CERN / LHCC / 2001-004

CERN / RRB-D 2001-3

Original: English

22 February 2001

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE

CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

REPORT OF THE STEERING GROUP*

OF THE LHC COMPUTING REVIEW

---

* S. Bethke (Chair), M. Calvetti, H.F. Hoffmann, D. Jacobs, M. Kasemann, D. Linglin

Acknowledgement

The Steering Group wishes to acknowledge the outstanding effort and contributions of all those who participated in the review, especially those in attendance at the Steering Group meetings. Sincere thanks are also due to all the members of the technical panels and in particular to the external experts.

The present report is the result of a large collaborative effort. Its contents are the responsibility of the Steering Group and represent a broad consensus of all parties involved.

MEMBERSHIP OF THE REVIEW

Steering Group

Members:	S. Bethke (MPI Munich)	Chair	
	H.F. Hoffmann (CERN)	CERN Director for Sc. Computing	
	D. Jacobs (CERN)	Secretary	
	M. Calvetti (INFN Florence)	Chair of the Mgmt and Resources Panel	
	M. Kasemann (FNAL)	Chair of the Software Project Panel	
	D. Linglin (CC-IN2P3/CNRS)	Chair of the Computing Model Panel	
In Attendance:		**Representative**	**Alternate**
	IT Division	M. Delfino (CERN)	L. Robertson (CERN)
	ALICE	F. Carminati (CERN)	K. Safarik (CERN)
	ATLAS	N. McCubbin (RAL)	G. Poulard (CERN)
	CMS	M. Pimia (CERN)	H. Newman (CALTECH)
	LHCb	J. Harvey (CERN)	M. Cattaneo (CERN)
Observers:	R. Cashmore (CERN)	CERN Director for collider programmes	
	J. Engelen (NIKHEF)	LHCC chairman	

Worldwide Analysis / Computing Model panel

	D. Linglin (CC-IN2P3/CNRS)	Chair	
	F. Gagliardi (CERN)	Secretary	
Expt. Reps:		**Representative**	**Alternate**
	ALICE	A. Masoni (INFN Rome)	A. Sandoval (GSI Darmstadt)
	ATLAS	A. Putzer (U. Heidelberg)	L. Perini (U. Milan)
	CMS	H. Newman (CALTECH)	W. Jank (CERN)
	LHCb	F. Harris (U. Oxford)	M. Schmelling (MPI Heidelberg)
Experts:	Y. Morita (KEK)	C. Michau (UREC-STIC/CNRS)	

Software Project panel

	M. Kasemann (FNAL)	Chair	
	A. Pfeiffer (CERN)	Secretary and CERN-IT representative	
Expt. Reps:		**Representative**	**Alternate**
	ALICE	R. Brun (CERN)	A. Morsch (CERN)
	ATLAS	D. Barberis (U. Genoa)	M. Bosman (U.A. Barcelona)
	CMS	L. Taylor (Northeastern U.)	T. Todorov (IN2P3 Strasbourg)
	LHCb	P. Mato (CERN)	O. Callot (LAL Orsay)
Experts:	V. White (FNAL)		

Management and Resources panel

	M. Calvetti (INFN Florence)	Chair	
	M. Lamanna (INFN Trieste and CERN)	Secretary	
Expt. Reps:		**Representative**	**Alternate**
	ALICE	P. Vande Vyvre (CERN)	K. Safarik (CERN)
	ATLAS	J. Huth (U. Harvard)	H. Meinhard (CERN)
	CMS	P. Capiluppi (INFN Bologna)	I. Willers (CERN)
	LHCb	J. Harvey (CERN)	J.P. Dufey (CERN)
Experts:	F. Étienne (IN2P3 Marseilles)	J. Gordon (RAL)	L. Robertson (CERN)
	F. Ruggieri (INFN Bari)	T. Wenaus (BNL)	K. Woller (DESY)
	G. Wormser (IN2P3 Paris)		

CONTENTS

1 EXECUTIVE SUMMARY

The requirements to ensure the storage, management, simulation, reconstruction, distribution and analysis of the data of the four LHC experiments (ALICE, ATLAS, CMS and LHCb) constitute an unprecedented challenge to the High Energy Physics (HEP) and Information Technology (IT) communities. Fulfilment of these requirements will, on the same basis as the successful construction of the LHC accelerator and particle detectors, be crucial for the success of the LHC physics programme.

The LHC Computing Review evaluated the current situation, plans and prospects of data management and computing at the LHC. Based on the detailed work of three independent panels, the *Software Project Panel*, the *Worldwide Analysis and Computing Model Panel* and the *Management and Resources Panel*, this report reaches the following main conclusions and makes associated recommendations:

1.1 The LHC Computing Model:

1) After critical assessment, the review accepts the scale of the resource requirements as submitted by the four experiments and as summarised in Table A2.1 on page 61. The details of this table are explained in Chapter 5.1.2 below.

2) A *multi-Tier hierarchical model* similar to that developed by the MONARC project[1] should be the key element of the LHC computing model. In this model, for each experiment, raw data storage and reconstruction will be carried out at a Tier0 centre. Analysis, data storage, some reconstruction, Monte-Carlo data generation and data distribution will mainly be the task of several (national or supra-national) "Regional" Tier1 centres, followed by a number of (national or infra-national) Tier2 centres, by (institutional) Tier3 centres or workgroup servers, and by end-user workstations (Tier4). The CERN-based Tier0+Tier1 hardware for all LHC experiments should be installed *as a single partitionable facility*.

---

[1] MONARC Phase 2 report CERN / LCB 2000-001, March 2000 http://monarc.web.cern.ch/MONARC/docs/phase2report/Phase2Report.pdf

3) *Grid Technology* [2] will be used to attempt to contribute solutions to this model that provide a combination of efficient resource utilisation and rapid turnaround time.

4) Estimates of the required *bandwidth* of the wide area network between Tier0 and the Tier1 centres arrive at 1.5 to 3 Gbps for a single experiment. The traffic between other pairs of nodes in the distributed systems will be comparable. While technology will certainly be able to deliver such rates, it is vital that a well-supported Research Networking infrastructure with sufficient bandwidth is available, *at affordable costs*, by the year 2006. Continued close monitoring of the trends in the ways that providers in this dynamic market supply bandwidth will contribute much to such an aim.

1.2 Software:

5) *Joint efforts* and *common projects* between the experiments and CERN / IT are recommended to minimise costs and risks. Because, however, the experiments are very different from one another, it cannot be hoped that a single set of methods and tools will meet all needs. Pending the final choices being made, *support* of existing widely used products (cf. Table A1.2) should be provided (even if the products are not used by all the experiments).

6) *Data Challenges* of increasing size and complexity must be performed as planned by all the experiments until LHC start-up.

7) CERN should sponsor a coherent programme to ease the transition of the bulk of the physics community from Fortran to Object Oriented (OO) programming.

8) Further identified areas of concern are the limited maturity of current planning and resource estimates, the development and support of simulation packages and the support and future evolution of analysis tools.

1.3 Management and Resources:

9) Current cost estimates are based on the forecast evolution of price and performance of computer hardware (cf. Appendix 3.5.2).

---

[2] See e.g. The GRID, Blueprint for a New Computing Infrastructure, I. Foster and C. Kesselmann ed., ISBN 1-55860-475-8

10) On this basis, the *hardware costs* for the initial set-up of the LHC distributed computing centres, including Tier0 to Tier2 structures for all experiments, are currently estimated to be *240 MCHF*. The CERN-based common Tier0+Tier1 centre is estimated to make up about 1/3 of the overall computing capacity.

11) The total investment for the initial system is due to be spent – in approximately equal portions – in the years 2005, 2006 and 2007, assuming LHC starts up in 2006 and reaches design luminosity in 2007.

12) The *Core Software teams* of all four experiments are currently *seriously understaffed* (see Chapter 6.5.1 below). Their contribution will be just as vital to the experiments as major sub-detectors and the senior Management of the Collaborations must seek solutions to this problem with extreme urgency.

13) The *staffing level of CERN/IT* as envisaged under the current CERN-wide staff reduction plan is incompatible with an efficient running of the CERN-based LHC computing system and software support (cf. Chapters 6.5.2-6.5.3 below).

14) The approach to *Maintenance and Operation* of the LHC computing system includes the strategy of rolling replacement within a constant budget. Maintenance and Operation will require within each three-year operating period an amount roughly equal to the initial investment.

15) The construction of a *common prototype* of the distributed computing system should be launched urgently as a joint project of the four experiments and CERN / IT, along with the major Regional Centres. It should grow progressively in complexity and scale to reach ~50% of the overall computing and data handling structure of one LHC experiment in time to influence the acquisitions of the full-scale systems.

16) An agreement should be reached amongst the partners in this project, in which the construction, the cost sharing, the goals and the technical solutions of the common prototype are laid down.

1.4 General recommendations:

17) An *LHC Software and Computing Steering Committee* (*SC2*), composed of highest level software and computing management in experiments, CERN / IT and Regional Tier1 Centres, must be established to oversee the deployment of the entire LHC hierarchical computing structure.

18) Under the auspices of this committee, *Technical Assessment Groups* (TAGs) must be established to prepare, for example, agreement on a common data management strategy, to permit a choice of common data management system components, and to initiate and monitor projects for persistency, data store, mass storage system and data handling system design.

19) Each collaboration must prepare, on a common pattern, a *Memorandum of Understanding* (MoU) for LHC computing, describing the funding of and responsibilities for the hard- and software, the human resources and the policy for access to the computing systems, to be signed between CERN and the funding agencies. Interim MoU's or software agreements should be set up and signed by the end of 2001 to ensure appropriate development of the software.

Details of these and other conclusions and recommendations are described in the following chapters and are further quantified in the appendices to this report.

1.5 Further Observations

The resource estimates used here represent today's best knowledge of the conditions expected at the time that LHC will run. Due to current uncertainties in the time structure of the LHC start-up, the initial LHC performance, background contributions, trigger and data reduction efficiencies, actual physics interests and detector performance, these estimates carry significant *intrinsic uncertainties.* The numbers given in this report should therefore be taken with necessary care. They must be continuously monitored and updated according to actual knowledge.

Initial funding is needed basically now, in order to provide resources for prototype developments. The serious shortfall of currently available human resources for software development and support of the computing infrastructure is a major concern expressed in this report. The experiments, CERN / IT Division, CERN as a whole and all the various funding agencies concerned are asked to undertake the necessary steps to ensure timely availability of sufficient computing resources to permit the LHC to realise its enormous potential in terms of physics outcome.

2 THE LHC COMPUTING CHALLENGE

Starting in 2006 the LHC accelerator will produce proton-proton collisions at 14 000 GeV (centre of mass) with an eventual rate of 10^9 events / s at design luminosity, which, for a multipurpose detector, corresponds to more than 10^{11} particles / s to be recorded. Events of fundamental scientific interest like the production of Higgs particles decaying into detectable decay modes are predicted to occur at an approximate rate of 1 in 10^{12} of the average proton-proton collisions. Operating as a Heavy Ion accelerator, LHC will produce collisions at 5 700 GeV per nucleon (centre of mass). Although the rate (10^4 events / s at design luminosity) will be lower than for the multipurpose detectors, the heavy ion detector will see charged multiplicities for a single event that can reach 8000 particles per unit of rapidity.

Events follow each other in a completely uncorrelated way and any event may be a good one. Thus all events must be recorded with precision by the detector elements, to allow significant online suppression and offline rejection of less interesting events when analysing one of the many different physics questions.

The recording rate of the events after online selection is of the order of 100 Mbytes / s in a multipurpose detector, corresponding to about 100 events / s. The rate for the heavy ion detector can reach 1.25 Gbytes / s. Each experiment will record about 10^9 events per year of operation, corresponding to a mass-storage accumulation rate of around 7 Petabytes / year for all experiments. To this must be added the requirements for reconstructed and simulated data, and all the calibration information, leading to a total mass storage at CERN in excess of 11 Petabytes / year for all experiments. This corresponds to the storage volume of 16 Million of today's CD-ROMs.

To extract precise physics results, extensive reconstruction and analysis codes are required, as well as complete simulation codes that take into account the detailed geometry of the detector, along with its physical properties and the readout characteristics of the electronics. Such programs contain several million lines of code, developed *ad hoc* for each detector by large, distributed teams of physicists and software engineers from the corresponding Collaboration. Because of the complexity and size of these programs, common efforts for their fabrication, especially non-detector-specific utilities and subsystems for software engineering, data processing, access and analysis, have been encouraged from the beginning. In addition, the programs must be optimised for efficient use of simple, basic, commodity building blocks in the underlying computing fabric.

The basic CPU building-block will correspond to the processor(s) of a PC. A total CPU capacity exceeding 2 M SI-95 will be required at CERN, corresponding to more than 100 000 PCs of today's compute capacity.

The LHC community contains more than 5000 physicists, residing in about 300 institutes in ~ 50 countries. Moving around the very large amounts of data so that they have transparent and rapid access to it will be a major challenge for networking.

A suitable co-ordinated worldwide computing fabric must then be created, from the first data recording and near on-line data reduction down to the desktop of the physicist doing analysis. This fabric must be able to cope with the experiments' full analysis chains and with all of the very large data sets accumulated over many years. Each physics subject will be treated by geographically distributed groups of interested physicists within the Collaborations.

This worldwide fabric embodies vast amounts of data and of reconstruction and analysis capacity. Its organisation and use by a multitude of distributed user groups, pose very novel demands on scalability, authentication of persons and co-ordinated collaboration amongst many computer centres. The effective use of the distributed system also requires the application of a new combination of local and global policies for prioritising use, as well as new methods and strategies for cost / performance optimisation. In terms of recent computer science terminology, this would be the first and ambitious realisation of a worldwide data-, compute- and user-intensive "Grid" structure. In the LHC community this model has been designed, simulated and optimised in years of collaborative work in the MONARC study group.

The complexity of the overall task, the likely cost of the very substantial computing infrastructure required, and the need for worldwide accessibility of the data and the applications, have made necessary a comprehensive review of the overall task.

Concerning the timing of the review: this point, at mid-term between the submission of the Technical Proposals and the start-up of the accelerator with the experiments taking first data, seemed the appropriate moment. Driving factors behind this decision were the exponential fall with time of cost / performance for computing hardware[3] and the necessity to understand the software issues adequately.

The structure of the review followed naturally the main areas of work enumerated above, namely panels on software, the worldwide computing model and resources, with the organisation co-ordinated by a steering group. Partners to the review were the experiments and IT Division. The reviewers were rather typical high-level users or organisers of such computing services outside CERN, while experiments and IT Division were represented by their top specialists.

---

3 Over the running period of LEP the compute power available per experiment increased by a factor of approximately 1000, within an essentially constant annual budget.

3 PURPOSE, MANDATE AND STRUCTURE OF THE REVIEW

3.1 Mandate

The mandate of the review as formulated in December 1999 by H.F. Hoffmann, the CERN Director responsible for Scientific Computing, is appended in Annex 1.

3.2 Review Structure

3.2.1 Composition and timing

It was decided to put in place Technical Panels to address each of the three main issues raised in the mandate (a - Software Project, b - Worldwide Analysis / Computing Model and c - Management and Resources), with the panels being co-ordinated by a Steering Group charged with producing the final report. Membership of the Steering Group comprised the nominated chairperson and those nominated as chairpersons of each of the three Technical Panels, along with the CERN Director responsible for Scientific Computing and a scientific secretary. The four main LHC experiments and CERN / IT Division were invited to name representatives (and alternates) to attend Steering Group meetings. The CERN Director responsible for Collider Programmes and the Chairman of the LHC Committee were invited as observers.

Initial discussions with CERN / IT Division, the experiments and the reviewers served to highlight points for consideration by the Technical Panels and these were passed to the panels as initial guidelines. In the knowledge of the subject matter to be addressed, the experiments then appointed representatives (and alternates) to each of the panels and the panel chairpersons co-opted external experts as they saw fit. The list of those who participated in the review is appended in Annex 2.

3.2.2 Working methods and meetings held

Bearing in mind the need to investigate thoroughly a very complex set of issues, the panels set their own pace of work, the necessary co-ordination being achieved by regular meetings of the Steering Group. Communication was facilitated by a private Web site for the review and regular use was made of videoconferencing to enable participation in meetings by those unable to be physically present.

- The *Steering Group* met at CERN twenty-two times (see Annex 3.1).
- The *Software Project panel* met fourteen times between March and September 2000 (see Annex 3.2).
- In addition to numerous informal contacts, the *Worldwide Analysis / Computing Model panel* met five times in March-November 2000, with each experiment individually and with all together (see Annex 3.3).
- Although the work of the *Management and Resources panel* was naturally dependent to a great extent on the findings of the other two panels, it began to gather input already in February 2000. It met sixteen times in total (see Annex 3.4).

3.3 Geographical coverage

It was the aim of the review to survey the LHC experiments' computing needs no matter where in the world these occur, in the belief that the only reasonable way of satisfying these needs lies in the adoption of a distributed computing model. While the CERN-based resources will clearly be a vital component, an important aspect of the review was consideration of the plans being made for regional, national and infra-national centres elsewhere, and for provision of analysis capacity to Collaboration members in their home institutes.

4 SOFTWARE PROJECT

The Software Project panel considered the scope of its work to cover the design, development, support, maintenance and use of the software, as well as the software infrastructure and data management needed to simulate, reconstruct and analyse the experiment data.

The panel expressed its thanks to the experiments and CERN / IT Division for preparing responses to questions and presentations for the panel. The answers were thoughtful and valuable and the presentations were very informative. The sharing of information amongst experiments was a very useful exercise.

The panel recorded a great deal of information about the current status of the software and about the architecture and product choices of the experiments (summarised in Appendix 1). It identified several areas of concern and risk, each of which was discussed in detail in its report to the Steering Group together with suggestions on how they might be addressed. These points are summarised below.

4.1 Areas of concern and panel suggestions

4.1.1 Human resources shortfall - maturity and accuracy of estimates

Taking into account the uncertainties involved, the panel endorsed the required human resources specified by the experiments as far as such estimates were provided (ATLAS was still evaluating its needs). It was concerned about the projected shortfalls, which are significant, and recommended that CERN and the experiments together do all they can to provide the total effort needed. After the panel concluded its work ATLAS provided an estimation of its needs, while ALICE and LHCb revised their numbers. The data in Appendix 3, Table A3.12 reflects these new figures, which reinforce the panel conclusions on this point.

While understanding that the available resources at CERN are limited and fixed, the panel saw a strong inconsistency between a software infrastructure co-ordination and support role for CERN / IT, and the projected IT staffing levels in the future. It recommended that IT establish a statement of mission, a strategic plan, and a detailed assessment of staffing levels and types of staff needed to carry out that mission for the LHC experiments.

In addition, the panel suggested that experiment leadership look again at overall needs for human resources within each experiment and try harder to direct sufficient resources to software and computing issues, commensurate with its increasingly important role, and with the increased complexity of building software systems and deploying them worldwide.

During the earlier stages of the review, all the LHC experiments, and many others, were undergoing a massive transition from Fortran to OO programming. This led them to propose that, given the fact that converting the entire HEP community is a major undertaking, CERN as an institution should sponsor a more coherent programme to help bring about this change, including bringing in more C++ and OO experts explicitly to work closely with the experiments.

The panel suggested that careful tracking of the required and actual levels of human resources, as well as the basis and procedure for the estimation, is needed to validate the resource models used and to gain more confidence in predictions.

