CERN-LHCC-2004-038/G-086
04 February 2005

A L I C E

Computing Model

Computing Project Leader
F. Carminati

Offline Coordinator
Y. Schutz

(Editors on behalf of the ALICE Collaboration)

Foreword

This document is an updated version of the *ALICE Computing Model* document presented at the LHCC review of LHC experiments computing resources. This new version takes into account the recommendation given by the referees. The major changes are the following:

- Tier 1 and Tier 2 services are required at CERN;
- The CPU resources requested at CERN for the Tier 0, Tier 1 and Tier 2 complex does not show a peak structure;
- The model has been made less dependent on the availability of advanced GRID functionalities. A specific class of tasks has been identified for each Tier;
- The reconstructions in Tier 1's are more evenly spread over time;
- Analysis has been split into scheduled and end-user (*chaotic*) analysis. The former is targeted to be performed at Tier 1's and the latter at Tier 2s;
- The calibration and alignment tasks have been better defined;
- Use of the HLT farm for offline processing is considered in the model. The HLT farm will provide a contingency for the processing of the first reconstruction pass for Pb–Pb events at Tier 0;
- We have not assumed any sharing of resources at the T0 between experiments in our request.

These modifications have introduced a small increase in the overall required CPU resources in Tier 1s and Tier 2s, part of it being absorbed by the Tier 1 and Tier 2 at CERN.

Executive Summary

The principal objective of this document is to present the computing needs of the ALICE collaboration in terms of offline computing and to substantiate them with a brief description of the ALICE computing model. The content of this document has been discussed with the ALICE Collaboration in a workshop held at CERN on December 9 and 10, 2004 and will be a major input to the ALICE Computing Technical Design Report.

We are, at the moment of writing, just more than two years away from the first collisions at LHC. This is still a long lapse of time for Information Technology, due to the fast evolution pace of this field. On the other hand, the anticipated needs for LHC computing are very large, and therefore the acquisition, deployment and organization of the material and human resources needed to satisfy them cannot be improvised. This is particularly true since these resources will be distributed in many centers in different countries, which will have to work together as a single entity.

Hence comes the difficulty of writing this document, which has to contain enough details to justify our requests, ideally without basing ourselves on elements that can and will still change in the course of the next few years.

The computing model concerns essentially what we call a *Standard Data Taking Year* (SDTY). During a SDTY, ALICE will take heavy-ion data for 10^6 effective seconds per year (one month), while for the rest of the time, 10^7 effective seconds, it will collect proton-proton data. During the initial phase we

imagine that the effective beam time may be less, and with progressively increasing luminosity. However, what will happen exactly during these first three years, i.e. the so called *initial running conditions*, is being periodically redefined. Our model covers this period via a staging of the deployment of the computing resources, i.e. 20% to be available in 2007 for the first proton–proton (p–p) run, 40% in 2008 and 100% at the end of 2008 for the Pb–Pb run. This responds to the requirement to delay the hardware acquisition as much as possible, every year bringing a reduction in cost of 30-40%.

The computing model for the p–p data is similar to the one of the other experiments. The data are recorded at a speed of 100 MB/s and they are immediately reconstructed at the CERN Tier 0 facility and exported to the different Tier 1's outside CERN (hereon *external Tier 1's*). This ensures that there will be two copies of the raw data, one at the CERN Tier 0 and another one shared by all the external Tier 1's to ensure that data are safely stored. The Tier 1's shall have collectively enough resources to perform a second and third reconstruction.

For Pb–Pb data this model is not viable, as data are recorded at up to 1.25 GB/s, and this would require a prohibitive amount of resources for quasi real-time processing. We therefore require that these data are reconstructed at the CERN Tier 0 and exported over a four month period after data taking. This should leave enough time for a second and third reconstruction pass at the Tier 1's.

It is customary to assume that scheduled and unscheduled analysis and simulation will go on at the Tier 2 centers. We adopt this model but we consider that, on the basis of our experience with the Data Challenges, this hierarchical model based on the MONARC work will be progressively replaced by a more *democratic* and model often indicated as *cloud model*. In this model the only distinctive features of the Tier 1's, apart from their size, are their service levels and their commitment to safely store the data, most probably on mass storage systems.

The final model will be decided by the quality and functionality of the Grid middleware. Should the middleware have a limited functionality in deciding dynamically where to perform the calculations and where to direct the data, a hierarchical model will be useful in organizing *by hand* the computing activity. A middleware implementing a set of functionalities closer to the *Grid vision* could profit from some more freedom of choice, leading to a usage pattern of the resources similar to the one predicted by the cloud model.

At the moment of writing it is not yet clear which functionality will be offered by the Grid middleware that will be deployed on the LCG resources. In writing this computing model we have implicitly assumed that a functional middleware will exist, optimizing to some extent the storage and workload distribution. Based on our experience with the ALICE-developed AliEn system we believe that this is technically possible. However, in this moment it is not sure whether this or an equivalent product will indeed be deployed by the time LHC starts. Therefore, to reduce the dependency on the middleware, we have indicated for each Tier, which are the specific tasks that have to be performed on it. It is clear that if the distribution and monitoring of these tasks require manual intervention, the system will be used less efficiently and will need more human and possibly hardware resources to fulfill the computing needs of the experiment. Throughout the document we will use the MONARC terminology to discuss the different elements.

In this document we will not provide information on the resources pledged to ALICE by the funding agencies. While we have some preliminary information, indicating that the level of resources we request is commensurate with the intentions of the funding agencies, a proper allocation of resources will happen only upon consideration of the needs of all experiments after their validation by the review in January 2005.

We have not indicated any cost for the resources that we are requesting. We have obtained figures based on the costing model developed by the LCG project, however we are requesting computing resources and services and not financial resources. It is up to the funding agencies to decide on the financial model with which to provide these resources.

Finally it is important to note that all the information contained in this document is provided to the best of our knowledge. This document is the result of a long consensus building process within the

ALICE Collaboration and depends on a number of human and technological factors that are in rapid evolution. We anticipate a qualitative as well as quantitative evolution of the ALICE computing model. The final model described in the ALICE Computing Technical Design Report, due to appear in June 2005, may therefore differ from this document, even if the changes are expected to be small.

Contents

1 Distributed computing and the GRID

1.1 Introduction

The ALICE computing model makes the assumption that there will be a functional Grid allowing efficient access to the resources. We believe that this is technically feasible thanks to our experience during the Data Challenges with the ALICE-developed AliEn system. However, in order to make our model less dependent on this assumption, we have clearly indicated which are the specific classes of tasks that should be performed by each Tier. However, should the functionality of the middleware that will be deployed on the final system be reduced with respect to what we have already used, these specific tasks will have to be manually directed and monitored, and the usage of the computing resources, both CPU and storage, will be less efficient. As a result we expect that their operation will require more manpower. This will eventually increase the cost of the entire system.

In this chapter we give a short description of the ALICE distributed computing environment aimed at explaining why we believe that an advanced Grid solution for LHC computing is necessary and feasible.

