9

Grid Computing in High Energy Physics Experiments

Dagmar Adamová[1] and Pablo Saiz[2]
[1]*Nuclear Physics Institute, Řež near Prague*
[2]*CERN*
[1]*Czech Republic*
[2]*Switzerland*

1. Introduction

The High Energy Physics (HEP) [1] – often called Particle Physics – is one of the research areas where the accomplishment of scientific results is inconceivable without the infrastructure for distributed computing, the Computing Grid. The HEP is a branch of Physics that studies properties of elementary subatomic constituents of matter. It goes beyond protons and neutrons to study particles which existed only a fraction of a second after the Big Bang and quarks and gluons in the so-called Quark Gluon Plasma (QGP) [2]. These studies are based on experiments with particles colliding at very high energies, at speeds almost equal to the speed of light.

The world's leading Particle Physics research laboratory is CERN [3], the European Center for Nuclear and Particle Physics near Geneva, Switzerland. The CERN latest particle accelerator (see Fig. 1), the Large Hadron Collider (LHC) [4], installed in a 27 km long tunnel located about 100 m underground and crossing the Swiss - French border, uses counter-rotating beams of protons or lead ions (Pb) to collide at 4 points, inside large particle detectors: ALICE [5], ATLAS [6], CMS [7] and LHCb [8]. There are also another two smaller experiments, TOTEM [9] and LHCf [10]. These are much smaller in size and are designed to focus on

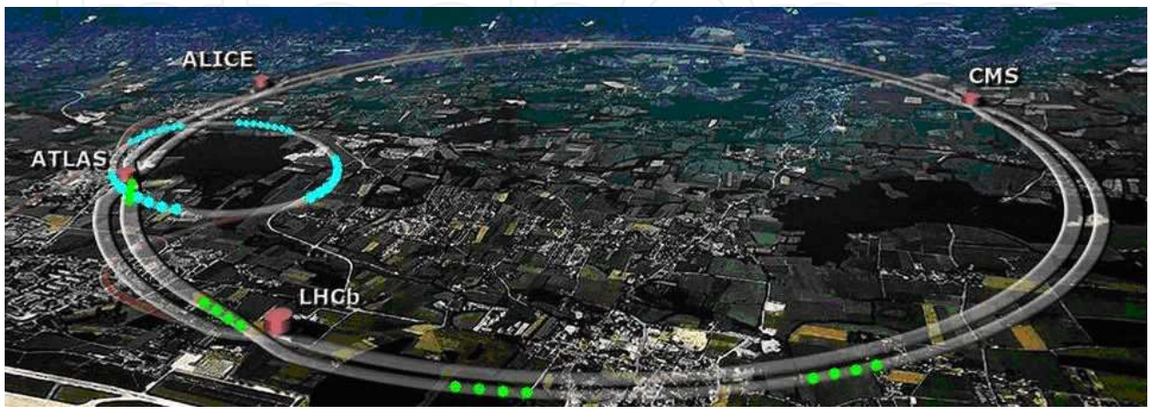


Fig. 1. LHC@CERN

"forward particles". These are particles that just brush past each other as the beams collide, rather than meeting head-on.

The energy of the protons is currently 3.5 TeV (1 Tera eV= 1 million MeV) and that of the Pb ions is 1.38 TeV, so the collision energies are 7 TeV for the protons and 2.76 TeV for the Pb ions.

The phrase often used to summarize the mission of the LHC is, that with the LHC we are going back in time very close to the Big Bang, as close as about 10^{-10} seconds. In terms of length it represents about 10^{-16} cm (compared to the dimensions of the Universe of about 10^{28} cm). At this scale, the matter existed in a form of a "soup" made of the quarks and gluons, the Quark Gluon Plasma. The quarks are objects protons and neutrons are made of, so the LHC represents in a sense a huge extremely complicated microscope enabling the study of the most basic elements of matter.

There are several major questions which scientists hope to get answered with the help of the LHC.

- What is the origin of mass, why do elementary particles have some weight? And why do some particles have no mass at all? At present, we have no established answers to these questions. The theory offering a widely accepted explanation, the Standard Model [11], assumes the existence of a so-called Higgs boson, a key particle undiscovered so far, although it was first hypothesized in 1964. One of the basic tasks of the LHC is to bring an established statement concerning the existence of the Higgs boson.
- Where did all the anti-matter disappear? We are living in the World where everything is made of matter. We suppose that at the start of the Universe, equal amounts of matter and antimatter were produced in the Big Bang. But during the early stages of the Universe, an un-known deviation or in-equilibrium must have happened, resulting in the fact that in our world today hardly any antimatter is left.
- What are the basic properties of the Quark-Gluon Plasma, the state of the matter existing for a tiny period of time after the Big Bang? Originally, we thought it would behave like a plasma, but the latest scientific results including those delivered by the LHC suggest that it behaves like a perfect liquid [2], which is somewhat surprising for us.
- What is the universe made of? At the moment, the particles that we understand create only 4 % of the universe. The rest is believed to be made out of dark matter and dark energy. The LHC experiments will look for supersymmetric particles, which would confirm a likely hypothesis for the creation of dark matter.

From the experiments analyzing the data from the LHC collisions, ATLAS and CMS are the largest. They were nominally designed to look for the Higgs boson but in fact these are general purpose detectors for the study of all kinds of Physics phenomena at the LHC energy range. The ALICE detector is a dedicated heavy ions detector to study the properties of the Quark Gluon Plasma formed in the collisions of lead ions at the LHC energies. The LHCb is much smaller detector and its mission is to study the asymmetry between matter and antimatter. Although all these experiments are designed for Particle Physics research, the scientific programs they follow actually cross a border between Particle Physics, Astrophysics and Cosmology.

Now, where does the Computing Grid show up in this scientific set-up? The LHC is the world's largest particle accelerator. The protons and lead ions are injected into the accelerator

in bunches, in counter-rotating beams. According to the original design proposal, there should be 2808 bunches per a beam. Each bunch of protons contains 10^{11} protons. the design beam energy is 7 TeV and the design luminosity is 10^{34} cm^{-2}s^{-1}. The bunch crossing rate is 40 MegaHz and the proton collisions rate $10^7 - 10^9$ Hz.

However, the new phenomena looked for by the scientists appear at a rate of 10^{-5} Hz. So the physicists must analyze 10^{13} collision events/sec to have a chance to discover a New Physics phenomenon. At present, the machine has not yet reached the full number of bunches per beam and is operating at half of the originally proposed energy, but the luminosity is getting rapidly to the goal value. The LHC team has been increasing the number of bunches gradually reaching 1380 bunches/beam at the time of writing. The full beam energy will be reached in 2014, after one year of a technical stop to arrange for this increase. The machine has already beaten some world records which we will mention in section 5. Let us just mention the one concerning the stored energy: at the end of 2010, the energy stored in the accelerator ring was about 28 MegaJoules (MJ). At the target full intensity, this energy will reach about 130 MJ which is an equivalent of 80 kg of TNT.

In any case, the volume of data necessary to analyze to discover New Physics was already in the original proposal estimated to be about 15 PetaBytes (PB, 1PB=1 million GB) per data taking year. The number of the processor cores, CPUs, needed to process this amount of data was estimated to be about 200 thousands. And here, the concept of a distributed data management infrastructure comes into the scenario, because there is no single computing center within the LHC community/collaboration to offer such massive computing resources, even not CERN. Therefore in 2002, the concept of the Worldwide LHC Computing Grid (WLCG) [12] was launched to build a distributed computing infrastructure to provide the production and analysis environments for the LHC experiments.

In the present chapter, we give a short overview of the Grid computing for the experiments at the LHC and the basics of the mission of the WLCG. Since we are members of the ALICE collaboration, we will also describe some specific features of the ALICE distributed computing environment.

In section 2, we will describe the architecture of the WLCG, which consist of an agreed set of services and applications running on the Grid infrastructures provided by the WLCG partners. In section 3, we will mention some of the middleware services provided by the WLCG which are used for the data access, processing, transfer and storage. Although WLCG depends on the underlying Internet - computer and communications networks, it is the special kind of software, so-called middleware, that enables the user to access computers distributed over the network. It is called “middleware” because it sits between the operating systems of the computers and the Physics applications that solve particular problems. In section 4, the Computing model of the ALICE experiment will be briefly described. It provides guide lines for the implementation and deployment of the ALICE software and computing infrastructure over the resources within the ALICE Grid and includes planning/estimates of the amount of needed computing resources. Section 5 will be devoted to the ALICE-specific Grid services and the ALICE Grid middleware AliEn. It is a set of tools and services which represents an implementation of the ALICE distributed computing environment integrated in the WLCG environment. In section 6, an overview will be given of the experience and performance of the WLCG project and also of the ALICE Grid project in particular during the real LHC data taking. The continuous operation of the LHC started in November 2009.