4.1.2 Simulation package development and support

Continued CERN / IT contributions to support Geant4 are expected by all experiments. The initial period of validity of the Geant4 MoU was two years. It is subject to tacit renewal in two-year steps. The first period expired at the end of 2000. According to the MoU, amendments are discussed in the Geant4 Collaboration Board where CMS, ATLAS and LHCb are represented. The panel suggested that the many concerns it heard (from all experiments except CMS) be brought to the attention of this board and CERN Management.

The role and support of FLUKA (requested by three experiments) needs to be further clarified and agreed on by CERN / IT, CERN / EP, the CERN Director responsible for Scientific Computing, the experiments and the authors.

4.1.3 Software and System Architectures

While most seem to agree that the fundamental requirements for a framework and architecture are the same for all experiments, the implementations chosen by ALICE and CMS differ from each other and from that now being developed and maintained by ATLAS and LHCb in common. The frameworks are still evolving but substantial progress in building and developing both the framework and the application software for each experiment has already been achieved. This area thus does not seem to be one where trying to force / forge a common approach would be likely to meet with success.

The panel suggested that all experiments pay close attention to taking a layered and modular approach to building their software systems.

4.1.4 Common projects, experiment-CERN/IT interactions

The panel suggested that the RD45 project be considered as terminated. It expected an immediate technical assessment and the formulation of some new project or projects for data management, without compromising ongoing activities.

It urged CERN / IT Management to take steps to understand what staffing levels, types of staff and forums for communication would be needed to improve the working relationship and trust between experiments and IT for software development and support.

Despite all the difficulties identified, the pursuit of common approaches was considered to be indeed a worthwhile goal. It must, however, be done with care and with understanding of the pitfalls and difficulties involved. In some cases a support policy statement from CERN Management might be required, in order to provide incentives for experiments to participate in a common effort.

The panel reaffirmed its belief that CERN / IT has a very important and special role to play in assuring the success of taking a common approach, or carrying out a common project. On the other side, experiment leadership and experiment software and computing management has an important responsibility to support fully and to provide human resources to contribute effectively, on the agreed time-scales, to any common effort in which the experiment participates.

4.1.5 Data management

CERN / IT should consider forming a team which designs the data-management system together with experiments and helps, in so doing, to propagate the best ideas and practices. IT will be instrumental in the set-up and management of a common prototype computing system, a concept put forward also by the other two panels and taken up in Chapter 7. The Division will play a large part in maintaining and operating the Tier0+Tier1 centre[4] at CERN, together with the infrastructure and operational load imposed by serving numerous other Tier1 (and Tier2) centres, and receiving their contributed computing resources and processed data. This effort will not be small. Experiments and IT must work together closely.

The existence of reliable, robust, long-term, well-supported solutions for both persistency and experiment data management systems is absolutely fundamental for each experiment. Furthermore the data management systems must work, to some extent, right now – not only from 2005 – and must evolve under the strain of real users attempting to get their work done in the interim years.

The panel expressed the belief that, prior to the decisions on the baseline data storage system by CMS and ATLAS in 2001 / 2002, the underlying requirements and cost-benefit equations need to be re-evaluated. The experience at BaBar and other experiments should be taken into account.

One must also understand better the realistic possibilities for the highly-valued requirement of random access to data on-demand, as apparently provided seamlessly, but not without cost, by a true object database.

Some experiments wish to retain the possibility of storing run conditions and calibration data using a different mechanism than for the event data. CERN / IT will not be able to operate production Conditions Database Services for four different flavours of database management system. Run and Conditions Databases should thus be addressed in a data management or special Technical Assessment Group (see panel recommendations 1 and 2 in 4.3 below) and a common approach, or approaches, should emerge.

A high level CERN and Experiment joint strategy must be developed (not necessarily a unique solution) on the approach to data handling. Significant projects, with experiment involvement and strong project leadership, must be started at this time to implement the strategy. This will involve reassessment of requirements and definition of one or more product deliverables, as well as realistic long-term plans for development, operations and support.

---

4 See 5.1.1 for a discussion of the hierarchical computing model and a definition of Tiers

4.1.6 Support for software packages and future evolution

While conceding it may not be easily achievable, the panel urged CERN / IT to play a leadership role in bringing together the community to get its work done. In doing this it should take advantage of, learn from, and help to evolve, much of the work that has been done. It should also form the anchor of an effort, with broad experiment participation, and in some cases broad HEP participation, that ensures appropriate new tools and ideas will be found or built for the future (The AIDA project, in which CERN / IT is involved, has started to take this approach).

CERN / IT clearly plays an important role in the support of software packages. It is vital for the experiments that support is continued for the packages that they use widely, including the Geant4 simulation package, object database and general libraries.

Most of the LHC experiments stated that they would like to see ROOT officially supported by CERN. The majority of the panel expressed the belief that it is high priority to consolidate the existing support and ensure that the immediate needs of the experiments using this package are met. Regarding the longer term, its future should be considered, together with other related on-going projects, as part of the overall programme of work of the laboratory (see also the panel recommendations 1 and 4 in 4.3 below).

The majority of the panel suggested that CERN seriously consider taking on ROOT core-team support as a mainline activity.

4.1.7 Quality assurance of physics reconstruction/filter/trigger code

The panel suggested that plans to assure, on a continuous basis, the correctness and stability of reconstruction code for online filter farms should be developed.

4.2 Common Projects

The panel identified several areas where joint efforts amongst one or more experiments and CERN / IT, resulting in common projects and products, might lead to cost savings, or decreased risk, or both.

The way in which common actions should be initiated, staffed, managed, overseen and terminated is addressed in the panel recommendations (4.3 below). While projects that fall within the purview of the proposed SC2 committee must be formally proposed (and agreed to by the experiments), a number of areas were identified that currently are seen as either most urgent to address in common or most likely to lead to success and optimal use of resources:

- Data Grid[5] system (building on and in collaboration with existing Grid projects)
- Data management system
  - Persistency and the data store
  - Mass Storage System
  - Common prototype for Data Challenges
- Interactive data analysis tools
- Simulation packages
- Common Tools projects
  - Configuration management and release tools
  - Coding rules and checking tools
  - Geometry specification tools
  - Run conditions databases
  - Production Farm management

Common Grid Projects, including the DataGrid in Europe, and GriPhyN and the Particle Physics Data Grid, are already underway. Other common activities, rather than defined projects, are taking place in some areas, such as simulation and data analysis. Developing common Grid systems, common policies and procedures for managing wide area and local area networks, and a common data management approach and strategy are among the first priorities. They could potentially have a considerable effect on the costs, operations and interoperability of computing and software systems not only at CERN, but also at the other Tier1 and Tier2 centres worldwide. Some of the potential common projects identified are rather small, but should nevertheless be pursued, although less urgently and possibly not all at the same time. Small projects with minimal set-up and overheads may also serve to foster better communications amongst experiments (and IT).

The panel also noted that currently projected CERN / IT staffing levels will not be compatible with successfully carrying out the above common projects, since this requires significant CERN / IT participation in the short term and a commitment to the support of products in the long term. It recommended an in-depth review of the current IT projects and plans to determine the current

[5] See e.g. The GRID, Blueprint for a New Computing Infrastructure, I. Foster and C. Kesselmann ed., ISBN 1-55860-475-8

human resources availability for common projects and quantify the shortfall (cf. Appendix 3.3).

The panel stressed the vital role of timely Data Challenges and encouraged the experiments to schedule increasingly rigorous and realistic challenges, in close collaboration with CERN / IT.

4.3 Recommendations

After taking into account the risks and concerns identified and the evaluations for each software topic examined, the panel made four major recommendations concerning a specific suggestion to implement immediately a formal process to foster, staff and manage common efforts. Similar input came from the other panels and was followed up in the Steering Group, resulting in the overall recommendation made in Chapter 7.

The Software Project panel believed that it would not be a helpful or tenable position for it to make recommendations to experiments about specific products or technologies, or to recommend that they immediately modify their current strategies and choices. Rather, it suggested a course of action expected to lead to additional common efforts and, where appropriate, reassessment of choices by the experiments themselves. Given the time still before experiment turn-on, and the very rapid rate of change of technology and software products, experiments must plan for a high degree of change. The panel suggested that they do this together, as far as possible, in the context of ongoing working groups and projects.

The four recommendations were as follows:

1. Establish an LHC Software and Computing Steering Committee (SC2)

- Convened by the CERN Director responsible for Scientific Computing;
- Composed of highest level software and computing management in experiments and IT;
- Appoint, only as needed, focused Technical Assessment Groups to formulate common approaches and projects, defining their scope and tasks. Although many of these TAGs will have short-term mandates, others will have roles that are needed throughout the LHC programme;
- Oversee and staff common projects – demand and review work breakdowns, including for R&D and prototyping phases.

2. Establish a Data Management Technical Assessment Group

- Immediately appoint a Data Management Technical Assessment Group with a charge to report rapidly;

- Agree on as many components of a common data management strategy as possible. This may permit a choice of basic data management system components and will allow for evolution;
- Initiate and closely monitor well defined projects for persistency and data store, mass storage system and data handling system design.

3. Validate and present needed human resources and current availability

- Rapidly develop a validated human resources profile, use informal software agreements;
- Refine and track needed and used human resources frequently, prior to formal MoUs;
- Provide as further input to funding agencies and CERN Management.

4. Initiate Technical Assessment Groups in a number of areas

- Address areas of concern or opportunity identified by the panel;
- Identify and define common strategies and common projects between experiments and IT and amongst experiments (where deemed necessary by the SC2) with a definite task and timetable.

The panel expressed the hope that some of the specific suggestions and comments contained in its report, together with the information gathered in the course of the review, would be useful starting points for the work of Technical Assessment Groups.

The panel was asked not only to make an assessment of the current status and plans, but also to “recommend actions and, in particular, common actions between experiments and IT Division that will help achieve the goals within existing resources”. As described above, it in fact reached the conclusion that it is not possible to meet the aims as charged "within existing resources". It also noted that achieving economies through common actions requires an effective process to ensure that beneficial common actions are identified and then properly executed.

5 WORLDWIDE ANALYSIS / COMPUTING MODEL

This panel viewed its goals as three-fold:

- to review the LHC off-line computing needs from raw data to the physics plots (calibration, reconstruction, simulation, analysis);
- to review the work on overall worldwide analysis and computing models and endorse a direction to be followed;
- to enumerate the main parameters of this model, with associated hardware needs.

Their report to the Steering Group summarised the main conclusions and recommendations of the panel and was intended as an input to the Management & Resources panel to allow it to estimate costs and human resources requirements. Since some of the work was done together with the Management & Resources panel, the report also reflected the outcome of joint meetings and of many discussions amongst members of the two panels.

Several of the issues and questions raised in this panel were also addressed by the Software Project panel, the findings and recommendations of which were strongly endorsed.

An overall positive conclusion was that the four Collaborations all agree on common solutions, and sharing of resources and efforts. The general atmosphere of the sessions was collaborative and the panel expressed its thanks to the four experiments and CERN / IT Division for their contributions.

5.1 Main considerations

5.1.1 Worldwide Analysis and Computing Model

From the start, the panel recognised that the deceptively simple option of placing all the computing capacity at CERN was entirely impractical for a whole set of reasons including the necessity to access funding that would not be available at CERN and to exploit established computing expertise and infrastructure in national labs and universities. There is also a clear wish to devolve control over the computing resources. Other large organisations that are themselves distributed have also rejected a centralised approach.

The panel strongly endorsed the adoption of the model developed by MONARC[6] (MOdels of Networked Analysis at Regional Centres for LHC experiments), a collaborative effort of all four experiments. This multi-tier hierarchical model has in fact already been adopted by the four Collaborations (one Tier0 and one Tier1 at CERN, a few Tier1's or Regional Centres, Tier2, Tier3...). The panel wished to record its recognition of the very valuable work performed by this R&D project since 1998, while noting that many of the concepts have been refined by the Grid projects and the US Software and Computing Projects. In particular those that relate to the Tier2 centres were originated at the initiation of the GriPhyN project.

From additional information on the MONARC model received by the panel, Tier Computer Centres are – in first approximation – defined as:

- Tier0 Raw data storage; first calibration & reconstruction; v. large storage capacity
- Tier1 Further calibration and reconstruction passes; large fraction of simulation and analysis; large storage capacity; associated support
- Tier2 (and lower levels) - The balance of simulation and analysis. Tier2 access would be limited to one country or a subset of countries, with typically ~50 active users. Each Tier2 centre would depend upon one dedicated Tier1 (the Tier1 at CERN in some cases) for co-ordination and optimisation purposes. Although Tier2's are smaller in size than Tier1, all experiments agree that their larger number can make them as important in computing power as the set of Tier1's.

Tier0 and Tier1 centres are basically open to all members of a Collaboration (some 200-500 active users per experiment) under conditions to be specified in MoUs. This automatically implies that the Tier1 centres must be available for the lifetime of LHC and leads to the expectation that, for each experiment, the Tier0 centre and all the Teir1's together will be managed coherently. This is currently conceived to be a Data Grid system (see footnote 5 on page 19). It is now possible to envisage such a model thanks to the ongoing and foreseen rapid growth of network bandwidths and associated services in the next years.

This faith in the distributed model is underpinned by confidence that the efforts related to the Grid concept will provide the necessary technological

[6] MONARC Phase 2 report CERN / LCB 2000-001, March 2000
http: / / monarc.web.cern.ch / MONARC / docs / phase2report / Phase2Report.pdf

infrastructure on the required time-scale. In this respect the HEP-proposed Grid projects, both GriPhyN and PPDG in USA, and the DataGrid[7] Project, recently approved by the European Commission, are expected to make major contributions. The Grid activities represent a fortunate opportunity for the HEP community to solve several key LHC computing problems. It is the current major direction of investigation in this respect and significant resources will be invested in it. The Grid concept goes beyond a simple hierarchical model and in time may lead to the entire set of Tier1 and Tier2 centres being seen by the end-user as a single facility.

5.1.2 Estimation of needs

Following the distributed computing model, quantitative estimates of the needs of each of the four experiments were assembled in order to provide input for the Management and Resources panel from which it could derive the budgetary and human resource estimates. This information is attached in Appendix 2.

5.1.2.1 Event Sizes and Rates

ATLAS and CMS consider recording rates around 100 Hz, with raw event size in the 1 MB range. Effective running time has been normalised to 10^7 sec / y (~110 days). These values have been used as the baseline in p-p collision mode, but they still have basic uncertainties. For instance, studies in ATLAS during the course of this review have led them to propose ceiling values as high as 270 Hz and 2 MB, with associated resources listed in the last column of Table A2.1. They have been considered as a useful reference for further estimates and studies, with the main impact being on Tier0. Such an event size arises if zero-suppression cannot be done on-line, while almost 100 Hz of the increased trigger rate comes from the wish to perform B physics. It is these revised figures for ATLAS that have been used by the Management and Resources panel in their calculations, since they represent the current best estimate from the Collaboration. LHCb foresees a 200 Hz recording rate, but with a much smaller event size (0.125 MB). ALICE is planning for 10^6 sec / year Pb-Pb running with a trigger rate of 50 Hz and an average event size of 25 MB. Dimuon events would have negligible event size but large trigger rate, etc. Proper averaging of such

---

7 See e.g. "Grid Computing: the European DataGrid Project", proc. IEEE 2000 Nuclear Science Symposium and Medical Imaging Conference, Lyon, 15-20 October 2000. The project also has a Web page at http://www.cern.ch/grid

different numbers is, understandably enough, not easy to perform and the balance will vary with time.

5.1.2.2 Data Storage

The four experiments plan to keep all their raw data at the Tier0 (CERN) and perform there the first reconstruction. The modality of data export from Tier0 will depend on the DataGrid project results and on the performance / cost of network bandwidth. At present, the experiments do not foresee massive raw data export. To first order, larger bandwidth availability will not change drastically the distributed computing models, but it would facilitate import / export operations (Tier0⇔Tier1 and Tier1⇔Tier1). On the other hand, if traffic from the Tier0 to the regional centres (and back) were to increase by large factors, the balancing of tasks within the distributed computing model might have to be reconsidered. This would also imply looking again at the distribution of costs.

The question of raw data backup is mainly constrained by financial considerations and hence by the expected data volume. ALICE would like to make a backup if this turns out to be economically feasible and studies to this end are under way. LHCb (which expects significantly less data) plans one full copy, while ATLAS and CMS decided during this review to do the same. In all cases, the technical solution is a full copy of all cartridges. The option of RAIT technology (= "RAID on tape") was discarded, as it does not protect the data against major disaster in the tape vault. The cost of this operation, although important, is clearly small compared to the total investment. Due to technological change and limited shelf-life of current tertiary storage media, not only the backup but also the migration to new storage media must be planned for. In the long run, this will probably be a major task of the Tier0 facility.

5.1.2.3 CPU Estimates

The panel recognised that CPU estimates, which are in the million SI-95 range per experiment, are at this stage very difficult to make with great precision, especially for the "final analysis". Estimates of event size, both for the raw data (ATLAS) and reconstructed events (several experiments), have grown since the first estimates were made. The trigger rate is set by a combination of physics goals, the need to understand and monitor the detector, and the use of resources on- and off-line. It must be expected therefore that the estimates for

CPU and storage will change (up and down) as the Collaborations continue to refine them.

The figures given in Appendix 2 are the baseline to perform the physics programmes and can also be taken as the basis for setting up the working constraints to guide effective deployment of the analysis software components.

5.1.2.4 Connectivity

The experiments estimated the CERN-Tier1 bandwidth needed in 2006. This work should be extended to include Tier1-Tier1, Tier2-Tier1 and Tier2-Tier0 traffic. Bandwidth is needed not only for bulk data transfers from TierN to TierN+1 but also for individual analysis tasks, for traffic to CERN from production and analysis performed elsewhere, between Tiers, for distribution of re-reconstructed data, for collaborative activities (meetings, videoconferencing, etc). To avoid saturation, a network should rarely be occupied above the 50% level, implying that the installed leased-line bandwidth should be at least ~twice as large as the requested sustained throughput values.

Altogether, the single experiment requirements are some 1.5-3 Gbps, as stated by CMS. The other experiments have not all studied bandwidth needs beyond the simple MONARC estimate and provided only partial requirements. Nevertheless, after reflection, they agreed to endorse the CMS estimates as applicable to themselves. Experiments like BaBar have shown that exceeding the planned bandwidth gives more freedom for decentralised computing resources, leading each LHC experiment to foresee bandwidth in the Gbps range over the whole Tier0⇔Tier1⇔Tier2 set. The impact of each Tier on LHC computing will depend strongly on its connectivity. The presently foreseen connectivity levels are appropriate to implement the distributed computing model "à la MONARC".

There can be no doubt that the required bandwidth will be technically achievable by the time of LHC start-up. Undersea cables of 800 Gbps (80*10 Gbps) capacity per fibre pair and four fibre pairs are already scheduled for 2004-5 using current DWDM technology. Technology such as this should permit end-to-end Gigabit-Ethernet or other technical solutions on HEP backbone networks.

No estimate has been made yet for transparent access to data and resources as expected from an eventual LHC Data Grid. Network engineering, protocol stack optimisation, throughput monitoring throughout the network and availability of Grid middleware are, however, needed to make possible the efficient use of the high bandwidths quoted above.

Accordingly, the panel recommended that the evolution of the full cost per unit bandwidth (including network site equipment and the necessary engineering systems) be carefully monitored and the computing model adapted accordingly.