1.2 Distributed computing

The ALICE computing model is driven by the large amounts of computing resources which will be necessary to store and process the data generated by the experiment (see Section 2). The required resources cannot be consolidated in a single computing center like CERN – it is more natural, both as a substantial financial investment as well as to cover the need for expert human resources, that these are spread over the HEP computing facilities of the institutes and universities participating in the experiment. This situation has been recognized already at the time of the conceptual design of the LHC experiments and its technical side has been formalized in the so-called MONARC model [1] shown in Fig. 1.1. MONARC describes an assembly of distributed computing resources, concentrated in a hierarchy of centers called Tiers, where Tier 0 is CERN, Tier 1's are the major computing centers with mass storage capability, Tier 2's the smaller regional computing centers, Tier 3 the university departmental computing centers and Tier 4 the user workstations.

The basic principle underlying our model is that every physicist should have equal access to the data and computing resources. The resulting system will be very complex. We expect hundreds of components at each site with several tens of sites. 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.

The commonality of distributed resources is being realized under the currently ongoing development of the Grid [3]. It was conceived to facilitate the development of new applications based on high-speed coupling of people, computers, databases, instruments, and other computing resource by allowing "dependable, consistent, pervasive access to high-end resources". Although the MONARC model predates the appearance of the Grid concept, its terminology is well adapted to the distribution of resources that is present in HEP and we will use the terms it has introduced throughout this document.

In a well functioning Grid the distribution of tasks to the different centers will be performed dynamically, based on the resources and the services that they advertise. This introduces a more flexible and *democratic* model sometimes called *Cloud Model*. The MONARC model remains however very useful for discussing the organization and relations of the centers, and we will use its nomenclature in

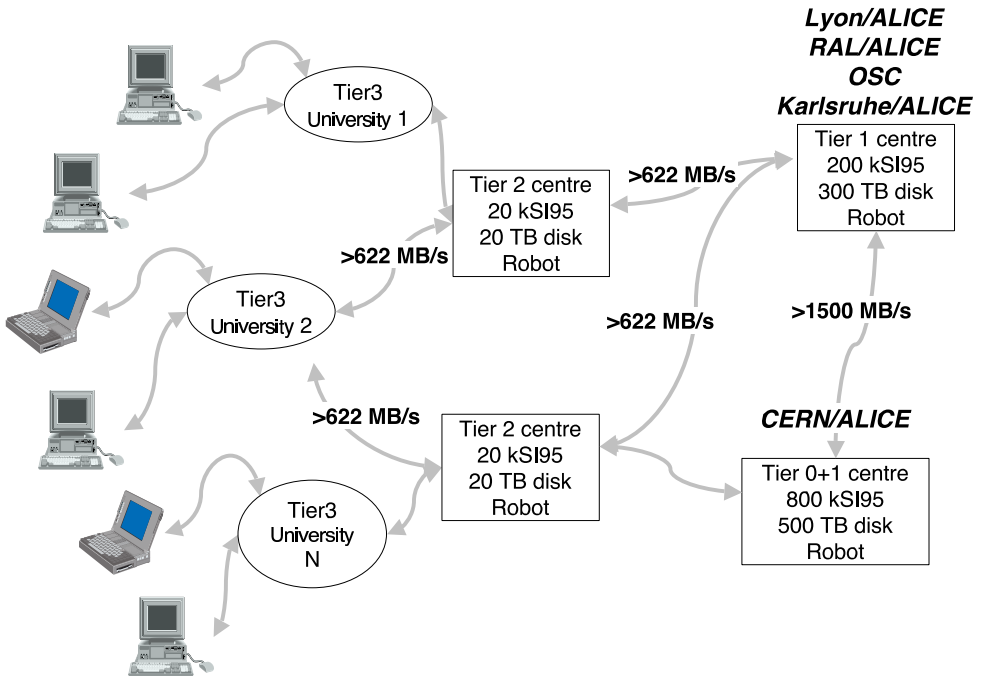

Figure 1.1: The MONARC model. The names of the centers in the figure are just examples.

the following. Note however that we will only discuss Tier 0, Tier 1 and Tier 2 in the computing model. The main difference between Tier 1 and Tier 2 is the quality of services and the amount of resources provided. In particular we suppose that Tier 1's are capable of storing a *safe* copy of the data, probably on a Mass Storage System (MSS).

Our computing model foresees that one copy of the ALICE raw data from the experiment will be stored at CERN (Tier 0) and a second copy will be distributed among the external (i.e. not at CERN) Tier 1 centers, thus providing a natural backup. Reconstruction will be shared by the Tier 1 centers, with the CERN Tier 0 responsible for the first reconstruction pass. Subsequent data reduction, analysis and Monte Carlo production will be a collective operation where all Tiers 1 and 2 will participate. The Tier 1's will perform reconstruction and scheduled analysis, while the Tier 2's will perform Monte Carlo and end-user analysis.

Grid technology holds the promise of greatly facilitating the exploitation of LHC data for physics research. Therefore ALICE is very active on the different Grid test-beds and worldwide project, where the Grid prototype middleware is deployed. The objective is to verify the functionality of the middleware, providing feedback to its authors, and to prototype the ALICE distributed computing environment.

This activity, both very useful and interesting in itself, is hindered by the relative lack of maturity of the middleware. This middleware is largely the result of leading-edge computer science research and is therefore still rather far from production quality software. Moreover, standards are missing and the middleware developed by different Grid projects is very different even if the functionality is similar. This makes it difficult for ALICE to exploit this software to run the productions that are more and more needed for the physics performance studies in a distributed environment, and to harness the resources of different computer centers. In particular the middleware provided by the Grid projects has shown to be unsuited for distributed analysis activities.

1.3 AliEn, the ALICE Grid

To alleviate these problems, while providing a stable and evolutionary platform, ALICE has developed the AliEn [2] (**Ali**CE **En**vironment) framework with the aim of offering to the ALICE user community transparent access to computing resources distributed worldwide. During the years 2001 - 2004 AliEn has provided a functional computing environment fulfilling the needs of the experiment in the preparation phase.

The system is built around Open Source components and uses Web Services model and standard network protocols. It 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. Only 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.

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, developed and deployed by other groups. This concept has been successfully demonstrated during the ALICE Physics Data Challenge '04 (PDC'04), where the resources of the LCG Grid were accessed through an AliEn-LCG interface.

AliEn Web Services play the central role in enabling AliEn as a distributed computing environment. The user interacts with them by exchanging SOAP messages and they constantly exchange messages between themselves behaving like a true Web of collaborating services. AliEn consists of the following key components and services: the authentication, authorization and auditing services; the workload and data management systems; the file and metadata catalogs; the information service; Grid and job monitoring services; storage and computing elements (see Fig. 1.2).

