LCG-TDR-001
CERN-LHCC-2005-024
20 June 2005

LHC Computing Grid

Technical Design Report

Version: 1.0
20 June 2005

The LCG TDR Editorial Board
Chair: J. Knobloch
Project Leader: L. Robertson

Printed at CERN

June 2005

ISBN 92-9083-253-3

Also available at: http://cern.ch/lcg/tdr

EDITORIAL BOARD

---

1 CERN, Geneva, Switzerland

2 NIKHEF, Amsterdam, Netherlands

3 University of Bristol, United Kingdom

4 Fermi National Accelerator Laboratory, Batavia, Illinois, United States of America

5 Brookhaven National Laboratory BNL, Upton, New York, United States of America

6 INFN Bologna, Italy

7 Forschungszentrum Karlsruhe, Institut für Wissenschaftliches Rechnen, Karlsruhe, Germany

8 School of Physics and Chemistry, Lancaster University, United Kingdom

9 CCLRC, Rutherford Appleton Laboratory, Chilton, Didcot, United Kingdom

10 Department of Physics, University of Oslo, Norway

11 INFN Milano, Italy

12 SUBATECH Laboratoire de Physique Subatomique et des Technologies Associées, Nantes, France

ACKNOWLEDGEMENTS

The Editorial Board wishes to thank all members of the LCG Project, and especially the many who have contributed directly to the work described in this Technical Design Report. We would particularly like to thank Jean-Philippe Baud and Erwin Laure for the contributions to Section 4.1 (EGEE Middleware); Charles Curran, Gordon Lee, Alessandro Marchioro, Alberto Pace, and Dan Yocum for their work on the Pasta report that was the basis for Section 4.5 (Fabric Technology); Alberto Aimar, Ilka Antcheva, John Apostolakis, Gabriele Cosmo, Olivier Couet, Maria Girone, Massimo Marino, Lorenzo Moneta, Witold Pokorski, Fons Rademakers, Alberto Ribon, Stefan Roiser, and Rob Veenhof for their contributions to Section 5 (Common Applications). Special thanks are also due to the members of the Computing Groups of the LHC experiments who either directly contributed to the work described here or have provided essential feed-back.

We would also like to thank the people who have helped in the editing and production of this document. We warmly thank Fabienne Baud-Lavigne and Susan Leech O'Neale for copy-editing the document, Rosy Mondardini and Christine Vanoli for their work on designing the cover, and the CERN Printshop for their friendly and efficient service.

EXECUTIVE SUMMARY

This Technical Design Report presents the current state of planning for computing in the framework of the LHC Computing Grid (LCG) Project. The mission of LCG is to build and maintain a data storage and analysis infrastructure for the entire high-energy physics community that will use the Large Hadron Collider (LHC) [1].

The principal material and human resources of the project will be provided by a collaboration including the LHC experiments, the participating computer centres and CERN formalized in a Memorandum of Understanding [2].

The requirements of the experiments have been defined in Computing Model documents [3] of each of the experiments and have been refined in individual Computing Technical Design Reports [4]–[7] appearing in parallel with the present paper. The requirements for the year 2008 sum up to a CPU capacity of 140 million SPECint2000[1], to about 60 PB[2] of disk storage and 50 PB of mass storage.

The data from the LHC experiments will be distributed around the globe, according to hierarchical model. The raw data emerging from the data-acquisition systems will be recorded on tape and initially processed at the Tier-0 centre of LCG located at CERN. The data will be distributed to a series of Tier-1 centres, large computer centres with sufficient storage capacity for a large fraction of the data, and with round-the-clock operation. Analysis tasks requiring access to large subsets of the raw, processed, and simulated data will take place at the Tier-1 centres.

The Tier-1 centres will make data available to Tier-2 centres, each consisting of one or several collaborating computing facilities, which can store sufficient data and provide adequate computing power for end-user analysis tasks and Monte Carlo simulation. Individual scientists will also access these facilities through Tier-3 computing resources, which can consist of local clusters in a University Department or even individual PCs.

Developing common applications software for all experiments is an important part of the LCG Project. This includes core software libraries, tools and frameworks for data management and event simulation as well as infrastructure and services for software development, and the support of analysis and database management.

The Project has undertaken regular technology studies, examining the evolution of technologies in the areas of processing, storage, and networking. The progression of the market is also being followed and estimates made of the evolution of prices. The systems architecture is designed with the flexibility to incorporate new technologies that can bring improvements in performance and costs.

Excellent wide-area networking is essential to the operation of the LHC Computing Grid. The national and regional research networking organizations are collaborating closely with the Tier-1 regional centres and CERN in a working group to devise an appropriate architecture and deployment plan to match LHC requirements.

Data challenges and service challenges probe and evaluate current software and hardware solutions in increasingly demanding and realistic environments approaching the requirements of LHC data taking and analysis. The service challenges together with a clear definition and implementation of the basic services required by the experiments form the basis of the overall plan for establishing the full-scale global LHC Grid service for the start-up of LHC in 2007.

---

[1] SPECint2000 is an integer benchmark suite maintained by the Standard Performance Evaluation Corporation (SPEC). The measure has been found to scale well with typical HEP applications. As an indication, a powerful Pentium 4 processor delivers 1700 SPECint2000.

[2] A petabyte (PB) corresponds to 10^{15} Bytes or a million gigabytes.

The LCG Project depends upon several other projects for the supply of much of the specialized software used to manage data distribution and access as well as job submission and user authentication and authorization, known as the *Grid middleware*. These projects include Globus, Condor, the Virtual Data Toolkit and the gLite toolkit.

The majority of the computing resources made available by the members of the collaboration are operated as part of the EGEE Grid [8], a consortium of national Grid infrastructures and computing centres from 34 countries. Other resources are operated as part of other grids, such as the Open Science Grid (OSG) [9] in the United States and the Nordic Data Grid Facility (NDGF) [10]. Achieving interoperability between different grids without compromising on the functionality constitutes a major challenge.

CONTENTS

EXECUTIVE SUMMARY I
1 INTRODUCTION 1
2 EXPERIMENTS' REQUIREMENTS 4
2.1 Logical Dataflow and Workflow 4
2.2 Event Simulation 9
2.3 Resource Expectations 10
2.4 Baseline Requirements 13
2.5 Online Requirements 22
2.6 Analysis Requirements and Plans 24
2.7 Start-up Scenario 27
3 BASIC LCG ARCHITECTURE 29
3.1 Grid Architecture and Services 29
3.2 Tier-0 Architecture 34
3.3 Tier-1 Architecture 40
3.4 Tier-2 Architecture 45
4 TECHNOLOGY AND INFRASTRUCTURE 48
4.1 EGEE Middleware 48
4.2 Grid Standards and Interoperability 53
4.3 Grid Operations and Regional Centre Service-Level Agreements 54
4.4 Life-Cycle Support – Management of Deployment and Versioning 59
4.5 Fabric Technology — Status and Expected Evolution 63
4.6 Databases – Distributed Deployment 70
4.7 Initial Software Choices at CERN 77
4.8 Initial Hardware Choices at CERN 79
4.9 Hardware Life-Cycle 81
4.10 Costing 81
4.11 Networking 81
4.12 Security 88
5 COMMON APPLICATIONS 90
5.1 High-Level Requirements for LCG Applications Software 91
5.2 Software Architecture 92
5.3 Operating System Platforms 93
5.4 Core Software Libraries 94
5.5 Data Management 97
5.6 Event Simulation 99
5.7 Software Development Infrastructure and Services 103
5.8 Project Organization and Schedule 104
6 EXPERIENCE: PROTOTYPES AND EVALUATIONS 106
6.1 Data Challenges 106
6.2 Service Challenges 117
6.3 ARDA 120
7 PLANS 124
7.1 Baseline Services 124
7.2 Phase-2 Planning 124
8 PROJECT ORGANIZATION AND MANAGEMENT 128
8.1 High-Level Committees: 128
8.2 Participating Institutes 129
8.3 Interactions and Dependencies 130

8.4 Resources 133
GLOSSARY – ACRONYMS – DEFINITIONS 136
REFERENCES 141

1 INTRODUCTION

The LHC Computing Grid Project (LCG) was approved by the CERN Council on 20 September 2001 [11] to develop, build and maintain a distributed computing infrastructure for the storage and analysis of data from the four LHC experiments. The project was defined with two distinct phases. In Phase 1, from 2002 to 2005, the necessary software and services would be developed and prototyped, and in Phase 2, covering the years 2006–2008, the initial services would be constructed and brought into operation in time for the first beams from the LHC machine. Towards the end of Phase 1 it was foreseen that a Technical Design Report (TDR) would be produced presenting the planning for the second Phase in light of the experience gained during Phase 1.

Over the past year a Memorandum of Understanding [2] has been developed defining the Worldwide LHC Computing Grid Collaboration. The members of the collaboration are CERN, as host laboratory, and the major computing centres that commit to provide resources for LHC data storage and analysis. It also defines the organizational structure for the management, deployment and operation of these resources as Phase 2 of the LCG Project.

Within this framework, the present paper is the TDR for Phase 2 of the LCG Project.

Section 2 summarizes the requirements of the LHC experiments.

The Large Hadron Collider (LHC), which will start to operate in 2007, will produce roughly 15 Petabytes (15 million Gigabytes) of data annually, which thousands of scientists around the world will access and analyse. The mission of the LHC Computing Project (LCG) is to build and maintain a data storage and analysis infrastructure for the entire high-energy physics community that will use the LHC.

When the LHC accelerator is running optimally, access to experimental data needs to be provided for the 5,000scientists in some 500 research institutes and universities worldwide who are participating in the LHC experiments. In addition, all the data needs to be available over the estimated 15-year lifetime of the LHC. The analysis of the data, including comparison with theoretical simulations, requires of the order of 100,000 CPUs at 2004 measures of processing power. A traditional approach would be to centralize all of this capacity at one location near the experiments. In the case of the LHC, however, a novel globally distributed model for data storage and analysis — a computing Grid — was chosen because it provides several key benefits. In particular:

- The significant costs of maintaining and upgrading the necessary resources for such a computing challenge are more easily handled in a distributed environment, where individual institutes and participating national organizations can fund local computing resources and retain responsibility for these, while still contributing to the global goal.
- Also, in a distributed system there are no single points of failure. Multiple copies of data and automatic reassigning of computational tasks to available resources ensures load balancing of resources and facilitates access to the data for all the scientists involved, independent of geographical location. Spanning all time zones also facilitates round-the-clock monitoring and support.

Of course, a distributed system also presents a number of significant challenges. These include ensuring adequate levels of network bandwidth between the contributing resources, maintaining coherence of software versions installed in various locations, coping with heterogeneous hardware, managing and protecting the data so that it is not lost or corrupted over the lifetime of the LHC, and providing accounting mechanisms so that different groups have fair access, based on their needs and contributions to the infrastructure. These are some of the challenges that the LCG Project is addressing.

Section 3 describes the basic LCG system architecture.

The LCG Project will implement a Grid to support the computing models of the experiments using a distributed four-tiered model.

- The original raw data emerging from the data acquisition systems will be recorded at the Tier-0 centre at CERN. The maximum aggregate bandwidth for raw data recording for a single experiment (ALICE) is 1.25 GB/s. The first-pass reconstruction will take place at the Tier-0, where a copy of the reconstructed data will be stored. The Tier-0 will distribute a second copy of the raw data across the Tier-1 centres associated with the experiment. Additional copies of the reconstructed data will also be distributed across the Tier-1 centres according to the policies of each experiment.
- The role of the Tier-1 centres varies according to the experiment, but in general they have the prime responsibility for managing the permanent data storage — raw, simulated and processed data — and providing computational capacity for re-processing and for analysis processes that require access to large amounts of data. At present 11 Tier-1 centres have been defined, most of them serving several experiments.
- The role of the Tier-2 centres is to provide computational capacity and appropriate storage services for Monte Carlo event simulation and for end-user analysis. The Tier-2 centres will obtain data as required from Tier-1 centres, and the data generated at Tier-2 centres will be sent to Tier-1 centres for permanent storage. More than 100 Tier-2 centres have been identified.
- Other computing facilities in universities and laboratories will take part in the processing and analysis of LHC data as Tier-3 facilities. These lie outside the scope of the LCG Project, although they must be provided with access to the data and analysis facilities as decided by the experiments.

Section 4 discusses the current status of and choices for initial deployment of computing technology and infrastructure.

The basic technology of computing, hardware and software, is in a continuous state of evolution. Over the past 20 years there have been enormous advances in performance and capacity of all the basic components needed for LHC data handling — processors, storage and networking. At the same time the adoption of a flexible computing architecture by the community has made it possible to benefit fully from high-performance, low-cost components developed for the mass market. In the software area also there have been major advances in computing languages, development tools, and user interfaces that have been adopted by HEP experiments. On the other hand, equipment rapidly becomes obsolete and computing systems must be designed to be capable of continuous change, with components being installed and removed as the service continues to operate. Technology tracking is therefore an important aspect of the project, to prepare for the rapid deployment of more cost-effective solutions, and to recognize and evaluate interesting new developments.

Section 5 describes software packages that are being developed for the common use of several experiments.

CERN and the HEP community have a long history of collaborative development of physics applications software. The unprecedented scale and distributed nature of computing and data management at the LHC require that software in many areas be extended or newly developed, and integrated and validated in the complex software environments of the experiments. The Applications Area of the LCG Project is therefore concerned with developing, deploying and maintaining that part of the physics applications software and associated supporting infrastructure software that is common to all LHC experiments.

Section 6 summarizes prototype implementations of system components.

The LCG system has to be ready for use with full functionality and reliability from the start of LHC data taking. In order to ensure this readiness, the system is planned to evolve through a set of steps involving the hardware infrastructure as well as the services to be delivered. At each step the prototype Grid is planned to be used for extensive testing. The experiments use it to perform their *Data Challenges*, progressively increasing in scope and scale, stressing the LCG system with activities that are more and more similar to the ones that will be performed when the experiments are running. The service providers at CERN and in the Tier-1 and Tier-2 centres also use the prototype Grid for a series of *Service Challenges*, aimed at identifying reliability, performance, and management issues at an early stage, to allow robust solutions to be developed and deployed.

Section 7 lays out the plan for the second phase of the project.

A set of baseline services to be deployed for the initial period of LHC running has been recently agreed. Further Service Challenges will lead to a full-fledged service to be in place in April 2007.

Section 8 deals with the organization and management of the project, describes interactions with collaborating projects, and summarizes the current state of resource planning at CERN and in the Tier-1 and Tier-2 centres.

The LCG Project will collaborate and interoperate with other major Grid development projects, network providers and production environments around the world. Formal relationships between LCG and these organizations are covered in Section 8.3. One of the Grid operations environments is mentioned here as it is very closely associated with the LCG project.

The EGEE (Enabling Grids for E-SciencE) [8] project, partially funded by the European Union, has the goal of deploying a robust Grid infrastructure for science. It interconnects a large number of sites in 34 countries around the world, integrating several national and regional Grid initiatives in Europe, such as INFN Grid in Italy, DutchGrid in the Netherlands and GridPP in the UK. EGEE was designed as a project that would extend the Grid that had been set up by Phase 1 of the LCG Project during 2003, with the aim of starting with a working Grid and demanding applications, and evolving it to support other sciences. The operation of the LCG Grid therefore effectively merged with that of EGEE when the latter began in April 2004. In return EGEE provides very substantial funding for Grid operations and support at its partner sites, the large majority of which provide computing resources for the LHC experiments. The majority of the sites in the Worldwide LCG Collaboration will be connected to EGEE. The current phase of the EGEE project is funded to March 2006. It is expected that a second phase will be funded for a further two years, until March 2008. It is clearly important that the LCG Project continues to be closely involved in the operation of EGEE until the longer term funding model is understood.

The LCG Project also has relationships with other Grid operations environments, such as the Open Science Grid (OSG) [9] in the US and the Nordic Data Grid Facility (NDGF) [10]. The LCG Project has a formal relationship with the US experiment projects whose resources are connected to OSG, but not with OSG itself.

2 EXPERIMENTS' REQUIREMENTS

This section summarizes the salient requirements of the experiments that are driving the LCG Project resource needs and software/middleware deliverables. These requirements are discussed in greater detail in the individual experiments' computing TDRs, Refs. [4]-[7].

2.1 Logical Dataflow and Workflow

2.1.1 ALICE

The ALICE DAQ system receives raw data fragments from the detectors through Detector Data Links and copies specified fragments to the High-level Trigger (HLT) farm. Fragments are then assembled to constitute a raw event taking into account the information sent by the HLT. The raw data format is converted into AliRoot data objects. Two levels of storage devices constitute the ALICE raw data archival system. A large disk buffer sufficient to save the equivalent of one day of Pb–Pb data is located on the experiment site. The archiving to mass storage is not synchronized with data taking. The CASTOR software provides a coherent and unified view of the mass storage system to the applications.

The maximum aggregate bandwidth from the DAQ system to the mass storage system required by ALICE is 1.25 GB/s. This allows the transfer of heavy-ion events with an average size of 12.5 MB at a rate of 100 Hz. The p–p raw data have an average size of 1.0 MB and are recorded at an average rate of 100 MB/s.

The p–p raw data are immediately reconstructed at the CERN Tier-0 facility and exported to the different Tier-1 centres. For heavy-ion data, the quasi real-time processing of the first reconstruction pass, as it will be done for p–p data, would require a prohibitive amount of resources. ALICE therefore requires that these data be reconstructed at the CERN Tier-0 and exported over a four-month period after data taking. During heavy-ion data taking only pilot reconstructions and detector calibration activities will take place. Additional reconstruction passes (on average three passes over the entire set of data are considered in the ALICE computing model) for p–p and heavy-ion data will be performed at Tier-1s, including the CERN Tier-1.

Raw data will be available in two copies, one at the Tier-0 archive and one distributed at the Tier-1 sites external to CERN. The reconstruction process generates Event Summary Data (ESD) with raw data size reduction coefficient of 10 for heavy-ion and 25 for p–p. Two copies of the ESD are distributed in the Tier-1 sites for archive. Multiple copies of active data (a fraction of raw data, the current set of ESD) are available on fast-access storage at Tier-1 and Tier-2 sites.

Analysis is performed directly on ESD or on Analysis Object Data (AOD). Two analysis schemes are considered: scheduled and chaotic. The scheduled analysis tasks are performed mainly at Tier-1 sites on ESD objects and produce various different AOD objects. These tasks are driven by the needs of the ALICE Physics Working Groups. The chaotic analysis tasks (interactive and batch) are launched by single users predominantly on AODs but also on ESD. These tasks are mainly processed at Tier-2 sites. It is the responsibility of the Tier-2 sites to make the data available on disk to the collaboration, the archiving being the responsibility of the Tier-1 sites.

To date, seven sites (including CERN) have pledged Tier-1 services to ALICE and about 14 sites (including CERN) have pledged Tier-2 services. The amount of resources provided by the various Tier-1 and Tier-2 sites is very uneven with a few Tier-1s providing a relatively small contribution compared to others. CERN will host the Tier-0, a comparatively large Tier-1 with no archiving responsibility for the raw data (this will be done by the associated Tier-0), and a Tier-2. During the first-pass reconstruction, when all the CPU resources installed at

CERN are required for the Tier-0, no computing tasks typically allocated to Tier-1 and Tier-2 will be processed at CERN. The first-pass reconstruction for heavy-ion data is scheduled to take place during the four months following the heavy-ion data-taking period.

2.1.1.1 Non-event Data

There are two types of non-event data required for reconstruction and simulation — static and dynamic. Static data are collected in the Detector Construction Data Base (DCDB) and regroup all information available on the various items pertinent to the construction of the detector (performance, localization, identification, etc.). Dynamic data are collected in the Condition Data Base (CDB) and regroup the calibration and alignment information needed to convert the raw signals collected by the DAQ into physics parameters. The Detector Control System (DCS), the Experiment Control System (ECS), the DAQ, the Trigger and HLT systems each collect parameters describing the run conditions. All run information relevant for event reconstruction and analysis is obtained from a metadata database. Most of the condition databases need to be available online for High-level Trigger processing, as well as offline.

2.1.2 ATLAS

2.1.2.1 Principal Real Datasets

The source of the data for the computing model is the output from the Event Filter (EF). Data passing directly from the experiment to off-site facilities for monitoring and calibration purposes will be discussed only briefly, as they have little impact on the total resources required. The input data to the EF will require approximately 10×10 Gbps links with very high reliability (and a large disk buffer in case of failures). The average rate at which the output data is transferred to the first-pass processing facility requires a 320 MB/s link.

The baseline model assumes a single, primary stream containing all physics events flowing from the Event Filter to Tier-0. Several other auxiliary streams are also planned, the most important of which is a calibration hot-line containing calibration trigger events (which would most likely include certain physics event classes). This stream is required to produce calibrations of sufficient quality to allow a useful first-pass processing of the main stream with minimum latency. A working target (which remains to be shown to be achievable) is to process 50% of the data within 8 hours and 90% within 24 hours.

Two other auxiliary streams are planned. The first is an express-line of physics triggers containing about 5% of the full data rate. This will allow both the tuning of physics and detector algorithms and also a rapid alert on some high-profile physics triggers. The fractional rate of the express stream will vary with time, and will be discussed in the context of the commissioning. The second minor stream contains pathological events, for instance those that fail in the event filter.

On arrival at the input-disk buffer of the first-pass processing facility (henceforth known as Tier-0) the raw data file

- is copied to CASTOR tape at CERN;
- is copied to permanent mass storage in one of the Tier-1s.
- The calibration and alignment procedures are run on the corresponding calibration stream events;
- the express stream is reconstructed with the best-estimate calibrations available.

Once appropriate calibrations are in place, first-pass reconstruction (‘prompt’ reconstruction) is run on the primary event stream (containing all physics triggers), and the derived sets archived into CASTOR (these are known as the ‘primary’ datasets, subsequent reprocessing giving rise to better versions that supersede them). Two instances of the derived ESD are exported to external Tier-1 facilities; each Tier-1 site assumes principal responsibility for its

fraction of such data, and retains a replica of another equal fraction of the ESD for which another Tier-1 site is principally responsible. Tier-1 sites make the current ESD available on disk.[1] ESD distribution from CERN occurs at completion of the first-pass reconstruction processing of each file. As physics applications may need to navigate from ESD to RAW data, it is convenient to use the same placement rules for ESD as for RAW, i.e., if a site hosts specific RAW events, it also hosts the corresponding ESD. The derived AOD is archived via the CERN analysis facility and an instance is shipped to each of the external Tier-1s. The AOD copy at each Tier-1 is replicated and shared between the associated Tier-2 facilities and the derived TAG is archived into CASTOR and an instance is copied to each Tier-1. These copies are then replicated to each Tier-2 in full.

The Tier-1 facilities perform all later re-reconstruction of the RAW data to produce new ESD, AOD and primary TAG versions. They are also potential additional capacity to be employed if there is a backlog of first-pass processing at the Tier-0.

Selected ESD will also be copied to Tier-2 sites for specialized purposes. The AOD and TAG distribution models are similar, but employ different replication infrastructure because TAG data are database-resident. AOD and TAG distribution from CERN occurs upon completion of the first-pass reconstruction processing of each run.

2.1.2.2 Non-Event Data

Calibration and alignment processing refers to the processes that generate 'non-event' data that are needed for the reconstruction of ATLAS event data, including processing in the trigger/event filter system, prompt reconstruction and subsequent later reconstruction passes. These 'non-event' data are generally produced by processing some raw data from one or more subdetectors, rather than being raw data itself. Detector Control Systems (DCS) data are not included here. The input raw data can be either in the event stream (either normal physics events or special calibration triggers) or can be processed directly in the subdetector readout systems. The output calibration and alignment data will be stored in the conditions database, and may be fed back to the online system for use in subsequent data-taking, as well as being used for later reconstruction passes.

Calibration and alignment activities will involve resource-intensive passes through large amounts of data on the Tier-1s or even the Tier-0 facility. Some calibration will be performed online and require dedicated triggers. Other calibration processing will be carried out using the recorded raw data before prompt reconstruction of that data can begin, introducing significant latency in the prompt reconstruction at Tier-0. Further processing will be performed using the output of prompt reconstruction, requiring access to AOD, ESD and in some cases even RAW data, and leading to improved calibration data that must be distributed for subsequent reconstruction passes and user data analysis.

All of the various types of calibration and alignment data will be used by one or more of the ATLAS subdetectors; the detailed calibration plans for each subdetector are still evolving.

2.1.3 CMS

2.1.3.1 Event Data Description and Flow

The CMS computing model is described in detail in Ref. [6]. The CMS DAQ system writes DAQ-RAW events (1.5 MB) to the *High-level Trigger* (HLT) farm input buffer. The HLT farm writes RAW events (1.5 MB) at a rate of 150 Hz. RAW events are classified in O(50) *primary datasets* depending on their *trigger history* (with a predicted overlap of less than 10%). The primary dataset definition is immutable. An additional *express-line* (events that will be reconstructed with high priority) is also written. The primary datasets are grouped into O(10) *online streams* in order to optimize their transfer to the *Offline* farm and the subsequent

[1] At least one Tier-1 site proposes to host the entire ESD. This is not precluded, but the site would nonetheless, like every other Tier-1, assume principal responsibility for its agreed fraction of the ESD.

reconstruction process. The data transfer from HLT to the Tier-0 farm must happen in real time at a sustained rate of 225 MB/s.

Heavy-ion data at the same total rate (225MB/s) will be partially processed in real-time on the Tier-0 farm. Full processing of the heavy-ion data is expected to occupy the Tier-0 during much of the LHC downtime (between annual LHC p–p running periods).

The first event reconstruction is performed without delay on the Tier-0 farm which writes RECO events (0.25 MB). RAW and RECO versions of each primary dataset are archived on the Tier-0 MSS and transferred to a Tier-1 which takes custodial responsibility for them. Transfer to other Tier-1 centres is subject to additional bandwidth being available. Thus RAW and RECO are available either in the Tier-0 archive or in at least one Tier-1 centre.

The Tier-1 centres produce Analysis Object Data (AOD, 0.05 MB) (AOD production may also be performed at the Tier-0 depending on time, calibration requirements etc.), which are derived from RECO events and contain a copy of all the high-level physics objects plus a summary of other RECO information sufficient to support typical analysis actions (for example, re-evaluation of calorimeter cluster positions or track refitting, but not pattern recognition). Additional processing (*skimming*) of RAW, RECO and AOD data at the Tier-1 centres will be triggered by Physics Groups requests and will produce custom versions of AOD as well as TAGs (0.01 MB) which contain high-level physics objects and pointers to events (e.g., run and event number) and which allow their rapid identification for further study. Only very limited analysis activities from individual users are foreseen at the Tier-1 centre.

The Tier-1 centre is responsible for bulk re-processing of RAW data, which is foreseen to happen about twice per year.

Selected skimmed data, all AOD of selected primary streams, and a fraction of RECO and RAW events are transferred to Tier-2 centres which support iterative analysis of authorized groups of users. Grouping is expected to be done not only on a geographical but also on a logical basis, e.g., supporting physicists performing the same analysis or the same detector studies.

CMS will have about 6 Tier-1 centres and about 25 Tier-2s outside CERN. CERN will host the Tier-0, a Tier-1 (but without custodial responsibility for real data) and a Tier-2 which will be about 3 times a standard Tier-2. The CERN Tier-1 will allow direct access to about $1/6^{th}$ of RAW and RECO data and will host the simulated data coming from about $1/6^{th}$ of the CMS Tier-2 centre. The CERN Tier-2 will be a facility useable by any CMS member, but the priority allocation will be determined by the CMS management to ensure that it is used in the most effective way to meet the experiment's priorities; particularly those that can take advantage of its close physical and temporal location to the experiment.

2.1.3.2 Non-Event Data

CMS will have four kinds of non-event data: construction data, equipment management data, configuration data and conditions data.

Construction data includes all information about the subdetector construction up to the start of integration. It has been available since the beginning of CMS and has to be available for the lifetime of the experiment. Part of the construction data is duplicated in other kinds of data (e.g., initial calibration in the configuration data).

Equipment management data includes detector geometry and location as well as information about electronic equipment. They need to be available at the CMS experiment for the online system.

Configuration data comprises the subdetector-specific information needed to configure the front-end electronics. They are also needed for reconstruction and re-reconstruction.

Conditions data are all the parameters describing run conditions and logging. They are produced by the detector's frontend. Most of the conditions data stay at the experiment and are not used for offline reconstruction, but part of them need to be available for analysis

At the CMS experiment site there are two database systems. The *Online Master Data Storage* (OMDS) database is directly connected to the detector and makes available configuration data to the detector and receives conditions data from the Detector Control System. The *Offline Reconstruction Conditions DB ONline subset* (ORCON) database has information from the OMDS but synchronization between the two is automatic only as far as it concerns conditions data coming from the detector. Configuration data are manually copied from ORCON to OMDS. ORCON is automatically replicated at the Tier-0 centre to and from the *Offline Reconstruction Conditions DB OFFlineOFfline subset* (ORCOFF), the master copy for the non-event data system. The relevant parts of ORCOFF that are needed for analysis, reconstruction, and calibration activities are replicated at the various CMS computing centres using technologies such as those being discussed in the LCG3D project, see Section 4.6.

Estimates for the data volumes of the non-event data are being collected based on the anticipated use cases for each subsystem. This will be addressed in the first volume of the CMS Physics TDR. Although the total data volume is small compared to event data, managing it carefully and delivering it effectively is essential.

2.1.4 LHCb

The LHCb computing model is described fully in Ref. [7].

2.1.4.1 RAW Data

The LHCb events can be thought of as being classified in four categories: exclusive b sample, inclusive b sample, dimuon sample and D* sample[2]. The expected trigger rate after the HLT is 2 kHz. The b-exclusive sample will be fully reconstructed on the online farm in real-time and it is expected that two streams will be transferred to the CERN computing centre: a reconstructed b-exclusive sample at 200 Hz (RAW+rDST) and the RAW data sample at 2 kHz. The RAW event size is 25 kB, and corresponds to the current measured value, whilst there is an additional 25 kB associated with the rDST. LHCb expect to accumulate 2×10^{10} events per year, corresponding to 500 TB of RAW data.

2.1.4.2 Reconstruction

LHCb plan to reprocess the data of a given year once, after the end of data taking for that year, and then periodically as required. The reconstruction step will be repeated to accommodate improvements in the algorithms and also to make use of improved determinations of the calibration and alignment of the detector in order to regenerate new improved rDST information. Since the LHCC review of the computing model, a prototype rDST has been implemented that meets the 25 kB/event estimate.

2.1.4.3 Data Stripping

The rDST is analysed in a production-type mode in order to select event streams for individual further analysis. The events that pass the selection criteria will be fully re-reconstructed, recreating the full information associated with an event. The output of the stripping stage will be referred to as the (full) DST and contains more information than the rDST.

LHCb plan to run this production-analysis phase (stripping) four times per year: once with the original data reconstruction; once with the re-processing of the RAW data, and twice more, as the selection cuts and analysis algorithms evolve.

---

[2] It is appreciated that there will be events that satisfy more than one selection criterion; for the sake of simplicity this overlap is assumed negligible.

It is expected that user physics analysis will primarily be performed from the output of this stage of data processing (DST+RAW and TAG.) During first data-taking it is foreseen to have at least four output streams from this stripping processing: two associated with physics directly (b-exclusive and b-inclusive selections) and two associated with 'calibration' (dimuon and D* selections). For the b-exclusive and b-inclusive events, the full information of the DST and RAW will be written out and it is expected to need 100 kB/event. For the dimuon and D* streams only the rDST information will be written out, with the RAW information added; this is estimated to be 50 kB/event.

2.1.4.4 LHCb Computing Model

The baseline LHCb computing model is based on a distributed, multi-tier, regional centre model. It attempts to build in flexibility that will allow effective analysis of the data whether the Grid middleware meets expectations or not; of course this flexibility comes at the cost of a modest requirement overhead associated with pre-distributing data to the regional centres. In this section we shall describe a baseline model but we shall comment on possible variations where we believe this could introduce additional flexibility.

CERN is the central production centre and will be responsible for distributing the RAW data in quasi-real-time to the Tier-1 centres. CERN will also take on a role of a Tier-1 centre. An additional six Tier-1 centres have been identified: CNAF (Italy), FZK (Germany), IN2P3 (France), NIKHEF (The Netherlands), PIC (Spain) and RAL (United Kingdom) and an estimated 14 Tier-2 centres. CERN and the Tier-1 centres will be responsible for all the production-processing phases associated with the real data. The RAW data will be stored in its entirety at CERN, with another copy distributed across the six Tier-1s. The 2nd pass of the full reconstruction of the RAW data will also use the resources of the LHCb online farm. As the production of the stripped DSTs will occur at these computing centres, it is envisaged that the majority of the distributed analysis by the physicists will be performed at CERN and at the Tier-1s. The current year's stripped DST will be distributed to all centres to ensure load balancing. To meet these requirements there must be adequate networking not only between CERN and the Tier-1s but also between Tier-1s; quantitative estimates will be given later.

The Tier-2 centres will be primarily Monte Carlo production centres, with both CERN and the Tier-1s acting as the central repositories for the simulated data. It should be noted that although LHCb do not envisage any analysis at the Tier-2s in the baseline model, it should not be proscribed, particularly for the larger Tier-2 centres.

2.2 Event Simulation

While ALICE and CMS intend to produce the same number of Monte Carlo (MC) events as the real p–p and heavy-ion data, ATLAS and LHCb plan for about 20%. MC events will be produced and reconstructed in a distributed way mainly at the Tier-2 sites. The subsequent archiving and distribution of MC data is a collective responsibility of the Tier-1 sites. The data flow for the analysis of MC data and the availability requirements are the same as for real data. The simulated data are stored on at least one Tier-1 centre, which takes custodial responsibility for them.

For ALICE, the size of a raw MC event is 0.4 MB and 300 MB for p–p and heavy-ion, respectively. The size of a reconstructed MC ESD object is identical to that of real data.

The size of a simulated event for ATLAS and CMS is about 2 MB.

The LHCb simulation strategy is to concentrate on particular needs that will require an inclusive b-sample and the generation of particular decay modes for a particular channel under study. It is anticipated that 2×10^9 signal events will be generated plus an additional 2×10^9 inclusive events every year. About 10% of these simulated events will pass the trigger simulation and will be reconstructed and stored on MSS. The current event size of the Monte Carlo DST (with truth information) is approximately 500 kB/event. LHCb are confident that

this can be decreased to 400 kB/event. TAG data will be produced to allow quick analysis of the simulated data, with ~1 kB/event.

2.3 Resource Expectations

For the purpose of this document, the luminosity is to be $L = 2 \times 10^{33}$ $cm^{-2}s^{-1}$ in 2008 and 2009 and $L = 10^{34}$ $cm^{-2}s^{-1}$ in 2010. The canonical beam time for proton-proton (p–p) operations is assumed to be 50 days for 2007 and 10^7 seconds for 2008 and 2009. For heavy-ion running a beam time of 10^6 seconds is assumed with $L = 5 \times 10^{26}$ $cm^{-2}s^{-1}$.

2.3.1 ALICE

The total amount of resources required by ALICE for the production of Monte Carlo data and the processing of real data in a standard year of running, are summarized in Table 2.1. The staging of the offline resources is dictated by the LHC heavy-ion schedule and is synchronized with the ramp-up of the DAQ online resources. Before the first heavy-ion run, 20% of the nominal total resources are required to process p–p data and to produce Monte Carlo data. ALICE foresees an early and short heavy-ion pilot run before the end of 2007 for which 40% of the resources required at CERN must be available already in mid 2007 and 40% of the external resources must be available at the same time to take up the tasks of the CERN Tier-1 and Tier-2 which will not be available during the heavy-ion run and during the first-pass reconstruction of the heavy-ion data. The first heavy-ion run is foreseen during the last quarter of 2008. Even though beams with reduced luminosity with respect to the nominal LHC luminosity might be available, data will be taken at the maximum rate allowed by the DAQ nominal bandwidth. Therefore, CERN must have 100% of its expected resources installed at this time. The external resources are also required at the 100% level at this same time for the reasons already discussed earlier.

In 2009 and 2010, a 30% increase is requested in Tier-1 and Tier-2 resources (external as well as at CERN) to be able to cope with the reconstruction of the data of the running year and previous years.

The computing resources required at CERN include Tier-0, Tier-1 and Tier-2 type resources. The first-pass heavy-ion reconstruction is performed during the four months after the heavy-ion run at Tier-0 and requests all resources (7.5 MSI2000) available at the CERN Tier-0. During the same time, no Tier-1/2 resources are requested at CERN, the associated tasks being transferred to the external Tier-1/2's.

Table 2.1: Computing resources requested by ALICE in 2009 in Tier-0, at CERN (includes CERN Tier-0, Tier-1 and Tier-2), at all Tier-1/2 including CERN, and at external Tier-1/2 ext. excluding CERN.

CPU (MSI2000)	**2007**	**2008**	**2009**	**2010**
CERN Tier-0	3.3	8.3	10.8	14.0
CERN Tier-1/2				
Ext Tier-1s	4.9	12.3	16.0	20.9
Ext Tier-2s	5.8	14.4	18.7	24.3
Total	14.0	35.0	45.5	59.2
Disk (TB)				
CERN Tier-0	95	238	309	402
CERN Tier-1/2	579	1447	1882	2446
Ext. Tier-1s	2941	7353	9559	12426
Ext. Tier-2s	2042	5106	6638	8629
Total	5658	14144	18387	23903
MSS (TB)				
CERN Tier-0	990	2475	3218	4183
CERN Tier-1	463	1158	1505	1957
Ext. Tier-1s	2779	6947	9031	11740
Total	4232	10580	13754	17880

2.3.2 ATLAS

It is proposed that by the end of 2006 a capacity sufficient to handle the data from first running needs to be in place, with a similar ramping of the Tier-1 facilities. During 2007, an additional capacity required for 2008 should be bought. In 2008, an additional full-year of capacity should be bought, including the additional, archive storage medium (tape) required to cope with the growing dataset. This would lead to a capacity, installed by the start of 2008, capable of storing the 2007 and 2008 data as shown in Table 2.2; the table assumes that only 20% of the data rate is fully simulated.

Table 2.2: ATLAS computing resource estimates

CPU(MSI2000)	**2007**	**2008**	**2009**	**2010**
CERN Tier-0	0.91	4.06	4.06	6.41
CERN T1AF	0.49	2.65	4.18	7.00
All Tier-1s	4.07	23.97	42.93	72.03
All Tier-2s	3.65	19.94	31.77	53.01
Total	9.12	50.61	82.94	138.45
Disk(TB)				
CERN Tier-0	82	388	388	530
CERN T1AF	376	1851	2464	3550
All Tier-1s	2771	14434	22449	40614
All Tier-2s	1607	8748	15905	25815
Total	4835	25420	41206	70509
MSS (TB)				
CERN Tier-0	978	5683	10388	16768
CERN T1AF	104	524	795	1245
All Tier-1s	1508	8992	17985	31095
Total	2590	15199	29168	49107

For the Tier-2s, a slightly later growth in capacity, following the integrated luminosity, is conceivable provided that the resource-hungry, learning-phase is mainly consuming resources in Tiers 0 and 1. However, algorithmic improvements and calibration activity will also require

considerable resources early in the project. As a consequence, we have assumed the same ramp-up for the Tier-2s as for the higher Tiers.

Once the initial system is built, there will be a linear growth for several years in the CPU required for processing, as the initial datasets will require reprocessing as algorithms and calibration techniques improve. In later years, subsets of useful data may be identified to be retained/reprocessed, and some data may be rendered obsolete. However, for the near future, the assumption of linear growth is reasonable. For storage, the situation is more complex. The requirement exceeds a linear growth if old processing versions are not to be overwritten. On the other hand, as the experiment matures, increases in compression and selectivity over the stored data may reduce the storage requirements.

2.3.3 CMS

CMS resource requirements are summarized in Table 2.3. These estimates assume the 2007 pilot run to be about 50 days. The 2008 numbers are derived in a similar way as in the Computing Model paper, but with a somewhat increased emphasis on the re-reconstruction at the Tier-1 centres. In 2009 the CMS RAW event size will reduce to 1 MB due to better understanding of the detector. In 2010 running at high luminosity leads to an increased CPU requirement, currently estimated to be about a factor of 5, but the event size is expected to remain at 1 MB. The CMS CERN Analysis Facility (CMS-CAF), which has a special role due to its temporal and geographic closeness to the running experiment, is calculated as a standard Tier-1 having taken into account that the raw data are already on tape at CERN Tier-0, plus an additional 2.5 standard Tier-2s to allow for the analysis-like activities at the CMS-CAF. Resources of Tier-1s and Tier-2s outside CERN are integrated. We anticipate that CMS will make use of 7–10 physical Tier-1 centres and 20–25 physical Tier-2 centres.

Table 2.3: CMS computing resource estimates

CPU(MSI2000)	**2007**	**2008**	**2009**	**2010**
CERN Tier-0	2.3	4.6	6.9	11.5
CMS-CAF	2.4	4.8	7.3	12.9
All Tier-1s	7.6	15.2	20.7	40.7
All Tier-2s	9.6	19.3	32.3	51.6
Total	21.9	43.8	67.2	116.6
Disk(TB)				
CERN Tier-0	100	400	400	600
CMS-CAF	500	1500	2500	3700
All Tier-1s	2100	7000	10500	15700
All Tier-2s	1500	4900	9800	14700
Total	4100	13800	23200	34700
MSS (TB)				
CERN Tier-0	1100	4900	9000	12000
CMS-CAF	400	1900	3300	4800
All Tier-1s	3800	16700	29500	42300
Total	5400	23400	41500	59500

.

2.3.4 LHCb

Unlike the other experiments, LHCb assumes a luminosity of $L = 2 \times 10^{32}$ $cm^{-2}s^{-1}$ independent of year, achieved by appropriate de-focussing of the beam. It is anticipated that the 2008 requirements to deliver the computing for LHCb are 13.0 MSI2000-years of processing, 3.3 PB of disk and 3.4 PB of storage in the MSS. The CPU requirements will increase by 11% in 2009 and 35% in 2010. Similarly the disk requirements will increase by 22% in 2009 and 45% in 2010. The largest increase in requirements is associated with the MSS where a factor

2.1 is anticipated in 2009 and a factor 3.4 for 2010, compared to 2008. The requirements are summarized in Table 2.1. The estimates given in 2007 reflect the anticipated ramp-up of the computing resources to meet the computing requirements need in 2008; this is currently 30% of needs in 2006 and 60% in 2007. This ramp-up profile should cover the requirements of any data taken in 2007.

Table 2.4: LHCb computing resource estimates

CPU(MSI2000.yr)	**2007**	**2008**	**2009**	**2010**
CERN Tier-0	0.34	0.57	0.60	0.91
CERN T1/T2	0.20	0.33	0.65	0.97
All Tier-1s	2.65	4.42	5.55	8.35
All Tier-2s	4.59	7.65	7.65	7.65
Total	7.78	12.97	14.45	17.88
Disk(TB)				
CERN Tier-0	163	272	272	272
CERN T1/T2	332	454	923	1091
All Tier-1s	1459	2432	2897	3363
All Tier-2s	14	23	23	23
Total	1969	3281	4015	4749
MSS (TB)				
CERN Tier-0	300	500	1000	1500
CERN T1/T2	525	859	1857	3066
All Tier-1s	1244	2074	4285	7066
Total	2069	3433	7144	11632

2.4 Baseline Requirements

2.4.1 ALICE

The ALICE computing model makes the assumption that there will be a number of Grid services deployed on the centres providing resources to ALICE. Moreover, the model assigns specific classes of tasks to be performed by each class of Tier.

2.4.1.1 Distributed Computing

The ALICE computing model is driven by the large amounts of computing resources that will be necessary to store and process the data generated by the experiment and by the ALICE-specific requirement for data processing and analysis:

- large events in heavy-ion processing;
- wide variety of processing patterns, from progressive skimming of rare events to high-statistics analysis where essentially most of the events are read and processed.

The required resources will be spread over the HEP computing facilities of the institutes and universities participating in the experiment.

A large number of tasks will have to be performed in parallel, some of them following an ordered schedule, reconstruction, large Monte Carlo production, and data filtering, and some being completely unpredictable: single-user Monte Carlo production and data analysis. To be used efficiently, the distributed computing and storage resources will have to be transparent to the end user, essentially looking like a single system.

2.4.1.2 AliEn, the ALICE Distributed Computing Services

During the years 2000–2005 ALICE developed the AliEn (**Ali**Ce **En**vironment) framework with the aim of offering to the ALICE user community transparent access to computing resources distributed worldwide.

This system has served the ALICE user community very well for simulation and reconstruction, while a prototype for analysis has been implemented but not widely tested. AliEn implements a distributed computing environment that has been used to carry out the production of Monte Carlo data at over 30 sites on four continents. Less than 5% (mostly code in PERL) is native AliEn code, while the rest of the code has been imported in the form of open-source packages and PERL modules. The user interacts with the AliEn Web Services by exchanging SOAP messages and they constantly exchange messages between themselves behaving like a true Web of collaborating services.

AliEn has been primarily conceived as the ALICE-user entry point into the Grid world. Through interfaces it could use transparently resources of other Grids (such as LCG) that run middleware developed and deployed by other groups.

The AliEn architecture has been taken as the basis for the EGEE middleware, which is planned to be the source of the new components for the evolution of LCG-2, providing the basic infrastructure for Grid computing at the LHC.

Following this evolution, a new version of AliEn (AliEn II) has been developed. Like the previous one, this system is built around open-source components and uses the Web Services model and standard network protocols.

The new system has increased modularity and is less intrusive. It has been designed to solve the main problems that ALICE is facing in building its distributed computing environment, i.e., the heterogeneity of the Grid services available on the computing resources offered to ALICE. It will be run as ALICE application code complementing the Grid services implemented at the different centres. Wherever possible, maximum use will be made of existing basic Grid services, provided that they respond to the ALICE requirements.

2.4.1.3 AliEn Components

AliEn consists of the following key components: the authentication, authorization and auditing services; the workload and data management systems; the file and metadata catalogues; the information service; Grid and job monitoring services; storage and Computing Elements (CEs). These services can operate independently and are highly modular.

AliEn maintains a central ‘state-full’ task queue from where tasks can be ‘pulled’ either by the AliEn workload management system, or by the AliEn jobwrapper, once a job has been scheduled to run on a Computing Element (CE) by another Workload Management System (WMS).

The AliEn task queue is a central service that manages all the tasks, while Computing Elements are defined as ‘remote queues’ and can, in principle provide an entry into a single machine dedicated to running a specific task, a cluster of computers, or even an entire foreign Grid. When jobs are submitted, they are sent to the central queue. The queue can be optimized taking into account job requirements based on input files, CPU time, architecture, disk space, etc. This queue then makes jobs eligible to run on one or more Computing Elements. The active nodes get jobs from the queue and start their execution. The queue system monitors the progress of the job and has access to the standard output and standard error.

Input and output associated with any job are registered in the AliEn file catalogue, a virtual file system in which one or more logical names are assigned to a file via the association to its GUID. Unlike real file systems, the file catalogue does not own the files; it only keeps an association between the Logical File Name (LFN), file GUID (unique file identifier) and (possibly more than one) Physical File Names (PFN) on a real file or mass storage system.

The system supports file replication and caching and uses file location information when it comes to scheduling jobs for execution. These features are of particular importance, since similar types of data will be stored at many different locations. The AliEn file system associates metadata with GUIDs.

2.4.1.4 TAG Databases

ALICE is planning an event-level database. Work is being done in collaboration with ROOT and the STAR experiment at RHIC on this subject.

2.4.2 ATLAS

The ATLAS requirements are all centred on those exposed by the Baseline Services Working Group. This section gives a brief encapsulation of the required services and tools. They are also expressed in the HEPCAL and HEPCAL II reports. More details may be found in the ATLAS Computing TDR.

2.4.2.1 Interoperability

ATLAS computing must work in six continents in a coherent way. Such coherence comes in part from the experiment itself layering tools on top of native Grid deployments and in part from the interoperability of those deployments. Clearly, placing the burden completely on the experiment would make unsupportable demands on the experiment in terms of manpower for maintenance and development. It is therefore vital that

- the deployments for HEP ensure the highest possible degree of interoperability at the service and API level;
- changes in the APIs be well advertised in advance, and technical support be available to experiment developers where applicable.

It is expected that a primary focus of the LHC Computing Grid project will be in promoting the interoperability between deployments and encouraging convergence wherever possible.

2.4.2.2 Virtual Organizations

All members of the ATLAS Collaboration are authorized to become members of the ATLAS Virtual Organization (VO). It is important that a single VO operates across the HEP Grid deployments for ATLAS.

ATLAS will set up an internal system of priority allocation for the usage of CPU and data storage resources. The problem is multidimensional, as one can define at least three dimensions that can affect a job priority on a given site.

The VOMS (Virual Organization Management Service) middleware package must allow the definition of user groups and roles within the ATLAS Virtual Organization. We are initially setting up a VOMS group for each Physics Working Group, Combined Performance group and Detector System, as well as a generic one, and another one for software testing, validation and central production activities.

In order to implement efficiently the allocation sharing and priority system in the Grid access tools, it is necessary that the majority of the jobs be submitted through a central job queue and distribution system. The current implementation of the production system does not yet support relative job priorities, but this extension of functionality must be available soon.

The computing centres must allocate the bulk of the resources dedicated to ATLAS to be usable only according to the ATLAS policies for the various VO groups and roles. Those resources only will be accounted as having been provided to the ATLAS Collaboration, as there is no way to prioritize the use of independently provided (and used) computing resources.

Beyond the definition of groups and subgroups and the allocation of resources on that basis, uniform monitoring and accounting tools must work on a similar basis.

2.4.2.3 Data Management

The data management system fulfils two functions: global cataloguing of files and global movement of files. ATLAS initially opted to realize the global catalogue function by building on the existing catalogues of the three Grid flavours (Globus RLS in the case of NorduGrid and Grid3, EDG-RLS in the case of the LCG). The data management system acted as a thin layer channelling catalogue requests to the respective Grid catalogues and collecting/aggregating the answers. At the same time it presented the users with a uniform interface on top of the grid-native data management tools, both for the catalogue functions and the data movement functions. This has been proven not to scale to the current needs of the ATLAS experiment.

The future data management system will be based on a multitiered cataloguing system. The basic concept is to have each dataset (a coherent set of events, recorded or selected according to the same procedure) subdivided into a number of 'data blocks'; each data block can consist of many files, but is considered a single indivisible unit for what concerns data storage and transfer. Central catalogues only need to know the location of the Storage Elements where each data block is present (in addition of course to the relevant metadata), whereas the mapping to the physical file names is known only to local catalogues. All local catalogues have to present the same interface to the ATLAS data management system, but could in principle be implemented using different Grid-specific technologies.

