CERN/Data Handling Division
DD/88/1

The MUSCLE Report

The Computing Needs of the LEP Experiments

B. Carpenter/DD
C. Jones/DD
G. Kellner/EP (Aleph)
R. Mount/EP (L3)
S. O'Neale/EP (Opal)
L. Pape/EP (Delphi)
L. Robertson/DD
D. Ward/EP (Opal)
D. Williams/DD - Editor

January 15, 1988

Contents

Chapter 2

General elements of LEP computing

It has been known since the early days of planning for the LEP experiments that there was a "LEP computing problem". There has been a common misconception that the problem is entirely financial, and indeed much of the discussion in 1984 and 1985 concentrated on the mismatch between the processor (CPU) resources that the experiments believed that they needed for "batch" processing, and the resources that CERN felt that it would be able to afford to provide at the CERN computer centre on the LEP timescale.

In this context, at least three important things have happened in the past three to four years:-

1. The cost of computer hardware has continued to fall at a rather spectacular speed, and this, together with active negotiation, has permitted CERN to acquire much more capacity in the computer centre than seemed probable in early 1984.

2. All of the experiments took the message from the CERN Directorate in 1984 – that CERN would not be able to provide all of the necessary capacity, and that the experiments had better make their own arrangements to deal with any shortfall – very seriously. Indeed the first installation of a major "private" computing facility for a LEP experiment is now taking place, in the "Barn" of the computer centre.

3. The experiments and the CERN specialists involved have come to a clearer understanding of the needs of the offline computing. These needs turn out to be much more complex than the provision of a given number of units of batch processing power. They also involve questions of data storage technology; of the rôle that can be played by "cheap Fortran farms" (whether these are composed of emulators or other special-purpose architectures); of links between the LEP pits and the Meyrin site; of the connections between the CERN computer centre mainframes and any mainframes belonging to the experiments; and of the extensive use of workstations for interactive physics analysis.

We can perhaps summarise by saying that at this point in time, less than two years from first beam, we are now able to form a much clearer picture of the offline computing needs of the LEP experiments. We now have to explain these needs to the managements of CERN, the outside laboratories and the experiments. They will then be able to understand the technical approaches that seem to us to be dictated by the state of computer hardware and software products that either exist already or will be available on a very short timescale. In this way the necessary money and manpower can be made available in time.

2.1 Monte Carlo simulation

There are a number of tasks in the process of analysing LEP data which will use Monte Carlo simulation. One such task deals with the prediction of those particles (hadrons etc.) that will be observable, according to a given theoretical model of the interactions of quarks, leptons and gluons, after a positron and an electron collide. Other tasks deal with the response of the detectors in the experiment to measurable particles with different characteristics - for example, what fraction of 15 GeV/c charged pions will be mis-identified as electrons?

When a physicist has a set of measured events, and (s)he wishes to investigate a theory concerning the production of these events, then the strongest test requires that a Monte Carlo simulation should be performed which generates raw data, according to the theoretical model, in the same format as that used for real data acquired from the experiment. These data are then processed and analysed using the normal programs, in order to demonstrate that the measured and simulated data are statistically indistinguishable.

We know that many more simulated than real events are needed in order to understand the errors in a given sample. One of the main challenges of LEP computing will be to do enough simulation, but not so much as to waste time and effort, to convince physicists that the experiment has understood the response of the detectors to events that would be produced by a given theoretical model. However, the processor time needed to simulate the behaviour of all secondary particles as they decay, interact, spiral and shower in all the elements of all the detectors is so enormous that it is generally impractical to test physics models using only such detailed techniques.

As a consequence, faster simulation techniques ("Fast Monte Carlo") must freqently be adopted. In this case the fully simulated output of the physics process is coupled to a simple parameterisation of the detector response. If this approach gives inadequate results then, for some detectors or detector components, the simple parametric model must be replaced by a more or less full simulation. For certain applications the fast Monte Carlo simulation programs can also conserve computing resources by producing their desired output directly in the form of histograms or n-tuples, rather than generating intermediate disk files for subsequent interactive analysis.

A typical application of fast Monte Carlo simulation would be to explore one particular physics process at a level where the finest details of the detector response are not important. Another use would be to study the sensitivity or stability of a preliminary physics result to the parameters of the physics model. In this way existing simulations can be better understood, and optimum parameters selected before large processor resources are committed to a full Monte Carlo simulation.

It should be remembered that, in addition to the processor intensive load coming from the Monte Carlo simulation itself, there is a non-negligible computing load resulting from the "normal" processing of events that are simulated at the raw data level, including the processor (CPU) and data storage aspects, and from the subsequent manipulation of all simulated events at the DST level.

2.2 Acquiring the raw data

It is the task of the teams dealing with the online systems to make sure that the detectors are working correctly, and to make available for further processing a sample of events containing a minimum amount of "background", while ensuring that no "interesting" events are discarded in a biased way. This process, normally called "data acquisition", is complex and difficult, but it was not the task of MUSCLE to look into this area. There are, however, two choices to be made here which will have a significant influence on the offline processing:-

1. The technology selected for recording the raw data is obviously a very strong candidate to be chosen for recording information at all subsequent stages of the experiment's data processing.

2. The decision on whether or not to install a dedicated high-speed link between the online computers at the LEP experiment, and the computers on the Meyrin site, potentially has a strong impact on the feasibility of generating the master DST in near-real time.

2.3 Compressing the raw data

One experiment is investigating the possibility of running a filter/compress pass after acquiring the raw data, but prior to generating the master DST. In this pass they would hope to remove background (non-Z^0) events, to compress data from certain detectors, and to apply calibration corrections. They feel that compression by at least a factor of 4 would be needed in order to make such a pass worthwhile.

2.4 DST terminology

When an event has been captured by the online system, or generated by a simulation program, it must be processed to produce a concise summary of all of the information that is likely to be useful in the subsequent physics analysis. The conventional name for this concise summary information is DST, for Data Summary Tape. Even though these data are unlikely to be stored on *tape* in the future we have retained the name DST in this paper.

The volume of the summary data produced by the LEP experiments will be so large that it will be necessary to make several successive selections in order to extract the physics from the data. In the next four sections we discuss some terminology for the different stages of DST selection. Further information, especially on data volumes and processing times, is presented in Chapter 5.

2.4.1 Master DST

This is the reference copy from which all other DSTs are derived. The experiments are studying carefully the information that will need to be retained in the master DST. It will, in general, be important to be able to carry out some reprocessing on these data, since the cost of having to re-access the raw data in a random way is likely to be prohibitive on any large scale. The sort of reprocessing that it will be

possible to carry out includes modification of the association between tracks and calorimeter clusters, and removal of poor measurements in the central detector.

In our estimates (see Section 5.1) we have assumed that the master DST contains 20 Kbytes of data for each event, giving a reasonably concise description of the processed detector measurements, of the particle trajectories and of the particle identifications. Some experiments express strong reservations with respect to this assumption, and believe that it might be necessary to retain much more of the raw data.

We should also point out that many previous experiments at CERN, including both UA1 and UA2, have retained all of the raw data as part of the information stored for each event at and even beyond the master DST stage. Indeed, for many experiments, the total number of tapes used has typically been a factor 3-5 times higher than the number of raw data tapes.

It is possible that the approach of retaining the raw data in the master DST will have to be adopted at LEP for a few special filtered channels (essentially searches for rare particles), where there is a strong probability that the processing of the raw data will have to be re-run several times. However, we must emphasise that the problem of handling the volume of the LEP DSTs is so severe that every possible effort must be made to avoid this.

2.4.2 Team DSTs

We expect that within each LEP experiment there will be many teams of physicists, each working on an area of the physics analysis. Once each team is convinced that the events with which it is working have been correctly measured, and that there will be no need for further re-processing, then they will have to use more compact formats for their further studies.