The AliEn workload management system is based on a pull approach. A central service manages all the tasks, while computing elements are defined as 'remote queues' and can, in principle, provide

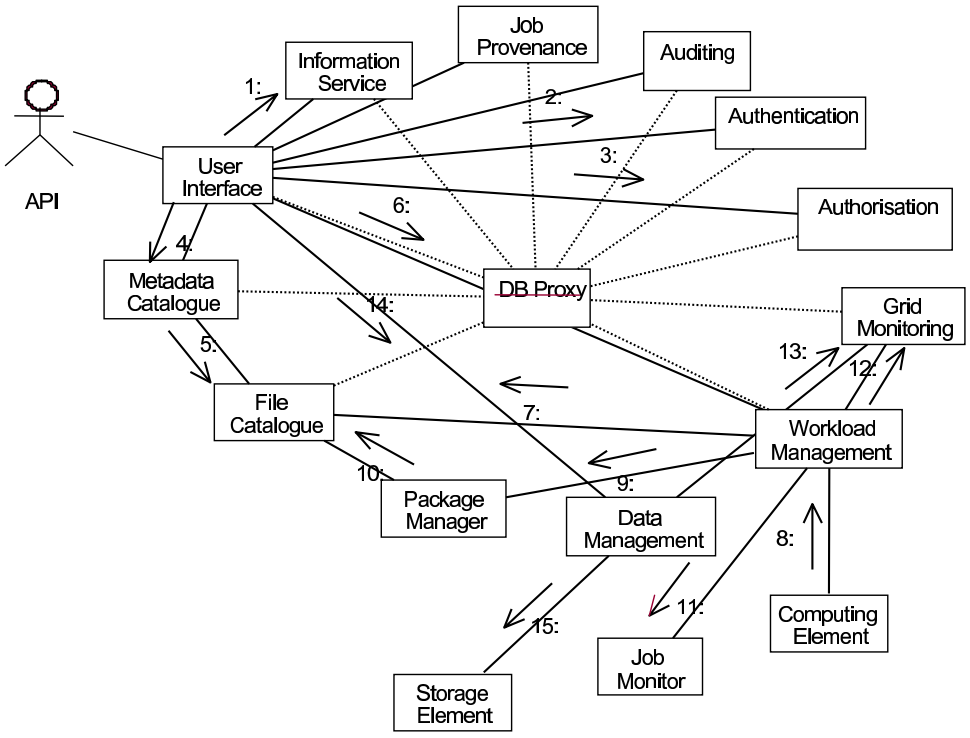

Figure 1.2: Interaction diagram of key AliEn components for typical analysis use case.

an outlet to 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 workload manager optimizes the queue taking into account job requirements based on input files, CPU time, architecture, disk space, etc. It then makes jobs eligible to run on one or more computing elements. The active nodes then get jobs from the queue and start their execution. The queue system monitors the job progression and has access to the standard output and standard error.

Input and output associated with any job are registered in the AliEn file catalog, a virtual file system in which a logical name is assigned to a file. Unlike real file systems, the file catalog does not own the files; it only keeps an association between the Logical File Name (LFN) 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 location and the necessary data replication is assumed to be provided transparently and automatically by the Grid middleware. The AliEn file system associates metadata with LFN's.

ALICE has used the system for distributed production of Monte Carlo data, reconstruction and analysis at over 30 sites on four continents (see Fig. 1.3). The round of productions run during the last year (PDC'04) is aimed at providing data for this report. During this period more than 400,000 jobs have been successfully run under AliEn control worldwide producing 40 TB of data. Computing and storage resources are available in both Europe and the US. The amount of processing needed for a typical production is in excess of 30 MSI2k×s to simulate and digitize a central Pb–Pb event. Some 100k events are generated for each major production. This is an average over a very large range since peripheral events may require one order of magnitude less CPU, and p–p events two orders of magnitude less. The Pb–Pb events are then reprocessed several times superimposing known signals, in order to be reconstructed and analyzed. Again there is a wide spread in the time this takes, depending on the event, but for a central

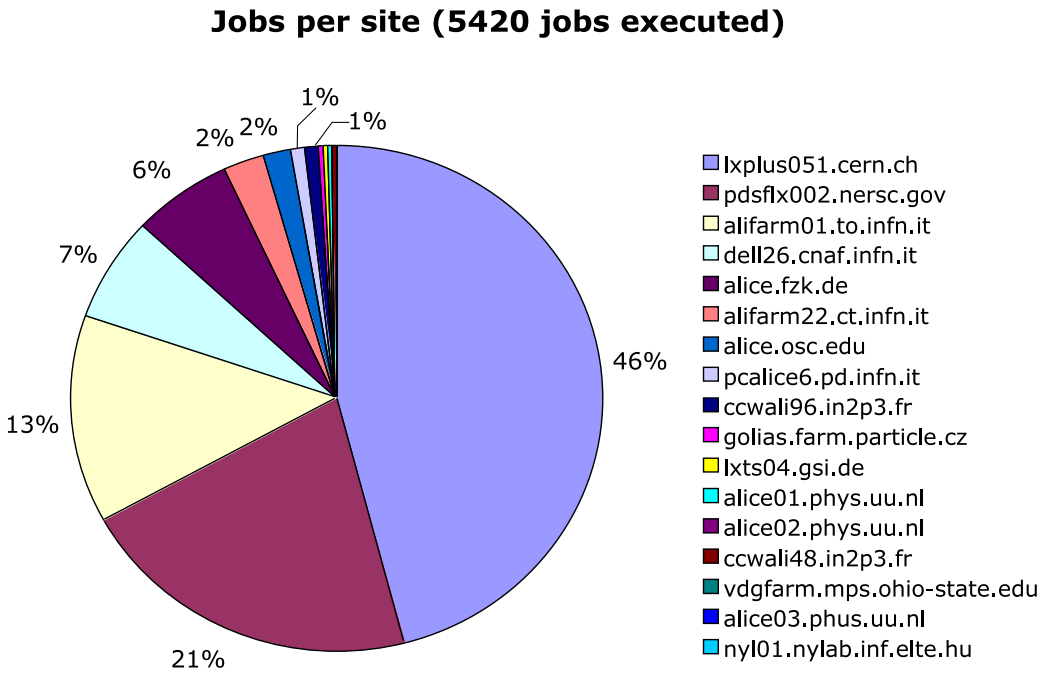

Figure 1.3: Job distribution around participating sites during typical production round.

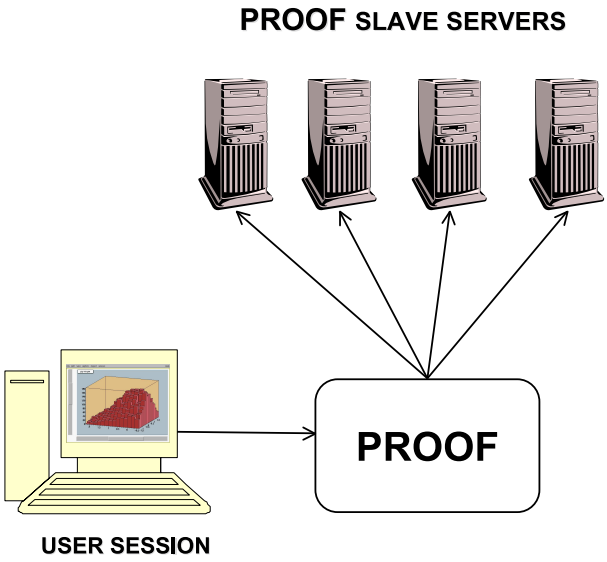

Figure 1.4: Conventional setup of a super-PROOF farm.

event this needs a few MSI2k×s. Each Pb–Pb central event occupies about 2 GB of disk space, while p–p events are two orders of magnitude smaller.

Using AliEn and the ARDA End-to-End ALICE analysis prototype [11] we can solve today the ALICE simulation and reconstruction use cases as well as tackle the problem of distributed analysis on the Grid where we follow two approaches: the asynchronous (interactive batch approach) and the synchronous (true interactive) analysis model.

The asynchronous model has been realized by using AliEn as a Grid framework 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 virtual file catalog 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 AliEn workload management with precise job descriptions. The framework collects and merges available results from all terminated sub-jobs on request.