Experience also showed that a large fraction of job failures were due to lack of robustness of the native Grid tools used for data cataloguing and data transfer, as well as to a lack of appropriate Storage Element space management tools. In order to build a robust Distributed Data Management system for ATLAS, it is necessary that:

- all sites deploy Storage Elements presenting the same interface (SRM), backed up by a reliable disk pool or mass storage management system, such as DPM, dCache or CASTOR;
- Grid middleware implements an infrastructure for storage quotas, based on user groups and roles[3];
- Grid middleware implements hooks for file placement policies depending on users' roles;
- all mass-storage based SEs provide information on file stage-in status (disk or tape) to be used for job scheduling and file pre-staging.

2.4.2.4 Job Submission

In order to be able to optimize job submission and distribution on the Grid, a certain number of middleware services must be provided. First of all, one needs a reliable information system so that job submission can take place to sites that support ATLAS, have the required system and software configuration, and have available capacity and network connectivity. Depending on the Grid flavour, other workload management services may be present, such as the Resource Broker on the LCG Grid; in all cases these services, in order to be useful, must be extremely reliable and scalable with large number of jobs and/or bulk job submission.

Information for job monitoring, accounting, and error reporting must be provided consistently by Grid middleware. It would help considerably if the reporting could use a uniform schema for all Grids. In particular, Grid middleware must:

- have the local job identifier returned at the time of submission to the submitter, to allow easy access to job status and matching of log files;

[3] This requirement will need further discussion with the storage system developers.

- make access to the standard job output possible while the job is running, to give single users as well as production managers a way to find out what happens in specific cases;
- report errors as they occur and not hide them behind a generic error type.

Job distribution on the Grid is done by a combination of Grid services (such as the Resource Broker for the EGEE Grid) and directives given by the submission system. In order to optimize job distribution and minimize data transfer and network activity, it would be useful to have ways to send jobs 'reasonably close' to where their input data reside. An exact, and dynamically updated implementation of a connectivity matrix between all Computing Elements and all Storage Elements is not needed, as network bandwidths can differ by orders of magnitude between local clusters and trans-oceanic links. Instead, a simple matrix of the 'distance' between CEs and SEs, even in rough terms (local, close, same continent, far-away), could be used in match-making and improve, at least on average, network traffic and job throughput.

2.4.3 CMS

The CMS computing system is geographically distributed. Data are spread over a number of centres following the physical criteria given by their classification into primary datasets. Replication of data is driven more by the need of optimizing the access to most commonly accessed data than by the need to have data 'close to home'. Furthermore Tier-2 centres support users not only on a geographical basis but mainly on a physics-interest basis.

CMS intends as much as possible to exploit solutions in common with other experiments to access distributed CPU and storage resources.

2.4.3.1 Access to Resources

The Computing Element (CE) interface should allow access to batch queues in all CMS centres independently of the User Interface (UI) from which the job is submitted. Mechanisms should be available for installing, configuring and verifying CMS software at remote sites. In a few selected centres CMS may require direct access to the system in order to configure software and data for specific, highly demanding processes such as digitization with pile-up of simulated data. This procedure does not alter the resource access mechanisms.

The Storage Element (SE) interface should hide the complexity and the peculiarities of the underlying storage system, possibly presenting to the user a single logical file namespace where CMS data may be stored. While we will support exceptions to this, we do not expect them to be the default mode of operation.

The technological choices to implement policies for disk space and CPU usage (including quotas and priorities) need to be flexible enough to reflect the structure of CMS as an organization, i.e., the definitions of groups and of roles.

The scheduling procedure should perform well enough to be able to keep all the CPUs busy even with modest duration of jobs. Given the foreseen number of processors ($O(10^4)$ in 2007), an average job duration of O(1) hours translates into a scheduling frequency of a few Hz.

2.4.3.2 Data Management

This section deals with management of event data only since non-event data will be discussed in detail in the CMS Physics TDR as anticipated in a previous section.

CMS data are indexed not as single files but as *Event-Collections*, which may contain one or more files. Event-Collections are the lowest granularity elements that may be addressed by a process that needs to access them. An Event-Collection may represent, for instance, a given data-tier (i.e. RAW or RECO or AOD, etc.) for a given primary dataset and for a given LHC fill. Their composition is defined by CMS and the information is kept in a central service provided and implemented by CMS: the *Dataset Book-keeping System* (DBS). The DBS

behaves like a Dataset Metadata Catalogue in HEPCAL [12] and allows all possible operations to manage CMS data from the logical point of view. All or part of the DBS may be replicated in read-only copies. Copies may use different back-ends depending on the local environment. Light-weight solutions like flat files may be appropriate to enable analysis on personal computers. In the baseline solution the master copy at the Tier-0 is the only one where updates may be made, we don't exclude that in future this may change. Information is entered in the DBS by the data-production system. As soon as a new Event-Collection is first made known to DBS, a new entry is created. Some information about production of the Event-Collection (e.g., the file composition, including their Globally Unique IDentifiers, GUIDs, size, checksum, etc.) may only be known at the end of its production.

A separate *Data Location System* (DLS) tracks the location of the data. The DLS is indexed by *file-blocks*, which are in general composed of many Event-Collections. The primary source of data location information is a *local index* of file-blocks available at each site. A *global data location index* maintains an aggregate of this information for all sites, such that it can answer queries on which file-blocks exist where. Our baseline is that information is propagated from the local index to the global one asynchronously. The queries against the global index are answered directly by the global index without passing the query to the local indices, and vice versa. Information is entered into DLS at the local index where the data are, either by the production system after creating a file-block or by the data transfer system (see below) after transfer. In both cases only complete file-blocks are published. Site manager operations may also result in modification of the local index, for instance in case of data loss or deletion. Once the baseline DLS has been proven sufficient we expect the DLS model to evolve.

Access to local data never implies access to the global catalogue; if data are found to be present locally (e.g., on a personal computer), they are directly accessible.

Note that the DLS provides only names of sites hosting the data and not the physical location of constituent files at the sites, or the composition of file-blocks. The actual location of files is known only within the site itself through a *Local File Catalogue*. This file catalogue has an interface (POOL [see paragraph 5.5]) which returns the physical location of a logical file (known either through its logical name which is defined by CMS or through the GUID). CMS applications know only about logical files and rely on this local service to have access to the physical files. Information is entered in the local file catalogue in a way similar to that of the local catalogue of the DLS, i.e. by the production system, by a data transfer agent or by the local site manager. Note that if the local SE may be seen as a single logical file namespace, the functionality of the catalogue may be implemented by a simple algorithm that attaches the logical file name as known by the CMS application to a site-dependent prefix that is provided by the local configuration. In this case no information needs to be entered when file-blocks are added or removed. This is the case for instance when data are copied to a personal computer (e.g., a laptop) for iterative analysis.

CMS will use a suitable DLS implementation able to co-operate with the workload management system (LCG WMS [see paragraph 4.1.2.2]) if it exists. Failing that, a CMS implementation will be used, with certain consequences on the job submission system (see below in the analysis section). An instance of the local index must operate on a server at each site hosting data; the management of such a server will be up to CMS personnel at the site. There may be a need to be able to contact a local DLS from outside the site, however, the local file catalogue conforming to the POOL API needs to be accessible only from within the site.

2.4.3.3 Data Transfer

Data transfers are never done as direct file copy by individual users. The data transfer system, Physics Experiment Data Export (PhEDEx [13]) consists of the following components:

- Transfer management database (TMDB) where transfer requests and subscriptions are kept.

- Transfer agents that manage the movement of files between sites. This also includes agents to migrate files to mass storage, to manage local mass storage staging pools, to stage files efficiently based on transfer demand, and to calculate file checksums when necessary before transfers.
- Management agents, in particular the *allocator* agent which assigns files to destinations based on site data subscriptions, and agents to maintain file transfer topology routing information.
- Tools to manage transfer requests, including interaction with local file and dataset catalogues as well as with DBS when needed.
- Local agents for managing files locally, for instance as files arrive from a transfer request or a production farm, including any processing that needs to be done before they can be made available for transfer: processing information, merging files, registering files into the catalogues, injecting into TMDB.
- Web-accessible monitoring tools.

Note that every data transfer operation includes a validation step that verifies the integrity of the transferred files.

In the baseline system a TMDB instance is shared by the Tier-0 and Tier-1s. Tier-2s and Tier-3 may share in the same TMDB instance or have site-local or geographically shared databases. The exact details of this partitioning will evolve over time. All local agents needed at sites hosting CMS data are managed by CMS personnel and are run on normal LCG user interfaces. The database requirements and CPU capacity for the agents is not expected to be significant. Between sites the agents communicate directly with each other and through a shared database. The amount of this traffic is negligible.

2.4.3.4 Production

Physics Groups submit data production requests to a central system (RefDB [14]), which behaves like a virtual data catalogue, since it keeps all the information needed to produce data. RefDB also has the information about the individual jobs that produced the data. Most of the information currently in RefDB will be moved to the DBS, leaving to RefDB only the management of information specific to the control of the production system and the data quality.

Data productions may happen on distributed or on local (e.g., Tier-1) resources. Once production assignments are defined by the CMS production manager, the corresponding jobs are created at the appropriate site, according to information stored in the RefDB. The tool that performs this operation is RunJob [15], but CMS is currently evaluating the possibility to use the same tool for data production and data analysis. Detailed job monitoring is provided by BOSS [16] at the site where the jobs are created. A summary of the logged information is also stored in RefDB.

Publication of produced data implies interaction with the DBS and with the local components of the DLS and the file catalogue at the site where the data are stored. Note that for Grid production this implies running the publication procedure at the site where the data are stored and not by the job that performed the production. Part of the publication procedure is the validation of the produced data, which is performed by the production system itself.

2.4.4 LHCb

The LHCb requirements for the LCG Grid are outlined in this section. LHCb expects to leverage from all the developments that were made in the past on its components in distributed computing, in particular DIRAC [17] and GANGA [18]. The baseline for GANGA is that it will use the services provided by DIRAC for job submission. The requirements of GANGA on the DIRAC services are an integral part of its design. Hence only DIRAC will

rely on externally provided Grid services. The details of DIRAC and GANGA are given in the LHCb Computing TDR [7].

2.4.4.1 Data Management Services

It is necessary to have a standard interface for all storage systems such that jobs can make use of them independently of where they are. We expect that SRM [19] will be the standard interface to storage, and hence a Grid Storage Element (SE) should be defined uniquely as an SRM front-end. As a consequence, Physical File Names (PFNs) are identified with Site URLs (SURLs).

In addition to storage, there is a need for a reliable file transfer system (*fts*). This reliable *fts* will thus permit the transfer between two SRM sites, taking care of the optimization of the transfer as well as of the recovery in case of failure (e.g., network).

At a higher level, replicas of files need to be registered in a File Catalogue (FC). Normal reference to a file by an application is via its Logical File Name (LFN). The FC fulfils two main functions:

- retrieve the SURL of a specific replica of a file at a given SE,
- information provider for the Workload Management System (WMS).

2.4.4.2 SRM Requirements

From the experience of the LHCb DC04, it is clear that the functionality of SRM v1.1 that is currently implemented on most storage systems is not sufficient. Hence we require that the SRM implementations be based on the protocol v2.1. The most urgent features needed in SRM are

- directory management,
- file management facilities (*get, put…*) with possibilities to define a lifetime for files on disk in case there is a MSS (pinning),
- space reservation (in particular in case of bulk replication),
- access control, allowing user files to be stored.

2.4.4.3 File Transfer System Requirements

As described in the LHCb computing TDR, the DIRAC system already has capabilities of reliable file transfer. The DIRAC transfer agent uses a local database of transfer requests from a local SE to any external SE(s). In addition it takes care of registration in the LHCb file catalogue(s). Currently the DIRAC transfer agent can use several transfer technologies, but GridFTP is the most commonly used. The LCG deployment team has provided a lightweight deployment kit of GridFTP in order to use it even on non-Grid-aware nodes.

When an *fts* is available and fully operational, LHCb is interested in replacing the current direct use of the GridFTP protocol by this *fts.* An implementation with a central request queue as currently implemented in the gLite FTS would be adequate, even if DIRAC keeps the notion of local agents for ensuring file registration.

2.4.4.4 File Catalogue

The requirements of LHCb in terms of the FC are fulfilled by most current implementations [20], [21]. They all differ by minor details as concerns the functionality, but we would like to have the opportunity to select the most suitable after appropriate tests of the access patterns implied by our Computing Model. In particular, the scalability properties of the FC services will be carefully studied.

We have developed an LHCb interface that the transfer agent uses and implemented it against several FCs. The aim is to populate all FCs with the few million entries we currently have and

continue populating them from the transfer agents. Performance tests will be performed with a central instance of a FC and with local read-only catalogue replicas, e.g., on Tier-1s. The most efficient and reliable FC will be selected as a first baseline candidate for the LHCb FC.

We do not consider there to be a need for standardization of all VOs on a single implementation provided the FC implements the interfaces needed by the WMS and the transfer service. A good candidate for WMS interface is one of the two currently available in gLite (LFC and FireMan) though only one will be selected.

2.4.4.5 Workload Management System

A lot of investment has gone into the LHCb production system, as well as into the analysis system (GANGA) for submitting and monitoring jobs through the DIRAC WMS. LHCb would like to keep DIRAC as the baseline for WMS.

The WMS needs interfacing to both the file catalogue and the Computing Element. The fact that DIRAC needs to interface to the Computing Element implies that some of the agents need to be deployed on the sites. This creates a number of requirements that are described below.

2.4.4.6 Computing Element Requirements

The definition adopted of a CE is that of a service implementing a standard interface to the batch system serving the underlying fabric. Jobs will be submitted, controlled, and monitored by the local DIRAC agent through this interface. Hence the following capabilities need to be implemented:

- Job submission and control, including setting CPU time limit.
- Proper authentication/authorization: the user credentials provided by the DIRAC agent should be used to allow jobs to be submitted with a mapped local userid.
- Batch system query: the DIRAC agent needs to have the possibility to query the batch system about its current load for the specific VO. Depending on the CPU sharing policy defined by the site, this may lead to fuzzy information that the agent should, however, use to determine if it is worth while requesting a job of a given type to the central WMS queue.

2.4.4.7 Hosting Computing Element

In order to be able to run local agents on the sites, we need to be able to deploy them on local resources at each site. The deployment is under the responsibility of LHCb. Deployed agents will run in user space without any particular privilege. However, proper authorization with a VO administrator role would be required for any action to be taken on the agents (launching, stopping, downloading).

In case agents need a particular infrastructure (e.g., local FCs), this infrastructure needs to be negotiated with the resource providers (e.g., if a specific database service is required). Similarly, the local storage on the node on which agents run will have to be negotiated.

We believe that a specialized instance of a CE limited to specific VOMS roles and giving access to its local CPU would be adequate provided it can be accessed from outside the site. The deployed agents would run under the VO responsibility and not require any particular intervention from the site besides regular fabric maintenance and survey. The VO would take responsibility for keeping the agents running.

The agents do not require incoming connectivity as they do not provide services to outside the site. The hosting node, however, needs outgoing connectivity in order to contact the central WMS, file catalogues, monitoring central services, etc.

For sites where a hosting CE would not be available, LHCb envisages to use, as it currently does on the LCG, pilot-agents submitted through a third party WMS (e.g., gLite RB) to the

sites. This is in particular valid for sites not connected to LHCb formally but which would grant resources to LHCb. It can also be applied to Grids not directly part of the LCG infrastructure. In this specific case, specific issues of authentication/authorization need to be addressed, in order for the job to be accounted to the actual owner of the job that is running, which could differ from the submitter of the pilot-agent.

2.4.5 Summary

All experiments require a single VO operating across all flavours of Grid, along with a well-defined way to specify roles, groups and subgroups via VOMS. Services to support long-lived agents running in user-space for file transfer or job submission, for example, are expected to be provided at the sites. ALICE would require limited inbound connectivity to support some of their services.

The experiments expect a standard interface for Grid access to SEs; it is envisaged this will be provided by SRM. Solutions must be provided for both large and small files, although the majority usage will be for large files. The interface must support usage quotas. In addition, it is expected that an SE will provide POSIX-like access to files for applications and present a single, logical file namespace. All experiments expect to have an experiment-specific central data catalogue containing the LFN with corresponding metadata. In addition, three experiments expect that there will be a local file catalogue that will map the LFN to a PFN, LHCb are currently planning to use a centralized catalogue (though expect replica(s) to be hosted to provide redundancy) to fulfil this functionality. CMS foresees it will perform data location through a two-tier service with a local component that publishes local file blocks and a global component that caches their locations. All experiments expect a reliable file transfer to be provided by the LCG Project, although other mechanisms may need to be supported.

The experiments require a common interface to the CE, with the possibility to autonomously install application software and to publish VO-specific information. WNs must present a standard environment configuration. LHCb, in particular, wish the WNs to have outgoing local and global network access for the site. There is a strong preference, from all experiments, that there should be an outbound network connection from the Worker Nodes, or at least the appropriate tunnelling mechanism. Some experiments request that a ‘pilot agent’ mechanism (an agent that runs basic tests on a WN before downloading a job from WMS to execute) be allowed.

It is required that a reliable Workload Management System be provided. The main requirements are that it must be able to efficiently exploit the distributed resources; that it is able to cope with the foreseen job submission frequency; that it is able to handle bulk job submission; and that it supports usage quotas and prioritization.

It is important that the WMS has a reliable information system for the WMS matching and resource monitoring service that includes a connectivity matrix between all CEs and SEs to be available for match making. There should be a consistent schema across Grids for job monitoring, accounting and error reporting and the WMS shoud reliably report at least the exit code of the application.

2.5 Online Requirements

2.5.1 ALICE

The ALICE DAQ has its own database based on MySQL. The DBMS servers will be located at the experimental area and operated by the experiment. It is not envisaged that the DAQ computing will be used for reprocessing or Monte Carlo production. Outside of the data-taking period they will be used for calibration as well as tests.

The maximum aggregate bandwidth from the DAQ system to the mass storage system required by ALICE is 1.25 GB/s, which allows transferring the trigger mix required by

ALICE. It corresponds to an average size of 12.5 MB at an average rate of 100 Hz. The p–p raw data have an average size of 1.0 MB and are recorded at an average rate of 100 MB/s.

2.5.2 ATLAS

The similar, and very large, overall CPU capacities of the ATLAS Event Filter (EF) and of the ATLAS share of the Tier-0 centre suggest that one should explore the technical feasibility of a sharing between the two. In steady-state running, if both systems have been designed correctly, this will not be relevant, as both systems will be running at close to full capacity. In particular, there will be times when the EF farm will be under-used, for example in LHC shutdowns.

It is envisaged, after the initial start-up, that data will typically be reprocessed a couple of months after the initial reconstruction, and after a year again. This work will be done primarily at Tier-1 sites. If these reprocessing periods, which will last many months, coincide with long LHC shutdowns, the Tier-0 site could also be able to assist with reprocessing, and the EF nodes could provide a valuable additional resource. However, various caveats have to be made though most of these questions seem to be tractable; ATLAS is currently keeping the option open to use the EF for data reprocessing: this potentially has implications for network and EF CPU/memory configurations, but we expect the impact to be manageable.

Given that the EF would be available for part of the time only, that it is unclear how often reprocessing will be scheduled in shutdowns that it is not known how much and how often EF nodes will be needed for development work; and that a clean switch-over with automated system management tools has not been demonstrated for this application, we do not assume for computing model calculations that the EF will be available for reprocessing. It is important to plan for a full capacity for reprocessing even in the event that the EF will not be available.

ATLAS will make extensive use of LCG database software and services to satisfy online database requirements, starting with subdetector commissioning activities in spring 2005. The LCG COOL conditions database software will be used as the core of the ATLAS conditions database, to store subdetector configuration information, detector control system data, online book-keeping, online and offline calibration and alignment data, and monitoring information characterizing the performance of the ATLAS detector and data acquisition system.

As well as the COOL core software, online use will be made of the POOL Relational Access Layer (RAL) to provide a uniform interface to underlying database technologies both online and offline, and POOL streamed file and relational database storage technologies for calibration and alignment data. Some use of these technologies has already been made in the 2004 ATLAS combined test beam, and their use will ramp up rapidly as subdetector commissioning gets under way.

The online conditions database will be based on Oracle server technology, deployed in collaboration with CERN-IT, and linked through the tools being developed by the LCG 3D project to the worldwide ATLAS conditions database, replicated to Tier-1 and Tier-2 centres as necessary. A first online conditions database server is being deployed in Spring 2005, and we expect the service to rapidly increase in capacity to keep pace with ATLAS commissioning needs. An important issue for ATLAS online will be to ensure scalability and high-enough performance to serve the many 1000s of online and high-level-trigger processors needing configuration and calibration data, and ATLAS is closely following the progress of 3D to see if the scalability and replication tools being developed can also be utilized in the online context.

2.5.3 CMS

The CMS databases located at CERN will be based on Oracle. The data model foresees a highly normalized online database (OMDS-Online Master Data Storage) at the experiment holding all data needed for running the detector and receiving status information created by

the detector. In the same physical instance a structural copy of the offline database called ORCON (Offline ReConstruction ONline copy) will be located, acting as a cache between OMDS and the Tier-0 conditions DB (ORCOFF - Offline ReConstruction OFFlineOFfline copy). The data needed offline will be projected from the OMDS onto a denormalized flat view in ORCON.

Depending on the API chosen to retrieve the conditions in the offline software, the ORCON/ORCOFF schema and the OMDS-ORCON transfer mechanism will be adapted accordingly.

A first Oracle DBMS has been set up at the experiment's site to serve for the combined detector test foreseen in early 2006. The server will be filled with a realistic CMS dataset to study the access patterns and the resulting performances. These tests will be used to define the hardware layout of the final DBMS.

It is hoped that the Oracle service could be supported centrally by CERN. It is not currently foreseen to use the CMS online system for reprocessing or Monte Carlo production.

2.5.4 LHCb

The LHCb online system will use a database based on Oracle with the server based in the experimental area. Tools will be needed to export the conditions database from the online system to the Tier-0 and subsequently dissemination of the information to the Tier-1 centres. Tools are also needed that allow the configuration data that are produced outside the online system (e.g., alignment data) to be imported into the configuration database. LHCb will develop the final tools (as only LHCb will know the configuration) but it is hoped that LCG could provide a certain infrastructure, such as notification mechanisms, etc. The online software will rely on a proper packaging of the LCG software, such that the LCG-AA software does not have Grid dependencies.

A b-exclusive sample will be fully reconstructed on the online farm in real-time and it is expected two streams will be transferred to the CERN computing centre: a reconstructed b-exclusive sample at 200 Hz (RAW+rDST) and the RAW data sample at 2 kHz. This would correspond to a sustained transfer rate of 60MB/s, if the data is transferred in quasi real-time. The CPU capacity that will be available from the Event Filter Farm corresponds to a power of ~5.55 MSI2000 so LHCb anticipate using the online farm during reprocessing outside of the data-taking period. This will allow 42% of the total reprocessing and subsequent stripping to be performed there. Hence the RAW data will also have to be transferred to the pit; similarly the produced rDST and stripped DSTs will have to be transferred back to the CERN computing centre and then distributed to the Tier-1 centres. Given the compressed timescale of 2 months, the transfer rate between the Tier-0 and the pit is estimated to be ~90 MB/s.

2.6 Analysis Requirements and Plans

The LCG Grid Applications Group (GAG) has provided an extensive discussion on the definition of analysis on the Grid. This activity has been summarized in the HEPCAL2 document [12].

A first scenario, analysis with fast response time and high level of user influence, is important for interactive access to event-data event display programs, for example. Here the user can interact effectively with the system because the size of the relevant data is minimal and all the computation can be performed locally.

Rather different is the well-known batch-system model where the user is faced with long response times and a low level of influence. The response time is the sum of three components: the initial submission latency, the queuing time and the actual job execution time.

As discussed in Ref. [12] interesting scenarios are within the transition area between the two extremes.

Detailed use cases are described in the Particle Physics Data Grid (PPDG) CS11 document 'Grid Service Requirements for Interactive Analysis' [22].

A common view is that the location of the data will determine the place where analysis is performed. This means that user jobs will be executed where a copy of the required data is already present at submission time. The problem is then *just* the normal fair-share of a batch system, which is by no means a trivial task if multiple users and groups are active at the same time. Even in this minimalist scenario, the experiments must be allowed to place datasets on a given set of Tiers. Conversely, users working at a given site should be able to stage-in data from other facilities prior to starting important analysis efforts.

In addition, 'random access' to relatively rare events out of a large number of files through event directories or TAGs must be provided. These schemes will allow fast pre-filtering based on selected quantities such as trigger flags or selection criteria, or reconstructed quantities. Technology choices have not been finalized yet but solutions such as POOL collections exist and should be considered in the baseline solution.

The existing experiments' frameworks (e.g., Cobra in CMS) allow users to navigate across different event components (AOD→RECO→RAW). It should be possible to implement control mechanisms to prevent unintentional excessive resource consumption such as the reloading of large datasets. It is assumed that such mechanisms be provided and implemented by the experiments.

Collaborative tools integrated with the analysis tools have recently met considerable interest. Some of the tools developed for detector operation could be also of interest for physics analysis. In this category are, for example, tools developed in the context of the WP5 of the GRIDCC project. Although analysis is likely to remain an individualistic activity, tools will be needed to enable detailed comparisons inside geographically dispersed working groups.

The experiments have to adapt to the coexistence of different operating systems and hardware architectures. In the case of the production activities, the experiments are already providing solutions to handle heterogeneous environments. For analysis, it will be important to shield the users from the infrastructures details.

A considerable effort is being put into understanding the analysis scenarios by both the experiments and the project itself. The LCG ARDA project is developing concrete prototype systems with the experiments and exposing them to pilot users as explained in Section 6.3.

It is expected that the choices for initial systems will be made in the near future, using the current experience.

2.6.1 Batch-Oriented Analysis

All experiments will provide frameworks for batch-oriented analysis simplifying the task to submit multiple jobs to a large set of files. As an example, GANGA (a common project of ATLAS and LHCb) is providing this functionality by allowing the user to prepare and run programs via a convenient user interface. The (batch) job can already be tested on the local resources. At user request, through the same interface, the user can take advantage of the available Grid resources (data storage and CPU power) typically to analyse larger data samples: GANGA is providing seamless access to the Grid and identifies all necessary input files, submits them to run in parallel, and provides an easy way to control the full set of jobs and to merge outputs at the end.

Such tools are necessary owing to the very large number of input files required by even simple studies. For example, with a raw data rate of 100 MB/s about 100 files of 1 GB size are produced in 15 minutes of data taking. The tools should also provide efficient Grid access to data for bulk operations.

The tools will also be important in the preparation of large skims, where the new data have to be shared across large working groups. Users or working groups must be able to publish their data without impacting the production catalogues holding the repository of all official data.

2.6.2 Interactive Analysis

Interactive tools for analysis such as PAW and ROOT have proven to be very powerful and popular inside the user community since the LEP era. All experiments are already using this approach for analysing simulated samples and test beam data. PROOF, a prototype extending ROOT to run on a distributed, heterogeneous system is described in Section 6.3.5. Another approach for the high-level analysis services, DIAL, is are under consideration in ATLAS.

2.6.3 ALICE

ALICE uses AliEn services and the ARDA end-to-end to realize distributed analysis on the Grid. Two approaches are being followed: the asynchronous (interactive batch approach) and the synchronous (true interactive) analysis model.

The asynchronous model has been implemented using the AliEn services and by extending the ROOT functionality to make it Grid-aware. As the first step, the analysis framework has to extract a subset of the datasets from the file catalogue using metadata conditions provided by the user. The next part is the splitting of the tasks according to the location of datasets. Once the distribution is decided, the analysis framework submits sub-jobs to the workload management with precise job descriptions. The framework collects and merges on request available results from all terminated sub-jobs.

The synchronous analysis model requires a tighter integration between ROOT, the Grid framework and the AliEn services. This has been achieved by extending the functionality of PROOF — the parallel ROOT facility. Rather than transferring all the input files to a given execution node, it is the program that is transferred to the nodes where the input data is locally accessible. The interface to Grid-like services is currently being developed.

2.6.4 ATLAS

The analysis activity is divided into two components. The first one is a scheduled activity analysing the ESD and other samples and extracting new TAG selections and working group enhanced AOD sets or ntuple equivalents. The jobs involved will be developed at Tier-2 sites using small subsamples in a chaotic manner, but will be approved for running over the large datasets by physics group organizers. The second class of user analysis is chaotic in nature and run by individuals. It will be mainly undertaken in the Tier-2 facilities, and includes direct analysis of AOD and small ESD sets and analysis of Derived Physics Datasets (DPDs). We envisage ~30 Tier-2 facilities of various sizes, with an active physics community of ~600 users accessing the non-CERN facilities. The CERN Analysis Facility will also provide chaotic analysis capacity, but with a higher-than-usual number of ATLAS users (~100). It will not have the simulation responsibilities required of a normal Tier-2.

2.6.5 CMS

While interactive analysis is foreseen to happen mainly locally at Tier-2/3 or on personal computers, in general, batch processing of data happens on the distributed system. The mechanism that CMS foresees to use is similar to the one described as “Distributed Execution with no special analysis facility support” in the HEPCAL-II document [12].

A user provides one or more executables with a set of libraries, configuration parameters for the executables (either via arguments or input files), and the description of the data to be analysed. Additional information may be passed to optimize job splitting, for example an estimation of the processing time per event. A dedicated tool running on the User Interface, the CMS Remote Analysis Builder (CRAB) [23], queries the DBS and produces the set of jobs to be submitted. In the baseline solution an additional query to the DLS selects the sites hosting the needed data. This translates to an explicit requirement to the WMS for a possible

set of sites in the job description (JDL file). In future the query to the DLS may be replaced by the WMS itself if a compatible interface between the DLS and the WMS is provided. Jobs are built in a site-independent way and may run on any site hosting the input data. CRAB takes care of defining the list of local files that need to be made available on the execution host (*input sandbox*) and those that have to be returned to the user at the end of execution (*output sandbox*). The user obviously has the possibility to specify that the data are local and that the job has to be submitted to a local batch scheduler or even forked on the current machine. In this case CRAB has the responsibility to build the jobs with the appropriate structure. Given the possibly large number of jobs resulting from the job-splitting procedure, it should be possible to submit the job cluster to the LCG WMS as a unique entity, with optimization in the handling of the input sandboxes. Single-job submission should also be possible. The WMS selects the site at which to run each job depending on load balancing only. As anticipated in the Data Management section the translation of logical file names to physical file names happens through a POOL catalogue interface on the Worker Node (WN).

Job cluster submission and all interactions with the cluster or with its constituent jobs happen through an interface (Batch Object Submission System, BOSS [16]) which hides the complexity of the underlying batch scheduler, in particular whether it is local or on the Grid. This layer allows the submitting and cancelling of jobs and clusters, the automatic retrieving of their output, getting information about their status and history. Furthermore it logs all information, either related to running conditions or specific to the tasks they performed, in a local database. The book-keeping database back-end may vary depending on the environment (e.g., performing RDBMS like Oracle for production systems, SQLite for personal computers or laptops). If outbound connectivity is provided on the WNs or if a suitable tunnelling mechanism (e.g., HTTP proxy, R-GMA *servlets*, etc.) is provided on the CE, a job submitted through BOSS may send information to a monitoring service in real-time and be made available to the BOSS system. Otherwise logging information is available only at the end of job execution (through the job output sandbox). Note that BOSS does not require any dedicated service on the sites where the jobs run.

2.6.6 LHCb

The LHCb physicists will run their physics analysis jobs, processing the DST output of the stripping on events with physics-analysis event-tags of interest and processing algorithms to reconstruct the B decay channel being studied. Therefore it is important that the output of the stripping process be self-contained. This analysis step generates quasi-private data (e.g., ntuples or personal DSTs), which are analysed further to produce the final physics results.

Since the number of channels to be studied is very large, we can assume that each physicist (or small group of physicists) is performing a separate analysis on a specific channel. These ntuples could be shared by physicists collaborating across institutes and countries, and therefore should be publicly accessible.

2.7 Start-up Scenario

The data processing in the very early phase of data taking will only slowly approach the steady-state model. While the distribution and access to the data should be well prepared and debugged by the various data challenges, there will still be a requirement for heightened access to raw data to produce the primary calibrations and to optimize the reconstruction algorithms in light of the inevitable surprises thrown up by real data.

The steady-state model has considerable capacity for analysis and detector/physics group files. There is also a considerable planned capacity for analysis and optimization work in the CERN Analysis Facility. It is envisaged that in the early stages of data-taking, much of this is taken up with a deep copy of the express- and calibration-stream data. For the initial weeks, the express data for ATLAS or CMS may be upwards of 20 Hz, but it is clear that averaged over the first year, it must be less than this, about 10 Hz. Given the resource requirements,

even reprocessing this complete smaller sample will have to be scheduled and organized through the physics/computing management. Groups must therefore assess carefully the required sample sizes for a given task. If these are small enough, they can be replicated to Tier-2 sites and processed in a more *ad hoc* manner there. Some level of *ad hoc* reprocessing will of course be possible on the CERN Analysis Facility.

The CERN Analysis Facility resources are determined in the computing model by a steady-state mixture of activities that include AOD-based and ESD-based analysis and steady-state calibration and algorithmic development activities. The resources required for the initial year of data taking are given in the tables in Section 2.3 for each experiment. This resource will initially be used far more for the sort of RAW-data-based activity described in Sections 2.1 and 2.2, but must make a planned transition to the steady state through the first year. If the RAW data activities continue in the large scale for longer, the work must move to be shared by other facilities. The Tier-1 facilities will also provide calibration and algorithmic development facilities throughout, but these will be limited by the high demands placed on the available CPU by reprocessing and ESD analysis.

There is considerable flexibility in the software chain in the format and storage mode of the output datasets. For example, in the unlikely event of navigation between ESD and RAW proving problematic when stored in separate files, they could be written to the same file. As this has major resource implications if it were adopted as a general practice, this would have to be done for a finite time and on a subset of the data. Another option that may help the initial commissioning process is to produce DRD, which is essentially RAW data plus selected ESD objects. This data format could be used in the commissioning of some detectors where the overhead of repeatedly producing ESD from RAW is high and the cost of storage of copies of RAW+ESD would be prohibitive. In general, the aim is to retain flexibility for the early stage of data taking in both the software and the processing chain and in the use of the resources available.

In order that the required flexibility be achievable, it is essential that the resources be in place in a timely fashion, both in 2007 and 2008.

3 BASIC LCG ARCHITECTURE

3.1 Grid Architecture and Services

The LCG architecture will consist of an agreed set of services and applications running on the Grid infrastructures provided by the LCG partners. These infrastructures at the present consist of those provided by the Enabling Grids for E-scienceE (EGEE) project in Europe, the Open Science Grid (OSG) project in the U.S.A. and the Nordic Data Grid Facility in the Nordic countries. The EGEE infrastructure brings together many of the national and regional Grid programmes into a single unified infrastructure. In addition, many of the LCG sites in the Asia-Pacific region run the EGEE middleware stack and appear as an integral part of the EGEE infrastructure. At the time of writing (April 2005) each of these projects is running different middleware stacks, although there are many underlying commonalities.

The essential Grid services should be provided to the LHC experiments by each of these infrastructures according to the needs of the experiments and by agreement between LCG, the sites, and the experiments as to how these services will be made available. The set of fundamental services are based on those agreed and described by the Baseline Services Working Group [24]. Where a single unique implementation of these services is not possible, each infrastructure must provide an equivalent service according to an agreed set of functionalities, and conforming to the agreed set of interfaces. These services and other issues of interoperability are discussed in this section and also in the discussion on Grid operations (Sections 4.2 and 4.3).

In the discussion below, LCG-2 refers to the set of middleware currently deployed on the EGEE Grid.

3.1.1 Basic Tier-0–Tier-1 Dataflow

The dataflow assumed in this discussion is that described in the experiment computing models. Data coming from the experiment data acquisition systems is written to tape in the CERN Tier-0 facility, and a second copy of the raw data is simultaneously provided to the Tier-1 sites, with each site accepting an agreed share of the raw data. How this sharing will be done on a file-by-file basis will be based on experiment policy. The File Transfer Service (FTS) will manage this data copy to the Tier-1 facilities in a reliable way, ensuring that copies are guaranteed to arrive at the remote sites. As this data arrives at the Tier-1, it must ensure that it is written to tape and archived in a timely manner. Copies arriving at the Tier-1 sites should trigger updates to the relevant file and data location catalogues.

Raw data at the Tier-0 will be reconstructed according to the scheme of the experiment, and the resulting datasets also distributed to the Tier-1 sites. This replication uses the same mechanisms as above and again includes ensuring the update of relevant catalogue entries. In this case, however, it is anticipated that all reconstructed data will be copied to all of the Tier-1 sites for that experiment.

3.1.2 Grid Functionality and Services

The set of services that should be made available to the experiments have been discussed and agreed in the Baseline Services Working Group set up by the LCG Project Execution Board in February 2005. The report of the group [25] identified the services described here. The full details of the services, the agreed set of functionality, and the interfaces needed by the experiments is described fully in the report of the working group.

3.1.3 Storage Element Services

A Storage Element (SE) is a logical entity that provides the following services and interfaces:

- Mass storage system, either disk cache or disk cache front-end backed by a tape system. Mass storage management systems currently in use include CASTOR, Enstore-dCache, HPSS and Tivoli for tape/disk systems, and dCache, LCG-dpm, and DRM for disk-only systems.
- SRM interface to provide a common way to access the MSS no matter what the implementation of the MSS. The Storage Resource Manager (SRM) defines a set of functions and services that a storage system provides in an MSS-implementation independent way. The Baseline Services working group has defined a set of SRM functionality that is required by all LCG sites. This set is based on SRM v1.1 with additional functionality (such as space reservation) from SRM v2.1. Existing SRM implementations currently deployed include CASTOR-SRM, dCache-SRM, DRM/HRM from LBNL, and the LCG dpm.
- GridFTP service to provide data transfer in and out of the SE to and from the Grid. This is the essential basic mechanism by which data is imported to and exported from the SE. The implementation of this service must scale to the bandwidth required. Normally the GridFTP transfer will be invoked indirectly via the File Transfer Service or through srmcopy.
- Local POSIX-like input/output facilities to the local site providing application access to the data on the SE. Currently this is available through rfio, dCap, aiod, rootd, according to the implementation. Various mechanisms for hiding this complexity also exist, including the Grid File Access Library in LCG-2, and the gLiteIO service in gLite. Both of these mechanisms also include connections to the Grid file catalogues to enable an application to open a file based on LFN or GUID.
- Authentication, authorization and audit/accounting facilities. The SE should provide and respect ACLs for files and datasets that it owns, with access control based on the use of extended X509 proxy certificates with a user DN and attributes based on VOMS roles and groups. It is essential that a SE provide sufficient information to allow tracing of all activities for an agreed historical period, permitting audit on the activities. It should also provide information and statistics on the use of the storage resources, according to schema and policies to be defined.

A site may provide multiple SEs providing different qualities of storage. For example, it may be considered convenient to provide an SE for data intended to remain for extended periods and a separate SE for data that is transient — needed only for the lifetime of a job or set of jobs. Large sites with MSS-based SEs may also deploy disk-only SEs for such a purpose or for general use.

3.1.4 File Transfer Services

Basic-level data transfer is provided by GridFTP. This may be invoked directly via the globus-url-copy command or through the srmcopy command which provides 3rd-party copy between SRM systems. However, for reliable data transfer it is expected that an additional service above srmcopy or GridFTP will be used. This is generically referred to as a reliable file transfer service (rfts). A specific implementation of this — this gLite FTS has been suggested by the Baseline Services Working group as a prototype implementation of such a service. The service itself is installed at the Tier-0 (for Tier-0– Tier-1 transfers) and at the Tier-1s (for Tier-1– Tier-2 transfers). It can also be used for 3rd-party transfers between sites that provide an SE. No service needs be installed at the remote site apart from the basic SE services described above. However, tools are available to allow the remote site to manage the transfer service.

For sites or Grid infrastructures that wish to provide alternative implementations of such a service, it was agreed that the interfaces and functionality of the FTS will be taken as the current interface.

File placement services, which would provide a layer above a reliable file transfer service (providing routing and implementing replication policies), are currently seen as an experiment responsibility. In future such a service may become part of the basic infrastructure layer.

3.1.5 Compute Resource Services

The Computing Element (CE) is the set of services that provide access to a local batch system running on a compute farm. Typically the CE provides access to a set of job queues within the batch system. How these queues are set up and configured is the responsibility of the site and is not discussed here.

A CE is expected to provide the following functions and interfaces:

- A mechanism by which work may be submitted to the local batch system. This is implemented typically at present by the Globus gatekeeper in LCG-2 and Grid/Open Science Grid. NorduGrid (the ARC middleware) uses a different mechanism.
- Publication of information through the Grid information system and associated information providers, according to the GLUE schema, that describes the resources available at a site and the current state of those resources. With the introduction of new CE implementations we would expect that the GLUE schema, and evolutions of it, should be maintained as the common description of such information.
- Publication of accounting information, in an agreed schema, and at agreed intervals. Presently the schema used in both LCG-2 and Grid3/OSG follows the GGF accounting schema. It is expected that this be maintained and evolved as a common schema for this purpose.
- A mechanism by which users or Grid operators can query the status of jobs submitted to that site.
- The Computing Element and associated local batch systems must provide authentication and authorization mechanisms based on the VOMS model. How that is implemented in terms of mapping Grid user DNs to local users and groups, how roles and subgroups are implemented, may be through different mechanisms in different Grid infrastructures. However, the basic requirement is clear — the user presents an extended X509 proxy certificate, which may include a set of roles, groups, and subgroups for which he is authorized, and the CE/batch system should respect those through appropriate mappings locally.

It is anticipated that a new CE from gLite, based on Condor-C, will also be deployed and evaluated as a possible replacement for the existing Globus GRAM-based CEs within LCG-2 and Open Science Grid.

3.1.6 Workload Management

Various mechanisms are currently available to provide workflow and workload management. These may be at the application level or may be provided by the Grid infrastructure as services to the applications. The general feature of these services is that they provide a mechanism through which the application can express its resource requirements, and the service will determine a site that fulfils those requirements and submit the work to that site.

It is anticipated that on the timescale of 2006–2007 there will be different implementations of such services available, for example, the LCG-2 Resource Broker, and the Condor-G mechanism used by some applications in Grid3/OSG, and new implementations such as that coming from gLite implementing both push and pull models of job submission.

The area of job workflow and workload management is one where there are expected to be continuing evolutions over the next few years, and these implementations will surely evolve and mature.

3.1.7 VO Management services

The VOMS software will be deployed to manage the membership of the VOs. It will provide a service to generate extended proxy certificates for registered users which contain information about their authorized use of resources for that VO.

3.1.8 Database Services

Reliable database services are required at the Tier-0 and Tier-1 sites, and may be required at some or all of the Tier-2 sites depending on experiment configuration and need. These services provide the database back-end for the Grid file catalogues as either central services located at CERN or local catalogues at the Tier-1 and Tier-2 sites. Reliable database services are also required for experiment-specific applications such as the experiment metadata and data location catalogues, the conditions databases and other application-specific uses. It is expected that these services will be based on scalable and reliable hardware using Oracle at the Tier-0, Tier-1 and large Tier-2 sites, and perhaps using MySQL on smaller sites. Where central database services are provided, replicas of those databases may be needed at other sites. The mechanism for this replication is that described by the 3D project in the applications section of this report.

3.1.9 Grid Catalogue Services

The experiment models for locating datasets and files vary somewhat between the different experiments, but all rely on Grid file catalogues with a common set of features. These features include:

- Mapping of Logical file names to GUID and Storage locations (SURL)
- Hierarchical namespace (directory structure)
- Access control
    - At directory level in the catalogue
    - Directories in the catalogue for all users
    - Well-defined set of roles (admin., production, etc.)
- Interfaces are required to:
    - POOL
    - Workload Management Systems (e.g., Data Location Interface /Storage Index interfaces)
    - POSIX-like I/O service.

The deployment models also vary between the experiments, and are described in detail elsewhere in this document. The important points to note here are that each experiment expects a central catalogue which provides look-up ability to determine the location of replicas of datasets or files. This central catalogue may be supported by read-only copies of it regularly and frequently replicated locally or to a certain set of sites. There is, however, in all cases a single master copy that receives all updates and from which the replicas are generated. Obviously this must be based on a very reliable database service.

ATLAS and CMS also anticipate having local catalogues located at each Storage Element to provide the mapping for files stored in that SE. In this case the central catalogue need only provide the mapping to the site, the local catalogue at the site providing the full mapping to the local file handle by which the application can physically access the file. In the other cases, where there is no such local catalogue, this mapping must be kept in the central catalogue for all files.

The central catalogues must also provide an interface to the various workload management systems. These interfaces provide the location of Storage Elements that contain a file (or

dataset) (specified by GUID or by logical file name) that the workload management system can use to determine which set of sites contain the data that the job needs. This interface should be based on the StorageIndex of gLite or the Data Location Interface of LCG/CMS. Both of these are very similar in function. Any catalogue providing these interfaces could be immediately usable by, for example, the Resource Broker or other similar workload managers.

The catalogues are required to provide authenticated and authorized access based on a set of roles, groups and sub-groups. The user will present an extended proxy-certificate, generated by the VOMS system. The catalogue implementations should provide access control at the directory level, and respect ACLs specified by either the user creating the entry or by the experiment catalogue administrator.

It is expected that a common set of command-line catalogue management utilities be provided by all implementations of the catalogues. These will be based on the catalogue-manipulation tools in the lcg-utils set with various implementations for the different catalogues, but using the same set of commands and functionalities.

3.1.10 POSIX-Like I/O Services

The LHC experiment applications require the ability to perform POSIX-like I/O operations on files (open, read, write, etc.). Many of these applications will perform such operations through intermediate libraries such as POOL and ROOT. In addition, other solutions are being deployed to allow such operations directly from the application. The LCG Grid File Access Library, the gLite I/O service, and aiod in Alien are examples of different implementations of such a service.

It is anticipated that all such applications and libraries that provide this facility will communicate with Grid file catalogues (local or remote), and the SRM interface of the SE in order that the file access can be done via the file LFN or GUID. Thus these libraries will hide this complexity from the user.

It is not expected that remote file I/O to applications from other sites will be needed in the short-term, although the mechanisms described above could provide it. Rather, data should be moved to the local storage element before access, or new files be written locally and subsequently copied remotely.

3.1.11 VO Agents

The LHC experiments require a mechanism to allow them to run long-lived agents at a site. These agents will perform activities on behalf of the experiment and its applications, such as scheduling database updates. No such general service currently exists, but solutions will be prototyped. Currently such actions are performed by experiment software running in the batch system, but this is not a good mechanism in the longer term as it could be seen as a misuse of the batch system. It is better to provide a generic solution which is accepted by the sites, but which provides the facilities needed by the applications.

3.1.12 Application Software Installation Facilities

Currently each Grid site provides an area of disk space, generally on a network file system, where VOs can install application software. Tools are provided in LCG-2, or by the experiments themselves to install software into these areas, and to later validate that installation. Generally, write access to these areas is limited to the experiment software manager. These tools will continue to be provided, and will be further developed to provide the functionalities required by the experiments.

3.1.13 Job Monitoring Tools

The ability to monitor and trace jobs submitted to the Grid is an essential functionality. There are some partial solutions available in the current systems (e.g., the LCG-2 Workload Management system provides a comprehensive logging and book-keeping database),

however, they are far from being full solutions. Effort must be put into continuing to develop these basic tools, and to provide the users with the appropriate mechanisms through which jobs can be traced and monitored.

3.1.14 Validation

The programme of service challenges started in December 2004 and continuing through to the fourth quarter of 2006 are the mechanism through which these services will be validated by the experiments as satisfying their requirements. It is anticipated that continual modification and improvement of the implementations will take place throughout this period.

The process for certifying and deploying these (and other ancillary) services and tools is described in Chapter 5 (life-cycle support).

3.1.15 Interoperability

This section has outlined the basic essential services that must be provided to the LHC experiments by all Grid implementations. The majority of these deal with the basic interfaces from the Grid services to the local computing and storage fabrics, and the mechanisms by which to interact with those fabrics. It is clear that these must be provided in such a way that the application should not have to be concerned with which Grid infrastructure it is running on.

At the basic level of the CE and SE, both EGEE and Grid3/OSG use the same middleware and implementations, both being based on the Virtual Data Toolkit. In addition, both use the same schema for describing these services, and have agreed to collaborate in ensuring that these continue to be compatible, preferably by agreeing to use a common implementation of the information system and information providers. Common work is also in hand on other basic services such as VOMS and its management interfaces. In addition, both EGEE and OSG projects are defining activities to ensure that interoperability remain a visible and essential component of the systems.

The EGEE Resource Broker, as it is based on Condor-G, can submit jobs to many middleware flavours including ARC (NorduGrid). When the Glue2 information system schema, being defined jointly by several Grid projects, is available this will enable the EGEE Resource Broker to schedule resources at sites running ARC. Further steps towards interoperability in the areas of workload management and data management are planned by the NorduGrid Collaboration. Other activities are being undertaken by the developers of ARC to foster and support standards and community agreements. These include participation in the Rome Compute Resource Management Interfaces initiative and in the Global Grid Forum.

These activities will improve interoperability between different middleware implementations, and in the longer term we can expect standards to emerge and be supported by future versions of the software. For the medium term, however, the approach taken by the LCG Project was to set up the *Baseline Services Working Group* to define a set of basic services that can be deployed in all of the existing Grid infrastructures, taking account of their different technical constraints. In some cases the services are defined in terms of standard interfaces, while in other cases a specific implementation is identified. In this way sites providing resources to LCG will be able to provide these essential services in a transparent way to the applications.

3.2 Tier-0 Architecture

The experience with previous experiments is that the first two years of operation is a crucial period where data access patterns, and therefore the underlying architecture, only become clear after the start of data taking. Though the LHC experiments have defined the data flow for analysis in quite some detail, changes and adjustments can be expected after the whole system has stabilized and the detectors are fully understood. As an example, it can be expected that random access to the raw data will be much higher in the first few years than

later, affecting the data flow performance. Thus we have to be prepared for major changes in 2007 and 2008.

It is, however, important to maintain stability in the computing fabric during the first two years, so that the physicists can concentrate on debugging and analysis. This implies stability of operating systems, network infrastructure and basic hardware choices.