These formats will normally contain a set of data extracted from the master DST, consisting either of a subset of the events, or of a subset of the summary data for each event. When the size of this team DST exceeds a certain threshold, which we have assumed to be 2 Gbytes, it will be necessary to split the team DST into a full team DST and a condensed team DST.

2.4.3 Personal DSTs

Each individual physicist working on an analysis will have a personal set of summary data selected from the team DST and dealing with events of interest. Probably at this level, and certainly at the next level down (personal active event sample), most of the data retained will be interactively selected components (or "n-tuples") of the summary data.

In preparing our estimates we have assumed that this personal DST has a volume of 100 Mbytes, corresponding to roughly 200K events with 500 bytes of data per event, or 2 Mevents with 50 bytes of data each.

2.4.4 Personal active event sample

Each physicist will also have access, for detailed interactive work (which we believe should be carried out on a workstation), to a very condensed set of data (n-tuples) of about 5 Mbytes. This corresponds to roughly 10K events with 500 bytes of data per event, or to 100K events with 50 bytes each.

2.5 Generating and maintaining the master DST

The phase of the processing from raw to master DST has often been thought of, by earlier experiments, as the main batch production load. It has involved tape to tape processing with both heavy CPU and I/O requirements. After an initial phase of debugging and tuning, the running of the programs for all of the events could be carried out, on a "production" basis, by a few people on behalf of the whole experiment. Although it will still be a major component of the LEP computing load it will not be such a dominating influence as in the past, as the load from other areas grows in importance.

There are a number of issues that seem crucial for the LEP experiments:-

1. How soon after data taking will this processing be performed?

   This is a complicated question to answer. There will be a clear advantage available in the early days of LEP to any experiment that is able to produce its master DST in near-real time, say within a few hours of the data taking. However, there are reasons related to the physics goals and design of the experiments which will make it hard for some of them to process the data so quickly and in a definitive way.

2. Where will this processing be carried out - at the experiment or at (or close to) the computer centre?

   If it can be done at the experiment, using processors connected to the online system, then the data volume to be transferred to the computer centre should be much smaller than if all of the raw data have to be transferred. This factor might make it feasible to use links, rather than to physically transfer the data storage medium holding the raw data.

3. Do the raw data have to be recorded using the same storage medium as the master DST?

   We later recommend that, in view of the heavy manpower effort that will have to be devoted to mastering this area (where particle physics usage is outside the main stream of computing usage), the answer to the question should be "Yes".

4. Will it be necessary to re-process the raw data, and where would such re-processing be carried out?

   Experience from previous experiments shows that it will be necessary to re-process at least part of the raw data several times. Depending on the availability of processor power and peripherals, and on the relative ease of using links or physically transporting the raw data, such re-processing might be car-

ried out either at the experiment, or elsewhere at CERN, or offsite.

2.6 Accessing the DSTs

Each LEP experiment will have about two orders of magnitude more summary data than has been the case for many earlier experiments, and of the order of 1000 tapes are likely to be needed to hold the master DST. We cannot over-emphasise the problems that will be posed by this increase in volume.

Economic factors will impose the use of a two-level storage hierarchy for handling this huge volume of summary data. The bulk of the data will have to be held on relatively cheap, slow access media, such as magnetic cartridges with automated handlers. Only the most critical summary information will then be extracted and held on fast access disks.

The discussion of the computing resources needed to access the summary data has, therefore, to be made in two parts. The first part deals with the extremely long batch jobs which run through the master DST in order to generate team DSTs, which are, in turn, processed to generate personal DSTs. The second part, dealing with the interactive workload of the individual physicist, is treated in the following section.

Because of the processor resources and long real-time needed to run the batch jobs it will be important for the experiments to optimise their working methods. If too many people try to run simultaneously through a 200 Gbyte master DST then no amount of computing resources will be sufficient!

2.7 Extracting the physics

When the information on events of interest has been reduced to a manageable size it can be manipulated by the individual physicist with a view to understanding and extracting the physics information from the events. The activities involved are many, and include data histograming, data fitting, simulation of physics processes (vital at this stage), consideration of detector and analysis program responses, event scanning, assembly of graphic and textual results to generate publications, and so forth. The sequence cannot be foreseen in advance and forms a text-book example of the need for a powerful interactive system.

Our sub-group on interactive computing has looked in some detail at the relative merits of mainframes and workstations for carrying out these tasks. We agree with their conclusion that the time has come for the LEP experiments to benefit from the much better functionality and performance of workstations, and especially from the much improved working environment that they provide for program development, program testing, and graphics, areas which are clearly crucial for this aspect of LEP computing. Workstations offer a good deal of flexibility to the LEP experiments since they can, at rather short notice, install more, and adjust the mix of simple and advanced graphics capabilities that they obtain, while remaining with a range of compatible products.

Note that we have proposed, in the previous section, that the heavy searches through the DSTs should be carried out using batch jobs (which could be generated on the workstation) running on a mainframe. This implies that sufficient communication capacity must be made available to transfer the personal active event samples between the mainframe and the workstation. We recommend that each experiment should install a cluster of workstations attached via their Local Area Network and a direct channel connection to the central "IBM" mainframe holding the DSTs. Using today's technology this approach can offer a sustained bandwidth of some 50 Kbytes/sec, which calculations show may just be adequate, although rather on the low side.

At CERN there is work in progress to exploit the software commonly used for physics analysis in an integrated fashion in the PAW (Physics Analysis Workstation) project. It will be important that PAW is well implemented for use on the primary CERN interactive systems, including the central VM/CMS service, and Apollo and VAXstation workstations. Although we believe that it is time for physicists to make the transition to workstations, it will be essential that good PAW facilities are also available for simple graphics terminals under VM/CMS and VAX/VMS.

Chapter 3

Data storage

Chapter 5 of this report sets the scale of the requirements and resources needed for the LEP experiments. Without anticipating its contents, one may situate the importance of data storage by quoting the estimate of the accumulated data storage per experiment, namely about 7000 Gbytes by the end of 1991. In other terms this represents 30 times the present capacity of the CERN IBM Mass Storage System (MSS), 50 times the total online disk storage in the CERN computer centre, or about 50 000 conventional 6250 bpi tapes *for each experiment*. The aim of this chapter is to review the choices available for this storage.

3.1 Magnetic disks

The computer industry sees data storage on disk as a major growth area. Over the past few years disk technology has been progressing at least as fast as processor technology, and there is a general feeling that we can count on significant improvements in the price per Gbyte during the next few years. The recent demonstration in the laboratory of a magnetic disk with an increase of a factor 50 in density over present commercialised disks is an indication that the technology has a long way still to go.

When evaluating the cost of disk configurations it is important to look not just at the capacity, but also at the number of access paths needed for the data in question, and hence the number of heads, controllers, and channels required. It is reasonable to assume that the various kinds of DST's discussed in the previous chapter do not have any special access path requirements, and that they can therefore be stored on the highest capacity disks. Such disks have a cost of about 30 KSF per Gbyte today, and this might drop by a factor 3 to roughly 10 KSF per Gbyte in three years time.

3.2 Magnetic tapes and cartridges

Twenty years ago reel-to-reel tape units were the main storage devices seen in any computer centre, whether commercial, industrial, or scientific. Data recording at physics experiments has been based on the use of magnetic tape for as long as many of us can remember, and the last ten years have seen considerable stability, with the use of 6250 bits per inch devices giving a very wide *de facto* transportability between data acquisition computers and mainframes all over the world.

This period of stability in the computer industry is clearly coming to an end. In commercial terms the importance of removable storage has decreased considerably

as disk capacity has increased, and the main thrust of the computer manufacturers is to use removable storage for backing-up online disk storage, rather than for data transfer. They feel, therefore, free to optimise its use for backup on their machines in any way that seems appropriate to them, including, for example, the use of data compression techniques discussed below.

