomputer of its day, the Control Data Corporation 7600. There were three large disks connected to the system providing about 2 Gigabytes (GB) of online storage. Another 600 Megabytes (MB) of disk storage was available on the "front-end" computers mainly used for holding data which was required to be permanently online such as executable files, but the physics data were stored on magnetic tape which had to be copied to the 7600 disks before being accessed. By the end of 1974 there were about 60,000 magnetic tapes in the tape vault, each with a nominal capacity of 20 to 40 MB [11]. In order to manage the movement of data between tape and the 7600 a two-tier storage system was developed in 1973. The user referred to data by the tape number and the system arranged for the tape to be recovered from the vault, mounted on a tape drive on one of the front-end computers and copied into a cache on the 7600 disks. The cache was managed using criteria such as size of file, age, and time since last access. A third tier was introduced a few years later as a tape reel cache, enabling frequently used tapes to be available for immediate

mounting. The system was called FIND [12] and at the time it was not described as a *Hierarchical Storage Management* (HSM) system, a term that was coined only later with the arrival of robotic tape handling devices, but it already had all of the characteristics of an HSM.

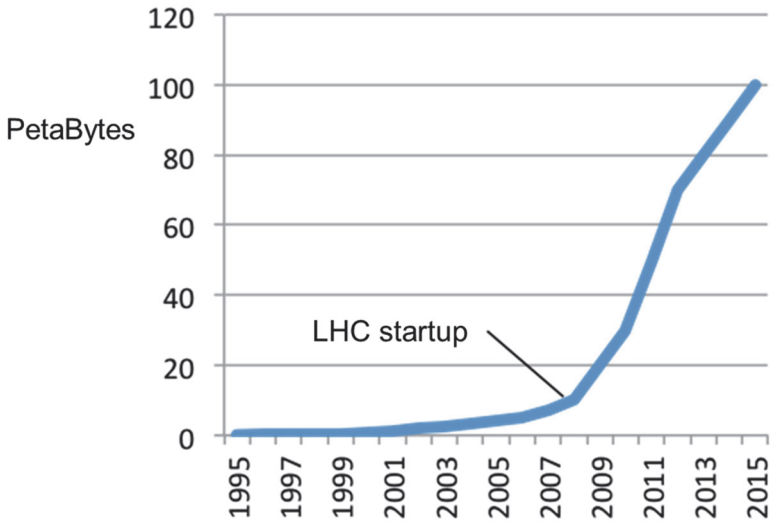

Fig. 9.2. Growth of mass storage usage at CERN.

The basic concept developed in 1973 has remained at the heart of storage management at CERN until the present day. There have been several generations of software integrating new technologies like robotic tape handlers and databases as the volume of data exploded (Fig. 9.2), and evolving to the cluster model with a fully distributed cache [13].

BETEL — Extending cluster, processors and data, across the wide area

In 1993 CERN led a project which implemented the first international demonstration of a computer cluster extended to the wide area using network technology with comparable performance to the local area networks of the time. The project, BETEL (Broadband Exchange over Trans-European Links) [14], used 34 Mbit/s ATM[a] (Asynchronous Transfer Mode) links provided by France Telecom and Telecom Switzerland, to interconnect the 100 Mbit/s FDDI (Fiber Distributed Data Interface) networks at CERN, EPFL (Lausanne, Switzerland), IN2P3 (Lyon, France) and EUROCOM (Nice, France).

[a]ATM was a standard used by telephone companies during the 1990s to provide high throughput data traffic combined with low latency characteristics suitable for voice and video. By the end of the decade it was being displaced.

Using this infrastructure the CERN physics cluster was extended to a similar cluster in Lyon, including access to the magnetic tape services at both sites. After enhancing the RFIO (Remote File Input Output) protocol [15] used for data access to take account of the long network latency, the performance goals were achieved enabling seamless access to all disk and tape resources from both sites. The project also demonstrated interactive physics analysis using CERN's Physics Analysis Workstation (PAW) system and a tele-teaching application used by users at EPFL and EUROCOM.

SHIFT — An early implementation of cluster computing

The computational needs of the experiments planned for the LEP collider from its start in 1989 far exceeded the capacity that could be provided within CERN's computing budget using the mainframe technology that was the backbone of the computer centre. In HEP data processing the major part of computational resources are used to process very large numbers of independent events. Events can therefore be processed in parallel, and using systems with different performance characteristics: the requirement is *high throughput* rather than *high performance*. A small team was given the job of looking for a distributed solution that could exploit low cost components integrated with the central computers and with the storage infrastructure essential for efficient data analysis.