In order to prepare changes without impacting the production system, a parallel and independent test/R&D facility must be provided. This must be integrated into the fabric facilitating the move from test to production.

Figure 3.1 shows a schematic view of the data flow in the Tier-0 system.

More details can be found in a paper on the sizing and costing of the Tier-0 [26]. From our current estimates 2/3 of the total costs and resources of the CERN Facility will be needed for the installation of the Tier-0 system.

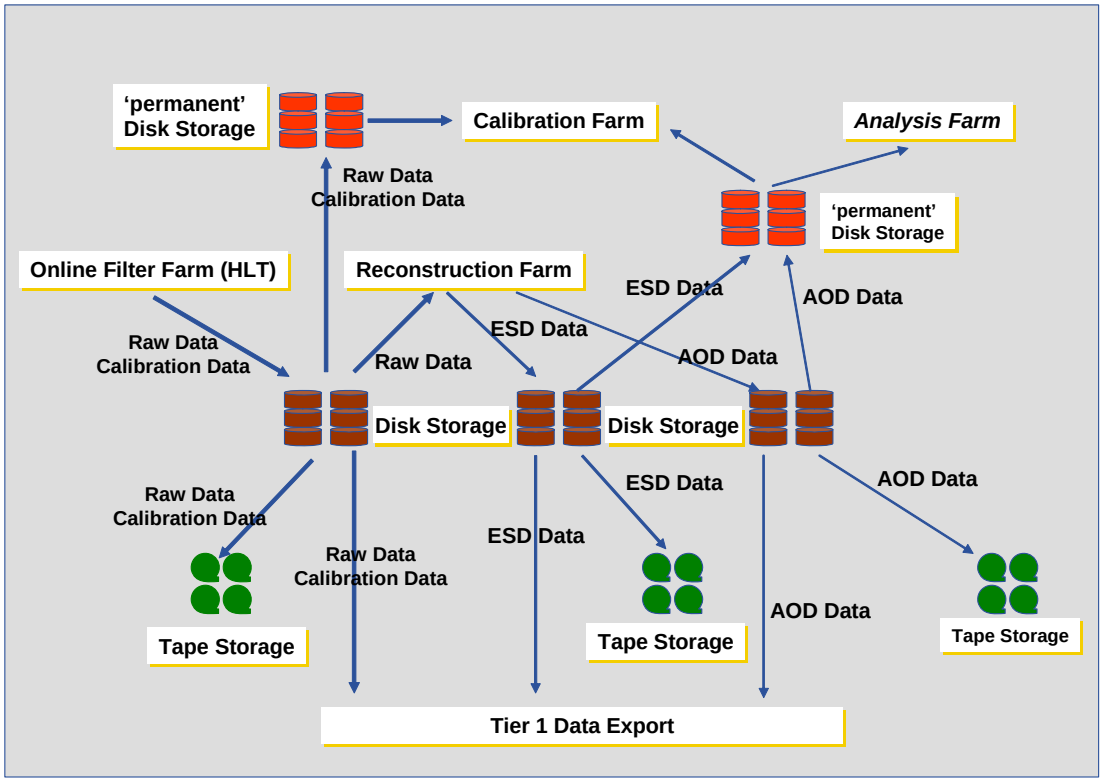


Figure 3.1: Data flow in the Tier-0 system

The general architecture is based on three functional units providing processing (CPU) resources, disk storage and tape storage. Each of these units contains many independent nodes which are connected on the physical layer with a hierarchical, tree-structured, Ethernet network. The application gets its access to the resources via software interfaces to three major software packages which provide the logical connection of all nodes and functional units in the system:

- a batch system (LSF) to distribute and load-balance the CPU resources,
- a medium-size, distributed, global, shared file system (AFS) to have transparent access to a variety of repositories (user space, programs, calibration, etc.),
- a disk pool manager emulating a distributed, global, shared file system for the bulk data and an associated large tape storage system (CASTOR).

The system is managed by a low-level node management system (ELFms) and a small set of sophisticated software components (batch system, mass storage system, management system). Figure 3.2 shows the dependency between the different items.

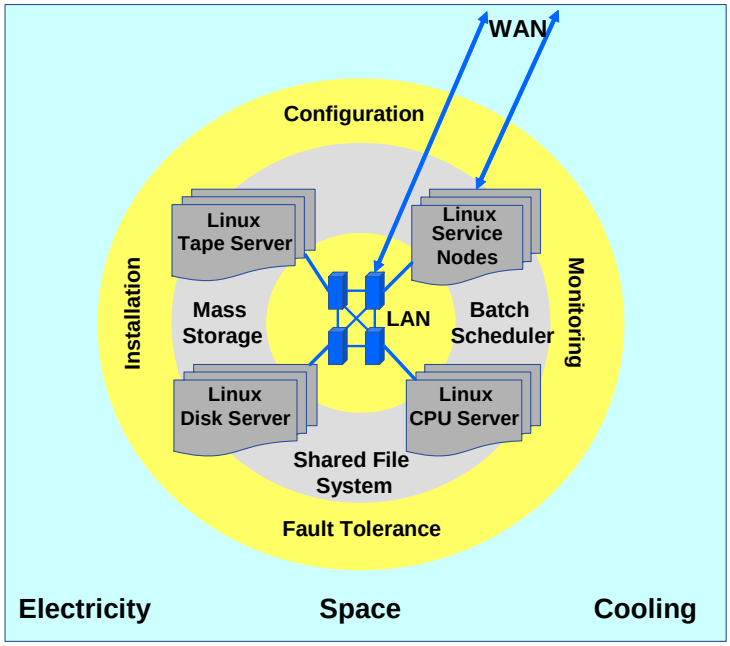


Figure 3.2: Schematic dependency between Tier-0 components

Figure 3.3 shows the structure of the hierarchical Ethernet network infrastructure. The heart of the set-up is based on a set of highly redundant and high throughput routers inter-connected with a mesh of multiple 10 Gbit connections. From the computing models and the cost extrapolation for the years 2006-2010 one can estimate the number of nodes (CPU, disk, tape, service) to be connected to this system to be about 5-8 thousand.

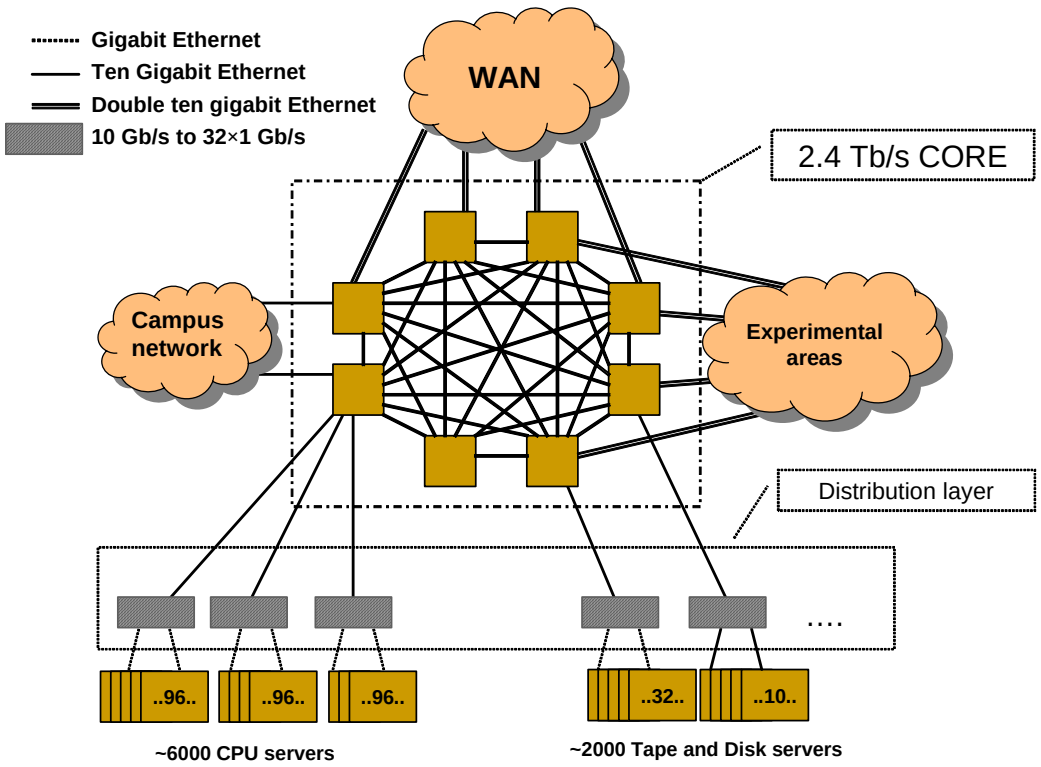


Figure 3.3: Layout of the Tier-0 network

The system provides full connectivity and bandwidth between any two nodes in the tree structure, but not full bandwidth between any set of nodes. Today, for example, we have 96 batch nodes on fast Ethernet (100 Mbit/s) connected to one Gb/s (1,000 Mbit/s) uplink to the backbone, that is a ratio of 10 to 1 for the CPU server. The ratio is about 8 to one for disk servers. Up to now we have not experienced a bottleneck in the network. The expected ratios for 2008 will be 9 to 1 for CPU servers and 3 to 1 for disk servers.

The configuration is based on experience and an analysis of the requirements of the experiments. We expect that this configuration will offer the flexibility to adjust critical parameters such as the bandwidth ratios as the analysis models evolve.

Figure 3.4 shows the aggregate network utilization of the Lxbatch cluster during 2004. The system is mainly used by the LHC experiments and the running fixed-target experiments. The jobs do mainly reconstruction of real or Monte Carlo data as well as analysis work on extracted datasets. Lxbatch was growing from about 1100 nodes in the beginning of the year towards 1400 nodes today containing about 4 different generations of CPU server. A very rough calculation using 600 high-end nodes and a peak data rate of 300 MBytes/s gives an average speed per node of 0.5 MByte/s. This relatively low speed approximately matches the projected requirements for 2008. A current dual processor CPU server has a total performance of ~ 2,000 SPECint2000, and about 8,000 SPECint2000 per node are expected in 2008.

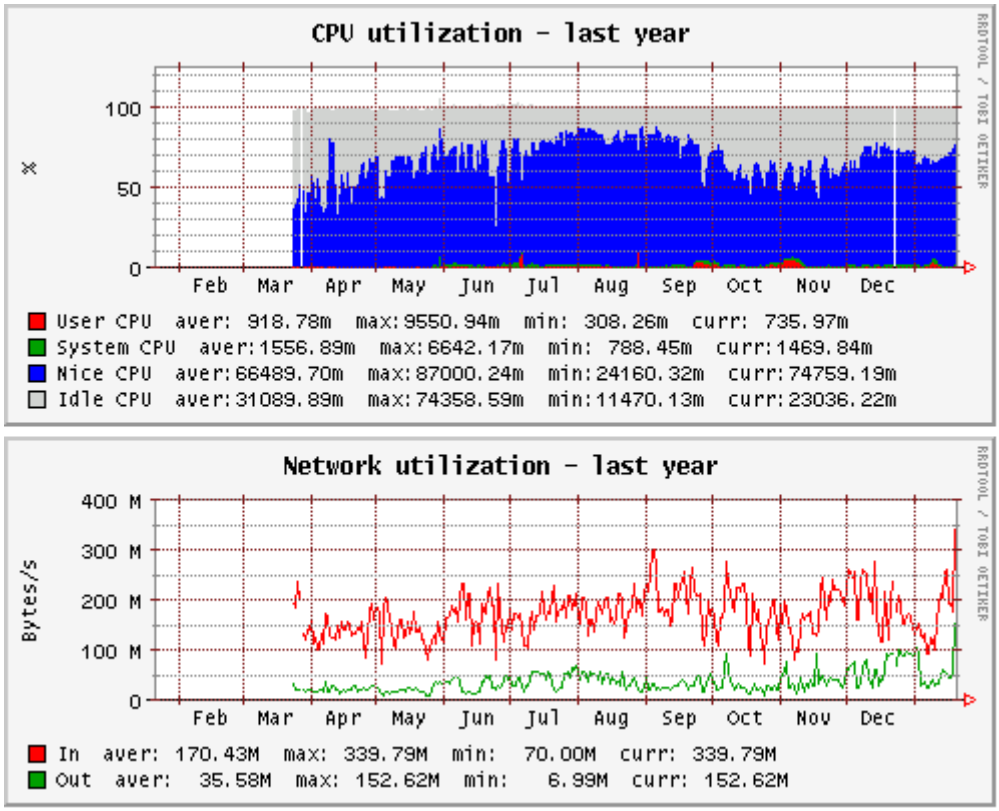


Figure 3.4: Aggregate network utilization of the Lxbatch cluster in 2004

Table 3.1: Reconstruction of raw data → producing ESD and AOD

	Raw data event size [MB]	CPU resource for one event [SPECint2000]	I/O value for a 8000 SPECint2000 CPU server [MB/s]
ALICE p–p	1.0	5,400	1.5
ALICE HI	12.5	675,000	0.1
ATLAS	1.6	15,000	0.9
CMS	1.5	25,000	0.5
LHCb	0.025	2,400	0.1

Table 3.2: Analysis of AOD data

	AOD event size [MB]	CPU resource for one event [SPECint2000]	I/O value for a 8000 SPECint2000 CPU server [MB/s]
ALICE p–p	0.05	3,000	0.1
ALICE HI	0.25	350,000	0.01
ATLAS	0.1	500	1.6
CMS	0.05	250	1.6
LHCb	0.05 – 0.1	300	1.0

The CPU servers are connected at 1 Gb/s and aggregated at 9:1, i.e., 90 servers are connected to 1×10 Gb/s uplink. Each server can communicate at approximately 100 Mb/s before saturating the uplink. The data rates in Table 3.1 and Table 3.2 are in the range of 10–15Mb/s leaving a comfortable margin.

The disk servers are connected at 1 Gb/s and aggregated at 3:1, i.e., 30 servers are connected to 1 × 10 Gb/s 4-35n2. So, each disk server can communicate at approximately 300 Mb/s before saturating the uplink. With 4,000 CPU servers and 1,000disk servers the ratio on average is 4:1 corresponding to an average load of 60 Mb/s to any disk server capable of running at 300 Mb/s.

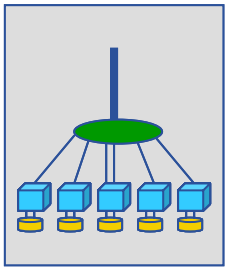
In case of 'hot spots' (i.e. many more than 4 CPU servers accessing the same disk server), the CASTOR disk pool manager will replicate the data across more disk servers but the existing CPU servers will compete for access.

Efficient data layout and strategies for submitting jobs that will use the system efficiently given these constraints have yet to be studied in detail.

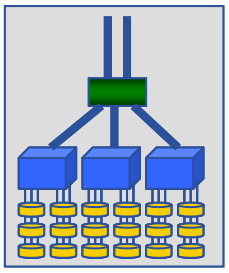
In the disk storage area we consider the physical and the logical view. On the physical side we will follow a simple integration of NAS (Network Attached Storage — each disk server implements an independent file system) boxes on the hierarchical Ethernet network with single or multiple (probably 3 max.) gigabit interconnects. The basic disk storage model for the large disk infrastructure (probably 2 PB in 2008) is assumed to be NAS with up to 1,000disk servers and locally attached disks. This amount of disk space is assumed to grow considerably between 2008 and 2012 whereas the number of servers could decrease substantially. However, the overall structure permits also the connection of different implementations of disk storage. Different levels of caching can be implemented if needed (e.g., as a front-end to the tape servers).

The following list shows some examples starting with the simple NAS storage solution. We are evaluating the other solutions to understand their benefits compared with the simple NAS solution.

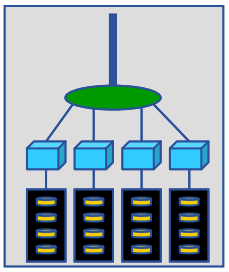
Simple Network Attached Storage boxes connected via gigabit Ethernet (one or several) and 10 gigabit Ethernet uplinks



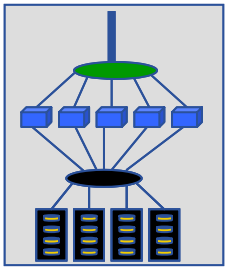
High-end multi-processor servers (> = 4 CPU) with large amounts of space per box connected to 10 gigabit Ethernet switches



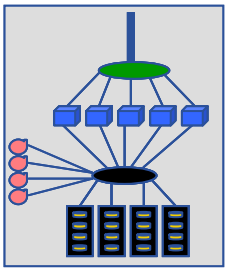
Separation of CPU part and the disk space itself, CPU server with Fibre Channel attached SATA disk arrays



Small Storage Area Network (SAN) set-ups linked with front-end file server nodes that are connected into the gigabit network — disks can be reconfigured to be connected to any of the file servers, but disks are not shared between servers



Combination of the SAN set-up with tape servers, locality of disk storage to tape drives



On the logical level the requirement is that all disk storage systems (independent of their detailed physical implementation) present file systems as the basic unit to higher level applications (e.g., Mass Storage System).

3.3 Tier-1 Architecture

Within the computing model for the LHC Computing Grid, the Tier-1 centres are the principle centres of full-service computing. These centres, in addition to supplying the most complete range of services, will supply them via agreed Grid interfaces and are expected to supply them with specified high levels of availability, reliability and technical backing. Each Tier-1 centre will have specific agreed commitments to support particular experiment Virtual Organizations (VOs) and their user communities and to supply required services to specific Tier-2 centres. There may also be specific agreements with other Tier-1 centres by which complete datasets are made available to user communities and/or by which data is backed up or other back-up services are supplied by the alternate centre during planned or unplanned facility outages.

The underlying elements of a Tier-1 consist of online (disk) storage, archival (tape) storage, computing (process farms), and structured information (database) storage. These elements are supported by a fabric infrastructure and using software and middleware packages are presented as Grid services meeting agreed interface definitions. While details may depend on the particular experiment supported, many services will be common.

3.3.1 Archival Storage

Archival storage systems in general consist of an automated tape library with a front-end disk cache running a Hierarchical Storage Management (HSM) system. Common HSMs within the LHC Grid are CASTOR, Enstore and HPSS. A Tier-1 is responsible for the storing and general curation of archived data through the life of the experiments it supports. This implies the need to retain capabilities in the technology in which the data is originally recorded or to migrate the data to new storage technologies as they are adopted at that site in the future. For archival storage, with its inherent tape mount and position search latency, the primary performance issue is one of long-term average throughput rather than latency or peak throughput. The level of sustain throughput is determined by the speed and number of tape drives, the size and speed of the disk cache, and the number and speed of the server machines to which these peripherals are connected. This level of throughput must be adequate to satisfy the simultaneous needs of the various specific archival activities described below. Depending on the mix and size of the reads and writes, the number of mounts and the time spent in search

mode on tapes, the effective performance is significantly less than the maximum streaming I/O rate the tape is capable of. This factor must be estimated with reasonable accuracy and taken into account in determining the number of tape drives required. In general, while access to data in archival storage is permitted by individual users, access rights to archival storage for the purpose of storing data are likely to be granted on a programmatic or policy-determined basis.

The services presented by such a system will at minimum be based on an agreed Storage Resource Management (SRM) interface specification. This SRM layer is above a scalable transport layer consisting in general of GridFTP servers. The SRM specification will evolve with time and the Tier-1s are committed to supporting this evolution. In addition, it is expected that there will be additional layers of protocol above this SRM layer which will meet LCG or experiment-specific agreed specifications. These added layers will improve the reliability, efficiency, and performance of data transfers and the coupling of these transfers to databases and other higher level data constructs, such as datasets or datastreams. Archival storage will appear as a Storage Element on the Grid and will supply a number of specific services.

3.3.2 Raw Data Archiving Service

Tier-1 centres are required to archivally store, reprocess, and serve as necessary, a fraction of an experiment's raw data. This data is to be accepted promptly so that it is still in a disk buffer at CERN at the time of transfer and thus does not require additional access to the CERN mass storage system, where an additional copy will be maintained. There must also be sufficient I/O capacity to retrieve data, perhaps simultaneously with archiving new data, from the archive for additional reconstruction passes as agreed upon with the experiments supported.

3.3.2.1 Monte Carlo Data Archiving Service

Tier-1 centres are required to archivally store, process, reprocess, and serve as necessary, a fraction of an experiment's Monte Carlo data. This data is that which is produced at an agreed set of Tier-2 centres and possible named non-Tier-2 Additional Facilities (AF). It must be accepted and recorded on a time-scale which will make it unnecessary that such Tier-2 centres or AFs will themselves need to maintain archival storage systems.

3.3.2.2 Derived Data Archiving Service

Tier-1 centres, as the primary archival sites within the LCG computing model, will also be expected to archivally store, for those experiments it supports, some fraction of those derived datasets which, while no longer required online, may be needed at some point in the future and for which regeneration in a timely manner may not be possible or practical.

3.3.3 Online Storage

The technologies by which such online storage systems are likely to be implemented can be divided into two categories. First, there are relatively costly, robust, centralized, commercial systems and second, there are less expensive, more distributed systems based on commodity hardware and public domain or community supported software. The first category includes FibreChannel connected RAID 5 systems running behind conventional NFS file servers and also custom network attached storage appliances such as Blue Arc, Panasas, etc. The second category includes such systems as dCache, RFIO and Lustre which run on arrays of Linux nodes mounting inexpensive commodity disk. Unlike archival storage, at least locally, record level access to online storage is required and so a POSIX or POSIX-like interface is in general required. In the case of online storage, issues of latency and peak transfer rate are much more important then in archival storage.

Again the services presented by such systems will at minimum be required to support the same LCG wide Storage Resource Management (SRM) interface specification as the archival storage system discussed above. Again this will be running above a scalable transport layer

and below expect higher level protocols. Online storage will also appear as a Storage Element on the Grid and will supply the following specific services.

3.3.3.1 Reconstructed Data Service

While the details of the plan for reconstruction passes and the output of reconstruction are experiment-dependent, all experiments plan to make multiple reconstruction passes and to keep one or more copies of the output of the most recent reconstruction pass available online at Tier-1 centres. Typically, a reconstruction pass produces multiple levels of output including large very inclusive sets, the Event Summary Data (ESD), set which is more concise but still relatively comprehensive for analysis purposes, the Analysis Object Data (AOD), and very compact highly structured sets, the TAG data. While the AOD and TAG sets are sufficiently compact that they can be stored online at multiple locations including the Tier-1 centres, the ESD set is, in general, very large and so its online storage is a specific responsibility of the Tier-1 centres. Depending on the experiment, the complete online storage of the ESD set may be accomplished by distributing it across multiple Tier-1 centres. Some experiments may also require that certain ESD sets corresponding to previous reconstruction passes also be maintained online at Tier-1 centres, though perhaps in fewer copies. In general, the availability of this ESD data is most important to the programmatic regeneration of derived datasets, including the AOD and others, done at the request of individual physics analysis groups. In addition, physicists doing their own individual chaotic analysis based on higher level, more concise, datasets may find it necessary for certain select events to refer back to this more complete output of reconstruction. Sustained high-bandwidth access is very important to assure that programmatic passes to select data subsets and regenerate derived data are accomplished quickly and efficiently as measured in hours or days. Reasonably low latency is also important to meet the requirements of users doing chaotic analysis who need to selectively reference back into this more complete dataset.

3.3.3.2 Analysis Data Service

Analysis data is typically being accessed in support of chaotic analysis done by individual physicists. The emphasis put on such analyses at Tier-1 centres is experiment dependent. For this service it is AOD, TAG and other relatively concise derived datasets that are being served. Since there is typically a physicist in real, or near real-time, waiting for results, the issue of access performance is more one of peak bandwidth and latency, which is likely to be measured in minutes or possibly even seconds, rather than long-term sustained bandwidth. It is also possible that particular datasets will become very popular at certain points in time and so be accessed very frequently my many users. This kind of access pattern can seriously impact performance from the user perspective and so strategies need to exist to deal with such hot spots in the analysis data. Hot spots are dealt with by replicating the target data and so distributing the access across multiple systems. Such contention-driven replication can be done within the storage service relatively automatically by products such as dCache or will need to be addressed at higher levels within the overall analysis system.

3.3.4 Computation

The technology used at the Tier-1 centres to supply computation services is the Linux processor farm coupled to a resource management system, typically a batch scheduler such as LSF, PBS or Condor. Intel processors are generally used, though AMD and PowerPC processors may come into common usage in the near future. At the moment processors are most commonly packaged two to the box and connected by 100 or 1,000Mb/s Ethernet. Grid access to computing resources is via a Globus Gatekeeper referred to as a Computing Element which serves as the Grid interface to the batch queues of the centre. While Globus-level details of this interface are well defined, it is likely that there will be LCG-agreed higher level interface layers which will be defined to guarantee the effective distribution of workload across Grid-available compute resources. Tier-1 centres will be responsible for presenting compute services with such agreed interfaces. There are a number of specific compute

services supplied by Tier-1 centres depending on the computing models of the experiments they support.

3.3.4.1 Reconstruction

Reconstruction is generally a CPU-intensive programmatic activity requiring an extended period of time, several weeks to a few months for a complete pass through a year's raw data. The effective utilization of CPU for reconstruction requires that the raw data be pre-staged from tape to a location offering direct access by the processor. Assuming adequate de-synchronization of input/output activity across a farm of processors doing reconstruction, modern networking should be able meet data transfer needs in and out without difficulty. During reconstruction it is necessary that there be access to condition and calibration information appropriate to the particular raw data undergoing reconstruction. This implies that there is either access to a database containing that information or that the need for that information has been externally anticipated and that the required information has been packaged and shipped to the reconstructing node. Given the general level of I/O to CPU in event reconstruction, while not optimal, the movement of data across the wide-area to the location of available compute resources is not likely to place an unacceptable load on the intervening WAN.

3.3.4.2 Programmatic Analysis

Programmatic analysis refers to passes made through the more inclusive output datasets of reconstruction, typically ESD, to select data subsets and/or to regenerate or generate additional derived data, such as AOD. Such programmatic analysis is typically done at the formal request of one or more physics groups and takes periods measured in days, perhaps only a couple but possibly more. In general such an activity is quite I/O intensive, with only modest calculations being done while accessing the complete ESD or some selected stream of it. The Tier-1 centre must be configured so that the CPU used for such a pass has excellent connectivity to the online storage on which the ESD is stored. If the CPU which is to perform this service were located at a Tier-1 centre which was Wide-Area-separated from the location of the ESD, such that the data had to be moved via WAN, this activity would likely place an excessive load on the network.

3.3.4.3 Chaotic Analysis

Individual user analysis, so-called chaotic analysis, is generally characterized by jobs consuming modest amounts of resources running for relatively short times, minutes to a few hours. The amount of this analysis which is expected to be done at a Tier-1 centre compared to that done at Tier-2 & 3 sites differs from experiment to experiment. Such analysis is often an iterative process with the result of one analysis pass being used to adjust conditions for the next. For this reason turn-around is in general important. Such analyses can be either quite I/O intensive, for example when establishing optimal selection criteria to apply to a relatively large dataset to reveal a signal, or can be quite CPU-intensive as in the case of doing a numerically sophisticated analysis on a modest number of events. In either case, such chaotic analysis tends to subject computing system to very spiky loads in CPU utilization and/or I/O. For this reason such chaotic analyses can be quite complementary to long-running programmatic activities utilizing the same resources. An analysis job interrupts the ongoing programmatic activity for a brief period of time, measure in minutes, across a large number of processors and so gets good turn-around while leaving the bulk of the time and thus integrated capacity to the programmatic activity whose time-scale is measured in days or weeks. Since the datasets used in chaotic analysis tend to be of small to modest scale and are generally accessed multiple times, moving the data and caching it at the wide-area location of the available Computing Elements is a useful strategy.

3.3.4.4 Calibration Calculation

The calculation of calibration and alignment constants can vary greatly in the ratio of CPU to I/O required, the absolute scale of the CPU required, and the latency which can be tolerated. Some calibration calculation may be almost interactive in nature with iterative passes through a modest dataset involving human evaluation of results and manual direction of the process. Depending on the scale of the computation and the immediacy of human intervention, a subset of the analysis resources, either those for programmatic or those for chaotic analysis, may be well suited to this type of calibration work. For other calibrations, the process may involve a very large scale production pass over a fairly large amount of data requiring very substantial compute resources done in a fairly deterministic way. In general, the calculation of calibration constants is an activity which precedes the performance of a reconstruction pass through the raw data. These make practical the use in a time-varying way of the same compute resources as are used for reconstruction to perform large-scale-production-pass-type calibration calculations.

3.3.4.5 Simulation

Simulation is in general a very CPU-intensive activity requiring very modest I/O. The amount of simulation done at Tier-1 centres as compared to that done at Tier-2 sites is again experiment-dependent. Most simulation is done as a programmatic production activity generating datasets required by various physics groups each frequently requiring several days or even weeks. The fact that the amount of output per unit of CPU is small, and the input is typically even smaller, means that the CPU need not be particularly well network connected to the storage it uses, with wide-area separation being quite acceptable.

3.3.5 Information Services

Relational database technology is the primary technology underlying the delivery of structured information by Tier-1 centres. The most commonly used database management system is likely to be MySQL but Oracle is likely to also be used and there may also be servers running other database managers. Depending on the detailed requirements of individual experiments, various specialized database operating modes may be required including distributed and/or replicated databases. Again, depending on the requirements of individual experiments, various catalogue technologies built upon these databases may need support, including for file catalogues Firemen and/or LFC. In some cases, information service services will require the gathering and publishing, in very specific LCG-agreed formats, information regarding the local site such as resource availability, performance monitoring, accounting and security auditing. A major information service which Tier-1s must support is that of serving the metadata which describes the data of the experiments it supports. While in detail this service will be experiment-specific, it is expected that there will be considerable commonality across experiments in terms of underlying tools and these will be ones agreed to and coherently supported by the LCG. Another major information service is that of the conditions and calibrations required to process and analyse an experiment's data. Again the details of how this is done will be experiment-specific. In general Tier-1 centres will be required to deploy, optimize, and support multiple database managers on multiple servers with appropriate levels of expertise. The services supplied will be interfaced to the Grid according to interface definitions agreed by LCG or specific experiments.

3.3.6 Cyber Security

While cyber security might naturally be regarded as part of the fabric and Grid infrastructure, it is today a very important and dynamic area requiring special attention. There are clearly many policy issues which must be dealt with between sites which are tightly coupled by the Grid and so very interdependent in terms of cyber security. This is especially true of the Tier-1s which are very large, prominent computing sites within their own countries and whose mission typically extends beyond the LCG. It is beyond the scope of this section to deal with cyber security in a substantive way, however, one high-profile cyber security element of an

architectural nature which impacts many of the services discussed above and is worth some discussion is the firewall. Many, if not most, of the Tier-1 centres include in the arsenal of tools used to strengthen their cyber security, a firewall. Its effectiveness against direct intrusion by random external threats is clearly quite high. However, it can have major negative impacts on the services discussed above. First, if not properly configured, it can block the communications required to supply the service at all. Second, even if the firewall is properly configured, it can slow the service unless its throughput is sufficiently high.

One function important to the services discussed above requiring firewall configuration is database access. The appropriate configuration of firewall conduits to permit needed database access by a modest number of systems does not in general represent a problem. However, sites are often uncomfortable with opening access through a firewall for a farm of Linux systems, perhaps numbering thousands of machines. Especially if the application of the latest security patches for such a farm is on occasion delayed by the scale of the effort and disruption involved in doing so for so many machines. An option in this case is to run a sufficiently frequently updated replica of the required remote database server behind the local firewall, thus requiring firewall conduits for only the replica server. One is thus trading the complexity of running such a replica service against the risk of exposing a large number of systems.

Another function important to the services discussed above that is affected by a firewall is high-speed data transfer where the issue is whether or not the firewall, even properly configured, has sufficient throughput. To the extent that such transfers are point to point via dedicate circuits, switched light path or routed, the possibility of bypassing the firewall altogether is a reasonable option. This is the plan for connections between Tier-1s and the Tier-0. The situation is not so clear in cases where the Tier-1 is using the general Internet for transfers to/from Tier 3's and perhaps Tier-2 and other Tier-1s as well. Depending on the rate of advance in firewall technology, the need to find suitably secure general techniques to bypass them for very high speed transfers may be necessary.

In the two examples discussed above, decisions will probably have to be made independently at each Tier-1 on the basis of local policy in the context of the requirements of the experiments it supports and the available personnel resources. With respect to many cyber security issues a one-solution-fits-all approach is unlikely.

3.4 Tier-2 Architecture

The primary roles of the Tier-2 sites are the production and processing of Monte Carlo data and end-user analysis, although these roles vary by experiment.

As Tier-2s do not typically provide archival storage, this is a primary service that must be provided to them, assumed via a Tier-1. Although no fixed relationship between a Tier-2 and a Tier-1 should be assumed, a pragmatic approach for Monte Carlo data is nevertheless to associate each Tier-2 with a 'preferred' Tier-1 that is responsible for long-term storage of the Monte Carlo data produced at the Tier-2. By default, it is assumed that data upload from the Tier-2 will stall should the Tier-1 be logically unavailable. This in turn could imply that Monte Carlo production will eventually stall, if local storage becomes exhausted, but it is assumed that these events are relatively rare and the production manager of the experiment concerned may in any case reconfigure the transfers to an alternative site in case of prolonged outage.

In the case of access to real data for analysis purposes, a more flexible model is required, as some portions of the data will not be kept at the 'preferred' Tier-1 for a given Tier-2. Transparent access to all data is required, although the physical data flow should be optimized together with the network topology and may flow between the Tier-1 hosting the data and the 'preferred' Tier-1 for a given Tier-2 site, or even via the Tier-0.

In order to provide this functionality, the Tier-2s are assumed to offer, in addition to the basic Grid functionality:

- client services whereby reliable file transfers maybe initiated to / from Tier-1/0 sites, currently based on the gLite File Transfer software (gLite FTS);
- managed disk storage with an agreed SRM interface, such as dCache or the LCG DPM.

Both gLite FTS and the LCG DPM require a database service. In the case of the former, it is currently assumed that the file transfer database be hosted at the corresponding Tier-1 site in an Oracle database. For the LCG DPM, its internal catalogue is also hosted in a database, which in this case is assumed to reside at the Tier-2, typically in a MySQL database. For dCache, a local PostgreSQL database is similarly required.

3.4.1 Tier-2 Network

The Computing Model papers of the experiments have been analysed and the resulting bandwidth requirements are depicted in Table 3.3. The bandwidth estimates have been computed assuming the data are transferred at a constant rate during the whole year. Therefore, these are to be taken as very rough estimates that at this level should be considered as lower limits on the required bandwidth. To obtain more realistic numbers, the time pattern of the transfers should be considered, but this is still very difficult to estimate today in a realistic manner. Furthermore, it is also very difficult to estimate the efficiency with which a given end-to-end network link can be used. In order to account for all these effects, some safety factors have been included. The numbers have been scaled up, first by a 50% factor to try to account for differences between 'peak' and 'sustained' data transfers, and second by a 100% factor on the assumption that network links should never run above their 50% capacity.

Table 3.3: Bandwidth estimation for the Tier-1 to Tier-2 network links.

	ALICE	**ATLAS**	**CMS**	**LHCb**
Parameters:				
Number of Tier-1s	6	10	7	6
Number of Tier-2s	21	30	25	14
Real data 'in-Tier-2':				
TB/yr	120	124	257	0
Mbit/s (rough)	31.9	32.9	68.5	0.0
Mbit/s (w. safety factors)	**95.8**	**98.6**	**205.5**	**0.0**
MC 'out-Tier-2':				
TB/yr	14	13	136	19
Mbit/s (rough)	3.7	3.4	36.3	5.1
Mbit/s (w. safety factors)	**11.2**	**10.2**	**108.9**	**15.3**
MC 'in-Tier-2':				
TB/yr	28	18	0	0
Mbit/s (rough)	7.5	4.9	0	0.0
Mbit/s (w. safety factors)	**22.5**	**14.7**	**0.0**	**0.0**

The Tier-1 and Tier-2 centres located in Europe will be computing facilities connected to the National Research and Educational Networks (NRENs) which are in turn interconnected through GÉANT. Today, this infrastructure already provides connectivity at the level of the Gb/s to most of the European Tier-1 centres. By the year the LHC starts, this network infrastructure should be providing this level of connectivity between Tier-1 and Tier-2 centres in Europe with no major problems.

For some sites in America and Asia the situation might be different, since the trans-Atlantic link will always be 'thin' in terms of bandwidth as compared to the intra-continental connectivity. Tier-1 centres in these countries might need to foresee increasing their storage capacity so that they can cache a larger share of the data, hence reducing their dependency on the inter-continental link. Tier-2 centres will in general depend on a Tier-1 on the same continent, so their interconnection by the time LHC starts should also be at the Gb/s level with no major problems.

According to the above numbers, this should be enough to cope with the data movement in ATLAS, CMS and LHCb Tier-2 centres. On the other hand, those Tier-2 centres supporting ALICE will need to have access to substantially larger bandwidth connections, since the estimated 100 MB/s would already fill most of a 11 Gb/s link.

It is worth to noting as well that the impact of the network traffic with Tier-2 centres will not be negligible for Tier-1s as compared to the traffic between the Tier-1 and the Tier-0. The bandwidth requirements have recently been estimated [27]. The numbers presented in this note indicate that, for a given Tier-1, the traffic with a Tier-2 could amount to ~10% of that with the Tier-0. Taking into account the average number of Tier-2 centres that will depend on a given Tier-1 for each experiment, the overall traffic with Tier-2s associated with a given Tier-1 could reach about half of that with the Tier-0. On the other hand, it should also be noted that the data traffic from Tier-1 into Tier-2 quoted here represents an upper limit for the data volume that a Tier-1 has to deliver into a given Tier-2, since most probably there will be Tier-2-to-Tier-2 replications that will lower the load on the Tier-1

4 TECHNOLOGY AND INFRASTRUCTURE

4.1 EGEE Middleware

The EGEE middleware deployed on the EGEE infrastructure consists of a packaged suite of functional components providing a basic set of Grid services including job management, information and monitoring and data management services. The LCG-2.x middleware, currently deployed in over 100 sites worldwide originated from Condor, EDG, Globus, VDT and other projects. It is anticipated that the LCG-2 middleware will evolve in summer 2005 to include some functionalities of the gLite middleware provided by the EGEE project. The architecture of gLite is described in Ref.

[28][4]. This middleware has just been made available as this report is being written, and has not yet passed certification . The rest of this chapter will describe, respectively, the LCG-2 middleware services and the gLite ones.

The middleware can in general be further categorized into site services and Virtual Organization (VO) services as described below.

4.1.1 Site Services

4.1.1.1 Security

All EGEE middleware services rely on the Grid Security Infrastructure (GSI). Users get and renew their (long-term) certificate from an accredited Certificate Authority (CA). Short-term proxies are then created and used throughout the system for authentication and authorization. These short-term proxies may be annotated with VO membership and group information obtained from the Virtual Organization Membership Services (VOMS). Access to (site) services is controlled by the Java authorization framework (Java services) and LCAS (C services). When necessary, in particular for job submission, mappings between the user Distinguished Names (DN) and local account are created (and periodically checked) using the LCAS and LCMAPS services. When longer-term proxies are needed, MyProxy services can be used to renew the proxy. The sites maintain Certificate Revocation Lists (CRLs) to invalidate unauthorized usage for a revoked Grid user.

VOMS and VOMS administrator documentation are available at Refs. [29] and [30].

4.1.1.2 Computing Element

The Computing Elements (CEs), often dubbed head nodes, provide the Grid Interfaces to Local Resource Managers (a.k.a. site batch systems). They normally require external network connectivity.

LCG-2 Computing Element

The LCG-2 Computing Element (CE) handles job submission (including staging of required files), cancellation, suspension and resume (subject to support by the Local Resource Management System — LRMS), job status inquiry and notification. It only works in push mode where a job is sent to the CE by a Resource Broker (RB). Internally the LCG-2 CE makes use of the Globus gatekeeper, LCAS/LCMAPS and the Globus Resource Allocation Manager (GRAM) for submitting jobs to the LRMS. It also interfaces to the logging and book-keeping Services to keep track of the jobs during their lifetime.

[4] An updated version is due in summer 2005.

The LCG-2 CE interfaces with the following LRMS: BQS, Condor, LSF, PBS and its variants (Torque/Maui), and many others.

gLite Computing Element

The gLite Computing Element (CE) handles job submission (including staging of required files), cancellation, suspension and resume (subject to support by the LRMS), job status inquiry and notification. The CE is able to work in a push model (where a job is pushed to a CE for its execution) or in a pull model (where a CE asks a known Workload Manager — or a set of Workload Managers — for jobs). Internally the gLite CE makes use of the new Condor-C technology, GSI and LCAS/LCMAPS, as well as the Globus gatekeeper. The CE is expected to evolve into a VO-based scheduler that will allow a VO to dynamically deploy their scheduling agents. The gLite CE also make use of the 3ogging and book-keeping services to keep track of the jobs during their lifetime.

The gLite CE interfaces with the following LRMS: PBS and its variants (Torque/Maui), LSF and Condor. Work to interface to BQS (IN2P3) and SUN Grid Engine (Imperial College) is under way.

4.1.1.3 Storage Element

The Storage Element (SE) provides the Grid interfaces to site storage (can be Mass Storage or not). SEs normally require external network connectivity.

LCG-2 Storage Elements

The LCG-2 SE can either by a ‘classic’ SE or an SRM SE. The classic SE provides a GridFTP (Efficient FTP functionality with GSI security) interface to disk storage. The RFIO protocol can be used for accessing directly the data on a classic SE. An SRM SE provides the GridFTP interface to a Storage Resource Manager (SRM), a common interface to Mass Storage Systems such as the CERN Advanced Storage Manager (CASTOR) or dCache/Enstore from DESY and FNAL.

Recently, a more lightweight and simpler SRM has been made available, the LCG Disk Pool Manager (DPM), which is targeted at smaller disk pools. The DPM is a natural replacement for the classic SE.

GFAL

The Grid File Access Library (GFAL) is a POSIX-like I/O layer for access to Grid files via their Logical Name. This provides open/read/write/close style of calls to access files while interfacing to a file catalogue. GFAL currently interfaces to the LFC and the LCG-RLS catalogs. A set of command line tools for file replication called lcg-utils have been built on top of GFAL and catalogue tools supporting SRMs and classic SEs.

gLite Storage Element

A gLite Storage Element consists of a SRM (such as CASTOR, dCache or the LCG Disk Pool Manager) presenting a SRM 1.1 interface, a GridFTP server as the data movement vehicle and gLite I/O for providing a POSIX-like access to the data. gLite itself does not provide a SRM nor a GridFTP server which must be obtained from the standard sources.

gLite I/O

The gLite I/O is a POSIX-like I/O service for access to Grid files via their Logical Name. This provides open/read/write/close style of calls to access files while interfacing to a file catalogue. It enforces the file ACLs specified in the catalogue if appropriate. gLite I/O currently interfaces to the FiReMan and the LCG-RLS catalogs.

An overview of gLite data management can be found at Ref. [31], while detailed usage of gLite I/O command lines and programmatic interfaces are available from Ref. [32].

4.1.1.4 Monitoring and Accounting Services

The monitoring and accounting services retrieve information on Grid services provided at a site as well as respective usage data, and publish them. User information (in particular related to job execution progress) may be published as well.

LCG-2 Monitoring and Accounting Services

The LCG-2 monitoring service is based on information providers which inspect the status of Grid services and publish their data into the LDAP based BDII system. Accounting data is collected by the Accounting Processor for Event Logs (APEL) system which publishes its data into the R-GMA system. R-GMA requires a server running at a site to produce and consume information.

gLite Monitoring and Accounting Services

gLite relies on the same services as described in Section 0. In addition, an R-GMA-based service discovery system is provided. The gLite accounting system (DGAS) is subject to evaluation.

DGAS collects information about usage of Grid resources by users, groups of users (including VO). This information can be used to generate reports/billing but also to implement resources quotas. Access to the accounting information is protected by ACLs. More information on DGAS is available at Ref. [33].

4.1.2 VO or Global Services

4.1.2.1 Virtual Organization Membership Service

The Virtual Organization Membership Service (VOMS) annotates short-term proxies with information on VO and group membership, roles and capabilities. It originated from the EDG project. It is in particular used by the Workload management System and the FireMan catalogue for ACL support to provide the functionality identified by LCG. The main evolution from EDG/LCG is support for SLC3, bug fixes and better conformance to IETF RFCs.

A single VOMS server can serve multiple VOs. A VOMS Administrator Web interface is available for managing VO membership through the use of a Web browser.

There is no significant functional difference between the VOMS in LCG-2 and in gLite. VOMS 1.5 and higher supports both MySQL and Oracle.

For a detailed description of VOMS and its interfaces, see Refs. [34] and [35].

4.1.2.2 Workload Management Systems

LCG-2 Workload Management System

The Workload Management System in LCG-2.x originated from the EDG project. It essentially provides the facilities to manage jobs (submit, cancel, suspend/resume, signal) and to inquire about their status. It makes use of Condor and Globus technologies and relies on GSI security. It dispatches jobs to appropriate CEs, depending on job requirements and available resources. BDII and RLS are used for retrieving information about the resources.

The user interfaces to the WMS using a job Description Language based on Condor Classads is specified at Ref. [36].

gLite Workload Management System

The Workload Management system in gLite is an evolution of the one in LCG-2. As such, it relies on BDII as an information system. It is interoperable with LCG-2 CEs.

The Workload Management System (WMS) operates via the following components and functional blocks:

The Workload Manager (WM) or Resource Broker, is responsible of accepting and satisfying job management requests coming from its clients. The WM will pass job submission requests to appropriate CEs for execution, taking into account requirements and preferences expressed in the job Description. The decision as to which resource should be used is the outcome of a matchmaking process between submission requests and available resources. This not only depends on the state of resources, but also on policies that sites or VO administrators have put in place (on the CEs).

Interfaces to Data Management allowing the WMS to locate sites where the requested data is available are available for LCG RLS, the Data Location Interface (DLI — used by CMS) and the StorageIndex interface (allowing for querying catalogs exposing this interface — a set of two methods listing SEs for a given LFN or GUID, implemented by the FiReMan and AliEn catalogs).

The WMproxy component, providing a Web service interface to the WMS as well as bulk submission and parametrized job capabilities is foreseen to be available before the end of the EGEE project.

The user interfaces to the WMS using a Job Description Language based on Condor Classads is specified at Ref. [37] . The user interacts with the WMS using a Command Line Interface or APIs. Support of C++ and Java is provided (for a detailed description of the WMS and the interfaces, see Refs. [38] and [39])

4.1.2.3 File Catalogs

Files on Grids can be replicated in many places. The users or applications do not need to know where the files actually are, and use Logical File Names (LFNs) to refer to them. It is the responsibility of file catalogs to locate and access the data. In order to ensure that a file is uniquely identified in the universe, Global Unique Identifiers (GUIDs) are usually used.

EDG RMS

The services provided by the RMS, originating from EDG, are the Replica Location Service (RLS) and the Replica Metadata Catalogue (RMC). The RLS maintains information about the physical location of the replicas. The RMC stores mappings between GUIDs and LFNs. A last component is the Replica Manager offering a single interface to users, applications or Resource Brokers. The command line interfaces and APIs for Java and C++ are respectively available from Refs. [40] and [41]. It is anticipated that the EDG RMS will gradually be phased out.

LCG File Catalogue

The LCG File catalogue (LFC) offers a hierarchical view of logical file name space. The two functions of the catalogue are to provide Logical File Name to Storage URL translation (via a GUID) and to locate the site at which a given file resides. The LFC provides Unix style permissions and POSIX Access Control Lists (ACL). It exposes a transactional API. The catalogue exposes a so-called Data Location Interface (DLI) that can be used by applications and Resource Brokers. Simple metadata can be associated with file entries. The LFC supports Oracle and MySQL databases. The LFC provides a command line interface and can be interfaced through Python.

gLite Fireman catalogue

The gLite File and Replica Catalogue (FiReMan) presents a hierarchical view of a logical file name space. The two main functions of the catalogue are to provide Logical File Name to Storage URL translation (via a GUID) and to locate the site at which a given file resides. The catalogue provides Unix-style permissions and Access Control Lists (ACL) support via Distinguished Names or VOMS roles. File access is secured via these ACLs. The Fireman catalogue provides Web Services Interfaces with full WSDL availability. Bulk operations are supported. The catalogue exposes to so-called Storage Index interface used by the gLite Workload Management System to dispatch jobs at the relevant site. The DLI interface will be added soon. Metadata capabilities are supported through the use of key/value pairs on directories. FireMan supports Oracle and MySQL database back-ends. An overview of gLite data management can be found at Ref. [31], while the Fireman catalogue command line interface, Java and C++ API's are at Ref. [42].

4.1.2.4 Information Services

Information services publish and maintain data about resources in Grids. This information in LCG is modelled after the Grid Laboratory Uniform Environment schema (GLUE).

BDII

The Berkeley Database Information Index (BDII) is an implementation of the Globus Grid Index Information Service (GIIS), but allowing for more scalability. Information provided by the BDII adheres to the GLUE information model. Interfacing with BDII is made of ldap operations for which commands and API exist. Both LCG-2 and gLite currently rely on BDII for proper operation.

R-GMA

R-GMA is an implementation of the Grid Monitoring Architecture of GGF and presents a relational view of the collected data. It is basically a producer/consumer service with command line interfaces as well as an API for Java, C, C++ and Python and a Web interface. R-GMA models the information infrastructure of a Grid as a set of consumers (that request information), producers (that provide information) and a central registry which mediates the communication between producers and consumers. R-GMA (via GIN) can use the same information providers as used by BDII.

Recently, a Service Discovery mechanism using R-GMA has been implemented. Detailed information is available at Ref. [43].

R-GMA is currently also used to collect LCG accounting records.

The R-GMA and Service Discovery command line interface, Java, C, C++, Python APIs are available at Ref. [44].

Logging and Book-keeping

The Logging & Book-keeping services (LB), which tracks jobs during their lifetime in term of events (important points of job life, such as submission, starting execution, etc.) gathered from the WM's and the CE's (they are instrumented with LB calls). The events are first passed to a local logger then to book-keeping servers. More information on the Logging and Book-keeping services are available at Ref. [45].

Job Provenance

Job Provenance Services, whose role is to keep track of submitted jobs (completed or failed), including execution conditions and environment, and important points of the job life-cycle for longs periods (months to years) are being prototyped. This information can then be

reprocessed for debugging, post-mortem analysis, and comparison of job execution and re-execution of jobs. More information on Job Provenance Services is available at Ref. [46].