IBM has introduced 3480 cartridge technology, recording at 38000 bits per inch on 18 tracks across the tape, to replace the older 3420 reel-to-reel devices, which recorded at 6250 bits per inch. This 3480 technology has proved very reliable[1] and cost-effective, and the commercial world has converted to use it to a very considerable extent. The cartridge is four times smaller in volume and weight than a conventional tape, which leads to important improvements in storage, handling, mounting and transport. The maximum data rate on writing or reading is 3 Mbytes per second to a buffered controller, compared with a maximum of 1.25 Mbytes per second, usually un-buffered, to the best of the conventional magnetic tape units.

The standard 3480 cartridge available today has an effective capacity of 200 Mbytes (cf. the 150 Mbytes of standard 3420 tape), but the potential to increase the capacity while retaining the same external cartridges and/or cartridge dimensions clearly exists. Improvements to linear or track density could be accomplished with the same physical tape drives merely by changing read/write heads. It is commonly speculated in the industry that 1 or 2 Gbytes per cartridge is technically achievable today. While it is not clear on what timescale such increases might become available, the decision being of a commercial and marketing rather than technical nature, we would be surprised if 400 Mbyte effective capacity is not available by 1990 or 1991.

The 3480 technology is available from at least five manufacturers, (Aspen, Fujitsu, Hitachi, IBM, StorageTek), and through them on a wide range of systems with other labels. IBM has stopped production and sale of conventional 3420 drives. The price of the IBM 3480 cartridge itself has dropped in two years from 46 SF to 12 SF, thus demonstrating the substantial volume being used in the commercial world.

In view of the importance of DEC computers in the high energy physics world, it is unfortunate that they have not followed this change as quickly as we would have expected.

Finally a remark on pricing is necessary. Computer centres see 3480 units as cheaper, more convenient, more reliable and more performant than 3420 units. The substantial 3420 maintenance costs are halved by 3480s. Clearly, this technology is a guaranteed and obvious way to transport data between such centres. On the other hand, it would be incorrect not to remark that, given the present controller prices, the entry price for a 3480 configuration can be a serious investment for a small installation.

---

[1] One industry measure of computer performance, the Reliability Plus data-base, demonstrates a good factor 20 for the reliability improvement over the former IBM 3420 technology.

3.3 Data compression

A number of manufacturers appear to be moving towards the introduction of hardware for data compression into disk, tape and cartridge controllers. This hardware invokes well-understood mathematical techniques to compress the data that are being written by the processor before it arrives on the disk/tape, and to regenerate the original data as they are read back. These techniques seem to offer at least a doubling in the effective data storage capacity for the future.

However, it is most important to note that, while the use of data compression techniques on disks has no obvious disadvantages, that is not true for the use of data compression on cartridge and other long-term storage media. Unless the data compression techniques are standardised across the industry, cartridges written in compressed format by one supplier will not be readable on computers produced by other suppliers. Thus, the "600 Mbyte" effective capacity offered today on 3480 drives of Hitachi origin, (e.g. BASF, Comparex, NAS etc.), is in all probability completely incompatible with the compression techniques suggested for future announcement by other manufacturers.

All is not entirely black for data compression. Some initial tests on six CERN experimental binary data samples by one of the above manufacturers showed encouraging and unexpected results, of up to a factor two in space gained.

3.4 Automated cartridge handlers

At present the active tape vault under the computer centre holds about 100K tapes, of which about 70K are in the "experimental" category owned and managed by the experiments. There are about thirty 3420 tape units serving the different mainframes in the centre (IBM, Siemens, VAXcluster and Cray). Contract staff search for tapes requested on the mainframes, and roughly one tape mount is made each minute, with an average delay between request and mount of 4-5 minutes, provided the tape is not already in use elsewhere.

We believe that it would be very unwise to continue to expand this present system indefinitely. In other industries, such as banking and telephone systems, where there is such a large volume of data that it cannot all be held on disk, there is a big push to automate the handling of this second level of data storage.

At least three companies are very active in the field of automated handling of 3480 cartridges, although they have not delivered many units to customers yet, and we are investigating the suitability of their products for the CERN situation. Another two companies have announced their intention to enter this market in the near future. All of these companies store the cartridges in racks and then use robots to retrieve the requested cartridge from its expected position, verify its label, and present it to a cartridge reader. The average time between a request and reading the first block is a few tens of seconds.

Several other HEP laboratories are actively seeking cartridge robot systems. Indeed, DESY and SLAC would appear to be further committed than CERN because their needs are more urgent. It seems inevitable in view of the numbers in Chapter 5 of this report that CERN acquires such a system during 1988, with an initial capacity of the order of 5000 cartridges, that is capable of being substantially expanded.

3.5 Optical disks

Optical disks are an enigma. For at least ten years now they have looked as though they might offer a solution to at least some of the data storage problems of particle physics, but the promise has never really been fulfilled. UA1 invested a considerable effort in the Thomson Gigadisc, and concluded that it did not match their requirements. StorageTek made substantial investments in a WORM (write once read multiple) optical disk as a major computer centre peripheral without getting beyond the beta-test phase. Many people expect DEC to introduce a reasonably priced optical disk within the next few months.

There are at least three worries about the potential use of optical disks for physics:-

- *TRANSFER RATE*

  Present optical disks have a nominal transfer rate of around 250 Kbytes/sec, which can be compared to roughly 750 Kbytes/sec for tapes on most data acquisition computers, 1.25 Mbytes/sec for tapes on a mainframe, and up to 3 Mbytes/sec for 3480 cartridges on a mainframe. The major reasons for this are said to be quality of the medium, and the absence up to now of GaAs diode lasers of sufficient power. In most systems, the writing speed is half the above reading speed if the data are checked as they are written.

- *MEDIA COST*

  2 Gbyte (12″) optical disks cost of the order of 600 SF today, which is about a factor 3 more expensive per byte than 3420 tapes, and a factor 6 more expensive per byte than 3480 cartridges. This cost is, not unnaturally, expected to fall as the level of production increases, but there are good reasons to believe that optical media will remain considerably more expensive than magnetic media for many years to come, e.g. the absence of serious volume production of any of the many types on the market that could drive the price down.

  To situate the numbers, if each LEP experiment had to purchase optical disks to hold the expected 4000 Gbytes of raw data at today's prices, and not making any allowance for volume purchase discounts, then they would be faced with a bill for about 1.2 MSF, whereas the bill using cartridges would be for just over 200 KSF.

- *IBM CONNECTIONS*

  The message from IBM is rather clear - they do not expect to attach optical disks as IBM mainframe peripherals until the technology is ready for disks that can be re-written very many times. The current belief is that they will introduce a mass storage device based on re-writeable optical disks in the early 1990s. Their reluctance to use WORM (write once read multiple) optical disks is due to the extensive software changes that are needed to use these as "systems" peripherals. As a result, it is highly unlikely that experimental data recorded on optical disk will be re-readable on IBM mainframes in the various HEP computer centres without an extraordinary interfacing effort.

One is left then with a device that is too expensive and too slow for raw data acquisition, and which cannot replace disk storage in the computer centres for the DSTs. It is good for archival storage but that is not really the problem we face. It can pro-

vide random access to large volumes of fairly constant data and this could be very useful at some stages of the analysis. Unfortunately, as seen at the moment, the overall problems of cost and speed are dominant.

On the other hand, when and if DEC do demonstrate reasonably priced optical disk storage systems, including automated handlers, then these devices may become very attractive for holding rather long-lived DSTs at DEC-based computer centres. Note, however, that in this scenario some rather tedious conversion would probably be necessary when moving DSTs between centres using magnetic cartridges and those using optical disks.