5.1.2.5 Prototyping

All the experiments plan "Data Challenges", tackling different critical points of their computing model with samples of simulated data and studying the physics, trigger and detector performance. Several Data Challenges have already been performed. In the near future, they will include, besides CERN resources, some or all of the Tier1's available (and possibly Tier2's) to solve the deployment issues of the entire system, from DAQ flow and on-line farm to the physics analysis. With the proviso that improved co-ordination of Data Challenges amongst the experiments might bring resource-sharing benefits, the panel endorsed the plans to perform a sequence of Data Challenges of increasing size and complexity until LHC start-up.

Prototyping of distributed computing system components and interconnections, leading on smoothly to the deployment of the production systems, is clearly required. Both this panel and that for Management and Resources made strong recommendations in this respect, which are taken up in Chapter 7.7.

5.1.3 Operating Systems and Persistency solutions

The choices made concerning Operating Systems and Object Persistency solutions can strongly influence the hardware needs and the human resources required to support the production installations.

Distributed computing models and the emergence of a hierarchical structure of Tier centres force a review of the traditional free-choice attitudes to these aspects. Not only CERN but also most other Tier1 centres will provide resources for several or all LHC experiments, together with other HEP experiments or non-HEP needs. For a given level of human resources, each centre could provide more services if one could limit the proliferation of Operating Systems and of solutions to handle the event store.

In what concerns Operating Systems, there is already a well-established trend towards one single Unix-like system, namely LINUX, or at most two, consisting of LINUX and one commercial UNIX OS. There are in fact good arguments to have a second platform for code and result verification.

For event store Persistency, the number of solutions, especially commercial ones, has a strong impact on resources and should be kept to a minimum. The Software Project panel discussed at length the technical and qualitative aspects of such a choice but this panel felt it important to return to the point, because of the consequences of any additional choice. Criteria include portability on different platforms and OS, open source for maintainability, interfaces with other products, licences if commercial, human support and expertise, etc. This is indeed a highly sensitive issue, given the investments already made by some experiments.

In view of the above, the recommendation that emerged after several meetings was to accept that CERN should support a maximum of two choices for object persistent systems. This was felt to be a good compromise between freedom and limited resources. In addition, the panel wished to stress that the existing products / solutions (ROOT and Objectivity) should not necessarily be considered as the final choice, provided that choosing a new product would mean phasing out an existing one.

5.1.4 Resource Sharing

Although the panel members and experiments agreed that the cost sharing between CERN and elsewhere is a matter of policy and subject to negotiation, the actual physical location of resources will have a strong impact on the efficiency of their use.

There was agreement that resources located at CERN should include three linked infrastructures and their associated human support:

- at the experiments (on-line farm etc. outside running periods);
- Tier0 for raw data storage and reconstruction;
- Tier1 mainly for analysis needs of CERN-based researchers + some simulation.

Concentrating too many resources at CERN could, however, create bottlenecks (WAN and LAN, etc.), as shown by MONARC and other simulation studies, and is impractical due to financial, managerial and working-efficiency considerations.

With this in mind, the panel arrived after discussion at the capacity sharing proposal shown in detail in Table A2.1. This indicates (using the canonical 100 Hz trigger / 1 MB event size figures for ATLAS) that {CERN Tier0+1} and {Tier1+2 not at CERN} should provide respectively:

- 11.5 and 17 PB of tape storage (40% - 60%)
- 2.3 and 8.1 PB of disk space (22% - 78%)
- 2.4 and 5 M SI-95 CPU (33% - 67%)

Although Tier2's have been included in recent MONARC simulations and CMS has also worked on the specifications for workload sharing between Tier1 and Tier2's in the US, the relative importance and roles of Tier2, Tier3, etc. have yet to be studied in depth in most cases. They will, however, clearly be country- and experiment-dependent. Nevertheless, it must be emphasised that, for this review, experiments have estimated total needs and have in fact made allowance in "Tier1" for the contribution of Tier2.

It is interesting to observe that, in the above figures, the ratios CERN / non-CERN are not far from the 1/3-2/3 traditionally adopted by CERN policy makers. Furthermore, it is expected that the Tier2's together will contribute as much CPU power as the Tier1's. One thus gets to first approximation three thirds equally shared between a) CERN, b) {Tier1} and c) {Tier2}, namely:

- Tape storage (40% - 30% - 30%)
- Disk space (22% - 39% - 39%)
- CPU (33% - 33% - 33%)

It is not desired to account for resources below the Tier2 level since in most cases these will not have the level of management and support required for adequate central co-ordination.

5.2 Recommendations

In the light of the above the panel made the following recommendations:

1. A Steering Committee for Software and Computing (SC2) should be created. (This is taken up in Chapter 7.6 below, where the proposed charge of the committee is developed in detail).
2. The multi-tier hierarchical model is recommended as being a key element of the LHC computing model.

3. Grid technology should be developed and tested as it is currently the best candidate to provide the basic software tools needed for efficient use of resources in a system of distributed processing and distributed data.
4. All experiments should carry out their programmes of Data Challenges of increasing size and complexity until LHC start-up.
5. A prototype of the planned distributed computing facility should be built by the four experiments, CERN / IT Division and the set of Tier0+Tier1+Tier2 centres as a common project, with the goal of reaching a significant fraction of the overall computing and data handling capacity of one LHC Experiment.
6. The Tier0+Tier1 prototype hardware to be installed at CERN should be planned as one partitionable facility.
7. It is highly desirable that MONARC Phase 3 and this prototype project should rapidly co-ordinate their efforts (also suggested by the Software Project panel).
8. There should be a maximum of only two persistency tools for handling the event data stores.
9. Support of existing products should be continued for those using them but they should not be necessarily considered as the final choice.
10. The evolution of the cost of wide-area networking per unit of useable bandwidth should be monitored on an ongoing basis and the computing model adapted accordingly.

6 MANAGEMENT AND RESOURCES

The Management and Resources panel's mandate was to evaluate the resources necessary to run the off-line analysis of the four major LHC experiments, to estimate the cost and the necessary human resources, and to propose plans for addressing the issues.

The panel made extensive use of the conclusions and recommendations of the other two panels. The meetings were open to all interested experts. On many occasions members of the other panels were present and made substantial contributions, bringing to this panel information on their deliberations.

The panel reviewed the expected necessary computing infrastructures in terms of CPU, tape and disk storage, human resources and organisation, as presented by the four Collaborations. It took care to expose the requirements of the four experiments in such a way that they can easily be compared (see Appendix 3).

A difficult part of the panel's mandate was to estimate costs. For this it used the results of the PASTA[8] committee (on which some of the panel members had also served) to predict component costs in the future years. In addition, however, it asked external experts for their estimates, in order to obtain independent evaluations of the cost and human resources necessary to run a Regional Centre. The comparison of these independent evaluations, from centres such as RAL in the UK, Lyon (IN2P3) in France, INFN in Italy, DESY in Germany, and DOE and NSF in USA, allowed the panel's understanding of the cost uncertainties to be verified.

The panel was unanimous in its opinion that the complexity of the computing systems required is such that a common initiative, involving all LHC experiments, CERN / IT Division and several Regional Centres, to build a shared prototype of the LHC computing system, should start soon (taken up in detail in Chapter 7.7).

---

8 PASTA, the Technology Tracking Team for Processors, Memory, Storage and Architectures, was set up by IT Division and the LHC Computing Board (LCB) to follow the progress of some of the basic technologies required for LHC. In 1996 a first report was issued: http://wwwinfo.cern.ch/di/pasta.html. In 1999 a second report was issued: http://tilde-les.home.cern.ch/~les/pasta/run2/pasta_report_2/report.html

As a final caveat, the panel warned that many factors pointed to the need for a flexible strategy to optimise the cost / performance ratio. Chief amongst these are the rapid evolution of the performance and costs of computing, the possible luminosity growth of the LHC machine with time and the necessity to learn how to use the apparatus at higher trigger rates.

6.1 Computing Model

As explained in Chapter 5.1.1, the baseline LHC computing model is founded on a distributed multi-tier architecture involving Regional Centres and is a direct outcome of the MONARC research activity.

Many site-specific resources will be connected to the Tier1 regional centres, extending the distributed computing hierarchy to lower levels (Tier2 etc.). Tier2 centres will participate in the overall system via their reference Tier1 centres, providing capacity for both scheduled batch activities (co-ordinated by the Tier1) and user-driven analysis activities. Lower levels are also foreseen, in the form of Institution- and desktop-resources, each level down being progressively more dedicated to individual analysis tasks. This hierarchical scheme will provide the user community with access to appropriate resource levels, corresponding to the needs of the individual user, the physics group, and the experiment.

The Collaborations gave their current understanding concerning the provision of Tier1 and Tier2 centres around the world. This information, the best currently available, is reproduced in Appendix 3.1 and was used by the panel for its resource calculations, but it is obvious that the figures are likely to evolve with time.

6.1.1 The Tier0 and Tier1's at CERN

Each experiment requires a dedicated Tier0 and Tier1 located at CERN. The main use of the Tier0 component of this system will be the proper collection and processing (calibration, reconstruction, monitoring, and reprocessing) of the raw data. The CERN Tier1 will be devoted mainly to the CERN-based researchers and to some of the collaborators not having a Tier1 in their country.

6.1.2 The Tier1's at the Regional Centres

Each experiment will use several Regional Centres outside CERN. ALICE currently plans 4, ATLAS 6, CMS 5 and LHCb 5. "Regional" may here imply a geographical zone *larger* than a single country. There will be two Regional Centres in USA, one (BNL) for ATLAS and one (FNAL) for CMS. There will be several Regional Centres in Europe, at least one in each of the UK, France, Italy, Germany, and Russia. Those in the UK (at RAL), France (at Lyon), Italy and Germany will serve all four experiments and will be dedicated to their needs.

6.1.3 The Tier2 distributed analysis

As noted in Chapter 5.1.4, the resource sharing amongst Tier1 and Tier2 (or smaller) facilities could vary significantly from region to region. Tier2's are viewed loosely as "national" centres. In some regions, groups of institutes may find it in their interest to provide Tier2 centres in common. In other cases a Tier1 centre complemented only by the local (institute-level) computing facilities may be the chosen solution.

6.1.4 Tier3 and lower levels

Institute-level facilities (desktops, local servers, etc.) will certainly be necessary too, and could indeed constitute significant computing power, but these facilities are assumed to be provided by institute-level funding and are not included in any estimates now given.

6.1.5 The CERN-based Computing Centre

The CERN-based computing facility, as it will result from merging the requirements of the four experiments, will be a common system (CPU and data storage) partitionable amongst all experiments. CERN / IT Division should have the responsibility to run it together with the four Collaborations. The four Tier0's and Tier1's, thus merged together into a single system, must clearly work in close collaboration with the external Tier1's. The CERN computing system will suffice to take data, reconstruct them and perform selected analyses, at the nominal luminosity of LHC (10^{33} for p-p and 10^{27} for Pb-Pb runs). Its expected performance, summarising the requests of the experiments, is shown in Appendix 3.2.

6.2 Full initial system investment costs and schedule

The panel discussed further with all parties the detailed capacity requirements prepared by the Worldwide Analysis / Computing Model panel and used these, along with studies of cost evolution and input from other sites, to prepare costing estimates. These estimates are shown in Appendix 3.5.

The estimates clearly depend strongly on the construction schedule for the systems. The figures given correspond to matching the presently understood start-up timing of the LHC machine and experiments (first beam in early 2006).

6.3 Prototype investment costs and schedule

The panel strongly endorsed the proposal of the other two panels that a realistic shared prototype of the final systems should be built. By realistic is meant that it should embody all the main components, including aspects related to distribution, and should grow progressively to reach a capacity corresponding to some 50% of the needs of a single experiment. The schedule for construction of this prototype is obviously dictated by the need to use the results of the prototyping work to influence the deployment of the full system. The corresponding costing estimates for the portion of the prototype at CERN are given in Appendix 3.5.5.

6.4 Maintenance and operation costs

Maintenance and operation costs were only evaluated for the CERN-based facility. The associated model and figures are shown in Appendix 3.5.3.2. Other centres may well use somewhat different models but the estimate shown may be taken as indicative. It is important to note that while, for the foreseeable future, the one-for-one replacement costs of individual components can be expected to continue to fall, this trend is counterbalanced by a natural growth in capacity requirement as the LHC machine and experiments mature.

6.5 Human Resources

The panel considered the human resources required for LHC computing under five headings: deployment of the prototype and full initial systems, systems administration, physics software in the experiments, Core Software[9] in the experiments, support effort from CERN / IT Division.

No particular concerns were raised about human resources for system deployment although it is evident that in a resource-constrained environment, both at CERN and elsewhere, there will also be pressures on this aspect.

---

[9] The Core Software requires software engineers just as construction projects require mechanical and electrical engineers. These people generally have formal training and experience in Computer Science although they may in exceptional cases be physicists who have undergone some years of appropriate re-training. Typical skills include, but are not restricted to: architectural design, software frameworks, database administration, quality assurance, code management, software configuration, software build and distribution systems. Software engineers provide a well-designed, functional, efficient, reliable and maintainable environment for use in the critical on-line event selection applications and within which physicists are able to develop the physics software required to analyse the LHC data.

System administration was only considered for CERN, where it is looked on as a fully outsourced activity, which thus becomes a cash expense (see Appendix 3.5.3.2). The resultant amounts have been accounted under Maintenance and Operation costs for CERN. The input received from elsewhere indicates that they are likely either to cope with their foreseen staff profiles for this activity or adopt an approach similar to CERN's.

Concerning the physics software, the Collaborations were reasonably confident (see Appendix 3.4b) that enough human resources would be available from their physicist communities to carry out the required work.

For the other two headings, however, the picture that emerged was extremely alarming and, compared with the needs, very significant shortfalls in the currently planned human resource levels were identified. This highly critical aspect for the success of the LHC experiments is expanded-on further in the following sections.

6.5.1 Core Software human resources needed by the Collaborations

The estimations given by the Collaborations for the human resources necessary to build and run the Core Software are summarised in Appendix 3.4. The panel endorses these estimates, which it finds to be consistent and reasonable. It must be emphasised that one is talking here of expert software engineers and not *ad hoc* help from enthusiastic physicists. The discrepancy between the needs and the actually available level of human resources is striking. Averaged over all the experiments, the human resources in the year 2000 were only 70% of those required, a figure that will drop to 60% by 2002 in the absence of positive action.

6.5.2 Human resources from CERN/IT Division for support

CERN / IT Division provides the computing infrastructure of the laboratory, including general file services, campus networking, wide area networking, desktop computing support, Web servers, relational database support, batch computing services, telephone services, etc. These services form an important base for more specialised physics and engineering computing services, the details and evolution of which are agreed in a number of formal committees. The specialised services that IT expects to provide for LHC physics computing are:

- Co-ordination and base support of the Geant4 simulation framework

- Provision of a set of data analysis and visualisation tools
- Provision of standard libraries, including common mathematical libraries
- Data management interfaces
- HEP class libraries
- Support of tools for software development and software process
- Data recording and mass storage services
- Support for object persistency for experimental data
- Support of common controls solutions (JCOP - the Joint Controls Project)
- Support for computing tools widely-used by the LHC Collaborations
- Provision of the computing services for physics data handling at CERN (Tier0+ Tier1). This includes the development of the service model, and the acquisition, installation, operation and maintenance of the computing equipment – processors, storage, databases and local area networking.
- Provision of high bandwidth wide-area networking to support the regional centres computing model
- The worldwide co-ordination of the Tier1 computing infrastructure
- Participation in and co-ordination of selected common software and computing projects, when permitted by available resources. Current examples include the DataGrid project; common Tier0+Tier1 prototype; participation in Data Challenges with ALICE, ATLAS, CMS and LHCb, in which IT provides object database and mass storage expertise and services.

Although it is very difficult to put precise numbers on the human resources that will be needed in all these different categories, the panel does not believe that IT will be able to cope with the demands being put on it for LHC computing with a staff level significantly below that at present. This is true not only because of the vastly increased scale of the exercise but also because of additional specific support demands identified by the Software Project panel (cf. Chapter 4.1.2 & 4.1.6 above).

CERN is, however, currently committed, for budgetary reasons, to significant progressive staff reductions through to 2006 (a fact that can also only exacerbate the problems for Core Software outlined above). For IT, the current plan foresees by 2006 a reduction in strength of 50 on a 1999 complement of 187, a measure that, if carried out, would render the Division totally incapable of fulfilling its mission in the Organization. CERN Management has already, in fact, proposed to Council that the reduction be limited to 30. The panel considers that even this level would imply a significant risk for LHC computing. To be acceptable, it would require more standardisation of common software and services than currently seems to be achievable and a significant lowering of user expectations.

Considering the needs and relative priorities of its various tasks, IT has prepared two sets of projected staffing figures by activity, one for a model with 30 fewer staff than today ("-30 model") and the other with the same number as today ("zero model"). These are shown in Appendix 3.3.

The panel has verified (Table A3.11) that the "zero model" IT estimate for operation of the CERN-based centre (23 staff plus the outsourced effort) is significantly lower (on an equivalent capacity basis) than the estimates being made by three prospective European Regional Centres. Even allowing for economies of scale, this, together with the pivotal role of the CERN centre in the LHC distributed computing system, seems to indicate that the IT estimate is not over-generous.

In an effort to cross-check the "zero model" of IT provision for support to simulation, analysis and visualisation, databases, data management, common libraries, tools and base support, the experiments were invited to say what, from their point of view, is needed. This exercise was, of course, fraught with all the dangers that attend efforts to merge partially overlapping sets of needs. Nevertheless, the results are clearly not compatible with a staffing level significantly lower than the "zero model".

6.5.3 Additional demands on CERN for human resources

While recognising that the support CERN can give for software packages must necessarily be resource-limited, the panel sympathised with the idea that, considering the likely technology evolution between now and LHC start-up, there are benefits in maintaining effort for parallel, complementary approaches in a very limited set of cases. Simulation, object persistency and analysis / visualisation are considered to be three such cases.

For simulation, it would be beneficial to be able to make crosschecks between two different Monte-Carlo programs and the panel supported the proposal made by the Software Project panel in Chapter 4.1.2 above that, in addition to ongoing support for Geant4, the role and support of FLUKA should be clarified and agreed. It is estimated that one additional permanent CERN-resident person would be enough to help the FLUKA authors to provide support for the program and to make, in due course, a release of the code. The panel thus proposes that a CERN staff member be allocated to this task.

Concerning the other two areas, the majority of the Software Project panel proposed in Chapter 4.1.6 above that ROOT, which (separately) addresses them both, should receive more support from CERN. It is estimated that its basic support and development would need a minimum of three permanent CERN-resident persons, possibly complemented by visitors from the normal CERN programme. The current development team includes one CERN / IT staff member. The panel proposes that one more CERN staff member be allocated to this task in the "-30 model" for support and two more in the "zero model".

6.6 Recommendations

In summary, the panel made the following recommendations:

1. The panel strongly endorses the proposals made by the other two panels concerning the setting up of a Steering Committee (SC2) for LHC computing. Its mandate should include the co-ordination of the computing system construction activities. Since this system, including the Regional Centres, has to be considered a single worldwide facility, representatives of the Regional Centres must be involved from the beginning in its design and in the construction of the prototypes.
2. The CERN-based facility (Tier0 + Tier1) should be implemented in a single computing system, partitionable amongst the four Collaborations. CERN / IT should have the responsibility for its construction, running, and usage co-ordination under conditions to be specified in the Computing MoU of each Collaboration.
3. The predicted shortfall in human resources for all four Core Software teams is extremely alarming. It is imperative that the Collaborations identify the resources needed to make the Core Software teams efficient. The building of the Core Software infrastructure should be addressed with priority equivalent to that for the construction of a major sub-detector. The necessary human and financial resources must be found within the collaborating institutions. This issue must be tacked urgently by each Collaboration with the solutions arrived at being underwritten on paper (see recommendation 6 below). The Core Software teams must co-ordinate their efforts effectively

with CERN / IT to make sure that their work takes adequate account of resource optimisation over the life-cycle of the experiments.