4.1.2.5 File Transfer Services

LCG-2 File Transfer Services

LCG-2 did not provide File Transfer Service *per se*. Rather it was up to the user to issue the relevant commands to replicate the files from one Storage Element to another. During the Service Challenge 2 in 2004, however, a set of *ad hoc* tools (Radiant) were developed for managing the huge amount of files to be moved from site to site.

gLite Transfer Services

The gLite File Placement Service (FPS) takes data movement requests and executes them according to defined policies. It maintain a persistent transfer queue thus providing reliable data transfer even in the case of network outage and interacts fully with the Fireman catalogue. The File Placement service can be used without the interaction with the catalogue and is then referred to as the File Transfer Service (FTS). It is planned to use the gLite File Transfer Service for Service Challenge 3 in summer 2005. The FTS command line interface and API are available at Ref. [47].

4.1.3 VDT

The Virtual Data Toolkit (VDT) is an ensemble of Grid middleware that can be easily installed and configured. The VDT was originally created to serve as a delivery channel for Grid technologies developed and hardened by the NSF-funded GriPhyN [48] and iVDGL [49] projects, and these two projects continue to be the primary sources of funding for the VDT. However, the role of the VDT has expanded and now supports the LHC Computing Grid Project (LCG) [50] and the Particle Physics Data Grid (PPDG) [51].

Both LCG-2 and gLite middleware components rely on the VDT versions of Condor, Globus, ClassAds and MyProxy. VDT provides direct support to LCG for those packages. LCG-2 and gLite components such as VOMS and information-providers, are also being added to VDT.

4.2 Grid Standards and Interoperability

During the past few years, numerous Grid and Grid-like middleware products have emerged. Examples include UNICORE [52], ARC [53], EDG/LCG-2/gLite [54], Globus [55], Condor [56], VDT [57], SRB [58]. They are capable of providing some of the fundamental Grid services, such as Grid job submission and management, Grid data management and Grid information services. The emergence and broad deployment of different middlewares has raised the problem of interoperability and standards. Unfortunately, so far the Grid community has not met the expectations of delivering widely accepted, implemented and usable standards.

We will therefore have to accept for some time, at least for the duration of Phase 2 of the LCG Project, more than one middleware implementation and find *ad hoc* solutions to provide a certain level of interoperability.

In the short-term the approach taken by the project is to define a set of basic services that can be deployed in all of the existing Grid infrastructures, taking account of their different technical constraints.

In the longer term we must continue to encourage the developers of middleware technology to work towards solutions that coexist and interoperate using well-defined interfaces and, where possible, community or more general standards.

4.3 Grid Operations and Regional Centre Service-Level Agreements

The operational control of the Grid services are the responsibility of each of the different Grid infrastructure projects. The Grid infrastructures involved in providing services to LCG are EGEE covering Europe and many sites in the Asia-Pacific region and Canada; Open Science Grid in the United States of America; and the Nordic Data Grid Facility in the Nordic countries.

The Grid Infrastructure projects provide two support services — Grid operational monitoring and support, and user support.

Grid Operations Centres (GOC) are responsible for providing essential Grid services, such as maintaining configuration databases, operating the monitoring infrastructure, providing pro-active fault and performance monitoring, provision of accounting information, and other services that may be agreed. Each GOC will be responsible for providing a defined subset of services, agreed by the project. Some of these services may be limited to a specific region or period (e.g., prime shift support in the country where the centre is located). Centres may share responsibility for operations as agreed by the project.

The user support activities provide a first- and second-level support structure and cover both the Grid and computing service operations. The first-level (end-user) helpdesks are assumed to be provided by LHC experiments and national or regional centres, depending on the structure agreed in that region. The second-level support is assured by Grid Call Centres. These centres function as service helpdesks and their role includes pro-active problem management. These centres would normally support only service staff from other centres and expert users. Each call centre is expected to be responsible for the support of a defined set of users and regional centres and will provide coverage during specific hours.

The remainder of this section describes the currently deployed operational support services and their expected evolution over the next few years. At the time of writing, formal operations and support infrastructures are in service within EGEE (including the participation of ASCC Taipei) and within Grid3/Open Science Grid. The Nordic Data Grid Facility infrastructure does not yet provide these services in a formal way.

4.3.1 EGEE Grid Operations

Within the Enabling Grids for E-SciencE project (EGEE) structure there are several different organizations currently defined to provide Grid operations and user support. These are:

- Operations Management Centre (OMC)
- Core Infrastructure Centres (CICs)
- Regional Operations Centres (ROCs)
- Grid User Support Centre (GGUS)

The Operations Management Centre is located at CERN and provides co-ordination for the EGEE Grid operation. In addition to management co-ordination it also provides the co-ordination of the deployment of the middleware distributions, the integration, certification, and documentation of the middleware releases, and the co-ordination of the deployment of those distributions. It provides support for problems found in the middleware, both directly through a small team of expert analysts, and also as a co-ordination point with the middleware developers and projects that supply the software. This is discussed in more detail in the section on life-cycle management.

The Core Infrastructure Centres (CICs) have two roles. The first is to run essential core Grid services such as database and catalogue services, VO management services, information services, general usage resource brokers, etc. In addition these operations centres provide resource and usage monitoring and accounting. The second role is to act as the front line Grid operators, and manage the day-to-day Grid operation. Here the CICs take a week as the

primary Grid operator, the responsibility being handed between the CICs in rotation. The responsibilities include active monitoring of the Grid infrastructure and the resource centres (Tier-1 and Tier-2), taking the appropriate action to avoid or recover from problems. Part of the responsibility includes the development and evolution of tools to manage this activity. The CICs also must ensure that recovery procedures for critical services are in place.

There is a tight coupling between the CICs, including the shared operational responsibility and close communication between the operations teams. Each CIC manager reports to the CIC co-ordinator at CERN. There is a weekly operations meeting where the current operational issues are discussed and addressed, and where the hand-over between the on-duty CICs takes place. This meeting also ensures that issues and problems in operations get reported back to the middleware developers, deployment teams, applications groups, where necessary and as appropriate.

CICs are currently operational at CERN, RAL (UK), CNAF (Italy), CCIN2P3-Lyon (France), and just starting (April 2005) at MSU in Russia. In addition, ASCC-Taipei provides resources to the operations monitoring activities and expects to also participate in the Grid operations shifts in the fourth quarter of 2005.

The Regional Operations Centres (ROCs) provide the front-line support in each of the geographical regions of the EGEE project. In the regions also operating CICs these functions overlap within the same organizations. In other regions the ROCs are distributed activities with staff in several physical locations. The roles of the ROCs include:

- Co-ordinating the deployment of the middleware releases within the region;
- Providing first-level support to resolve operations problems at sites in the region. The ROC must have the needed expertise and information on the operational state in order to diagnose problems as originating in the operation of the site, a service, or in the middleware itself;
- Providing support to the sites in the region in all aspects of the Grid operation, including providing training to the staff in the sites;
- Taking ownership of problems within a region, ensuring that they are addressed and resolved. They may refer them to the CICs or OMC for second-level support;
- Providing first-line support to users to resolve problems arising from the operation of the services in the region or within the middleware. It involves the VO support teams where necessary;
- Negotiating agreed levels of service and support with the sites in the region, and monitor them to ensure delivery of those levels of service.

The ROC co-ordinator is responsible for ensuring coherent and consistent behaviour of the several ROCs, and reports to the OMC.

Both the CICs and the ROCs generally are located at LCG Tier-1 sites, and in regions with no Tier-1 they take the role of the support functions of the Tier-1 centres.

The User support centre (GGUS) is currently located at Forschungszentrum Karlsruhe (FZK), with additional resources also provided by ASCC-Taipei. This centre provides a central portal for user documentation and support, providing access to a problem ticketing system. The system should ensure that problem tickets are dispatched to the appropriate support teams and that problems are addressed and resolved in a timely manner, according to an advertised policy. Although this provides a single point of entry for support for a user, the support teams are many and widely distributed, and cover all aspects of the operations, services, and the middleware. It is expected that each experiment also provides application support personnel who can address problems determined to be application-specific.

In building this Grid operations infrastructure it was clear that a hierarchical support structure was needed to support the Grid sites, since with a large number of sites a central organization

would not scale. The ROCs form the core of this hierarchy, with the CICs as second-level support and operational oversight, and the OMC at CERN as the co-ordinating and management centre.

The operations infrastructure described above is that built up during the first year (2004–2005) of the EGEE project. In the preparation for the second phase of EGEE (anticipated to be for two years beginning Spring 2006), the distinction between the ROCs and CICs will probably become less pronounced. Since the sites that operate a CIC are also Regional Operations Centres, with the operations and support teams shared between both sets of roles and responsibilities, the distinction is in any case somewhat blurred. (CERN is an exception to this, although it operates a CIC it does not formally have a ROC, but it does, however, act as the ultimate support centre for all unresolved issues and for other sites that are not part of EGEE but nevertheless participate in the same Grid infrastructure.) What seems reasonable for the longer term is to have a hierarchy of Regional Operations Centres, some of which take additional responsibility for operations oversight and management, and some of which run some of the essential core Grid services according to an agreed service level. It is essential to maintain the hierarchical structure of the operations activity to ensure that no single support centre has to manage more than a reasonable number of sites. In this way the operation of the Grid can scale in a manageable and predictable way to a much larger number of sites.

4.3.2 OSG Grid Operations

The operational model for Open Science Grid is somewhat different from that described above for EGEE. It provides a distributed support structure between the VOs, the sites, resource providers, service and technology providers, and the OSG operations. In particular, user support is fully the responsibility of the VOs who arrange for support to be provided directly or through collaborating organizations. The model described below is presently (May 2005) being implemented, although the iGOC has been in use during the previous year.

The operations infrastructure provides the following services: provisioning, distribution and configuration management of the Grid middleware; publishing accounting and monitoring information; providing communication channels for incident response, fault handling etc.; and support services including problem tracking and management systems.

The OSG Support Centres Technical Group provides co-ordination and guidance to the development of the operations structure. Various support centres provide support for VOs, resources, Grid services, middleware, and operations for the overall OSG Operations activity. This is the entity responsible for the daily operation of the OSG infrastructure.

In addition to the Support Centres, there will be one or more OSG-wide Operations Support Centres providing support for operational activities that have impact on the full infrastructure, providing grid-wide monitoring and accounting, change management, etc. At the present time there is one such entity — the iGOC (iVDGL Grid Operations Centre) in Indiana, co-located with, and leveraging the infrastructure of the Network Operations Centre.

Figure 4.2 illustrates the relationships between the various groups contributing to the OSG Operations activity. An example of its implementation is shown in Figure 4.2.

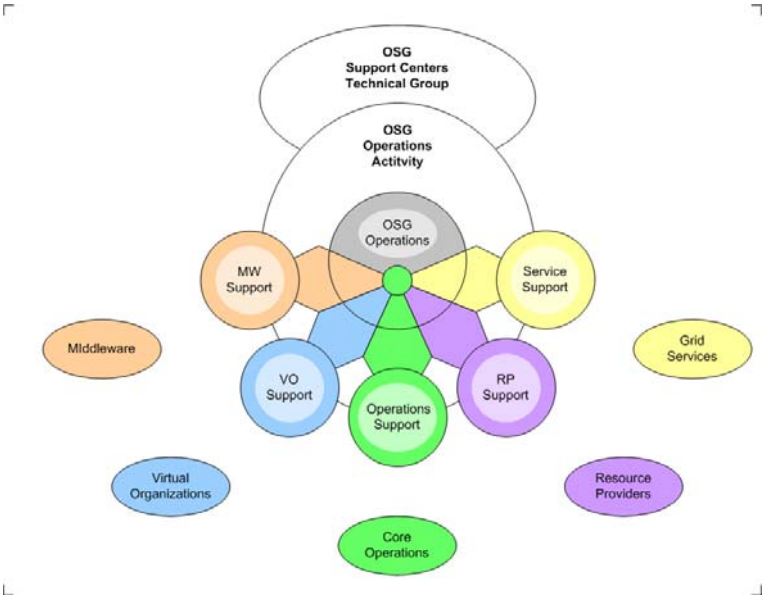


Figure 4.1: OSG Operations. The logical view of the support roles, the second shows an example of how these may be implemented by real support centre organizations.

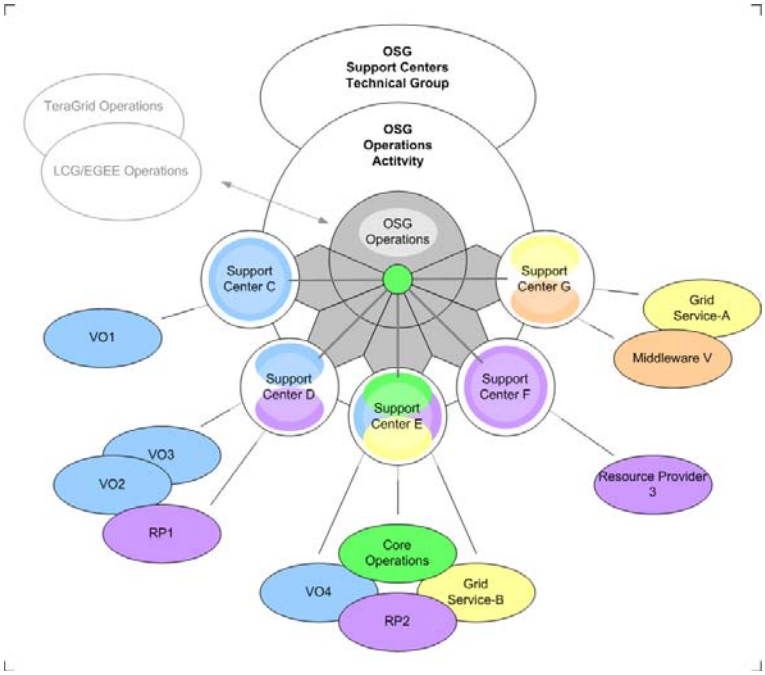


Figure 4.2: An example of implementation by real support centre organizations.

4.3.3 Nordic Data Grid Facility Operations

There is no formal operations activity within the NDGF. However, the sites providing resources to the LHC experiments are in countries that are part of EGEE, and consequently are covered by the Northern European ROC. The majority of those sites run the ARC middleware stack, not the EGEE one. This means that different information systems, services, and policies are in place.

4.3.4 Co-ordination of Operations

At the moment each Grid infrastructure manages its own operation as described in the preceding sections. Within EGEE, however, there is a move to more than prime-shift operations support by including Operations Support centres from different time zones. The first of these is ASCC-Taipei in Asia. It is to be hoped that additional centres in Asia–Pacific

will eventually contribute so that within Europe and within Asia–Pacific several centres can take shared responsibility for operational oversight for different hours of the day, and that between the centres in each region the responsibility is shared on a weekly rotation (for example).

Discussions are in progress between EGEE operations and the OSG operations activity on how such a sharing might also be implemented between the two projects. In this way a 24-hour operational oversight service can be envisioned for the full LCG Grid infrastructure. There are many aspects to such collaboration, since the middleware stacks, and the operational and management policies of the two Grid infrastructure projects are different. However, there are many commonalities and a willingness to collaborate. The details of how this sharing can be achieved are still to be understood, but a series of joint operations workshops will be sponsored by both projects with a goal of bringing about this co-ordination.

4.3.5 Security Operations

The operational aspects of Grid security include security monitoring, intrusion detection, security incident response, auditing, and the co-ordination of the production and deployment of required fixes and security patches. In Europe this activity is managed by the EGEE SA1 Operational Security Co-ordination Team (OSCT). This body consists of at least one representative per ROC and a small number of additional experts. These regional representatives are then charged with organizing security operations within their region. Links to other Grid operations, e.g., Open Science Grid, are also essential as security incidents can easily span multiple Grids. These links are being established.

At this time, most effort has been put into the definition of policy and procedures for Security Incident Response. The aim here is not to replace existing site and network procedures, but rather to supplement these with speedy information exchange between the Grid operations centres and site security contacts following a security incident. All Grid sites are required by policy to inform all other security contacts of any actual or suspected incident that has the potential to attack or affect other Grid sites. LCG/EGEE has agreed to base its Incident Response on the earlier work in this area by Open Science Grid and the procedures for exchanging incident information between Grids is also being explored.

4.3.6 Accounting

It is essential that usage of the compute and storage resources provided by the collaborating sites be accounted for and reported to the stakeholders — the experiments, the funding agencies, and the LCG Project. The project, and each site, must be able to demonstrate that it is providing the resources that it has committed to provide to the experiments, and that it has been done in a way consistent with the agreed shares of each experiment. This should be done both site-by-site, country-by-country, and project-wide. It is also important that the experiments be able to understand what resources it has consumed and who within the experiments has used those resources.

Both EGEE and Open Science Grid have accounting systems, with sites publishing the accounting data through similar schema (both based on the GGF schema). Currently each publishes the information independently, but there are discussions under way to agree the technical mechanisms, and the policies by which these data can be published into a common system to provide a full view of Grid accounting to the LCG stakeholders.

4.3.7 Service Level Agreements

The main service parameters such as availability, levels of support, response times for the Tier-0, Tier-1 and Tier-2 centres are laid out in the LCG Memorandum of Understanding, although the details of these will evolve as a better understanding of the requirements and the services is gained. The service levels set out in the MoU cover the essential services and problem response expectations for the various Tiers. In addition, both the EGEE and OSG

projects will document the actual levels of staff, response commitments, and so on with each of the centres providing resources to the Grid infrastructures. These service level definitions will also set out the commitments required of the participating centres in terms of security incident response, operational response, and commitment to appropriate service quality.

4.4 Life-Cycle Support – Management of Deployment and Versioning

The experience in Phase I of LCG and in other Grid deployment efforts has shown that in order to be able to deploy a service at anything like production quality, it is essential to have in place a managed process to integrate middleware components into a coherent distribution, to test and certify those components; and during deployment and operations to provide adequate feedback mechanisms based on experience to the appropriate development and deployment teams.

The elements of the managed process are the following:

- Integration into a coherent middleware distribution of middleware services and components from the various middleware suppliers.
- Testing and certification of the middleware distribution.
- Managed deployment process, including procedures for updates, security fixes, configuration management.
- Feedback loops from each stage to the appropriate teams.
- Commitments for maintenance agreements for all components must be in place.

There must also be adequate mechanisms in place to ensure that feedback on the usage of the middleware, Grid services, and common application layers, must be directed to the developers to ensure that required changes are included in a timely manner. The general process is shown in Figure 4.3 below.

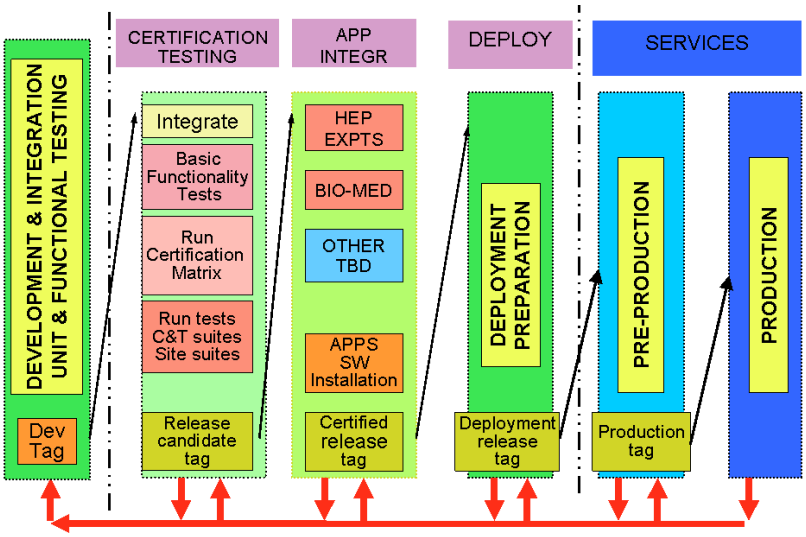


Figure 4.3: Deployment life-cycle

4.4.1 Elements of the Life-cycle Process

The overall deployment life-cycle is shown in Figure 4.3. It is assumed that middleware components, application libraries and tools, etc. are delivered from the developers having undergone a reasonable level of unit and functional testing, including testing of essential integration with other components on which they rely. The elements of the life-cycle are described in the following.

4.4.1.1 Testing and Certification

At this stage there are several activities that culminate in a coherent set of middleware, services, and tools which are certified as working together in a reasonably reliable and robust manner.

- Integration of the various components from the different middleware and tools suppliers, ensuring that the tools and services are able to work together as expected, and that they can co-exist with their various external dependencies;
- To test the integration a basic set of functional tests are run – ideally including the tests provided by the middleware developers, and this testing should be able to verify the test results of the developers.
- A set of functional tests is run, to test the system as a whole, or to test various sub-systems. This should include regression testing, in order to determine that new releases do not break existing functionality or degrade the performance. At this stage, the basic test suite that is run daily on the deployed system (the Site Functional Test (SFT) suite) is also run.
- Once the basic integration, functional, and regression testing is done, the candidate release is subjected to a week of stress-testing, which includes tests such as large job storms, data movement storms, etc. which attempt to overload the system and test its robustness.

Once these tests have been run successfully, or to a point where the remaining problems are at an acceptable level, a tag is made on each component to label the set of consistent versions. Of course, during this entire process, problems are fed back to the component and service suppliers, and fixes to the problems included into the tested versions. This tight feedback is essential to the success of this activity, and it is vital that all of the suppliers have committed to appropriate levels of responsiveness and support.

The testing and certification test-bed for LCG is large, containing close to 100 machines. It attempts to simulate most of the environments into which the middleware will be deployed — various batch systems, different architectures, etc. At this point the certification test-bed does not include remote sites, so cannot test issues related to wide-area networks, but current experience has shown that it is more expedient to run the certification as a centrally managed process.

4.4.1.2 Application Integration

Once a potential release has been tagged in the certification process, it is deployed to a small test-bed set-up for the applications to begin their validation tests, and if new services or tools are provided, to be able to begin the integration of those tools with the experiment software stacks. A small team (the Experiment Integration Support team) aids the experiments in this task and ensures that problems encountered at this stage are fed back to the certification process or to the developers as appropriate. At this stage also integration of common application layer components (such as POOL) is tested and verified. Once these tests are complete a candidate release is again tagged together with any additional certification testing.

4.4.1.3 Release Preparation

In parallel with the certification and application integration procedures the work to prepare the middleware release for deployment is started. This includes updating the installation tools, the configurations, and the associated installation instructions. These are tested through test deployments on a small test-bed. In addition, the user documentation (user guides, example jobs, etc.) is completely revised to ensure consistency with the new release. Finally a complete set of release notes is compiled to describe the release, including changes, known problems, and important differences to previous versions, as well as points that must be

considered before the system is installed. The release notes are provided both for the sites installing the release and for the users.

4.4.1.4 Pre-Production Service

Once a release has been built as described above it is first deployed onto the Pre-Production Service. This is a small number of sites (~10) who provide resources to this service. This pre-production deployment allows several things:

- Testing of the deployment and release itself in real situations.
- Testing new functionality of the applications before they move into the production service.

Of course, there is the possibility to provide feedback and subsequent updates to the release based on the pre-production experience, before moving to production. To enable scalability testing, it is expected that resources can be dynamically assigned to the pre-production service by the participating sites when the need arises and is scheduled.

4.4.1.5 Production Deployment

Finally, once the release is accepted by the stakeholders, it can be scheduled for deployment into the production system. This acceptance should be based on the experience on the pre-production service.

4.4.2 Layered Services

The process described above treats all services and tools at the same level. Of course, in reality we can distinguish three sets of components:

- Core services that must run at each site, and for which coherent and scheduled updates must be done. These include the Computing Elements, the Storage Elements, local Grid catalogues, and other services which may affect access to resources.
- Other services, such as information providers, components of the information system, monitoring services, Resource Brokers, central Grid catalogues, etc. Many of these components can be updated independently of the core services.
- Client tools and libraries. These can be installed on the worker nodes by a user-level process, and do not require privileged access.

While all of these of course require testing and certification, those components that are not critical for applications (such as monitoring components) should not hold up the release of critical components or core services. Different release and deployment time-scales for these three layers can be foreseen, with the core components requiring a scheduled upgrade across the entire infrastructure, the other services can be upgraded asynchronously with the agreement of the sites and applications affected, while the client tools can be updated at will since the existence of them at a site is published through the information system and they can coexist with previous versions. Of course, issues of compatibility with other deployed services must be tested and advertised.

This layered view is essential for managing a large Grid infrastructure, core services can only be coherently updated on very long and well-scheduled time-scales, while the other services and tools require more frequent changes to fix problems, and provide new functionality.

Figure 4.4 and Figure 4.5 show different views of these processes. The first illustrates the process to produce the releases; the second shows the release process itself.

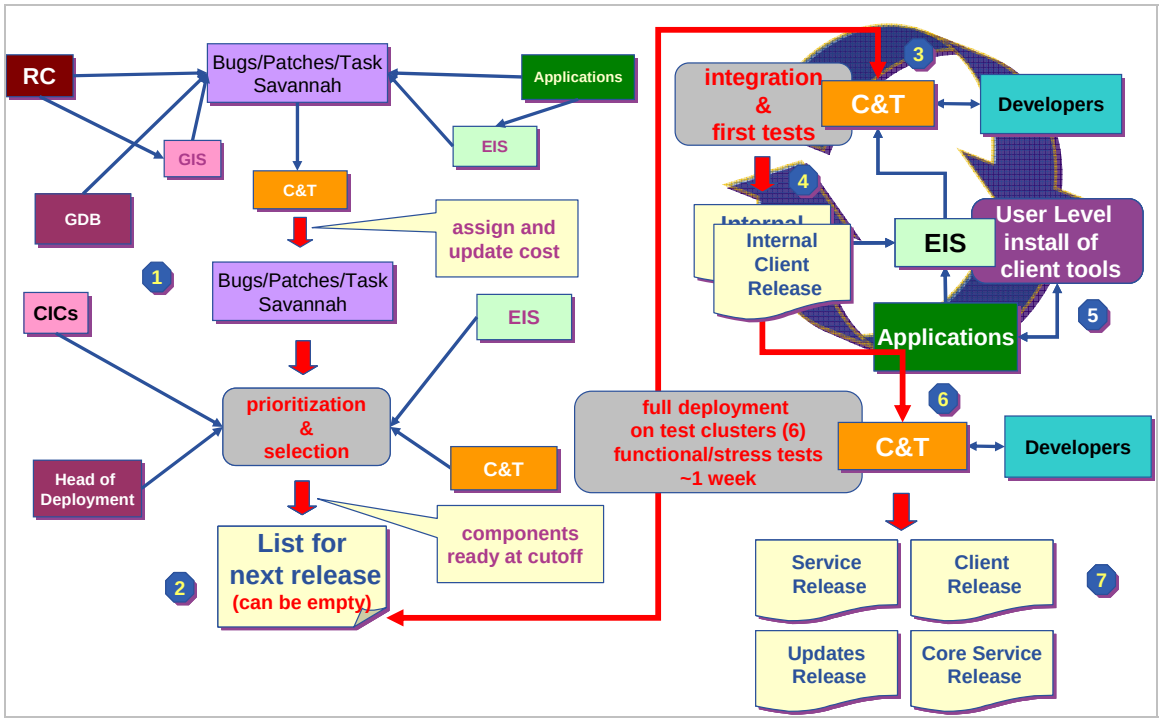


Figure 4.4: Middleware distribution preparation process

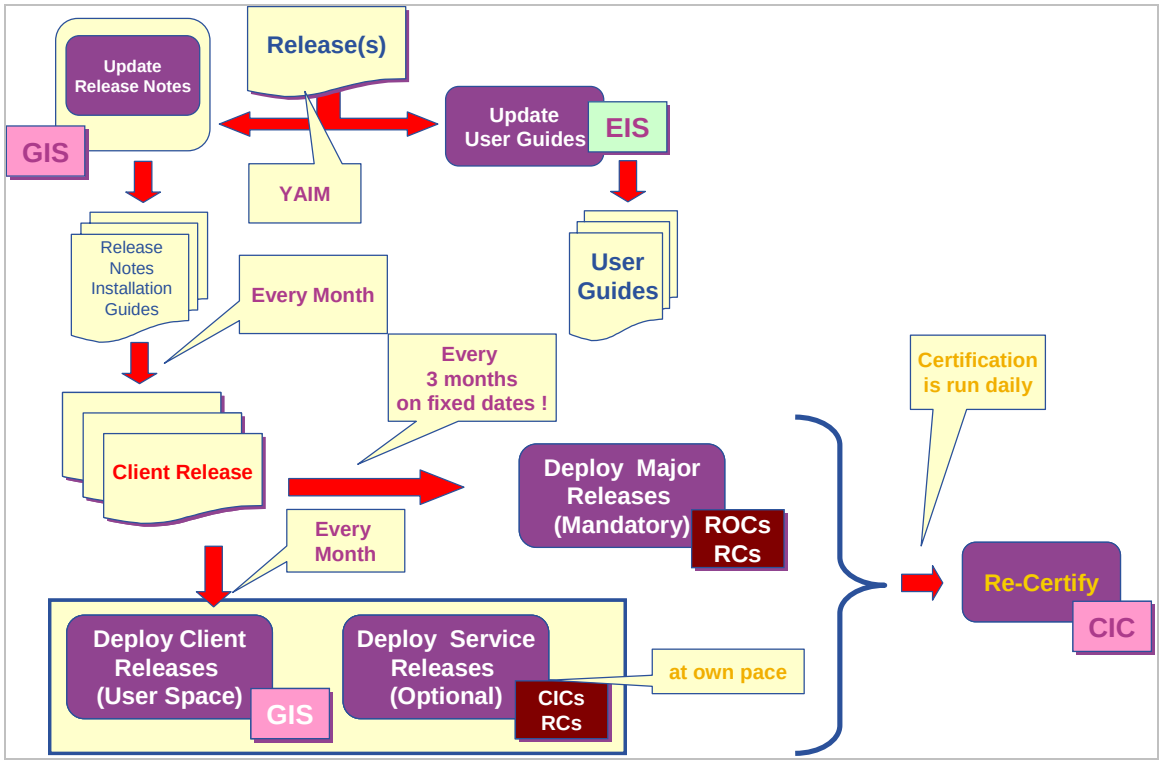


Figure 4.5: LCG release process

4.4.3 Management of the Process

The processes described in the preceding sections involve a number of groups with different roles and responsibilities.

4.4.3.1 Middleware Suppliers

The middleware suppliers are responsible for providing the basic Grid services and tools to respond to the needs of the experiments. Maintenance and support agreements must be in place with all of the groups providing these tools to ensure that problems reported from the testing, deployment, and use of the software are addressed appropriately.

4.4.3.2 Certification and Deployment Team

The certification team manages the testing and certification process, preparation for deployment, and co-ordinates the deployment to the sites providing resources. The team also acts as the conduit for problem reporting for all of these activities.

4.4.3.3 Resource Providers

The resource providers are responsible for installing the appropriate middleware releases, services, and tools, and for ensuring that these installations are done in a timely manner according to agreed schedules. They should report all problems back to the deployment team. Some of the resource providers will also provide resources as part of the pre-production service. The resource providers are ultimately responsible for ensuring that security updates are applied when required.

4.4.3.4 Regional Operations Centres

The regional operations centres are responsible for co-ordinating the deployment within a region, and for providing support to the sites in their region during this process. They are responsible for ensuring that security-related updates are applied at all sites.

4.4.3.5 Core Infrastructure Centres

The Core Infrastructure Centres are responsible for reporting problems encountered during operation to the appropriate teams so that they may be addressed for future releases or updates.

4.4.4 Change Management

In a production system it is essential that new or upgraded services do not break compatibility between services, such changes must be backwards compatible, unless a full migration plan and upgrade can be scheduled and implemented with the agreement and participation of the stakeholders that are affected. It is vital that middleware developers, tools suppliers, etc. understand this point, and that essential changes that do break compatibility should address these issues as part of the development and the proposed change.

4.4.4.1 Security Patches

Patches, updates, and changes may be required urgently to address security concerns. In these cases, the full process described above may be very much reduced or steps avoided altogether, with patches being provided as rapidly as possible, with an expectation that the affected services or software be patched as soon as possible in the deployed system by the sites in co-ordination with the ROCs.

4.5 Fabric Technology — Status and Expected Evolution

4.5.1 Processors

4.5.1.1 Microprocessor Process Technology

The HEP community started relying heavily on commodity x86 PCs in the late 1990s. This period represented a ‘golden’ expansion period during which the manufacturers were able to introduce new semiconductor processes every two years. Increased transistor budgets allowed more and more functionality to be provided and the size reduction itself (plus shortened

pipeline stages) allowed a spectacular increase in frequency which led to more performance for lower cost.

Nevertheless, the industry has now been caught by a problem that was almost completely ignored ten years ago, namely heat generation from leakage currents. As the feature size decreased from hundreds of nanometres to today's 90 nm (and tomorrow's 65 nm) the gate oxide layer became only a few atom layers thick with the result that leakage currents grew exponentially.

Moore's law, which only stated that the transistor budget grows from one generation to the next, will continue to be true, but both the problems with basic physics and the longer verification time needed by more and more complex designs may start to delay the introductions of new process technology. The good news for HEP is that the transistor budget will from now on mainly be used to produce microprocessors with multiple cores and we are already starting to see the first implementations (see Section 4.5.1.4).

4.5.1.2 64-bit Capable Processors

64-bit microprocessors have been around for a long time as exemplified by, for instance, the Alpha processor family. Most RISC processors, such as PA-RISC, SPARC and Power were also extended to handle 64-bit addressing, usually in a backwards-compatible way by allowing 32-bit operating systems and 32-bit applications to continue to run natively. In 1999 Intel came out with IA-64, now called the Itanium Processor Family (IPF), with a strong focus only on 64-bit addressing.

AMD chose an alternative plan and decided to extend x86 with native 64-bit capabilities. This proved to be to the liking of the market at large, especially since the revision of the architecture brought other improvements as well. The architectural 'clean-up' gives a nice performance boost for most applications. After the introduction of the first 64-bit AMD Opterons, Intel has realized that this was more than a 'fad', and, within a short time window we are likely to see that almost all x86 processors integrate this capability. During a transition period it is unavoidable that our computer centres will have a mixture of 32-bit hardware and 32/64-bit hardware, but we should aim at a transition that is as rapid as possible by acquiring only 64-bit enabled hardware from now on.

4.5.1.3 Current Processors and Performance

Today, AMD offers single-processor Opteron processors at 2.6 GHz whereas Intel offers Pentium 4 Xeon processors at 3.6 GHz. Both are produced in 90 nm process technology and, as far as performance measurements are concerned, both provide market-leading SPECINT2000 results.

AMD has just announced the first series of dual core Opteron processors with speeds between 1.8 and 2.2 GHz. Intel has announced a dual-core P4 'Extreme Edition' at 3.2 GHz. They are expected to have dual-core Xeons available by the end of this year (in 65 nm technology). In general a dual-core processor is likely to run 10–15% slower than the uni-processor equivalent but should offer throughput ratings that are at least 50% higher.

Intel's current 1.6 GHz Itanium processor, although it has an impressive L3 cache of 9 MB, currently offers less performance than the x86 processors. However, this is a processor produced in 130 nm and performance results from the forthcoming 90 nm Montecito processor, which is dual core with dual threads, are not yet available.

IBM has offered dual-core processors for some time. The current 90 nm 1.9 GHz Power-5 processor (with a 36MB L3 off-chip cache!) offers acceptable (but not great) SPECInt results. A more popular version of the Power-based processors is the G5 processor used in Apple Macintosh systems. Frequencies now reach 2.7 GHz. There is little doubt that the Apple systems are growing in popularity as witnessed by the recent uptake of desk-/laptop Macintosh systems by the physicists. A third initiative from IBM is the Cell processor which is destined for game systems as well as other (more traditional) computer systems. This is a

novel design with several 'attached' processors linked to a central management unit. It is too early to say whether the Cell processor will have an impact on LHC computing or not, but the evolution of this processor (with its required software environment) should be watched closely.

SUN is another player in the processor market. Their UltraSPARC processor is usually targeted at the high-end server market, and in any case, its SPECint results are rather lacklustre. A new development will be the 'Niagara' processor that is said to come out with eight cores, each core capable of running four threads. Such a 32-way engine will be entirely focused on throughput processing (in Web services and similar environments), and each core will be relatively simple. A second generation is said to contain additional features needed by HEP jobs, and may be an interesting development.

All in all, the attractive x86 performance results, combined with the competitive structures in today's market, leave little opportunity for non-x86 contenders to make themselves relevant. The situation is not likely to change in the near-term future since both AMD and Intel will continue to fight for market leadership by pushing their x86 offerings as far as they can (to our great benefit).

4.5.1.4 Multicore Future

For the HEP community it would be great if the semiconductor world would agree to push a geometric expansion of the number of cores (from 2, to 4, or even 8). The main problem will be the 'mass-market acceptance' of such a new paradigm and some sceptics believe that large-scale multicore processors will gradually limit themselves to the 'server niche' which may not be dominated by commodity pricing in the same way as today's x86 market with its underlying one-size-fits-all mentality.

4.5.1.5 Summary and Conclusions

- All LHC experiments should make a real effort to ensure that their offline software is '64-bit clean'. This should be done in such a way that one can, at any moment, create either a 32-bit or a 64-bit version of the software.
- The LCG sites should concentrate their purchases on the x86-64 architecture. The 32-bit-only variants should be avoided since they will act as a roadblock for quick adoption of a 64-bit operating system and application environment inside the LHC Computing Grid.
- Should IPF, Power, or SPARC-based systems become attractive in the future, our best position is to ensure that our programs are 64-bit clean under Linux/Unix.

4.5.2 Secondary storage: hard disks and connection technologies

4.5.2.1 Hard Disk Drives — Market Survey

The rationalization and consolidation of disk manufacturing observed in the last couple of years has continued and the market is now dominated by the four companies Seagate, Maxtor, Western Digital and Hitachi which account for 84% of the market. In the 3.5 inch sector used for data storage in HEP, the market is divided into desktop and enterprise drives with different performance characteristics.

4.5.2.2 Hard Disk Drives — Capacity and Performance

In the 2002—2004 timeframe, there was a small slowdown in the increase in aerial density as the physical limits of longitudinal recording were revealed. Nonetheless, the capacity of a 3.5 inch platter has continued to double roughly every 18 months. At the time of writing, the largest drive is a Hitachi 500 GB unit made from five 100 GB platters. 3.5 inch drives

dominate the bulk storage market with few units produced exclusively in the 1 inch form factor.

In the same time period, disk rotation speeds have remained constant. Desktop drives operate at 5400 or 7200 r.p.m. whereas high-end drives operate at rates up to 15,000 r.p.m. reflecting the need for data access performance as well as transfer rate. Disks are now produced with as few platters as possible. Such an approach is in the interests of simplification and reduced cost. Disk manufactures expect the current longitudinal recording technology to reach its limit with platters of 160 GB. Beyond this, developments based on perpendicular recording will be used and Maxtor have demonstrated a platter of 175 GB using this technology.

4.5.2.3 Desktop and Enterprise Drives

Desktop and *Enterprise* are two terms commonly used to characterize separate market segments for 3.5 inch drives. Enterprise drives are designed for incorporation into storage systems which offer high levels of data availability. Desktop drives, as the name implies are aimed at the PC market where low price/GB is the primary factor and the drives are generally higher capacity units. Enterprise drives are offered with Fibre Channel (FC) and SCSI interfaces and in the future will have Serial Attached SCSI (SAS) interfaces. Desktop drives are marketed with Serial ATA (SATA) interfaces. Desktop drives are assumed to have a daily utilization of about 8 hours whereas enterprise systems operate 24 hours. This factor of 3 in duty cycle is reflected in MTBF figures that are quoted for the two categories.

4.5.2.4 Disk Connection Technologies for Commodity Storage

For the commodity storage likely to be used for LHC experiments, the important connection technologies are SATA and to a lesser extent, SAS.

Serial ATA - SATA

SATA technology has seen a rapid uptake for several reasons. It uses a simple four-wire cable which is lighter and has simple, more reliable connectors than the parallel ATA ribbon cable. The integration and development of industry chipsets to support serial ATA is facilitated by the lower voltages and a reduced pin count when compared to parallel ATA. Finally, its high-performance transfer speed starts at 150 MB/s with an evolution to 600 MB/s.

Serial Attached SCSI - SAS

SAS is a technology that has emerged in the last couple of years. It is aimed at the enterprise storage market. It is seen as a natural follow-on to parallel SCSI and in terms of cost, will be similar to current SCSI or FC storage. SAS shares a lot of commonality with SATA at the hardware level including the longer cabling distances and small connectors. In terms of end-user benefits, SAS maintains backwards compatibility with the established base of SCSI system and driver software. SAS supports dual porting of disks which is the key to multi-pathing and high availability RAID storage.

The first SAS products, hard disks and host adapters, are expected in 2005 and the plan is for SAS to replace both parallel SCSI and FC Arbitrated Loop for disk connectivity. SAS is seen as complementary to SATA by adding dual porting needed for high availability/reliability environments. SATA, however, is targeted to cost-sensitive, non-mission-critical applications.

Integrated PC-Based File Server

The integrated PC-based disk server is a cost-effective architecture in terms of GB/$ and is widely deployed in the HEP community for staging space.

At CERN, for instance, the units are usually 4U form factor and comprise 20 SATA disks connected with three 3Ware 9,000controllers. The disk drives in the latest purchases are 400 GB and provide several TB of space in either a mirrored or RAID5 configuration. Operational experience with the type of system at CERN has been mixed. In only about 50%

of hardware interventions is it possible to resynchronize a broken mirror without impacting the end-user service: power cycle, reboot, component exchange.

Low-Cost External RAID Storage

An alternative to the integrated file server is to physically separate the physical storage from the PC server. Low-cost RAID controllers are populated with 16 SATA drives and connected via a high-speed FC link to a server. One option is to build 3 RAID5 volume elements, each with capacity of 1.6TB, leaving 1 disk in 16 assigned as a hot spare. For environments like CASTOR where the workload and access patterns are chaotic, the simplest approach would be to build file systems from striped mirrors. This storage model does have the advantage that the storage connects over a standard FC link and therefore is more loosely coupled to the release of the Linux operating system. Issues of firmware compatibilities are handled by the RAID controller.

Conclusions

- The hard-disk-drive market has seen consolidation and specialization and profit margins remain low. In spite of this, technology developments have meant that raw storage costs have continued to fall by a factor of 1.4 per year.
- Developments in drive technology, particularly aerial density and capacity, have far exceeded earlier predictions.
- In the interest of simplified head structures, disks are produced with fewer platters as the recording density increases.
- With longitudinal recording technology, manufacturers expect to produce platters of 160 GB in Q4 of 2005. Higher densities will be achieved using perpendicular recording technology and the first laboratory products are emerging. In practical terms, the so-called 'Super Paramagnetic' effect will not limit disk storage densities in the LHC time frame and a PB disk storage analysis facility will be feasible at acceptable costs.
- The critical issues as far as LHC data analysis is concerned are likely to remain data access techniques and operational availability of large farms of disk servers.
- SATA disk drives are in widespread use for both PC and server storage systems. SATA2 which supports transfer rates of 320MB/s has addition command queuing techniques similar to SAS/SCSI. These features are targeted at disks in storage systems rather than purely PC desktop applications.

Given the current industry trends, SATA drives will continue for several years to be the storage technology that minimizes the cost per gigabyte at reasonable speed. This fact would indicate that LCG should continue to invest in this technology for the bulk physics storage requirements.

4.5.3 Mass storage – Tapes

There will still be a large amount of local tape-based storage in the LHC era. At present HEP still uses 'tertiary storage' on tape for relatively active data, because an entirely disk-based solution cannot be afforded. Unfortunately for HEP, the driving force in the tape storage market is still back-up and archiving of data. This implies that most data is mostly written just once and never read back. Drives are often designed to run for considerably less than 24 hours a day, and to stream data continuously at full speed to tape. HEP reading patterns are less efficient, however, and show an efficiency of use of ~5–10% because of their 'chaotic' access pattern. Tapes used for back-up and archive usually have a short lifetime, so little if any data needs to be carried forward from one generation of equipment or media to its successor. In contrast, the long life-time of HEP data implies at least one if not two migrations of data to

new media and devices over the useful life of the data. Unlike disk server equipment, tape equipment is definitely not a commodity. Linear Tape Open (LTO) and Super DLT drives approach 'commodity' status, but the LTO robotic unit, for example, is still well above commodity pricing.

4.5.3.1 Relevant Tape Equipment

Table 4.1 and 4.2 summarize the robotic tape libraries and drives of potential interest to the HEP community:

Table 4.1: Tape drives

Company	Drive	Capacity (GB)	Peak speed (MB/s)	Comments
Multiple	LTO 3	400	80	
Quantum	SDLT 600	600	32	
IBM	3592	300	40	
IBM	Next gen	~500	~100	
STK	9940 B	200	30	
STK	'Titanium'	~500	~100	
Sony	SAIT	500	30	Cartridge in 3480 form factor. Helical.

Table 4.2: Robotic tape libraries

Company	Library	Max cartridges	Max drives	Comments
ADIC	AML J	7560	226	Mixed media (supports 20 drive types)
	AML 2	76,608	256	Mixed media (supports 20 drive types)
	Scalar 1000	~1000	48	Mixed media
	Scalar 10K	15,885	865	Mixed media
IBM	3494	6,240	60	Supports 3590 and 3592 drives
	3584	5,500	180	Supports 3590x and LTO drives
StorageTek	Powerhorn	96,000	640	Up to 16 interconnected libraries.
	Streamline 8500	300,000	2048	Up to 32 interconnected libraries. Supports most drives except IBM's.

4.5.3.2 Tape Technology Considerations

The peak recording rates of current/imminent drives is quite sufficient for capturing data at anticipated LHC rates, since parallel tape transfers using reliable automated equipment is not a serious problem for the 2–4 GB/s data rates required.

Magnetic tape head technology now benefits directly from disk technology advances. Advanced-technology heads and disk head movement as well as servo tracking technologies are 'free' for use. Thus there is still no technical barrier to increasing cartridge capacities or data transfer rates — it is only a question of market demand.

One change that could still provoke an architectural shift in HEP is the use of Fibre Channel attached tape drives. This makes the classical 'tape server layer' potentially redundant. Remote systems attached to an FC fabric could address the tape drive directly at full speed, without any intermediate layers of tape servers and disk servers. This approach has been tried at CASPUR and demonstrated to work over hundreds of kilometres.

If LCG moves closer to the 'all data on disk' paradigm with tape systems used only for mass archive and recall, lower-cost LTO systems might answer the LHC requirements in 2007. This is, however, still unclear.

4.5.3.3 Tape Equipment Costs

Robotics: Today robotic costs are still low at the high-capacity end owing to continued production and support of the 3480 form-factor STK Powderhorn. The cost (~50 CHF/slot) is not likely to change much, so the question is the capacity of the cartridge.

Units: The cost of an LTO 3 drive is presumed to be ~20 kCHF, compared to ~50 kCHF for the high-end products.

Media: The 'standard cost' of a cartridge at the high end of the market seems to be quite steady at ~150 CHF at the time of its introduction to the market. Over time this cost may drop by 20–30%. In the future, however, the profitability of media manufacturers will be under pressure. Production runs are smaller than in the past, and the lifetimes of particular products quite short. One the other hand, the numerous competing suppliers of LTO media may make this an increasingly cost-effective option. If the price of tape storage does not fall significantly by 2006, massive disk storage could look relatively inexpensive from the capital cost viewpoint.

4.5.3.4 Conclusions

- CERN should expect not to be using current Powderhorns for LHC start-up. The latest generation of libraries should be installed and evaluated on at least a '1 Powderhorn' scale.
- Several drives can be seen as candidates suitable for LHC. The linkage between drives and libraries must obviously be kept in mind. Investment costs are high, but media represent ~50% of the costs.
- For estimating total drive capacity one should use conservative rates of ½ the peak speed (or less). CERN, as the Tier-0 centre, should expect to need at least ~100 drives in 2006, but note that drive costs are only ~30% of the overall total costs.
- Expect slow drifts downwards for drive and media costs.
- Be prepared to replace all existing media, which implies a repack of several petabytes (probably by 1Q2007).

4.5.4 Networking

4.5.4.1 Ethernet

From its origin more than 30 years ago, Ethernet has evolved to meet the ever-increasing demands of packet-switched networks. Its popularity has grown to the point where nearly all traffic on the Internet originates or ends with an Ethernet connection. Consequently, Ethernet will govern all LCG networking, be it as LAN, MAN or WAN. (A competing cluster interconnect technology, Infiniband, is reviewed in the next section.)

Ten-Gigabit Ethernet

In the last few years ten-gigabit Ethernet has moved from the early prototype stage (with prohibitive pricing) to a more solid production environment (with more acceptable price levels). Multiple vendors (Enterasys, Force 10, Foundry, CISCO, and possibly others) now offer powerful core routers with backplanes capable of switching traffic at the rate of about one terabit per second. Costs per port are now in the $10,000 range.

Network Interface Cards (NICs) at ten Gb/s have also matured and in a series of land speed record tests between CERN and Caltech single severs have been capable of speeds up to 7.4 Gbps, mainly limited by the PCI-X bus in the servers. Servers with the next generation bus

technology, PCI-X2 or PCI-Express, will undoubtedly be able to drive ten-gigabit NICs at peak rates.

One-Gigabit Ethernet

Full commoditization has occurred with this technology. Most, if not all, PCs come with 10/100/1000 Mb/s ports directly on the motherboard. Switches at the edge of the network typically provide dozens of one gigabit port aggregated via a couple of ten-gigabit uplinks allowing sites easily to decide the desired oversubscription (if any).

Future Evolution

Both 40- and 100-gigabit standards have been proposed as the next step in the evolution of high-speed Ethernet. The proposal for 40-gigabit aligns itself with the SONET hierarchy which moves in multiple of four from OC-192 to OC-768. Companies are already working on solutions in the category.

A jump directly to 100 Gb/s would follow the Ethernet historical trajectory but products are definitely further away from realization.

In any case, as far as HEP is concerned, we can expect the current ten-gigabit technology to satisfy our needs for many years to come.

4.5.4.2 Infiniband

Infiniband (IBA) is a channel-based, switched fabric which can be used for inter process communication, network and storage I/O. The basic link speed is 2.5 Gb/s. Today, the common link width is 4X (10 Gb/s), bidirectional but higher speed implementations will soon enter the market. Copper cables can be used for distances up to ≈15 m. Fibre optics cables are available for long-distance connections, however, prices are still high.