There was good experience with microprocessors for online use of the experiments, but these had limited floating-point performance. Specialized processors had also been developed, implementing an instruction set sufficient for certain physics codes but without general I/O and network capability [16]. More promising were the "personal workstations" being used at CERN in a limited role for interactive graphics applications, and based on the new single-chip reduced instruction set (RISC[b]) processors. Personal workstations with the graphics functionality removed gave an order of magnitude improvement in price/performance over the mainframes. The only problem was how to integrate them into the physics analysis service.

The project SHIFT (*Scalable Heterogeneous Integrated Facility Testbed*) [17] developed a straightforward architecture providing scalability and accepting heterogeneity that would enable the service to expand smoothly as the underlying data storage and computational resources grew. It was seen as essential to construct the system with off-the-shelf hardware components, in order to be able to exploit new technologies and cost opportunities as soon as they arose. The key idea was to define separate services for mass storage, data cache management (disk) and

[b]Compared with traditional processors of that period RISC processors provided higher performance by using a simplified instruction set that enabled execution to be completed in fewer microprocessor cycles per instruction.

computation loosely integrated across a high performance network backbone. Each service would be implemented as a set of independent servers to ensure scalability. The software was designed to be portable across systems providing a Unix-like system interface and a TCP/IP[c] (Transmission Control Protocol/Internet Protocol) network service. The fundamental components that had to be developed were: a distributed cache management system integrated with the site mass storage service; a remote file access system [15], later, re-routing in the event of a link failure; a job scheduler to manage the workload across the cluster, initially based on a system developed by NASA.

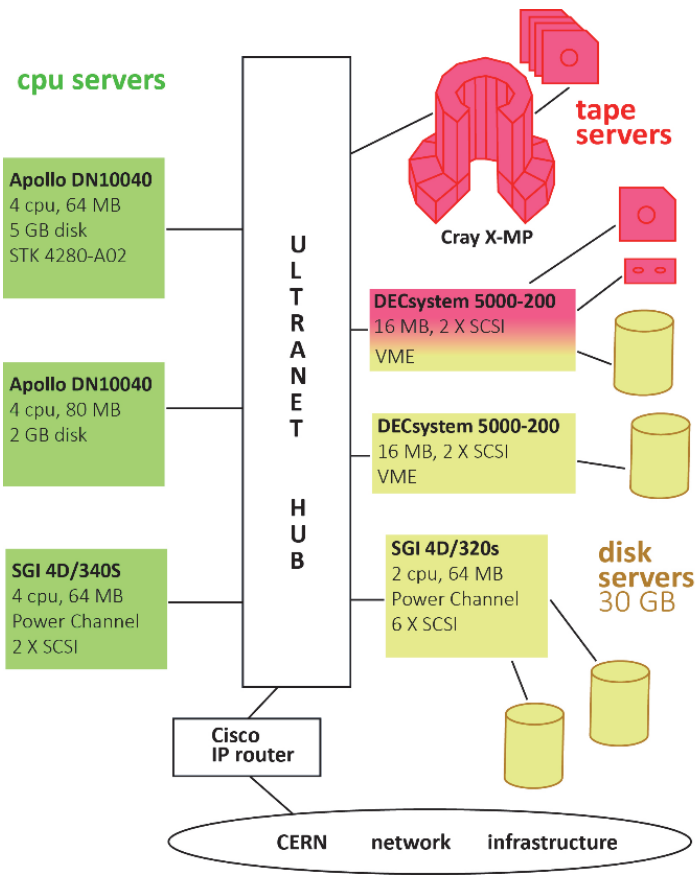

Fig. 9.3. Pilot SHIFT configuration at CERN in 1991.

[c]TCP/IP is the most widely used communications protocol. It was invented and developed in the 1970s at DARPA (Defense Advanced Research Projects Agency), USA.

The initial implementation of the architecture using Hewlett-Packard Apollo DN10.000 computers as disk and computer servers, could already rival the largest mainframe in terms of aggregate processing capacity, but lacked efficient access to mass storage and was only used for simulation work. The cluster was soon enhanced with the addition of systems from Silicon Graphics Inc., Digital Equipment Corporation, SUN Microsystems and a Cray X-MP as the mass storage server, all interconnected by a high-speed network from UltraNet. Figure 9.3 shows the configuration at the beginning of 1991. The unusual choice of a supercomputer as a mass storage server was made because, in common with the other systems in the cluster, it used a version of Unix as operating system, had a very good implementation of TCP/IP network protocols, and could be connected to the UltraNet backbone. From this early start, the cluster grew rapidly, and within a few years supplanted the mainframes as the workhorse of physics simulation and analysis. The architecture, in large part because of its simplicity, proved its flexibility, absorbing successive generations of disk and processor technologies. PCs were introduced in 1997 as soon as their performance on HEP codes became acceptable and they in turn soon displaced the RISC-based workstations.