4. It is essential that CERN / IT Division be adequately staffed to carry out its roles in the Organization. In what concerns support to the LHC physics programme, it is clear that the 2006 staffing level implied by the current CERN-wide staff reduction plan is incompatible with this objective. On the evidence reviewed, it would be courting disaster to reduce the IT human resources significantly below the present level. If despite urgent consideration it still appears that it will not be possible to maintain the human resources from within the CERN staff budget, then other sources must be sought.
5. A CERN prototype facility should be built up progressively towards full functionality, reaching the scale of some 50% of the needs of a single experiment in time to influence the acquisitions for the production system. The prototyping work at CERN must be co-ordinated with similar efforts at other centres so that the distributed aspects can be tested. The four Collaborations should share this prototype facility in common. This work must be launched in 2001, with a proposal being made to the LHCC in order to support funding requests. The project work should be covered by a written agreement signed by the Collaborations, CERN / IT and the main Regional Centres.
6. Interim MoU's or software agreements should be written by each Collaboration soon (by end-2001) to clarify the situation for software development.
7. Computing MoU's, established for each Collaboration on a common pattern, are required well ahead of the main acquisitions (probably in 2003). These must describe the funding and responsibilities for the hardware and the software, along with the human resources committed, and will be signed between CERN and funding agencies. They should also describe, *inter alia*, the policy for access to the computing systems (in particular via the Regional Centres) for all collaborating institutions.

In conclusion, the panel wished to stress that the construction and use of the computing system for LHC analysis is a formidable project and hence a formidable opportunity to build a powerful computing facility and innovative software. The four experiments and CERN / IT must work in collaboration to optimise the cost / performance ratio and make best use of the available human resources.

7 OVERALL CONCLUSIONS AND RECOMMENDATIONS

The review has already had a considerable beneficial effect in that it has served to clarify the issues in the minds of all parties and (re-)establish common views where these were lacking. In several cases, it has endorsed views that were already widespread in the community: Examples are the basing of the software on OO methods, the adoption of a multi-tier distributed computing model and the use of low-cost PC farms as the main workhorses of data processing.

Having frequently debated the work of the Technical Panels and in the light of their reports, the review team has assembled what it feels to be the most important issues to carry forward as the key messages from this review. These are spelled out in the following sections.

An important message must be that the fifteen-year future of LHC is a very long time in the evolution of computing. The maximum effort must be made to remain flexible and modular - plan for change. It will be vital to examine carefully the support implications of any choices of solution.

The review strongly endorses the view that much is to be gained in cost-effectiveness, throughout the entire life cycle of LHC, by executing as much as possible of the development work in common amongst the experiments. In this, they should be aided by CERN / IT Division and the computing support organisations of the other centres that will be contributing to LHC computing. This applies equally to software and to system-related aspects.

While CERN / IT Division must play a central role in this respect, it is also important that the leadership of the Experiments, both at the top level and that of the software and computing projects, recognise their responsibility to staff adequately, on the agreed time-scales, any common effort in which they participate.

7.1 The Computing Model for LHC

The review strongly endorses the adoption of the multi-tier hierarchical distributed model developed by the MONARC[10] (MOdels of Networked Analysis at Regional Centres for LHC experiments) project. In this model, for each experiment, raw data storage and reconstruction will be carried out mainly

---

[10] MONARC Phase 2 report CERN / LCB 2000-001, March 2000
http: / / monarc.web.cern.ch / MONARC / docs / phase2report / Phase2Report.pdf

at a Tier0 centre. Analysis, data storage, Monte-Carlo data generation, some reconstruction and data distribution will be the main tasks of several regional Tier1 centres, followed by a number of (national) Tier2 centres. Below this come (institutional) Tier3 centres and end-user work stations (Tier4). It is important to note that "regional" may here imply a geographical zone *larger* than a single country.

While it fully recognises the important role played by Tiers 3 and 4, the present review has confined its considerations to Tiers 0, 1 and 2 since it felt that the levels below this are not amenable to central management. In terms of the capacity reviewed, it is expected that, for each Collaboration, one third of the total will be located at CERN, one third in the Tier1 centres away from CERN and one third in the Tier2 centres.

The review considers that, in the implementation for LHC, there are several important points regarding the Tiers:

- In addition to the Tier0 centre, there should be a Tier1 centre at CERN. The combination of CERN Tier0+Tier1 should be operated by CERN / IT Division as a common partitionable facility for all LHC experiments, with sharing determined along the lines of past practice (the COCOTIME committee).
- Each Collaboration plans on several (4-6) Tier1 regional centres, some common amongst experiments, away from CERN. These centres, while predominantly serving their geographical regions, should in each case be open to the entire Collaboration. All the Tier1 centres, including CERN, should co-ordinate their efforts at an early stage.
- The foreseen capacity contribution (1/3) at the level of Tier2 is considerable. Special attention must be paid to catering for the needs of Tier2 centres that are not associated with any Tier1 regional centre other than CERN. The limited size of an individual centre at this level may mask the fact that there is a host organisation (university or national lab) with considerable expertise behind it.

The MONARC model goes beyond defining a hierarchy of centres and also comments extensively on the manner of their interconnection. In this respect the present review underlines strongly that its endorsement of the distributed model is underpinned by confidence that the efforts related to the Grid[11]

[11] See e.g. The GRID, Blueprint for a New Computing Infrastructure, I. Foster and C. Kesselmann ed., ISBN 1-55860-475-8

concept will provide the necessary technological infrastructure on the required time-scale. In this respect the HEP-proposed Grid projects, both GriPhyN and PPDG in USA, and the DataGrid[12] Project, recently approved by the European Commission, are expected to make major contributions. Grid activities represent a fortunate opportunity for the HEP community to solve several key LHC computing problems and deserve firm support from all the agencies concerned by LHC computing. The Grid concept may ultimately lead to the entire set of Tier1 and Tier2 centres being seen by the end-user as a single facility.

At a fundamental level, the distributed model depends on good wide-area networking facilities. The whole LHC computing enterprise will depend on the existence of a well-supported, cost-effective high bandwidth Research Networking infrastructure.

7.2 Software

Given the overall goal of producing reliable component software, the review has two major concerns:

- The apparent shortfall of human resources, especially for Core Software development and for support from CERN / IT Division, and the maturity and accuracy of the related resource estimates.
- The need for a well-understood mechanism to guide future development work.

The former is addressed further in 7.4 and the latter in 7.6.

The review is concerned about the development and support of simulation packages and the support and future evolution of analysis tools, as well as the work still to be accomplished in areas such as object persistency, data store and the system design for mass storage and data handling. The policy for establishing the common software base should take into account software already in widespread use in the LHC experiments, be it commercial or specifically developed in the HEP context. In the areas where final choices have yet to be made, support of such existing software must be continued in the meantime. Table A1.2 lists the software in this category that was identified during the review. As described in detail by the Software Project panel in

---

[12] See e.g. "Grid Computing: the European DataGrid Project", proc. IEEE 2000 Nuclear Science Symposium and Medical Imaging Conference, Lyon, 15-20 October 2000. The project also has a Web page at http: / / www.cern.ch / grid

Chapters 4.1.2 and 4.1.6, and by the Management and Resources panel in Chapter 6.5, the review wishes to stress to CERN Management the particular importance of finding satisfactory solutions for the ongoing CERN support of Geant4, Anaphe, object database and the general libraries, along with the introduction of CERN support for FLUKA and ROOT.

Efforts to develop and support software in common amongst the experiments must be pursued vigorously.

Nevertheless, the experiments are in fact very different from each other. The general-purpose experiments produce a huge data rate with a wide variety of interesting physics channels, bringing the special problem of efficient data selection and data access, along with certain bookkeeping challenges. For heavy ions, the event size is enormous but the selection is more straightforward, with many fewer analysis channels. The B-physics requires precision measurements, where excellent simulation is of vital importance, and this may dominate their computing needs. Due to these differences it is clear that the software needs cannot be met with one set of methods and tools for everybody. It would be unwise to start today a new list of tools needed; instead, a process is needed to deal with the issues as they arise. This is detailed in 7.6 below.

7.3 Capacity, Schedule and Costs

The review has compiled extensive sets of information concerning the estimated computing capacity required for Tiers 0-2 and has, with the aid of costing models elaborated by the PASTA committee (see footnote 8 on page 31), derived the likely associated investment and operating costs. In this it was also guided by independent estimations from other HEP computing centres around the world. The detailed results are presented in Appendix 3. The global investment cost of some 240 MCHF for the full initial deployment considered in this report underlines the need for all parties, including funding agencies, to pay close attention to resourcing correctly this aspect of the LHC, essential for success of the overall programme.

The estimates given are the best that can be made now (five years before turn-on). They are subject to several sources of uncertainty and must be refined as time goes by. Key elements here are:

- The initial performance of the LHC machine, influencing event rates and annual data volume.

- The trigger and data reduction efficiencies of the experiments, along with their detector performance, background levels and physics interests at the time.
- The achieved performance and reliability of the software.
- The actual, as opposed to predicted, price-performance evolution of computer hardware and wide-area networking. Closely related to this is the actual schedule for LHC start-up, which determines the time at which the elements must be purchased.

In connection with the last bullet it must be stressed that a three-year acquisition profile, centred on the start-up year of LHC, is proposed. It is primarily motivated by practical considerations and the need to work within a realistic funding profile. In the first running year, however, there will be pressure to relax trigger conditions and to reprocess data many times, which will place heavy demands on the data processing system. Additionally, LHCb expects to receive nominal luminosity shortly after start-up and certainly within the first year.

Concerning the sharing of investment costs between CERN and other agencies, the review remarks that the planned fraction of Tier0-Tier2 capacity to be installed at CERN is one third, corresponding to established past practice for the funding of computing by CERN. It recommends that this past practice be broadly continued, while noting that the detailed decisions will be for negotiation in the writing of the computing Memoranda of Understanding (see 7.5 below).

The maintenance and operation (M&O) model adopted by the review includes the strategy of rolling replacement. Given the size of the initial investment and the need for subsequent capacity upgrades to cope with improvements in machine and experiment performance, the amounts of resources required under this heading will also be substantial. Concerning capacity growth, it should be noted that the computing capacity available to the LEP experiments increased by a factor of about 1'000 over their lifetime, within an essentially constant budget. While recognising that attitudes to replacement / upgrade may well differ across the various participating sites, the review considers that the model proposed for the CERN site (Appendix 3.5.3.2) is reasonable. Under these conditions, for the foreseeable future the community will have to spend within each three-year operating period an amount roughly equal to the initial investment. This covers all materials aspects, including storage for each year's data, but does not include human resources.

To place the computing M&O expenditure in context, the review wishes to stress that the sum quoted bears broadly the same relationship to the M&O costs of the detectors themselves as was the case for previous generations of experiments.

In the past, CERN has charged tapes to the users but has covered all other M&O costs of off-line computing itself. In view of the plans to handle on-site mass storage in an integrated way, a better scheme for the future, at least for the automated tape, will be to replace the tape charging by a mass storage levy calculated by a formula to be determined.

7.4 Human Resources

The Management and Resources panel drew together the information on needed human resources assembled by the other two panels and compared these with the currently foreseen reality. There is inevitably some uncertainty in these numbers at the detailed level. For example, the Collaborations were not all using exactly the same definition of tasks and, in the case of CERN / IT, were giving estimations of human resources for their own experiment's needs, without reference to the internal organisation of that unit. Nevertheless, the review considers the picture that emerges (cf. Chapter 6.5) to be extremely worrying.

For Core Software development (taking together the figures for the four experiments) the available human resources in 2000 (64 m-y) were just 70% of the level deemed to be necessary (92 m-y) that year and only 60% of the sustained level needed over the coming four years (about 105 FTEs). CERN and the Collaborations together must do all that they can to provide the human resources that are needed.

On the side of the Collaborations, as was pointed out by the Software panel (Chapter 4.1.1) and underlined by the Resources panel in their Recommendation 3 (Chapter 6.6), their leadership must look again at overall needs for human resources and try harder to direct sufficient resources from their institutes, including CERN as a collaborating institute, to software and computing issues. This is justified by the increasingly important role of these issues, and the growing complexity of building software systems and deploying them worldwide.

In what concerns CERN, it is primarily IT Division that is implicated and here the worries of the review extend to all aspects of computing support. The IT staff level foreseen in the current CERN planning (a reduction of 50 by 2006) falls far short of what is required to support the lab's computing infrastructure and the LHC programme. The slightly higher level proposed by CERN to Council (equivalent to a reduction of 30 by 2006) was also considered by the Management and Resources panel to be insufficient. Details on this point are contained in Chapter 6.5.2-6.5.3, and in the associated Recommendation 4 of the panel. The review wishes to emphasise most strongly that it endorses the recommendation that CERN human resources for computing support should not be reduced below its present level. Staff effort for LHC that becomes available as a result of the present plan being changed should be allocated to IT and / or the experiments, taking into account the recommendations of the committee proposed in 7.6 below.

It is estimated that the major multi-experiment Tier1 centres away from CERN will each require total human resources of about 40 FTEs for their operation.

In view of the uncertainties alluded to at the start of this section and the comments of the technical panels regarding the maturity of the estimates, the review thinks it especially important that the human resource issues for LHC computing be followed up urgently, using the mechanism proposed in 7.6 below.

7.5 Definition of Goals and Responsibilities

As with the other aspects of detector construction, the provision of computing for the LHC experiments must be the subject of formal agreements amongst the parties concerned. For this purpose, three aspects of the work can be distinguished: development and integration of software and system elements; provision of a prototype facility for test purposes (see 7.7 below); provision and operation of the production systems. For each there is an optimum time at which the formal agreements should be made.

The review has identified a severe potential lack of human resources for software and a lack of maturity in the associated planning. In order at least to cover the development work, it is thus necessary for the Collaborations to define rapidly interim understandings on responsibilities, human resources and sometimes money (ATLAS is in the process of preparing "software agreements"). Such documents are sometimes referred to as interim Memoranda of Understanding (IMoUs).

It will also be important to write in 2001 an agreement amongst the four experiments, some of the regional centres and CERN / IT Division that covers construction of the distributed computing prototype. This agreement should specify costs, goals, technical solution and organisation and is a necessary prerequisite to asking support from the appropriate funding agencies.

By 2003 a much clearer picture should have emerged concerning the parameters of the production systems and this will be the moment to write the formal MoUs covering all aspects of their provision and operation, including hardware and software.

A necessary prerequisite for the formal MoUs will be the submission to the LHC Committee of the Collaborations' Computing Technical Design Reports (TDRs), as already foreseen.

7.6 Guidance of LHC Computing Activities in the Future

This review gives only a snapshot of the issues involved in LHC computing. Given the complexity of the endeavour and the remaining uncertainties, the review considers it essential that a formal mechanism for ongoing monitoring and guidance is set up immediately.

The review recommends that an LHC Software and Computing Steering Committee (SC2) should be convened by the CERN Director responsible for Scientific Computing. Its membership should comprise the highest level software and computing management in the experiments and CERN / IT Division, as well as including representation of the Tier1 centres. Its role will be to advise the Director and the leadership of the Collaborations on all aspects (technical and resources) of the development work for both computing and networking, and to establish, oversee and terminate common projects as thought necessary.

The review identified early and important tasks for the SC2:

- The committee should function with the aid of focused Technical Assessment Groups (TAGs) that it will create to look into well-defined issues. In the short term, it is important that such a TAG be put in place for Data Management as recommended by the Software Project panel (Chapter 4.3 Recommendation 2).

- The committee should closely track the evolution of requirements, given the uncertainties that were alluded to in 7.3 above.
- Another early task is to validate further the needed human resources, and to track and report regularly on the correspondence between needs and reality.
- The committee should be instrumental in ensuring coherence in the structure and provisions of the Computing MoUs that will be written.
- Concerning the common projects that the SC2 oversees, it is vital for them to be assigned full-time leaders to oversee their entire scope, to drive the various activities and to ensure cohesion amongst them. Those working on the projects should report to the project management.
- Common projects that should be launched urgently:
  - Construction of prototype LHC computing facility (see 7.7 below)
  - Development of LHC common software items, along with support for their maintenance, as proposed by the Software Project panel (cf. Chapter 4.2).

7.7 Data Challenges and Prototyping

All the Collaborations plan a sequence of progressively more complex Data Challenges to test their software systems in the years before LHC start-up. This process is essential and is strongly endorsed by the review.

In what concerns the computing facilities themselves, a clear path from the present situation to the final systems is needed. A common project leading to a realistic common prototype is highly desirable. Its main goals should be to:

- Provide a clean test for the different software and technical options;
- Follow the market and the evolution of technology;
- Provide a convincing prototype before the production investment;
- Be the common test set-up where all the experiments may share resources and solutions as much as reasonably possible.

This prototype should be installed progressively over three years to reach by 2004 about 50% of the CERN-based computing and data handling capacity needed by one LHC experiment. Most of the capacity should be added in the last year in order to minimise the cost. The results of this prototyping exercise will then be available to guide the main acquisitions starting in 2005.

The prototype should make use of the middleware developed and tested by the Grid projects (DataGrid time-scale is 2001-3). It must involve most elements of the final distributed computing systems. This implies installing Tier0+Tier1 hardware at CERN and also proving the inter-working with Tier1 and Tier2 centres elsewhere. The prototype should be set up and exploited as a common facility by CERN / IT Division, the four experiments and existing Tier1 centres. These partners must plan the detailed configuration to address all the critical aspects:

- Complexity (number of CPU boxes, etc.);
- Computing power;
- Data access (hierarchy, data locality, etc.);
- Heterogeneity across the different centres involved;
- Management issues;
- Test of the full chain, including the main physics triggers.

A cost estimate for the CERN part of such a prototype is shown in Appendix 3.5.5.

In view of the strong interdependencies, it is highly desirable that MONARC Phase 3 and the various Grid initiatives co-ordinate their efforts effectively with the common prototype project described above.

Close monitoring of the experience being gained by other experiments with large computing needs (such as BaBar, D0 and CDF) will be a necessary complement to the prototype work.

7.8 Heterogeneity

CERN / IT Division is planning the Tier0+Tier1 centre at CERN within a model where the only sources of heterogeneity are the different software configurations required by the different experiments and the continuously changing technology standards that will force different hardware to coexist even in the same installation. The review supports this view of limited heterogeneity, which should only be relaxed if the LHC Collaborations demonstrate, e.g. as a result of the prototype work, that their computing models are incompatible with this idea.

ANNEX 1

MANDATE OF THE REVIEW

Review of the progress and planning of the computing efforts at CERN and of the LHC-experiments for the LHC start-up

Purpose and Mandate

At about mid-term between the publication of the technical proposals of the experiments and the start-up of LHC and well before the submission of the computing TDRs of the experiments it is an appropriate moment to review the computing plans of the LHC experiments and the corresponding preparations of IT Division with the following aims:

a) Update the assessment of individual, experiment specific, regional and CERN facilities and their relative roles required to perform the computing of the LHC experiments and update the estimates of the corresponding resources required, taking into account the evolution of the underlying technologies. Identify the software packages to operate in the various places. Identify services to be provided centrally. Identify activities which have to be done at CERN.

b) Assess the analysis software projects and their organisational structures and interfaces between parts, the corresponding role of CERN and possible common efforts.

c) Review and comment about the overall and individual computing project management structures and review the resources required.