IBA host channel adapters (HCAs) are available as PCI-X and PCI-Express versions. Several companies offer modular switch systems from 12 4X-ports up to 288 4X-ports as well as higher speed uplink modules and Fibre-Channel and gigabit-Ethernet gateway modules to provide connectivity to other networks.

With its RDMA (Remote Direct Memory Access) capabilities, current 4X IBA hardware allows data transfer rates up to ≈900 MB/s and latencies of 5 μs and below.

Several upper layer protocols are available for Inter Process Communication (IPC) and network as well as storage I/O. Additionally, a prototype implementation of RFIO (as used by CASTOR) is available which allows the transfer of files at high speed and very low CPU consumption.

Infiniband drivers are available for Linux, Windows and some commercial UNIX systems. The low-level drivers have recently been accepted for inclusion into the Linux kernel starting with version 2.6.11.

IBA prices have been falling rapidly over the last few years. 4X switches can be purchased for ≈$300/port and less, dual-4X HCAs are ≈$500, and cables are available for ≈$50–150. These prices fall drop quickly since the first manufacturers announced implementing IBA on the mainboard directly connected to the PCI-Express bus. Other developments with a direct IBA-Memory connection are under way. These developments will not only ensure further price reductions and a wider market penetration of IBA, but also enable lower latency making it more suitable for very low latency dependent applications.

4.6 Databases – Distributed Deployment

LCG user applications and middleware services rely increasingly on the availability of relational databases as a part of the deployment infrastructure. Database applications like the

conditions database, production workflow, detector geometry, file-, dataset- and event-level metadata catalogs will be deployed from online and offline components of the physics production chain. Besides database connectivity at CERN Tier-0, several of these applications also need a reliable and (grid)-location-independent service at Tier-1 and 2 sites to achieve the required availability and scalability. LCG addresses these requirements with a scalable and open architecture taking into account the existing experience and the available resource at CERN and outside sites. In particular the recent deployment experience from FNAL RUN2 experiments has in many areas provided an important starting point for the proposed set-up for LCG.

4.6.1 Database Services at CERN Tier-0

The database services for LCG at CERN Tier-0 are currently going through a major restructuring to be ready for the LCG start-up. The main challenges are the significant increase in database service requests from the application side together with the significant remaining uncertainties of the experiment computing models in this area. To be able to cope with the needs at LHC start-up, the database infrastructure needs to be scalable not only in terms of data volume (the storage system) but also in server performance (number client sessions, server CPU, memory and I/O bandwidth) available to the applications. During the ramp-up phase (2005/2006) with several key database applications still under development, a significant effort in application optimization and service integration will be required from the LCG database and application development teams. Given the limited available manpower this can only be achieved by planning of the application life-cycle and adhering to a strict application validation procedure.

Based on these requirements at Tier-0 a homogenous database service based on the existing Oracle experience is proposed. Even though scalability in data volume into the multi-petabyte area may be required in the medium term, this will not be the main challenge in the early start-up phase. In contrast to traditional database deployment for relatively stable administrative applications, the database deployment for LCG will face significant changes of access patterns and will (as most other areas of physics data management) typically operate close to resource limitations. Automated application and database resource monitoring and the provision of guaranteed resource shares (in particular server CPU, I/O and network connections) to high-priority database applications are of crucial importance to ensure stable production conditions. As automated monitoring and throttling are still new to the database service area, a significant service development effort during the first deployment phase has to be expected to insure a controlled environment by LCG ramp-up.

4.6.1.1 Database Technologies for Tier-0 Services

The recent Oracle 10g release offers several technologies for setting up a flexible, scalable, and highly available infrastructure. As the application requirements are not well known and the database deployment is still ramping up to a realistic deployment scenario, these technologies will still have to be validated. The discussion here therefore contains the main elements of a service strategy and validation plan rather then already proven components as in other sections of this document.

Oracle 10g Real Application Clusters (RAC) promises to provide the technology for building database clusters that provide higher availability than single database and at the same time allow to scale the server CPU with the application demands. By adding nodes to a cluster, the number of queries and concurrent sessions can be increased together with the total amount of database cache memory, which is shared across all cluster nodes. How far a RAC set-up will be able to scale for a given application depends on the application design. Limiting factors are typically inter-node network communication (cache coherency) and application contention on shared resources, which need to be identified in validation tests and can often be avoided by application design. To control these scalability limitations, a close interaction between application developers and database administration team is required.

In a RAC set-up the Oracle 10g 'service' concept allows one to structure larger clusters into groups of nodes that are allocated to particular database applications (e.g., online configuration DB, Grid file catalogue). This pre-allocation of cluster resources is required to limit inter-node communication and to isolate key applications from lower-priority tasks executing on the same cluster.

In addition to CPU scalability, RAC does also provide increased availability. In case of unavailability of a server node (e.g., because of a hardware or software problem or a planned service intervention) the system will redirect incoming client connections automatically to other cluster nodes. Open transactions may still be affected in case of a node failover and will be rolled back. This needs to be taken into account in the retry logic of database applications.

Not all of the Oracle patches can be performed as 'rolling-upgrades' on individual nodes transparent to the service. Therefore more experience will be required to estimate the service availability which realistically can be achieved with a RAC set-up. Of particular importance for the service availability are the Oracle security patches, which have increased in frequency during the last year.

As of today clusters of 16 Linux server nodes are in production at other Oracle sites and a prototype system with several RAC set-ups (two to eight nodes) is being evaluated at CERN. The server nodes in this set-up consist of mid-range dual CPU machines under Linux, to achieve cost efficiency and integration into the existing fabric infrastructure. The database nodes are connected to a shared storage system based on Fibre Channel attached disk arrays as shown in Figure 4.6. This fulfils the requirements of the 'shared-everything' architecture of Oracle and allows scaling of the storage system and CPU requirements independently.

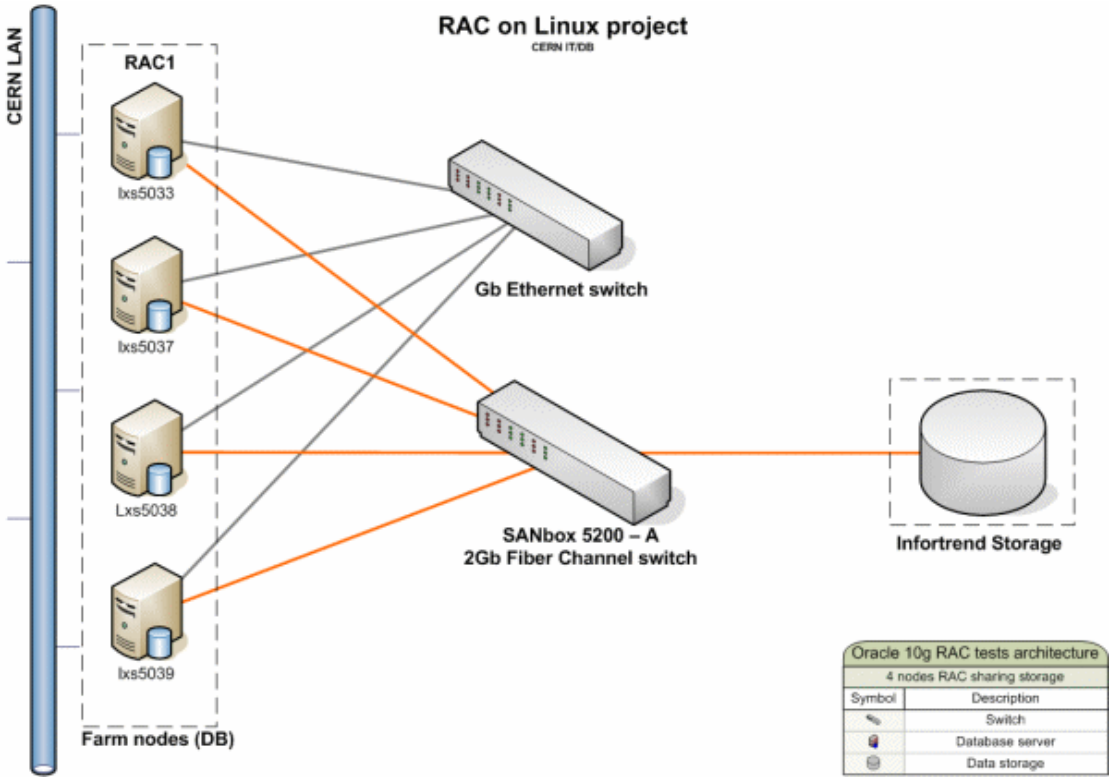


Figure 4.6: The schematic database test cluster set-up and its connection to SAN-based storage

Another component aiming to increase the database service availability is Oracle Data Guard, which complements the CPU redundancy of RAC. Data Guard allows keeping copies of the database data on disk to avoid unavailability as a result of disk media faults. Changes between a writable master database and read-only slave copies can be applied introducing a time lag between master and slave which may be used to recover to a previous database state, for

example, in case of human error. Oracle Data Guard has been deployed as part of the LCG RLS service and allowed to perform database upgrades transparent to service users.

4.6.1.2 Database Back-up Requirements

The back-up volume requirements on the Tier-0 are closely related to the database volume requirements and share their remaining uncertainties. The physics database service at CERN uses the Oracle RMAN to implement a redundant hierarchy of full and incremental back-ups without introducing the need for service downtime. The default retention policy always keeps two full database back-ups available. Back-ups are scheduled based on the update activity for a particular database. For active databases full back-ups are typically created every week and are accompanied by incremental back-ups on database change logs generated every 10 minutes. The back-up files (archive logs) are stored on tape via the Tivoli Storage service and recent files are kept in a disk pool to decrease the recovery latency.

To allow the database back-up infrastructure to scale to large database volumes at LHC it will be essential to mark completed table spaces as read-only as early as possible to avoid multiple transfers of unchanged data to the tape storage. This will require appropriate design on the application development and deployment sides in order to keep the active (writable) data physically well clustered.

Based on the experience with current physics applications we estimate the factor between database volume and the required back-up volume to be around 2.5.

A rough estimate of the total database volume and back-up volume can be obtained from the computing model and data volumes of pre-LHC experiments. The COMPASS experiment for example uses for some 400 TB of event data about 4 TB of database data (1%) and 10 TB (2.5%) of database back-up data). Assuming a similar split we would estimate based on 15 PB/year of event data (all 4 LHC experiments), some 150 TB/year of database data and a database back-up volume of 375 TB/year.

4.6.2 Database Services at Tier-1 and Higher

Building on database services at the CERN Tier-0 and other LCG sites, the 3D project (http://lcg3d.cern.ch) has been set up to propose an architecture for consistent distributed deployment of database services at LCG Tiers. The main goals of this infrastructure are:

- Provide location-independent database access for Grid user programs and Grid services
- Increased service availability and scalability for Grid application via distribution of application data and reduction of data access latencies
- Reduced service costs by sharing the service administration between several database teams in different time zones.

This service aims to handle common database requirements for site local or distributed database data in LCG. Given the wide-area distribution of the LCG resources, this cannot be achieved by a single distributed database with tight transactional coupling between the participating sites. The approach proposed is rather based on independent database services, which are only loosely coupled via asynchronous data replication or data copy mechanisms.

For several reasons, including avoidance of early vendor binding and adaptation at the available database services at the different tiers, a multi-vendor database infrastructure has been requested by the experiments. To allow to focus the limited existing database administration resources on only one main database vendor per site, it is proposed to deploy Oracle at Tier-0 and 1 and MySQL at Tier-2.

4.6.2.1 Requirement Summary

The 3D project has based its proposal on submitted database requirements from participating experiments[5] (ATLAS, CMS, LHCb) and software providers (ARDA, EGEE, LCG-GD). Experiments have typically submitted a list of 2–5 candidate database applications, which are planned for distributed deployment on LCG worker nodes. Many of these applications are still in the development phase and their volume and distribution requirements are expected to be finalized only after first deployment at the end of 2005. The volume requirements for the first year are rather modest compared with existing services and range from 50 GB to 500 GB at Tier-0/1. The foreseen distribution is compatible with a fan-out scheme originating from Tier-0. As data at Tier-1 and higher is considered to be read-only (at least initially) the complex deployment of multi-master replication can be avoided.

The distributed database infrastructure is in an early test phase and expected to move into first pre-production in autumn 2005. Based on experiment requirements and available experience and manpower at the different tiers, it is proposed to structure the deployment into two different levels of service, as shown in Figure 4.7.

1. Consistent database backbone (at Tier-0 and Tier-1)
   - Read/write access for Tier-0, read access for Tier-1 (initially)
   - Reliable database service including media recovery and back-up services based on a homogenous Oracle environment
   - Consistent asynchronous replication of database data is provided as an option (some application may decide on application-specific replication mechanism and deployment infrastructure)
2. Local database cache (at Tier-2 and higher)
   - Read-only database access
   - All data can be obtained from Tier-0/1 in case of data loss
   - Low latency access to read-only database data either through live database copies or data caches
   - Local write access for temporary data will be provided but should not be relied on for critical data.

[5] The ALICE experiment has been contacted, but at project start did not plan deployment of databases for their applications outside of Tier-0. ALICE requirements have therefore only been taken into account for the calculation of Tier-0 requirements.

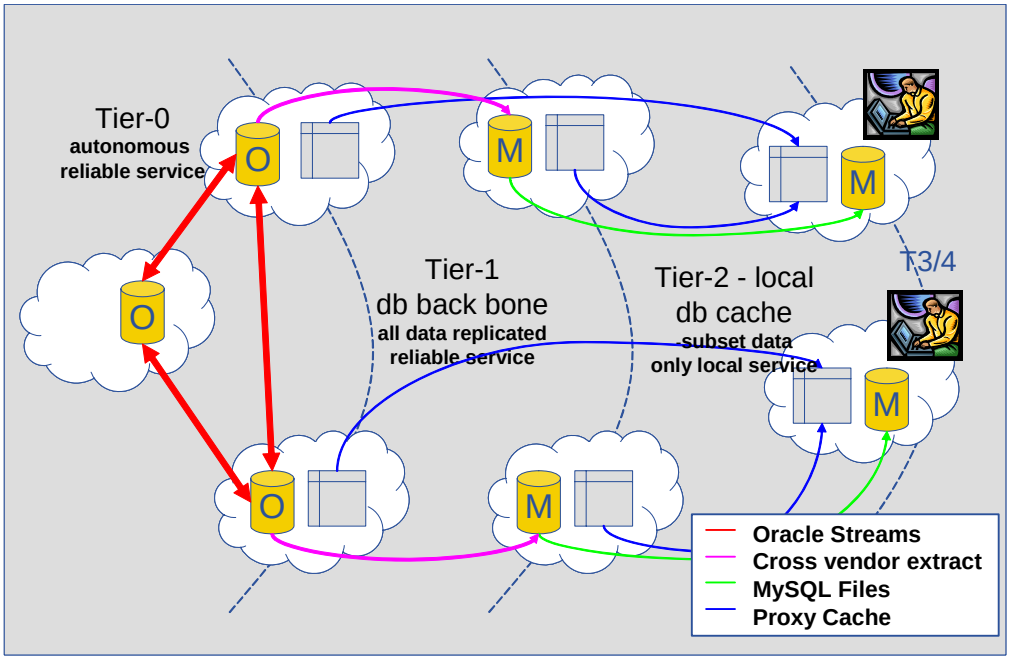


Figure 4.7: Proposed service architecture and service level at LCG Tiers

4.6.2.2 Database Service Requests and Available LCG Database Resources

The 3D project has collected initial database requirements for a first production service in autumn 2005 for online, Tier-0, Tier-1 and higher tiers. The evolution of these requirements will be reviewed regularly; in particular after some experience has been obtained with the new service and its applications in a first production phase has been obtained. The initial data volume requirements for 2005 should not create unexpected demands for the participating Tier-1 sites. The requested server capacity (CPU, I/O bandwidth and memory requirements) though is difficult to predict and can only be obtained from a validation test in the Tier-0 and 3D test beds using realistic workloads. A test plan for scheduling application validation tests has been proposed. Several of the experiment applications and reference workloads are still under development. Also several of the LCG Grid services will deploy new software implementations, which are exposed to realistic workloads, for the first time during the service challenges in 2005. The late availability of the software components together with the uncertainties of their access patterns will likely result in service resource mismatch and contention on validation hardware and associated support during first deployment.

One area that is still open is what commitment in terms of server hardware and database administration services needs to be expected from the Tier-1 and Tier-2 sites. Currently several of the participating database Tier-1 sites are tightly associated with individual experiments. The proposed model is therefore that Tier-1 sites foresee their hardware acquisitions and service staffing based only on the requests of their associated experiment(s). Database support for baseline Grid services (FTS, VOMS, local file catalogs) is required at all Tier-1 sites and needs to be taken into account. As Grid jobs running at Tier-2 sites will, according to the proposed architecture, access the database data from their closest Tier-1 or even Tier-0 services, this additional service load needs also to be taken into account.

4.6.3 Integration with Application Software

A reference integration between the distributed database services and LCG application software will be provided as part of the LCG RAL relational abstraction layer (RAL). This includes the use of a logical service look-up, a common infrastructure to gather client-side monitoring information (detailed timing of the main queries issued by a complex user application etc.) support certificate based for Oracle and MySQL and consistent connection

retry in case of network problems. Applications that do not use POOL/RAL will need to be adapted by their developers to use these features.

4.6.3.1 Database Service Look-Up

In order to achieve location independent access to local database services a database location service is proposed, similar to the existing (file) replica location service (RLS). This service would map logical database names into a physical database connection strings and avoid the hard coding of this information into user applications. As this service is similar to the file cataloguing service it could re-use the same service implementation and administration tools. A prototype catalogue is being integrated into POOL/RAL, which allows using any POOL supported file catalogue. In contrast to the LCG file replica catalogue, which refers to physically identical copies of a given file, we propose to allow for an abstraction from the concrete database vendor through this service; i.e. an application would detect the database back-end (Oracle, MySQL, etc.) at runtime and load the appropriate database-specific connection module. This would simplify the database deployment so that at Tier-2 database vendor heterogeneity could be allowed.

4.6.3.2 Database Authentication and Authorization

To provide secure access to the database service and keep at the same time the database user administration scalable, we propose to integrate database authentication with LCG certificates and to authorize database users based on role definition from the Virtual Organization Membership Service (VOMS). This will provide for a consistent Grid identity for file and database data and a single VO role administration system, which also controls the Grid user rights for database access and data modification.

Oracle provides for this purpose an external authentication mechanism between database and a LDAP-based authentication server. This protocol can also be used to determine which database roles a particular user may obtain based on his credentials, e.g., a X.509 certificate. The authentication can be done either by connecting directly from the client application to the database server or it can be mediated via a proxy server at the boundary of a local sub-network containing the database server. The latter approach is expected to provide more flexibility, for example, for the integration of LCG specific certificate validation procedure and to allow the running of a set of database servers behind a firewall without exposing their service ports directly. Also MySQL provides X.509-certificate-based authentication methods which have been used, for example, by ATLAS. For both database vendors, complete end-to-end integration of authentication and authorization still needs to be proven and the performance impact of secure (SSL based) network connections for bulk data transfers needs to be evaluated.

4.6.3.3 Database Network Connectivity

One implication of the proposed database service architecture is that a single Grid program may need to access both databases at Tier-2 (for reading) and at Tier-1 or Tier-0 for writing. This implies that appropriate connectivity for service TCP ports of database servers at Tier-1 and Tier-0 should be provided to worker nodes at Tier-2. In addition the database servers at Tier-0 and Tier-1 need to be able to connect to each other in order to allow the database replication between servers to function. This will require some firewall configuration at all tiers but as the number of individual firewall holes in this structure is small and contains only well-defined point-to-point connections, this is currently not seen as a major security risk or deployment problem.

4.6.3.4 Service Responsibilities and Co-Ordination

To define the split of service responsibilities between the database teams at the different sites we propose to differentiate between ‘local services’ which are performed as required by a local database team and ‘shared services’ which can be provided on a rotational shift basis by one of the teams for the LCG database infrastructure.

Local services include server operating system and database software installation, application of patches and security upgrades, support for database back-up and recovery and larger-scale data migration between database servers at one site.

Shared services include common administration tasks such as the routine maintenance of database accounts (quota and role management), monitoring and tuning of the server status, monitoring of application resource consumption, and identification of resource-consuming database sessions and basic storage management.

To perform the necessary administration tasks and to obtain an overview of the performance parameters of the distributed system we evaluate on the Tier-0 and Tier-1 level the Oracle Enterprise Manager tool, a Web-based administration and diagnostic tool, which is used at several LCG sites. This tool has been installed in the 3D test bed. For the information exchange between the participating database teams we propose to use a wiki-site [59] to log the daily service interventions accompanied by regular operations meetings (phone/VRVS) to plan deployment changes with the experiment and site representatives.

4.7 Initial Software Choices at CERN

4.7.1 Batch Systems

For about five years CERN has been using very successfully the LSF Batch scheduler from Platform Computing in the CERN computing farm. This has evolved considerably during the years and copes with the current work-load without any bottleneck. The system is used by more than 100 groups running up to 3,000concurrently executing user jobs. There can be more than 50,000 jobs in the queues. The support relationship with Platform Computing is very good and feedback from CERN experts is taken into account. There is currently no reason for a change.

4.7.2 Mass Storage System

The mass storage system has two major components: a disk space management system and a tape storage system. We have developed the CASTOR Mass Storage System at CERN and by the middle of 2005 the system contained about 35 million files with an associated 4 PB of disk space. The system uses files and file systems as the basic operation unit.

The new improved and re-written CASTOR software is in its final phase and will be deployed during the second half of 2005.

The new CASTOR system implements a completely different architecture with the vision of a Storage Resource Sharing Facility. The system has a database centric architecture with stateless components where the locking is provided through the DB system. All requests are scheduled to achieve predictable loads of the different components (disk servers, tape servers, DB). The scheduler is implemented as a pluggable module and already two different systems (LSF and Maui) have been used. Also the other components (data transfer modules, policies, garbage collectors, etc.) are implemented as pluggable modules. Orders of magnitude better scalability and a much better redundancy and error-recovery capability have already been shown in several tests and the ALICE Data Challenge VI.

The problem of large numbers of small files in the system can only partly be addressed by the new CASTOR implementation, as the major obstacles are not CASTOR-specific but rather arise from limitations in the tape technology

The CASTOR MSS software is the CERN choice for the foreseeable future.

4.7.3 Management System

The Extremely Large Fabric management system (ELFms) [60] was developed at CERN based on software from the EU DataGrid project. It contains three components:

1. **quattor** [61**]**, a system administration toolkit provides a powerful, portable and modular suite for the automated installation, configuration and management of clusters and farms running Linux or Solaris.

2. The LHC Era Monitoring (**LEMON**) [62] monitoring system is server/client based. On every monitored node, a monitoring agent launches and communicates using a push/pull protocol with sensors which are responsible for retrieving monitoring information. The extracted samples are stored on a local cache and forwarded to a central Measurement Repository.

3. The LHC-Era Automated Fabric (**LEAF**) [63] toolset, consist of a State Management System (SMS), which enables high-level commands to be issued to sets of quattor-managed nodes, and a Hardware Management System (HMS), which manages and tracks hardware workflows in the CERN Computer Centre and allows equipment to be visualized and easily located.

The system is now dealing with more than 2500 nodes in the centre with varying functionality (disk, CPU, tape, service nodes) and multiple operating systems. It has been in full production for a year and provides a consistent full-life-cycle management and high automation level. This is the CERN choice for the foreseeable future.

4.7.4 File System

The Andrew File System (AFS) is an integral part of the user environment of CERN.

It serves as

- repository for personal files and programs,
- repository for the experiment software,
- repository for some calibration data,
- repository for some analysis data,
- as well as common shared environment for applications.

AFS provides worldwide accessibility for about 14,000 registered users. The system has a constant growth rate of more than 20% per year. The current installation (end 2004) hosts 113 million files on 27 servers with 12 TB of space. The data access rate is ~ 40 MB/s during daytime and has ~ 660 million block transfers per day with a total availability of 99.8 %.

During 2004 (and ongoing) an evaluation of several new file systems took place to judge whether they could replace AFS or even provide additional functionality in the area of data analysis.

Missing redundancy/error recovery and weaker security were the main problems in the investigated candidates. So far the conclusion is that the required functionality and performance for the next ~3 years can only be provided by keeping the AFS file system.

4.7.5 Operating System

All computing components (CPU, disk, tape and service nodes) are using the Linux operating system. For a couple of years the CERN version has been based on the RedHat Linux Distribution. RedHat changed their licensing policies in 2003 and they have been selling since then their different Linux RH Enterprise versions on a profitable basis. After long negotiations in 2003/2004 CERN decided to follow a four-way strategy:

- collaboration with FNAL on Scientific Linux, a HEP Linux distribution based on the re-compiled RH Enterprise source code, which RH has to provide without charge due to the GPL obligations.
- buying RH Enterprise licences for the Oracle on Linux service
- having a support contract with RH
- pursuing further negotiations with RH about possible HEP-wide agreements.

An investigation about alternative Linux distributions came to the conclusion that there was no advantage in using SUSE, Debian or others. SUSE, for example, is still available free of charge. However, the rather different implementation would require significant effort in adapting our management tools. In addition there are question marks about the community support.

CERN will continue with the described Linux strategy for the next couple of years. The situation will be kept under continuous review.

4.8 Initial Hardware Choices at CERN

4.8.1 CPU Server

CERN has been purchasing 'white boxes' from assemblers in Europe for more than five years now.

CERN has exclusively used INTEL processors from their IA32 production line in dual processor nodes. The 2005/2006 issues are the integration of the 64bit processor architecture and the advent of multicore processors.

The road to 64bit is easier now that INTEL is also providing an intermediate processor family (EM64T, Nocona) which can run 32bit and 64bit code. AMD has been doing this for more than a year with the Opteron processor line.

The AMD processors currently have a price/performance advantage, but this is of course varying over time (competition between INTEL and AMD) and one also has to consider that the processors are only about 30% of the costs for a CPU node. The major problem is the introduction of a second architecture and the corresponding extra costs, e.g., more systems administration effort and extra software build and interactive login facilities. The investigation about these consequences is ongoing. Several sites are getting now experience with larger scale Opteron installations and detailed comparisons are on the way, e.g., code stability between platforms, compiler effects, performance benchmarks, etc. More details about expected cost developments and the influence of multi-core processors on some fabric issues can be found in Ref. [64].

The stability of the CPU nodes is high enough to achieve efficiencies of 99% in terms of availability of the required CPU resources, thus the frequency and repair time of broken hardware leads to an average 'loss' of hardware resources of about 1%. The amount of software (operating system, application, etc.) errors is a factor of ten higher, but as the time to 'fix' problems is much shorter, it leads to resource 'losses' of the order of only 0.3%. The stability of the hardware is therefore good enough to continue the strategy of buying white boxes from assembler companies.

CERN will continue with the current strategy of buying white boxes from assembler companies and is considering including AMD in the new purchases if the price/performance advantages become significant.

4.8.2 Disk Storage

CERN is using the NAS disk server model with 400 TB of disk space currently installed. There are in addition some R&D activities and evaluations ongoing for a variety of alternative

solutions like iSCSI servers (in connection with file systems), Fibre Channel attached SATA disk arrays, large multiprocessor systems, USB/firewire disk systems.

Besides the performance of the nodes it is important to understand the reliability of the systems. In this section the current state of the failure rate of disk servers and components is described and the effect of this on the service.

Monitoring of failures at CERN shows that the MTBF (Mean Time Between Failures) of the ATA disks is in the range of 150,000 hours. This means there is one genuine disk failure per day with the currently installed 6,000 disks. The vendors quote figures in the range of 300,000 to 500,000 hours, but these apply to usage patterns common on home PCs, while our disks run continuously. We have observed similar MTBF figures for SCSI and Fibre Channel disks.

Disk errors are 'protected' by using mirrored (RAID1) or RAID5 configurations. In case of a disk replacement, the performance is degraded by rebuilding the RAID system. To cope with these negative effects one has to rely on the good redundancy and error recovery of the Mass Storage System CASTOR and also on similar efforts in the applications themselves.

With our current failure rate of disks and servers we have an average 'loss' of resources in the 1–2% range: i.e., we have bought 100% of the required resources, but on average only 98–99% are available to the users.

Simple NAS servers still deliver the best price/performance ratio and have an acceptable error rate. We will continue with this strategy but make some effort to work with the vendors on the improvement on the quality of hardware.

4.8.3 Tape Storage

An STK installation of 10 robotic silos with 5,000cartridges each and 50 tape drives from STK (9940B) has been in production at CERN since several years. The technology plans in industry are described in Section 4.5.3. It is planned to move to the next generation in the middle of 2006.

Today the technology of choice for 2006 and onwards cannot be predicted. Currently, the three technologies, IBM, STK, and LTO, are valid candidates. The real issue is to minimize the risk and total cost; this has several ingredients:

- Cost of the drives. This is linked to the expected efficiencies which we are currently evaluating (depends on computing models).
- Cost of silos and robots. These are highly special installations and the prices depend heavily on the package and negotiations.
- Cost of the cartridges.

Each of these items is about 1/3 of the costs over 4 years, but with large error margins, and support costs for the hardware and software need to be included.

4.8.4 Network

The current CERN network is based on standard Ethernet technology, where 24-port fast Ethernet switches and 12-port gigabit Ethernet switches are connected to multigigabit port backbone routers (3Com and Enterasys). The new network system needed for 2008 will improve the two involved layers by a factor 10 and the implementation of this will start in the middle of 2005. This will include a high-end backbone router mesh for redundancy and performance, based on 24 or more 10-gigabit ports, and a distribution layer based on multiport gigabit switches with one or two 10-gigabit uplinks. Later the performance in latency and throughput can be further improved by using Infiniband products as described in Section 4.5.4.2 which have the possibility to add conversion modules to Fibre Channel; later this year this will also be the case for 10-gigabit Ethernet modules.

4.9 Hardware Life-Cycle

The strategy for the hardware life-cycle of the different equipment types (CPU server, disk server, tape drives, network switches) is rather straightforward. The equipment is usually purchased with a three year warranty, during which the vendors provide for the repair of equipment. At CERN the equipment is not replaced systematically at the end of the warranty period, but is left in the production environment until:

- the failure rate increases,
- there are physical limitations, e.g., the PCs cannot run jobs anymore, because of too little memory or disk space or too slow CPU speed,
- the effort to handle this equipment becomes excessive.

This 'relaxed' replacement model has been successful so far. These extra resources are not accounted for in the resource planning, because the availability cannot be guaranteed. The cost model assumes that standard PC equipment is replaced in its fourth year and disk systems, tape drives and network switches have a five-year useful life.

4.10 Costing

The planning exercise for the CERN fabric uses the following input parameters to calculate the full cost of the set-up during the years 2006-2010:

1. the base computing resource requirements from the experiments (CPU, disk and tape),
2. derived resources (tape access speed, networking, system administration) from the combination of the base resources and the computing models,
3. the reference points of the equipment costs ,
4. the cost evolution over time of the different resources

The detailed list of base resource requirements has already been given in Section 2.3 and in Section 3.2 discusses some of the derived resources are discussed.

The cost evolution uses a formula for each resource that assumes a smooth evolution over time. In the case of processors and disks this is close to Moore's Law, a reduction of a factor of two in 18 months. With this steep price/performance ratio, significant cost penalties can arise when certain purchasing procedures impose the fixing of the price several months before the actual acquisition.

More details about the cost calculations for CPU, disk and tapes can be found in [65].

4.11 Networking

In the context of this section the following terms are defined to have the meaning as stated.

LHC network traffic: The data traffic that flows between Tier-0, the Tier-1s, and the Tier-2s.

Light path: (i) a point-to-point circuit based on WDM technology, or (ii) a circuit-switched channel between two end-points with deterministic behaviour based on TDM technology, or (iii) concatenations of (i) and (ii).

This section describes the high-level architecture for the LHC network. It is structured in such a way as to emphasize the 'end-to-end' path between communicating systems as this typically crosses a number of network infrastructures and network management domains.

With this in mind, the following aspects are considered:

- Provision of basic network bandwidth

- Technical implementation of IP connectivity

The end-to-end path may contain elements from the following network infrastructures depending on the specific systems involved:

1. The CERN campus network
2. The CERN computer centre network infrastructure
3. The connectivity to a remote Tier-1 or Tier-2 centre
    - Tier-0–Tier-1
    - Tier-1–Tier-2
4. The Tier-1 or Tier-2 campus infrastructure

4.11.1 The CERN Campus Network

Not much detail will be included in the TDR concerning the CERN campus infrastructure as this is considered to be part of the overall CERN infrastructure and not specific to the LCG TDR. However, it is worth noting that as of the time of writing, a project has been proposed to upgrade the core and starpoint infrastructure that will ensure basic desktop connectivity in the 100 to 1,000 Mb/s range as opposed to the 10–100 Mb/s available today.

As the desktop networking is a starred infrastructure the actual throughput obtained may vary considerably but the infrastructure is designed to provide a 'fair share' of the available bandwidth from the desktop to the campus backbone at CERN.

4.11.2 The CERN Computer Centre Network

This is considered in Section and is not repeated here. It involves the core network architecture that provides connectivity between the CERN Campus, the wide-area networks and the experimental areas.

4.11.3 Remote Connectivity Tier-0-Tier-1

With respect to Tier-0-Tier-1 networking this document proposes a detailed architecture based on permanent 10-gigabit light paths. These permanent light paths form an Optical Private Network (OPN) for the LHC Grid.

An overall picture of the relationship between the Tier-0, Tier-1 and Tier-2 networking is shown in Figure 4.8.

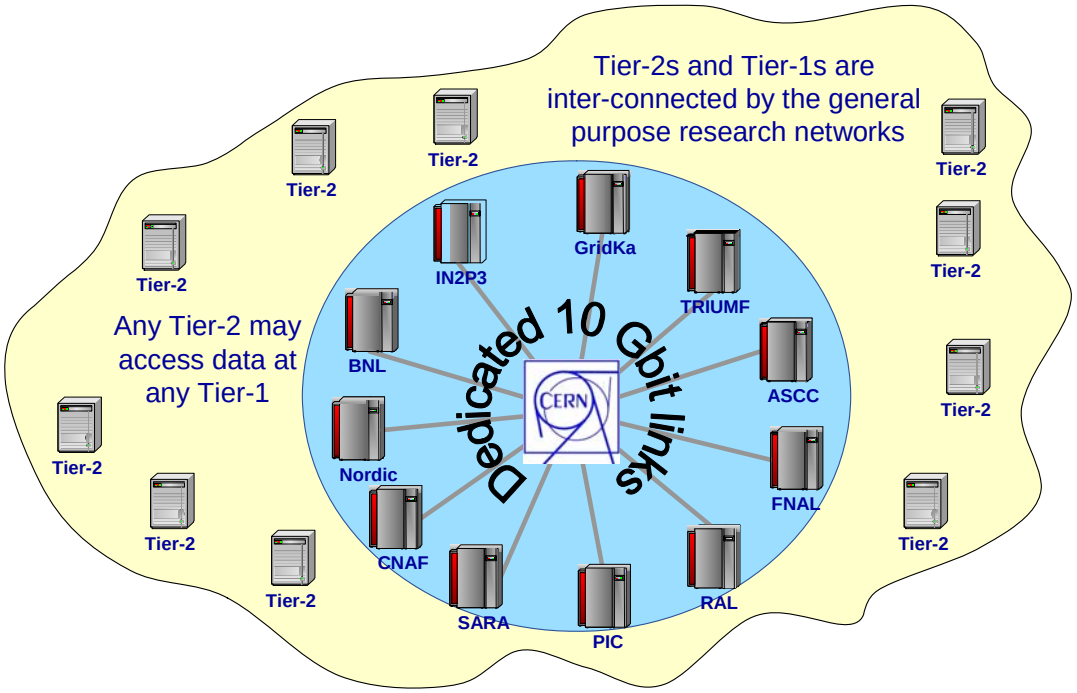


Figure 4.8: Tier-0–Tier-1–Tier-2 interconnectivity

The resources available at the Tier-1s will not all be the same and therefore the average network load is expected to vary. In addition, the anticipated peak load is an important factor as it is this peak load that the network should be capable of sustaining. As the computing models continue to be refined this is becoming clearer. For the moment the agreed starting point is the provisioning of at least one 10 Gb/s transmission path between each Tier-1 and Tier-0.

The path from CERN to a particular Tier-1 may take on a number of variants:

- CERN to GÉANT to a remote National Research Network (NREN) to a Tier-1
- CERN to a remote NREN
- CERN to a Tier-1.

The 12 envisaged Tier-1s and their associated NRENs are given in Table 4.3.

Table 4.3: Tier-1 centres and their associated NRENs

Tier-1 name	***Tier-1 location***	***NRENs involved***
ASCC	Taipei, Taiwan	ASnet
BNL	Upton, NY, USA	LHCnet/ESnet
CERN	Geneva, Switzerland	
CNAF	Bologna, Italy	GARR
FNAL	Batavia, IL, USA	LHCnet/ESnet
IN2P3	Lyon, France	RENATER
GridKa	Karlsruhe, Germany	DFN
SARA/NIKHEF	Amsterdam, The Netherlands	SURFnet6
NDGF	Nordic countries	NORDUnet
PIC	Barcelona, Spain	RedIRIS
RAL	Didcot, United Kingdom	UKERNA
TRIUMF	Vancouver, Canada	CANARIE

4.11.3.1 Network Provisioning

The responsibility for providing network equipment, physical connectivity and manpower is distributed among the cooperating parties.

CERN will provide the interfaces to be connected to each Tier-1s link termination point at CERN. Furthermore, CERN is available to host equipment belonging to a Tier-1 used for Tier-0–Tier-1 link termination at CERN, if requested and within reasonable limits. If this is the case, Tier-1 will provide CERN with the description of the physical dimensions and the power requirements of the equipment to be hosted.

The planned starting date for the production traffic is June 2007, but Tier-1s are encouraged to proceed with the provisioning well before that date, and in many cases this will be achieved in 2005. In any case the links must be ready at full bandwidth not later than Q1 2006. This is important as the Service Challenges now under way need to build up towards the full capacity production environment exercising each element of the system from the network to the applications. It is essential that the full network infrastructure be in place, in time for testing the complete environment.

Each Tier-1 will be responsible for organizing the physical connectivity from the Tier-1's premises to Tier-0, according to the MoU [1] LHC Homepage, http://cern.ch/lhc-new-homepage/

- [2] between the Tier-0 and the Tier-1s.
- Each Tier-1 will make available in due course the network equipment necessary for the termination point of the corresponding Tier-1–Tier-0 transmission path at the Tier-1 side.
- Tier-1s are encouraged to provision direct Tier-1-Tier-1 connectivity whenever possible and appropriate.
- Tier-1s are encouraged to provision back-up Tier-0–Tier-1 links on alternate physical routes with adequate capacity.

4.11.3.2 Planning

The CERN connectivity planning is currently:

- Two dedicated 10 Gbit circuits to New York (ManLan) with transit to Chicago (Starlight) by Sept 1 2005. This will provide capability to connect from New York to Brookhaven at 10 Gb/s. Connectivity from Starlight to FNAL is already in place. These links are provided by the US Department of Energy and CERN consortium (USLiC) under the name 'LHCNet'.
- A dedicated 10 Gbit circuit to New York via the Netherlight optical switch in Amsterdam. This will provide a peering connection with CANARIE and transit to TRIUMF. This link is expected to be in place in summer 2005 and is provided by CANARIE from New York to Amsterdam with extension to CERN provided by SURFnet.
- Two 2.5 Gbit circuits to Taipei via Netherlight. This connectivity is in place and is not expected to be upgraded in 2005. These links are provided by the Academica Sinica Taipeito Amsterdam with extension to CERN provided by SURFnet.
- A 10 Gb circuit to the Computing Centre of IN2P3. This link is a single wavelength on a dark fiber provided and lit by RENATER. This link is in place and is expected to be operational by Summer 2005.
- Six 10 Gbit circuits to GÉANT-2 for connections to the national research networks of the remaining Tier-1 centres. The expected dates by which the national research networks will be able to provide onward transit to the respective Tier-1s is given in

the following table. GÉANT-2 is expected to start implementation in June 2005 and to complete by December 2005. These links will be provided by the GÉANT-2 project partners, and CERN, as part of their cost-sharing model.

The current connectivity from Netherlight to CERN provided by SURFnet is 2 × 10 Gb/s and will be replaced by transit circuits from Netherlight to CERN provided by GÉANT-2 in the future. The current connectivity is over subscribed as a number of 1Gb/s circuits are used for various testing purposes. However, the plan will be to provide at least 2 × 10 Gb/s connections via GÉANT-2 to Netherlight for connections to Taipei and TRIUMF by early 2006.

The Planning for the NREN transit circuits from the GÉANT point of presence to the Tier-1s is summarized in Table 4.4.

Table 4.4: Planning for the NREN transit circuits from the GÉANT point of presence to the Tier-1

Tier-1	***NRENs involved***	***Date for 10Gb/s circuit***
CNAF	GARR	H2 2005
IN2P3	RENATER	H1 2005
GridKa	DFN	H2 2005
SARA/NIKHEF	SURFnet6	Now
Nordic Data Grid Facility	NORDUnet	H1 2006
PIC	RedIRIS	Not yet known
RAL	UKERNA	2006

4.11.3.3 Technical Implementation

The proposed networking strategy is to use at least one dedicated 10 Gb/s connection between Tier-0 and each Tier-1. This provides basic connectivity but does not yet constitute a network capable of exchanging IP traffic.

How this will be achieved is currently under study in a subgroup of the Grid Deployment Board (GDB) that has not yet concluded, but it seems likely that the recommendations will be along the following lines.

The networking equipment on both ends of a light path should be capable of speaking BGP4 (BGP – Border Gateway Protocol). An eBGP (extended Border Gateway Protocol) peering will be established between the equipment of the Tier-0 and the Tier-1 using the following parameters:

- Tier-0 will use the CERN Abstract Syntax Notation (ASN): AS513.
- For a Tier-1 site that has its own ASN, this ASN will be used in the peering.
- For a Tier-1 site that has no ASN, the ASN of the intermediate NREN will be used instead, in the case that the 10-gigabit light path terminates on equipment of the NREN.

The Tier-1 will announce its own prefixes and possibly any of the prefixes of Tier-1s and Tier-2s directly connected to it. From the architecture point of view, every Tier-0–Tier-1 link should handle only production LHC data. This can be accomplished by making the appropriate BGP announcements.

On the Tier-0's networking equipment, for connecting the Tier-1s' access links, 10GE LAN PHY ports will be available. Ports of flavour 10GE WAN PHY or STM-64/OC-192 can be negotiated between CERN and an individual Tier-1 on request.

4.11.3.4 Security Considerations

It is important to address security concerns during the design phase. The fundamental remark for the security set-up proposed below is that because of the expected network traffic data rates across 10 Gb/s links, it is not possible to interpose firewalls without considerable expense.

It is also assumed that the overall number of systems exchanging LHC traffic is relatively low given that these links are for high-speed bulk data transfer. These links then do NOT provide a general interconnection between all Tier-0 and Tier-1 systems resident on their respective internal networks.

While Access Control List (ACL)-based network security is not sufficient to guarantee enough protection for the end-user applications, it can considerably reduce the risks involved with unrestricted internet connectivity at relatively low cost.

The architecture will be kept as protected as possible from external access, while, at least in the beginning, access from trusted sources (i.e. LHC prefixes) will not be restricted.

Incoming traffic from Tier-1s will be filtered using ACLs on the Tier-0's interfaces connected to the Tier-1s. Only packets with LHC prefixes in the source-destination pair will be allowed. The default behaviour will be to discard packets.

Tier-1s are encouraged to apply equivalent ACLs on their side. Otherwise outgoing filters at the Tier-0's level can be considered.

At least initially, filtering will be at the IP level (permit IP or deny IP). Later restrictions to allow only some specific ports may be considered, in cooperation with the application managers.

4.11.3.5 Operations

It is clear that all entities contributing to the LHC OPN have a level of responsibility in ensuring the smooth operation of the network. Fault detection, diagnosis, resolution and reporting are all complex functions that require disciplined co-ordination and good communication channels among the parties involved. Similarly, day-to-day configuration of the infrastructure to add new locations or functionality also requires co-ordination. This chapter proposes a *Keep It Simple* approach by introducing the LHC-OPN Helpdesk.

The network elements (routers, switches) of the LHC-OPN are procured, owned and managed by the respective Tier-1 and Tier-0 centres. The LHC-OPN helpdesk will have read-only access to the network elements at Tier-0 and Tier-1 sites and will through this access proactively monitor the status of the infrastructure. It will provide a single point of contact for the users at Tier-0 and the Tier-1s of the LHC-OPN for fault reporting and correction. The helpdesk will liaise with all parties contributing to the infrastructure, i.e. Tier-0, Tier-1s, GÉANT-2, and the NRENs, in order to diagnose faults and to ensure they are resolved. The helpdesk will not resolve configuration and equipment faults, but will rely on the intervention of the appropriate partner in the overall infrastructure. The helpdesk will issue periodic reports to the LHC user community on the resolution of the faults and will provide periodic usage data.

4.11.4 Remote Connectivity Tier-1–Tier-1 and Tier-1–Tier-2

As Figure 4.9 illustrates, the Tier-1 centres are connected to dedicated links to ensure high reliability, high-bandwidth data exchange with the Tier-0 but are also connected to what is described as the ‘General-Purpose Research Networks’. In a very real sense this is a world-wide ‘Research Internet’ providing IP communication between systems.

This is required to ensure that there is good connectivity between Tier-2s and Tier-1s as well as Tier-1 to Tier-1 communication. Of course, the OPN could be used to provide paths for

Tier-1–Tier-1 data transfer and this is a capability that will be studied if such a possibility would ensure effective use of the dedicated links.

As with the general-purpose Internet, the 'Research Internet' is in fact a set of interconnected networks that link together the national and international networking initiatives through bi-lateral agreements. A few simple examples of this are GÉANT that links together the NREN networks of the European countries as well as providing some interconnects with other networks, and ESNet the Energy Sciences Network of the USA that provides a US backbone linking metropolitan area networks in the USA.

In an effort to understand the initiatives that are taking place worldwide that can be potentially used as connectivity for research purposes, the Global Lambda Integrated Facility was created. This brings together the major players funding and installing connectivity for research purposes.

Figure 4.9: Map of the GLIF infrastructure

The importance of this is to understand that the general–purpose connectivity between Tier-2s and Tier-1s will be comprised of a complex set of research initiatives world-wide that, as with the general Internet, will provide global connectivity permitting Tier-2–Tier-2 and Tier-2–Tier-1 communications to take place.

It will certainly be the case that if the bandwidth costs continue to drop as research network initiatives continue to acquire dark fibre there will be an increasing number of high-speed (10 Gb/s or more) direct links between Tier-2s and Tier-1s in the near future.

4.11.5 The Tier-1 and Tier-2 Campus Infrastructure

It is assumed that the Tier-1 and Tier-2 centres already have plans in place to implement the required infrastructures for connecting at sufficient bandwidth according to the experiments' models.

The service challenges under way will continue to exercise the total end-end infrastructure as the level of service challenge activity increases and the number of sites involved increases.

4.11.6 Future Plans

It is anticipated that the infrastructure and design being implemented will be adequate for LHC start-up according to the experiments' computing models but that this is assumed to be only the beginning. As an increasing volume of data is created there will always be the requirement to transfer these data between sites 'as fast as possible'.

Table 4.5 (provided by H. Newman) indicates the expected capability based on what we understand today of cost and technology considerations.

Year	Production	Experimental	Remarks
2001	0.155	0.622 – 2.5	SONET/SDH
2002	0.622	2.5	SONET/SDH; DWDM; GigE Integration
2003	2.5	10	DWDM; 1 + 10 GigE Integration
2005	10	2-4 × 10	λ switch, λ provisioning
2007	2–4 × 10	~10 × 10; 40	1st gen. λ grids
2009	~10 × 10 or 1–2 × 40	~5 × 40 or ~20–50 × 10	40, λ switching
2011	~5 × 40 or ~20 × 10	~25 × 40 or ~100 × 10	2nd gen. λ grids, terabit networks
2013	~terabit	~multi-terabit	~Fill one fibre

Table 4.5: Bandwidth roadmap (in Gb/s) for major HENP network links

As of the time of writing, this table has shown to be quite accurate given that in 2005 we are implementing switched lambda circuits for the Tier-0 to Tier-1 connectivity.

4.12 Security

There are many important challenges to be addressed in the area of computer and network security for LCG. Today's public networks are becoming an increasingly hostile environment, where sites and systems connected to them are under constant attack. Individual sites have gained extensive experience at coping with this enormous problem via the use of many different aspects of a co-ordinated approach to security. The components of the site security approach include firewalls, security monitoring and auditing, intrusion detection, training of system administrators and users, and the speedy patching of systems and applications. The collaboration of a large number of independent sites into one Grid computing infrastructure potentially amplifies the security problems. Not only do Grids contain large computing and data storage resources connected by high-speed networks, these being very attractive to potential hackers, but the connectivity and ease of use of the Grid services means that a successful compromise of one site in the Grid now threatens the Grid infrastructure in general and all of the participating sites.

The Grid services used by LCG must be secure, not only in terms of design and implementation, but they also need to be deployed, operated and used securely. LCG must constantly strive to attain the most appropriate balance between the functionality of its services and applications and their security. The decisions taken in reaching this balance must protect the LCG resources from attack thereby ensuring their availability to meet the scientific aims of the project. The setting of priorities will be informed by an ongoing threat and risk analysis and appropriate management of these risks to mitigate their effects. Sufficient resources need to be available for various aspects of operational security, e.g., in security incident response and forensic analysis, to limit and contain the effect of attacks whenever they happen, as they surely will.

The LCG security model is based on that developed and used by EDG, EGEE and the first phase of LCG. Authentication is based on the Grid Security Infrastructure from Globus using a Public Key Infrastructure (PKI) based on X.509 certificates. An essential component of the PKI is the Certification Authority (CA), this being the trusted third-party that digitally signs the certificate to confirm the binding of the individual identity to the name and the public key. The CAs used by LCG are accredited by the three continental Grid Authentication Policy Management Authorities, namely the European, the Americas and the Asia-Pacific, under the auspices of the International Grid Federation. The PMAs define the minimum acceptable standards for the operation of these accredited CAs. Users, host and services apply for a certificate from one of the accredited CAs and this can then be used for single sign-on to the Grid and is accepted for the purposes of authentication by all resources.