3.6 Helical recording devices (videotapes)

A number of smaller peripheral manufacturers are starting to offer devices that provide backup storage on videotape. This technology typically offers high capacity (1-2.5 Gbytes), low to medium transfer rates (100-500 Kbytes/sec), low media cost (about 10 SF/Gbyte), and reasonable pricing (10-30 KSF per unit). However, we are aware of little practical user experience in the areas of reliability, robustness, lifetime, compatibility of data written on the same or different manufacturer's drives, or software. Since the helical heads are normally in direct contact with the tape, and do not fly on an air-cushion, some of these issues are not trivial.

No major computer supplier that we know of is planning to offer this technology as a major feature of their product line, and, in particular, neither IBM nor DEC seem to be doing any work in this area. So, although these devices might appear to have some interest for physics data recording, we are worried about the overall level of engineering support that will be required to turn them into reliable and well integrated peripherals on the computers that we typically use.

3.7 Summary

Especially in view of the need for automated handling of LEP data, we conclude that 3480-style cartridge tapes should be used for the storage of LEP data. The manpower required to master and manage this technology effectively will be such that we will be unable to afford to devote any significant effort to following the development of optical disks or videotapes. We hope that it will not be necessary to review this situation until re-writeable optical storage becomes widely available.

Chapter 4

On-site links

In this chapter we look at the different types of links that will be needed by the LEP experiments on the CERN site. We should remember that the members of the LEP collaborations will find themselves working at many different locations around the CERN site, and that these locations will be separated by a distance of over 10 kilometres. They will include offices (there are likely to be a few main concentrations plus many scattered ones); the pit where the experiment is installed; laboratories, halls and test-beams where detector tests and calibrations are being carried out; and the places where the computer systems used by the experiment have been installed (again there are likely to be several).

4.1 General purpose links

It will be vital for the LEP experiments that there is a good onsite network infrastructure providing basic facilities such as electronic mail and file transfer between *all* of these locations, including the LEP pits.

The CERN Management Board has endorsed the recommendation of the CERN Technical Board on Communications that Ethernet/Cheapernet should be the Local Area Network that is installed on a general site-wide basis. Provided that the necessary money and staff are made available, and the Ethernet infrastructure is extended to the LEP pits, this will meet the needs of the experiments for general purpose links. In terms of capacity it seems clear to us that there will be a need for Ethernet bridging to the LEP pits at 8 Mbits/sec by early 1989. We would find it normal that this infrastructure should be funded by CERN.

4.2 Fast links between computers at the pit and the Meyrin site

The general purpose links described above will offer a throughput of roughly 30-50 Kbytes/sec between any two computers. The experiments believe that it will be either essential or desirable for them to be able to transfer large volumes (Gbytes) of data at high speed (up to 1 Mbyte/sec) between the data acquisition computers at the LEP pit and the computer or computers on the Meyrin site where they will carry out the subsequent processing.

These links will probably run over the time division multiplexing (G.703) infrastructure provided by the LEP machine, and they will terminate either at "private" mainframes on the Meyrin site, or at the "IBM" system in the computer centre. Commercial products are under development which aim to provide adequate performance, and which would cost some 300-400 KSF per experiment if ordered

today.

We recommend that both DD and the experiments should watch this area carefully, with a view to taking a definite decision on the approach to be adopted by mid-1988. Central funding for these fast links would encourage the use of uniform technology, rather than four different *ad hoc* solutions.

4.3 Dedicated LANs

In Section 2.7 we recommend that much of the job of extracting physics from the main team DSTs should be carried out on workstations. There is a related problem of providing access from these workstations to the personal active event samples, which would be held at the "IBM" centre.

We think that each experiment should ensure that their workstations are linked into a private cluster, with appropriate management. One or more dedicated LAN connections should then be provided between the "IBM" machines holding the DSTs and the workstation cluster. On the "IBM" side this connection would be channel-connected. There are indications that the necessary bandwidth of 50 Kbytes/sec per experiment can be reached with this approach, using the same technology as will be used for the general purpose links.

We believe that the LAN connections are infrastructure, which CERN should provide, while the experiments should be responsible for funding the workstations.

Chapter 5

Estimating the requirements for LEP computing

5.1 Setting the scale

There was much confusion in the early discussions on LEP computing caused by differing estimates of LEP luminosity in the first few months; by differing views of the total number of events that could or should be recorded and processed; by differing estimates of the time needed to process each event; and by differing views of how quickly various components of the LEP computing load could be moved offsite. However, we believe that we now have a consensus view as to the overall numerology. There are, of course, differences between the individual experiments, but the agreements are more striking than the discrepancies.

LEP will start working at some point in the second half of 1989, and a relatively small amount of data will be taken that year. During the subsequent years there will be about 3000 hours/year of physics time, split into a few main periods of several months of running. All of the experiments aim to have acquired 10^7 hadronic Z^0 events after about 2 to 2.5 years of data taking, say by end-1991.

The average size of the raw data for each Z^0 event will be in the range of 100-250 Kbytes - we use 200 Kbytes for our estimates - and the average size of each event on the master DST will be in the range of 10-120 Kbytes - we use 20 Kbytes for our estimates. The time to process one event from the raw data to the master DST stage is in the range of 20-30 sec - we use 30 sec for our estimates. The time to fully simulate an event is presently well above 300 sec and might be reduced by hard work and improved programming techniques to 60-120 sec by 1989. It is hoped to be able to generate "fast" Monte Carlo events in a time of 0.3-5 sec.

There is also agreement that each experiment will have, worldwide, 100-200 collaborators working heavily on computers during the period mid-1989 to end-1991 - we use 100 physicists in our estimates. The physicists will be working in about 20 small teams, which can be translated in our terms to mean that there will be about 20 different team DSTs (see Section 2.4.2). We expect that roughly half of these physicists and software specialists will be at CERN during this time, and the other half spread around the world. We accordingly expect 10 of the teams to operate at CERN, and 10 to operate offsite - the team DSTs will be located accordingly.

The estimates of requirements that follow are always given on a per experiment basis.

5.2 Uncertainties in the estimate

There are at least five factors that complicate all estimates of LEP computing requirements:-

- This is *experimental* physics and it is impossible to plan for the computing requirements of the LEP experiments in too much detail. All sorts of things may change between now and 1991, including ideas about data rates, detector performance, accelerator schedules, and the physics topics of crucial interest.

- The rate of accumulation of Z^0 events is likely to increase significantly between 1989 and 1991, and there will be a steady growth in the number of events accumulated (which governs the size of the master DST). It is important to always understand the period for which any estimate has been made.

  We have found it useful to develop estimates for the computing load in the "years" "1989", "1990" and "1991". These "years" (which might not coincide exactly with the calendar!) correspond to the periods when the total number of Z^0 events that each experiment has accumulated reaches 1 million, 4 million and 10 million.[2]

- The place where the various computing tasks should be carried out is always open to discussion – is a given task more suited to the CERN computer centre, to workstations, to a "private" mainframe, or to an offsite centre? As far as possible we have developed global estimates for the total worldwide load from an experiment, and only dealt with the distribution of tasks to locations in Chapter 7.

- Some of the estimates appear so high that the people making them have been tempted to understate some expansion factors which experience tells us will be required. For example:-

  — Our estimates for the number of events to be simulated using Monte Carlo techniques are very conservative.

  — Although we have made estimates for the processing power needed to derive summary data from the fraction of the simulated events that are generated in raw data format, we have neglected the power needed to access the simulated (as opposed to the real) DST events.

  — We have made no allowance for the time needed to test these major production jobs, which can be reasonably estimated at 10% of the overall load.

  As a consequence, most of us would not be surprised to see the total computer usage by the LEP experiments exceed these estimates by a factor of at least two.

- The system is not without feedback. As people start to understand some of the problems of dealing with the data they may want to take different approaches

---

[2] On the assumption that these "years" will have 3000 hours of beam time, we obtain an average data taking rate for Z^0 events during beam periods of 330 events/hour in "1989", 1000 events/hour in "1990", and 2000 events/hour in "1991".

to the physics analysis.