When the data started to flow from the detectors, the distributed data handling machinery was performing almost flawlessly as a result of many years of a gradual development, upgrades and stress-testing prior to the LHC startup. As a result of the astounding performance of WLCG, a significant number of people are doing analysis on the Grid, all the resources are being used up to the limits and the scientific papers are produced with an unprecedented speed within weeks after the data was recorded.

Section 7 contains a short summary and an outlook. This chapter is meant to be a short overview of the facts concerning the Grid computing for HEP experiments, in particular for the experiments at the CERN LHC. The first one and half a year of the LHC operations have shown that WLCG has built a true, well functioning distributed infrastructure and the LHC experiments have used it to rapidly deliver Physics results. The existing WLCG infrastructure has been and will be continuously developing into the future absorbing and giving rise to new technologies, like the advances in networking, storage systems, middleware services and inter-operability between Grids and Clouds.

2. WLCG

As mentioned in section 1, the LHC experiments are designed to search for rare events with the signal/noise ratio as low as 10^{-13}. This Physics requires a study of enormous number of pp and Pb-Pb collisions resulting in the production of data volumes of more than 10 PetaBytes per one data taking year. The original estimates elaborated when the LCG TDR [13] was put together were about $\sim$ 15 PetaBytes (PB) of new data each year which translates into $\sim$ 200 thousands of CPUs/processor cores and 45 PB of disk storage to keep the raw, processed and simulated data.

Nowadays, 200 thousands cores does not sound like much and one can find them in large computer centers. 50 PB of a disk storage is however not that common. In any case, at the time the LHC Computing Grid was launched there was no single site within the LHC community able to provide such computing power. So, the task of processing the LHC data has been a distributed computing problem right from the start.

2.1 WLCG mission

The Worldwide LHC Computing Grid (WLCG) project [13] was launched in 2002 to provide a global computing infrastructure to store, distribute and process the data annually generated by the LHC. It integrates thousands of computers and storage systems in hundreds of data centers worldwide, see Figure 2. CERN itself provides only about 20% of the resources needed to manage the LHC data. The rest is provided by the member states' national computing centers and research network structures supported by national funding agencies. The aim of the project is the "collaborative resource sharing" between all the scientists participating in the LHC experiments, which is the basic concept of a Computing Grid as defined in [14]. The infrastructure is managed and operated by a collaboration between the experiments and the participating computer centers to make use of the resources no matter where they are located.

The collaboration is truly worldwide: it involves 35 countries on 5 continents and represents 49 funding agencies having signed the WLCG Memorandum of Understanding on Computing (WLCG MoUC) [15]. The distributed character has also a sociological aspect: even if the contribution of the countries depends on their capabilities, a member of any institution involved can access and analyze the LHC data from his/her institute.

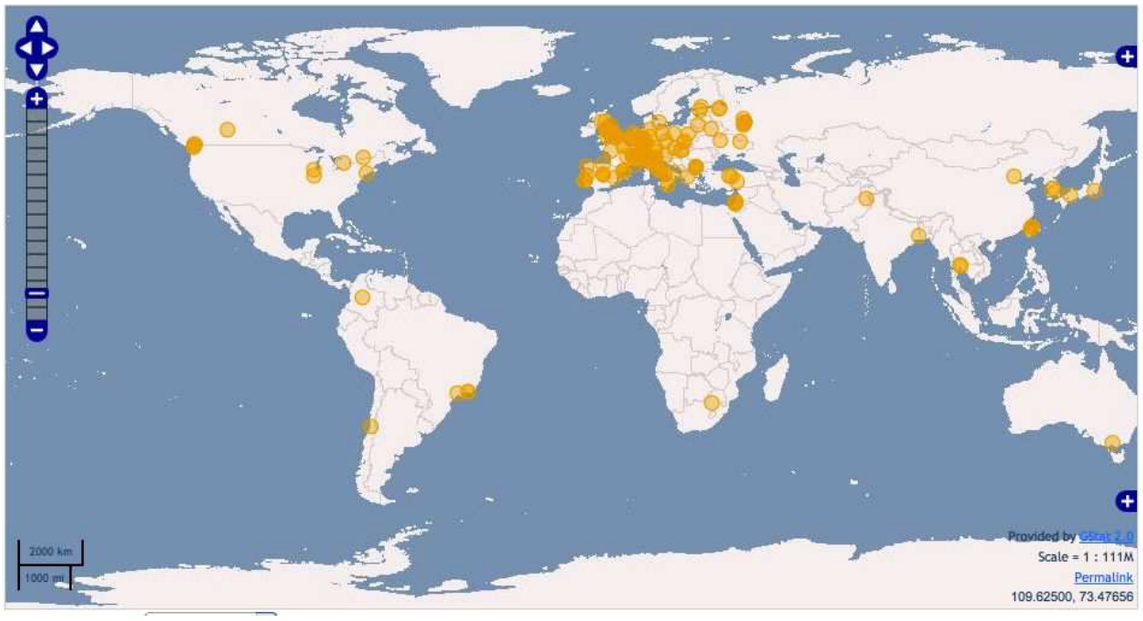


Fig. 2. Distribution of WLCG computing centers

Currently, the WLCG integrates over 140 computing sites, more than 250 thousands CPU cores and over 150 PB of disk storage. It is now the world's largest computing grid: the WLCG operates resources provided by other collaborating grid projects: either the two main global grids, EGI [16] and OSG [17], or by several regional or national grids.

2.2 Hierarchical (Tier) structure, the roles of different Tier-sites

The WLCG has a hierarchical structure based on the recommendations of the MONARC project [18], see Figure 3. The individual participating sites are classified according to their resources and level of provided services into several categories called Tiers. There is one Tier-0 site which is CERN, then 11 Tier-1 centers, which are large computing centers with thousands of CPUs, PBs of disk storage, tape storage systems and 24/7 Grid support service (Canada: TRIUMF, France: IN2P3, Germany: KIT/FZK, Italy: INFN, Netherlands: NIKHEF/SARA, Nordic countries: Nordic Datagrid Facility (NDGF), Spain: Port d'Informació Científica (PIC), Taipei: ASGC, United Kingdom: GridPP, USA: Fermilab-CMS and BNL ATLAS). Then there are currently about 140 Tier-2 sites covering most of the globe. The system also recognizes Tier-3 centers, which are small local computing clusters at universities or research institutes.

The raw data recorded by the LHC experiments (raw data) is shipped at first to the CERN Computing Center (CC) through dedicated links. CERN Tier-0 accepts data at average of 2.6 GBytes(GB)/s with peaks up to 11 GB/s. At CERN, the data is archived in the CERN tape system CASTOR [19] and goes through the first level of processing - the first pass of reconstruction. The raw data is also replicated to the Tier-1 centers, so there are always 2 copies of the raw data files. CERN serves data at average of 7 GB/s with peaks up to 25 GB/s [20]. The Tier-0 writes on average 2 PB of data per month to tape in pp running, and double that in the 1 month of Pb-Pb collisions, (cf. Figures 4,5). At Tier-1 centers, the raw data replicas are permanently stored as mentioned before and several passes of the data re-processing are performed. This multiple-stage data re-processing is performed using methods to detect

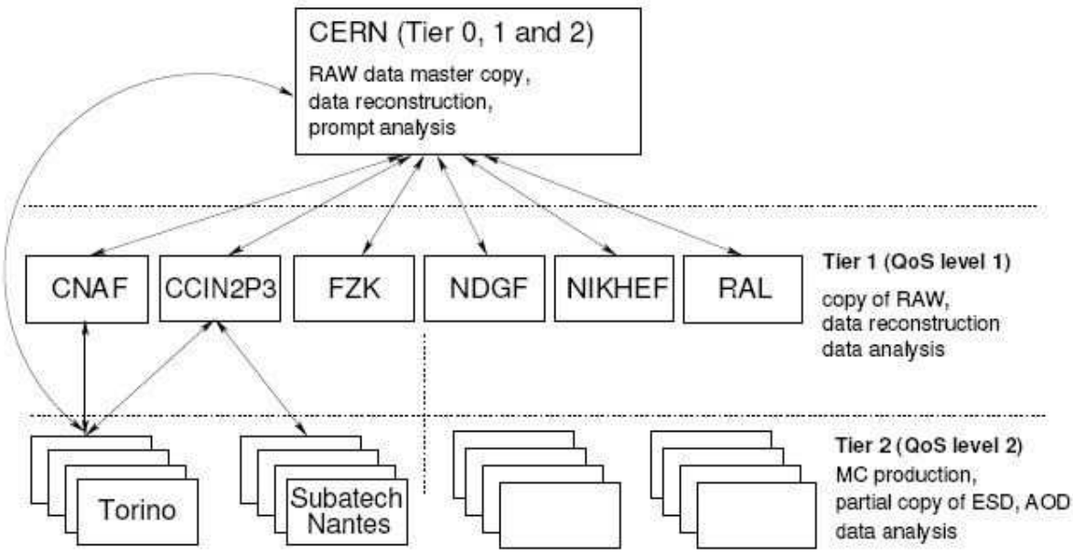


Fig. 3. Schema of the hierarchical Tier-like structure of WLCG

interesting events through the processing algorithms, as well as improvements in detector calibration, which are in continuous evolution and development. Also, the scheduled analysis productions as well as some of the end user analysis jobs are performed at Tier-1s.