Authorization to use LCG services and resources is managed via the use of VOMS, the Virtual Organization Membership Service, and local site authorization Grid services, such as LCAS and LCMAPS. The registered users of a VO are assigned roles and membership of groups within the VO by the VO manager. Access to LCG resources is controlled on the basis of the individual user's VOMS authorization attributes, including their roles and group membership.

Operation of the LCG infrastructure requires the participating institutes providing resources and the four LHC experiment VOs to define and agree robust security policies, procedures and guides enabling the building and maintenance of 'trust' between the various bodies involved. The user, VO and site responsibilities must be described together with a description of the implications and actions that will be taken if a user, a VO or a site administrator does not abide by the agreed policies and rules.

The production and maintenance of LCG security policies and procedures will continue to be the responsibility of the Joint (LCG/EGEE) Security Policy Group. The approval and adoption of the various policy documents will continue to be made by the LCG GDB or other appropriate senior management body on behalf of the whole project. The existing set of documents, approved for use in LCG in 2003, is currently under revision by the JSPG with the aim of having security policy and procedures which are general enough to be applicable to both LCG and EGEE but also compatible with those of other Grid projects such as OSG. This aim is helped by the active participation of representatives from OSG in JSPG and by the use of common text for policy and procedures wherever possible.

The operational aspects of Grid security are also important. It is essential to monitor Grid operations carefully to help identify potential hostile intrusions in a timely manner. Efficient and timely Incident Response procedures are also required. Appropriate audit log files need to be produced and stored to aid incident response. More details of Operational Security are given in Section 4.3.5, while details of the planned security service challenges are presented in Section 6.2.2. The Security Vulnerability analysis activity recently started in GridPP and EGEE is considered to be an important contribution to the identification and management of security vulnerabilities both in terms of Grid middleware and deployment problems.

Special attention needs to be paid to the security aspects of the Tier-0, the Tier-1s and their network connections to maintain these essential services during or after an incident so as to reduce the effect on LHC data taking.

5 COMMON APPLICATIONS

CERN and the HEP community have a long history of collaborative development of physics applications software. The unprecedented scale and distributed nature of computing and data management at the LHC require that software in many areas be extended or newly developed, and integrated and validated in the complex software environments of the experiments. The Applications Area of the LCG Project is therefore concerned with developing, deploying and maintaining that part of the physics applications software and associated supporting infrastructure software that is common among the LHC experiments.

This area is managed as a number of specific projects with well-defined policies for co-ordination between them and with the direct participation of the primary users of the software, the LHC experiments. It has been organized to focus on real experiment needs and special attention has been given to maintaining open information flow and decision-making. The experiments set requirements and monitor progress through participation in the bodies that manage the work programme. Success of the project is gauged by successful use, validation and deployment of deliverables in the software systems of the experiments. The Applications Area is responsible for building a project team among participants and collaborators; developing a work plan; designing and developing software that meets experiment requirements; assisting in integrating the software within the experiments; and providing support and maintenance.

The project started at the beginning of 2002 and recently completed the first phase in its programme of work. Detailed information on all Applications Area activities can be found on the project website [66]. The scope and highlights of Phase 1 activities may be summarized as follows:

- The establishment of the basic environment for software development, documentation, distribution and support. This includes the provision of software development tools, documentation tools, quality control and other tools integrated into a well-defined software process. The Savannah project portal and software service has become an accepted standard both inside and outside the project. A service to provide ~100 third-party software installations in the versions and platforms needed by LCG projects has also been developed.
- The development of general-purpose scientific libraries, C++ foundation libraries, and other standard libraries. A rather complete set of core functionality has already been made available in public releases by the SEAL and ROOT projects, and has been used successfully in both LCG and experiment codes. The SEAL and ROOT project teams have recently joined forces and are working on a combined programme of work with the aim of producing a single deliverable on a time scale of 1–2 years.
- The development of tools for storing, managing and accessing data handled by physics applications, including calibration data, metadata describing events, event data, and analysis objects. The objective of a quickly-developed hybrid system leveraging ROOT I/O and an RDBMS was fulfilled with the development of the POOL persistency framework. POOL was successfully used in large scale production in ATLAS, CMS and LHCb data challenges in which >400 TB of data were produced.
- The adaptation and validation of common frameworks and toolkits provided by projects of broader scope than LHC, such as PYTHIA, GEANT4 and FLUKA. GEANT4 is now firmly established as baseline simulation engine in successful ATLAS, CMS and LHCb production, following validation tests of physics processes and by proving to be extremely robust and stable.

The work of the Applications Area is conducted within projects. At the time of writing there are four active projects: Software Process and Infrastructure (SPI), core software common libraries and components (CORE), persistency framework (POOL), and simulation (SIMU).

We begin the detailed description of Applications Area activities by recalling the basic high-level requirements. Architectural considerations and domain decomposition are described in Section 5.2. All Applications Area software is developed and tested on a selected number of platforms and the considerations that led to the choice of these are described in Section 5.3. There then follows a description of the software components under development grouped by domain. Finally we give an overview, and links to more detailed information, on project organization, plans and schedule.

5.1 High-Level Requirements for LCG Applications Software

A basic set of high-level requirements were established at the start of Phase 1 of the project. Here we recall those that have guided development work so far.

It is evident that software environments and optimal technology choices evolve over time and therefore LCG software design must take account of the >10 year lifetime of the LHC. The LCG software itself must be able to evolve smoothly with it. This requirement implies others on language evolution, modularity of components, use of interfaces, maintainability and documentation. At any given time the LCG should provide a functional set of software with implementations based on products that are the current best choice.

The standard language for physics applications software in all four LHC experiments is C++. The language choice may change in the future, and some experiments support multilanguage environments today. LCG software should serve C++ environments well, and also support multilanguage environments and the evolution of language choices.

LCG software must operate seamlessly in a highly distributed environment, with distributed operation enabled and controlled by components employing Grid middleware. All LCG software must take account of distributed operation in its design and must use the agreed standard services for distributed operation when the software uses distributed services directly. While the software must operate seamlessly in a distributed environment, it must also be functional and easily usable in ‘disconnected’ local environments.

LCG software should be constructed in a modular way based on components, where a software component provides a specific function via a well-defined public interface. Components interact with other components through their interfaces. It should be possible to replace a component with a different implementation respecting the same interfaces without perturbing the rest of the system. The interaction of users and other software components with a given component is entirely through its public interface.

Already existing implementations which provide the required functionality for a given component should be evaluated and the best of them used if possible (re-use). Use of existing software should be consistent with the LCG architecture.

LCG software should be written in conformance to the language standard. Platform and operating system dependencies should be confined to low-level system utilities. A number of hardware/operating system/compiler combinations (platforms) will be supported for production and development work. These will be reviewed periodically to take account of market trends and usage by the wider community.

Although the Trigger and DAQ software applications are not part of the LCG scope, it is very likely that such applications will re-use some of the core LCG components. Therefore, the LCG software must be able to operate in a real-time environment and it must be designed and developed accordingly, e.g., incorporating online requirements for time budgets and memory leak intolerance.

5.2 Software Architecture

Applications Area software must conform in its architecture to a coherent overall architectural vision; must make consistent use of an identified set of core tools, libraries and services; must integrate and inter-operate well with other LCG software and experiment software. This vision was established in a high-level 'blueprint' for LCG software which provided the guidance needed for individual projects to ensure that these criteria are met [67].

LCG software is designed to be modular, with the unit of modularity being the software component. A component internally consists of a number of collaborating classes. Its public interface expresses how the component is seen and used externally. The granularity of the component breakdown should be driven by that granularity at which replacement of individual components (e.g., with a new implementation) is foreseen over time.

Components are grouped and classified according to the way the way in which they interact and cooperate to provide specific functionality. Each group corresponds to a domain of the overall architecture and the development of each domain is typically managed by a small group of 5-10 people. The principal software domains for LCG Applications Area software are illustrated schematically in Figure 5.1. Software support services (management, packaging, distribution etc.) are not shown in this figure.

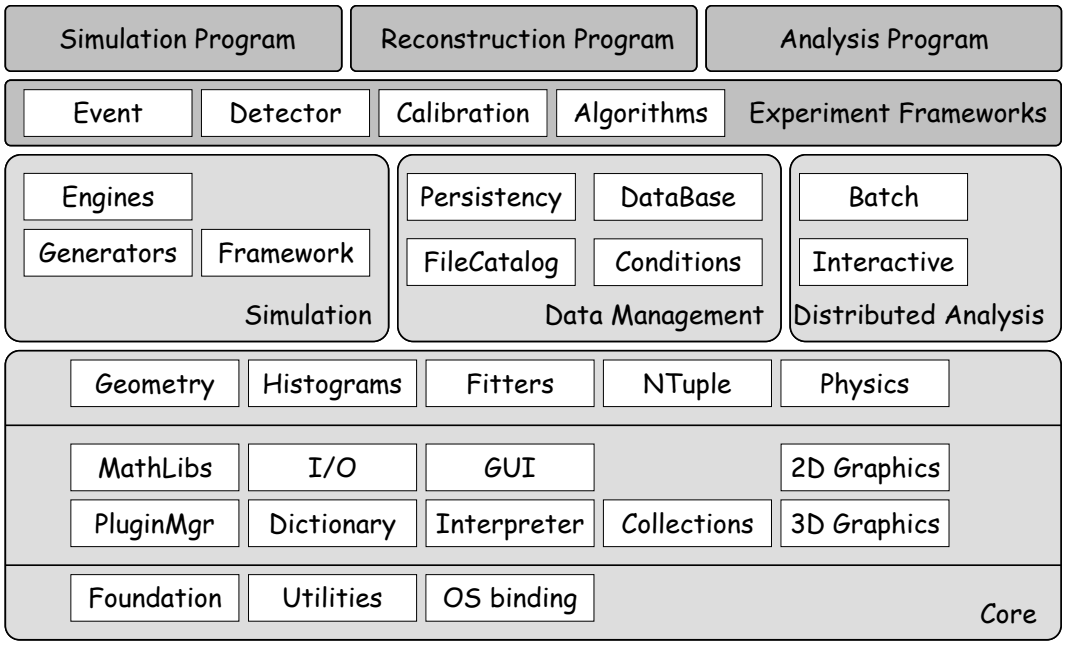


Figure 5.1: Physics applications software domain decomposition

The **Core Software Domain** provides basic functionality needed by any application. At the lowest level we identify the foundation libraries, utilities and services employed that are fairly independent class libraries (e.g., STL, or a library providing a Lorentz vector). Above this are core services supporting the development of higher-level framework components and specializations such as the plug-in manager and object dictionary by which all parts of the system have knowledge of, and access to, the objects of the system. Other Core software services include command line environments for interactive and batch (script) access to the functionality of the system, as well as general graphics and GUI tools that can be used to build experiment-specific interfaces but which are not themselves experiment-specific. Histogramming, ntuples, fitting, statistical analysis, and data presentation tools also contribute to Core functionality.

Above the Core software are a number of specialized frameworks that offer services specific to particular domains. The **Data Management Domain** covers object persistency, file cataloguing, event-specific data management, and detector-conditions-specific data

management. In general, the domain of expertise stays in the area of relational databases applications development. Support and LCG-directed development of simulation toolkits such as GEANT4 and FLUKA and ancillary services and infrastructure surrounding them are part of the **Simulation Domain**. Ancillary services surrounding event generators (e.g., standard event and particle data formats, persistency, configuration service), and support and distribution of event generator software, are also in the scope of common project activities. The **Distributed Analysis Domain** is the area where the physicist and physics application software interface to Grid middleware and services in order to support job configuration, submission and monitoring, distributed data management and Grid-enabled analysis. The scope of common activities in this area has still to be specified

Experiment applications are built on top of specialized frameworks which are specific to the experiment and not in LCG scope.

5.3 Operating System Platforms

The LHC experiments and the computer centres of universities and laboratories need to run LCG software on a large variety of platforms and operating systems, in several flavours and versions. Therefore, in order to guarantee portability, the software must be written following the most common standards in terms of programming languages and operating systems. Applications Area software is being routinely developed and run on a number of different compilers and operating systems, including Red Hat Linux, Microsoft Windows, and Apple Mac OSX, both with gcc and with their C++ proprietary compilers. This approach helps to ensure conformance to language standards and allows the project to manage dependencies on platform-specific features, both on 32-bit and 64-bit hardware architectures. Applications Area projects are involved in the certification and in the verification of new versions of compilers or operating systems at CERN.

The ‘production’ platforms currently supported are

- Red Hat 7.3 with gcc 3.2 and gcc 3.2.3 - the Linux reference platform for the LHC experiments and for the main computer centres. Red Hat 7.3 will be stopped by end 2005;
- Scientific Linux 3 with gcc 3.2.3, and in the near future also with gcc 3.4.3 - the new Linux reference platform for CERN and other large HEP laboratories. This is binary compatible with Red Hat Enterprise 3.

In addition, ‘development-only’ platforms are supported that have better development tools and are therefore used by many programmers and users to increase productivity and assure software quality:

- Microsoft Windows, with the Visual C++ 7.1 compiler and CygWin;
- Mac OSX 10.3 with gcc 3.3, and soon 10.4 probably with gcc 4.

Any changes to the list of supported platforms or compilers is discussed and approved at the Architects Forum, where all the LHC experiments are represented. When a new platform is a candidate to become supported, firstly all LCG software and external packages are re-compiled and re-built in order to assess the implications and changes needed for the new platform to become fully supported.

Platforms that will likely be supported in the near future are

- SLC3 Linux on AMD 64-bit processors as an additional production platform;
- gcc 3.4.3 compiler on all Linux platforms to take advantage of better performance.
- Mac OSX 10.4 as development platform, to resolve issues related to loading of dynamic libraries.

5.4 Core Software Libraries

The Core Software Project addresses the selection, integration, development and support of a number of foundation and utility class libraries that form the basis of typical HEP application codes. Its scope includes the development of dictionary and scripting services, facilities for statistical analysis and visualization of data, storage of complex C++ object data structures, and distributed analysis. In Phase 1 a number of implementations of core libraries were already made in public releases by the SEAL project. The ROOT analysis framework also contained a rather complete set of core functionality.

The SEAL and ROOT project teams have recently joined forces and are working on a combined programme of work with the aim of producing a single coherent set of deliverables on a timescale of 1–2 years. This initiative is a continuation of the work started in 2004 on convergence around a single dictionary and math library. By focusing efforts on a single set of software products we expect to project a more coherent view towards the LHC experiments and to ease considerably the long-term maintenance of the software. Another consequence has been that the ROOT activity has now become fully integrated in the LCG organization. The programme of work is being elaborated together with the LHC experiments in order to define priorities and to ensure user-level compatibility during the period of change.

5.4.1 Foundation Libraries

This provides a large variety of useful utility classes and operating system isolation classes that supply the low-level functionality required in any software application. Libraries are mainly in C++ and exist for basic functions (e.g., string manipulation), timers, networking, file system access, stream-oriented I/O, and for data compression and file archiving. As a consequence of the SEAL and ROOT project merge a number of features from SEAL in the area of plug-in management and support for software components will be added to the ROOT plug-in and component managers. The new features will be introduced in such a way as to be backward compatible for current ROOT users and to cause minimal changes for the SEAL users. Tasks involve maintenance and support of the available classes, and adding new functionality when the need arises.

5.4.2 C++ Dictionary and Reflection System

The ability of a programming language to introspect, interact and modify its own data structures at run time is called reflection. This functionality is required in two fundamental areas of the software:

- Object persistency: The meta description of objects allows the persistency libraries to write and read back C++ objects in an automatic way.
- Scripting: introspection allows users to interact with C++ objects from scripting languages at run-time.

However, unlike other languages such as Java and Python, the C++ language standard does not currently support reflection. A new reflection system for C++ that supports complete introspection of C++ types at run-time has therefore been developed in this project [68]. This involved developing a model for describing the reflection information that conforms as closely as possible to the C++ standard. Information about the types of the system that conform to this model is stored in dictionary modules.

Currently the LCG reflection system consists of several software packages including:

- Reflex, which is the library implementing the reflection model and API
- The LCGDICT package, which provides for the production of dictionaries in an automatic way from header files in a non-intrusive manner
- Cintex, a package that allows cross population of Reflex and CINT (ROOT) dictionaries.

In the future the reflection capabilities of the Reflex package will be adopted by ROOT, which implies converging on a single common dictionary making Cintex unnecessary. This will bring several advantages:

- Data files written with POOL can be natively accessed from within ROOT.
- Only one code base has to be developed and maintained.
- POOL users will observe a smaller memory allocation as only one dictionary system will be loaded into memory, and ROOT users will benefit from the smaller footprint of the Reflex dictionaries.
- Better support of C++ constructs within Reflex will allow more operations through the CINT interpreter.
- Reflex will stay a modular package and users needing only reflection capabilities will be able to use Reflex in a stand alone manner.

A workshop, organized at CERN at the beginning of May 2005, showed the feasibility of this approach and a detailed work plan to achieve this final goal has been agreed.

5.4.3 Scripting Services

Scripting is an essential ingredient in today's software systems. It allows rapid application development to produce quick prototypes, which can be readily exploited in physics analysis and other applications. The Applications Area has chosen to support the use of two languages:

- CINT, an interpreted implementation of C++ developed in the context of the ROOT project
- Python, which is also an interpreted, interactive, object-oriented programming language.

Extension modules are being developed to bind existing custom-written C or C++ libraries to the Python and C++ environments. These bindings to basic services may be viewed as a 'software bus' that allows easy integration of components, implemented in a variety of languages and providing a wide range of functionality.

The use of Python as a language for steering scientific applications is becoming more widespread. Python bindings to C++ objects can be generated automatically using dictionary information and a package (PyLCGDict) has been developed to enable the Python interpreter to manipulate any C++ class for which the dictionary information exists without the need to develop specific bindings for that class [69]. PyLCGDict is already used, for example, to provide bindings to physics analysis objects developed in the ATLAS and LHCb experiments. Another package, PyROOT, allows interaction with any ROOT class by exploiting the internal ROOT/CINT dictionary.

More recently work has led to a new API (Reflex package) for the reflection model in collaboration with the ROOT developers. The goal is to achieve a common API and common dictionary between LCG and ROOT, which will automatically give access and communication between the two environments without the need to develop dedicated gateways. A new package has been under development, PyReflex, to deal with the new reflection model API.

The final goal is to provide symmetry and interoperability between Python and CINT such that the end-user has the freedom to choose the best language for his/her purpose. To date Python courses have been prepared and delivered to more than 70 people as part of the CERN technical training programme.

5.4.4 Mathematical Libraries and Histograms

The provision of a validated and well-documented set of mathematical and statistical libraries is essential for the development of the full range of LHC physics applications spanning analysis, reconstruction, simulation, calibration etc. The primary goal of this project (Mathlib) is to select, package and support libraries that together provide a complete and coherent set of functionality to end-users and to ease the maintenance load by avoiding unnecessary duplication [70].

A thorough evaluation of the functionality offered by existing HEP libraries, such as CERNLIB and ROOT, has already been made and compared to that provided by general-purpose mathematical libraries such as the open source GNU Scientific Library (GSL) and the commercial NagC library. From this study a rather complete inventory of required mathematical functions and algorithms was compiled and made available on the project website. The various components of the required library may be classified as follows:

- Mathematical functions: special functions and statistical functions needed by HEP.
- Numerical algorithms: methods for numerical integration, differentiation, function minimization, root finders, interpolators, etc.
- C++ Function classes: generic functions, parametric functions or probability density functions used in conjunction with the numerical algorithms.
- Linear algebra: vector and matrix classes and their operations.
- Random number generators: methods for generating random numbers according to various distributions.
- Fitting and minimization libraries, including the minimization package MINUIT.
- Vector libraries describing vectors in 3D and in 4D (Lorentz Vectors).
- Statistical libraries for data analysis.
- Histogram library.

The activities of the last year have concentrated on providing a core mathematical library using implementations of the mathematical functions and the numerical algorithms contained in the GNU Scientific Library. The library includes an implementation of the special functions which conforms to the proposed interface to the C++ Standard. This involved making a detailed evaluation of the GNU Scientific Library in order to confirm the accuracy of the numerical results and therefore its quality. In addition the MINUIT minimization package was re-written in C++ and its functionality enhanced. MINUIT has also been completed with a generic fitting package (FML), to provide a convenient way of using it in fitting problems.

Currently work is on-going in order to integrate what has been produced inside the ROOT framework. The ROOT mathematical activities are being re-organized to facilitate the integration with the SEAL packages and to satisfy the requirements imposed by the LHC experiments. The first deliverable will be to produce a core mathematical library, which will include a new package for random numbers and for geometry and Lorentz Vectors. These new packages will result from a merge between the existing CLHEP and ROOT versions.

5.4.5 User Interface and Visualization Components in ROOT

ROOT is an object-oriented data analysis framework that provides an interface for users to interact with their data and algorithms. Data can be analysed using many different algorithms and results can be viewed using different visualization techniques. The Applications Area is participating in the development and support of basic GUI and visualization components of ROOT.

ROOT's graphical libraries provide support of many different functions including basic graphics, high-level visualization techniques, output on files, 3D viewing etc. They use well-known world standards to render graphics on screen (X11, GDK, Qt, and OpenGL), to produce high-quality output files (PostScript, PDF), and to generate images for Web publishing (SVG, GIF, JPEG, etc.). This ensures a high level of portability and a good integration with other software available on the market. These graphical tools are used inside ROOT itself but are also executable in experiment applications such as those for data monitoring and event display.

Many techniques allow visualization of all the basic ROOT data types (e.g., histograms, Ntuples, 'trees', graphs, analytic functions, cuts), projected in different dimensions and co-ordinate systems (2D, pseudo 3D, full 3D, 4D) and can be produced in high quality for publication purposes. Work is ongoing to support the existing tools, to improve their functionality and robustness, and to provide documentation and user support. 3D visualization must be enhanced to make sure it will be able to visualize and interact with the very complex detector geometries at LHC.

The Graphical User Interface (GUI) consists of a hierarchy of objects, sometimes referred to as window gadgets (widgets), that generate events as the result of user-actions. The Graphical User Interface is a bridge between the user and a software system — it provides methods that detect user actions and that react to them. The user communicates with an application through the window system which reports interaction events to the application.

The ROOT GUI classes are fully cross-platform compatible and provide standard components for an application environment with Windows 'look and feel'. The object-oriented, event-driven programming model supports the modern signals/slots communication mechanism as pioneered by Trolltech's Qt. This communication mechanism is an advanced object communication concept that replaces the concept of call-back functions to handle actions in GUIs. It uses ROOT dictionary information to connect signals to slots in ROOT. The ROOT GUI classes interface to the platform-dependent, low-level graphics system via a single abstract class. Concrete versions of this abstract class have been implemented for X11, Win32, and Qt.

A well-designed user interface is extremely important as it provides the window to users to view the capability of their software system. Many tasks remain to be done in the future in order to provide missing components such as undo/redo features, a set of object editors, and improvements to the tree viewer application, etc. The GUI design and integration are primary elements that have a direct impact on the overall quality and the success of the interactive data analysis framework.

5.5 Data Management

The POOL project (acronym for POOL Of persistent Objects for LHC) provides a general-purpose persistency framework to store experiment data and metadata in a distributed and Grid enabled way. This framework combines C++ object streaming technology (ROOT I/O, ref [71]) for the bulk data with transactionally safe Relational Database Management Systems (RDBMS) such as MySQL or Oracle for file, collection and event-level metadata. The POOL project was started mid 2002, as a common effort between CERN and the LHC experiments [72], [73], [74]. The strong involvement of the experiments has facilitated the implementation of their requirements as well as the integration of POOL into their software frameworks [75].

The POOL framework is structured into three main areas, which expose technology-independent (abstract) component interfaces. Each main interface is provided by several technology specific implementations as shown in Figure 5.2. This allows POOL to adapt to the requirements of very different environments (ranging from a small development system to a fully Grid connected production set-up) without imposing code changes on the user side.

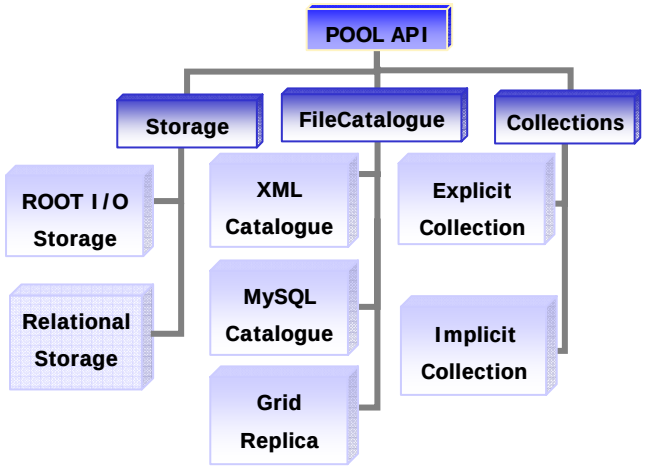


Figure 5.2: POOL components breakdown

The POOL Storage Service components are responsible for streaming C++ transient objects to and from disk. The POOL File Catalogue components maintain a consistent lists of data files (or databases connections) mapping the unique and immutable file identifiers to the physical locations of file or database replicas, which are used by the POOL storage system. Finally, a Collection component provides the tools to manage large ensembles of objects stored via POOL and provides the base for a technology-independent implementation of event collections.

5.5.1 Storage Components

The POOL storage system consists of multiple software layers, which define the store semantics and implement the basic functionality using different back-end technologies. The top layer from the user point of view is the Data Service, which is responsible for providing a client-side cache of C++ objects and transparent navigation among them via smart pointers (POOL Refs). The Data Service uses the Persistency Service to orchestrate the transactionally consistent access to any persistent storage (file or database). The persistency service delegates individual object access to the available Storage Services. Two implementations are supported today, one based on ROOT I/O for the storage of bulk event data and on RDBMS for metadata. The interactions between these storage components and the LCG dictionary are documented in Ref. [76].

The foundation for POOL database access is provided via a Relational Abstraction Layer (RAL), which defines a vendo- independent interface to several back-end databases [77]. At the time of writing Oracle, MySQL and SQLight are supported and are choosen at runtime via the LCG plug-in mechanism. The POOL RAL components also connect POOL-based components to the distributed database infrastructure, which is currently set up among the LCG sites [78].

5.5.2 Catalogues and Grid Integration

The basic model of a File Catalogue, shown in Figure 5.3, assumes the standard many-to-many mapping between logical file names (LFN) and physical file names (PFN) implemented by Grid middleware. POOL has introduced a system-generated file identifier, based on so-called Globally Unique Identifiers (GUID) [79], to insure stable inter-file references in an environment were both logical and physical file names might change. In addition POOL optionally supports file-level metadata to support queries on large file catalogues. This has been utilized by some experiments production systems to define catalogue fragments, e.g., for transfer to other sites or decoupled production nodes [80].

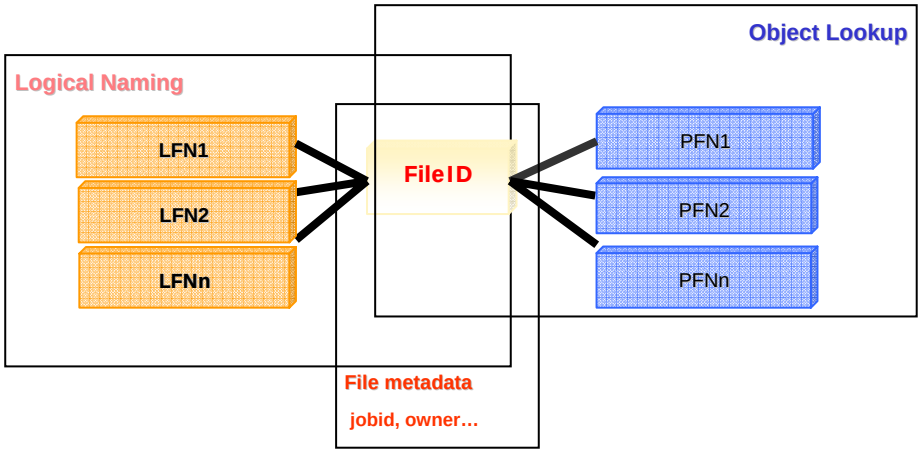


Figure 5.3: Logical view of the POOL file catalogue

The file catalogue component provides both a C++ API and a set of command-line tools, which can be used outside the application process for catalogue management operations. Concrete catalogue implementations are provided in collaboration with the Grid middleware providers for the LCG-RLS, LCG-LFC, GLOBUS-RLS, gLite catalogues. To complement these Grid-enabled catalogues, POOL provides local catalogue implementations based on XML and RAL as storage technologies for grid-decoupled use-cases. Different catalogue implementations can be deployed together (e.g., a read-only Grid catalogue with a writable user catalogue) and cross population and catalogue management are supported via a set of uniform command line tools. Most experiment production environments make use of more than one file catalogue implementation to decrease the runtime dependencies on central cataloguing services [81].

5.5.3 Collections

POOL provides several implementations of persistent object collections, which expose a coherent access and iteration interface independent of the concrete collection store. Supported back-end implementations are provided for ROOT-file-based collections (e.g., ROOT trees or ROOT directories) and RDBMS-based collections (relational tables containing collection elements in either Oracle, MySQL or SQLight). Similar to the file catalogue area, also here a set of technology-independent management tools is provided to administer collection data.

5.5.4 Conditions Database

The second project in the persistency framework area is the COOL project, which provides a common framework for access to conditions data such as detector alignment and calibration, environment conditions (e.g., measured temperatures and pressures). This time-versioned data is typically accessed based on the interval of validity (IOV) and needs to be versioned as more precise calculations of calibration values become available. To insure consistency between versions of different conditions quantities a CVS-like tagging mechanism is implemented.

The COOL system provides like the other LCG application area projects a common interface to support a conditions data independent of the back-end database used. The implementation of COOL uses the same relational abstraction as the POOL project and therefore shares the same monitoring and security infrastructure with the rest of the persistency framework. For a more detailed description of the COOL package please refer to [82].

5.6 Event Simulation

The simulation of an LHC experiment is an important element to allow the understanding of the experimental conditions and its performance, both in the optimization and design phase as well as during future data taking and analysis. The simulation project of the LCG Applications Area encompasses common work among the LHC experiments and is organized

into several subprojects which report to the Simulation Project leader. The principle activities in the various subprojects are described in the following.

5.6.1 Event Generator Services

The LCG Generator project collaborates with the authors of Monte Carlo (MC) generators and with LHC experiments in order to prepare validated code for both the theoretical and experimental communities at the LHC. Tasks include sharing the user support duties, providing assistance for the development of the new object-oriented generators and guaranteeing the maintenance of existing packages on LCG-supported platforms.

The Generator library (GENSER) is the central code repository for MC generators and generator tools, including test suites and examples. This new library is intended to gradually replace the obsolete CERN MC Generator library. It is currently used in production by ATLAS, CMS and LHCb. The current version of GENSER (1.0.0) includes most of the popular event generators in use at the LHC, including PYTHIA, HERWIG, JIMMY, ISAJET, EVTGEN, ALPGEN, COMPHEP, LHAPDF, PDFLIB, PHOTOS, GLAUBER and HIJING.

The LCG generator project also contributes to the definition of the standards for generator interfaces and formats, collaborating in the development of the corresponding application program interfaces. One example is the Toolkit for High Energy Physics Event Generation framework (THEPEG), a common effort between authors of MC generators that eases the integration of the new object-oriented MC generators in the experiment simulation frameworks. The first test of ThePEG integration in Herwig++ has been set for Q3 2005.

The project also works on the production of 'certified' event files that contain the data output by the generators and that can be used by all LHC experiments for benchmarks, comparisons and combinations. Three different developments have been started

- a simple production and validation framework at generator level,
- a dedicated production centre to provide the LHC experiments and other end-users with a transparent access to the public event files,
- a public database for the configuration, book-keeping and storage of the generator level event files (MCDB).

5.6.2 Detector Simulation

The LCG Project provides the context for supporting the use of GEANT4 for detector simulation by the LHC experiments. The team of GEANT4 developers based at CERN provides contact persons for LHC experiments and undertakes parts of the support, maintenance and development in a number of key areas of the toolkit (in particular geometry, integration testing and release management, electromagnetic and hadronic physics, and software/repository management). Requirements for new capabilities and refinements are received from LHC experiments at different times. Simple requirements are addressed directly, often with the assistance of other GEANT4 collaborators. Requirements that are complex, have large resource needs or broad impact are discussed at the quarterly meetings of the GEANT4 Technical Forum, and the work is evaluated and planned by the GEANT4 Steering Board in consultation with the concerned users.

GEANT4-based detector simulation programs entered production between November 2003 and May 2004 in CMS, ATLAS and LHCb and have demonstrated very low crash rates (less than one crash per ten thousand events) and computing performance comparable to GEANT3 (the latest within a factor of 1.5 to 2) [83]. The considerable set of physics validations in test beam set-ups has provided a measure of the physics performance achieved. These GEANT4-based simulation programs continue to evolve, utilizing new capabilities of the GEANT4 toolkit, and continue to provide regular and important feedback. The widespread and growing use of these simulations in productions for physics studies is enabling further comparisons and validation tests of the GEANT4 toolkit under realistic conditions.

The latest developments in the toolkit [84] have included robustness improvements, a number of new hadronic models addressing primarily the interactions of ions, as well as improvements in the Photo Absorption Ionization (PAI) and multiple scattering models. CMS and ATLAS developers contributed a new module for performing fast shower parametrization using the techniques of the GFLASH package for GEANT3. A configurable calorimeter set-up has been created for use in a suite for making statistical regression tests. Different calorimeter set-ups are defined, spanning simplified versions of LHC experiment calorimeters.

Of the current work, a large part of the effort involves the support and maintenance of existing classes and functionality, identifying issues and improvements required, and addressing problem reports on key components. Work is ongoing to extend the verification of physics models for thin-target tests and to follow up on issues arising from experiment test-beam studies. Geometry improvements are addressing issues related to surface boundaries of complex Boolean volumes, which have been seen infrequently in large productions. A new shape has been created, a general twisted trapezoid with different parallel trapezoidal caps, to address a requirement from the ATLAS EM calorimeter endcap. An improved facility for parallel navigation will enable the calculation of radiation flux tallies on arbitrary surfaces.

Planned refinements in electromagnetic (EM) physics include improvements in ionization processes at small production thresholds, a prototype model for the multiple scattering of electrons addressing effects at low energies, a review of the LPM effect and additional channels for high-energy positron annihilation. Planned hadronic physics developments include a propagator interface in Binary Cascade [85] to string models, to enable use of this promising intermediate energy model in sensitive applications. Refinements in the Chips model, enabling its use in the capture of negatively charged particles, and for the treatment of string fragmentation are ongoing.

Continued improvements in testing, will include identifying and extending the power of current regression tests for shower shape, and refining the Bonsai tool for choosing and steering integration tests. Work on monitoring and improvement of computing performance is ongoing.

5.6.3 Simulation Framework

The general task of the Simulation Framework subproject is to provide flexible infrastructure and tools for the development, validation and usage of Monte Carlo simulation applications. The purpose of this is to facilitate interaction between experiments and simulation toolkits developers as well as to eliminate duplication of work and divergence. The Simulation Framework subproject consists of several work packages addressing particular areas of detector simulation, such as the geometry description exchange mechanisms, geometry persistency, Python interfacing, Monte Carlo truth handling as well as a generalized interface to different simulation toolkits for application in physics validation studies.

The Geometry Description Markup Language (GDML) has been adopted as the geometry exchange format. Its purpose is to allow the interchange of detector geometries between different applications (simulation and/or analysis). GDML processors have been implemented in C++ and in Python.

In addition, effort is being devoted to address direct object persistency in GEANT4. It is planned to perform a feasibility study of the usage of POOL for that purpose. Such a mechanism would be useful for running detector simulation of complex detectors, as well as for storing GEANT4 geometries that are constructed interactively.

Python interfaces to C++ applications have already proven their usefulness in adding flexibility, configurability as well as facilitating the 'gluing' of different elements together. This technology has also clear benefits in the context of detector simulation. The effort undertaken so far demonstrates the usage of Reflex and its Python binding for running GEANT4 applications from the Python prompt. Several examples have been implemented

and documented on the Simulation Framework Web page [86]. An example has been implemented demonstrating GEANT4 simulation interfaced to ROOT visualization, all in Python and using GDML as the geometry source. This uses the existing ROOT Python binding (PyRoot).

Monte Carlo truth handling is a difficult task, especially for large multiplicity events found at the LHC. There are a large number of particles produced in the showers and the selection criteria for filtering out unwanted particles are often complicated. All of the LHC experiments have come up with their own solutions, but improvements in performance and flexibility can still be envisaged. A feasibility study for a common mechanism for MCTruth handling is under consideration.

Finally, a more general approach to interfacing different simulation engines has been adopted by the Virtual Monte Carlo project [87]. A complete interface to a generalized simulation toolkit has been implemented, isolating the user from the actual simulation engine. Both the geometry as well as the simulation workflow is treated in a toolkit-independent way. This approach has been developed by ALICE and will also be used in the physics validation studies described below.

5.6.4 Physics Validation

The goal of the Physics Validation project is to compare the main detector simulation engines for LHC, GEANT4 and FLUKA, with experimental data, in order to understand if they are suitable for LHC experiment applications. The main criterion for validating these simulation programs is that the dominant systematic effects for any major physics analyses should not be dominated by the uncertainties coming from simulation. This approach relies on the feedback provided by the physics groups of the LHC experiments to the developers of these simulation codes.

Two classes of experimental set-ups are used for physics validation: calorimeter test-beams, and simple benchmarks. These provide complementary information, because the observables in calorimeter test-beam set-ups are of direct relevance for the experiments but are the macroscopic result of many types of interactions and effects, whereas with simple benchmark set-ups it is possible to make microscopic tests of single interactions.

The electromagnetic physics has been the first large sector of the physics models that have been carefully validated, with excellent agreement with data at the per cent level. Over the last couple of years, most of the physics validation effort has been focused on hadronic physics, which is a notoriously complex and broad field, owing to the lack of predictive power of QCD in the energy regime of relevance for tracking hadrons through a detector. This implies that a variety of different hadronic models are needed, each suitable for a limited selection of particle type, energy, and target material.

The results of these studies have been published in a number of LCGAPP notes [88]. The software infrastructure has been set up to compare FLUKA and GEANT4 with data for simple geometries and ‘single interactions’. Firstly studies of 113 MeV protons on thin Al targets, and comparisons to Los Alamos data, were performed. The study of double differential cross-sections for (p, xn) at various energies and angles has also been completed. Radiation background studies in the LHCb experiment, aiming at comparing GEANT4/FLUKA/GCALOR, have started. Physics validation of FLUKA using ATLAS Tilecal test-beam data is also in progress. Comparisons of test-beam data with GEANT4 have concentrated on hadronic physics with calorimeters, both in ATLAS and CMS, as well as with special data collected with the ATLAS pixel detector. One interesting result is that corrections to the pion cross-section in GEANT4 have yielded significant improvements in the description of the pion shower longitudinal shape in the LHC calorimeters

The conclusions of the first round of hadronic physics validations are that the GEANT4 LHEP and QGSP Physics Lists, currently in use by three LHC experiments (ATLAS, CMS, LHCb), are in good agreement with data for the hadronic shower energy resolution and e/π

ratio. For the hadronic shower shapes, both longitudinal and transversal, the comparisons between data and simulation are less satisfying. In particular, the GEANT4 QGSP Physics List seems to produce hadronic showers slightly too short and narrow with respect to those seen in the data. Work is ongoing in order to address this discrepancy.

Physics validation activities will continue in order to take advantage of new data currently being taken in the ATLAS and CMS test beams. The calorimeter test-beam data will also be used for validating the hadronic physics of FLUKA, similarly to what has already been done for the simple benchmark tests. A new simple benchmark test, relevant for LHC detector applications, has started, and others are foreseen for the future. Background radiation studies with GEANT4 are in progress, and comparisons with FLUKA results will be made available. A longer term goal is to make all the data useful for validating detector simulations properly organized and available from a central repository, in such a way as to be routinely utilized at each new release by the code developers. This will also provide users with a consistent and documented monitor of the precision of the various physics models, allowing a more effective and clear choice of the best simulation for their applications.

5.6.5 Simulation of Gaseous Detectors with Garfield

Garfield is a computer program for the detailed simulation of two- and three-dimensional chambers made of wires and planes, such as drift chambers, TPCs and multiwire counters. For most of these configurations, exact fields are known. This is not the case for three-dimensional configurations, not even for seemingly simple arrangements like two crossing wires. Furthermore, dielectric media and complex electrode shapes are difficult to handle with analytic techniques. Garfield therefore also accepts two- and three-dimensional field maps computed by finite element programs such as Maxwell, Tosca and FEMLAB as a basis for its calculations. Work is ongoing to upgrade interfaces to all these finite element programs.

An interface to the Magboltz program is provided for the computation of electron transport properties in nearly arbitrary gas mixtures. Garfield also has an interface with the Heed program to simulate ionization of gas molecules by particles traversing the chamber. New releases of both Heed and Magboltz are in the process of being interfaced and the cross-sections of Magboltz have been documented. The integration of the new release of Heed will also mark a major change in the programming aspects of Garfield since Heed is now written in C++. Garfield, already containing significant portions of C, will at that point probably have a main program in C++.

Transport of particles, including diffusion, avalanches and current induction is treated in three dimensions irrespective of the technique used to compute the fields. Currently Monte Carlo simulations of drift with diffusion assume Gaussian spreads. This is not applicable in detectors such as GEMs where, indeed, the calculated diffusion spread depends on the step length. Correction of this is in progress.

Negative-ion TPCs are being considered as detectors in the search for dark matter. To simulate these devices needs not only attachment processes, which are already available, but also dissociation processes. These are in the process of being written.

5.7 Software Development Infrastructure and Services

The LCG Applications Area software projects share a single development infrastructure; this infrastructure is provided by the SPI project. A set of basic services and support are provided for the various activities of software development. The definition of a single project managing the infrastructure for all the development projects is crucial in order to foster homogeneity and avoid duplications in the way the AA projects develop and manage their software.

5.7.1 Build, Release and Distribution Infrastructure

A centralized software management infrastructure has been deployed [89]. It comprizes solutions for handling the build and validation of releases as well as providing a customized

packaging of the released software. Emphasis is put on the flexibility of the packaging and distribution procedure as it should cover a broad range of needs in the LHC experiment, ranging from full packages for developers in the projects and experiments to a minimal set of libraries and binaries for specific applications running on CPU nodes.

Configuration management support is provided for all LCG projects in both CMT and SCRAM configurations such that LCG software can be used in the various build environments of the experiments. LCG software is distributed using Web-downloadable tarfiles of all binaries. In the near future 'pacman' repositories of both sources and binaries will be provided.

5.7.2 External Libraries

The External Software Service provides open source and public domain packages required by the LCG projects and experiments [90]. Presently, more than 50 libraries and tools are provided on the set of LCG-supported platforms. All packages are installed following a standard procedure and are documented on the web. A set of scripts has been developed to automate new installations.

5.7.3 Software Development and Documentation Tools

All the tools used for software development in the Applications Area projects are either standard on the platform used or provided as part of the External Libraries Service. Compilers, test frameworks, documentation tools (e.g., Doxygen, LXR) are made available on all supported platforms. Support is provided for all these tools.

5.7.4 Quality Assurance and Testing

Software Quality Assurance is an integral part of the LCG software development process and includes several activities such as automatic execution of regression test suites, and automatic generation of test coverage reports [91]. Several software metrics are used to measure quality and reports are generated giving information on the number of defects, code complexity, usage statistics and compliance to build, code and release policies.

Test frameworks (CppUnit, PyUnit, Oval, and QMTest) are provided in order to support unit and regression tests. Tools are also provided for handling specific development issues, such as Valgrind for memory leak detection.

5.7.5 Savannah Web-based Services

A Web-based 'project portal' based on the Savannah open source software has been deployed and has been put in production [92]. It integrates a bug tracking tool with many other software development services. This service is now in use by all the LCG projects and by more than 100 projects in the LHC experiments. The portal is based on the GNU Savannah package which is now developed as 'Savane' by the Free Software Foundation. Several features and extensions were introduced, in collaboration with the current main developer of Savannah, to adapt the software for use at CERN and these were merged back into the Savannah open source. Work is ongoing to maintain the system and to implement new features according to requests from users.

5.8 Project Organization and Schedule

Applications Area work in the various activity areas described above is organized into projects, as shown in Figure 5.4, each led by a Project Leader with overall responsibility for the management, design, development and execution of the work of the project. The Applications Area Manager has overall responsibility for the work of the Applications Area.

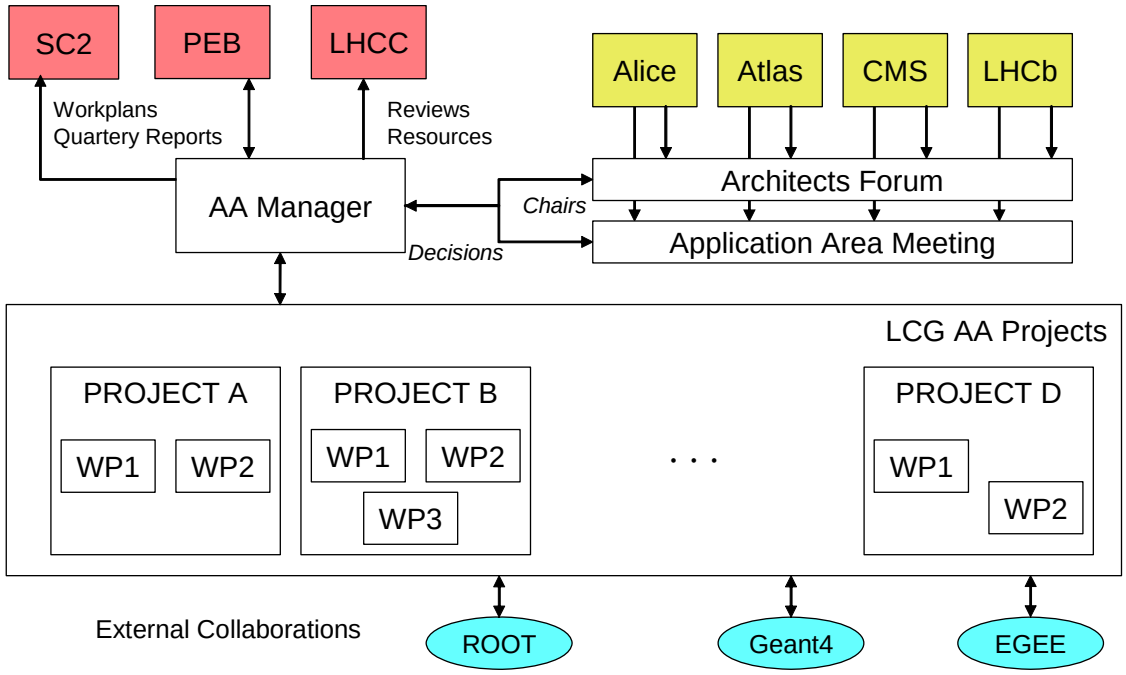


Figure 5.4: Applications Area organization

Work in the projects must be consistent and coherent with the architecture, infrastructure, processes, support and documentation functions that are agreed Applications Area-wide. Larger projects may in turn be divided into work packages with ~1–3 FTE activity levels per work package.

An Architects Forum (AF) consisting of the Applications Area Manager (chair), the software architects of the four LHC experiments, the leaders of the various AA software projects and other invited members provides for the formal participation of the experiments in the planning, decision-making and architectural and technical direction of applications area activities. Architects represent the interests of their experiment and contribute their expertise. The AF meets every two weeks and makes decisions about the difficult issues that cannot be resolved in open forums such as the Applications Area meeting. The Applications Area Meeting takes pace fortnightly and provides a forum for information exchange between the project and the LHC experiments.

The Applications Area work breakdown structure, milestones and deliverables for all aspects of the project are documented on a Web page [93]. The work breakdown structure maps directly onto the project breakdown of the Applications Area. The schedule of milestones for the completion of deliverables is similarly organized. Milestones are organized at three levels:

- Level 1: the highest level. A small, select number of very important milestones are at this level. These milestones are monitored at the LHCC level.
- Level 2: the ‘official milestones’ level. Milestones at this level chart the progress of applications area activities in a comprehensive way. Each project has a small number of milestones per quarter at this level. These milestones are monitored at the LCG Project level.
- Level 3: internal milestones level. Milestones at this level are used for finer-grained charting of progress for internal applications area purposes. These milestones are monitored at the AA level.

Milestones include integration and validation milestones from the experiments to track the take-up of AA software in the experiments.

6 EXPERIENCE: PROTOTYPES AND EVALUATIONS

The LCG system has to be ready for use with full functionality and reliability from the start of LHC data taking. In order to ensure this readiness, the system is planned to evolve through a set of steps involving the hardware infrastructure as well as the services to be delivered.

At each step the prototype LCG is planned to be used for extensive testing:

- By the experiments, that perform their Data Challenges, progressively increasing in scope and scale. The aim is to stress the LCG system with activities that are more and more similar to the ones that will be performed when the experiments will be running. The community of physicists is also involved more and more, and gives the feedback necessary to steer the LCG evolution according to the needs of the users.
- By the service providers themselves, at CERN and in the outside Tier-1 and Tier-2 centres that perform Service Challenges, aimed at stressing the different specific services. The Service Challenges involve for the specific services a scale, a complexity, and a level of site co-ordination higher than that needed at the same time by the Data Challenges of the experiments.

Part of the plan of the Challenges has already been executed, and has provided useful feedback. The evaluation of the results of the Challenges and the implementation of the suggestions coming from this evaluation will provide an important contribution towards reaching the full readiness of the LCG system on schedule.

6.1 Data Challenges

The LHC Computing Review [94] in 2001 recommended that the LHC experiments should carry out Data Challenges (DC) of increasing size and complexity. A Data Challenge comprises, in essence, the simulation, done as realistically as possible, of data (events) from the detector, followed by the processing of that data using the software and computing infrastructure that will, with further development, be used for the real data when the LHC starts operating.

All Data Challenges are to prepare for LHC running and include the definition of the computing infrastructure, the definition and set-up of analysis infrastructure, and the validation of the computing model. They entail each year an increase in complexity over the previous year, leading to a full-scale test in 2006.

Even though their primary goal is to gradually build the computing systems of the experiments in time for the start of LHC, they are tightly linked to other activities of the experiment and provide computing support for production and analysis of the simulated data needed for studies on detector, trigger, and DAQ design and validation, and for physics system set-up.

6.1.1 ALICE

The specific goals of the ALICE Physics Data Challenges are to validate the distributed computing model and to test the common LCG middleware and the ALICE developed interfaces which provide all the functionalities required by distributed production and analysis.