5.3 Monte Carlo simulation

The trouble with estimates of the Monte Carlo simulation required by each experiment is that there are good scientific arguments why this workload can expand to fill any available computing resources, up to and beyond 100 units per experiment. The real limit is economic.

We have developed a model, which is only a model, but which seems to us to be reasonable. It allows each experiment to generate, each year, twice as many Monte Carlo events as they acquire real events, for a total of 2 million in "1989", 6 million in "1990", and 12 million in "1991". Of these events, one half would be produced directly in team DST format (assumed to need 20 Kbytes/event), while the other half would be in raw data format (200 Kbytes/event), and would have to be passed through the normal analysis programs in order to generate the DSTs. This would correspond, in "1989", to a quota of 50,000 simulated raw data events and 50,000 simulated DST events for each of the 20 teams of physicists carrying out an analysis, a quota which some of us feel is far too low. We assume an average processor time of 60 seconds per event for the generation, being a balance between the 300 seconds or more that a full simulation typically takes today, and the times for "fast" Monte Carlo generators.

These assumptions lead to a requirement for 5 units of processor power in "1989", 15 units in "1990" and 30 units in "1991". The storage requirement per experiment is for 220 Gbytes in "1989", 660 Gbytes in "1990", and for 1320 Gbytes in "1991".

5.4 Acquiring the raw data

We assume that the experiments will succeed in keeping the total volume of events that they record down to twice the volume of the hadronic Z^0 events. In this case the accumulated volume of raw data will be 400 Gbytes at end-"1989", 1600 Gbytes at end-"1990", and 4000 Gbytes at end-"1991". 4000 Gbytes corresponds to 26000 full reels of 6250 bpi tape, or 20000 full cartridges holding 200 Mbytes each.

For the sake of completeness we note that we do not make any estimates of the processor power needed for data acquisition.

5.5 Compressing the raw data

If a filter/compress pass is carried out in near-real time at the experiment then it will have little effect on the computing resources needed. On the other hand, if the raw data are moved to a computer centre to run the filter/compress programs, and we assume that 5 sec/event processing time is required and that the data from one beam period is completely processed before the next one starts, then the processor requirements are still rather modest - 0.15 units in "1989", 0.5 units in "1990" and 1 unit in "1991".

5.6 Generating and maintaining the master DST

Provided that the experiments can really manage to keep the volume of the data maintained in the master DST down to 20 Kbytes/event then the overall volume of each experiment's master DST will be about 20 Gbytes at end-"1989", 80 Gbytes at end-"1990" and 200 Gbytes at end-"1991". 200 Gbytes will require 1000 cartridges for storage, if the capacity of an individual cartridge is 200 Mbytes. We should re-emphasise that many previous major CERN experiments have retained much of their raw data in the DST. That approach may not be feasible at LEP.

We have based our estimates of the processor power needed to generate the master DST from the raw data on the assumption that this will be done in near-real time, say a few hours after the data have been recorded. Since the time for physics will be limited to some 3000 hours/year this processing capacity needed to generate the DST in near-real time will be potentially available, outside beam periods, to re-run the processing roughly twice. We expect that such re-processing will be necessary in "1989" and "1990". By "1991" there is a chance that the detectors will be better understood, and that the need for re-processing will diminish. If that is the case, then the spare capacity will be available to help with the simulation load, which will be growing strongly at that stage.

Using the 30 sec processor time estimate from Section 5.1 and the average data taking rates from Section 5.2, it can be seen that the requirement for processor power is 3 units in "1989", 8 units in "1990", and 16 units in "1991".

5.7 Accessing the DSTs

Very long batch jobs will be needed to select information or event subsets from the master DST, and to finalise physics analyses. We have prepared a model of the overall resources required to select the subset DSTs and to finalise the physics analyses as the volume of the master DST grows during the "years" "1989" to "1991". We would like to re-emphasise that this is only a model which we have used to obtain an estimate of the computing resources needed for this task. In view of the importance of efficient DST access for high-statistics physics we would certainly expect the LEP experiments to devote considerable attention to organising this access. We believe, however, that the experiments are very unlikely to need less resources for this stage of the processing than we have predicted from our model.

Tests show that making a first simple selection from among the events in a master DST requires about 30 minutes[3] of processor time per Gbyte, corresponding to a processing time of about 35 msec to deal with each event. The rather more complex criteria needed for subsequent selections typically require 60-90 minutes of processor time per Gbyte of DST. We have used 1 hour per Gbyte for all of our estimates.

- *"1989"*

  By the end of "1989" the volume of the master DST will be about 20 Gbytes. Each of the 20 teams carrying out an analysis will have a team DST of 2 Gbytes, and the experiment can make two runs through the master DST each

---

[3] The tests read ZEBRA data structures and use HBOOK to fill a few histograms.

week. This means that, on average, each team can expect a run through the master DST in order to update the team DST or to make a final analysis once every 10 weeks. This operation will require 40 hours/week of processing time. Each of the 100 physicists working on the analysis can expect to regenerate his or her personal DST, assumed to be 100 Mbytes, once per day (or rather, 5 times per week) from the team DST. This will require 1000 hours/week of processing time. Finally each of the 100 physicists will be able to prepare, twice per day, a 5 Mbytes active event sample from his or her 100 Mbytes personal DST, and to load it into a workstation for physics analysis using PAW. This will require 100 hours/week of processing time.

This produces a total processor load for each experiment of 1140 hours/week, say 8 units, in "1989". The total data storage capacity is 70 Gbytes (20 for the master DST, 40 for the team DSTs, and 10 for the personal DSTs) without allowing for any duplicates. If all of the data are held on cartridges then the cartridge mounting load per experiment will average about 40/hour. If roughly 100 Gbytes of disk space is available to hold the team DSTs and personal DSTs (allowing a factor 2 for duplicates), so that only the master DST remains on cartridges, then the cartridge mounting load will be negligible (under 2/hour).

The data transfer between the personal DSTs and workstations averages some 25-30 Kbytes/sec for each experiment, assuming that each physicist is allowed to reload his or her 5 Mbytes active event sample twice daily during the period from 0800 to 2000.

- *"1990"*

By the end of "1990" the volume of the master DST will be about 80 Gbytes. Each of the 20 teams carrying out an analysis will have a full team DST of 8 Gbytes and a condensed team DST of 2 Gbytes. Each experiment can make two runs through the master DST each week. This operation will require 160 hours/week of processing time. Each team can prepare its condensed team DST from the full team DST once per week. This operation will also require 160 hours/week of processing time. The assumptions concerning the preparation of the personal DSTs and the active event sample are as in "1989", and require 1100 hours/week of processing time.

This produces a total processor load for each experiment of 1420 hours/week, say 9 units, in "1990". The total data storage capacity is 290 Gbytes (80 for the master DST, 200 for the team DSTs, and 10 for the personal DSTs) without allowing for any duplicates. If all of the data are held on cartridges then the cartridge mounting load per experiment will average about 50/hour. If roughly 100 Gbytes of disk space is available to hold the compressed team DSTs and personal DSTs (allowing a factor 2 for duplicates), so that only the master DST and full team DSTs remain on cartridges, then the cartridge mounting load will drop to 10/hour.

- *"1991"*

By the end of "1991" the volume of the master DST will be about 200 Gbytes. Each of the 20 teams carrying out an analysis will have a full team DST of 20 Gbytes and a condensed team DST of 2 Gbytes. Each experiment can make two runs through the full DST each week. This operation will require 400

hours/week of processing time. Each team can prepare its condensed team DST from the full team DST once per week. This operation will also require 400 hours/week of processing time. The assumptions concerning the preparation of the personal DSTs and the active event sample are as in "1989" and "1990", and require 1100 hours/week of processing time.