Tier-2 centers (more than 130 in the WLCG, integrated within 68 Tier-2 federations) are supposed to process simulation (Monte Carlo simulations of the collision events in the LHC detectors) and end-user analysis jobs. The load of simulations needed to correctly interpret the LHC data is quite sizeable, close to the raw data volume. The number of end users regularly using the WLCG infrastructure to perform analysis is larger than expected in the beginning of the LCG project, it varies from about 250 to 800 people depending on the experiment. This is certainly also a result of the experiments' effort to hide the complexity of the Grid from the users and make the usage as simple as possible. Tier-2 sites deliver more than 50% of the total CPU power within the WLCG, see Figure 6.

2.3 Network

The sustainable operation of the data storing and processing machinery would not be possible without a reliable network infrastructure. In the beginning of the WLCG project there were worries that the infrastructure would not be able to transfer the data fast enough. The original estimates of the needed rate were about 1.3 GB/s from CERN to external Tiers. After the years spent with building the backbone of the WLCG network, CERN is able to reach rates about 5 GB/s to Tier-1s, see Figure 7. The WLCG networking relies on the Optical Private Network (OPN) backbone [21], see Figure 8, which is composed of dedicated connections between CERN Tier-0 and each of the Tier1s, with the capacity of 10 Gbit/s each. The original connections proliferated into duplicates or backroutes making the system considerably reliable. The OPN is then interconnected with national network infrastructures like the GEANT [22] in Europe or the US-LHCNet [23] and all the National Research and Education Networks (NRENs) in other countries.

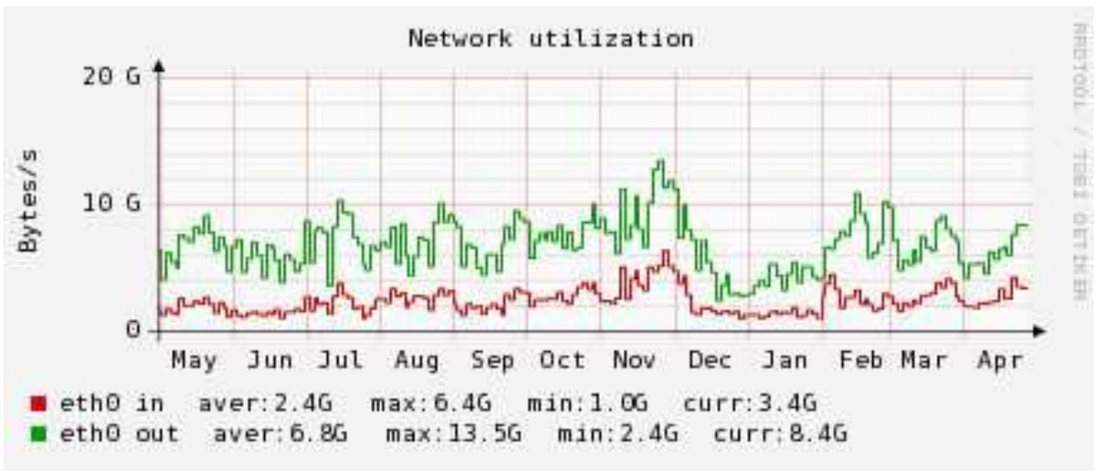


Fig. 4. CERN Tier-0 Disk Servers (GB/s), 2010/2011

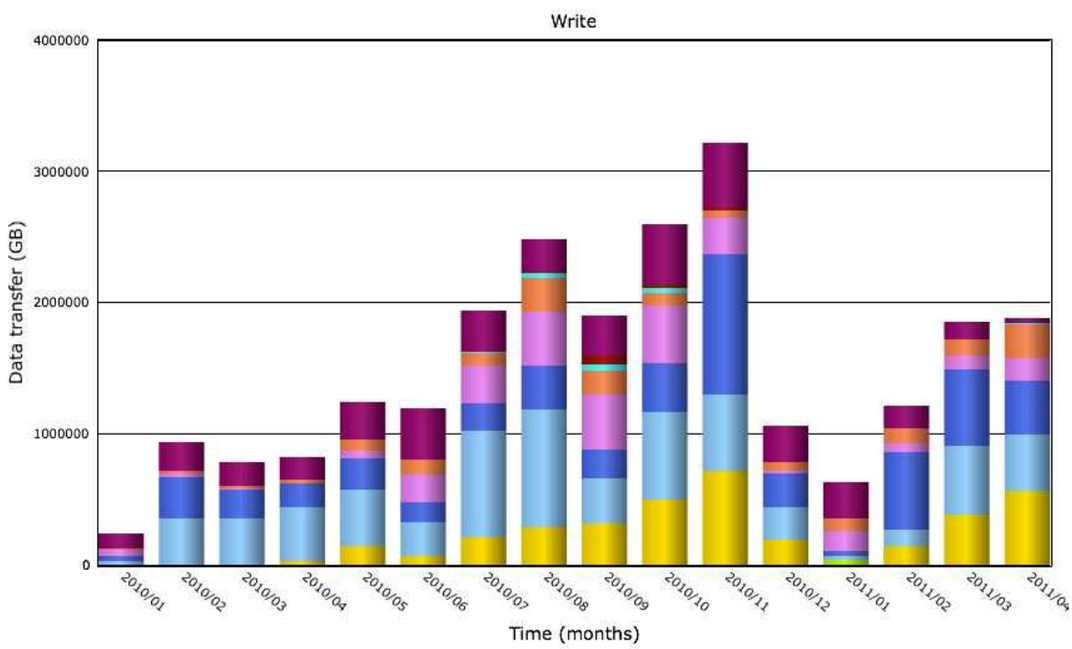


Fig. 5. Data written to tape at the CERN Tier-0 (GB/month)