The synchronous analysis model requires a much tighter integration between the ROOT and AliEn frameworks. This has been achieved by extending the functionality of PROOF [12] – the parallel ROOT facility. Rather than transferring all the input files to a single execution node (farm), it is the program to be executed that is transferred to the nodes where the input is locally accessible and then run in parallel. The interface to Grid-like services is presently being developed, focusing on authentication and the use of file catalogs, in order to make both accessible from the ROOT shell.

In the conventional setup (see Fig. 1.4), PROOF worker servers are managed by a PROOF master server, which distributes tasks and collects results. In a multi-site setup each site running a PROOF environment will be seen as a SuperPROOF worker server for a SuperPROOF master server running on the user machine. The PROOF master server has therefore to implement the functionality of a PROOF master server and a SuperPROOF worker server at the same time. AliEn classes used for asynchronous analysis as described earlier can be used for task splitting in order to provide the input data sets for each site that runs PROOF locally.

1.4 Off-line detector alignment and calibration model

Another major application of the Grid file catalog and ROOT is their use in the ALICE model for alignment and calibration conditions databases. The model stipulates that these objects will be read-only ROOT files, registered in the distributed file catalog. The file content, validity range and versioning will be described as metadata in the file catalog. By adopting this lightweight model, ALICE does not have to develop condition database structure in the 'traditional' sense (using RDBMS technology), thus avoiding functional duplication already provided in ROOT (I/O, persistency) and in the Grid distributed file catalog (metadata capability, local registration and world wide access).

At the moment, we are collecting requirements from the detector groups. These requirements will determine the parameters of the objects: volume, granularity, access patterns, update frequency, runtime environment and the metadata tags needed for the object description. The AliROOT classes needed to access the metadata catalog and retrieve the alignment and calibration objects are being developed. In addition we are finalizing the picture of the relations and access methods of the off-line conditions DB to the databases used in the other ALICE groups: DAQ, Trigger, DCS, ECS and HLT.

1.5 Future of Grid in ALICE

The experience with the AliEn system has been instrumental in shaping the ALICE computing model. We now know that it is technically feasible to run large distributed productions with a very limited number of people and to perform distributed analysis with optimal use of the resources. Of course not all the problems are solved, however we see no stumbling blocks to the realization of a system that would implement a large portion of the *Grid vision*.

AliEn and its architecture has been taken as one the fundamental components on which to build the EGEE middleware, and ALICE has great hopes to be able to transfer its expertise and applications to this new middleware in order to continue the development of its Grid environment and computing model.

At the moment of writing all the former AliEn developers are working for the EGEE project. While this offers good perspectives both for ALICE and for EGEE, it makes ALICE computing partly dependent on the success of this large European project, and this has to be considered a risk factor for our computing model. For the same reasons AliEn is now a "frozen system" and, while providing still an excellent service, it is scheduled to be replaced as soon as possible with the EGEE middleware.

2 Computing model and capacity requirements

2.1 Introduction

This chapter describes a computing model for the processing of the offline data to be produced by ALICE in p–p and A–A every year. As it will be seen, there are several changes with respect to the last official evaluation done during LHC Computing Review (LHCCR) [1]. The model has been refined and developed thanks to the experience gained during the Physics Data Challenges. The permanent storage has increased because a duplication of raw data at external Tier 1's is now foreseen in the model. The disk storage has substantially increased; the new estimate is deduced from the disk usage made during the Physics Data Challenges. This document contains also an evaluation of the Tier 2's contributions, which was not present in the ALICE LHCCR estimates.

2.2 Input parameters

Input parameters for the present computing model are derived from information contained in the ALICE DAQ, HLT and Trigger TDR [2] and the ALICE Physics Performance Report [3]. All input parameters are to be considered as our best estimates at the moment. They are based on the nominal figures of a standard data taking year. The resource requirements during the commissioning of the machine and the two first years of running have been estimated as a percentage of a standard data taking year, taking into account the information available at this moment.

The following values for the LHC design parameters and trigger rate for p–p have been adopted.

Center of mass energy	14 TeV (two 7 TeV proton beams)
Luminosity	0.5×10^{33} cm^{-2}s^{-1} in 2007
	2×10^{33} cm^{-2}s^{-1} in 2008 and 2009
	10^{34} cm^{-2}s^{-1} (design luminosity) from 2010 onward
Total reaction cross section	100 mb
Nominal collision rate	10^{9} Hz p–p collisions at design luminosity

Independently from the LHC p–p luminosity, the beam parameters will be adjusted to achieve a luminosity between 10^{29} and 5×10^{30} which will lead to event rates from 10 kHz to 200 kHz, depending on the events pile up during the ALICE TPC drift time.

For Pb–Pb interactions the following parameters have been adopted:

Center of mass energy	5.5 TeV/n (two 2.7 TeV per nucleon beams)
Luminosity	5×10^{25} cm^{-2}s^{-1} in 2008
	5×10^{26} cm$^{-2}s^{-1}$ (design luminosity) from 2009 onward
Total reaction cross section	8 b
Nominal collision rate	4×10^{3} Hz Pb–Pb collisions at design luminosity

The trigger rate and RAW data bandwidth are independent of the luminosity as trigger thresholds and conditions will be adjusted to saturate the bandwidth. During p–p collisions, more than one bunch crossing will happen during the gating of the TPC due to its long drift time (100 μsec), and therefore more than one collision may be recorded (*pileup*). The entity of this pileup depends on luminosity, and

therefore the event size will increase with the luminosity. Studies are underway to remove these pile-up events on line with the HLT system. However, given the criticality of this operation, we do not foresee to do it in the first years of p–p running.

During the first year of operation, it is further assumed that the first p–p run will take place over a reduced time, from first of July 2007 until end of September 2007. Assuming the maximum luminosity allowed for ALICE, this first run will deliver 40% of the p–p data during a standard data taking year. This run will be followed by a short Pb–Pb pilot run with reduced luminosity which will deliver at most 30% of the Pb–Pb data we will collect in one standard data taking year. The first Pb–Pb run with nominal luminosity will occur at the end of the 2008, the second year of LHC operation. These assumptions justify the rampup of resources deployment required by ALICE: 20% in 2007, 40% in 2008 and 100% installed end of 2008 for the heavy-ion run.

In the following we will discuss about the following type of data:

- **RAW**: raw data as recorded by the DAQ (real) or simulated (Monte Carlo, MC);
- **ESD**: Event Summary Data as produced by the reconstruction program;
- **AOD**: Physics Analysis Object Data derived from the ESD and similar to today's n-tuples;
- **TAG**: Event tags for event selection.

As far as the basic RAW data parameters are concerned, both for p–p and Pb–Pb an average acquisition rate is considered. During p–p run, the rate can go up to 500 Hz, however only in special cases and for a short period of time only. Since for Pb–Pb the event size changes significantly with the centrality of the collision, a normalized rate, taking an average event size of 12.5 MB, is considered. This event size has been calculated assuming a charged particle density of dN/dy = 4000. The actual trigger rate will be the sum of several event types with different sizes and rates.

	p–p	Pb-Pb
Event recording rate (Hz)	100	100
Event recording bandwidth (MB/s)	100	1250
Running time per year (Ms)	10	1
Events per year	10^9	10^8

Table 2.1: Data taking parameters

2.3 The computing model

The computing model is subject to change with time. During the first years of running, even if luminosities might be below the nominal luminosity, only slightly less data will be recorded than during runs with nominal luminosities. The reason is that selective triggers will become operational only after a necessary training period needed to understand the various detection systems. Data from this early period will be processed offline rather differently than data from later runs. Understanding the detectors and training of the alignment, calibration, reconstruction and analysis algorithms will required many processing passes by many physicists on a subset of raw data. This phase must be limited in time and followed by an early fast processing of the total set of available data to guarantee early access to physics results. This fast processing must however be complete to preserve the richness of the data and the discovery potential in the early run. Therefore, adequate computing must already be available when LHC will provide the first p–p collisions.