This produces a total processor load for each experiment of 1900 hours/week, say 13 units, in "1991". The total data storage capacity is about 650 Gbytes (200 for the master DST, 440 for the team DSTs, and 10 for the personal DSTs) without allowing for any duplicates. If all of the data have to be held on cartridges then the cartridge mounting load per experiment will average about 65/hour. If roughly 100 Gbytes of disk space is available to hold the compressed team DSTs and personal DSTs (allowing a factor 2 for duplicates), leaving only the master DST and full team DSTs on cartridges, then the cartridge mounting load will drop to 30/hour.

These DST data must be managed automatically, and it would be highly desirable for all of the subset DSTs to be held on direct access storage (disk). For comparison the CERN computer centre is currently equipped with a total of about 150 Gbytes of disk space, and the total tape mounting presently averages some 60/hour.

We would also like to point out that the elapsed time to run the job or jobs that process the full master DST will be very long, many days or even several weeks, depending on the power of the processor that is being used.

5.8 Extracting the physics

The load during a 12-hour working day (0800 to 2000) can be very conservatively estimated to 1 hour of processor time per physicist (say 0.36 sec for each of 10000 events), or 8 units to support 100 physicists. This matches reasonably with opinions on how many users can be supported on an interactive system - such as very few (say 1,2 or 3) "serious" (CPU-intensive) users on a 0.25 unit VAX-11/780 or the present 400-plus general-purpose users on the 12-unit central VM/CMS service.

We have made it clear that we recommend that workstations should be used for this part of the work. If the number of workstations per experiment reaches 50 by "1991", which seems perfectly reasonable to us, then the processor power that they represent is likely to total at least 20-25 units. Experiments should seriously evaluate the use of this capacity during non-peak periods for activities such as Monte Carlo simulation.

5.9 Summary

We repeat that we are making estimates per experiment, on a global (worldwide) basis, for the "years" "1989", "1990" and "1991" when the number of Z^0 events accumulated by each experiment reach 1 million, 4 million and 10 million respectively.

5.9.1 Processor power

The following table summarises the estimates for the different types of processor power that will be needed by each experiment. The attentive reader will notice the question mark attached to the Total? line. It is, indeed, rather unclear (and perhaps not even very meaningful to ask) how the different types of processor power should be added in order to come to a total. This is because there are many ways in which this power can be provided; some options permit sharing the power between different tasks, while for other options such sharing is effectively excluded. The major sharing that we believe might be effective would be to carry out some of the simulation work either using the interactive capacity (workstations etc.) during the middle of the night, or using the capacity needed to generate the master DST outside beam periods.

We discuss in following Chapters how much of the power should be provided on the different systems (onsite and offsite) that are under consideration.

"Year"	"1989"	"1990"	"1991"
Monte Carlo generation	4	12	24
Processing of MC events	1	3	6
Generating the master DST	3	8	16
Accessing the DSTs	8	9	13
Extracting the physics	8	8	8
Total?	24	40	67

Processor Power per Experiment (CERN units)

Table 1

5.9.2 Data storage

The following table summarises the estimate for the total volume of data, in Gbytes, that will have to be stored by each experiment. Note that the expansion needed for making duplicate copies is left to the appreciation of the reader. By the end of 1991, if the data are held on cartridges of 200 Mbyte capacity and no cartridge can be recycled, then each experiment will need a total of at least 35000 cartridges to hold these data.

"Year"	"1989"	"1990"	"1991"
Raw data	400	1600	4000
Master DST (real events)	20	80	200
Team DSTs	40	200	440
Personal DSTs	10	10	10
Simulated raw data	200	800	2000
Simulated DST events	20	80	200
Duplicate copies	?	?	?
Total	690	2770	6850

Accumulated Data Volume per Experiment (Gbytes)

Table 2

5.9.3 Data manipulation

The following table gives the number of cartridges that each experiment will have to mount per hour in order to access the DSTs. The data are assumed to be stored on 200 Mbyte cartridges. One entry deals with the (hopefully hypothetical) case that no disk space is available, while the other assumes that each experiment has 100 Gbyte of disk space available to hold the compressed team DSTs and personal DSTs. 20 cartridge mounts per hour corresponds roughly to an average data rate of 1 Mbyte/sec.

"Year"	"1989"	"1990"	"1991"
With no disk space	40	50	65
With 100 Gbytes disk space	2	10	30

Cartridge Mounts per Hour per Experiment

Table 3

- 10 Gbytes of disk space by "1989" and 20 Gbytes of disk space by "1991";
- Facilities to access some hundreds of 3480-style cartridges with up to 10 mounts/hour.

Such centres should also provide:-

- A link of 64 Kbits/sec to CERN by "1989", upgraded to 2 Mbits/sec by "1991";
- An operating system and general software environment as identical as possible to that found at CERN;
- The normal range of professional support.

It will be worthwhile for the LEP experiments to arrange to carry out general-purpose processing at such centres, since the resources available will be in balance with the resources required, and it will be possible to use them efficiently.

From past experience we would certainly expect some centres in France, Italy and the U.K. to meet these criteria. These centres should certainly hold several team DSTs (which will each occupy 20 Gbytes by "1991") for subsequent physics analysis, and it would be desirable for them to be able to hold copies of the master DSTs (each 200 Gbytes by "1991"), and to take a full share in the whole range of LEP computing.

There are a number of other centres, some of them university centres and some more closely attached to particle physics, which are potentially quite powerful. We recommend that they should be upgraded to be able to hold team DSTs that have been generated at CERN or at another major regional centre.

Again we suggest that the HEP-CCC must review the situation, with a view to ensuring that each major country, and as many of the smaller ones as possible, offers the facilities of a powerful computer centre for general-purpose LEP computing.

6.2.2 Individual institutes

Many individual physics institutes will have limited resources of staff and money, and will not want to run a powerful general-purpose centre for LEP computing, of the style outlined above. These institutes will nevertheless play a vital, if more specialised, rôle in LEP computing.

We imagine that most, if not all, will want to install sufficient workstations and data storage facilities, so that they can take a full share in the physics analysis. They should upgrade their communications links with CERN (via their nearest regional centre) to at least 64 Kbits/sec.

In addition we recommend that some institutes should specialise in the vital job of simulating events, physics processes, and detectors. This is an area well adapted to specialised computer architectures, such as vector-processors, emulators, transputers, or other cost-effective "Fortran farms". Once again, it will be vital to ensure that there is adequate communication between these facilities and the CERN-based component of the experiment.

6.3 Location of the DSTs

It will be a challenging job for each experiment to manage hundreds of Gbytes of DSTs. By the end of "1991" they can expect to have to deal with at least 200 Gbytes of master DST, a similar amount of data simulated at DST level, and about 450 Gbytes of team and personal DSTs. Economics probably exclude that all of these data can be held on disk - at today's prices 850 Gbytes would cost at least 25 MSF and by 1991, when the full volume would be needed, the price will probably still be in the region of 10 MSF. If 200 Mbyte cartridges are used, instead of disks, to store the master and full team DSTs, and 100 Gbytes of disk space are available, then the likely cartridge mounting requirement will reach about 30 cartridges/hour for each experiment. We believe that these facts imply that the master DST can only reasonably be held on a computer that is equipped both with a significant amount of disk space (in the range of 50-100 Gbytes) and with automated cartridge handling equipment.

In addition the book-keeping involved in the management of the ten million events in the master DST will be a problem in itself, however it is carried out. One technique would involve maintaining a few items of key information ("tags") about each event in a database. Provided that such a database is held on the same machine as the one processing events for the master DST, then the tags can be entered and updated automatically by the processing programs. Any such automatic database update will be more complicated to achieve when the processing is carried out remotely.

In conclusion we would recommend that the master version of each experiment's DST should be held on the "IBM" systems in the CERN computer centre, and that full copies should only be maintained on a very few other centres. It should be possible to maintain team DSTs on a wider range of other adequately configured computers.