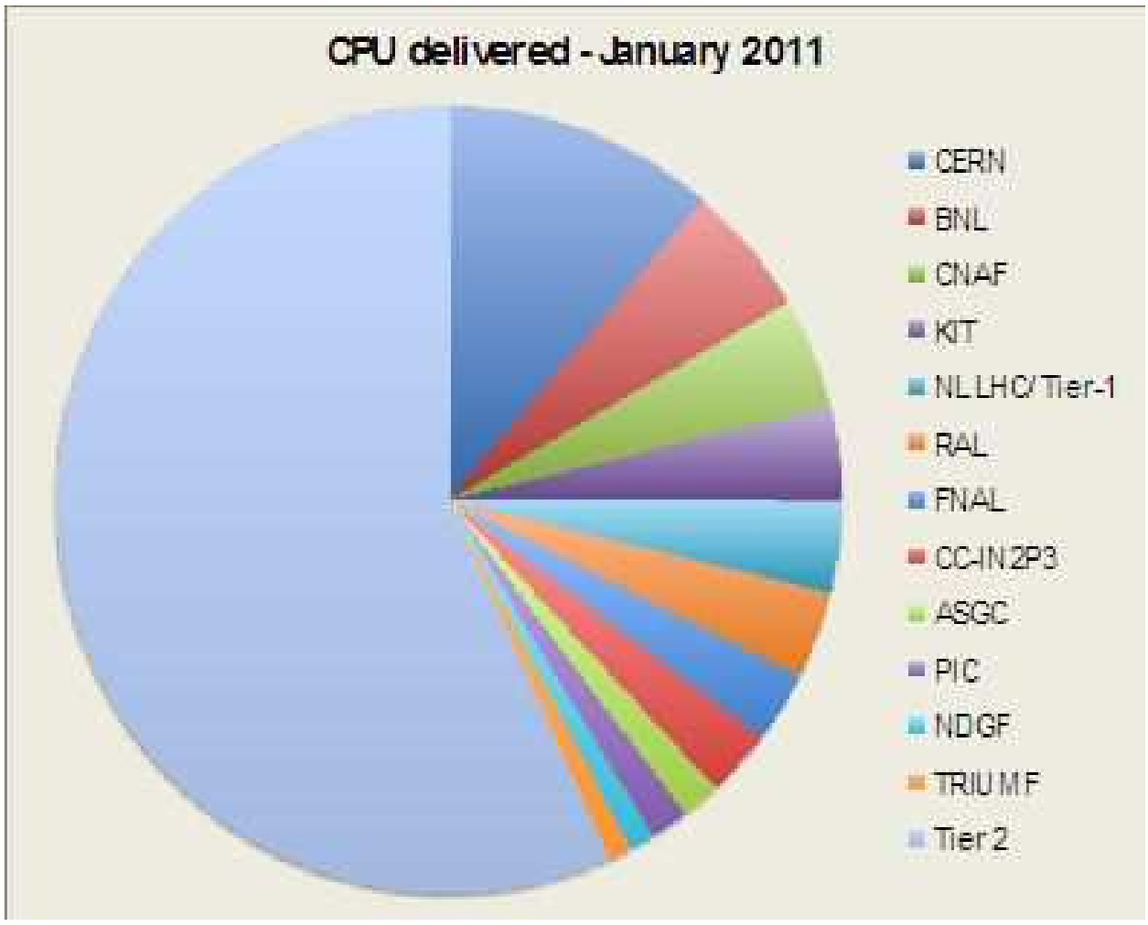


Fig. 6. CPU resources in WLCG, January 2011. More than 50% was delivered by Tier-2s.

There exists a concept of so-called LHCONE [24], which should enable a good connectivity of Tier-2s and Tier-3s to the Tier-1s without overloading the general purpose network links. It will extend and complete the existing OPN infrastructure to increase the interoperability of all the WLCG sites.

2.4 Data and Service challenges

As we will describe in section 6, the WLCG data management worked flawlessly when the real data started to flow from the detectors in the end of 2009. This was not just a happy coincidence. There were over 6 years of continuous testing of the infrastructure performance. There was a number of independent experiments' so-called Data Challenges which started in 2004, when the "artificial raw" data was generated in the Monte Carlo productions and then processed and managed as if it was the real raw data. Moreover, there was a series of WLCG Service Challenges also starting in 2004, with the aim to demonstrate WLCG services aspects: data management, scaling of job workloads, security incidents, interoperability, support processes and all was topped with data transfers exercise lasting for weeks. The last test was the Service Challenge STEP'09 including all experiments and testing full computing models. Also, the cosmic data taking which started in 2008 has checked the performance of the data processing chain on a smaller scale.

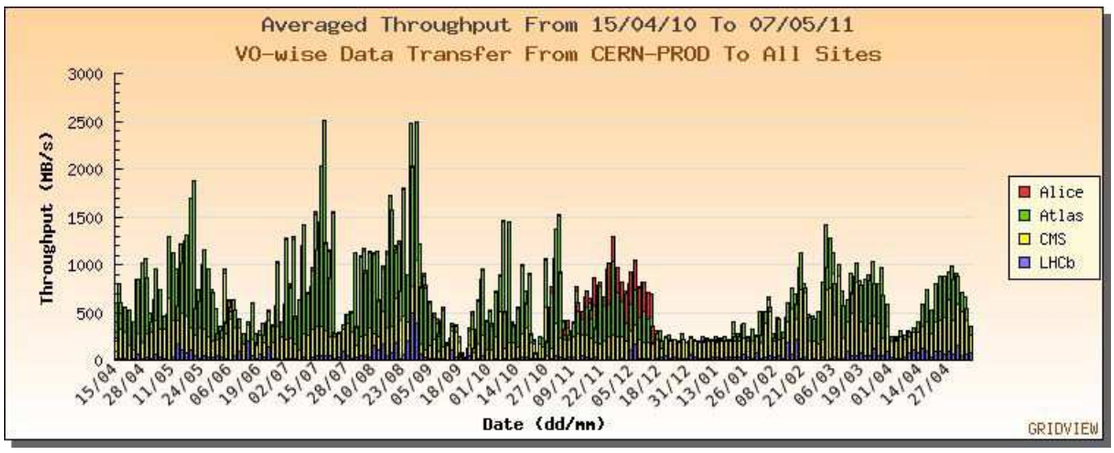


Fig. 7. CPU resources in WLCG, January 2011. More than 50% was delivered by Tier-2s.

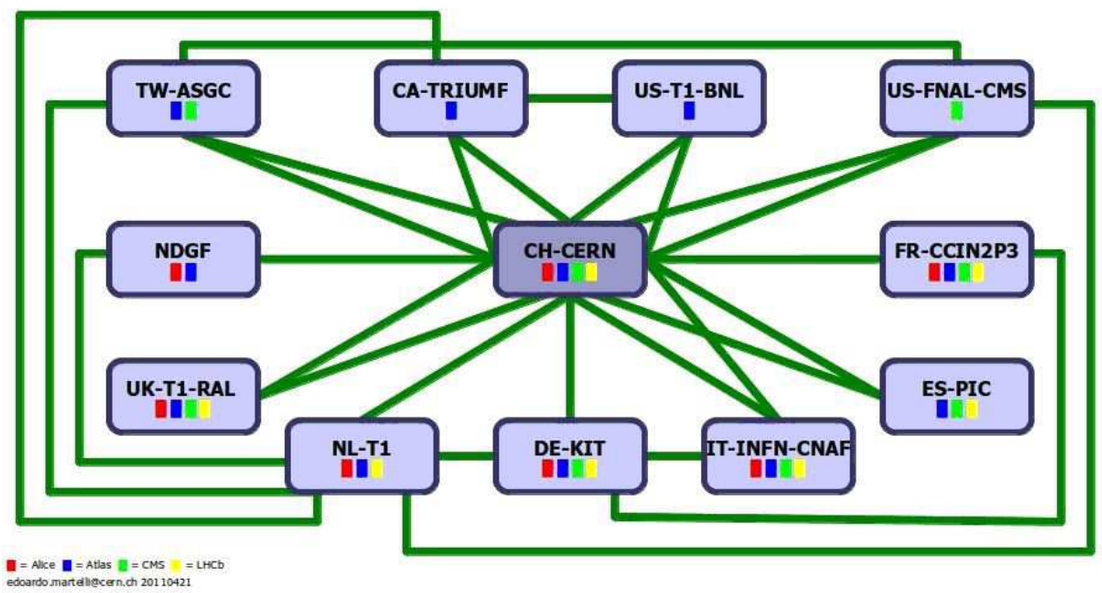


Fig. 8. LHCOPN

Currently, whether the data taking is going on or not, the network, especially the OPN, and the sites are under continuous checking: there are automatically generated test jobs periodically sent over the infrastructure to test the availability and functioning of the network and on-site services.

2.5 Concluding remarks

The WLCG integrates and operates resources distributed all over the world and its task is to make all these resources accessible and usable for the LHC experiments to distribute, archive and process the data produced by the LHC. This task is done using a specialized software called "middleware" because it sits between the operating systems of the computers at the WLCG sites and the Physics applications software used for the reconstruction, analysis and simulation of the LHC data (or any other application software layer). The middleware is a collection of protocols, agents and programs/services which we describe in the following section.

3. Middleware

As we already mentioned, the Worldwide LHC Computing Grid is a distributed computing infrastructure that spans over five continents managing resources distributed across the world (due to funding, operability and access reasons). The resources operated by the WLCG belong either to the two main global grids, EGI [16] and OSG [17], or to other collaborating regional or national grids. To make this diverse variety of resources globally available for all the WLCG users, the WLCG has been developing its own middleware, a software layer that "brings all the resources together": a collection of programs, services and protocols to manage and operate the entire WLCG infrastructure (see Figure 9).

