Computing Trends & Strategy at CERN

Les Robertson

CERN, IT Division, 1211 Geneva 23, Switzerland
les.robertson@cern.ch

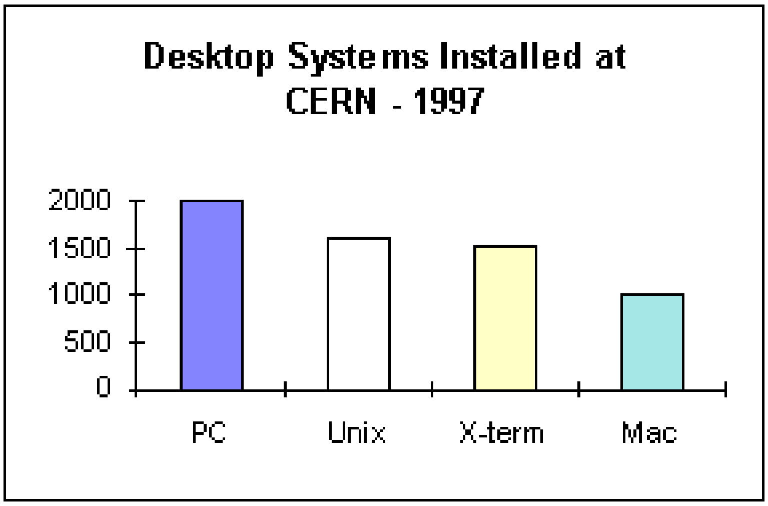
This paper summarises current trends and strategies at CERN in a number of areas of computing services: desktop; local networking; physics computing farms; data management.

1 Introduction

Computing strategy at CERN is the product of many interacting committees and processes:

- The Information Technology Division makes proposals for new services, policies and strategies, and has the final responsibility for implementation of the results of the decision making process.
- The COCOTIME committee (chair: Lorenzo Foa) undertakes a peer review of physics collaboration resource requests and defines annual budget allocations.
- The FOCUS committee of physics users (chair: Manuel Delfino) reviews proposals made by IT Division and others and makes policy recommendations for physics services.
- The Desktop Forum representing CERN users (chair - Horst Wenninger) makes policy recommendations for desktop services.
- The LHC Computing Board (LCB - chair: Mirco Mazzucato) nurtures common development projects between LHC experiments, and thereby exerts an important influence on long term strategies.

As a background to all this, the most important factor influencing CERNs computing strategy is the evolution of the consumer/commodity computing market that provides most of the components from which computing services are constructed.



2 Desktop

The desktop environment at CERN exists in three distinct flavours: PCs under Windows 95, Apple Macintosh, and Unix. In a minority of cases Unix runs directly on the desktop in a private workstation, but more commonly a work-group cluster or central time-sharing service is used, viewed from a desktop X-terminal or an X-Window in a PC, Mac or ageing RISC workstation. The trend, proceeding at the rate at which obsolete systems are replaced, is towards a domination of PCs, with RISC workstations becoming limited to those with a real need for a private Unix system for program development or interactive physics data analysis. The Mac is officially deprecated and very few new systems are being purchased.

The most important factor affecting the choice of desktop is the total cost of ownership - not only the direct cost of purchase, warranty and hardware repairs, but also the cost of the network and file service infrastructure and, most importantly, the management costs. The PCs at CERN are supported by the NICE[1] system, developed over the past five years. NICE implements a low administration cost environment, with software and user files maintained on central servers. Individual PCs have the minimum set of programs and initialisation files installed locally, and the environment is brought up to date each time the PC is restarted. A user can log in to the service from any PC, which is then personalised using his private profiles. NICE has kept the costs of support and administration well under control. The 2,000 PCs at CERN are supported by a central team of 4 people, assisted by one person in each of the user divisions who installs the hardware and provides first-line user support. The administrative duties of the PC owner are very close to zero. While keeping the management costs low, NICE imposes common choices on

[1] For more information see http://nicewww.cern.ch/doc

software versions and desktop style which produce both positive and negative reactions from the community. A new version of NICE has been developed for the Windows NT environment, serving a small but growing number of users.

In the longer term CERN's desktop strategy for hardware and software will be dictated by market trends. The NICE environment will evolve towards use of Microsoft's "Zero Administration Windows" feature when that becomes available. The integration of portable PCs, carried between home, CERN and the hundreds of institutes within the High Energy Physics community will be an important problem to solve, but a problem that we have in common with many other communities. Network Computers (NCs) and Java-only Stations may have a limited role for some applications, but are unlikely to be flexible enough for the needs of the general user.