7.7 Summary of resources needed onsite at CERN

We recall that we have assumed that 100 Gbytes of disk space are available to each experiment in order that the cartridge mounting rate can be held to a reasonable level. In view of the fact that copies of all DSTs should be held at CERN, we think that each experiment will need at least 50 Gbytes of disk space at CERN in "1989", rising to 100 Gbytes by "1991".

The following table summarises the processor resources that each experiment needs to identify at CERN, either at the computer centre, or at the experiment, or on "private" mainframes, or on workstations. We have recommended that the task of accessing the DSTs should be carried out on the central "IBM" systems, and that the task of extracting the physics should be performed on workstations. Especially if efforts to vectorise GEANT3 prove to be successful, much of the simulation load of Monte Carlo generation and processing would fit well on the Cray X-MP/48. We have discussed how the different experiments aim to generate their master DST in Section 7.4 above.

"Year"	"1989"	"1990"	"1991"
Monte Carlo generation and processing	2.5	7.5	15
Generating the master DST	3	8	16
Accessing the DSTs	4	5	8
Extracting the physics (workstations)	4	4	4
Total	13.5	24.5	43

"Private" and "Public" Processor Power per Experiment at CERN
(CERN units)

Table 4

7.8 Summary of resources needed outside CERN

The following table summarises the processor power that will have to be provided outside CERN for each experiment. It can be seen that this represents a substantial fraction of the total LEP computing load. We repeat that provision of generous network capacity will be necessary to exploit this power.

"Year"	"1989"	"1990"	"1991"
Monte Carlo generation and processing	2.5	7.5	15
Accessing the DSTs	4	4	5
Extracting the physics (workstations)	4	4	4
Total	10.5	15.5	24

Processor Power per Experiment offsite
(CERN units)

Table 5

Chapter 8

The impact of "private" mainframes

Now that we understand the magnitude of the LEP computing problem, especially the data manipulation aspects, many of us are concerned and even perhaps apprehensive about having to cope with the additional complexity of "private" mainframes on the CERN site. In terms of optimising the use of expert staff it would have been better to have adopted a fully centralised solution, such as the one planned for HERA, but that is not the situation that we are now called on to deal with.

There will be at least one, and probably more, "private" mainframes close to the CERN computer centre. In many ways these mainframes, owned and operated by the experiments or institutes, are rather similar to offsite computer centres. However, the analogy cannot be pushed too far, and isolating these machines from the CERN computer centre as a matter of policy would result in an extremely inefficient use of expensive resources.

The general CERN networking facilities, to which we can expect all machines to be attached, will offer higher bandwidth than international connections, and it is much easier to move tape cartridges between the centre and the "Barn" than it is to send them to and from Abingdon, Bologna, Munich or even nearby Lyon. The technical reality is that these computers will have to be integrated at some level with the CERN computer centre, and that this will require considerable efforts in order to provide compatibility of operating systems, networking, and automated cartridge handling.

We feel that it is becoming urgent to have clearer political guidelines on the use of these machines. The processor power that they represent is clearly required in order to handle the LEP data. When the political guidelines have been given we can clarify which components of the LEP computing workload will be carried out "privately" and which "publicly", and how the configurations should evolve.

If the "private" mainframes are adequately configured, and they are equipped with adequate methods for accessing data at the CERN computer centre, and their workload is selected carefully as a function of their configuration, then they will be able to take a useful part in the LEP data analysis.

Chapter 9

Resources

The continual progress in computing technology does not look likely to be checked in the next five years - indeed there are signs that in some areas of interest it might even accelerate. If a 20% per year improvement in price-performance can be sustained in some area, then the price for equivalent performance will be reduced by a factor 3 between 1988 and 1992. It is also perhaps possible that new products will be introduced in the area of data storage which could allow us to completely review our approach in a few years time.

We must, however, now take decisions for LEP startup on the basis of products that are either already installed or very close to first delivery, and to plan the resources needed as a consequence.

9.1 Manpower

The informatics staff responsible for running the existing services find themselves under considerable pressure, and are overworking to the point where their personal efficiency is impaired. This applies both to the CERN staff, and to the specialists working in the experiments.

Furthermore, as far as we can tell, the experiments have not yet been able to identify the staff, such as systems managers, who will be needed to organise the operation of their computing activities. Insufficient CERN staff are available for this work, and the experiments prefer to use their limited budgets to help physicists to come to CERN, rather than use it for these essential "operations-style" specialists.

In addition, this report has identified several areas where existing services need to be expanded, sometimes substantially, and where new services must be started.

It is now very urgent for the CERN Directorate and the experiment spokesmen to decide how this serious situation should be tackled.

9.2 Money

Tables 6 and 7 show our estimates of the likely expenditure that will be required from CERN and the experiments in the period "1989" to "1991". As far as we are aware none of the CERN items is explicitly budgeted for. The network infrastructure figures are in addition to those for the general-purpose infrastructure proposed by the CERN Technical Board on Communications. Although some money (about 4 MSF/year) is foreseen in the long term CERN budgets for computer centre invest-

ment, this has been almost entirely committed for the years in question in order to pay for already authorised extensions of the "IBM" capacity.

"Year"	"1989"	"1990"	"1991"
Cartridge handling	2	2	2
Disk storage	4	2	2
Extra processor power	-	6	10
Network infrastructure (links to pits and workstations, etc.)	2	1	-
Total	8	11	14

CERN computer centre - financial requirements (MSF)

Table 6

As a cross-check we list a set of items which we believe the experiments will need to budget for. In some cases this may have been done already.

"Year"	"1989"	"1990"	"1991"
Cartridge purchase	40	120	220
Workstations at CERN	250	250	250
Workstations offsite	250	250	250
"Private" system to generate DST	?	?	?
Other "private" systems	?	?	?
Total	?	?	?

Experiments - financial requirements (KSF)

Table 7

Chapter 10

Recommendations

10.1 Scale of the problem

Many of us, the so-called experts, are only now starting to understand the enormity of the problem of LEP computing. The essential point is that this is high-statistics, high-complexity physics, with four experiments competing under tight time pressures. While we clearly must communicate our understanding of the scale of the problem to the management of the experiments and laboratories concerned, no one should be overly alarmed, since it will be possible to achieve good results with adequate but not excessive resources of manpower and money.

10.2 LEP computing outside CERN

Physicists should not find themselves in a situation where they have to come to CERN to obtain adequate computing resources. Two things need to be done:-

1. *COMPUTING FACILITIES*

   (See Section 6.2) The HEP-CCC should lead a vigorous program aimed at ensuring that the computing facilities outside CERN available to LEP experiments are upgraded. We recommend that:-

   - Most, if not all, of the collaborating institutes should be able to carry out physics analysis based on the use of groups of workstations attached to a computer which holds some set of DSTs.

   - Some institutes should specialise in the simulation techniques, which will form an essential component of the exploitation of LEP data.

   - A few regional centres should be upgraded so that they are able to deal with a full copy of the master DST, or even with a significant fraction of the raw data.

2. *NETWORK FACILITIES*

   (See Section 6.1) It will be vital that the HEP-CCC leads a similar drive to upgrade the networking facilities between CERN and the home institutes. This will require the installation of at least 2 Mbits/sec connections to the main regional centres, which should themselves have 64 Kbits/sec connections to the individual institutes. Reinforced coordination and management of physics networking will be essential if the improved external computing facilities are to be fully exploited.

10.3 "Private" computers

(See Chapter 8) Directorate guidelines concerning the use of "private" computers at CERN are needed rather urgently. These guidelines should deal with the conditions under which these machines can be brought to CERN, the ways in which they can be connected to the CERN computer centre and networks. We also need to understand what impact the installation of "private" capacity will have on the share of the "public" resources that the owners are likely to obtain, and, more generally, how the shares of the "public" resources will be distributed across the different services (Cray, "IBM" and VAXcluster).

10.4 The size of the master DST

(See Section 5.6) It is vital that all experiments arrange that they use the raw data as infrequently as possible, and that the master DST is kept as small as possible.