3.1 Overview of Grid services

The WLCG middleware is a complex suite of packages which includes (see also Figure 10):

- Data Management Services:
  - Storage Element
  - File Catalogue Service
  - Grid file access tools
  - File Transfer Service
  - GridFTP service
  - Database and DB Replication Services
  - POOL Object Persistency Service
- Security Services:
  - Certificate Management Service
  - Virtual Organization [25] Management Registration Service (VOMRS)
  - Authentication and Authorization Service (the X509 infrastructure)
- Job Management Services:
  - Compute Element
  - Workload Management

Fig. 9. Grid Layers

  - Service VO Agent Service
  - Application Software Install Service
- Information Services:
  - Accounting Service
  - Site Availability Monitor
  - Monitoring tools: experiment dashboards; site monitoring

The WLCG middleware has been built and further developed using and developing some packages produced by other projects including, e.g.:

- EMI (European Middleware Initiative) [26], combining the key middleware providers of ARC, gLite, UNICORE and dCache

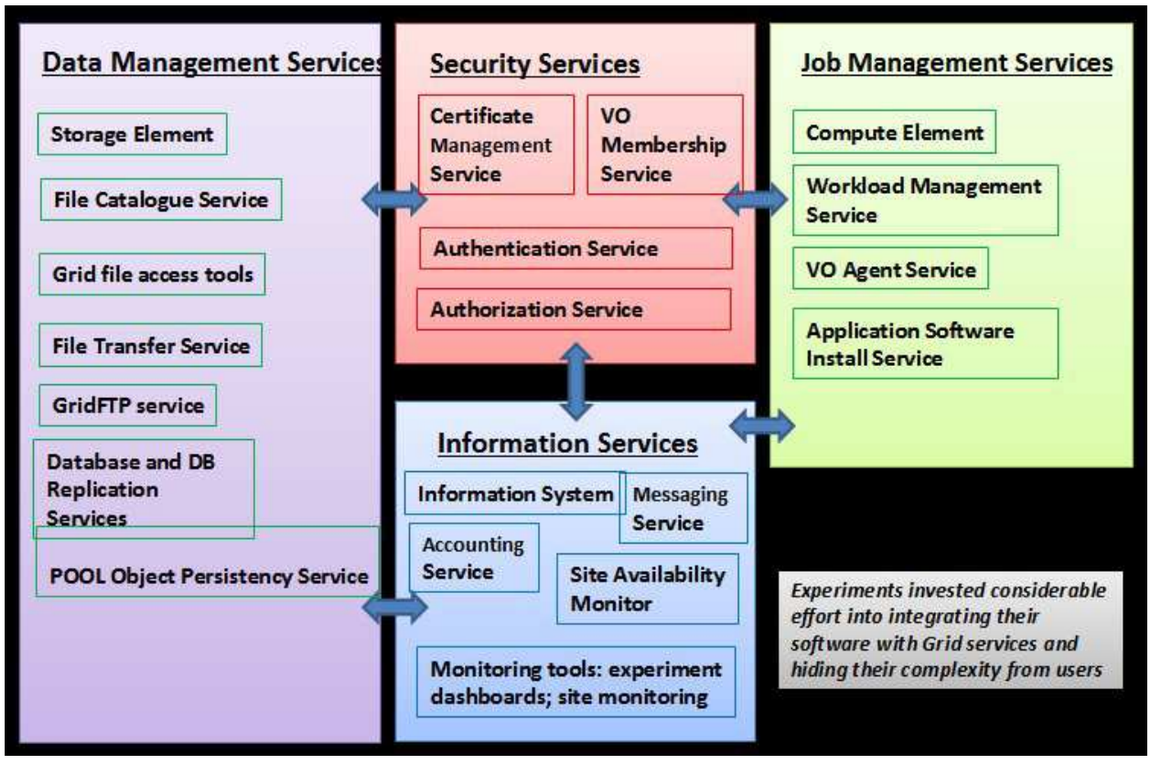


Fig. 10. Schema of Grid services

- Globus Toolkit [27] developed by the Globus Alliance
- OMII from the Open Middleware Infrastructure Institute [28]
- Virtual Data Toolkit [29]

3.2 Experiments' specific developments

All the LHC experiments created their own specific Computing models summarized in the individual Computing Technical Design Reports (TDRs). They do not rely only on the WLCG-provided middleware packages but are also developing some specific components tailored to better comply with their Computing models.

For example, the ALICE experiment has developed a grid middleware suite AliEn (AliCE Environment [30]), which provides a single interface for a transparent access to computing resources for the ALICE community. AliEn consists of a collection of components and services which will be described in the next section. AliEn, together with selected packages of the WLCG-provided middleware, gives a complete framework to the ALICE community to manage and process the data produced by the LHC according to the ALICE Computing model.

All the LHC experiments invested a considerable effort into shielding the users from the underlying complexity of the Grid machinery, trying to provide relatively simple entry points into the Grid. This effort has payed off and is reflected in a considerable number of physicists actually using the WLCG for their analysis.

3.3 Selected WLCG-provided services

In the following section, we will describe as an example the Computing model of the ALICE experiment. The WLCG services used in this model include the Computing Element (CE), the Storage Element (SE) and the VOBOX.

3.3.1 Computing Element

The Computing Element (CE) [31] is a middleware component/grid service providing an entry point to a grid site. It authenticates users and submits jobs to Worker Nodes (WN), aggregates and publishes information from the nodes. It includes a generic interface to the local cluster called Grid Gate (GG), Local Resource Management System (LRMS) and the collection of Worker Nodes.

Originally, the submission of jobs to CEs was performed by the Workload Management System (WMS) [32], a middleware component/grid service, that also monitors jobs status and retrieves their output. WLCG (gLite) CE is a computing resource access service using standard grid protocols. To improve the performance, the CREAM (Computing Resource Execution And Management) Computing Element [33] has replaced the gLite-CE in production since about 2009. It is a simple, lightweight service for job management operation at the Computing Element level. CREAM-CE accepts job submission requests (described with the same files as used for the Workload Management System) and other job management requests like, e.g., job monitoring. CREAM-CE can be used by a generic client, e.g., an end-user willing to directly submit jobs to a CREAM-CE, without the WMS component.

3.3.2 Storage Element (XRootD)

The Storage Element (SE) [34] provides storage place and access for data. Important variables apart from available storage space, read/write speeds and bandwidth concern reliability against overload, percentage of failed transfers from/to SE and percentage of lost/corrupted files.

WLCG (gLite) provides dCache [35] and DPM [36] storage management tools used by the LHC experiments. However within the ALICE infrastructure, the preferred storage manager is the Scalla/XRootD package [37] developed within a SLAC [38] - CERN collaboration (originally, it was a common project of SLAC and INFN [39]). After CERN got involved, the XRootD was bundled in ROOT [40] as a generic platform for distributed data access, very well suited for the LHC data analysis.

The primary goal has been the creation of data repositories with no reasonable size limit, with high data access performance and linear scaling capabilities. The framework is a fully generic suite for fast, low latency and scalable data access, which can serve any kind of data, organized as a hierarchical filesystem-like namespace, based on the concept of directory.

“xrootd” is just the name of the data access daemon. Although fundamental, it is just a part of the whole suite. The complete suite is called Scalla/XRootD, Scalla meaning Structured Cluster Architecture for Low Latency Access.

The manager exhibits important features including:

- High speed access to experimental data

- High transaction rate with rapid request dispersement (fast open, low latency)
- Write once read many times processing mode
- Fault tolerance (if servers go, the clients do not die)
- Fault tolerance (able to manage in realtime distributed replicas)
- Integrated in ROOT

From the site administrator point, the following features are important:

- No database requirements (no backup/recovery issues, high performance)
- Resources gentle, high efficiency data server (low CPU/byte overhead, small memory footprint)
- Simple installation
- Configuration requirements scale linearly with site complexity
- No 3rd party software needed (avoids messy dependencies)
- Low administration costs
- Self-organizing servers remove need for configuration changes in big clusters

Additional features:

- Generic Mass Storage System Interface (HPSS, CASTOR, etc)
- Full POSIX access
- Server clustering for scalability, supports large number of clients from a small number of servers
- Up to 262000 servers per cluster
- High WAN data access efficiency (exploit the throughput of modern WANs for direct data access, and for copying files as well)

3.3.3 VOMS

VOMS [41] stands for Virtual Organization Management Service and is one of the most commonly used Grid technologies needed to provide user access to Grid resources. It works with users that have valid Grid certificates and represents a set of tools to assist authorization of users based on their affiliation. It serves as a central repository for user authorization information, providing support for sorting users into a general group hierarchy - users are grouped as members of Virtual Organizations (VOs). It also keeps track of users' roles and provides interfaces for administrators to manage the users. It was originally developed for the EU DataGrid project.

3.3.4 VOBOX

The VOBOX [42] is a standard WLCG service developed in 2006 in order to provide the LHC experiments with a place where they can run their own specific agents and services. In addition, it provides the file system access to the experiment software area. This area is shared between VOBOX and the Worker Nodes at the given site. In the case of ALICE, the VOBOX is installed at the WLCG sites on dedicated machines and its installation is mandatory for sites

to enter the grid production (it is an "entry door" for a site to the WLCG environment). The access to the VOBOX is restricted to the Software Group Manager (SGM) of the given Virtual Organization. Since 2008, this WLCG service has been VOMS-aware [42]. In the following section, we will describe the services running on the VOBOX machines reserved at a site for the ALICE computing.