3 Local Networking

Over the past two years two major changes have been made to CERN's local networking infrastructure. The topology of the general purpose network has evolved from a single bridged network to a routed configuration with sub-nets organised on a geographical basis. At the same time a new structured cabling infrastructure has been installed providing several outlets in each office, each with the potential of carrying over 150 Mbps. In practice the current service is 10baseT Ethernet on the desk, with CDDI[2] available for selected high performance servers (in particular the physics data processing servers in the Computer Centre). The backbone of the general purpose network is based on multiple FDDI rings bridged with a DEC GigaSwitch.

The Data Network interconnects the components of the CORE[3] processing farms in the Computer Centre, with extensions to private farms operated by physics collaborations and to data acquisition systems located at the experimental sites. The basic service provided by the data network is FDDI/CDDI in the form of a bridged network based on a second DEC GigaSwitch. A scalable bandwidth connection to the general purpose network is provided using multiple routers. The data network also provides a high performance service using HiPPI[4] for systems that require higher aggregate bandwidth than available

[2] CDDI is the copper version of FDDI - an IEEE standard for 100 Mbps local area networking. For more information see -
http://sholeh.nswc.navy.mil/x3t12/overview.html
[3] CORE - The Centrally Operated RISC Environment - the collective term for the s ervices provided in the CERN Computer Centre for processing and storage of physics data.
[4] HiPPI - the High Performance Parallel Interface - a network standard supportin g an 800 Mbps data rate. For more information see the HiPPI Networking Forum

with FDDI. Examples are large multi-processor Silicon Graphics Challenge computers, and servers supporting high performance StorageTek tape drives. The basic (FDDI) and high performance (HiPPI) services are integrated using Ascend GigaRouters.

A significant recent development involving local networking at CERN has been the growth in central recording of experimental data. Almost all current experiments now send their raw data over the network to the Computer Centre to be written to automated magnetic tape storage.

In summary the infrastructure and topology of the local network are in good shape, with sufficient flexibility and scalability to satisfy the requirements for the next few years. Looking to the future, 100baseT Ethernet is ready for deployment as required, and two potential technologies are being looked at for use in the backbone and data network. ATM is being evaluated in several pilot projects, and an evaluation of GigaBit Ethernet [5] is being planned.

4 Physics Farms

The physics data processing services at CERN are organised as a series of large processor "farms", constructed from about 250 separate computers containing over 500 processors with 8 TeraBytes of attached disk. The computers vary from simple and inexpensive "headless" workstations, through shared memory multi-processor (SMP) systems with 2 to 32 processors in a box, to Scalable Parallel Processors (SPPs) such as the QSW [6] /Meiko CS2 and the IBM SP2. In all cases the operating system is Unix. The simpler systems are mainly used for computation intensive applications such as simulation, while applications that require substantial storage capacity and high I/O performance use SMP or SPP systems with high performance network connections.

The evolution from the mainframe-based computing services of the 1980s to today's environment using RISC processor technology and inexpensive disks developed for desktop scientific workstations has enabled us, over the 9 years since the start of LEP operation, to provide 150 times more processing capacity and 100 times more disk space within an annual budget that has decreased by over 40% in real terms. Good work perhaps, but with a continuing demand for increased capacity and a budget under even greater pressure we must keep up the search for less expensive technology. Two pilot services are being set up

---

home page at - http://www.esscom.com/hnf/index.html

[5] GigaBit Ethernet is in process of becoming an IEEE standard providing a bandw idth of 1000 Mbps. See http://www.gigabit- ethernet.org/

[6] Quadrics Supercomputing World

using PC components: the Intel Pentium Pro processor and the Windows NT operating system. One of the services is a compute intensive simulation farm, testing out the compilers, processor performance, and basic usability of the PC environment. The second will tackle the more demanding requirements of an interactive analysis service.

An important and related development has been the contracting out of the basic systems management of the physics farms. There were two motivations for this. Firstly, CERN's staff policy foresees a continued reduction in staffing-levels over the next five years, without any corresponding reduction in the requirements for computing services. Secondly, the overall cost of providing computing facilities in this environment is in most cases minimised by the use of larger numbers of inexpensive boxes rather than a few big integrated but more expensive multi-processor computers, even when full account is taken of the management and administration costs. It is important that the financial argument is not distorted by the inflexible staffing policy. Outsourcing systems management has freed CERN staff to concentrate on wider aspects of service management and introduction of new technologies, and has introduced more objectivity into the assessment of overall costs.