The outcome of the assessment will help CERN and the experiments to formulate resource loaded work-plans with objectives, detailed schedules and milestones between now and LHC start to be described finally in their computing TDRs.

The results of the review will be the basis for CERN, the collaborating institutes and their funding agencies, for the formulation of Computing Memoranda of Understanding, which will describe the commitments of institutes inside the collaborations towards their computing goals and commitments of CERN to provide computing infrastructure and central facilities for the LHC era. The MoUs should be put in place in 2001.

The review team should recommend actions and in particular common actions between experiments and IT Division that will help to achieve these goals within the existing resources.

The review reports to the Research Board and the Director General. Interim status reports will be given to FOCUS, LHCC and other appropriate bodies.

ANNEX 2

MEMBERSHIP OF THE REVIEW

Steering Group

Members:	S. Bethke (MPI Munich)	Chair	
	H.F. Hoffmann (CERN)	CERN Director for Sc. Computing	
	D. Jacobs (CERN)	Secretary	
	M. Calvetti (INFN Florence)	Chair of the Mgmt and Resources Panel	
	M. Kasemann (FNAL)	Chair of the Software Project Panel	
	D. Linglin (CC-IN2P3 / CNRS)	Chair of the Computing Model Panel	
In Attendance:		**Representative**	**Alternate**
	IT Division	M. Delfino (CERN)	L. Robertson (CERN)
	ALICE	F. Carminati (CERN)	K. Safarik (CERN)
	ATLAS	N. McCubbin (RAL)	G. Poulard (CERN)
	CMS	M. Pimia (CERN)	H. Newman (CALTECH)
	LHCb	J. Harvey (CERN)	M. Cattaneo (CERN)
Observers:	R. Cashmore (CERN)	CERN Director for collider programmes	
	J. Engelen (NIKHEF)	LHCC chairman	

Worldwide Analysis / Computing Model panel

	D. Linglin (CC-IN2P3 / CNRS)	Chair	
	F. Gagliardi (CERN)	Secretary	
Expt. Reps:		**Representative**	**Alternate**
	ALICE	A. Masoni (INFN Rome)	A. Sandoval (GSI Darmstadt)
	ATLAS	A. Putzer (U. Heidelberg)	L. Perini (U. Milan)
	CMS	H. Newman (CALTECH)	W. Jank (CERN)
	LHCb	F. Harris (U. Oxford)	M. Schmelling (MPI Heidelberg)
Experts:	Y. Morita (KEK)	C. Michau (UREC-STIC / CNRS)	

Software Project panel

	M. Kasemann (FNAL)	Chair	
	A. Pfeiffer (CERN)	Secretary and CERN-IT representative	
Expt. Reps:		**Representative**	**Alternate**
	ALICE	R. Brun (CERN)	A. Morsch (CERN)
	ATLAS	D. Barberis (U. Genoa)	M. Bosman (U.A. Barcelona)
	CMS	L. Taylor (Northeastern U.)	T. Todorov (IN2P3 Strasbourg)
	LHCb	P. Mato (CERN)	O. Callot (LAL Orsay)
Experts:	V. White (FNAL)		

Management and Resources panel

	M. Calvetti (INFN Florence)	Chair	
	M. Lamanna (INFN Trieste and CERN)	Secretary	
Expt. Reps:		**Representative**	**Alternate**
	ALICE	P. Vande Vyvre (CERN)	K. Safarik (CERN)
	ATLAS	J. Huth (U. Harvard)	H. Meinhard (CERN)
	CMS	P. Capiluppi (INFN Bologna)	I. Willers (CERN)
	LHCb	J. Harvey (CERN)	J.P. Dufey (CERN)
Experts:	F. Étienne (IN2P3 Marseilles)	J. Gordon (RAL)	L. Robertson (CERN)
	F. Ruggieri (INFN Bari)	T. Wenaus (BNL)	K. Woller (DESY)
	G. Wormser (IN2P3 Paris)		

ANNEX 3

MEETINGS OF THE REVIEW

3.1 Steering Group

8 Dec 1999	Initial discussions with panel chairpersons on the Review format
24 Jan 2000	Founding meeting. Discussion of panel work-plans
22 Feb 2000	Panel progress reports. Lessons from DOE review, CHEP 2000 and the Grid Initiative meeting in Padua.
27 Mar 2000	Panel progress reports.
19-20 Apr 2000	Detailed progress reports from the panels.
29 May 2000	Progress of Software and Model panels. Agreement that Resources panel will mainly work in the light of their reports.
13 Jun 2000	Preliminary reports from the Software and Model panels.
4 Jul 2000	New draft report from Software panel. Further discussion of Model panel preliminary report. Assembly of missing points.
18 Jul 2000	Panel progress discussion. First draft conclusions from Resources panel.
29 Aug 2000	Towards finalisation of Software panel report. Plans of the Model and Resources panels. Proposal for LHC Software and Computing Steering Committee (SC2).
13 Sep 2000	Updates on finalisation of the Software and Model panel reports. Plans of the Resources panel to report. Further discussion on the proposed SC2.
26 Sep 2000	Presentation of final Software panel report. Updates from the other two panels.
3 Oct 2000	Discussion towards finalisation of the review and on draft proposal for the SC2.
6 Oct 2000	Core Steering Group: Preparation of oral RRB presentation.
12 Oct 2000	Core Steering Group: Review of material for RRB presentation.
24 Oct 2000	Core Steering Group: Plans for convergence to final report.
16 Nov 2000	Progress of the Model and Resources panels. Final report skeleton.
6 Dec 2001	Finalisation of data from the Model panel (and elsewhere) to be used by the Resources panel in reporting. Establishment of final reporting schedule.
18 Jan 2001	Final input to the draft Model and Resources panel reports.
31 Jan 2001	Discussion and editing of the draft final review report.
14 Feb 2001	Resolution of last open points, final input from those in attendance.
22 Feb 2001	Finalisation.

3.2 Software Project panel

27 Jan 2000	Questions to experiments and CERN / IT Division
13-17 Mar 2000	Presentations to panel
28 Mar 2000	Closed session panel meeting by video
4 Apr 2000	Closed session panel meeting by video
12 / 30 Apr 2000	RFI to experiments / Written responses
3 May 2000	Closed meeting with experiments
18 May 2000	Closed meeting with experiments
24 May 2000	*Draft 1 of panel report (panel internal)*
26 May 2000	Closed panel session and CERN / IT Division meeting
28 May 2000	Closed panel session and CERN / IT Division meeting
8 Jun 2000	*Draft 2 of panel report (panel internal)*
29 Jun 2000	Panel 'phone meeting
30 Jun 2000	*Draft 3 of panel report – to experiments and Steering Group*
5 Jul 2000	Closed panel session

7 Jul 2000	Meetings on Draft 3 with each experiment and CERN / IT Division. *Panel agrees to restructure its report*
15 Jul 2000	*Written Draft 3 re-wording from panel members*
19-20 Jul 2000	Panel meeting & Draft 4 (restructured)
30 Jul 2000	*Draft 5 of panel report available*
24-25 Aug 2000	Panel meeting. *Draft 6.1 of panel report available*
29 Aug 2000	Panel meeting
12 Sep 2000	Panel meeting. *Draft 6.2 of panel report available*
14 Sep 2000	*Draft 6.3 of panel report available*
21 Sep 2000	*Draft 6.4 of panel report available*
2 Oct 2000	Final panel report sent to Steering Group

3.3 Worldwide Analysis / Computing Model panel

22 Mar 2000	Separate meetings with ALICE and with CERN / IT Division.
23 Mar 2000	Separate meetings with LHCb and with ATLAS.
24 Mar 2000	Meeting with CMS.
12 May 2000	Combined meeting with all four experiments and CERN / IT, along with the Resources panel.
27 Nov 2000	Meeting with all four experiments.
5 Feb 2001	Final panel report sent to Steering Group.

3.4 Management and Resources panel

16 Feb 2000	Separate presentations from the representatives of ALICE and LHCb (together) and from those of ATLAS.
17 Feb 2000	Presentations from ATLAS and CMS representatives (together).
24 Mar 2000	Reports from the other panels and all experiment representatives.
14 Apr 2000	Consideration of desirable content of the computing MoUs.
12 May 2000	Combined meeting with all four experiments and CERN / IT, along with the Model panel.
3 Jul 2000	Further presentations from the experiment representatives.
4 Jul 2000	Core panel meeting to consider the input received.
17 Jul 2000	Further presentations from the experiment and IT representatives.
18 Jul 2000	Core panel meeting to consider the further input.
22 Aug 2000	Consideration of first report draft and plans for further work.
20 Sep 2000	Further input from ATLAS. INFN plans for a Tier1 centre.
14-15 Nov 2000	Further editing of the draft panel report.
8 Dec 2000	Refinement of the draft panel report.
9 Jan 2000	Further refinement of the draft panel report.
26 Jan 2000	More work on the draft panel report.
31 Jan 2000	Finalisation of the panel report.
21 Feb 2001	Final panel report sent to Steering Group.

APPENDIX 1

KEY SOFTWARE PARAMETERS

TABLE A1.1

Major experiment software milestones (DC = Data Challenge)

Date	ALICE	ATLAS	CMS	LHCb
Feb-99		Internal Computing Rev. and "Action Plan"		
Apr-99	DC 1 - 10 TB			
May-99		Physics TDR using Fortran/ Zebra/GEANT3		
Jun-99			Fill OODBMS at 100 MB/s	
Mar-00	DC 2 - 100 MB/s + LV3. 30 TB			DC 0 - 3 M evts/y in 2000 2001
Jun-00	Full detailed simulation & full sub-detector recon.	Framework prototype ready	Integration of OODBMS and MSS	
Jul-00	DC 2 - LV4 added			
Sep-00		1st resource loaded WBS		
Oct-00			Prototype user analysis environment	
Nov-00	Full sub-detector recon. & 1st resource loaded WBS	Integration of OODBMS and MSS		
Dec-00			Simulation of data access pattern. DC 0 - 5TB (0.5%)	
Mar-01	DC 3 - 80 TB			
Apr-01		Physics wkshp - G3 data analyzed with new sw		
Jun-01	Physics performance report		Fully functional GEANT4 simulation of CMS, ORCA prodn for HLT studies, 10 M events, 20 TB out	
Dec-01		TDAQ TDR - event filter using new OO sw +framework. DC 0 - 100GB	Fully functional reconstr./ analysis framework, choice of vendor for OODBMS, data acquisition TDR	
Jun-02	5%*3PB Data Challenge in 2002 (DC 4)		Fully functional detector reconstruction	
Jul-02		DC 1-1TB (0.1%) completed		Fully functional OO sw - Recon/Sim/Analysis, DC 1 - 6 M events
Dec-02	Computing TDR	Computing TDR	Computing TDR, install ODBMS and MSS, fully functional user analysis env., fully functional physics object recon., DC 1 - 50TB (5%)	Computing TDR.
Jul-03				DC 2 - 6 M events
Sep-03	10% DC in 2003 (DC 5)	DC 2 - 100TB (10%) completed		
Nov-03	Decision on MSS/HSM			
Dec-03			Production SW: GEANT4 sim. of CMS, recon./analysis framework. Physics TDR	
Jun-04	25% DC in 2004 (DC 6)	Physics Readiness Report		
Dec-04			Prodn sw: detector recon., physics object recon., user analysis environment. DC 2 - 200TB (20%)	Prodn readiness review for data processing software
2005	50% DC in 2005 (DC 7)			DC 3 - 10 M events

TABLE A1.2

Common software products in use by experiments (excluding software development and methodology tools)

					X=yes			*used by*	
Project/Product		ALICE	ATLAS	LHCb	CMS	IT	as	HEP outside LHC	non-HEP
GEANT4	**Detector Simulation**	**X**	**X**		**X**	**X**	**developer**	**X**	**X**
	package written by GEANT		**X**	**X**	**X**	**X**	**maintainer**	**X**	**X**
	collaboration	**X**	**X**	**X**	**X**		**user**	**X**	**X**
GEANT3	**Detector Simulation**						**developer**	**X**	
	package written in Fortran	**X**			**X**	**X**	**maintainer**	**X**	
		X	**X**	**X**	**X**		**user**	**X**	
FLUKA	**MC for radiation studies**						**developer**	**X**	
							maintainer	**X**	
		X	**X**	**X**	**X**		**user**	**X**	
Event Generators	**Many - including Pythia,**						**developer**	**X**	
	Herwig, QQ,etc.						**maintainer**	**X**	
		X	**X**	**X**	**X**		**user**	**X**	
Objectivity DB +	**Commercial Object**				**X**	**X**	**"developer"**	**X**	**X**
tools/knowledge	**Database**				**X**	**X**	**administrator**	**X**	**X**
	tools+knowledge		**X**		**X**	**X**	**user**	**X**	**X**
ROOT	**ROOT objects streamed**	**X**					**developer**	**X**	**X**
persistency,CINT	**to files for either data**	**X**					**maintainer**	**X**	**X**
and file format	**or conditions**	**X**		**X**			**user**	**X**	**X**
Mass Storage	**HPSS, Castor, other…**					**Castor**	**developer**		
System						**X**	**administrator**		
		X	**X**	**X**	**X**	**X**	**user**		
Relational DB	**ORACLE or MySQL**						**developer**	**X**	**X**
for data handling						**X**	**administrator**	**X**	**X**
		X		**X**			**user**	**X**	**X**
ANAPHE	**Replacement for CERNLIB**				**X**	**X**	**developer**	**X**	
	several commercial and					**X**	**maintainer**	**X**	
	many HEP packages		**X**	**X**	**X**		**user**	**X**	
ROOT	**PAW replacement**	**X**					**developer**	**X**	**X**
as an Analysis	**OO analysis tool**	**X**					**maintainer**	**X**	**X**
Tool		**X**	**X**	**X**	**X**		**user**	**X**	**X**
GAUDI	**Basis of Frameworks for**		**X**	**X**			**developer**		
	ATLAS and LHCb		**X**	**X**			**maintainer**		
			X	**X**			**user**	**X**	
CERNLIB/PAW	**Fortran based libs**						**developer**	**X**	
						X	**maintainer**		
		X	**X**	**X**	**X**	**X**	**user**	**X**	**X**
WIRED	**Event Display**					**X**	**developer**	**X**	
						X	**maintainer**	**X**	
			X	**X**			**user**	**X**	
Conditions DB	**evolved from BaBar**					**X**	**developer**		
	Objectivity conditions db					**X**	**maintainer**		
			X	**X**			**user**	**X**	
XML Parser	**Open source product**						**developer**		**X**
							maintainer		**X**
			X	**X**	**X**	**X**	**user**		**X**

TABLE A1.3

Not-in-common software products in use by experiments

					X=yes			*used by*	
Project/Product		ALICE	ATLAS	LHCb	CMS	IT	as	HEP outside LHC	non-HEP
ATLAS code	**Commercial tool**					**X**	**developer**		
checker	**(Code Wizard)**					**X**	**maintainer**		
	customized by IT		**X**				**user**		
ALICE Code	**developed externally**						**developer**		**X**
checker							**maintainer**		**X**
		X					**user**		
DepUty	**Dependency code**				**X**		**developer**		
	checker				**X**		**maintainer**		
					X		**user**		
StyUty	**Style violation code**				**X**		**developer**		
	checker				**X**		**administrator**		
					X		**user**		
CMT	**Configuration**						**developer**	**X**	
	management						**maintainer**	**X**	
	builds and releases			**X**			**user**	**X**	
SCRAM	**Configuration**				**X**		**developer**		
	management				**X**		**maintainer**		
	builds and releases				**X**	**X**	**user**		
SRT	**Configuration**		**X**				**developer**		
	management		**X**				**maintainer**		
	builds and releases		**X**				**user**		
CARF	**CMS Framework**				**X**		**developer**		
					X		**maintainer**		
					X		**user**		
GAUDI	**LHCb-specific additions**			**X**			**developer**		
	to GAUDI Framework			**X**			**maintainer**		
				X			**user**		
Athena	**ATLAS-specific additions**		**X**				**developer**		
	to GAUDI Framework		**X**				**maintainer**		
			X				**user**		
AliROOT	**ALICE framework**	**X**					**developer**		
	based on ROOT	**X**					**maintainer**		
		X					**user**	**X**	
PROOF	**parallel distributed data**	**X**					**developer**	**X**	
	processing	**X**					**maintainer**	**X**	
		X					**user**	**X**	
IGUANA	**Interactive**				**X**		**developer**		
	visualisation				**X**		**maintainer**		
	toolkit				**X**		**user**		
CVS-pm	**Access management**				**X**		**developer**		
	and code organization				**X**		**maintainer**		
	on top of CVS				**X**		**user**		

APPENDIX 2

KEY SYSTEM PARAMETERS

The key numbers and computing needs of the four large LHC experiments are summarised in Table A2.1. This table is normalised to nominal running periods of 10^7 seconds / year. Here follow additional explanations and comments to aid the interpretation of this table.

LHC design parameters

E = 14 TeV (two 7 TeV proton beams)
L = 10^{34} cm^{-2} sec^{-1} (design luminosity)
σ = 100 mb = 10^{-25} cm^2
Collision rate = L. σ = 10^9 Hz p-p collisions
4 large experiments approved: ALICE, ATLAS, CMS and LHCb.
CMS will host the TOTEM dedicated experiment (to run at low luminosity)

Approximate numbers to remember (per experiment):

- 10^6 1 MB = size of a recorded p-p event (up to 40MB for a Pb-Pb event in ALICE)
- 10^2 Data taking rate (Hz), down from 10^9 Hz p-p collisions, after several trigger levels.
- 10^7 recorded p-p events per day (out of 10^{14})
- 10^7 data taking seconds / y, or ~116 days (except Ion runs for ALICE ~15 days / y). To ease comparisons, this number has been fixed equal for all experiments (p-p collisions)
- 10^9 recorded p-p events per year

Tiers

A first order definition of tier functionalities is given in Chapter 5.1.1. Each experiment plans of the order of 5 Tier1's. In reality, the boundaries between tier levels will become blurred. Although CMS is the only experiment to give explicit estimates for Tier2 (see Table A2.1), all other experiments have taken account of their contributions in giving the numbers labelled as Tier1. The dividing line between Tier1 and Tier2 will clearly differ greatly in different countries and, very likely, in different experiments. It is out of the scope of this review to attempt a country-by-country analysis in this respect. Several Tier1's will indeed serve several or all LHC experiments, in addition to other experiments. A qualitative distinction, however, is the open nature of Tier1's to all members of a Collaboration. Also, following the Grid concept, it is expected that a Tier1, especially that at CERN, would store data that are nowhere else. This is already a key element of the BaBar "TierA" definition.

Types of data and volumes of data sets

Tiers of data corresponding to analysis process milestones:
RAW Real Raw Data, as recorded after the on-line Farm

SIM	Simulated Raw Data
ESD or REC	Event Summary Data (after production reconstruction)
AOD	physics Analysis Object Data (after production analysis)
DPD	Derived Physics Data (similar to today's N-tuples) (after production or individual analysis)
TAG	Event tags

RAW / SIM ⇨ ESD / REC ⇨ AOD ⇨ DPD

Experiments would like to back up RAW and REC, as those sets will not be fully replicated. This will depend on availability and cost of *ad hoc* archival media. In general, backup is not excluded if the experiment can afford it. Average sizes for these data sets are given in Table A2.1.