6.1.1.1 Physics Data Challenge 2004

ALICE has used the AliEn services either in native mode or interfaced to the LCG-Grid for distributed production of Monte Carlo data, reconstruction and analysis at over 30 sites on four continents. The Physics Data Challenge 2004 (PDC04) aimed at providing data for the ALICE Physics Performance Report. During this period more than 400,000 jobs were

successfully run under AliEn control worldwide producing 40 TB of data. Computing and storage resources were provided both in Europe and the US. The amount of processing needed for a typical production is in excess of 30 MSI2000×s to simulate and digitize a central Pb–Pb event. Some 100k heavy-ion (underlying) events were generated for each major production. This required CPU time varying over a very large range since a peripheral Pb–Pb event may require one order of magnitude less CPU than a central event, and a p–p event two orders of magnitude less. Each underlying Pb–Pb event was reprocessed several times superimposing known signals which were subsequently reconstructed and analysed. Again there is a wide spread in the required CPU time this takes depending on the event type. For a central event a few MSI2000×s are needed. Each Pb–Pb central event occupies about 2 GB of disk space, while p–p events are two orders of magnitude smaller.

The asynchronous distributed analysis of the produced data is starting at the time of writing of the present document.

6.1.1.2 Physics Data Challenge 2005

The goals of the Physics Data Challenge 2005 (PDC05) are to test and validate parts of the ALICE computing model. These include the quasi-online reconstruction, without calibration, of raw data at CERN (Tier-0), export of the data from CERN to Tier-1 for remote storage, delayed reconstruction with calibration at Tier-1 sites, asynchronous and synchronous analysis.

The PDC05 will be logically divided into three phases:

- resource-dependent production of events on the Grid with storage at CERN;
- quasi-online first pass reconstruction at CERN, push data from CERN to Tier-1 sites, second-pass reconstruction at Tier-1 sites with calibration and storage;
- analysis of data: batch and interactive analysis with PROOF.

For this exercise, ALICE will use the Grid services available from LCG in the framework of the LCG Service Challenge 3 and AliEn (Alice Environment) for all high-level services.

The AliEn framework has been developed with the aim of offering to the ALICE user community a transparent access to distributed computing resources through a single interface, shielding the users from the complexity of the Grid world. Through interfaces it uses resources of different Grids developed and deployed by other groups, transparently. In addition, AliEn has been engineered to be highly modular, and individual components can be deployed and used in a foreign Grid, which is not adapted to the specific computational needs of ALICE. The system uses a Web Services model and standard network protocols. The user interacts with them by exchanging SOAP messages.

AliEn consists of the following components and services: authentication, authorization an auditing services; workload and data management systems; file and metadata catalogues; the information service; Grid and job monitoring services, storage and Computing Elements.

The AliEn workload management system is based on a ‘pull’ approach. A service manages a common task queue, which holds all the jobs of the ALICE VO. On each site providing resources, CE services act as ‘remote queues’ giving access to computational resources that can range from a single machine, dedicated to running a specific task, to a cluster of computers in a computing centre, or even an entire foreign Grid. When jobs are submitted, they are sent to the central queue. The workload manager optimizes the queue taking into account job requirements such as the input files needed, the CPU time and the architecture requested, the disk space request and the user and group quotas. It then makes jobs eligible to run in one or more CEs. The CEs of the active nodes get jobs from the central queue and deliver them to the remote queues to start their execution. The queue system monitors the job progress and has access to the standard output and standard error.

Input and output associated with jobs are registered in the AliEn file catalogue, a virtual file system in which LFNs are assigned to files and which keeps an association between LFN and PFN. The file catalogue supports file replication and caching and it provides the information about file location to the RB.

ALICE has used the system for distributed production of Monte Carlo data, reconstruction and analysis to realize PDC04. The Grid user data analysis has been tested in a limited scope using tools developed in the context of the ARDA project. Two approaches were prototyped and demonstrated: the asynchronous (interactive batch approach) and the synchronous (true interactive) analysis.

The asynchronous model has been realized by extending the ROOT functionality to make it Grid-aware. As the first step, the analysis framework has to extract a subset of the datasets from the file catalogue using metadata conditions provided by the user. The next part is the splitting of the tasks according to the location of datasets. Once the distribution is decided, the analysis framework splits the job into sub-jobs and inserts them in the AliEn task queue. The jobs are then submitted to the local CEs for execution. Upon completion, the results from all sub-jobs are collected, merged and delivered to the user.

The synchronous analysis model relies on extending the functionality of PROOF. The PROOF interface to Grid-like services is presently being developed, focusing on authentication and the use of the file catalogue to make both accessible from the ROOT shell. The AliEn-PROOF-based system for distributed synchronous analysis will be used for a rapid evaluation of large data samples in a time-constrained situation, for example the evaluation of the detector calibration and alignment at the beginning of a data-taking period.

6.1.2 ATLAS Data Challenges

The goals of the ATLAS Data Challenges are the validation of the ATLAS Computing Model, of the complete software suite, of the data model, and to ensure the correctness of the technical computing choices to be made.

A major feature of the first Data Challenge (DC1) was the development and the deployment of the software required for the production of large event samples required by the High-level Trigger and Physics communities, and the production of those large data samples involving institutions worldwide.

ATLAS intended to perform its Data Challenges using as much as possible Grid tools provided by the LHC Computing Grid project (EDG), NorduGrid and Grid3. DC1 saw the first usage of these technologies in ATLAS, where NorduGrid for example relied entirely on Grid. Forty institutes from 19 countries participated in DC1 which ran from spring 2002 to spring 2003. It was divided into three phases: (1) event generation and detector simulation, (2) pile-up production, (3) reconstruction. The compute power required was 21 MSI2000-days. 70 TB of data were produced in 100,000 partitions.

In order to handle the task of ATLAS DC2 an automated production system was designed. All jobs are defined and stored in a central database. A supervisor agent (Windmill) picks them up, and sends their definition as XML message to various executors, via a Jabber server. Executors are specialized agents, able to convert the XML job description into a Grid-specific language (e.g., the Job Description Language (JDL) for LCG and the Extended Resource Specification Language (XRSL) for NorduGrid). Four executors have been developed, for LCG (Lexor), NorduGrid (Dulcinea), GRID3 (Capone) and legacy systems, allowing the Data Challenge to be run on different Grids. The interplay between the components is shown in Figure 6.1.

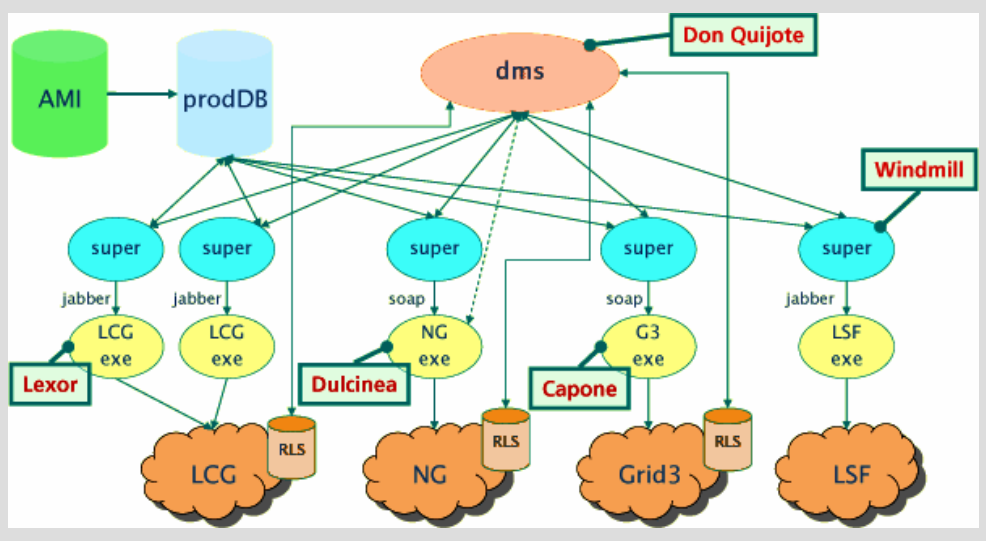


Figure 6.1: Components used in ATLAS Data Challenges

For data management, Don Quijote (DQ), a central server developed by ATLAS, offers a uniform layer over the different replica catalogues of the three Grid flavours. Thus all the copy and registration operations are performed through calls to DQ. The automatic production system has submitted about 235,000 jobs belonging to 158,000 job definitions in the database, producing around 250,000 logical files and reaching approximately 2500–3500 jobs per day, evenly distributed over the three Grid flavours. Overall these jobs consumed approximately 1.5 million SI2000 months of CPU (~ 5,000 present CPUs per day) and produced more than 30 TB of physics data.

When a LCG job is received by Lexor, it builds the corresponding JDL description, creates some scripts for data staging, and sends everything to a dedicated, standard Resource Broker (RB) through a Python module built over the workload management system (WMS) API. The requirements specified in the JDL let the RB choose a site where ATLAS software is present and the requested amount of computation (expressed in SpecInt2000 × Time) is available. An extra requirement is a good outbound connectivity, necessary for data staging.

Dulcinea, was implemented as a C++ shared library. This shared library was then imported into the production system's Python framework. The executor calls the ARC user interface API and the Globus RLS API to perform its tasks. The job description received from the Windmill supervisor in the form of an XML message was translated by the Dulcinea executor into an extended resource specification language (XRSL) [95] job description. This job description was then sent to one of the ARC-enabled sites, selecting a suitable site using the resource brokering capabilities of the ARC user interface API. In the brokering, among other things, the availability of free CPUs and the amount of data needed to be staged in on each site to perform a specific task is taken into account. The look-up of input data files in the RLS catalogue and the stage-in of these files to the site is done automatically by the ARC Grid Manager. The same is true for stage-out of output data to a storage element and the registration of these files in the RLS catalogue. The Dulcinea executor has to add only the additional RLS attributes needed for the Don Quijote data management system to the existing file registrations.

The Dulcinea executor also takes advantage of other capabilities of the ARC middleware. The executor does not have to keep any local information about the jobs it is handling, but can rely on the job information provided by the Grid information system.

GRID3 involved 27 sites with a peak of 2,800 processors.

The 82 LCG deployed sites in 22 countries contributed with a peak of 7269 processors and a total storage capacity of 65 TB. In addition to problems related to Globus Replica Location Services (RLS), the Resource Broker and the information system were unstable at the initial

phase. But it was not only the Grid software that needed many bug fixes, another common failure was the mis-configuration of sites.

In total 22 sites in 7 countries participated in DC2 through NorduGrid/ARC, with 700 CPUs out of 3,000 being dedicated to ATLAS. The amount of middleware-related problems was negligible, except for the initial instability of the RLS server. Most job failures were due to specific hardware problems.

6.1.3 CMS

All CMS Computing data challenges are constructed to prepare for LHC running and include the definition of the computing infrastructure, the definition and set-up of analysis infrastructure, and the validation of computing model. By design they entail each year a factor 2 increase in complexity over the previous year, leading to a full-scale test in 2006.

Even though their primary goal is to gradually build the CMS computing system in time for the start of LHC, they are tightly linked to other CMS activities and provide computing support for production and analysis of the simulated data needed for studies on detector, trigger and DAQ design and validation, and for physics system set-up.

The purpose of the 2004 Data Challenge (DC04) was to demonstrate the ability of the CMS computing system to cope with a sustained data-taking rate equivalent to 25 Hz at a luminosity of $2 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$ for a period of 1 month. This corresponds to 25% of the LHC start-up rate (or 5% of the LHC full-scale system).

The CMS Data Challenge in 2004 (DC04) had the following phases:

- Reconstruction of data on the CERN Tier-0 farm for a sustained period at 25 Hz.
- Data distribution to Tier-1 and Tier-2 sites.
- Prompt data analysis at remote sites on arrival of data.
- Monitoring and archiving of resource and process information.

The aim of the challenge was to demonstrate the feasibility of operating this full processing chain.

6.1.3.1 PCP04 Data Productions

About 50 million events were required to match the 25 Hz rate for a month. More than 70 millions events were requested by the CMS physicists. These were simulated during 2003 and the first months of 2004 and about 35 million of them were digitized in time for the start of DC04. This task is known as the Pre-Challenge Production for DC04 (PCP04). Simulation of other events and digitization of the whole sample continued after the end of DC04. All events are being used by CMS physicists for the analysis needed for the Physics Technical Design Report.

Data production runs in a heterogeneous environment where some of the computing centres do not make use of Grid tools and the others use two different Grid systems: LCG in Europe and Grid3 in the USA. A set of tools, OCTOPUS, provide the needed functionalities.

The work-load management is done in two steps. The first assigns production slots to regional centres. The brokering is done by the production manager who knows about validated sites ready to take work. The second step assigns the actual production jobs to CPU resources. Brokering is performed either by the local resource manager or by a Grid scheduler. In the case of LCG this is the Resource Broker and in the case of Grid3 it is the match-making procedure within Condor. RefDB is a database located at CERN where all information needed to produce and analyse data is kept. It allows the submission of processing requests by the physicists, the assignment of work to a distributed production centre and the browsing of the status of the requests. Production assignments are created by the production team and assigned to centres that have demonstrated the ability to produce data properly (via the

execution of a *validation assignment*). At each site, McRunJob is used to create the actual jobs that produce or analyse the data following the directives stored in RefDB. Jobs are prepared and eventually submitted to local or distributed resources. Each job is instrumented to send to a dedicated database (BOSS) information about the running status of the job and to update the RefDB in case the job finished successfully. Information sent to RefDB by a given job gets processed by a validation script implementing necessary checks, after that RefDB gets updated with information about the produced data. The RLS catalogue, also located at CERN, was used during PCP as a file catalogue by the LCG Grid jobs.

The Storage Resource Broker (SRB) has been used for moving data among the regional centres and finally to CERN where they have been used as input to the following steps of the data challenge.

6.1.3.2 DC04 Reconstruction

Digitized data were stored on the CASTOR Mass Storage System at CERN. A fake online process made these data available as input for the reconstruction with a rate of 40 MB/s. Reconstruction jobs were submitted to a computer farm of about 500 CPUs at the CERN Tier-0. The produced data (4 MB/s) were stored on a CASTOR stage area, so files were automatically archived to tape. Some limitations concerning the use of CASTOR at CERN due to the overload of the central tape stager were found during DC04 operations.

6.1.3.3 DC04 Data Distribution

For DC04 CMS developed a data distribution system over available Grid point-to-point file transfer tools, to form a scheduled large-scale replica management system. The distribution system was based on a structure of semi-autonomous software agents collaborating by sharing state information through a Transfer Management DataBase (TMDB). A distribution network with a star topology was used to propagate replicas from CERN to six Tier-1s and multiple associated Tier-2s in the USA, France, UK, Germany, Spain and Italy. Several data transfer tools were supported: the LCG Replica Manager tools, Storage Resource Manager (SRM) specific transfer tools, and the Storage Resource Broker (SRB). A series of 'export buffers' at CERN were used as staging posts to inject data into the domain of each transfer tool. Software agents at Tier-1 sites replicated files, migrated them to tape, and made them available to associated Tier-2s. The final number of file-replicas at the end of the two months of DC04 was ~3.5 million. The data transfer (~6 TB of data) to Tier-1s was able to keep up with the rate of data coming from the reconstruction at Tier-0. The total network throughput was limited by the small size of the files being pushed through the system.

A single Local Replica Catalogue (LRC) instance of the LCG Replica Location Service (RLS) was deployed at CERN to locate all the replicas. Transfer tools relied on the LRC component of the RLS as a global file catalogue to store physical file locations.

The Replica Metadata Catalogue (RMC) component of the RLS was used as global metadata catalogue, registering the file attributes of the reconstructed data; typically the metadata stored in the RMC was the primary source of information used to identify logical file collections. Roughly 570k files were registered in the RLS during DC04, each with 5 to 10 replicas and 9 metadata attributes per file (up to ~1 kB metadata per file). Some performance issues were found when inserting and querying information; the RMC was identified as the main source of these issues. The time to insert files with their attributes in the RLS — about 3 s/file in optimal conditions — was at the limit of acceptability; however, service quality degraded significantly with extended periods of constant load at the required data rate. Metadata queries were generally too slow, sometimes requiring several hours to find all the files belonging to a given 'dataset' collection. Several workarounds were provided to speed up the access to data in the RLS during DC04. However, serious performance issues and missing functionality, like a robust transaction model, still need to be addressed.

6.1.3.4 DC04 Data Analysis

Prompt analysis of reconstructed data on arrival at a site was performed in quasi real-time at the Italian and Spanish Tier-1 and Tier-2 centres using a combination of CMS-specific triggering scripts coupled to the data distribution system and the LCG infrastructure. A set of software agents and automatic procedures were developed to allow analysis-job preparation and submission as data files were replicated to Tier-1s. The data arriving at the Tier-1 CASTOR data server (Storage Element) were replicated to disk Storage Elements at Tier-1 and Tier-2 sites by a Replica agent. Whenever new files were available on disk, the Replica agent was also responsible for notifying an Analysis agent, which in turn triggered job preparation when all files of a given file set (run) were available. The jobs were submitted to an LCG-2 Resource Broker, which selected the appropriate site to run the jobs.

The official release of the CMS software required for analysis (ORCA) was pre-installed on LCG-2 sites by the CMS software manager by running installation Grid jobs. The ORCA analysis executable and libraries for specific analyses were sent with the job.

The analysis job was submitted from the User Interface (UI) to the Resource Broker (RB) that interpreted the user requirements specified using the job description language (JDL). The Resource Broker queried the RLS to discover the location of the input files needed by the job and selected the Computing Element (CE) hosting those data. The LCG information system was used by the Resource Broker to find out the information about the available Grid resources (Computing Elements and Storage Elements). A Resource Broker and an Information System reserved for CMS were setup at CERN.

CMS could dynamically add or remove resources as needed. The jobs ran on Worker Nodes, performing the following operations: establish a CMS environment, including access to the pre-installed ORCA; read the input data from a Storage Element (using the rfio protocol whenever possible otherwise via LCG Replica Manager commands); execute the user-provided executable; store the job output on a data server; and register it to the RLS to make it available to the whole Collaboration.

The automated analysis ran quasi-continuously for two weeks, submitting a total of more than 15,000 jobs, with a job completion efficiency of 90–95%. Taking into account that the number of events per job varied from 250 to 1000, the maximum rate of jobs, ~260 jobs/hour, translated into a rate of analysed events of about 40 Hz. The LCG submission system could cope very well with this maximum rate of data coming from CERN. The Grid overhead for each job, defined as the difference between the job submission time and the time of start execution, was on average around 2 minutes. An average latency of 20 minutes between the appearance of the file at CERN and the start of the analysis job at the remote sites was measured during the last days of DC04 running.

6.1.3.5 DC04 Monitoring

MonaLisa and GridICE were used to monitor the distributed analysis infrastructure, collecting detailed information about nodes and service machines (the Resource Broker, and Computing and Storage Elements), and were able to notify the operators in the event of problems. CMS-specific job monitoring was managed using BOSS. BOSS extracts the specific job information to be monitored from the standard output and error of the job itself and stores it in a dedicated MySQL database. The job submission time, the time of start and end execution, the executing host are monitored by default. The user can also provide to BOSS the description of the parameters to be monitored and the way to access them by registering a job-type. An analysis-specific job-type was defined to collect information like the number of analysed events, the datasets being analysed.

6.1.3.6 CMS DC04 Summary

About 100 TB of simulated data in more than 700,000 files have been produced, during the pre-production phase, corresponding to more than 400 kSPECint2000 years of CPU. Data

have been reconstructed at the Tier-0, distributed to all Tier-1 centres and re-processed at those sites at a peak rate of 25 Hz (4MB/s output rate). This rate was kept only for limited amount of time (one day); nevertheless the functionality of the full chain was demonstrated. The main outcomes of the challenge were:

- the production system was able to cope with an heterogeneous environment (local, Grid3 and LCG) with high efficiency in the use of resources
- local reconstruction at the Tier-0 could well cope with the planned rate; some overload of the CERN CASTOR stager was observed
- a central catalogue implemented using the LCG RLS, managing at the same time location of files and their attributes was not able to cope with the foreseen rate
- the data transfer system was able to cope with the planned rate and to deal with multiple point-to-point transfer systems
- the use of the network bandwidth was not optimal owing to the small size of the files
- the use of MSS at the Tier-1 centres was limited by the big number of files of small size it had to deal with; only about 1/3 of the transferred data was safely stored on Tier-1's MSS
- quasi-real-time analysis at the Tier-1 centres could cope well with the planned rate; a median latency of ~20 minutes was measured between the appearance of the file at CERN and the start of the analysis job at remote sites.

The main issues addressed after the end of DC04 are the optimization of file sizes and the re-design of the data catalogues.

6.1.4 LHCb

In this Section a description of the LHCb use of the LCG Grid during Data Challenge'04 is outlined. The limitations of the LCG at the time and the lessons learnt are highlighted. We also summarize the baseline services that LHCb need in LCG in order for the data to be processed and analysed in the Grid environment in 2007. The detailed implementation of these services within the LHCb environment is described earlier in this document.

6.1.4.1 Use of LCG Grid

The results described in this section reflect the experiences and the status of the LCG during the LHCb data challenge in 2004 and early 2005. The data challenge was divided into three phases:

- Production: Monte Carlo simulation
- Stripping: Event pre-selection
- Analysis

The main goal of the Data Challenge was to stress test the LHCb production system and to perform distributed analysis of the simulated data. The production phase was carried out with a mixture of LHCb dedicated resources and LCG resources. LHCb managed to achieve their goal of using LCG to provide at least 50% of the total production capacity. The third phase, analysis, has yet to commence.

6.1.4.2 Production

The DC04 production used the Distributed Infrastructure with Remote Agent Control (DIRAC) system. DIRAC was used to control resources both at DIRAC dedicated sites and those available within the LCG environment.

A number of central services were deployed to serve the Data Challenge. The key services are:

1. A production database where all prepared jobs to be run are stored
2. A Workload Management System that dispatches jobs to all the sites according to a 'pull' paradigm
3. Monitoring and accounting services that are necessary to follow the progress of the Data Challenge and allow the breakdown of resources used
4. A book-keeping service and the AliEn File Catalogue (FC) to keep track of all datasets produced during the Data Challenge.

Before the production commenced, the production application software was prepared for shipping. It is an important requirement for the DIRAC system to be able to install new versions of the LHCb production software soon after release by the production manager. All the information describing the production tasks is stored in the production database. In principle the only human intervention during the production by the central manager is to prepare the production tasks for DIRAC. The first step of production is the preparation of a workflow, which describes the sequence of applications that are to be executed together with the necessary application parameters. Once the workflow is defined, a production run can be instantiated. The production run determines a set of data to be produced under the same conditions. The production run is split into jobs as units of the scheduling procedure. Each DIRAC production agent request is served with a single job. When new datasets are produced on the worker nodes they are registered by sending a XML dataset description to the book-keeping service. The output datasets are then transferred to the associated Tier-1 and the replica is registered in the book-keeping service.

The technologies used in this production are based on C++ (LHCb software), Python (DIRAC tools), Jabber/XMPP (instant messaging protocol used for reliable communication between components of the central services) and XML-RPC (the protocol used to communicate between jobs and central services). Oracle and MySQL are the two databases behind all of the services. Oracle was used for the production and book-keeping databases, and MySQL for the workload management and AliEn FC systems.

On the LCG, 'agent installation' jobs were submitted continuously. These jobs check if the Worker Node (WN) where the LCG job was placed was configured to run a LHCb job. If these checks were in the affirmative, the job installed the DIRAC agent, which then executed as on a DIRAC site within the time limit allowed for the job, turning the WN into a virtual DIRAC site. This mode of operation on LCG allowed the deployment of the DIRAC infrastructure on LCG resources and uses them together with other LHCb Data Challenge resources in a consistent way.

A cron script submits DIRAC agents to a number of LCG resource brokers (RB). Once the job starts execution on the WN, and after the initial checks are satisfied, the job first downloads (using http) a DIRAC tarball and deploys a DIRAC agent on the WN. A DIRAC agent is configured and executed. This agent requests the DIRAC WMS for a task to be executed. If any task is matched, the task description is downloaded on the WN and executed. The software is normally pre-installed with the standard LCG software installation procedures. If the job is dispatched to a site where software is not installed, then installation is performed in the current work directory for the duration of the job. All data files as well as logfiles of the job are produced in the current working directory of the job. Typically the amount of space needed is around 2 GB plus an additional 500 MB if the software needs to be installed. The book-keeping information (data file 'metadata') for all produced files is uploaded for insertion into the LHCb Book-keeping Database (BKDB) At the end of the reconstruction, the DST file(s) are transferred by GridFTP to the SEs specified for the site, usually an associated Tier-1 centre. Once the transfer is successful, the replicas of the DST file(s) are registered into the LHCb-AliEn FC and into the replica table of BKDB. Both catalogues were accessed via the same DIRAC interface and can be used interchangeably.

By the end of the production phase, up to 3,000jobs were executed concurrently on LCG sites. A total of 211 k jobs were submitted to LCG, LHCb cancelled 26 k after 24–36 hours in order to avoid the expiration of the proxy. Of the remaining 185 k, 113 k were regarded as successful by the LCG. This is an efficiency of ~61%.

The Data Challenge demonstrated that the concept of light, customizable and simple-to-deploy DIRAC agents is very effective. Once the agent is installed, it can effectively run as an autonomous operation. The procedure to update or to propagate bug fixes for the DIRAC tools is quick and easy as long as care is taken to ensure the compatibility between DIRAC releases and ongoing operations. Up to now over 200 k DIRAC tasks have successfully executed on LCG, corresponding to approximately 60% of the total, with up to 60 different contributing sites and major contributions from CERN and the LHCb proto-Tier-1 centres.

To distribute the LHCb software, the installation of the software is triggered by a running job and the distribution contains all the binaries and is independent of the Linux flavour. Nevertheless, new services to keep track of available and obsolete packages and a tool to remove software packages should be developed.

The DIRAC system relies on a set of central services. Most of these services were running on the same machine that ended up with a high load and too many processes. With thousands of concurrent jobs running in normal operation, the services are approaching a Denial of Service regime, where you have a slow response and with services stalled.

In the future releases of the DIRAC system, the approach to error handling and reporting to the different services will be improved.

As LCG resources were used for the first time, several areas were identified where improvements should be made. The mechanism for uploading or retrieving the output sandbox should be improved, in particular to have information about failed or aborted jobs. The management of each site should be reviewed to avoid and detect that a misconfigured site becomes a ‘black-hole’. The publication of information about site intervention should be also provided to the Resource Broker or to the Computing Element. In particular, both DIRAC and the LCG need extra protection against external failures, e.g., network failures or unexpected system shutdowns.

The adopted strategy of submitting resource reservation jobs to LCG that only request a LHCb task once they are successfully running on a WN has proven to be very effective to protect the LHCb DIRAC production system against problems with LCG WMS. This approach allowed effective separation of the resource allocation (that is left to LCG) from the task scheduling (that is handled by DIRAC). Some improvement on the LCG scheduling mechanism has taking place but still further improvements are essential as far as concerns CPU and local disk space reservation for the jobs.

Another successful approach has been the inclusion, on the same LCG job, of the simulation task, the upload and the registration (including error recovery and retrial mechanisms) of the produced data. This assures that once the job is finished no further actions are needed. Again this has added extra redundancy against errors on the LCG scheduling (at the retrieval of the output sandbox step) that would otherwise have been considered as failed.

Other important lessons are the need for better logging and debugging tools that should allow a more efficient understanding of system misbehaviours, the need for bulk operations for large production activities where thousands of jobs need to be processed every day, and extreme care on the performance of basic commands that must always return (successfully or not) after a reasonable amount of time (simple edg-job-submit or globus-url-copy commands do, under some circumstances, hang for days until they are killed by the user or system administrator).

Running a production over months has shown that every possible hardware piece will eventually fail at some point (from the local disk of a WN to the mirrored load-balanced DB

server or a system administrator accidentally hitting a reset bottom) and all software pieces must be protected against these problems, retrying on alternate servers when possible or returning meaningful error messages otherwise.

6.1.4.3 Organized Analysis

The stripping process consists in running a DaVinci program that either executes the physics selection for a number of channels or selects events that pass the first two levels of trigger (L0+L1). The former will be run on all signal and background events while the latter will be run on minimum-bias events.

The DaVinci applications (including JobOptions files) were packaged as a standard production application such that they can be deployed through the standard DIRAC or LCG software installation procedures. For the handling of the stripping, a database separate from the LHCb Book-keeping Database (BKDB), called the Processing Database (PDB), was used.

Information was extracted from the BKDB based on queries on the type of data. New files were incrementally added to the PDB, upon production manager request, and initially marked as 'created'. This database, is scanned for a given event type with enough data to be stripped. The files are marked as 'grouped' and assigned a Group tag. Jobs are then prepared to run on all files with the same Group tag. The files are then marked as 'prepared.' The JDL of the job contains the logical file names (LFN) of all selected files and from the list of files a GaudiCatalog, corresponding to those files, was created and was shipped in the jobs' input sandbox.

SRM was used as a technology-neutral interface to the mass storage system during this phase of the LHCb data challenge. The original plan was to commence at CERN, CNAF and PIC (CASTOR based sites) before moving to non-CASTOR technologies at other proto-LHCb Tier-1 centres, such as FZK, IN2P3, SARA/NIKHEF and RAL. The SRM interface was installed at CNAF and PIC at the request of LHCb and we were active in aiding debugging these implementations.

The Grid File Access Library (GFAL) APIs were modified for LHCb to allow some of the functionality requirements described above to be available. The motivation of using GFAL was to hide any SRM implementation dependencies, such as the version installed at a site. From these APIs LHCb developed a number of simple command line interfaces. In principle the majority of the functionality required by LHCb was described in the SRM (version 1.0) documentation, unfortunately the implementation of the basic SRM interfaces on CASTOR did not match the functional design. Below we describe the missing functionality and a number of *ad hoc* solutions was used.

The inability to pin/unpin or mark file for garbage collection means it is possible that files for a SRM request are removed from the disk pool before being processed. A number of temporary solutions were considered:

- Throttle the rate at which the jobs were submitted to a site. This would be a large overhead for the production manager and needs detailed knowledge of the implementation of the disk pools at all sites. It also assumes that the pool in use is only available to the production manager; this is not the case. SRM used the default pool assigned to the mapped user in the SRM server.
- Each time a file status is checked, a new SRM request is issued. This protected against a file being 'removed' from the disk pool before being processed but it was not clear what the effect had on the staging optimization. This was the solution adopted.
- Use of technology-specific commands to (pin and) remove the processed file from disk. This assumes that such commands are available on the worker nodes (not always the case) and an information service that maps a site with a technology.

Originally there was no control over the stage pool being used. It is highly desirable to have separate pools for production activities and user analysis jobs to remove any destructive interference. Mapping the production users in a VO to a particular user account solved this problem but this required intervention at the LCG system level.

The stripping concept was proven by running on the LXBATCH system at CERN (but with submission through DIRAC.) This approach made making use of technology (CASTOR) specific stage commands. Over 20 million events were processed through the stripping with over 70 concurrent jobs running on this single site. Work has started to re-use SRM through LCG for this phase.

6.2 Service Challenges

So as to be ready to fully exploit the scientific potential of the LHC, significant resources need to be allocated to a series of Service Challenges. These challenges are an essential on-going and long-term commitment to achieving the goal of a production-quality worldwide Grid at a scale beyond what has previously been achieved.

Whilst many of the individual components that make up the overall system are understood or even deployed and tested, much work remains to be done to reach the required level of capacity, reliability, and ease-of-use. These problems are compounded not only by the inherently distributed nature of the Grid, but also by the need to get large numbers of institutes and individuals, all with existing, concurrent and sometimes conflicting commitments, to work together on an incredibly aggressive time scale.

The service challenges must be run in an environment that is as realistic as possible, which includes end-to-end testing of all key experiment use-cases over an extended period, demonstrating that the inevitable glitches and longer-term failures can be handled gracefully and recovered from automatically. In addition, as the service level is built up by subsequent challenges, they must be maintained as stable production services on which the experiments test their computing models.

The first two challenges — December 2004 and March 2005 — focused on the basic infrastructure and involved neither the experiments nor Tier-2 sites. Nevertheless, the experience from these challenges proved extremely useful in building up the services and in understanding the issues involved in offering stable production services around the clock for extended periods.

During the remainder of 2005, the Service Challenges will expand to include all the main offline use cases of the experiments apart from analysis and will begin to include selected Tier-2 sites. Additional components over the basic infrastructure will be added step by step, including experiment-specific solutions. It is important to stress that each challenge includes a set-up period, during which residual problems are ironed out, followed by a period that involves the experiments but during which the focus is on the ‘service’, rather than any data that may be generated and/or transferred (that is, the data are not necessarily preserved and the storage media may be periodically recycled). Finally, there is an extended service phase designed to allow the experiments to exercise their computing models and software chains.

The workplan continues to evolve with time: the current status including completed and future milestones is maintained in the LCG Project planning website [97].

6.2.1 Network Workplan

The network workplan is described elsewhere in this document. As far as the service challenges are concerned, the principle requirement is that the bandwidth and connectivity between the various sites should be consistent with the schedule and goals of the service challenges. Only modest connectivity is required between Tier-2 sites and Tier-1s during 2005, as the primary focus during this period is on functionality and reliability. However, connections of 10 Gb/s are required from CERN to each Tier-1 no later than the end of 2005.

Similarly, connectivity between the Tier-1s at 10 Gb/s is also required by summer 2006 to allow the analysis models to be fully tested. Tier-1–Tier-2 connectivity of at least 1 Gb/s is also required on this time scale, to allow both Monte Carlo upload and analysis data download.

6.2.2 Security Service Challenges

A number of security service challenges will be performed during the preparation for LHC start-up. These will test the various operational procedures, e.g., security incident response, and also check that the deployed Grid middleware is producing audit logs with appropriate detail. One important aim of these challenges is to ensure that ROC managers, site managers and security officers understand their responsibilities and that audit logs are being collected and maintained according to the agreed procedures. Experience from the service challenges and real security incidents, as and when they happen, will be used both to improve the content of the audit logs and the incident handling procedures and also to drive future security service challenges.

6.2.3 Results of Service Challenge 1 and 2

Service Challenge 1 was scheduled to complete in December 2004, demonstrating sustained aggregate 500 MB/s mass store to mass store between CERN and three Tier-1 sites. A data rate of 500 MB/s was sustained between FNAL and CERN for three days in November. The sustained data rate to SARA(NIKHEF) in December was only 54 MB/s., but this had been pushed up to 200 MB/s by the start of SC2 in mid-March. A data rate of 500 MB/s was achieved in January with FZK. Although the SC1 goals were not achieved a great deal was learned at CERN and other sites, and the experience was an important contribution towards achieving the SC2 goals.

Service Challenge 2 started on 14 March 2005. The goal was to demonstrate 100 MB/s reliable file transfer between CERN and seven Tier-1s (BNL, CNAF, FNAL, FZK, IN2P3, NIKHEF and RAL), with one week at a sustained aggregate throughput of 500 MB/s at CERN.

A regular series of monthly service challenge progress meetings had been held, together with weekly phone conferences to track progress and address technical issues. A set of software scripts, using a database to hold and schedule transfer requests was put in place to manage the transfers via GridFTP. Several sites, however, used SRM-SRM copies for the transfers. Both mechanisms were tested and used in the challenge itself. SARA/NIKHEF provided a set of tools to monitor the transfers and throughputs. The sites involved, other than CERN, were BNL, CNAF, FNAL, GridKa, Lyon, RAL, and SARA/NIKHEF.

The challenge was successful with all goals met: the transfers ran for 11 days, achieving around 600MB/s on average, with peaks above 800MB/s sustained for several hours. There were some service outages, but these were understood and fixed, with the service rapidly recovering.

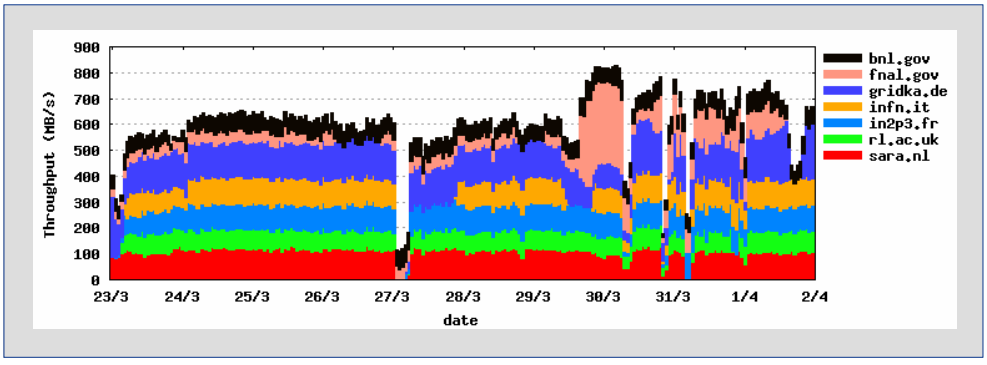


Figure 6.2: Throughput to Tier-1 sites from CERN (hourly averages)

6.2.4 Goals of Service Challenge 3

In terms of file transfer services and data rates, the goals of Service Challenge 3, to start in July 2005, are to demonstrate reliable transfers at rates of 150 MB/s per Tier-1 managed disk to managed disk and 60 MB/s to managed tape. The total aggregate data rate out of CERN that should be achieved is 1 GB/s. All Tier-1 sites will participate in this challenge, although a small number will not have the necessary network bandwidth installed to achieve the target data rate above. However, they will nevertheless be involved in testing the basic infrastructure and gaining experience with the corresponding services. A small number of Tier-2 sites will also be involved, focusing on those with good local support, both at the level of the required infrastructure services and from the relevant experiment. The file transfer goals between Tier-2 and Tier-1 sites are to show sustained transfers using 1 GB files of ~3 files / hour Tier-2 to Tier-1 over several days. These tests are scheduled for the end of July. In addition to building up the data rates that can be supported at both CERN and outside sites, this challenge will include additional components, such as catalogues, support for multiple VOs, as well as experiment-specific solutions. It is foreseen that the challenge will start with a phase that demonstrates the basic infrastructure, albeit with higher data rates and more sites, including selected Tier-2s. Subsequently, the data flows and access patterns of the experiments will be tested, initially by emulating the models described in the Computing Model documents and subsequently by running the offline frameworks themselves. However, during both of these phases the emphasis will be on the Service, rather than the Data, which will not normally be preserved. Finally, an extended Service Phase is entered, currently foreseen from September 2005 until the end of the year, during which the experiments validate their computing models using the facilities that have been built up during the Service Phase.

6.2.5 Service Challenge 3 Planning

SC3 consists of a set-up phase starting on 1st July 2005, during which a number of throughput tests will be performed, followed by a Service Phase from 1 September 2005 until the end of the year.

All data management components for SC3 were delivered ready for production by the end of May 2005.

Final testing and integration of these components and services must be completed by the end of June 2005.

The primary responsibility of the participating sites at the infrastructure level is to provide a conformant SRM 1.1 interface to their managed storage. A reliable file transfer service is being set up based on the gLite File Transfer Service (FTS) at CERN and is foreseen at those Tier-1s that will support Tier2s during the throughput tests of the set-up phase (see Table 2.1). A service based on the LCG File Catalogue (LFC) will be provided at CERN for ATLAS, CMS and LHCb. ATLAS and CMS require a local file catalogue at all sites, LHCb would like read-only replicas for reasons of availability at two external sites.

A detailed planning schedule is available in Ref. [98].

The primary sites that will participate in SC3 are the Tier-0 (CERN) and the following Tier1 sites: ASCC, BNL, CCIN2P3, CNAF, FNAL, GridKA, SARA/NIKHEF, PIC, RAL and TRIUMF.

The Nordic Data Grid Facility is expected to exercise file transfers but has not yet committed to participate in the Throughput Phase of the service challenge.

A restricted number of Tier2 sites will also participate. The names of these sites will be decided in agreement with the Tier1 site that will support them in terms of File Transfer and Storage services. The following Tier1 sites have stated that they will participate in this component of the challenge: BNL, CNAF, FNAL, PIC and RAL. CMS transfers in all cases will be driven by PhEDEx.

Table 6.1: The list of known Tier2 sites that will participate in SC3.

Site	Tier-1	Experiment
Legnaro, Italy	CNAF, Italy	CMS
Milan, Italy	CNAF, Italy	ATLAS
Turin, Italy	CNAF, Italy	Alice
DESY, Germany	FZK, Germany	ATLAS, CMS
CMS Tier-2, Spain	PIC, Spain	CMS
Lancaster, UK	RAL, UK	ATLAS
Imperial, UK	RAL, UK	CMS
Edinburgh, UK	RAL, UK	LHCb
US Tier2s	BNL / FNAL	ATLAS / CMS
U. Chicago	BNL	ATLAS
Boston (to be confirmed)	BNL	ATLAS
Florida	FNAL	CMS
Caltech	FNAL	CMS
UCSD	FNAL	CMS
Wisconsin	FNAL	CMS
Purdue (to be confirmed)	FNAL	CMS

6.2.6 Goals of Service Challenge 4

Service Challenge 4 needs to demonstrate that all of the offline data processing requirements expressed in the experiments' Computing Models, from raw data taking through to analysis, can be handled by the Grid at the full nominal data rate of the LHC. All Tier-1 sites need to be involved, together with the majority of the Tier-2s. The challenge needs to successfully complete at least 6 months prior to data taking. The service that results from this challenge becomes the production service for the LHC and is made available to the experiments for final testing, commissioning and processing of cosmic-ray data. In parallel, the various centres need to ramp up their capacity to twice the nominal data rates expected from the production phase of the LHC, to cater for backlogs, peaks and so forth. The analysis involved is assumed to be batch-style analysis, rather than interactive analysis, the latter is expected to be performed primarily 'off the Grid'. The total aggregate data rate out of CERN that needs to be supported in Sevice Challenge 4 is 1.6 GB/s to tape at the Tier-1s. The final service must be capable of running at twice this data rate.

6.3 ARDA

The ARDA project has been set up to investigate the area of distributed analysis together with the LHC experiments. Owing to the relative immaturity of this field at the time the project was started, it was decided to use end-to-end prototypes as tools to investigate the field while making real progress in contributing to the experiments' distributed analysis programs. The second goal of ARDA is to influence the evolution of the gLite middleware during the prototype activity.

The project was started following the recommendation of the LCG ARDA RTAG [99] and the subsequent ARDA workshop at CERN in January 2005.

Experience shows that additional functionality is expected from the Grid to fulfil the needs of various application domains. Following a common pattern in the evolution of computing technologies, the usage itself will stimulate further developments.

All the experiments currently have to provide their own (evolving) layer of high-level services to satisfy their requirements. Examples are:

- The impact on the generic services (in terms of security, scheduling and accounting policies) is not yet clear for the new experiment-specific services: for example for the interactivity services.
- A second-generation of book-keeping and work-flow systems will emerge from the experience of the complex data challenges with the needs and the peculiarities of the real data handling and user analysis.
- The interaction of users with the Grid is still an open field, with promising prototypes being exposed to the user community.
- The success of Grid technology will need very large communities using it on a daily basis on personal laptops and workstations. The Grid software must be integrated with the collaborative tools in daily use in the LHC community.

These *services* cannot be delivered outside the applications because they are still in the evolution phase. The model is not to develop a 'standard' solution and then implement it but to provide concrete implementations of useful services to later derive the general features and maybe propose them as a standard (e.g., CMS Phedex and gLite FTS/FPS). This approach will continue to make progress possible because these experiment-specific services are concrete instantiations of useful services and should therefore be supported following the model of LCG ARDA.

6.3.1 ARDA/ALICE End-To-End Prototype

The ALICE experiment is providing an analysis platform using the AliROOT framework based on ROOT.

A Grid-enabled version of ROOT is under development to allow a user-transparent way of analysis on the local machine or in a Grid environment. Users can select between a classical batch analysis style and interactive analysis using PROOF — the parallel ROOT facility. The user procedures should be kept identical for both analysis modes.

Within ARDA, ALICE is developing a generic C++ client/server application interface connecting to the Grid middleware services, the implementation of a generic Grid class and Grid plug-ins in the ROOT framework. The goals of this implementation are high performance, high security and to allow the users fast iterative prototyping of analysis applications.

The PROOF infrastructure together with the underlying Grid middleware provide the back-end of the ALICE system. With respect to PROOF a new connectivity scheme for the master/slave architecture is under development to cope with the restriction of outgoing (or proxy) connectivity from slave hosts to central master hosts.

The Grid middleware is used to allow the system to run on many distributed resources via standard access methods and agreed procedures (access control, accounting, access to CPU and data resources).

The C++ API access library allows, moreover, the provision of Grid commands inside a standard user shell. The current implementation is general enough to be considered of interest to the other experiments and middleware projects.

The system has been successfully presented several times both inside the ALICE Collaboration and to conferences (e.g., Super Computing 2004).

6.3.2 ARDA/ATLAS End-To-End Prototype

The ATLAS strategy for distributed analysis includes different systems. It includes the investigation of high-level services to set up dedicated analysis facilities at a site and as a

general way to access Grid resources (e.g., DIAL) and the production system (used to run the large data challenges on three different Grids).

Within the ARDA/ATLAS activity, several different components have been scrutinized collaborating with the different developers (DIAL, AMI, Production system, GANGA, users accessing data from the common test beam, from the recent data challenges etc…). Examples are: DIAL services have been demonstrated using the gLite prototype as backend; The production system has been studied, contributing to different parts (notably integrating gLite into the Don Quijote framework); Exploratory activity on executing user applications (Athena) in parallel is promising. The next goal is to provide an integration of the different component in a coherent system (benefiting from the activity on GANGA in the ARDA/LHCb prototype as well).

The ARDA contributions have been presented several times both inside the ATLAS collaboration and to conferences (e.g., CHEP 2004).

6.3.3 ARDA/CMS End-To-End Prototype

ARDA/CMS prototype activity explores the potential of the gLite middleware by delivering an end-to-end prototype and providing specific tests on gLite components of possible interest to CMS.

The activity focuses on contributing to the evolution of the RefDB/PubDB layer to open up the possibility for users to effectively perform their analyses in a distributed environment. In addition, ARDA/CMS investigates the potential of the LCG and gLite file catalogues. The key activity is to expose the prototype to end users and to evolve it according to the feedback and the Grid middleware functionality.

The CMS prototype (ASAP, ARDA Support for CMS Analysis Processing) has been exposed to several users, using it for their analysis activity. The system uses gLite as back-end and it is used (within ARDA) to prototype the different component for the final system. Experience is being discussed with CMS. The LCG2 analysis system in development in CMS is CRAB.

ASAP is used by several CMS physicists performing analysis for their daily work. It provides a simple but effective way to execute their applications on the Grid, benefiting from job splitting, JDL preparation and submission and results' retrieval. The starting point is a working ORCA application the user prepares and runs on a PC or a batch system.

ASAP offers as a key capability a good overview of the tasks submitted by the users — a task is the analysis of (a fraction of) a dataset. Status, log files and exit status are collected from gLite services (mainly logging and book-keeping) and by the Monalisa system and made conveniently available (Web pages). The monitor system is capable of keeping track of the running jobs on behalf of the user (using gLite myProxy)..

The ARDA contributions have been presented several times both inside the CMS Collaboration and at other events (e.g., CHEP 2004, LHCC).

6.3.4 LHCb End-To-End Prototype

The LHCb *end-to-end* prototype for EGEE is based on GANGA ([18]).

GANGA is a common project between ATLAS and LHCb. LHCb decided to expose their users to GANGA as a portal to deal with data and distributed resources. GANGA provides to the user a coherent set of utilities and an isolation layer allowing him/her to move his/her activity from developing the analysis with a few data samples on a personal computer to a distributed environment. The different back-ends — local PC, local batch, Grid — and disconnected operations are transparent for the user.

In parallel, ARDA and LHCb are investigating the potential of Grid metadata catalogues (the ARDA interface has been accepted by the EGEE project and now it is being evolved together with the middleware team).

GANGA has a great potential as common project in ATLAS and LHCb. In particular the recent evolution (GANGA4) is benefiting from a strong commitment from ARDA. The more mature architecture not only addresses the limitations observed in previous versions but also enables the integration of GANGA4 with other frameworks (e.g., sharing and contributing (Python) modules, adding functionality via plug-ins etc.)

The GANGA system has been presented several times both inside the LHCb collaboration and at other events (e.g., EGEE conferences).

6.3.5 Distributed Analysis Using PROOF

The work on the Parallel ROOT Facility, PROOF [100], is accelerating with several new developments. While initially designed as a purely interactive facility for short queries (interactive here means less than a few minutes) the system is now being extended to also support long(er) running, i.e., batch, queries. To be able to support this new mode PROOF must be able to support a so-called stateless mode which allows users to disconnect and later reconnect to retrieve the query results. This batch extension is very attractive since it allows the same 'simple' ROOT TSelector-based analysis model for very large datasets.

Work is ongoing to improve the functionality and robustness of the PROOF system. Analysis objects such as 'friend trees', event lists, and tree indices that were available in a local ROOT analysis session are now supported by PROOF. In addition, PROOF sessions, their status and query results, can be browsed using the ROOT browser. The installation of a PROOF cluster has been simplified to allow nodes able to run PROOF to register themselves with a central resource manager, instead of having to rely on a static cluster configuration file. Developments are also being made to adapt PROOF to use Grid middleware Services in order to distribute the PROOF master and slaves servers over multiple sites and clusters. This involves building interfaces to Grid Job Schedulers to start the PROOF agents on the Grid and to Grid File Catalogs to find the location of the files to be analysed.

Work is continuing on the authentication modules for the xrootd data server developed by SLAC. xrootd is an extensible, modular, robust, and scalable data server that has been specially optimized to serve ROOT files (any type of file can be served via a POSIX file access layer). The xrootd infrastructure will be extended to act as a PROOF front-end server (xproofd). A first prototype is looking very promising.

7 PLANS

7.1 Baseline Services

A working group met to forge an agreement between the experiments and the LHC regional centres on the baseline services to be provided to support the computing models for the initial period of LHC running. These services must therefore be in operation by September 2006.

The services concerned are those that supplement the basic services for which there is already general agreement and understanding (e.g., provision of operating system services, local cluster scheduling, compilers) and which are not already covered by other LCG groups.

The final report [101] of the working group is available.

7.2 Phase-2 Planning

This section summarizes the high-level plan for the deployment and commissioning of the LHC Computing Grid. Detailed plans are developed for each of these major activities by the groups and centres concerned. The overall planning at the project level includes two service challenges to co-ordinate the ramp-up of the Grid to the capacity and performance required for LHC sustained operation. The initial service is scheduled to be in operation, including all of the Tier-1 centres, a full six months before the first beams. The service must have demonstrated the capability of continuous operation at the full capacity and performance required in 2007 at least three months before the first collisions.