In the following we discuss an offline processing scenario, from which the required computing resources are estimated, and which we foresee to apply during standard data taking years at nominal luminosities. We assume that a normal run period is split into:

- seven months (effective 10^7 s) of p–p collisions,
- one month (effective 10^6s) of heavy-ion collisions, and
- four months of winter shutdown.

The overall organized processing (calibration, alignment, reconstruction and scheduled analysis) is scheduled and prioritized in the ALICE Physics Working Groups (PWG) and by the ALICE Physics Board. The chaotic processing (chaotic analysis) is organized within the PWGs and, in case of lack of resources, prioritized by the Physics Board.

Depending on the middleware that will be in use at the time LHC is starting, a more or less hierarchical organization of computing resources (Tier 0, Tier 1, Tier 2) will be in order. Although ALICE believes that most likely the democratic cloud model rather than the strict hierarchy of the MONARC model will prevail, the estimate of the required resources uses the concepts and names introduced by MONARC.

- Tier 0 provides long term storage for the raw data, distributes to Tier 1 raw data and performs the calibration and alignment task and the first reconstruction pass;
- Tier 1s provide long term storage of a copy of the raw data, perform the subsequent reconstruction passes, the scheduled analysis tasks, and the reconstruction of Pb–Pb MC data; they have the responsibility of the long term storage of data processed at Tier 1s and Tier 2s;
- Tier 2s generate and reconstruct the simulated Monte-Carlo data and perform the chaotic analysis;
- Tier 0, Tier 1s and Tier 2s provide short term storage with fast access of multiple copies of active data, raw (only a fraction) and processed.

Although the scenario may change to adjust to the available computing resources at a given time or to provide rapid feedback to physics results obtained during the initial reconstruction and analysis, it is unlikely that the estimated required computing resources will vary considerably. Changes of the order of up to 50% in processing, short term storage, or long term storage must be however anticipated. Unexpected features in the operation of the Grid resources, which might result in efficiencies lower than anticipated, and which could not be predicted during the Data Challenges could also require more or less important changes in the needed computing resources.

To estimate the computing resources required by the present model, the efficiency factors for the usage of processors (CPU), short term storage (disk), and long term mass storage (MS) listed in Tab. 2.1 have been adopted. These are factors agreed by the LCG project and used by all the LCG experiments.

Efficiency factors (%)	
Efficiency for scheduled CPU	85
Efficiency for chaotic CPU	60
Disk utilization efficiency	70
MS utilization efficiency	100

Table 2.2: Efficiency factors for the usage of processors (CPU), short term storage (disk), and long term mass storage (MS)

2.4 CPU requirements

2.4.1 Parameters values

The ALICE Data Acquisition system will record p–p data and heavy-ion data at a rate of 100 Hz during the previously discussed effective time of 10^7s and 10^6s respectively (see Tab. 2.1). The raw data size will vary with the trigger selection. The product trigger rate times the raw data size is limited by the maximum bandwidth available to transfer data to the mass storage system, 1.25GB/s.

The processing power required for the reconstruction of data, raw as well as Monte-Carlo, has been estimated from the performance of the reconstruction algorithms presently in use and tested on Monte-Carlo data. We have taken as a value 80% of the present reconstruction times, which takes into account the optimization of the code and its increase in complexity within a reasonable safety factor. For Pb–Pb the average charged particle multiplicity of 4000 has been adopted. This is a conservative estimation based on the recent results from RHIC and the theoretical extrapolations at the LHC energy. The real value might however be different as suggested by the RHIC data. The reconstruction algorithms are not yet entirely optimized, however, on the other hand, they are not complete as they do not yet include calibration or alignment procedures. The figures quoted for reconstruction try to take these factors into account, considering the evolution of the code performance during the past years.

During the first years there will be a full first reconstruction pass followed by a "second pass" which will in reality be composed of several short runs to tune the reconstruction programs based on the results of the first full pass. It is expected that the resources needed by this activity will correspond to a full reconstruction pass. Then we will run a second full reconstruction pass on all data of that year's run with final condition information. As time passes we expect to improve our capability to derive good condition information soon after the first reconstruction pass. However we also expect the need to reconstruct and analyze data coming from previous runs. So the estimation of three reconstruction passes is reasonable over the foreseeable lifetime of the experiment, if compound with the foreseen upgrade of the computing equipment (see 2.7).

The computing power required for analysis has been estimated based on a limited variety of Grid enabled analysis use cases performed in the Physics Data Challenges. Since the analysis algorithms will vary in complexity depending on the physics channel under study, the computing power for analysis will have to be revised once a larger set of analysis classes has been exercised on the Grid. A variation of a factor up to two must be anticipated.

The computing power required for the generation of Monte-Carlo data is better under control. It will be subject to changes only if the real charged particle multiplicity will differ substantially from the one predicted by the Hijing Monte-Carlo generator. Tab. 2.3 lists the values adopted for the various parameters entering our estimation. The CPU power quoted for simulation includes event generation, tracking of particles through the detectors (an average value for the GEANT3 and FLUKA transport codes has been adopted) and digitization of the generated signals.

Chaotic analysis tasks are well focused analysis performed by single users on a subset of the data, usually residing entirely at given Tier2 centers. Scheduled analysis tasks gather many analysis from users and thus require much more CPU power than single analysis tasks. They are performed on the entire set of reconstructed data, once per reconstruction pass. To optimize data transfer, scheduled analysis is best performed at Tier1 centers.

2.4.2 Raw data processing strategy

The scheduled processing of raw data consist in the reconstruction of the raw data and the creation of ESD objects.

Although the processing strategies vary between p–p and heavy-ion runs, they have in common that the first reconstruction pass must be fast and must start latest immediately after (for heavy-ion) or during (for p–p) data taking to allow on one hand for rapid discoveries and to establish quickly the

Processing power parameters			
		pp	HI
Reconstruction	KSI2k×s/event	5.4	68.0
Chaotic analysis	KSI2k×s/event	0.5	7.5
Scheduled analysis	KSI2k×s/event	15.0	230.0
Simulation	KSI2k×s/event	35.0	15000.0
Reconstruction passes		3	3
Chaotic analysis passes		3	3
Schedule analysis passes		20	20

Table 2.3: Computing power required for the processing (simulation includes event generation, tracking and digitization) of the ALICE real and Monte-Carlo data.

gross properties of the collisions. On the other hand, the first reconstruction pass must be finished well before (at least six months) the start of the next heavy-ion run to be able to define from the data analysis the running conditions. For heavy-ion, this first reconstruction is preceded by a computing intensive calibration and alignment task not exceeding one month in duration. The additional reconstruction passes will consist of a full fledged reconstruction including fine tuning of all the parameters of the algorithms.

Calibration and alignment tasks are performed quasi on line on the events stored on a disk buffer at Tier 0. The size of this buffer is the equivalent of 1 day of heavy-ion data taking, ie. 100 TB.