As a general principle, Raw Data will stay at CERN, except for *ad hoc* selected samples to be replicated at Tier1 (5-10%). All other data sets will be exported or exchanged between Tier1 and Tier0, partially also with lower level tiers.

Luminosity and Detector Calibration

Experiments have estimated the storage needs for these tasks as being small compared to the total, and so the strategy would have a small relative impact on resources. This explains why this point has not yet been thoroughly studied. As a first approximation, experiments envisage doing this work at CERN.

Storage per experiment:

3 to 10 PB on tape	Total ~28 PB (with 2 / 3 more per year beyond) Raw Data storage is ~ 1 / 3 of this total.
1 to 6 PB of disk	Total ~11 PB (with 1 / 3 more per year beyond)

CPU (off-line) per experiment:

Best guesses today range from ~1 M SI-95 in LHCb to ~2 M SI-95 for each of ALICE, ATLAS and CMS. But all experiments agree that uncertainties are at least a factor 2. As explained in Chapter 5.1.4, estimates are in fact the sum of Tier0, Tier1 and Tier2.

Numbers in Table A2.1 are based on the estimated number of recorded events and the reconstruction time per event of the present software. However, as most of the CPU will be used in the physics analysis stage and not by the reconstruction stage, estimates of the total CPU needed would depend on elements such as physics, sociology and organisation, that can hardly be deduced from a simple formula.

The LHC Computing Review has triggered further studies and estimates by all experiments and, as a consequence, figures for storage and CPU have evolved during the review, especially in experiments where only rough estimates had been done previously. Some numbers are up by 50% compared to previous estimates in early 2000 and by more than a factor of two when compared to figures in the Computing Technical Proposals. However, in the time since the proper evaluation by experiments to produce the first released version of Table

A2.1 in summer 2000, total estimates for CPU, disk and tape storage have not changed by more than 20%, which is well within the error bars.

LHCb is the only experiment where Monte-Carlo simulation will consume most of the resources (~80%) and the Collaboration plans to keep its MC data at the MC production centres.

Data Challenges

Experiments are considering several types and several levels of Data Challenges (DC) for testing software and computing models, as well as performing scalability tests for simulation, reconstruction and analysis. ALICE and CMS have already done quite significant DCs. Restricting DCs to those involving at least some Tier1's (in view of the 100% = 1PB Real Data Challenge) yields the following plans:

		Size	Date
ALICE	DC1	10 TB	Spring 1999
	DC2	30 TB	Spring 2000
	DC3	80 TB	Spring 2001
	DC4	5%*3PB	2002
	DC5	10%	2003
	DC6	25%	2004
	DC7	50%	2005
ATLAS	DC0	100 GB (0.01%)	end 2001
	DC1	1 TB (0.1%)	2002
	DC2	100 TB (10%)	2003
	DC3	...	
CMS	DC0	5 TB (0.5%)	Dec 2000
	DC1	50 TB (5%)	Dec 2002
	DC2	200 TB (20%)	Dec 2004
LHCb	DC0	$3*10^6$ evts / y	2000-2001
	DC1+2	$6*10^6$ evts / y	2002-2003
	DC3	10^7 evts / y	2004-2005

ALICE also plans three Physics Data Challenges, to test physics analysis tools with simulated data. ALICE Data Challenges from spring 2001 will involve distributed tiers.

TABLE A2.1

Computing Resources[13] planned by the four LHC Experiments in 2007 (*)

Parameter	Unit	ALICE p-p	ALICE Pb-Pb	ATLAS	CMS	LHCb	TOTAL	ATLAS (**)
# assumed Tier1 not at CERN		4		6	5	5		6
# assumed Tier2 not at CERN***					25			
Event recording rate	Hz	100	50	100	100	200		270
RAW Event size	MB	1	25	1	1	0.125		2
REC/ESD Event size	MB	0.1	2.5	0.5	0.5	0.1		0.5
AOD Event size	kB	10	250	10	10	20		10
TAG Event size	kB	1	10	0.1	1	1		0.1
Running time per year	M seconds	10	1	10	10	10		10
Events/year	Giga	1	0.05	1	1	2		2.7
Storage for real data	PB	1.2	1.5	2.0	1.7	0.45	6.9	8.1
RAW SIM Event size	MB	0.5	600	2	2	0.2		2
REC/ESD SIM Event size	MB	0.1	5	0.5	0.4	0.1		0.5
Events SIM/year	Giga	0.1	0.0001	0.12	0.5	1.2		0.12
Number of reconst. passes	Nb	2		2-3	2	2-3		2-3
Storage for simul. data	PB	0.1	0.1	1.5	1.2	0.36	3.2	1.5
Storage for calibration	PB	0.0	0.0	0.4	0.01	0.01	0.4	0.4
Tape storage at CERN T0+T1	PB (10**15 B)	3.23		2.86	4.17	1.22	11.5	9.00
Tape storage at each Tier1 (Avg.)		}0.37		}1.26	1.02	}0.32	}3.0	}1.80
Tape storage at each Tier2 (Avg.)***					0.05			
Total tape storage / year		**4.7**		**10.4**	**10.5**	**2.8**	**28.5**	**19.8**
Disk storage at CERN T0+T1	PB	0.53		0.31	1.14	0.33	2.3	0.41
Disk storage at each Tier1 (Avg.)		}0.27		}0.26	0.44	}0.15	}1.1	}0.36
Disk storage at each Tier2 (Avg.)***					0.10			
Total disk storage		**1.6**		**1.9**	**5.9**	**1.1**	**10.4**	**2.57**
Time to reconstruct 1 event	k SI-95 sec	0.4	100	0.64	3	0.25		0.64
Time to simulate 1 event	k SI-95 sec	3	2250	3	5	1.5		3
CPU for 1 rec. pass/y (real data)	k SI-95	20	250	200	434	50		385
CPU for 1 SIM pass/y (sim+rec)	k SI-95	19	269	30	200	660		30
CPU reconstruction, calib.	k SI-95	65	525	251	1040	50	1931	435
CPU simulation		19	269	30	587	660	1564	30
CPU analysis		880		1479	1280	215	3854	1479
Total CPU at CERN T0+T1	k SI-95	824		506	820	225	2375	690
Total CPU each Tier1 (Avg.)		}234		}209	204	}140	}787	}209
Total CPU each Tier2 (Avg.)***					43			
Total CPU		**1758**		**1760**	**2907**	**925**	**7349**	**1944**
WAN, Bandwidths								
Tier0 - Tier1 link, 1 expt.	Mbps	1500		1500	1500	310	4810	1500
Tier1 - Tier2 link		622		622	622			622

(*) or the first full year with design luminosity

(**) further estimates envisaged by ATLAS, see Chapter 5.1.2.1 for details.

(***) for all except CMS, the Tier1 and Tier2 needs are merged together.

[13] Including realistic usage efficiency factors

APPENDIX 3

MANAGEMENT AND RESOURCES

3.1 Computing needs of the experiments

The LHC experiments have described their computing plans in terms of CPU power, storage capacity, human resources, organisation and schedule.

It should be noted that, in what follows, only the off-line computing infrastructure downstream of the event-filter farm is considered. Furthermore, the predicted necessary computing resources have been estimated using as key input the actual measured performance of prototype LHC analysis, reconstruction and simulation programs executed on existing computer systems.

The needs quoted correspond to data-taking in the first year (10^7 s) of running time with a luminosity of 10^{33} $cm^{-2}s^{-1}$ with protons, plus one month of Ion running at 10^{27} $cm^{-2}s^{-1}$.

It is at once obvious that the requirements in terms of CPU, storage capacity, disk space, complexity, etc. are all many times bigger than what is available to any existing experiments. Beyond this, it has emerged that, while the event sizes are reasonably well known already, the estimated trigger rates are still preliminary and are the subject of intensive simulation activity concerning the physics channels under study and the background conditions.

As a reminder - All Collaborations propose that data recording, calibration, reconstruction, first reprocessing and data storage of all events will take place at CERN. On the other hand, Monte Carlo simulations will be mainly performed outside CERN, at the Tier1 and Tier2 Centres, and at dedicated facilities, together with most of the analysis work. Some fraction of the reprocessing can also be performed in the Tier1's, and part of the simulation at CERN.

3.1.1 ALICE

The collaboration plans to take data for about 10^6 s during four weeks every year with Lead ion beams, and 10^7 s with proton beams, together with the other experiments. In fact, ion running is not expected to start until the second year of data taking and so the acquisition profile for the computing system should be adapted accordingly.

Table A3.1 presents the estimates of the amount of data produced by the ALICE detector in one year of running at nominal conditions. Monte Carlo simulated events and the additional information produced by reconstruction are included.

Table A3.2 presents the estimated CPU power necessary to run ALICE and to make the analysis.

In heavy-ion mode (Pb-Pb) ALICE will take data with different triggers in parallel, for which the event size and rate varies significantly. During the first year of operation these will be: central Pb-Pb collisions (40 MB after compression), minimum bias Pb-Pb collisions (between 2 and 40 MB) and dimuon (at much higher rate but with negligible event size). Assuming an effective running time of 10^6 s, the total amount of collected raw data will be 1.25 PB, and together with reconstructed events 1.5 PB. In following years, other triggers (electrons) and higher event rates are envisaged, with L3 filtering and / or partial event readout.

In p-p mode, during the first year, ALICE will basically collect minimum bias events. As reference data for comparison with Pb-Pb, about 10^9 pp events will be needed in order to get comparable statistical resolution in the hadronic observables (due to significantly lower multiplicity). The event size of pp data depends crucially on the luminosity, due to event pile-up in the TPC. Up to $L=2*10^{29} cm^{-2}s^{-1}$ it is very small, about 0.2 MB per event. At the luminosity $L=3*10^{30}$ $cm^{-2}s^{-1}$ it rises to 1.0 MB per event in average, and at $L=10^{31}$ $cm^{-2}s^{-1}$ it reaches 3.5 MB per event (all after compression). Using the size for $L=3*10^{30}$ $cm^{-2}s^{-1}$ the total amount of raw and reconstructed data is estimated to be 1.2 PB in first year. If the pp ALICE running is spread over an effective time 10^7 s, the average throughput to mass storage will be 100 MB / s. Providing a proper buffering scheme is used, the maximum instantaneous throughput will not exceed largely the average one. If a special effort is made to keep the luminosity low and / or a successful filtering of the TPC pile-up is demonstrated, these requirements could be 5 times lower. During the following years ALICE will progressively use the pp running time for comparison of pp data with the hard-probe triggers and the requirements will stay on the same level, or will increase only moderately.

As far as the CPU is concerned, it is estimated that the offline computing power needed for pp is about 20% of that for heavy-ion running. The construction schedule of the full computing system should be adapted to the needs. As ALICE should prepare for a heavy-ion beam before the end of 2006, enough bandwidth and computing power should be installed during 2006. At very small additional cost, this scenario provides a good test environment before the challenging first heavy-ion run, which is expected at the start of the second year of data taking.

TABLE A3.1

ALICE Storage Requirements[14]

Tier0

		# of events	Evt-Size MB	Tape storage TB
Pb-Pb	Raw data	5.E+07	25.	1250
	Rec. Raw	5.E+07	2.5	125
p-p	Raw data	1.E+09	1.0	1000
	Rec. Raw	1.E+09	0.1	100

Total tape storage = 2850 TB. Disk pool of 260 TB.

For each Tier1

Pb-Pb & p-p	SIM. Out[1]	2.E+04	600.	12
Pb-Pb & p-p	SIM. Rec.[2]	1.E+07	5.	50
Copy from the other Tier1's	SIM. Rec.[2]	4.E+07	5.	200
Copy from Tier0	Rec. Raw			48

Total tape storage at a Tier1: 370 TB. Disk pool: 270 TB.

Grand total of Tier1's: Tape storage 1850 TB with a total disk space of 1350 TB.

[1] = Monte Carlo - production of the raw data.

[2] = Monte Carlo - production of the reconstructed data.

TABLE A3.2

ALICE CPU Requirements[14]

Data Processing	# events to Mass Store	CPU per event kSI95 s	CPU $2*10^7$ s / pass kSI95	CPU total (for rec. raw) kSI95
Tier0				
Raw Pb-Pb Reconstr.	5.E+07	100	250	500
Raw p-p Reconstr.	1.E+09	0.4	20	40
Calibration (assume 10%)				50

Total CPU at Tier0: 590 kSI95

Each Tier1

Pb-Pb Simul. (generation)	2.E+04	2250	54.5	54.5
Pb-Pb Simul. (reconstr.)[1]		500*100		
p-p Simul. (generation)	2.E+07	3.	3.5	3.5
p-p Simul. (analysis)				176

Total CPU at a Tier1: 234 kSI95

Grand total of Tier1's: CPU 1170 kSI95

[1] Each generated event will be re-used with different simple physics signals some 500 times, each reconstruction pass taking 100 kSI95-seconds.

3.1.2 ATLAS

For ATLAS, it is assumed that there will be (in addition to CERN) about 10 regional centres - Canada, France, Germany, Italy, Japan, Netherlands, Scandinavia, Russia, UK, and USA are currently candidates to host such

[14] Including realistic usage efficiency factors

centres. However, not all of them will be full-size Tier1 centres. Therefore, for the purposes of the computing resource calculations, it is estimated that these correspond to 6 full-size Tier1 centres.

During 2000, ATLAS performed detailed studies of the expected trigger rates and event sizes in the context of the Trigger / DAQ Technical Proposal. These studies have resulted in significantly higher trigger rate and event size (270 Hz and 2 MB) at a luminosity of 10^{33} than expected since the Computing Technical Proposal in 1996, which was based on the "canonical" assumption of 100 Hz and 1 MB. Work on understanding and reducing these numbers is going on. However, they represent the current best estimates and are thus used here (Table A3.3 and Table A3.4) for cost calculations, for a typical 10^7 seconds run every year. The scenario with 100 Hz (1 MB / event) is summarised for comparison in the Table A2.1.

The necessary storage capacity of the Tier0 and the Tier1 centres is presented in Table A3.3 and the necessary CPU power in Table A3.4.

Reconstruction and reprocessing assumes 640 SI95's per event, but better simulations are under way to evaluate how much the necessary CPU power will grow when the luminosity of the accelerator increases.

TABLE A3.3

ATLAS Storage Requirements[15]

Tier0 at CERN

(Most of the Disk space accounted for in Tier1)

One Year running	# of Events	Event size (MB)	Disk storage (TB)	Tape storage (TB)	Comment
Raw Data	$2.7*10^{9}$	2		5400	Backup not included
Rec. Raw (ESD)	$2.7*10^{9}$	0.5		2700	Current and previous versions for all recorded events
Calibration etc.			40	400	
Simul. (repository)	$1.2*10^{8}$	2.0		240	
ESD (Sim data)	$1.2*10^{8}$	0.5		120	Current and previous versions
Total			**40**	**8860**	

For each Tier1 (except CERN)

Simul. (MCRaw)	$0.2*10^{8}$	2.0		40	Only locally produced data
ESD (Sim data)	$1.2*10^{8}$	0.5	24	60	Current version
ESD (Sim data)	$1.2*10^{8}$	0.5	6	60	Previous version
General ESD for local analysis	$1.55*10^{9}$	0.5	194	775	Current version
Re-processed ESD for local analysis	$1.55*10^{9}$	0.5	78	775	Previous version
General AOD	$1.55*10^{9}$	0.01	16	16	
General TAG	$1.55*10^{9}$	0.002	3	3	
Local AOD, TAG etc.			20	20	
Raw data sample	$2.*10^{7}$	2.0	4	40	
User data			20	50	
Total			**365**	**1839**	

"Special" Tier1 at CERN

(Most of the tape storage accounted for in Tier0)

ESD (Sim data)	$1.2*10^{8}$	0.5	24		Current version
ESD (Sim data)	$1.2x10^{8}$	0.5	6		Previous version
General ESD for local analysis	$1.55*10^{9}$	0.5	194		Current version
Re-processed ESD for local analysis	$1.55*10^{9}$	0.5	78		Previous version
General AOD	$1.55*10^{9}$	0.01	16	16	
General TAG	$1.55*10^{9}$	0.002	3	3	
Local AOD, TAG etc			20	20	
User data			30	60	More users than the avg. Tier1
Total			**371**	**99**	

---

[15] Including realistic usage efficiency factors

TABLE A3.4

ATLAS CPU Requirements[16]

Tier0 at CERN

One Year running	# of Events	CPU / event (SI95 sec)	CPU power (kSI95)	Comment
Reconstruction	$2.7*10^9$	640	173	"quasi real-time"
Re-processing	$1.55*10^9$	640	212	3 month response time
Calibration			50	
Total			**435**	

For each Tier1 (except CERN)

Simulation (MCRaw)	$0.2*10^8$	3000	4	Events processed in 6 mth
Rec. Simulation	$0.2*10^8$	640	0.9	Events processed in 6 mth
Selection and re-definition of AOD			4	1 month response time
User Analysis			200	4 hours response time
Total			**209**	

Additional Tier1's and / or Tier2's will reduce the computing requirements for the average Tier1.

"Special" Tier1 at CERN

Selection and re-definition of AOD			5
User Analysis			250
Total			255

ATLAS Summary of Computing Resources (for 270 Hz x 2 MB)

Total CPU at CERN (Tier0+Tier1) = 411 + 255 = 690 kSI95
Total CPU at Tier1's (excluding CERN, 6 Tier1's) = 209x6 = 1254 kSI95
Total CPU for ATLAS = 690 + 1254 = 1944 kSI95

Total Tapes at CERN (Tier0+Tier1) = 8860 + 99 = 8959 TB
Total Tapes at external Tier1's (for 6 Tier1's) = 1839x6 = 11034 TB
Total Tapes for ATLAS = 8959 + 11034 = 19993 TB

Total Disk at CERN (Tier0+Tier1) = 40 + 371 = 411 TB
Total Disk at external Tier1's (for 6 Tier1's) = 365x6 = 2190 TB
Total Disk for ATLAS = 411 + 2190 = 2601 TB

---

16 Including realistic usage efficiency factors

3.1.3 CMS

The basic assumption for the CMS Collaboration is to run with 100 Hz of events stored, for a typical 10^7 seconds of physics data taking, during a run of about 200 days / year. The storage capacity necessary in the Tier0 and Tier1 centres for one year of data taking is presented in Table A3.5. Table A3.6 gives the expected CPU power needs to fulfil the various tasks at CERN and at the Tier1's.

The values quoted express the resources that should be available to users, without accounting for implementation efficiency. The implementation efficiency has two major contributions: the first being usage of the bare resources (e.g. the CPU cannot run at 100% and a single disk cannot be filled at 100% capacity), the second being the overhead for operation (e.g. downtime for faults or maintenance) and development. These efficiencies must clearly be taken into account appropriately when sizing the various systems (Tier0, Tier1 and Tier2).

At the moment the CMS collaboration has a preliminary commitment for 5 Tier1's located outside CERN (FNAL in the US, and RAL, IN2P3 Lyon, INFN and Moscow in Europe). CMS plans also to have many Tier2 centres, about 25, that will contribute substantially to the analysis of the data, the simulation and the reconstruction.