Table 7.1: Service Challenge milestones

Date	Description
31 July 05	**Service Challenge 3 Set-up:** Set-up complete and basic service demonstrated. Performance and throughput tests complete. See Section 6.2.4 for detailed goals.
1 Sept 05	**Service Challenge 3**: start of stable service phase, including at least 9 Tier-1 and 10 Tier-2 centres.
31 Dec 05	**Tier-0/1 high-performance network operational** at CERN and 8 Tier-1s.
31 Dec 05	**750 MBytes/s data recording demonstration at CERN:** Data generator→ disk → tape sustaining 750 MBytes/s for one week using the CASTOR mass storage system.
28 Feb 06	**All required software for baseline services deployed** and operational at all Tier-1s and at least 20 Tier-2 sites.
30 Apr 06	**Service Challenge 4 Set-up:** Set-up complete and basic service demonstrated. Performance and throughput tests complete: Performance goal for each Tier-1 is the nominal data rate that the centre must sustain during LHC operation (see Table 7.2 below) CERN-disk → network → Tier-1-tape. Throughput test goal is to maintain for three weeks an average throughput of 1.6 GB/s from disk at CERN to tape at the Tier-1 sites. All Tier-1 sites must participate. The service must be able to support the full computing model of each experiment, including simulation and end-user batch analysis at Tier-2 centres.
31 May 06	**Service Challenge 4:** Start of stable service phase, including all Tier-1s and 40 Tier-2 centres.
30 Sept 06	**1.6 GB/s data recording demonstration at CERN:** Data generator→ disk → tape sustaining 1.6 GB/s for one week using the CASTOR mass storage system.
30 Sept 06	**Initial LHC Service in operation:** Capable of handling the full target data rate between CERN and Tier-1s (see Table 7.2). The service will be used for extended testing of the computing systems of the four experiments, for simulation and for processing of cosmic-ray data. During the following six months each site will build up to the full throughput needed for LHC operation, which is twice the nominal data rate.
1 Apr 07	**LHC Service Commissioned**: A series of performance, throughput and reliability tests completed to show readiness to operate continuously at the target data rate and at twice this data rate for sustained periods.

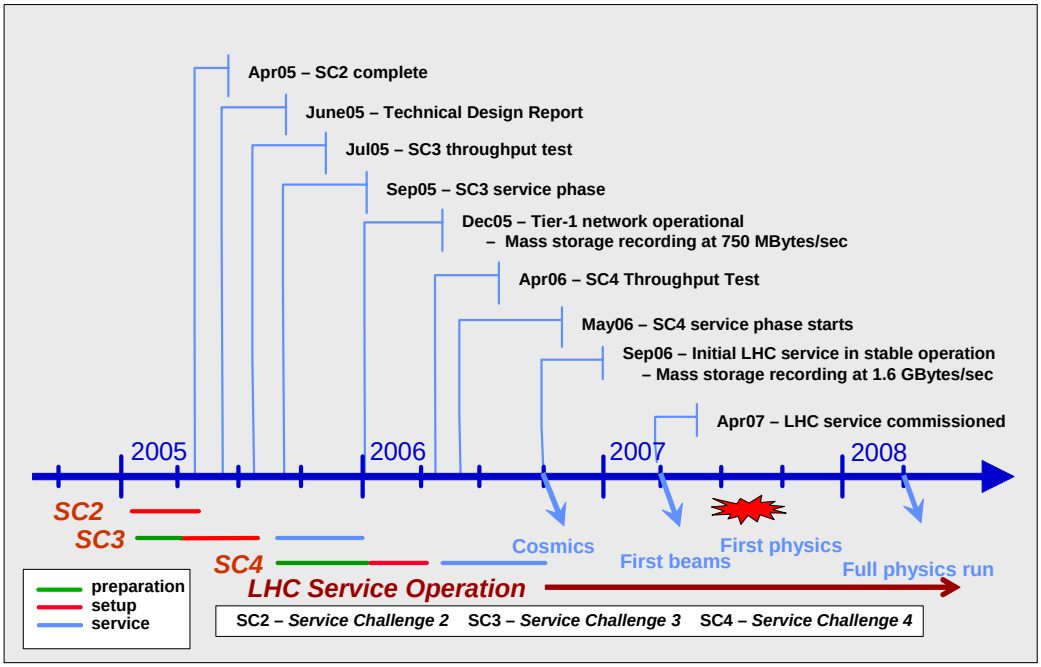


Figure 7.1: LHC Grid deployment schedule 2005 – 2008

Table 7.2 Target data rates for CERN and Tier-1 centres in SC4

Centre	ALICE	ATLAS	CMS	LHCb	Target Data Rate MBytes/s
ASCC		X	X		110
CNAF	X	X	X	X	220
PIC		X	X	X	200
CC-IN2P3	X	X	X	X	220
GridKA	X	X	X	X	220
RAL	X	X	X	X	220
BNL		X			65
FNAL			X		50
TRIUMF		X			65
SARA/NIKHEF	X	X		X	175
NDGF	X	X	X		90
Target data rate at CERN					1,600

Note that the *target data rate* is the data rate that must be sustained continuously during the normal operation of the LHC machine. These targets must be demonstrated during SC4. The Grid and its component Tier-0 and Tier-1 centres must be capable of sustained operation at twice this rate to allow for catching up after service interruptions, and to be able to absorb locally generated load from the Tier-1 or Tier-2 centres. A working group is at present

studying the computing model papers with a view to improving the estimate of Tier-1 I/O loads.

Further information on detailed planning can be found in the following documents:

- LCG planning page [102]
- Applications area plan [103]
- Service challenge planning paper [104].

8 PROJECT ORGANIZATION AND MANAGEMENT

The organizational structure described in this section is that which is defined in the MoU and will take effect when Phase 2 of the LCG Project is approved. A schematic view is shown in Figure 3.1.

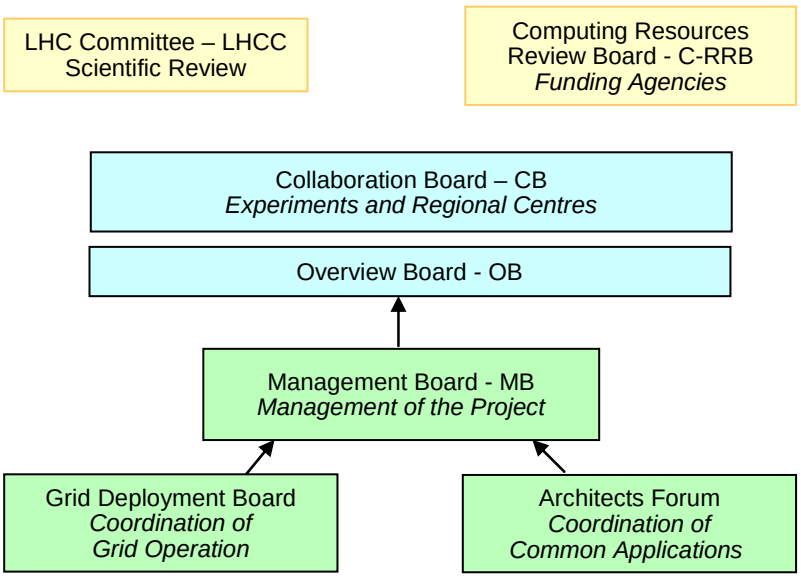


Figure 8.1: LCG organization

8.1 High-Level Committees:

8.1.1 Collaboration Board and Overview Board

Concerning its main technical directions, the Collaboration is governed by the Collaboration Board (CB). The CB is composed of a representative of each Institution or federation of Institutions that is a Member of the Collaboration, the LCG Project Leader and the Spokespersons of each LHC Experiment, with voting rights; and the CERN Chief Scientific Officer (CSO), and CERN/IT and CERN/PH Department Heads, as *ex-officio* members without voting rights, as well as a Scientific Secretary. The CB elects the Chairperson of the CB from among its Members. The CB meets annually and at other times as required.

A standing committee of the CB, the Overview Board (OB), has the role of overseeing the functioning of the Collaboration. It also acts as a clearing-house for conflicts that may arise within the Collaboration. The OB is chaired by the CERN CSO. Its other members include one person appointed by the agency/agencies that funds/fund each of the Tier-1 centres, the Spokespersons of the LHC Experiments, the LCG Project Leader, the CERN/IT and CERN/PH Department Heads, and a Scientific Secretary. It meets about four times per year.

Both the CB and the OB may co-opt additional non-voting members as they deem necessary. The non-voting members complement the regular members by advising on, for example, matters concerning the environment in which the Collaboration operates or specialist aspects within their areas of expertise.

8.1.2 LHC Computing Grid Management Board

The work of the Collaboration is organized and managed as the LHC Computing Grid Project. The Management Board (MB) supervises the work of the Project. It is chaired by the LCG Project Leader and reports to the OB. The MB organizes the work of the Project as a set

of formal activities. It maintains the overall programme of work and all other planning data necessary to ensure the smooth execution of the work of the Project. It provides quarterly progress and status reports to the OB. The MB endeavours to work by consensus but, if this is not achieved, the LCG Project Leader shall make decisions taking into account the advice of the Board. The MB membership includes the LCG Project Leader, the Technical Heads of the Tier-1 centres, the leaders of the major activities managed by the Board, the Computing Co-ordinator of each LHC Experiment, the Chair of the Grid Deployment Board (GDB), a Scientific Secretary and other members as decided from time to time by the Board.

8.1.3 Grid Deployment Board

The Grid Deployment Board (GDB) is the forum within the Project where the computing managements of the experiments and the regional computing centres discuss and take, or prepare, the decisions necessary for planning, deploying, and operating the LHC Computing Grid. Its membership includes: as voting members — one person from each country with a regional computing centre providing resources to an LHC experiment (usually a senior manager from the largest such centre in the country), a representative of each of the experiments; as non-voting members — the Computing Co-ordinators of the experiments, the LCG Project Leader, and leaders of formal activities and projects of the Collaboration. The Chair of the GDB is elected by the voting members of the board for a two-year term. The GDB may co-opt additional non-voting members as it deems necessary. The GDB reports to the LCG Management Board which normally meets immediately after the GDB ratifying the decisions prepared by the GDB.

8.1.4 Architects Forum

The Architects Forum manages the work of the Applications Area of the Project. See section 5.8.

8.1.5 LHCC

Concerning all technical matters, the Project is subject to review by the Large Hadron Collider experiments Committee (LHCC), which makes recommendations to the CERN Research Board (RB).

8.1.6 C-RRB

Concerning all resource and legal matters, the Collaboration is subject to the Computing-Resource Review Board (C-RRB). The C-RRB is chaired by CERN's Chief Scientific Officer. The C-RRB membership comprises a representative of each Funding Agency, with voting rights, and *(ex-officio)* members of the LHC Computing Grid Management and CERN Management, without voting rights.

The LCG Project Leader represents the Collaboration to the outside and leads it in all day-to-day matters. He/she is appointed by the CERN Director General in consultation with the CB.

8.2 Participating Institutes

The Tier-0 centre and analysis facility at CERN will be used by all experiments. The assignment of Tier-1 centres to experiments is given in Table 8.1.

Table 8.1: Tier-1 centres

Centre	Experiments served with priority			
	ALICE	ATLAS	CMS	LHCb
TRIUMF, Canada		X		
GridKA, Germany	X	X	X	X
CC, IN2P3, France	X	X	X	X
CNAF, Italy	X	X	X	X
SARA/NIKHEF, NL	X	X		X
Nordic Data Grid Facility (NDGF)	X	X	X	
ASCC, Taipei		X	X	
RAL, UK	X	X	X	X
BNL, US		X		
FNAL, US			X	
PIC, Spain		X	X	X

The list of more than 100 Tier-2 centres is still expanding. The current version is maintained at a website [105].

8.3 Interactions and Dependencies

The success of the Worldwide LCG Collaboration will depend on close cooperation with several major publicly funded projects and organizations, for the provision of network services, specialized Grid software, and the management and operation of Grid infrastructure. These three areas are considered separately in this section. In the case of Grid software and infrastructure it is expected that the situation will evolve rapidly during the period of construction and commissioning of the LHC computing facility, and so the LCG Project will have to remain flexible and review support and collaboration agreements at frequent intervals.

8.3.1 Network Services

In most cases the network services used to interconnect the regional computing centres participating in the LHC Computing Grid will be provided by the national research networks with which the centres are affiliated and, in the case of European sites, the pan-European backbone network, GÉANT. The architecture of these services is described in Section 4.5.4. While LCG is one of the many application domains served by these general purpose research networks it will, during the early years of LHC, be one of the most demanding applications, particularly between CERN, the Tier-1 and major Tier-2 centres. The formal service agreements will be made directly between the computing centres and the national research network organizations. However, in order to ensure that the individual service agreements will provide a coherent infrastructure to satisfy the LHC experiments' computing models and requirements, and that there is a credible solution for the management of the end-to-end network services, an informal relationship has been established between the major centres and research networks through the *Tier-0/1/2 Networking Group*, a working group of the Grid Deployment Board. It is expected that this group will persist throughout 2006 while the various components of the high-bandwidth infrastructure are brought into full operation. At this stage it is not clear what, if any, special relationship will be required between LCG and the research networks after this point.

8.3.2 Grid Software

The Grid software foreseen to be used to provide the Grid infrastructure for the initial LCG service has been developed by a number of different projects. Some of these are no longer in operation, some have funding for only a limited period, while others have longer-term plans. In the case of software developed by individual institutes, or by projects that have ceased operation, bilateral support agreements have generally been made between LCG and the

developers, with different levels of formality according to the complexity of the software involved. There are several cases, however, where it is necessary to have more complex relationships.

8.3.2.1 Globus, Condor and the Virtual Data Toolkit

Key components of the middleware package used at the majority of the sites taking part in LCG have been developed by the Globus [55] and Condor [56] projects. These are long-term projects that continue to evolve their software packages, providing support for a broad range of user communities. It is important that LCG maintains a good working relationship with these projects to ensure that LHC requirements and constraints are understood by the projects and that LCG has timely information on the evolution of their products. At present there are two main channels for this: key members of Globus and Condor take part in the Open Science Grid [9] and in the middleware development activity of the EGEE project [8]. Both of these projects and their relationships to LCG are described below.

The Virtual Data Toolkit (VDT) [57] group at the University of Wisconsin acts as a delivery and primary support channel for Globus, Condor and some other components developed by projects in the US and Europe. At present VDT is funded by the US National Science Foundation to provide these services for LCG. It is expected that this or a similar formal relationship will be continued.

8.3.2.2 The gLite Toolkit of the EGEE Project

The EGEE project (Enabling Grids for E-sciencE) [8] is funded on a 50% basis by the European Union to operate a multiscience Grid built on infrastructure developed by the LCG Project and an earlier EU project called DataGrid [106]. EGEE includes a substantial middleware development and delivery activity with the goal of providing tools aimed at the High Energy Physics and Biomedical applications. This activity builds on earlier work of the DataGrid and AliEn [107] projects and includes participation of the Globus and Condor projects. The EGEE project and the LCG Collaboration are closely linked at a management level: the middleware activity manager and the technical director of EGEE are members of the LCG Project Execution Board; the EGEE project director is a member of the LCG Project Oversight Board; the LCG Project leader is a member of the EGEE project management board. The EGEE project also provides some funding for the support of applications using the EGEE developed software.

8.3.2.3 The NorduGrid Project

NorduGrid [53] was initiated in 2001 by researchers at Scandinavian and Finnish academic institutes, with the goal of building a Grid-based computing infrastructure in the Nordic countries. The NorduGrid Collaboration has developed the ARC (Advance Resource Connector) Grid middleware, which is deployed in the pilot Nordic Data Grid Facility (NDGF). Although the NorduGrid project was initiated by the experimental High-Energy Physics community in the Nordic Countries, a growing number of scientists from other fields are now using the NDGF with the ARC software as their primary source of computer power and storage capacity.

The formal relationship with LCG will be through the NDGF.

8.3.3 Grid Operational Groupings

The computing resources will be committed by funding agencies through a Memorandum of Understanding, which will specify the capacity to be provided at individual centres. These centres are organized into three major operational groupings: the EGEE Grid, the Open Science Grid, and the Nordic Data Grid Facility. Each of these groups uses a specific base set of middleware tools and has its own Grid operations infrastructure. The body governing the overall operational policy and strategy for the LHC project is the Grid Deployment Board (GDB). This has national representation, usually from the major centre(s) in each country. The GDB will agree on the basic services to be provided at each centre with the aim of

providing a consistent environment for each experiment across the different operational groupings.

8.3.3.1 The EGEE Grid

This group is an evolution of the centres that took part in the DataGrid project, expanded during 2003–04 to include other centres involved in the LCG Project and centres receiving funding from or associated with the EGEE project. The EGEE Grid has at present over 130 centres, including all of the centres serving LCG in the participating countries (with the exception of the United States and the Nordic countries). This Grid includes many national Grid organizations with their own administrative and management structure, but all of the entities involved agree to install the same base middleware and cooperate in Grid operations.

The operational infrastructure at present receives, in Europe, important support from the EGEE project for *Core Infrastructure Centres* and *Regional Operations Centres*, but the infrastructure is also supported by significant national contributions in Europe, Asia and Canada. The centres that are partners in EGEE have contracts with the EU to provide these infrastructure and operations services. The centres involved in LCG will commit to provide the services through the LCG MoU.

The operation is managed at present by the LCG Grid Deployment Area manager, who also holds the position of operations manager of the equivalent activity of EGEE. This may cause some confusion, especially at those sites that are not members of both the LCG and EGEE projects, and could lead to potential conflicts because LCG and EGEE have different, though not incompatible, goals. The LCG Grid Deployment Board (GDB) at present serves as an effective organ for operations policy and strategy in this overlapping LCG/EGEE environment, which has thus far, through its national representation, been able to represent interests of computing centres outside of the physics community. The long-term idea is that EGEE will evolve into an organization that will provide core operation for a science Grid in Europe and perhaps further afield, rather akin to the role of GÉANT in research networking. However, the EGEE project is at present funded only until March 2006. It is expected that the project will be extended for a further period of two years, which means that it would terminate at the beginning of the first full year of LHC operation. It is therefore important that LCG maintains its role in the core operation, and prepares a fall-back plan in the event that the EU-subsidized evolution beyond EGEE does not materialize or does not fulfil LCG's requirements. This is clearly a difficult strategy, with significant risk, but the long-term advantages of a multiscience Grid infrastructure receiving significant non-HEP funding must be taken into account.

8.3.3.2 The Open Science Grid

The Open Science Grid is a common production Grid infrastructure built and maintained by the members of the Open Science Grid Consortium for the benefit of the users. Members of the consortium have agreements to contribute resources and the Users, who are members of the participating VOs, agree to abide by simple policies.

The US LHC programmes contribute to and depend on the Open Science Grid.

The US LHC signs both LCG MoUs and agreements with the OSG Consortium for the provision of resources and support. The OSG includes

- a common, shared, multi-VO national Grid infrastructure which interoperates with other Grid infrastructures in the US and internationally,
- a common Operations organization distributed across the members,
- the Publication of common interfaces and capabilities, and reference implementations of core and baseline services on the OSG.

OSG activities are co-ordinated through a series of Technical Groups each addressing a broad technical area and Activities with deliverables and developments. The OSG is operated by a distributed set of Support Centres operating through agreements and contracts in support of the infrastructure.

The OSG Consortium Council includes representatives of the LCG and EGEE in non-voting roles. Many of the OSG Technical Groups and Activities collaborate with and work on interoperability with the EGEE infrastructure. The Interoperability Activity has special responsibilities in this area.

The formal relationship between LCG and the US resources will be through the US ATLAS and US CMS Software and Computing Projects and their respective host labs, BNL and FNAL. These projects are represented in the GDB. Agreements on the provision of the basic Grid services and integration with other resources in LCG, both at the level of the operational management and infrastructure, and at the level of resource sharing, will be made in the GDB.

8.3.3.3 The Nordic Data Grid Facility

The Nordic Data Grid Facility (NDGF) [10] pilot project was established in 2003 jointly by the Nordic Natural Science Research Councils, NOS-N, to investigate the possibility of the Nordic countries joining into a common science Grid. The pilot project uses resources known at the NDGF prototype which runs the NorduGrid middleware, ARC. This led to a proposal, currently under international evaluation, to build a large-scale production facility common to all sciences. The facility will use the NorduGrid middleware and be funded from 2006. The NDGF will be able to represent the Nordic countries as one unit towards large international collaborations, including LCG. The Nordic resources available to LHC experiments will be incorporated in the NDGF when it begins production operation in 2006. This will include a distributed Nordic Tier-1 and Tier-2 centres.

An interim NDGF organization has been established and discussions are taking place on how the Nordic LCG resources will be integrated, and the formal relationship between NDGF and LCG.

8.4 Resources

Assuring stable and efficient access to computing resources, which are distributed in a truly global fashion, is an enormous technological undertaking. Its complexity is well documented in the preceding chapters of this document. Aligning the resource planning of over 100 institutions in more than 20 countries, involving more than 25 funding agencies with widely different budgeting procedures is not an easy task.

The Memorandum of Understanding for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid (MoU) [1] LHC Homepage, http://cern.ch/lhc-new-homepage/

[2] defines the procedures to enable funding agencies to pledge resources and follow-up on their delivery. The MoU, through various annexes, lists the Tier-1 and Tier-2 centres represented in the collaboration, defines the service levels to be provided at CERN and the Tier-1 and Tier-2 centres, and gives the resource-planning at each centre over a five year period.

These annexes will be updated at most twice per year for the spring and autumn meetings of the C-RRB. It is anticipated that the resources for the coming year will be committed at each autumn meeting, in time to allow the acquisition and installation to be completed in time for the accelerator running-period.

Only the data contained in the copies of the MoU signed by the funding agencies constitute pledged commitments, but provisional planning data provided by the representatives of the centres are maintained on the Web:

- Current list of Tier-1 centres (Ref. [108])
- Current list of Tier-2 centres (Ref. [109])
- Current planning of computing capacities for the Tier-0 and the Tier-1s (Ref. [110])
- Current planning of computing capacities for the Tier-2s (Ref. [111]).

Table 8.2 summarizes the requirements of the experiments for capacity at CERN, at external Tier-1 centres, and in Tier-2 centres in the years 2007-2010. The requirements for the first full year of data taking (2008) are compared with the current capacity planned at CERN and in the regional centres. Note that the site planning data has been prepared assuming that the ALICE experiment would run for only a limited period in 2008, whereas the ALICE requirements in Section 2 of the TDR now assume a full four week run in 2008. All of the Tier-1 centres have provided resource planning estimates for 2008, but only a small subset of the 39 Tier-2 centres and 'federations' currently identified have provided such data.

Table 8.2: Requirements and currently planned capacity at CERN, Tier-1s, and Tier-2s

	Requirements - all experiments				Current planned capacity		
CPU (MSI2K)	**2007**	**2008**	**2009**	**2010**	**2008**	**% of requirements**	**Notes**
CERN Total	10.0	25.3	34.5	53.7	20.0	79%	1, 2
CERN Tier-0	6.9	17.5	22.4	32.8	12.3	70%	
CERN T1/T2	3.1	7.8	12.1	20.9	7.7	99%	
All external Tier-1s	19.2	55.9	85.2	142.0	47.1	84%	1
All Tier-2's	23.6	61.3	90.4	136.6	15.8	26%	3
Total	**53**	**143**	**210**	**332**	**83**	58%	
Disk(TB)							
CERN Total	2'200	6'600	9'200	12'600	5'700	86%	1
CERN Tier-0	400	1'300	1'400	1'800	1'100	85%	
CERN T1/T2	1'800	5'300	7'800	10'800	4'600	87%	
All external Tier-1s	9'300	31'200	45'400	72'100	21'700	70%	1
All Tier-2's	5'200	18'800	32'400	49'200	3'300	18%	3
Total	**17'000**	**57'000**	**87'000**	**134'000**	**31'000**	54%	
MSS (TB)							
CERN Total	4'900	18'000	31'100	45'600	15'300	85%	1
CERN Tier-0	3'400	13'600	23'600	34'500	12'000	88%	
CERN T1/T2	1'500	4'400	7'500	11'100	3'300	75%	
All external Tier-1s	9'300	34'700	60'800	92'200	24'700	71%	1
Total	**14'000**	**53'000**	**92'000**	**138'000**	**40'000**	75%	

Notes

1. CERN and Tier-1 planning has not been reviewed after the announcement that ALICE assumes a full period of data taking in 2008
2. At CERN all ALICE processing requirements are assimilated in the Tier-0.
3. Planning for Tier-2s available only from France, Japan, Spain, Switzerland and UK

Estimates of the cost of providing the resources planned for the CERN facility (Tier-0, Tier-1 and analysis facility) have been made, as described in an earlier section of this TDR. The current cost estimates exceed the planning budget presented at the Computing Resource Review Board in April 2005 by about 10%. The funding allocated to this budget in the CERN Medium Term Plan at present will cover only about 70% of these estimated costs. On the

other hand, the funding for the personnel required at CERN through to the end of 2008 is already in the CERN Medium Term Plan.

The financial and human resources required to provide the planned capacity at the various centres other than CERN are not visible to the project — the MoU deals only with capacity and service levels.

GLOSSARY – ACRONYMS – DEFINITIONS

3D	Distributed Deployment of Databases for LCG Project
ACL	Access Control List
ADIC	Advanced Digital Information Corporation
AF	Architect Forum
AFS	Andrew File System
ALICE	A Large Ion Collider Experiment (LHC experiment)
AliEn	Alice Environment
AMD	Advanced Micro Devices (semiconductor company)
AOD	Analysis Object Data (LCG)
API	Application Program Interface
ARC	Advanced Resource Connector
ARDA	A Realisation of Distributed Analysis for LHC
ASAP	ARDA Support for CMS Analysis Processing
ASN	Autonomous System Number
ATA	Advanced Technology Attachment, 16 bit disk interface
ATLAS	A Toroidal LHC ApparatuS (LHC experiment)
AUP	Acceptable User Policy
BDII	Berkeley Database Information Index
BGP	Border Gateway Protocol
BKDB	Book-keeping Database
BOSS	Batch Object Submission System
BQS	Batch Queueing System
CA	Certificate Authority
CAF	CERN Analysis Facility
CASPUR	Consorzio interuniversitario per le Applicazioni di Supercalcolo Per Università e Ricerca - "Inter-university consortium for the application of super-computing for Universities and Research"
CASTOR	CERN Advanced STORage Manager
CB	Collaboration Board
CDB	Condition Data Base
CE	Computing Element: a Grid-enabled computing resource
CIC	Core Infrastructure Centres
CIDR	Classless Inter-Domain Routing
CINT	C/C++ INTerpreter
CMS	Compact Muon Solenoid (LHC experiment)
CMT	Configuration Management Tool
COOL	Conditions Database Project
CORE	
CPU	Central Processing Unit
CRAB	CMS Remote Analysis Builder
C-RRB	Computing Resource Review Board
DAQ	Data Acquisition System
DBMS	Bata Base Management System
DBS	Dataset Book-keeping System (CMS)
DC	Data Challenge
dCache	Hierachical storage manager (DESY, FNAL)
DCGC	Danish Centre for Grid Computing
DCS	Detector Control System
DGAS	DataGrid Accounting System
DIAL	Distributed Interactive Analysis of Large datasets
DIRAC	Distributed Infrastructure with Remote Agent Control
DLI	Data Location Interface
DLS	Data Location System (CMS)

DN	Distinguished Names
DPM	Disk Pool Manager
DQ	Don Quijote
DRD	RAW data plus selected ESD objects
DST	Data Summary Tape, LHCb
ECS	Experiment Control System
EDG	European Data Grid Project
EF	Event Filter
EGEE	Enabling Grids for E-sciencE
ELFms	Extremely Large Fabric management system
EM	Planned refinements in electromagnetic ??
Enstore	Mass storage system implemented at FNAL
ESD	Event Summary Data
FC	File Catalogue
FiReMan	File and Replica Catalogue
FLUKA	Simulation tool
FML	Fitting and Minimization Library
FPS	gLite Transfert Services
FTS	File Transfer Service
GAG	Grid Applications Group
GANGA	Gaudi / Athena and Grid Alliance
Gb	Gigabits
GB	Gigabytes
GDB	Grid Deployment Board
GDML	Geometry Description Markup Language
GÉANT	European networking backbone interconnecting national research networks (NRENs)
GEANT-2	European overlay platform of the NRNs
GEANT3	Simulation program in FORTRAN
GEANT4	Toolkit for the simulation of the passage of particles through matter.
GENSER	Generator Library
GFAL	Grid File Access Library
GFLASH	Fast shower parametrization package
GGF	Global Grid Forum
GGUS	Grid User Support Centre
GIIS	Grid Index Information Service
gLite	Lightweight middleware for Grid computing
Glue	Grid Laboratory Uniform Environment
GNU	GNU's Not Unix - recursive acronym
GOC	Grid Operation Centres
GridFTP	Grid Service for File Transfer
GridICE	Grid Monitoring middleware
GSI	Grid Security Infrastructure
GSL	GNU Scientific Library
gSOAP	Toolkit for the development of SOAP/XML Web services in C/C++
GT2	Globus Toolkit Version 2
GUI	Graphicla User Interface
GUID	Globally Unique Identifiers
HDD	Hard Disk Drive
HEP	High Energy Physics
HEPCAL	HEP Application Grid requirements
HLT	High-level Trigger
HMS	Hardware Management System
HPSS	High Performance Storage System
HSM	Hierarchical Storage Manager

HTTP	HyperText Transfer Protocol
I/O	Input/Output
IBA	Infiniband
IGP	Interior Gateway Protocol
IOV	Interval Of Validity
IPC	Inter Process Communication
iVDGL	International Virtual Data Grid Laboratory
Jabber	open, XML-based protocol for instant messaging
LAN	Local Area Network
LB	Logging and Book-keeping
LCAS	Local Centre Authorization Service
LCG	LHC Computing Grid
LCMAPS	Local Credential Mapping Service
LDAP	Lightweight Directory Access Protocol
LEAF	LHC-Era Automated Fabric
LEMON	LHC Era Monitoring
LFC	LCG File Catalogue
LFN	Logical File Name
LHCb	Large Hadron Collider beauty experiment
LHCC	Large Hadron Collider experiments Committee
LPM	Landau, Pomeranchuk, and Migdal suppression of bremsstrahlung
LRC	Local Replica Catalogue
LRMS	Local Resource Management System
LSF	Load Sharing Facility
LTO	Linear Tape-Open
LXBATCH	LinuX cluster for BATCH processing at CERN
LXR	Linux Cross Reference
MAN	Metro Area Network
MC	Monte Carlo
MCDB	Monte-Carlo Events Data Base
MINUIT	Fred James' function minimization and error analysis package
MonaLisa	MONitoring Agents using a Large Integrated Services Architecture
MoU	Memorandum of Understanding
MSI2000	Million SpecInt 2000 – see SpecInt
MSS	Mass Storage System
MTBF	Mean Time Between Failures
MySQL	Open source database
NAS	Network-Attached Storage
NDGF	Nordic Data Grid Facility
NIC	Network Interface Card
NOC	Network Operation Centre
NorduGrid	A collaboration that has developed the Advanced Resource Connector (ARC) middleware
NREN	National Research and Education Network
OCTOPUS	CMS production system
OMC	Operations Management Centre
OMDS	Online Master Data Storage (LHCb)
OpenLDAP	Open Source version of LDAP – Lightweight Directory Access Protocol
OpenSSL	Open source version of SSL - Secure Sockets Layer
OPN	Optical Private Network
ORCA	Object Reconstruction for CMS Analysis
ORCA	Official Release of the CMS software required for Analysis
ORCOF	Offline Reconstruction Conditions database OFFline (LHCb)
ORCON	Offline Reconstruction Conditions database ONline (LHCb)
OS	Operating System

OSCT	Operational Security Coodination Team
OSG	Open Science Grid
PAI	Photo Absorption Ionisation
PASTA	LHC technology tracking team for Processors, memory, Architectures, Storage and TApes
PBS	Portable Batch System
PDB	Processing Database
PDC	Physics Data Challenge
PFN	Physical File Name
PhEDEx	Physics Experiment Data Export
PKI	Public Key Infrastructure
PMA	Policy Management Authorities?
POOL	Pool Of persistent Objects for LHC
POSIX	Portable Operating System Interface
PPDG	Particle Physics Data Grid
PyLCGDict	Python binding to the LCG Dictionary
PyROOT	Access ROOT objects from Python
QCD	Quantum ChromoDynamics
QDR	Quad Data Rate
RAC	Real Application Clusters
RAID	Redundant Array of Independent Disks
RAL	Relational Access Layer
RAW	Raw data file
RB	Research Board
RDBMS	Relational Database Management System
RDMA	Remote Direct Memory Access
rDST	Reduced DST (LHCb)
RECO	Reconstructed events (CMS)
RefDB	Reference Database
RFIO	Remote File I/O
R-GMA	Relational Grid Monitoring Architecture
RH	Red Hat (Linux)
RIR	Regional Internet Registry
RLS	Replica Location Service
RMAN	Recovery Manager (Oracle)
RMC	Replica Metadata Catalogue
ROC	Regional Operations Centres
SAN	Storage Area Network
SAS	Serial Attached SCSI
SASL	Simple Authentication and Security Layer
SATA	Serial ATA
SC	Service Challenge
SCRAM	Software Configuration, Release And Management
SCSI	Small Computer System Interface
SDLT	Super Digital Linear Tape
SE	Storage Element
SFT	Site Functional Test
SIMU	Simulation
SMS	State Management System
SOAP	Simple Object Access Protocol
SPECint	Measure of CPU performance - see www.spec.org
SPECint2000	The geometric mean of 12 normalized benchmark ratios - see www.spec.org
SPI	Software Process Infrastructure
SQL	Structured Query Language

SQLite	SQL database engine
SRB	Storage Resource Broker
SRB	Storage Resource Broker
SRM	Storage Resource Manager
SSE	Smart Storage Element
STK	Storage Technology Corporation, StorageTek®
SURL	Storage Unique Resource Locator
TAG	Event index information
Tb	Terabits
TB	Terabytes
THEPEG	Toolkit for High Energy Physics Events Generation
TMDB	Transfer Management DataBase (CMS)
UI	User Interface
UNICORE	Uniform Interface to Computing Resources
VDT	Virtual Data Toolkit
VO	Virtual Organization
VOMS	Virtual Organization Management System
WAN	Wide-Area Network
WDM	Wavelength Division Multiplexer, increase optical channels per fibre
WLCG	Worldwide LHC Computing Grid
WM	Workload Manager
WMS	Workload Management System
WN	Worker Node
XML	eXtensible Markup Language
XMPP	eXtensible Messaging and Presence Protocol
xRSL	Extended Resource Specification Language

REFERENCES

[1] LHC Homepage, http://cern.ch/lhc-new-homepage/

[2] Memorandum of Understanding, CERN-C-RRB2005-01, http://lcg.web.cern.ch/LCG/C-RRB/2005-04/LCG_T0-2_draft4p5_f.pdf

[3] Computing Model documents of the LHC experiments, http://lcg.web.cern.ch/LCG/peb/LHCC/expt%5Freqts/

[4] ALICE Technical Design Report of the Computing, ALICE-TDR-012, CERN-LHCC-2005-018

[5] ATLAS Computing, ATLAS-TDR-017, CERN-LHCC-2005-022

[6] CMS Computing Technical Design Report, CMS-TDR-007, CERN-LHCC-2005-023

[7] LHCb Computing, LHCb-TDR-11, CERN-LHCC-2005-019

[8] EGEE Homepage, http://www.cern.ch/egee

[9] Open Science Grid (OSG) Web Page, http://www.opensciencegrid.org/

[10] Nordic Data Grid Facility (NDGF), http://www.ndgf.org/

[11] Proposal for Building the LHC Computing Environment at CERN, CERN/2379/Rev., 2001, http://cern.ch/LCG/PEB/Documents/c-e-2379Rev.final.doc

[12] F. Carminati, J. Templon *et al.*, Common use cases for a HEP Common Application layer, LHC-SC2-20-2002 (2002) and F. Carminati, J. Templon *et al.*, Common use cases for a HEP Common Application layer for analysis, LHC-SC2-2003-032 (2003)

[13] PhEDEx - Physics Experiment Data Export - Project Homepage, http://cms-project-phedex.web.cern.ch/cms-project-phedex/

[14] V. Lefebure *et al.*, RefDB: The Reference Database for CMS Monte Carlo Production., Proceedings of the CHEP03 Conference, La Jolla, California (2003)

[15] G. Graham *et al.*, McRunjob: A high Energy Physics Workflow planner for Grid Production., Proceedings of the CHEP03 Conference, La Jolla, California (2003)

[16] BOSS - Batch Object Submission System - Project Homepage, http://boss.bo.infn.it/ and C. Grandi, A. Renzi, Object Based system for Batch Job Submission and Monitoring (BOSS), CMS NOTE 2003-005 (2003)

[17] A. Tsaregorodstev *et al.*, DIRAC - Distributed Infrastructure with Remote Agent Conntrol, Proc. of CHEP2003
A. Tsaregorodstev *et al.*, DIRAC - The Distributed MC Production And Analysis For LHCb, Proceedings of CHEP04, Interlaken, Switzerland, September 2004

[18] U. Egede, LHCb Use Cases for Distributed Analysis, LHCb 2005-027,D. Adams *et al.*, and GANGA 4 architecture, http://ganga.web.cern.ch/ganga/documents/pdf/Ganga4Architecture.pdf

[19] SRM working group, http://sdm.lbl.gov/srm-wg/

[20] J-P Baud, J. Casey, Evolution of LCG-2 Data Management, Proceedings of CHEP04, Interlaken, Switzerland, September 2004

[21] FiReMan catalogue, http://egee-jra1-dm.web.cern.ch/egee-jra1-dm/

[22] D. Olson, J. Perl, Grid Service Requirements for Interactive Analysis (2002), http://www.ppdg.net/pa/ppdg-pa/idat/papers/analysis_use-cases-grid-reqs.pdf

[23] CRAB - CMS Remote Analysis Builder - Project homepage, http://cmsdoc.cern.ch/cms/ccs/wm/www/Crab/

[24] Baseline Services Working Group, http://cern.ch/lcg/PEB/BS

[25] LCG Baseline Services Group Report, http://cern.ch/LCG/PEB/BS/BSReport-v0.7.pdf

[26] B. Panzer-Steindel, Sizing and Costing of the CERN T0 centre, CERN-LCG-PEB-2004-21 http://cern.ch/LCG/planning/phase2_resources/SizingandcostingoftheCERNT0center.pdf
[27] LCG File Transfer Service Challenge Requirements, http://cern.ch/lcg/tdr/docs/SC_Requirements.pdf
[28] gLite Architecture, https://edms.cern.ch/document/476451
[29] User's Guide for the VOMS Core Services, http://edms.cern.ch/document/571991
[30] EGEE User's Guide, https://edms.cern.ch/document/572406
[31] EGEE gLite User's Guide, Overview Of Glite Data Management, https://edms.cern.ch/file/570643/1/EGEE-TECH-570643-v1.0.pdf
[32] EGEE gLite User's Guide, gLite I/O, https://edms.cern.ch/document/570771/1
[33] User's Guides for the DGAS Services, https://edms.cern.ch/document/571271
[34] User's Guide for the VOMS Core Services, https://edms.cern.ch/document/571991/1
[35] VOMS admin user's guide, https://edms.cern.ch/document/572406/1
[36] JDL Attributes, http://server11.infn.it/workload-grid/docs/DataGrid-01-TEN-0142-0_2.pdf
[37] JDL Attributes Specification, https://edms.cern.ch/document/555796/1
[38] WMS User's Guide, https://edms.cern.ch/document/572489/1
[39] LB Service User's Guide, https://edms.cern.ch/document/571273/1
[40] User Guide For Edg Replica Manager 1.5.4, http://cern.ch/edg-wp2/replication/docu/r2.1/edg-replica-manager-userguide.pdf
[41] Developer Guide For Edg Replica Manager 1.5.4, http://cern.ch/edg-wp2/replication/docu/r2.1/edg-replica-manager-devguide.pdf
[42] Fireman Catalogue User Guide, https://edms.cern.ch/document/570780
[43] Service Discovery User Guide, https://edms.cern.ch/document/578147
[44] gLite Release 1 Web Page, http://hepunx.rl.ac.uk/egee/jra1-uk/glite-r1
[45] LB Service User's Guide, https://edms.cern.ch/document/571273
[46] JP usage guide, http://egee-jra1-wm.mi.infn.it/egee-jra1-wm/jp_usage.shtml
[47] File Transfer Service User Guide, https://edms.cern.ch/document/591792/1
[48] Grid Physics Network (GriPhyN) Web Page, http://www.griphyn.org/
[49] International Virtual Data Grid Laboratory (iVDgL) Web Page, http://www.ivdgl.org/
[50] LHC Computing Grid (LCG), Web Page, http://lcg.web.cern.ch/LCG/
[51] Particle Physics Data Grid (PPDG) Web Page, http://www.ppdg.net/
[52] UNICORE Forum e.V. Web Page, http://www.unicore.org/
[53] NORDUGRID Web Page, http://www.nordugrid.org/
[54] gLite Web Page, http://glite.web.cern.ch/glite/
[55] Globus Alliance Web Page, http://www.globus.org/
[56] Condor Project Homepage, http://www.cs.wisc.edu/condor/
[57] Virtual Data Toolkit (VDT) Web Page, http://www.cs.wisc.edu/vdt//index.html
[58] Storage Resource Borker (SRB) Web Page, http://www.npaci.edu/DICE/SRB/
[59] 'What Is Wicki' website, http://wiki.org/wiki.cgi?WhatIsWiki
[60] ELFms, Extremely Large Fabric management system, http://elfms.web.cern.ch/elfms/
[61] quattor, system administration toolsuite, http://quattor.web.cern.ch/quattor/
[62] LEMON — LHC Era Monitoring, http://lemon.web.cern.ch/lemon/
[63] LHC-Era Automated Fabric (LEAF), http://leaf.web.cern.ch/leaf/

[64] B. Panzer-Steindel, Price extrapolation parameters for the CERN LCG Phase II Computing Farm, CERN-LCG-PEB-2004-20, http://lcg.web.cern.ch/LCG/planning/phase2_resources/Priceextrapolation.pdf

[65] B. Panzer-Steindel, Reference points for the cost calculations of LCG Phase 2, http://lcg.web.cern.ch/LCG/planning/phase2_resources/Referencepoints.pdf

[66] LCG Project Applications Area Website, http://lcgapp.cern.ch/project/

[67] J. Apostolakis *et al.*, Report of the LHC Computing Grid Project Architecture Blueprint RTAG, Oct 9, 2002

[68] S. Roiser and P. Mato, The SEAL C++ Reflection System, Proceedings of CHEP'04, Interlaken, Switzerland, 24 Sep – 1 Oct 2004, CERN-2005-02, Vol 1, p. 437 International Standard; Programming Languages – C++; ISO/IEC 14882:2003(E); Second edition 2003-10-15; ISO, CH-1211 Geneva 20, Switzerland

[69] J. Generowicz, P. Mato, W. Lavrijsen, and M. Marino, Reflection-based Python C++ bindings, Proceedings of CHEP'04, Interlaken, Switzerland, 24 Sep – 1 Oct 2004, CERN-2005-02, Vol 1, p. 441

[70] M. Hatlo *et al.* , Developments of mathematical software libraries for the LHC experiments, Proceedings of CHEP'04, Interlaken, Switzerland, 24 Sep – 1 Oct 2004, CERN-2005-02, Vol 1, p. 629

[71] R. Brun and F. Rademakers, ROOT-An Object Oriented Data Analysis Framework, Nucl. Inst.&Meth. in Phys.Res.A389(1997)81-86.

[72] D. Malon *et al.*, Report of the LHC Computing Grid Persistency Management RTAG, http://lhcgrid.web.cern.ch/LHCgrid/sc2/RTAG1

[73] D. Düllmann *et al.*, The LCG POOL Project – General Overview and Project Structure, CHEP03

[74] I. Papadopoulos *et al.*, POOL, the LCG Persistency Framework, IEEE Nuclear Science Symposium, Portland Oct 2003

[75] G. Govi *et al.*, POOL Integration into three Experiment Software Frameworks, CHEP04

[76] M. Frank *et al.*, The POOL Data Storage, Cache and Conversion Mechanism, CHEP03 and CHEP04

[77] D. Düllmann *et al.*, POOL Development Status and Plans, CHEP04

[78] D. Düllmann *et al.*, On Distributed Database Deployment for the LHC Experiments, CHEP04

[79] P. Leach, R. Salz, UUIDs and GUIDs, Internet-Draft, ftp://ftp.isi.edu/internet-drafts/draft-leach-uuids-guids-00.txt

[80] Z. Xie *et al.*, POOL File Catalogue, Collection and Metadata Components, CHEP03

[81] M. Girone *et al.*, Experience with POOL from the LCG Data Challenges of the three LHC experiments, CHEP04

[82] A. Valassi, LCG Conditions Database Project Overview, CHEP04

[83] J. Apostolakis *et al.*, GEANT4 simulation production in LHC experiments, CERN-LCGAPP-2005-02

[84] J. Apostolakis *et al.*, GEANT4: Status and recent developments, Proceedings of CHEP'04, Interlaken, Switzerland, 24 Sep – 1 Oct 2004, CERN-2005-02, Vol 1, p. 199

[85] G. Folger, J.P. Wellisch, The Binary Cascade', Proceedings of CHEP'04, Interlaken, Switzerland, 24 Sep – 1 Oct 2004, CERN-2005-02, Vol 1, p. 313

[86] Simulation Framework Web page: http://lcgapp.cern.ch/project/simu/framework/

[87] F. Carminati *et al.,* The Virtual Monte Carlo : status and applications, Proceedings of CHEP'04, Interlaken, Switzerland, 24 Sep – 1 Oct 2004, CERN-2005-02, Vol 1, p. 235

[88] J. Beringer, (p,xn) Production Cross Sections: A Benchmark Study for the Validation of Hadronic Physics Simulation at LHC , CERN-LCGAPP-2003-18,
F. Gianotti *et al.,* Simulation physics requirements from the LHC experiments , CERN-LCGAPP-2004-02
A. Ribon, Validation of GEANT4 and FLUKA Hadronic Physics with Pixel Test-Beam Data , CERN-LCGAPP-2004-09
F. Gianotti *et al.,* GEANT4 hadronic physics validation with LHC test-beam data: first conclusions , CERN-LCGAPP-2004-10
W. Pokorski, In-flight Pion Absorption: Second Benchmark Study for the Validation of Hadronic Physics Simulation at the LHC , CERN-LCGAPP-2004-11
A. Ribon, Physics validation of the simulation packages in a LHC-wide effort, Proceedings of CHEP'04, Interlaken, Switzerland, 24 Sep – 1 Oct 2004, CERN-2005-02, Vol 1, p. 203

[89] A. Pfeiffer *et al.,* Software management infrastructure in the LCG Application Area, Proceedings of CHEP'04, Interlaken, Switzerland, 24 Sep – 1 Oct 2004, CERN-2005-02, Vol 1, p. 644

[90] E. Poinsignon *et al.,* Managing third-party software for the LCG, Proceedings of CHEP'04, Interlaken, Switzerland, 24 Sep – 1 Oct 2004, CERN-2005-02, Vol 1, p. 583

[91] M. Gallas *et al.,* Quality Assurance and Testing in LCG, Proceedings of CHEP'04, Interlaken, Switzerland, 24 Sep – 1 Oct 2004, CERN-2005-02, Vol 1, p. 651

[92] Y. Perrin *et al.,* The LCG Savannah software development portal, Proceedings of CHEP'04, Interlaken, Switzerland, 24 Sep – 1 Oct 2004, CERN-2005-02, Vol 1, p. 609

[93] LCG Applications Area WBS and Schedule, http://atlassw1.phy.bnl.gov/Planning/lcgPlanning.html

[94] Report of the Steering Group of the LHC Computing Review, http://cern.ch/lhc-computing-review-public/Public/Report_final.PDF

[95] O Smirnova, Extended Resource Specification Language, http://www.nordugrid.org/documents/xrsl.pdf

[96] J. Kennedy, The role of legacy services within ATLAS DC2, CHEP 2004, Interlaken, contribution no. 234

[97] LCG Project Planning Web, http://lcg.web.cern.ch/LCG/planning/planning.html

[98] Service Challeng 3 Planning Document, http://cern.ch/LCG/planning/planningdoc/SC3-planning-May.doc

[99] ARDA RTAG Report, Architectural Roadmap for Distributed Analysis, http://lcg.web.cern.ch/lcg/PEB/arda/public_docs/ARDA_report_final.pdf

[100] M.Ballintijn *et al.,* PROOF Distributed Parallel Analysis Framework based on ROOT Proceedings of CHEP'03, La Jolla, California
M.Ballintijn *et al.,* Super Scaling PROOF to very large clusters, Proceedings of CHEP'04, Interlaken, Switzerland, 24 Sep – 1 Oct 2004, CERN-2005-02, Vol 2, p. 1111

[101] Baseline Services Working Group Report, http://lcg.web.cern.ch/LCG/PEB/BS/Baseline-Services-Report.pdf

[102] LCG Planning Page http://lcg.web.cern.ch/LCG/planning/planning.html

[103] LCG Applications Area Management and Planning, http://lcgapp.cern.ch/project/mgmt/

[104] LCG Service Challenges, Web Page with link to milestones, http://lcg.web.cern.ch/LCG/activities/servicechallenges.html

[105] List of Tier-2 centres, http://lcg.web/lcg/C-RRB/Tier-2/ListTier2Centres.pdf

[106] The DataGrid Project Web Page, http://cern.ch/eu-datagrid/

[107] Alien Web Page, http://alien.cern.ch/

[108] List of Tier-1 centres, http://lcg.web.cern.ch/LCG/C-RRB/Tier-1/ListTier1Centres.pdf

[109] List of Tier-2 centres, http://lcg.web.cern.ch/LCG/C-RRB/Tier-2/ListTier2Centres.pdf

[110] Computing capacities for the Tier-0 and the Tier-1s, http://lcg.web.cern.ch/LCG/planning/phase2_resources/draft_LCG_Tier0-1Res.pdf

[111] Computing capacities for the Tier-2s, http://lcg.web.cern.ch/LCG/planning/phase2_resources/draft_LCG_Tier2Res.pdf