2.4.2.1 p–p data processing

The data collected during pp runs, being less demanding in terms of computing resources than heavy-ion data, will be processed on line or quasi online during data taking (seven months according to our standard data taking year) at the CERN Tier 0. The data temporarily stored on the Tier0 disk buffer (one day of Pb–Pb running is equivalent of 10 days p–p data taking) are processed, successively calibrated and the reconstructed. Reconstructed data will be kept for long term storage locally at Tier 0 and distributed to the Tier 1's. The second and the third reconstruction passes will be distributed in Tier 1's and spread over a period of six months.

2.4.2.2 Heavy-ion data processing

Because of the much larger data recording rate in heavy-ion mode, on line reconstruction of the entire set of data would require unafordable computing resources. The first reconstruction pass will therefore last four months and can start immediately after the data taking ends or even during the run depending on the performance of the calibration and alignment task. It is crucial that this first reconstruction ends well before the next heavy-ion run starts to allow for enough time to analyze the data and elaborate the new running conditions. ALICE estimates that the time required for this is a minimum of six months. The second and third reconstruction passes will be distributed in Tier 1's and can be done over a period of six months. Obviously in such a scenario, p–p and A–A reconstruction passes will be concurrent. Two different reconstruction tasks, one for p–p and one for A–A will be permanently active in Tier1s and at Tier0 either the first p–p or the first A–A reconstruction will be active (see the sketch of a possible processing scenario on figure 2.1).

The resources required to perform the ALICE p–p and heavy-ion data reconstruction within such a scenario are listed in Tab. 2.4.

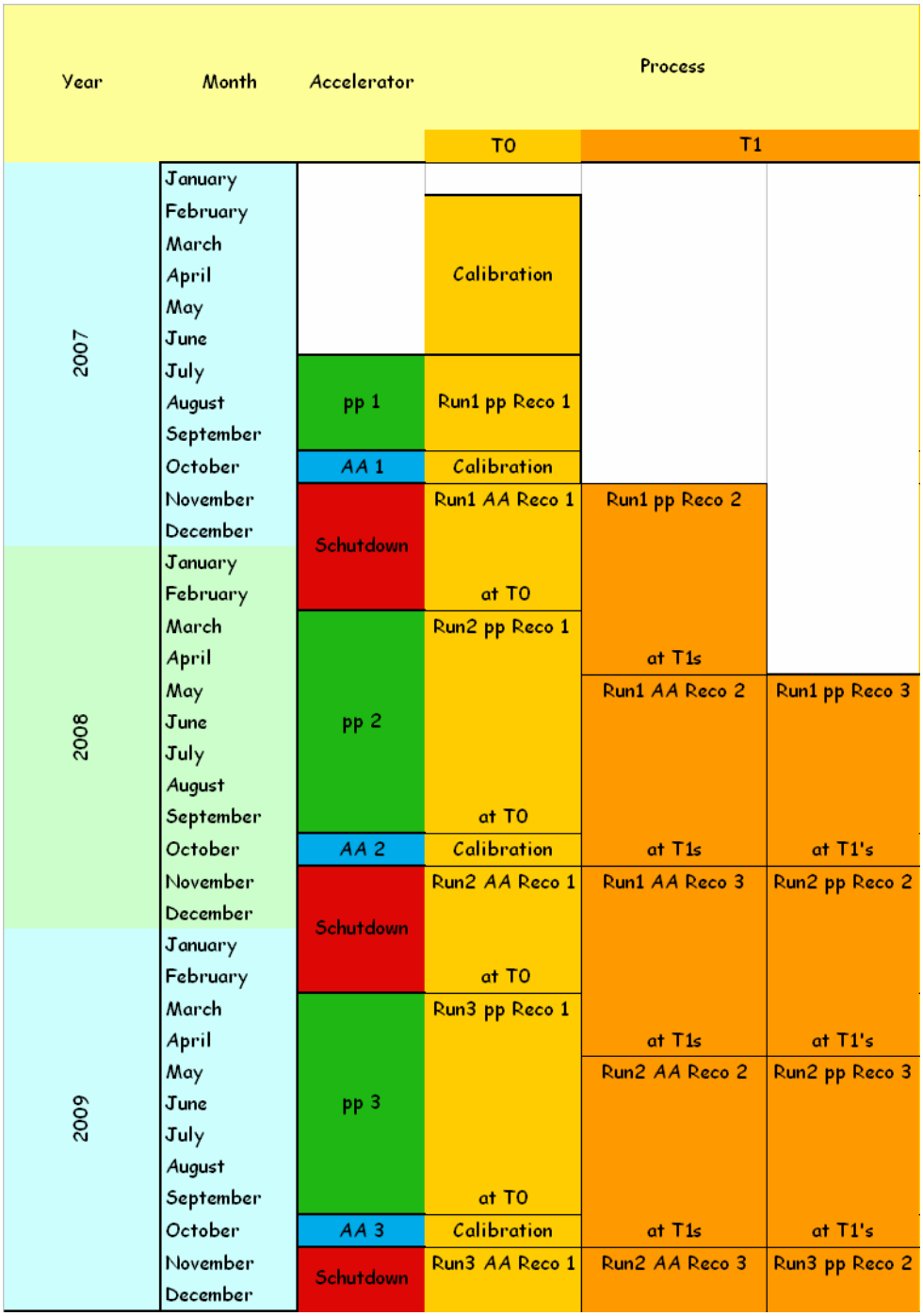

Figure 2.1: Processing scenario used to estimate the resources required at Tier 0, Tier 1's and Tier 2's

	CPU (MSI2K)		
	Integrated	Average	Max
MC (simulation+reconstruction)	150	12.0	12.0
Real reconstruction	94	7.8	13.0
Scheduled analysis	96	8.0	8.4
Chaotic analysis	16	1.3	1.3
Alignment & Calibration	3	0.3	3.0

Table 2.4: Yearly processing resources required to reconstruct and analyze the real p–p and AA data and to process (generation, reconstruction, and analysis) Monte-Carlo data

2.4.3 Monte-Carlo data simulation

The production scheme of Monte-Carlo differs for p–p and heavy-ion simulations. In case of p–p, Monte-Carlo data will be generated in an amount similar to the one of collected real data (10^9 events per year). To avoid requesting a prohibitive amount of resources and also to enrich generated events with physics signal of interest we have adopted a merging technique for heavy-ion simulation. Full fledged heavy-ion events are generated in a limited amount (10^7 events per year) in sets of different impact parameters (minimum bias, peripheral, and central): these events are called the "underlying events". The different signal events are then generated, merged into one underlying event, at the digits level, and the merged event is reconstructed in a single task. In doing so, the underlying events can be reused several times. From the data challenges experience we have fixed to 10, the number of times an underlying event can be reused and preserve realistic event to event fluctuations.

The Monte-Carlo p–p and heavy-ion events are typically generated and reconstructed at the Tier2 centers.

The yearly resources required to process the Monte-Carlo data (generation, reconstruction) are listed in Tab. 2.4.