The Worldwide LHC Computing GRID — A Cluster of Clusters

When planning the facilities for analysing the data from LHC began in 1998, once again it was clear that the budget for computing at CERN would fall far short of supplying the needs. The CERN cluster was by that time largely powered by PCs which seemed to offer a very promising future for raw performance and cost effectiveness — no better technology was on the horizon. However, many other HEP sites had also installed powerful clusters. A neat way had to be found of interconnecting them without the physicist having to worry about where the available resources were and where the data was located. A group of physicists and computing experts from a broad selection of HEP institutes got together to study possible solutions, producing a recommendation in 2000 for a multi-tier hierarchy of inter-connected regional centres, each with its own cluster and storage [18]. The tiers were defined by the type and volume of data that would be stored there: the top level, Tier-0 (CERN, and Budapest from 2013), would store a copy of all of the raw and simulated data and of the processed data required by all of the physicists in an experiment; Tier-1 sites would be large operations with mass storage services, and each Tier-1 would hold a copy of a fraction of the data held at CERN (ensuring at least two copies of all important data); lower tier sites would store only temporary copies of the main datasets, obtaining copies of datasets as required from a higher tier. The end-user would not have to know where the data was located or where there were free resources. The work would be automatically split into suitable processing units and directed to an appropriate regional centre.

Fig. 9.4. A grid of 170 interconnected computing clusters in 42 different countries provide the processing capacity for analysing LHC data (Sept. 2015). Budapest and CERN are Tier 0 sites.

While the responsibilities for operating services and managing data were defined as a *hierarchy* which would have implications for planning inter-site network bandwidth, it was important that the lower-level software should view the system as fully inter-connected, providing complete flexibility to job schedulers, data management services and other higher-level components: the logical architecture would be a *grid*. The *middleware* (software) required to operate the grid would be complex and rather than developing a special LHC solution it was decided to start with the *Computational Grid* technology being developed in the GLOBUS project in the USA [19]. The Italian Nuclear Physics Institute, INFN, already had experience with the Globus toolkit. The grid concept looked to be of general interest for scientific applications and CERN took the lead in launching an international multi-science project, the *European Data Grid* (EDG) [20], which received funding from the European Commission. EDG was a proof of concept project which developed an initial implementation of the middleware needed to complement the Globus toolkit and deployed this on a demonstration grid. A similar project, the Grid Physics Network (GriPhyN) was begun in the USA.

The grid concept and experience with early prototypes was promising and in September 2001 CERN set up the *LHC Computing Grid* (LCG) Project to move ahead with planning and deploying a grid on the scale required for LHC era. LCG collaborated closely with EDG, GriPhyN and successor grid projects, but with the clear goal of focusing on the functionality, delivery schedules, reliability metrics, interoperability and other factors needed to ensure an operational grid with the performance and timescales imposed by the LHC experiments. The first LCG

deployment was in September 2003 with 14 sites, followed by rapid growth in the number of sites and the installed resources. The differing goals of the R&D grid projects and LCG with its emphasis on service stability would be the cause of considerable conflict in the ensuing years, particularly concerning the functionality and timescales of the middleware. With the deadline of LHC start-up approaching, a compromise was worked out and the Worldwide LHC Computing Grid was in round-the-clock operation at 135 sites when data began to flow in 2008 [21].

A major concern in 2000 was the network bandwidth that would be available (and affordable) between the centres and much effort was invested in working with national and international bodies that provided network capacity for science to ensure that their solutions took account of the relatively high requirements of LHC. This work paid off and by the time that LHC began operation CERN was well connected to a large, mainly fibre, international infrastructure [Highlight 9.4]. Indeed, today, fifteen years after the first proposal for the distributed model, the available bandwidth enables a much more flexible model for data distribution: instead of scheduling work to where the data is located it is now possible in many cases to move data on demand to where there are free computational resources.

9.3 Local Area Networks: Organizing Interconnection

Ben Segal

The ever closer and increasingly important interconnection of particle experimentation and accelerators with Information Technology (IT) is the main theme of this chapter. In this highlight we illustrate one essential facet: the evolution of network concepts.

Innovating in the Early Years — 1960s and 1970s

CERN rarely develops computer networking technology *per se*, but is often an early adopter of the latest technology, operating it at its technical limits. This was different, however, during the 1960s and 1970s, when CERN had no choice but to innovate due to the limited offer of suitable commercial products. During these years several networking systems were developed in-house, presented here in chronological order, and reflecting the rapid evolution of IT and its growing importance for high energy physics. A comprehensive survey is given in [22].

In the early 1960s CERN's first computer, the Ferranti Mercury, was connected by a one-kilometre *data link* to electronic experiments in the PS experimental hall, making it the first computer at CERN to analyse on-line experimental data in "real-time". The complete link interfacing was designed and built in-house.