It is not possible at the moment to detail the exact number and site location of the Tier2's with the corresponding total resources that will be available. Some of the already identified Tier2 sites have, however, detailed plans for implementation and resource allocation (Table A3.5 and Table A3.6 also therefore give an average dimension of foreseen resources for Tier2's).

CMS expects that Tier2 centres will contribute substantially to the overall computing effort and in particular to the analysis activities. It thus confirms that they must be taken into account in the resource estimations.

TABLE A3.5

CMS Storage Requirements[17]

Tier0 + Tier1 at CERN

	# of events	Event-size (Mbytes)	Active tape / Archive tape / Disk (Tbytes)
Raw data	1.E+09	1.0	1000 / 1000 / 0
Rec.Raw	1.E+09	0.5	500 / 0 / 200
Calibration			0 / 10 / 10
Simulation (repository)	5.E+08	2.0	0 / 1000 / 0
Re-proc. ESD	1.E+09	0.5	0 / 200 / 200
Rec-simulation	5.E+08	0.4	40 / 200 / 30
Reprocessed ESD (Tier1)	2.E+08	0.5	0 / 100 / 40
Revised ESD	2.E+08	0.5	0 / 100 / 40
General AOD	1.E+09	0.01	0 / 10 / 10
Revised AOD	2.E+08	0.01	0 / 2 / 2
Local AOD,TAG, DPD	2.E+08		0 / 10 / 10
Cache Disk for active Tapes			0 / 0 / 154
User Data			0 / 0 / 100
Total			**1540/2632/796**

1540 TB of tape storage is online (robot, tape drives), the rest (2632 TB) is archive.
The disk cache is estimated of the order of 10% of active Tape storage.

For each Tier1 not at CERN

SIM. Out	1.E+08	2.	0 / 200 / 0
SIM. Rec.	1.E+08	0.4	40 / 0 / 30
Raw-sample	5.E+07	1.	50 / 0 / 0
Calibration			0 / 10 / 10
ESD	1.E+09	0.5	500 / 0 / 0
Re-proc. ESD	0.2E+09	0.5	0 / 100 / 40
Revised ESD	0.2E+09	0.5	0 / 100 / 40
General AOD	1.E+09	0.01	0 / 10 / 10
Revised AOD	2.E+08	0.01	0 / 2 / 2
TAG	1.E+09	0.001	0 / 1 / 1
Local AOD, TAG, DPD	2.E+08		0 / 10 / 10
User data			0 / 0 / 50
Cache Disk for active Tapes			0 / 0 / 120
Total			**590/433/313**

There are 6 Tier1's: CERN (in the first part of the table) + 5 (candidates currently are USA (FNAL), France (Lyon), UK (RAL), Italy (INFN), Russia (Moscow)).
The disk cache is estimated of the order of 20% of active Tape storage.

For each Tier2

Local cached data (real + simulated)	0 / 0 / 50
User data	0 / 50 / 20
Total	0 / 50 / 70

About 25 Tier2's are foreseen

[17] NOT including efficiency factors

TOTAL Active tape for CMS :
TIER0 / 1 CERN + 5 TIER1 = 1540+5*590 = 4490 TB
TOTAL Archive tape for CMS :
TIER0 / 1 CERN + 5 TIER1 + 25 TIER2 = 2632+5*433 +25*50= 6047 TB
TOTAL tape for CMS :
TIER0 / 1 CERN + 5 TIER1 + 25 TIER2 = 4172+5*1023 +25*50 = 10537 TB
TOTAL Disk for CMS :
TIER0 / 1 CERN + 5 TIER1 +25 TIER2 = 796+5*313 + 25*70 = 4111 TB

TABLE A3.6

CMS CPU Requirements[18]

Data Processing	# events	CPU / event (kSI95 s)	CPU total (kSI95)
Tier0 + Tier1 at CERN			
Reconstruction	1.E+09	3.	440
Reprocessing	1.E+09	3.	Included above
Selection	1.E+07-1.E+08	Up to 0.025	15
Analysis and DPD	1.E+07	0.010	160
TOTAL CPU			**615**
For each Tier1 not at CERN			
Simulation	0.25E+08	5.	5
Rec-Simulation	0.25E+08	3.	3
Re-Processing	0.1E+09	3.	50
Selection	1.E+07-1.E+08	Up to 0.025	15
Analysis	1.E+07	0.010	80
TOTAL			**153**
For each Tier2			
Simulation	0.5E+08	5.	10
Rec-Simulation	0.5E+08	3.	6
Analysis	1.E+07	0.010	16
TOTAL			**32**

About 25 Tier2's are foreseen

TOTAL for CMS:
TIER0 / 1 CERN + 5 TIER1 + 25 TIER2 = 615 + 5*153 + 25*32 = 2180 kSI95

[18] NOT including efficiency factors

3.1.4 LHCb

The data volumes and storage requirements corresponding to an annual data-taking period of 10^7 sec are shown in Table A3.7, together with the CPU power requirements in Table A3.8.

The reconstruction facility will be implemented as part of the online event filter farm. Reprocessing will take place on the same filter farm, outside of normal data-taking periods. This explains why no additional capacity is requested for reprocessing. By far the largest demands on the computing infrastructure come from the heavy simulation load. It is assumed that this load will be shared equally between the Tier1 and Tier2 computing centres. Some capacity will also be required at a Tier1 centre at CERN.

LHCb expects to have several Tier1 and Tier2 centres, although precise knowledge of the number and nature of the facilities is still evolving. Tier1 centres will exist at RAL, Lyon and INFN. Significant computing capacity will also be available in Liverpool, NIKHEF, and Switzerland. There are also on-going discussions in Germany, Poland and Moscow on the provision of regional centres for computing. For the purposes of calculating the installed capacity required outside CERN, it is assumed that the total computing load will be distributed over the equivalent of 5 Tier1 centres and CERN.

TABLE A3.7

LHCb Storage Requirements[19]

(a) Tier0 + Tier1 AT CERN

	Number of events / year	Event Size (MB)	Active Tape (TB)	Archive tape (TB)	Disk (TB)
Raw data	2.0 E+09	0.125	250	250	0
ESD	2.0 E+09	0.100	200	0	80
Repr. ESD			200	0	80
Analysis	2.0 E+09	0.020	160	0	80
MC RAW	0.2 E+09	0.200	40	40	0
MC ESD	0.2 E+09	0.100	20	0	20
MC AOD	1.2 E+09	0.035	42	0	20
Calibration			0	10	10
Physicist data			0	10	10
Disk cache			0	0	30
Total			**912**	**310**	**330**

[19] Including realistic usage efficiency factors

(b) For each Tier1 not at CERN

	Number of events / year	Event Size (MB)	Active Tape (TB)	Archive tape (TB)	Disk (TB)
AOD data	1.0 E+09	0.020	160	0	80
MC RAW	0.2 E+09	0.200	40	40	0
MC ESD	0.2 E+09	0.100	20	0	20
MC AOD	1.2 E+09	0.035	42	0	20
Calibration			0	10	10
Physicist data			0	5	5
Disk cache			0	0	15
Total			**262**	**55**	**150**

TOTAL Tier1 active tape (assuming 5 Tier1's) 1310 TB
TOTAL Tier1 archive tape (assuming 5 Tier1's) 275 TB
TOTAL Tier1 disk (assuming 5 Tier1's) 750 TB

TABLE A3.8

LHCb CPU Requirements[20]

	Data Processing	No. of events to Mass Storage	CPU / event (kSI95 s)	Total CPU (kSI95)
Tier0 + Tier1 AT CERN				
	Reconstruction	2.0 E+09	0.250	50
	Reprocessing	2.0 E+09	0.250	0
	Data analysis	2.0 E+09	0.005	20
	Simulation	0.2 E+09	0.550	110
	Physics analysis			45
	TOTAL CPU			**225**
For each Tier1 not at CERN				
	Simulation	0.2E+09	0.55	110
	MC AOD definition	0.2E+09		10
	Physics analysis			20
	TOTAL			**140**

TOTAL for LHCb: CERN + 5 Tier1 = 225 + 5*140 = 925 kSI95

20 Including realistic usage efficiency factors

3.2 The CERN-based Computing Centre

Table A3.9 shows the expected performance of the CERN computing system, summarising the requests of the four experiments. For Tape, the figures supersede those given earlier to the Computing Model panel and shown in Table A2.1.

TABLE A3.9

Computing power and storage capacity at CERN
(Realistic usage efficiency factors for the components taken into account)

	CPU (kSI95)	Tape (TB)	Disk (TB)	Tape I/O (MB/s)	Shelf Tape (TB)
ALICE	824	3'200	534	1'200	0
ATLAS	690	8'959	410	800	0
CMS	820	1'540	1'143	800	2'632
LHCb	225	912	330	400	310
Total	**2'559**	**14'611**	**2'417**	**3'200**	**2'942**

The large Tape storage demand from ATLAS is due to the possibility that they will run with a 270 Hz trigger rate, while the large CPU demand from ALICE is motivated by the complexity of Pb-Pb events.

3.3 Staffing for CERN/IT computing services

The estimate of the minimum staffing level for managing and operating the offline computing facilities for all experiments active in 2006 assumes the current model. In this, CERN staff are responsible for planning, development and management of the services, while the operation, systems administration and first-level user support are handled by a contractor. While the quantity of equipment to be handled will be much greater than today, it is expected that:

- the extent of the heterogeneity of hardware and systems software can be substantially reduced;
- agreement can be reached with the experiments on a much greater degree of standardisation across experiments than is present today;
- the level of automation of installation, maintenance, monitoring and management of the computing farms and storage systems can be increased.

An important step in this direction is the decision to reduce the number of supported platforms to two (Linux as mainstream platform and Solaris as second development platform) recently agreed within the FOCUS committee. It is assumed that four LHC experiments must be planned for in 2006 and that non-LHC experiments will account for only a small fraction of the service.

It is clear that the regional centre model will require a new set of services for the co-ordination of the facilities at CERN and in the Tier1 centres and the distributed management of the data. This will involve a significant level of management and operational attention, the level of which is not yet understood.

Table A3.10 gives estimates for the CERN staff required for direct support of the services for physics data handling, specialised support for engineering and the basic computing and networking infrastructure (e.g. desktop services, internet services, general networking infrastructure, etc.). It is assumed that, as at present, each experiment provides its own support for experiment-specific physics libraries and environment, data management, etc. As mentioned above, the computing farms will be administered and operated by contractors operating under service level agreements. The cost of this support is discussed in Appendix 3.5.

The “zero model” refers to a model where the support level is comparable with that at present. The “-30 model” is the bare minimum to run the computer centre with a strong reduction on the support level for users and experiments.

TABLE A3.10

IT Staff Estimates for LHC Computing Support

	FTEs 2006 -30 model	FTEs 2006 zero model
Dedicated Physics Computing Services		
Physics applications services		
Simulation	4	5
Analysis & visualisation	6	8
Common libraries, tools and base support	6	8
Controls for physics	8	8
Tier0 + Tier1 centre		
Basic farm mgt / planning	6	8
Operation & support for LHC experiments	10	12
High bandwidth WAN for LHC	3	3
Direct support for non-LHC experiments	4	5
Total - dedicated support for physics	**47**	**57**
Specialised support for engineering	**14**	**18**
Infrastructure and shared services (including base support for physics and engineering)		
Data management	5	8
Desktop support	19	24
Campus networking	10	11
Controls infrastructure	5	5
Database services (relational and object)	10	13
External networking	5	6
Internet applications	9	12
User support, operation and infrastructure	12	12
Management, administration	21	21
Total - infrastructure and shared services	**96**	**112**
Total IT staff	**157**	**187**

TABLE A3.11

Comparison of projected IT human resources for operation of the CERN-based computing facility in 2006 with what is foreseen at a number of other prospective European Regional Centres

	CPU (MSI95)	DISK (PB)	TAPE (PB)	#FTE ()=outsourced
CERN “zero model”	2.6	2.4	18	23+(20)
IN2P3	0.80	0.72	3.0	38
INFN	0.32	0.50	3.0	38
RAL	0.28	0.30	2.0	45

3.4 Evaluations of human resources needed by the Collaborations

The Collaborations have given their estimation of the human resources necessary to build and run the software. The software itself is subdivided into:

a) The Core Software. The complexity of LHC software requires a well-engineered architecture and infrastructure written by software experts in order that the Collaborations may contribute efficiently and that the result is manageable. Core Software comprises that software infrastructure which is required by detector-specific and analysis-specific software. It includes items like the 'event-loop' and associated control, database infrastructure, calibration infrastructure, data-management infrastructure etc. 'Core' does not include things like detector-specific reconstruction code or calibration code etc., nor of course the software of specific physics analyses (e.g. a neural-net analysis).

b) The physics analysis programs, the detector-specific programs, calibration, etc…, written by physicists. The four collaborations are confident that they will find the human resources to build the physics analysis software and to perform the physics analysis in connection with the Tier1 and Tier2 centres.

The Core Software teams will be composed of software professionals and expert physicists coming from the LHC community. They will not necessarily be located at CERN, nor in any single place. The team will write the software infrastructure to enable collection, calibration, reconstruction, and distribution of the data.

Table A3.12 presents the experiments' best estimates of the human resources needed to write the Core Software of the four Collaborations, as a function of time. The figures do not include experiment's activities in EU DataGrid Work Package 8 (or similar initiatives).

TABLE A3.12

Required human resources (FTEs) to write the Core Software

Year	2000 have(missing)	2001	2002	2003	2004	2005
ALICE	12(5)	17.5	16.5	17.0	17.5	16.5
ATLAS[1]	23(8)	36	35	30	28	29
CMS	15(10)	27	31	33	33	33
LHCb	14(5)	25	24	23	22	21
Totals	**64(28)**	**105.5**	**106.5**	**103**	**100.5**	**99.5**

[1] ATLAS used the following definition of 'Core' computing activities in preparing their data for Table A3.12. "These activities are seen as covering the development, deployment and maintenance of all infrastructure software (dBase, framework etc.), plus all associated management, co-ordination, and support, including the organisation of Data Challenges. 'Core' thus covers everything except the algorithmic parts of the code used for reconstruction, simulation, and physics analysis. However, human resources directly attributable to the Grid are not included."

3.5 Cost estimates and schedule

3.5.1 Introduction

The construction schedule of the whole LHC computing system must clearly be adapted to that of the overall LHC project, machine and detectors.

The importance of the computing facilities located at CERN, necessary for data taking, is stressed.

The rapid evolution of the computing market and the special requirements of LHC computing make the cost evaluation exercise difficult. The Management and Resources panel has nevertheless gone ahead in order to have a first-order approximation of the cost for both the final system and the prototypes.

The design of the software and hardware infrastructures will continue to evolve as the development activity progresses. A more precise evaluation of the costs will only be possible much nearer the time of LHC start-up.

This report addresses the estimated investment costs to build the system. The cost of maintenance and operation in the Tier1 centres is not addressed, since this depends on the local situation of each centre.

3.5.2 Evolution of the cost for CPU, Disk Space, Tapes

The cost and performance of individual components on the market changes with time and it is therefore necessary to extrapolate their costs. This extrapolation is done here using the conclusions of the PASTA committee (see footnote 8 on page 31).

The CPU prices are weighted for infrastructure (racking, power and vertical slice of the LAN), giving a 25% increase on the computer box price. The CPU prices are also a mean value among “low-end” and “high-end” servers. Presently most of the calculations rely on dual-CPU boxes for PC systems.

The disk prices include the disk server hardware (CPU, network) and are extrapolated taking as the starting point an existing system (fully mirrored IDE disks, Linux server currently deployed in the CERN computer centre for data-intensive applications).

In Table A3.13 the best estimates of the cost are presented as a function of time (see also Figure A3.1 and Figure A3.2). As can be seen from the table, unit prices are expected to continue to change as they did in recent years. The effective cost varies much with time.

FIGURE A3.1
CPU price evolution (PASTA report)

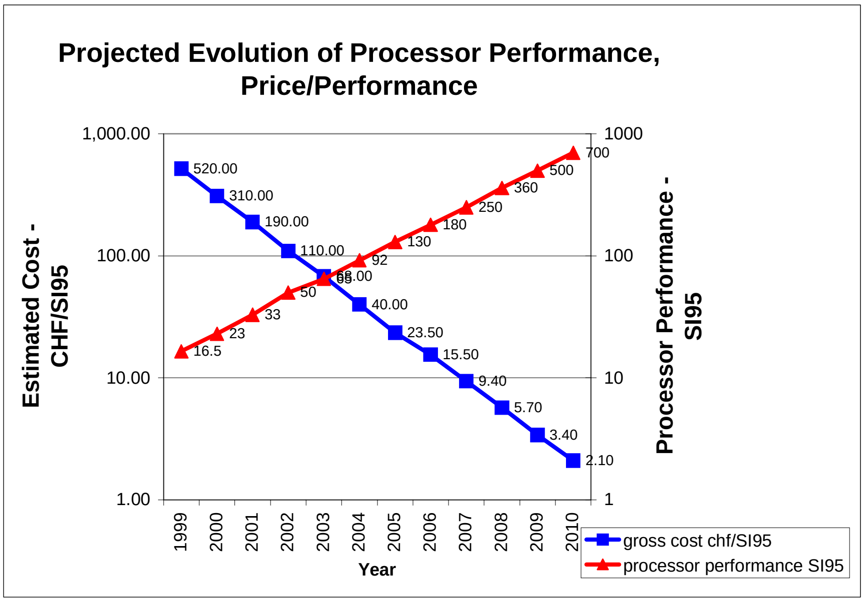
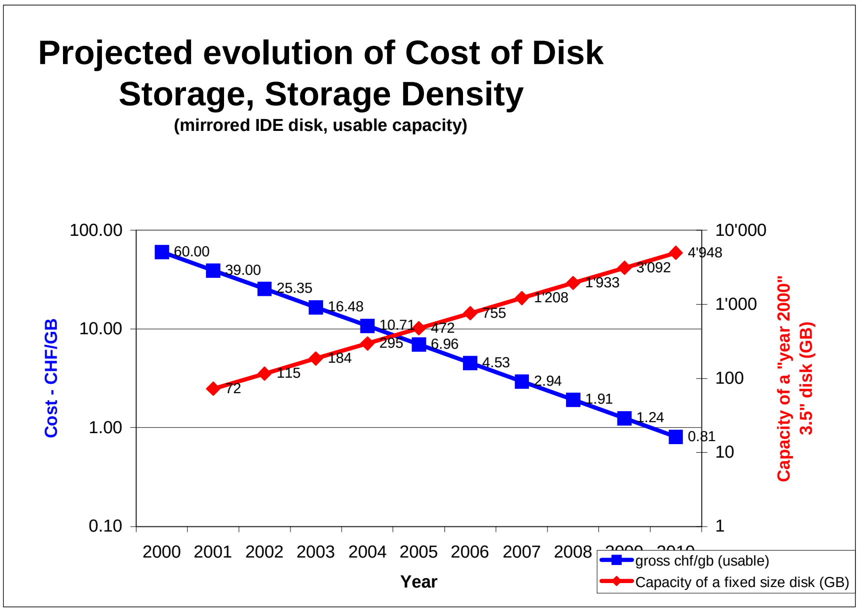


FIGURE A3.2
Disk price evolution (PASTA report)



3.5.2.1 Tape price evolution.

The starting price for tapes is a heavily discounted high-end unit (TC40, Magstar). The robot slot price is that for the Powderhorn silo. The robotics increases the cost of the tape by about 50%. One could consider shelving inactive data, after a certain time, to reduce the cost of the robot. The "Active Tape" price includes the cost of the appropriate number of automated slots (robot). A medium price is considered for the tapes on the shelves. The price evolution is the PASTA prediction on technology improvement.