2.4.4 Data analysis

Based on the experience of past experiments, the data analysis will consume a large fraction of the total amount of resources. At variance with the reconstruction and the simulation tasks, not all the analysis tasks can be easily scheduled as they will involve potentially all the physicists of the collaboration applying customized algorithms many times on subsets of the data of various sizes. This way of performing analysis is called chaotic analysis. However it is foreseen to perform scheduled analysis during the same production tasks as the reconstruction passes and therefore over the entire set of data. This scheduled analysis will group in each pass all the official algorithms. The time needed to analyze reconstructed events depends very much on the complexity of the analysis algorithm and on the overhead introduced by the Grid processing. The latter can only be predicted once the analysis on the Grid has been exercised with the final middleware and at the final scale. The processing power we have considered is an average value which can vary largely from one analysis to the other. To estimate the processing resources required for the analysis we have considered that 200 physicists will exercise many times their algorithms on a small fraction of the data and produce physics results on a larger subset of the data. We furthermore assumed that the various analyzes launched by the physics working groups are performed only once over the entire set of reconstructed data. Scheduled analysis regrouping many analysis tasks will require a larger amount of computing power per event than chaotic analysis. Individuals will also performed analysis using computing resources available at their home institutes. These resources are not accounted for in the present evaluation of the requirements.

The yearly required resources to analyze real data are listed in Tab. 2.4.

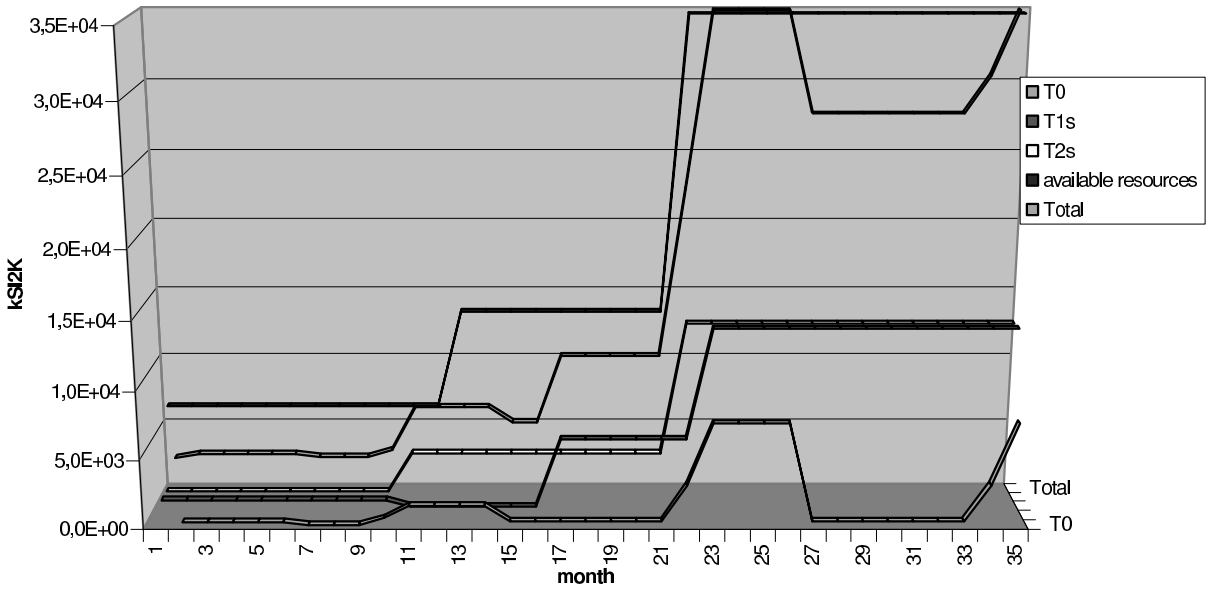

Figure 2.2: Profile of processing resources required at Tier 0, Tier 1's, and Tier 2's to process all ALICE data.

From the scenario sketched in Fig. 2.1, the profile of CPU resources, including the rampup scenario, presented in Fig. 2.2 is deduced.

The final values of resources required during every standard year of running at Tier 0, Tier 1's and Tier 2's are summarized in Tab. 2.6.

2.5 Storage requirements

All data produced by the ALICE detector will be stored permanently for the duration of the experiment. This includes the raw data, one second copy of the raw data and one set of Monte-Carlo data, reconstructed data from all reconstruction passes, analysis objects, calibration and condition data. A fraction of the produced data will be kept on short term storage, providing rapid access to data frequently processed and for I/O dominated tasks. The media for permanent and transient storage will be decided by the technology available. Presently the permanent storage media is magnetic tapes and the transient storage disks. The ratio between disk and tape storage will change with time and will be dictated by the price-performance ratio. The parameters used to estimate the storage requirements are derived from the Data Challenge experience and are reported in Tab. 2.1 and Tab. 2.5.

2.5.1 Permanent storage

Raw data from p–p and heavy-ion runs are copied onto permanent storage at the Tier 0. They are further exported to the external Tier 1's, each one keeping a fraction of raw data on permanent storage, thus providing a second copy. To prevent network interruptions, a disk buffer corresponding to the equivalent

	Real data (MBytes)				Monte-Carlo data (MBytes)	
	Raw	ESD	AOD	Event catalog	Raw	ESD
p–p	1.0[1]	0.04	0.004	0.01	0.4	0.04
Heavy ion	12.5	2.5	0.25	0.01	300	2.5

Table 2.5: Size of raw data produced by the ALICE experiment in p–p and in Pb–Pb runs and by the processing of raw and Monte-Carlo data

of one day of heavy-ion data taking is foreseen. One copy of each reconstruction pass is stored on permanent storage at Tier 0 and external Tier 1's, most likely where they have been generated. One copy of Monte-Carlo data is stored and distributed among the Tier 1's. The yearly requirement for permanent storage in each Tier category is summarized in Tab. 2.6. These values correspond to the needs during a standard year of running. Tier 2's are not required to provide mass storage.

2.5.2 Transient storage

The requirements for transient storage on fast access media depend very much on the computing model, on the balance of available distributed processing resources and on the network bandwidth and occupancy. In the absence of Grid simulation tools, our estimates on needed transient storage capacities are based on the requirement to store on disks data that will be frequently accessed primarily through non scheduled tasks. They include:

- a fraction of the yearly raw data stored at Tier 0 (2%) and at each external Tier 1 (10%);
- one copy of the calibration and condition data at Tier 0 and each external Tier 1;
- reconstructed data (ESD) are replicated with a factor 2 and distributed to Tier 0 and Tier 1's, 2 reconstruction passes
- all analysis objects (AOD, event catalog) distributed in Tier 2's and Tier 1's where they have been produced;
- one copy of generated Monte-Carlo data are distributed in Tier 1's;
- reconstructed Monte-Carlo data are replicated with a factor 2 and distributed to Tier 2's where they have been produced ;

The overall transient storage capacities required during a standard data taking year in each Tier category are listed in Tab. 2.6.

2.6 Network

2.6.1 Tier 0

Data from the ALICE DAQ are transfered to the Tier 0 at a rate of 1.25 GBytes/s during heavy-ion runs and 100 MBytes/s during p–p runs. The DAQ farm provides for sufficient disk storage to buffer the equivalent of 10 hours of data taking. The maximum network bandwidth into Tier 0 is therefore 10 Gb/s during the heavy-ion run and 0.8 Gb/s during the p–p run. The 10 Gb/s lines scheduled to be installed between the ALICE experiment and the Tier 0 and exclusively dedicated to the raw data transfer will be sufficient. A disk buffer of 50 TB is provided by the DAQ farm to store the equivalent of 10 hours of data taking.