4. ALICE computing model

In this section, we will briefly describe the computing model of the ALICE experiment [43].

ALICE (A Large Ion Collider Experiment) [5] is a dedicated heavy-ion (HI) experiment at the CERN LHC which apart from the HI mission has also its proton-proton (pp) Physics program. Together with the other LHC experiments, ALICE has been successfully taking and processing pp and HI data since the LHC startup in November 2009. During the pp running, the data taking rate has been up to 500 MB/s while during the HI running the data was taken with the rate up to 2.5 GB/s. As was already mentioned, during 2010 the total volume of data taken by all the LHC experiments reached 15 PB, which corresponds to 7 months of the pp running and 1 month of the HI running (together with 4 months of an LHC technical stop for maintenance and upgrades this makes up for one standard data taking year (SDTY)).

The computing model of ALICE relies on the ALICE Computing Grid, the distributed computing infrastructure based on the hierarchical Tier structure as described in section 2. ALICE has developed over the last 10 years a distributed computing environment and its implementation: the Grid middleware suite AliEn (AliCE Environment) [30], which is integrated in the WLCG environment. It provides a transparent access to computing resources for the ALICE community and will be described in the next section.

4.1 Raw data taking, transfer and registration

The ALICE detector consists of 18 subdetectors that interact with 5 online systems [5]. During data taking, the data is read out by the Data Acquisition (DAQ) system as raw data streams produced by the subdetectors, and is moved and stored over several media. On this way, the raw data is formatted, the events (data sets containing information about individual pp or Pb-Pb collisions) are built, the data is objectified in the ROOT [40] format and then recorded on a local disk. During the intervals of continuous data taking called runs, different types of data sets can be collected of which the so-called PHYSICS runs are those substantial for Physics analysis. There are also all kinds of calibration and other subdetectors' testing runs important for the reliable subsystems operation.

ALICE experimental area (called Point2 (P2)) serves as an intermediate storage: the final destination of the collected raw data is the CERN Advanced STORage system (CASTOR) [19], the permanent data storage (PDS) at the CERN Computing center. From Point2, the raw data is transferred to the disk buffer adjacent to CASTOR at CERN (see Figure 11). As mentioned before, the transfer rates are up to 500 MB/s for the pp and up to 2.5 GB/s for the HI data taking periods.

After the migration to the CERN Tier-0, the raw data is registered in the AliEn catalogue [30] and the data from PHYSICS runs is automatically queued for the Pass1 of reconstruction, the first part of the data processing chain, which is performed at the CERN Tier-0. In parallel with the reconstruction, the data from PHYSICS runs is also automatically queued for the

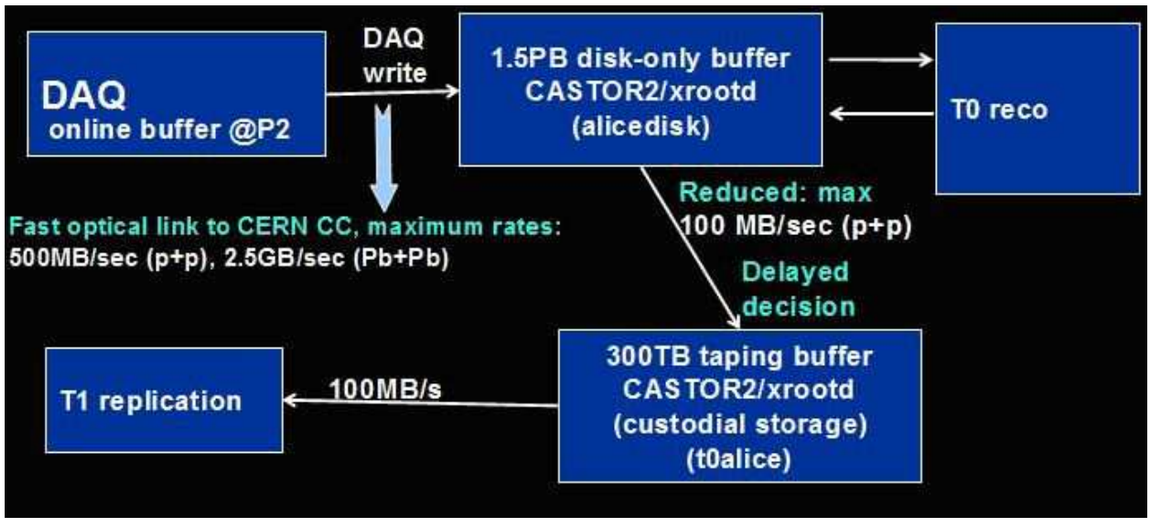


Fig. 11. Data processing chain. Data rates and buffer sizes are being gradually increased.

replication to external Tier-1s (see Figure 11). It may happen that the replication is launched and finished fast and the data goes through the first processing at a Tier-1.

The mentioned automated processes are a part of a complex set of services deployed over the ALICE Computing Grid infrastructure. All the involved services are continuously controlled by automatic procedures, reducing to a minimum the human interaction. The Grid monitoring environment adopted and developed by ALICE, the Java-based MonALISA (MONitoring Agents using Large Integrated Services Architecture) [44], uses decision-taking automated agents for management and control of the Grid services. For monitoring of raw data reconstruction passes see [45].

The automatic reconstruction is typically completed within a couple of hours after the end of the run. The output files from the reconstruction are registered in AliEn and are available on the Grid (stored and accessible within the ALICE distributed storage pool) for further processing.

4.2 AliRoot

AliRoot [46] is the ALICE software framework for reconstruction, simulation and analysis of the data. It has been under a steady development since 1998. Typical use cases include detector description, events generation, particle transport, generation of "summable digits", event merging, reconstruction, particle identification and all kinds of analysis tasks. AliRoot uses the ROOT [40] system as a foundation on which the framework is built. The Geant3 [47] or FLUKA [48] packages perform the transport of particles through the detector and simulate the energy deposition from which the detector response can be simulated. Except for large existing libraries, such as Pythia6 [49] and HIJING [50], and some remaining legacy code, this framework is based on the Object Oriented programming paradigm and is written in C++.

AliRoot is constituted by a large amount of files, sources, binaries, data and related documentation. Clear and efficient management guidelines are vital if this corpus of software should serve its purpose along the lifetime of the ALICE experiment. The corresponding policies are described in [51]. For understanding and improvement of the

AliRoot performance, as well as for understanding the behavior of the ALICE detectors, the fast feedback given by the offline reconstruction is essential.

4.3 Multiple reconstruction

In general, the ALICE computing model for the pp data taking is similar to that of the other LHC experiments. Data is automatically recorded and then reconstructed quasi online at the CERN Tier-0 facility. In parallel, data is exported to the different external Tier-1s, to provide two copies of the raw data, one stored at the CERN CASTOR and another copy shared by all the external Tier-1s.

For HI (Pb-Pb) data taking this model is not viable, as data is recorded at up to 2.5 GB/s. Such a massive data stream would require a prohibitive amount of resources for quasi real-time processing. The computing model therefore requires that the HI data reconstruction at the CERN Tier-0 and its replication to the Tier-1s be delayed and scheduled for the period of four months of the LHC technical stop and only a small part of the raw data (10-15%) be reconstructed for the quality checking. In reality, comparatively large part of the HI data (about 80%) got reconstructed and replicated in 2010 before the end of the data taking due to occasional lapses in the LHC operations and much higher quality of the network infrastructure than originally envisaged.

After the first pass of the reconstruction, the data is usually reconstructed subsequently more times (up to 6-7 times) for better results at Tier-1s or Tier-2s . Each pass of the reconstruction triggers a cascade of additional tasks organized centrally like Quality Assurance (QA) processing trains and a series of different kinds of analysis trains described later. Also, each reconstruction pass triggers a series of the Monte Carlo simulation productions. All this complex of tasks for a given reconstruction pass is launched automatically as mentioned before.

4.4 Analysis

The next step in the data processing chain is then the analysis. There are two types of analysis: a scheduled analysis organized centrally and then the end-user, so-called chaotic analysis. Since processing of the end-user analysis jobs often brings some problems like a high memory consumption (see Figure 12) or unstable code, the scheduled analysis is organized in the form of so-called analysis trains (see [52]). The trains absorb up to 30 different analysis tasks running in succession with one data set read and with a very well controlled environment. This helps to consolidate the end-user analysis.