Maintenance for the tape drives, at the level of 15% per year, is included in the total cost. After the first year of running we assume 4 years lifetime (i.e. the cost is increased by 45% to cover maintenance).

The cost of the drive (unit price) is considered to be constant over time while the performance follows the evolution of the tape size (i.e. constant time to read a full tape). An efficiency factor of 60% is used to take into account real operations (mount time, etc.).

TABLE A3.13

Unit Cost Estimate Evolution (from PASTA committee)

Year	CPU CHF / SI95	CPU per box SI95 / box	Disk CHF / GB	Autom.tape CHF / GB	Shelf tape CHF / GB	Tape I / O CHF / (MB / s)	Sys Adm CHF / box
2000	310.00	46	60.00	3.70	2.50	7'250	2'500
2001	190.00	66	39.00	2.80	2.00	4'833	1'875
2002	110.00	100	25.30	2.40	1.60	4'833	1'667
2003	68.00	130	16.50	1.90	1.20	3'625	1'250
2004	40.00	184	10.70	1.50	1.00	2'900	1'000
2005	23.50	260	7.00	1.30	0.80	2'900	833
2006	15.50	360	4.50	1.00	0.70	1'812	652
2007	9.40	500	3.00	0.90	0.60	1'812	500
2008	5.70	720	1.90	0.86	0.55	1'812	500
2009	3.40	1000	1.20	0.81	0.50	1'812	500
2010	2.10	1400	0.80	0.81	0.50	1'812	500

The cost estimates require the investigation of four extra items:

- Wide area networking cost estimates;
- system management estimates;
- provision for the tertiary storage for the years after the start-up;
- provision for system evolution (replacement of obsolete hardware and evolution of the facilities in terms of disk space and CPU power).

The prices for the wide area networking have been derived from the predictions reported by CERN / IT. It is worth noting here that, already now, CERN shares the cost of some dedicated links with external bodies, as in the case of the US connection, the cost of which is shared between CERN and USA with a 25% / 75% ratio. It is also worth noting that the price actually paid (by CERN, by other computing centres or research institutes) is normally a fraction of the real cost, due to subsidies (CERN gets a 40% subsidy from the European Community). The estimates provided here only try to guess a possible financial envelope for the CERN wide area networking costs and are not a detailed forecast of the price and technology evolution in the next few years.

For system administration, CERN is already managing a large part of its computer infrastructure (system administration, part of the end-user support for both software and hardware) via a Service Level Agreement (SLA) with different service providers. In the computer centre there are over 800 systems (subdivided in many separate farms or clusters), which are managed with these types of SLAs. The costing model uses the present contracts as a starting point. For the evolution a 20% per year improvement in the number of (dual-CPU) systems handled by one administrator is assumed. The cost of each administrator is fixed and estimated at 150 kCHF per year. The only cost evolution comes from the higher number of systems handled by one person. This “evolution” is stopped arbitrarily when an administrator can handle 300 systems. These forecasts are rather conservative, taking into account the evolution of the tools to provide automated fabric management both in the research sector (e.g. the WP4 of the EU DataGrid project) and in the growing commercial marketplace for service providers.

The number of new tapes for the storage of the data is considered constant, in terms of capacity, for the first few years after the start-up. Every year new tapes (and the necessary robotics) are purchased and the previous infrastructure is kept. No provision for data migration to different technology is made, although it will be needed.

The evolution of the amount of necessary disk space is not clear. The model used here is to add annually 50% of the required capacity (taken from Table A3.9) starting from the year 2008. This is to provide space for more complex analyses, requiring more and more user space to be performed efficiently. The CPU is constantly recycled and, starting in 2008, 33% of extra CPU (compared to the Table A3.9 figures) is added. The choice of a somewhat higher growth rate for disk capacity is based on experience.

3.5.3 Baseline cost estimate of the CERN-based computing system

3.5.3.1 Investment

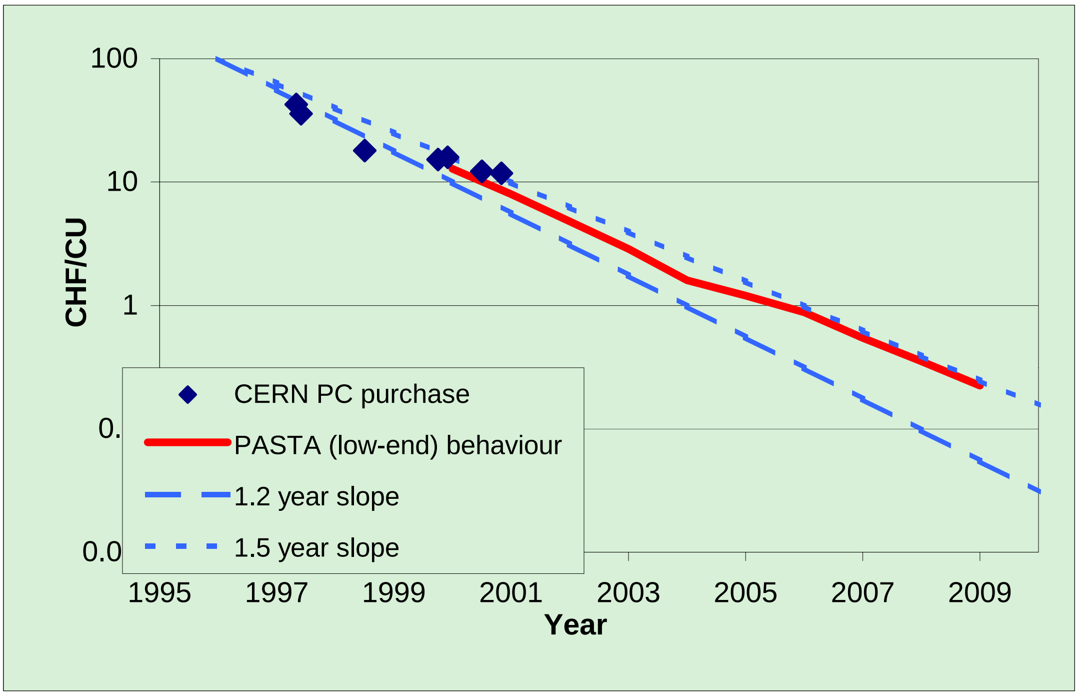
The following figures have been obtained using an educated guess for the price evolution, sometimes referred as Moore’s law. As a matter of fact, Moore’s law is only about technology improvements, while the price evolution is linked to external factors like the market trends. Other phenomena (e.g. the CHF - USD exchange rate) will clearly play a major role.

The input to the calculations comes mainly from the conclusions of the PASTA committee. These can be validated with the actual trends observed in the last years for the purchase of equipment in the CERN computer centre, where acquisitions of PCs in the years 1997-2000 basically confirm the trend of the PASTA reports. These acquisitions were similar in scale to those foreseen for the LHC farms (large quantities purchased via tender, the actual vendor and detailed configurations change, the whole PC set-up evolves - e.g. more memory): see Figure A3.3. The data available are insufficient to validate completely the model (and they cannot do this, due to the above-mentioned external factors), but give confidence that, on the time scale of the next years (until end-2003), the predictions are accurate to some 10-20%.

On the longer term, the error probably becomes too large for accurate financial planning. Nevertheless the present situation is enough to fix the magnitude of the effort and to be sure that the new predictions made after the prototype initiatives deliver their results (after 2003) will be a more solid basis for financial planning.

FIGURE A3.3

Comparison of recent CERN PC purchases with PASTA predictions



CERN data on PC costs (in CHF/CU) are plotted as solid squares. The solid line is the PASTA prediction for low-end CPU cost. Two exponential functions have been superimposed. They show the behaviour under two hypotheses discussed in the submission from FNAL, namely the doubling of the computing power every 1.2 and 1.5 years (dashed and dotted respectively).

The cost of the system is evaluated using Table A3.13 as the input and summing up the four major contributions to the investment cost, namely:

- CPU cost (including all infrastructure)
- Disk cost (including disk server and connectivity)
- Tape cost (including robot)
- Tape I / O (tape drivers and connectivity)

For all experiments, the cost has been computed integrating the investment in a three-year slice. Prices take into account guarantee and maintenance contracts, but within the first three year period there is no hardware recycling.

The selected acquisition profile depends very much on the performance of the LHC during the first year, plus the need to work within a realistic funding profile. For ATLAS, CMS, and LHCb, it has been agreed to use a deployment profile corresponding to 30%-60%-100% of the total capacity in the years 2005 - 2007. The same profile has been applied for all resources.

Note that the tape installation grows to 100% in three years for both tape media and infrastructure. Only after the third year are tapes purchased to increase the storage: this implies a high level of recycling of tapes in the first three years.

Since the ALICE collaboration has a two-fold physics programme with protons and ions, a slightly different deployment profile is proposed for them. The physics program with protons will start with the other three experiments, whilst the Pb beams are expected to start one year later. This is why ALICE suggests a ramp-up profile (20%-40%-100%) adapted to the commissioning of the detector and the start of the first physics programme.

Note that the ALICE requirement of 1200 MB / s of Tape I / O is concentrated in one month a year (the Pb-Pb period). The costing has thus been done assuming that 800 MB / s are purchased and guaranteed over all year, while the remaining 400 MB / s is borrowed from the rest of the infrastructure (25% of the infrastructure for the other three experiments).

The total three years investment for the computing at CERN (**not** including networking and personnel for operations) is shown in Table A3.14.

TABLE A3.14

Hardware Investment in Computing at CERN in 2005-7 (excluding networking)

	2005	2006	2007	Hardware cost (MCHF)
ALICE	20%	40%	100%	18
ATLAS	30%	60%	100%	24
CMS	30%	60%	100%	23
LHCb	30%	60%	100%	7.0

A one-year shift for the spending profile (to 2006-2008) would reduce the investment by about 30%, while a flat expenditure profile over three years can be obtained using a 15%-50%-100% profile for the CPU only, leaving the other components to follow the standard 30%-60%-100%. The expenditure profile of the construction period (sum of all experiments, hardware only) is 40%-30%-30%.

The sharing among different components is such that the investment for CPU is still the most important factor (>60% in the case of ALICE). Taking into account the years up to 2010, the CPU share remains generally dominant (>50% for ALICE). ATLAS, on the contrary, is dominated by the data storage. In the construction period, the CPU is 45% of the total cost while the tapes account for about 40%.

The complexity of these systems, at the end of the full deployment in 2007, requires about 500 boxes for LHCb, 2200 boxes for ALICE, 1500 boxes for ATLAS and 1500 boxes for CMS. Thus the total number of dual computers (boxes) at CERN will be about 5700 (to be compared with the 800 boxes presently installed).

3.5.3.2 Maintenance and operations

Equipment and tape storage

Although the price-performance ratio of all components can be expected to continue to improve for the foreseeable future, thus tending to reduce replacement costs, this trend will be counterbalanced by the increasing capacity needs of the experiments as they and the LHC machine mature.

The maintenance and operation costs of the CERN-based computing system are estimated on the following assumptions:

- After the construction period, the capacity of the system is increased by a constant amount every year.
- The amount of CPU is increased every year by 33% of the 2007 value. The obsolete equipment is replaced after the 3-year maintenance period.
- The tertiary tape storage capacity is increased every year by 100% of the 2007 capacity, to store the new data.
- The disk space is increased every year by 50% of the 2007 value, to make the analysis of bigger data samples possible.
- The costing for the WAN and for system administration is made for the full installation (i.e. no breakdown by experiment).

Wide Area Networking

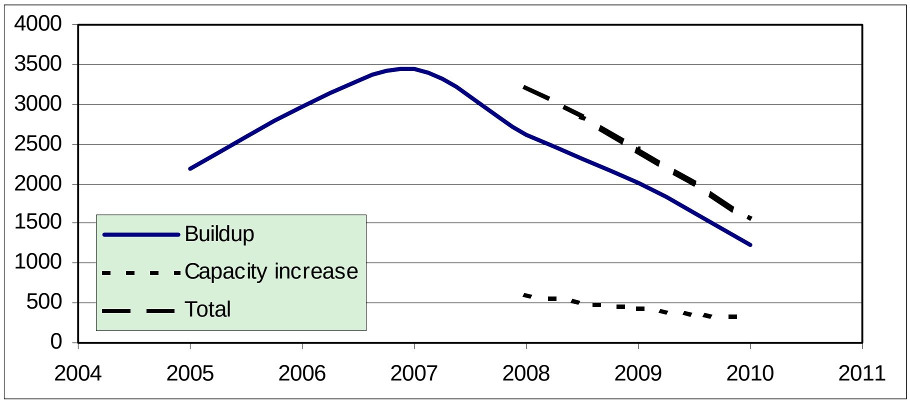
The bandwidth out of CERN that will be available for LHC computing is not yet completely clear, since it depends on the price of this commodity and the consequent detailed implementation of the computing model. The order of magnitude will be a few Gbps per experiment (a reasonable goal is ~10 Gbps). The price paid by CERN is estimated to be of the order of 4 MCHF / year during the ramp-up period and the first years of operation. This will yield ~7 Gbps for all experiments in 2005 with a yearly capacity increase of the order of 20%.

System Administration

System administration will be a largely outsourced activity and is thus accounted as a cash cost. This cost has a time evolution (see Table A3.13), which depends on the complexity of the deployed system. The overall result for the CERN-based computing facility is shown in Figure A3.4.

FIGURE A3.4

System Administration costs for the CERN-based System



System admin. costs (kCHF) as a function of time, as computed for the capacity available at start-up and its evolution due to recycling (solid line), the contribution of the 33% capacity increase (dotted line), and the total after 2007 (dashed line).

3.5.3.3 Summary of overall costs

Table A3.15 gives the hardware and tape costs for the construction period (2005-2007) broken down into the main categories for each experiment, as well as the percentage contributions of each category to the totals. Table A3.16 shows the expected maintenance and evolution costs of this installation during the subsequent 3 years. Note that, while the cost of automated tape is included, in order to facilitate maintaining the homogeneity of the system, shelf tape will be the responsibility of the experiments and is not included here.

TABLE A3.15

Overall hardware and tape costs of the initial CERN-based Computing Facility constructed in the years 2005-2007 (kCHF)

	ALICE	ATLAS	CMS	LHCb
CPU	11'069	10'667	12'667	3'479
Disk Pool	2'188	1'907	5'314	1'535
Automated Tape	3'200	9'407	1'617	958
Shelf Tape	0	0	1'816	214
Tape I/O	1'616	1'711	1'711	855
Total cost	**18'073**	**23'692**	**23'135**	**7'040**
%CPU Cost	61.2%	45.0%	54.8%	49.4%
%Disk Pool Cost	12.1%	8.0%	23.0%	21.8%
%Automated Tape Cost	17.7%	39.7%	7.0%	13.6%
%Shelf Tape Cost	0.0%	0.0%	7.0%	3.0%
%Tape I/O Cost	8.9%	7.2%	7.4%	12.2%

TABLE A3.16

Overall annual hardware replacement and evolution costs, including automated tape costs, of the CERN-based Computing Facility in the years 2008-2010 (kCHF)

	ALICE	ATLAS	CMS	LHCb
CPU cost evolution	1'870	1'680	1'996	548
Disk Pool cost evolution	375	299	834	241
Automated Tape cost evolution	2'741	7'675	1'319	781
Shelf Tape cost evolution	0	0	0	0
Tape I/O cost evolution	626	580	580	290
Total evolution	**5'613**	**10'234**	**4'730**	**1'860**
%CPU Cost	33.3%	16.9%	45.9%	32.4%
%Disk Pool Cost	6.7%	2.7%	17.1%	12.7%
%Automated Tape Cost	48.8%	77.4%	30.3%	46.3%
%Shelf Tape Cost	0.0%	0.0%	0.0%	0.0%
%Tape I/O Cost	11.1%	2.9%	6.7%	8.6%

To summarise, the cost of the hardware deployed in 2005-2007 is 72 MCHF, including the initial investment in tapes. In addition to this, during the same period, wide area networking costs are expected to total 12 MCHF and those for system administration 9 MCHF. For the subsequent three years, the annual hardware replacement and evolution costs, not including shelf tape, are expected to average 22 MCHF / year, while wide area networking continues at 4 MCHF / year. The system administration costs are expected to be of the order of 2-3 MCHF / year.

3.5.4 Cost estimate of the Tier1 computing systems

The specifications for the Tier1 used in this section have been obtained from the detailed data used to produce Table A2.1 (i.e. including realistic usage efficiency factors). Table A3.17 lists the specifications for typical Tier1's.

TABLE A3.17
Typical Tier1 Specifications

	CPU (kSI95)	Tape (TB)	Disk (TB)	Tape I/O (MB/s)	Shelf Tape (TB)
ALICE	234	400	273	1'200	0
ATLAS	209	1'839	360	800	0
CMS	417	590	943	800	683
LHCb	140	262	150	400	55
Total	**1'000**	**3'091**	**1'726**	**3'200**	**738**

The pricing for the construction of one Tier1 is obtained using the same algorithms as for the CERN-based computer centre.

The estimated total prices, hardware only, for a typical Tier1 are given in Table A3.18. The number of equivalent Tier1 centres is also listed. Note that in the case of CMS (5 Tier1's and 25 Tier2's), the average Tier1 is taken, for estimating the cost of the capacity installed outside CERN, to include explicitly 5 Tier2's. The other experiments have implicitly included in the Tier1's the resources that will be deployed across the Tier2's.

TABLE A3.18
Tier1 Hardware Costs (kCHF), Compared with the CERN-based System

	ALICE	**ATLAS**	**CMS**	**LHCb**	**Total**
#Tier1	4	6	5	5	
Avg. Tier1	7'095	8'547	13'638	4'030	
All Tier1's	28'381	51'281	68'189	20'152	168'002
CERN	18'073	23'692	23'135	7'040	71'940
CERN / Total	39%	32%	25%	26%	30%

The ratio CERN / Total in Table A3.18 is the ratio of the hardware cost of the CERN-based system over the total cost of the CERN-based system and all the Tier1's and Tier2's. The so-called 1/3-2/3 rule (ratio of installed capacity between CERN and elsewhere) is roughly verified.

Note that the Tape I/O requirements of the centres away from CERN have not been discussed and the CERN-based values are used.

Furthermore, no provision is made for the wide area networking costs between each of the Tier1's and Tier2's.

3.5.5 Cost estimate for a shared prototype system

The model to estimate the cost of the prototype, to be built in three years, is the same as that used to evaluate the cost of the CERN-based centre. The proposed final configuration is shown in Table A3.19.

The main differences concern the ramp-up profile: CPU is seen as growing with a 10%-20%100% profile, whilst the Tape I/O infrastructure follows a slightly more aggressive 10%-30%-100% profile. Table A3.20 shows the resultant cost estimates.

TABLE A3.19

Possible LHC computing prototype configuration

CPU	120 kSI95
Disk	300 TB
Tape	150 TB
Tape I/O	1200 MB/s

TABLE A3.20

Spending profile for the LHC computing prototype

2001	2002	2003	TOTAL
3 MCHF	3 MCHF	12 MCHF	18 MCHF

The complexity, in terms of the number of boxes, is comparable with the CERN Tier0+Tier1, although the compute power is about 20% of that. Note that the I/O requirements should be demonstrated at full scale, to be sure data can at least be collected (1200 MB/s is the ALICE requirement).