Several kind of transfer will contribute to the outgoing traffic from Tier 0 to external Tier 1's:

- the raw p–p data exportation during the seven months of data taking leads to a rate of 55 MBytes/s;
- during the same time ESD data are exported with a rate of 3 MBytes/s;
- the raw heavy-ion data exportation during the month of data taking and the following four months of shutdown leads to a rate of 120 MBytes/s;
- during the four months of shutdown time ESD data are exported with a rate of 26 MBytes/s;

Data will be transfered simultaneously to every Tier 1. This traffic results in a data transfer peak of 150 MBytes/s during the winter shutdown period. These values translate into a maximum outgoing network bandwidth of about 2 Gb/s (average over the year 1 Gb/s). A disk buffer of 100 TB is required at the Tier 0 to store the equivalent of 24 hours of exported Pb–Pb RAW data.

2.6.2 Tier 1

The maximum incoming bandwidth required from Tier 0 in each Tier 1 is deduced from the discussion in the previous section to be 0.3 Gb/s assuming raw data are distributed in six Tier 1s. To the traffic from Tier 0 has to be added the traffic from Tier 2 exporting Monte-Carlo data. Assuming that all Monte-Carlo data are processed in Tier 2s and that there on average 3.5 Tier 2s sending their locally produced data to a given Tier 1, this traffic amounts to 20 MBytes/s. The incoming bandwidth is therefore equal to about 2 Gb/s.

Tier 1 exports ESD data to Tier 2 for analysis. Assuming that all the analysis is done in Tier 2s and that each Tier 1 serves all the Tier 2s, the traffic from one Tier 1 to Tier 2s amounts to 3 MBytes/s and requires a maximum bandwidth of about 0.03 Gb/s.

2.6.3 Tier 2

The incoming and outgoing bandwidth from Tier 2 is deduced from the discussion in the previous section to be 0.01 Gb/s and 0.6 Gb/s respectively.

The requested network bandwidth for the various Tier categories is summarized in Tab. 2.6. It should be noted that we assumed that the network is 100% efficient and operational 100% of the time. The network infrastructure should make provision of upgrades permitted by an evolving technology.

One of the elements that make this evaluation difficult is the duplication factor for disk files. It is clear that we cannot keep all the active data on disk without requiring an unreasonable amount of disk space. Data will be copied from Tier 1's permanent storage onto Tier 2's disk buffers. When the disk buffer is full, data will be removed and then copied again when needed . This may affect the network load between Tier 1's and Tier 2's and also between Tier 2's according to the end-user analysis pattern and the functionality of the middleware performing the file movement.

2.7 Rampup of resources after 2008

ALICE plans to have installed 20% of resources in 2007, 40% in 2008 and 100% end of 2008 for the Pb–Pb run. From there on, as foreseen in the Hoffmann review, ALICE plans on an annual investment of 30% of the value of the installed resources in 2010. Moore's law gives predicts a 40% decrease in price for the same computing power, which means that the power can be increased by 67%, assuming a flat expenditure. In turn, this means that replacing 30% of the capital investment each year provides for a 20% increase on top of the 3-year renewal of capital equipment. This increase would satisfy the ALICE requirements. These arguments hold for CPU. For disks, the reduction is closer to 30% a year, which would provide for a 13% increase each year. This increase is reasonable, given that the data per year will not necessarily increase much, but there will be a global increase in the need to keep old data in disk. For mass storage the reduction is more erratic, with a stepwise behavior and an observed average

	Tier0	Tier 1s	Tier 2s	Total
CPU (MSI2k)	7.5	13.8	13.7	35.00
Transient storage (PBytes)	0.1	7.5	2.6	10.2
Permanent storage (PBytes/year)	2.3	7.5	-	9.8
Bandwidth in (Gb/s)	10	2	0.01	
Bandwidth out (Gb/s)	1.2	0.02	0.6	

Table 2.6: Summary of the computing resources required by the ALICE computing model during a standard data taking year. The data for Tier 1s and Tier 2s include the Tier 1 rand Tier 2 resources ar CERN.

decrease of 10% per year. This gives a modest 3% increase in the mass storage capabilities, over the linear increase. Again ALICE can cope with such a moderate increase, because this cannot be calculated yearly, as moving to a new technology would mean migrate all the data, and can be considered only a few times in the life of the experiment.

2.8 Summary

The computing resources requirements in each Tier category are summarized in Tab. 2.6. Within our resources deployment scenario the values during the first year of LHC (2007) and the second year (2008) the requirements represent 20% and 40% respectively of the final values required during the third year of LHC running. The peak value of the processing resources at Tier 0 is fixed by the requirement to perform the first reconstruction pass of the heavy-ion runs over a time period not exceeding 4 months after the end of the runs (see Fig. 2.1). The peak in the processing resources at Tier 1's occurs when heavy-ion data and p–p data are reconstructed in parallel (see Fig. 2.1). Tier1's perform in addition the scheduled analysis and the merging and reconstruction of heavy-ion Monte-Carlo events. The processing resources at Tier 2's are fixed by the requirement to process all the Monte-Carlo data (except the reconstruction of heavy-ion events) and the end-user analysis tasks. The repartition of resources between the different Tier 1's and the different Tier 2's will be adjusted later when the amount of resources pledged to ALICE will be known. The requirement for permanent storage is not likely to change with time whereas the disk storage capacity requirements might change depending on the performance of the upcoming mass storage systems on one hand and on the performance of the GRID on the other.

References

Chapter 1

[1] http://www.cern.ch/MONARC/

[2] P. Saiz *et al.*, Nucl. Instrum. Methods **A502** (2003) 437-440;
http://alien.cern.ch/

[3] The Grid: Blueprint for a New Computing Infrastructure I. Foster, C. Kesselmann (Eds), M. Kaufmann, 1999;
I. Foster, C. Kesselman, S. Tuecke, The Anatomy of the Grid Enabling Scalable Virtual Organizations, see, http://www.globus.org/research/papers/anatomy.pdf

[4] http://www.eu-datagrid.org/
http://www.gridpp.ac.uk/
http://www.ppdg.net/
http://www.griphyn.org/
http://www.ivdgl.org/
http://www.infn.it/grid/
http://www.eu-datagrid.org

[5] http://www.globus.org/

[6] P. Buncic, Proc. of Computing in High Energy and Nuclear Physics, La Jolla, California (2003), http://www.slac.stanford.edu/econf/C0303241/proc/papers/MOAT004.PDF

[7] Condor Classified Advertisements, http://www.cs.wisc.edu/condor/classad

[8] http://lcg.web.cern.ch/LCG/SC2/

[9] http://lcg.web.cern.ch/LCG/lw2002/

[10] http://lhc-computing-review-public.web.cern.ch/

[11] http://lcg.web.cern.ch/LCG/peb/arda/Default.htm

[12] M. Ballintijn, R. Brun, F. Rademakers and G. Roland, *Distributed Parallel Analysis Framework with PROOF*, Proc. of TUCT004.

[13] http://glite.web.cern.ch/glite/

[14] http://egee-intranet.web.cern.ch/egee-intranet/gateway.html

Chapter 2

[1] LHC Computing Review, CERN/LHCC/2001-004

[2] ALICE Collaboration, *Technical Design Report of Trigger, Data Acquisition, High-Level Trigger and Control System*, CERN/LHCC/2003–062.

[3] CERN/LHCC 2003-049
ALICE PPR Volume 1 (7 November 2003)