The computing model assumes that the scheduled analysis will be performed at Tier-1 sites, while the chaotic analysis and simulation jobs will be performed at Tier-2s. The experience gained during the numerous Data Challenges, the excellent network performance, the stable and mature Grid middleware deployed over all sites and the conditions at the time of the real data taking in 2010/2011 progressively replaced the original hierarchical scenario by a more "symmetric" model often referred to as the "cloud model".

4.5 Simulations

As already mentioned, ever since the start of building the ALICE distributed computing infrastructure, the system was tested and validated with increasingly massive productions

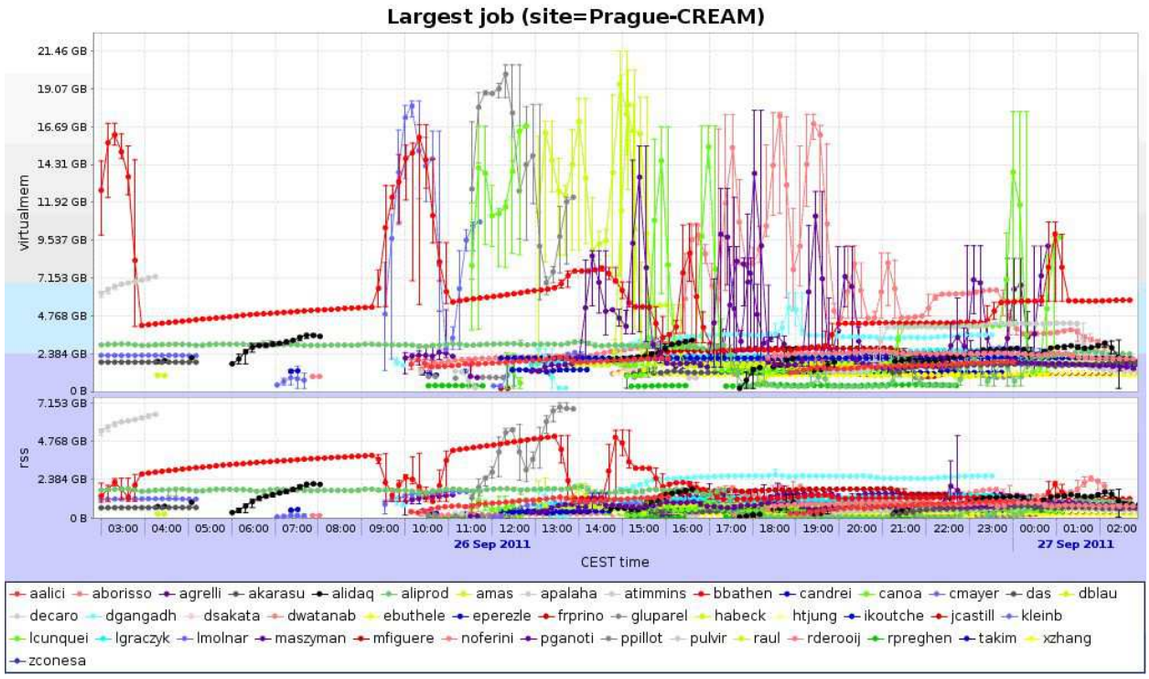


Fig. 12. End-user analysis memory consumption: peaks in excess of 20 GB

of Monte Carlo (MC) simulated events of the LHC collisions in the ALICE detector. The simulation framework [53] covers the simulation of primary collisions and generation of the emerging particles, the transport of particles through the detector, the simulation of energy depositions (hits) in the detector components, their response in form of so-called summable digits, the generation of digits from summable digits with the optional merging of underlying events and the creation of raw data. Each raw data production cycle triggers a series of corresponding MC productions (see [54]). As a result, the volume of data produced during the MC cycles is usually in excess of the volume of the corresponding raw data.

4.6 Data types

To complete the description of the ALICE data processing chain, we will mention the different types of data files produced at different stages of the chain (see Figure 13).

As was already mentioned, the data is delivered by the Data Acquisition system in a form of raw data in the ROOT format. The reconstruction produces the so-called Event Summary Data (ESD), the primary container after the reconstruction. The ESDs contain information like run and event numbers, trigger class, primary vertex, arrays of tracks/vertices, detector conditions. In an ideal situation following the computing model, the EODs should be of 10% size of the corresponding raw data files.

The subsequent data processing provides so-called Analysis Object Data (AOD), the secondary processing product, which are data objects containing more skimmed information needed for final analysis. According to the Computing model, the size of AODs should be 2% of the raw data file size. Since it is difficult to squeeze all the information needed for the Physics results in such small data containers, this limit was not yet fully achieved.

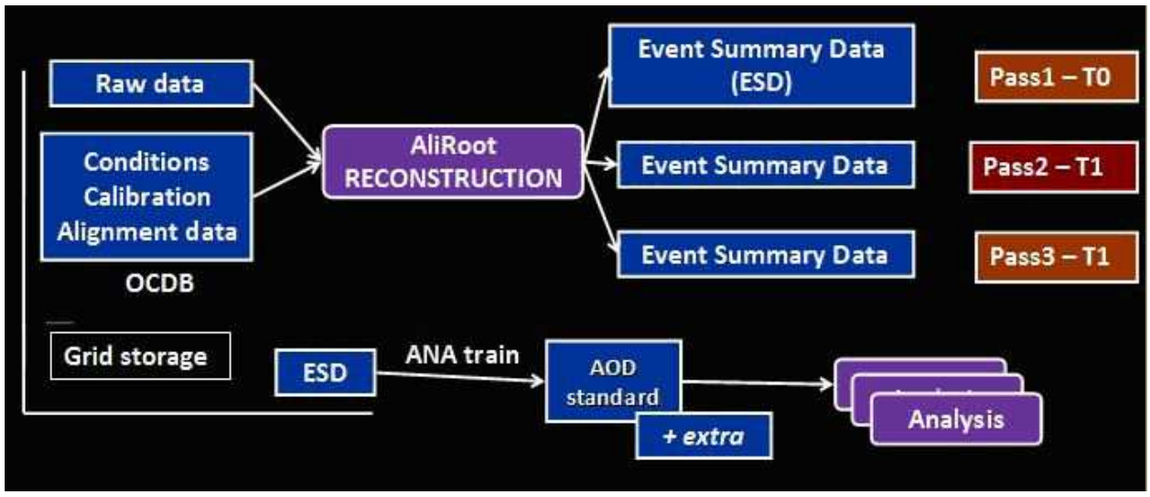


Fig. 13. Data types produced in the processing chain

4.7 Resources

The ALICE distributed computing infrastructure has evolved from a set of about 20 computing sites into a global world-wide system of distributed resources for data storage and processing. As of today, this project is made of over 80 sites spanning 5 continents (Africa, Asia, Europe, North and South America), involving 6 Tier-1 centers and more than 70 Tier-2 centers [55], see also Figure 14. Altogether, the resources provided by the ALICE