How do we expect the physics farms to evolve? It seems clear that the Intel Pentium Pro and its successors will be an important element. Today simple but complete systems based on these processors have a factor of 3 to 5 advantage in price/performance over RISC systems of similar power. PC technology must now be the prime candidate for building simulation and reconstruction farms where I/O is not the prime concern. We must, however, be a little careful in comparing costs. The total cost of ownership includes significant factors other than purchase and maintenance, such as management, provision of standard physics libraries, and the cost to the end user of learning a new environment and porting applications to it. We must also realise that the move from mainframes to RISC has been accompanied by a major change in the relative importance of the cost of the computer itself in the total cost of providing services. The pie chart on the next page shows the distribution of costs in 1997 for providing physics data processing services at CERN. Computers (including maintenance and systems management) account for only 37% of the total expenditure.

Despite the evident financial advantages of the PC technology there are a few important questions to be answered. When will PCs be able to handle storage intensive applications as well as RISC systems do today? The margins practised in the PC business do not allow for the level of expert support and advice available from the suppliers of scientific workstations - does it matter? PCs are sold as consumer electronics; design deficiencies/faults are only fixed in the next version. Similarly, most software problems will be fixed at best in a future release. Will this be a problem? We must at least learn to be much

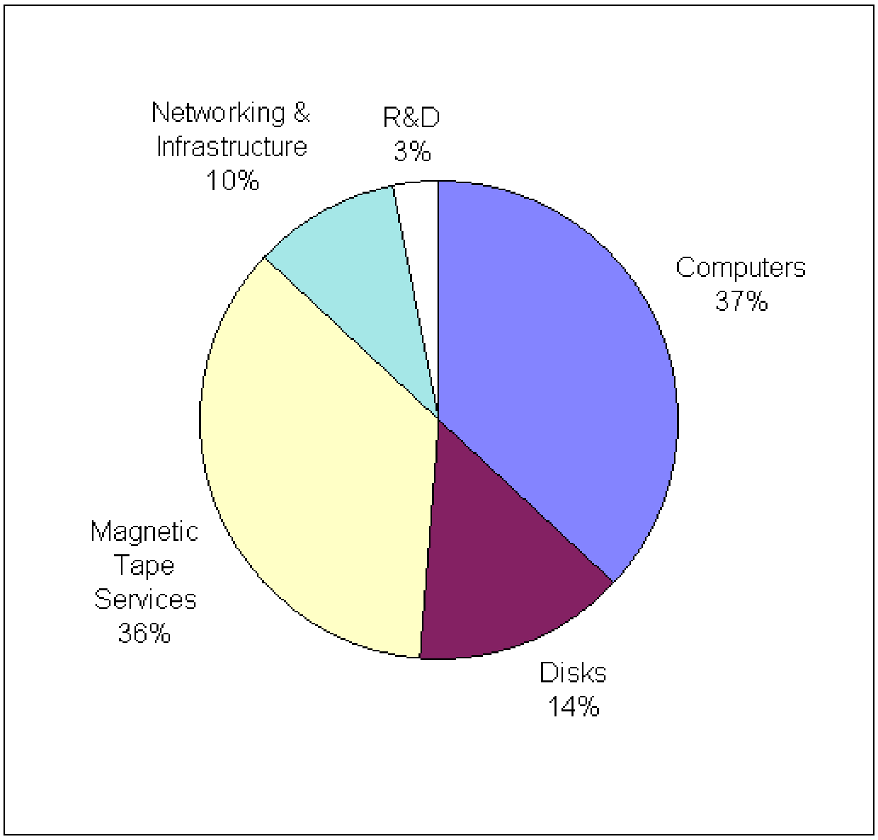


Fig. 1. Physics Data Processing Costs at CERN in 1997

more resilient than we have been in the past. Finally, there is at present a great deal of emotion about the choice of operating system for the PC in the physics environment - basically a debate over Linux and Windows NT. Both have their advantages and disadvantages, and we will see both of them in use in the shorter term. In the longer term it is going to be difficult to overlook the advantage of using Windows everywhere - at home, on the desk and while travelling.

5 Data Management

5.1 Objects

Practical experience has been acquired over the past two years in the area of persistent object storage within the RD45[7] project, which has the goal

[7] RD45 - A Persistent Object Manager for HEP - An R&D project including collabor ators from CERN, LHC and other experiments. More information can be found at: http://wwwinfo.cern.ch/asd/cernlib/rd45/

of demonstrating that an object database can meet the functionality, performance, scalability and reliability requirements of High Energy Physics (HEP). The experience of RD45 is the subject of several papers at this conference. The project has been using the Objectivity/DB [8] product with considerable success, and has shown that the performance is as good as obtained with classical HEP data access methods. This year the pilot installation at CERN is being used to store real physics data by the NA45 experiment and groups working on the CMS and ATLAS test beams. Dependence on a particular product is minimised by adherence as far as possible to the ODMG [9] standards. It is foreseen to increase progressively the scale of the CERN configuration over the next few years to enable extensive experience to be obtained with very large databases. An important recent development has been the adoption of Objectivity/DB by the BaBar experiment at SLAC, which will provide invaluable experience within the community.