10.5 Location of the master DST

(See Section 6.3) A complete copy of the master DST for each experiment should be held at the CERN computer centre.

10.6 The need for cartridges

(See Chapter 3) CERN should concentrate on the use of 3480-style magnetic cartridges for LEP data storage.

10.7 Storage hierarchy

(See Sections 3.4, 5.9.2 and 5.9.3) CERN should take urgent steps to make sure that an adequate hierarchy of mass storage systems, consisting of online disks and automated cartridge handlers, is available to all of the mainframes (Cray, "IBM" and VAXcluster) in the computer centre. The target for disk space should be a total of 200 Gbytes for the LEP experiments by end-1989 and an extra 200 Gbytes by end-1991. The target for cartridge mounting capacity should be 100 mounts/hour by end-1989 and 200 mounts/hour by end-1991.

10.8 Links to the experiments

(See Section 4.1) The general networking infrastructure (Ethernet and future upgrades) that is available on the CERN sites should be extended, at CERN's expense, to the LEP experimental areas.

(See Section 4.2) CERN and the experiments should collaborate to make high-speed links available between the data acquisition systems at the experiment and the main-

frames in the computer centre, or other private mainframes.

10.9 Interactive computing

(See Sections 2.7 and 4.3) Interactive physics analysis at LEP should be based as much as possible on the use of PAW software running on Apollo or VAXstation hardware. The experiments should plan to install a few clusters of their preferred workstation family, and CERN should install a high performance connection between these clusters and the "IBM" mainframes.

10.10 Processor power

(See Section 7.7) By LEP startup CERN should plan to provide at least 30 CERN units to be shared among the LEP experiments. An extra 50 units should then be provided within a further period of two years.

10.11 Finishing the job properly

The total cost of LEP and the experiments will be well over 10^9 Swiss francs, and the sustained efforts of thousands of people will soon generate large volumes of data from Z^0 collisions. We must make sure that we invest enough money and manpower in the computing facilities that will be needed in order to extract the physics contained in these events.

Appendix A

Comparison with other experiments

In order to develop confidence in the validity of our estimates we have tried to compare them with the computing resources used or planned to be used by a few other experiments. The fact that the computing for a modern experiment is very complicated and that it is carried out across the world on a variety of machines means that it is often difficult to obtain accurate data. We thank the people who provided us with the background material for the following paragraphs (A. Norton and F. Bernasconi for UA1, A. Parker for UA2, and H. Hoffmann for the HERA estimates). They and we would like it to be understood that the numbers given are only rough estimates.

A.1 Collider

When comparing computing at LEP and the $p\bar{p}$ collider it is important to understand that much collider physics to date has been a question of identifying and then studying the properties of rather rare and special events.

A.1.1 UA1

Until the Autumn 1987 run UA1 recorded all of their data on 6250 bpi magnetic tapes. Their events generate slightly more than 100 Kbytes of raw data, with 1000-1200 events typically fitting on a full tape. By the end of 1986 they had recorded some 12000 tapes (1800 Gbytes) of raw data, or 12-15 million events, and used a further 31000 tapes for subsequent processing, including generating and accessing the DSTs, extracting the physics, and all of the Monte Carlo simulation. We note in passing that UA1 themselves devoted considerable effort to managing the movement of these tapes, many of which were stored outside the CERN computer centre. Of these 31000 tapes some 9000 had been recycled, which means that their total tape consumption had been 34000 (12000 plus 22000). The total number of tapes written had been 43000 (12000 plus 31000), for a total data volume of over 6000 Gbytes. For the LEP experiments we are predicting a total data volume of about 7000 Gbytes by the end of "1991".

The first stage of "normal" processing of UA1 data was to run a program called BigMac at the CERN computer centre. This applied the calibration corrections to the raw data, and then performed a partial reconstruction in order to be able to filter the data into about four streams (typically jets, muons, W and Z candidates, and missing energy candidates). Typically only some 15-20% of all events were selected by BigMac, but, at the output stage, each one was already twice the size of the raw

data. BigMac needed about 4 Mbytes of main memory, and was normally run within about 2 weeks of the data being recorded at LSS5.

The subsequent processing of these "normal" events was carried out using the Bingo reconstruction program at outside laboratories, including RAL, Saclay and Harvard, as well as at CERN. It is estimated that Bingo processing, which required about 10-20 sec of processor time per event, was carried out twice for each event, on average. The UA1 approach was to retain the raw data for all events for a long time, certainly up to and including the stage that we have called the team DST in this report (though their overall approach to DST processing does not match exactly with the one we have used as our model). When the physicists were sure that the event had been well measured, then they proceeded to create summary tapes in their private format, without the raw data. This corresponds to the personal DST stage in our model. As a result the typical size of an event on what we have referred to as the master or team DSTs was about 300 Kbytes. This is in marked contrast to the 20 Kbytes of summary data that we have used in our estimates.

UA1 also used an "express line" scheme for fast processing. During the data acquisition all events were processed by 3081E emulators and up to 10% were signalled as being of special interest. These events were processed, typically within one day of being recorded, by a special version of the Bingo reconstruction program which refined the selection of "interesting" events, so that these could then be passed to the Megatek event viewing system. The maximum speed at which events could be scanned and processed by physicists on the Megatek was some 5 per hour, or 100 per day, which therefore defined the number of events that it was desirable to output from the express line system.

At CERN UA1 have used roughly 3-6 units at the computer centre during the past few years, at a time when the central batch capacity varied in the range 10 units (1981), through 33 units (1986) to 25 units (1987). In 1986 they caused some 65000 tapes to be mounted at the CERN computer centre, of which roughly 25000 during the last three months, corresponding to a rate of about 12 per hour during that busy period.

Outside CERN they have access to computer centres in several countries, and they have also been a driving force in developing and exploiting emulator "farms" of various flavours, but mainly 3081Es. At least in recent years they have used as much processor power outside CERN as onsite.

They emphasise the importance of good facilities for Monte Carlo simulation, and estimate that in 1986 they used more processor time for simulation than for processing real data. They would agree that Monte Carlo work needs considerable statistics, and that it generates a volume of data that is comparable to that coming from the real events.

For the present run UA1 have made radical changes to their data acquisition system, and they now record data on 3480 cartridges via an IBM 9370 computer. They are very happy with the performance and reliability of this technique.

A.1.2 UA2

UA2, until the recent upgrade, had a much simpler inner detector than UA1, and they typically produced 10-20 Kbytes of raw data. In addition there was no central magnetic field. As a consequence, the computing requirements were much simpler.

As far as we can tell the normal use of processor and I/O resources by UA2 scales, with respect to the UA1 values, roughly with the size of the raw data. By the end of 1986 UA2 had generated some 2800 tapes of raw data and had used a total of 9000 tapes overall. Processor time at CERN has been roughly 1-2 units per year, and they were the source of some 11000 tape mounts during 1986. UA2 also made strong use of outside computer centres, such as Heidelberg and Saclay.

It should be noted that, following the recent upgrade of the UA2 detector, which included the installation of scintillating fibres and flash ADC's in an inner drift chamber and in a transition radiation detector, the UA2 event size has increased to about 100 Kbytes per event.

Although there is no possibility, at present, for UA2 to record data on 3480-style cartridges, they do convert their raw data tapes as soon as they have been recorded, in order to benefit from the cheaper media costs and higher performance that this technology offers in computer centres.

A.2 HERA

Both of the HERA experiments, H1 and ZEUS, currently estimate their overall computing load at 36 units each, of which DESY plans to provide about 50%. See, for example, the minutes of the Hera Experiments Data Acquisition Committee HEDA, (HEDA 03-87).

DESY's plans, as expressed to the HEP-CCC, also foresee a move to 3480 cartridges and automated handling devices, probably the StorageTek jukebox.

Index