Data management is one of the few areas of computing where High Energy Physics has special needs that cannot be met by mass market products. For the foreseeable future HEP will be an important, although by no means the only, user of very large object databases. While this causes some worries about the long term availability of suitable products, it also brings the opportunity to influence the development of current products to satisfy future HEP requirements.

5.2 Storage Management

Storage management at CERN is based on the "SHIFT" [10] software developed in the early days of LEP. This implements a "staging" model where the master copy of the data is held on magnetic tape, and the active subset is cached on disk. The implementation is highly distributed, and therefore scalable in terms of performance and capacity. In particular the disk cache is organised as a series of logical "pools" each of which may be distributed across several nodes of the configuration, transparent to the application. The access to magnetic tape is also distributed across many independent tape servers, with the "data mover" transferring data directly between the tape and disk servers. The negative side is that the software is home made, with consequent long term maintenance issues, and there is no tape space management - that is left to the application.

[8] See http://www.objectivity.com/

[9] ODMG - The Object Database Management Group - see http://www.odmg.org/

[10] SHIFT - The Scalable Heterogeneous Integrated Facility for HEP Computing - Jea n-Philippe Baud et al. in Computing in High Energy Physics 91, Universal Academy Press, Tokyo, 1991.

The massive requirements for data storage of the LHC experiments are beyond the scope of the current software but the question is: what is the alternative? There appears to be only one product available that offers the correct scale of capacity and performance - the High Performance Storage System[11] (HPSS). This was developed by a consortium including IBM and several research laboratories and universities in the United States with very large requirements for storage. HPSS looks good - it has the right functionality, it is well designed and implemented, and it is the only mass storage management product with scalable/distributed disk and tape access. Its users include a number of Department of Energy and NASA laboratories with long term storage needs and assured funding. There are, however, a few open questions. Is there a "critical mass" of users sufficient to sustain the development and support of this complex software in the longer term? At present IBM is the only commercial company involved - does this mean that support for other manufacturers' hardware may not be assured? The first production release of the software took place only last autumn, and so it is still rather early to know just how well the product works. However, the most important question is: is there an alternative? The answer appears to be "no", and so CERN will acquire HPSS in the next few months and begin a project to use it to provide storage management support for the Objectivity/DB databases of the RD45 project. An effective collaboration of the HEP sites using or studying HPSS will be an important element in helping to guide the development of the product and ensure its long term success.

6 Hardware

CERN has recently completed a tender action to acquire a new automated magnetic tape storage system with the performance and capacity to handle requirements for the next five years. The tender was won by StorageTek with three Powderhorn cartridge handlers and 32 Redwood tape drives. The system can store about 18,000 cartridges, each with 10, 25 or 50 GBytes of data. The configuration is being installed in two phases, the first of which, about half of the full configuration, has just completed the acceptance test, and the remainder will be installed in January 1998. The tape drives are supported by DEC 4100 tape servers which have FDDI and HiPPI connections to the CERN data network. To help with the migration process the "legacy" data currently stored in the CERN tape vault on about 200,000 low capacity cartridges will be copied automatically into the new facility and made available for read access transparent to the applications. Partial "lights-out" operation of the tape service (with no manual tape mounting support during certain periods) is scheduled to begin at the end of the year.

[11] HPSS - see http://www.sdsc.edu/hpss/

The smaller existing automated facilities using IBM 3590 and DLT drives will be retained, and indeed these systems currently support about two thirds of the 18,000 weekly tape mounts. Data exchange with other laboratories and institutes continues to be supported on a variety of media types: predominantly DLT technology, but also using EXABYTE, DAT, 3480, 3490 and 3590 formats.

7 Summary

Desktop: The PC is rapidly becoming the standard desktop device, providing the basic interactive support (mail, Web access, office tools, X-Windows, ...).

Local networking: The routing topology and cable infrastructure are in very good shape now and it will be relatively easy to evolve towards higher performance desktop, backbone and physics links as required.

Physics Farms: Intel chips likely to become the predominant supplier of computational horsepower, but RISC workstations and Unix have some life left for storage-intensive applications.

Data Management:

- We can be optimistic about the use of Objectivity as a persistent object storage system.
- The only storage management system which comes close to meeting the long term requirements is HPSS, but its suitability in practice has yet to be proven.
- The data storage hardware configuration at CERN is in good shape for the next 5 years.