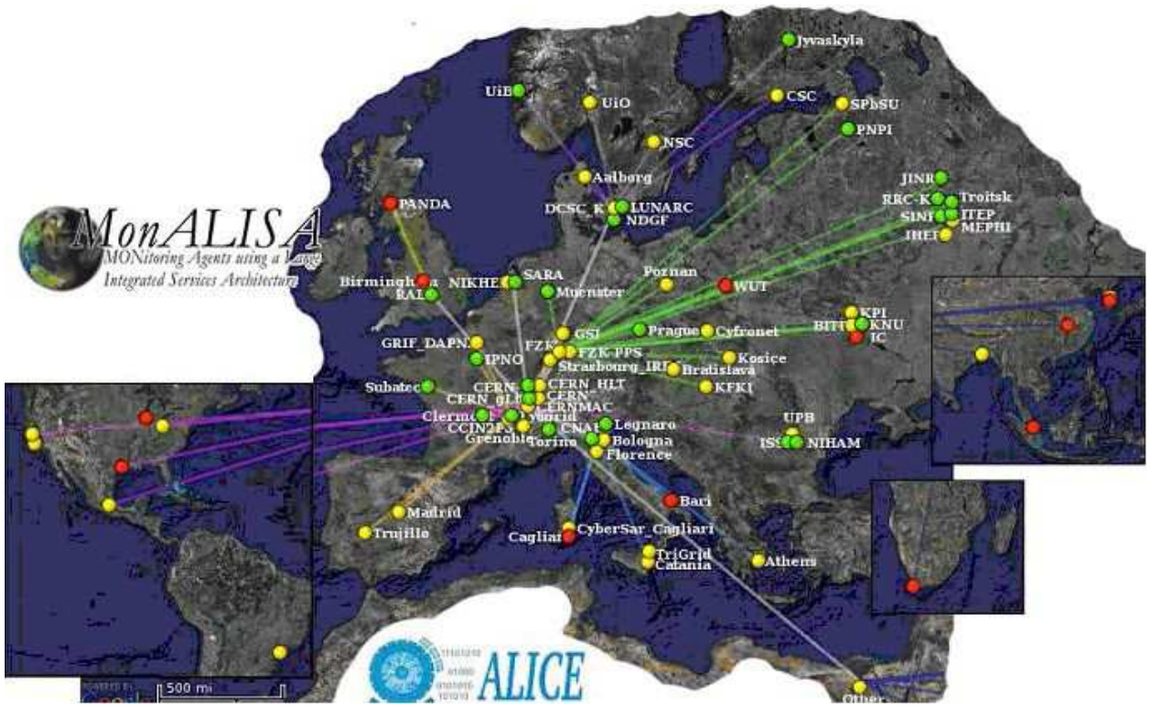


Fig. 14. ALICE sites

sites represent in excess of 20 thousands of CPUs, 12 PB of distributed disk storage and 30 PB of distributed tape storage, and the gradual upscale of this capacity is ongoing. Similar to other LHC experiments, about half of the CPU and disk resources is provided by the Tier-2 centers. For the year 2012, ALICE plans/requirements for computing resources within WLCG represent 211.7 of kHEP-SPEC06 CPU capacity, 38.8 PB of disk storage and 36.6 PB of tapes [56].

4.8 Concluding remarks

The concept of the ALICE computing model was officially proposed in 2005. Since then, it has been used for massive Monte Carlo event productions, for end-user analysis and for the raw data management and processing. The strategy has been validated under heavy load during a series of Data Challenges and during the real data taking in 2010/2011. The model provides the required Grid functionality via a combination of the common Grid services offered on the WLCG resources and the ALICE-specific services from AliEn. Today's computing environments are anything but static. Fast development in Information Technologies, commodity hardware (hardware being constantly replaced and operating systems upgraded), Grid software and networking technologies inevitably boosted also further development of the ALICE computing model. One of the main effects is a transformation of the model from the strictly hierarchical Tier-like structure to a more loose scenario, a "cloud-like" solution.

5. AliEn

AliEn [30] is a set of middleware tools and services which represents an implementation of the ALICE distributed computing environment integrated in the WLCG environment. AliEn has been under a constant development by ALICE since 2001 and was deployed over the ALICE Grid infrastructure right from the start. One of the important features is the set of interfaces to other Grid implementations like gLite [57], ARC [58] and OSG [17].

AliEn was initially developed as a distributed production environment for the simulation, reconstruction, and analysis of Physics data. Since it was put in the production in 2001, ALICE has been using AliEn before the start of the real data taking for distributed production cycles of Monte-Carlo simulated raw data, including subsequent reconstruction and analysis, during the regular Physics Data Challenges. Since 2005, AliEn has been used also for end-user analysis. Since December 2007, when the ALICE detector started operation taking cosmic data, AliEn has been used also for management of the raw data. Since the LHC startup in 2009, millions of jobs have been successfully processed using the AliEn services and tools.

AliEn developers provided the users with a client/interface - "alien shell" [59] and a set of plugins designed for the end users' job submission and handling. These tools together with the tools provided by the ALICE Grid monitoring framework MonALISA [44], hide the complexity and heterogenity of the underlying Grid services from the end-user while facing the rapid development of the Grid technologies.

AliEn is a lightweight Open Source Grid framework built around Open Source components using the combination of standard network protocols, a Web Service and Distributed Agent Model. The basic AliEn components include:

- AliEn File Catalogue with metadata capabilities
- Data management tools for data transfers and storage

- Authentication, authorization and auditing services
- Job execution model
- Storage and computing elements
- Information services
- Site services
- Command line interface - the AliEn shell aliensh
- ROOT interface
- Grid and job monitoring
- Interfaces to other Grids

AliEn was primarily developed by ALICE, however it was adopted also by a couple of other Virtual Organizations like PANDA [60] and CBM [61].

5.1 File Catalogue (FC)

The File Catalogue is one of the key components of the AliEn suite. It provides a hierarchical structure (like a UNIX File system) and is designed to allow each directory node in the hierarchy to be supported by different database engines, running on different hosts. This building on top of several databases allows to add another database to expand the catalogue namespace and assures scalability of the system and allow growth of the catalogue as the files accumulate over the years.

Unlike real file systems, the FC does not own the files; it is a metadata catalogue on the Logical File Names (LFN) and only keeps an association/mapping between the LFNs and (possibly multiple) Physical File Names (PFN) of real files on a storage system. PFNs describe the physical location of the files and include the access protocol (rfio, xrootd), the name of the AliEn Storage Element and the path to the local file. The system supports file replication and caching.

The FC provides also a mapping between the LFNs and Globally Unique Identifiers (GUID). The labeling of each file with the GUID allows for the asynchronous caching. The write-once strategy combined with GUID labeling guarantees the identity of files with the same GUID label in different caches. It is possible to automatically construct PFNs : to store only the GUID and Storage Index and the Storage Element builds the PFN from the GUID. There are two independent catalogues: LFN->GUID and GUID->PFN. A schema of the AliEn FC is shown in Figure 15.

The FC can also associate metadata to the LFNs. This metadata is a collection of user-defined key value pairs. For instance, in the case of ALICE, the current metadata is the software version used to generate the files, number of events inside a file, or calibration files used during the reconstruction.

5.2 Job execution model

AliEn's Job execution model is based on the pull architecture. There is a set of central components (Task Queue, Job Optimizer, Job Broker) and another set of site components (Computing Element (CE), Cluster Monitor, MonALISA, Package Manager). The pull

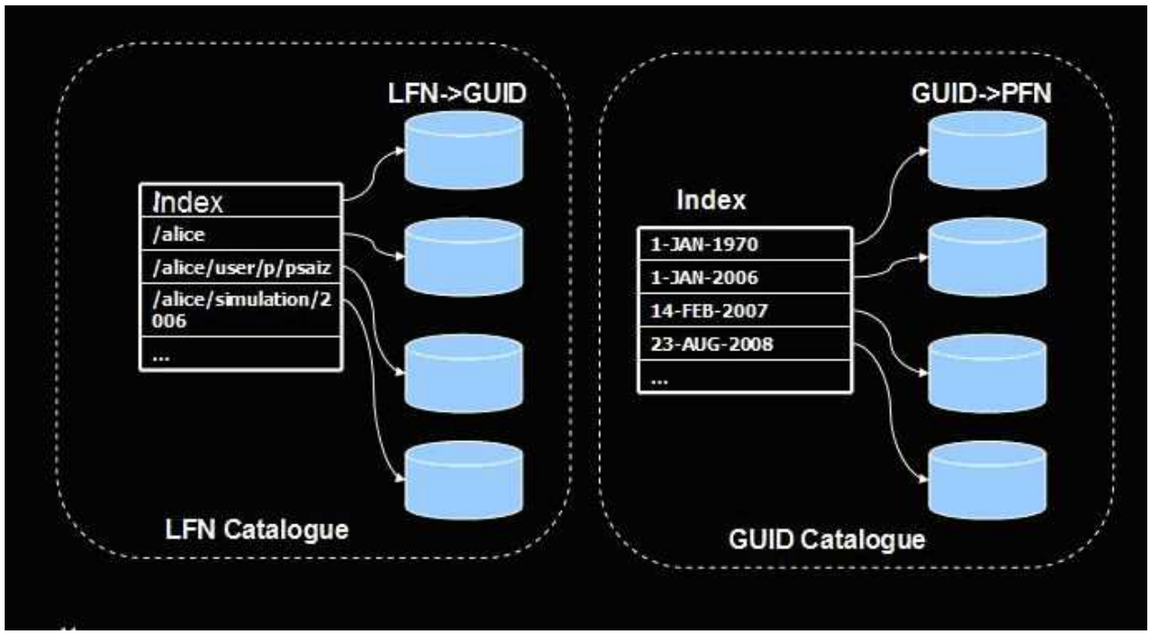


Fig. 15. AliEn File Catalogue

architecture has one major advantage with respect to the push one: the system does not have to know the actual status of all resources, which is crucial for large flexible Grids. In a push architecture, the distribution of jobs requires to keep and analyze a huge amount of status data just to assign a job, which becomes difficult in the expanding grid environment.

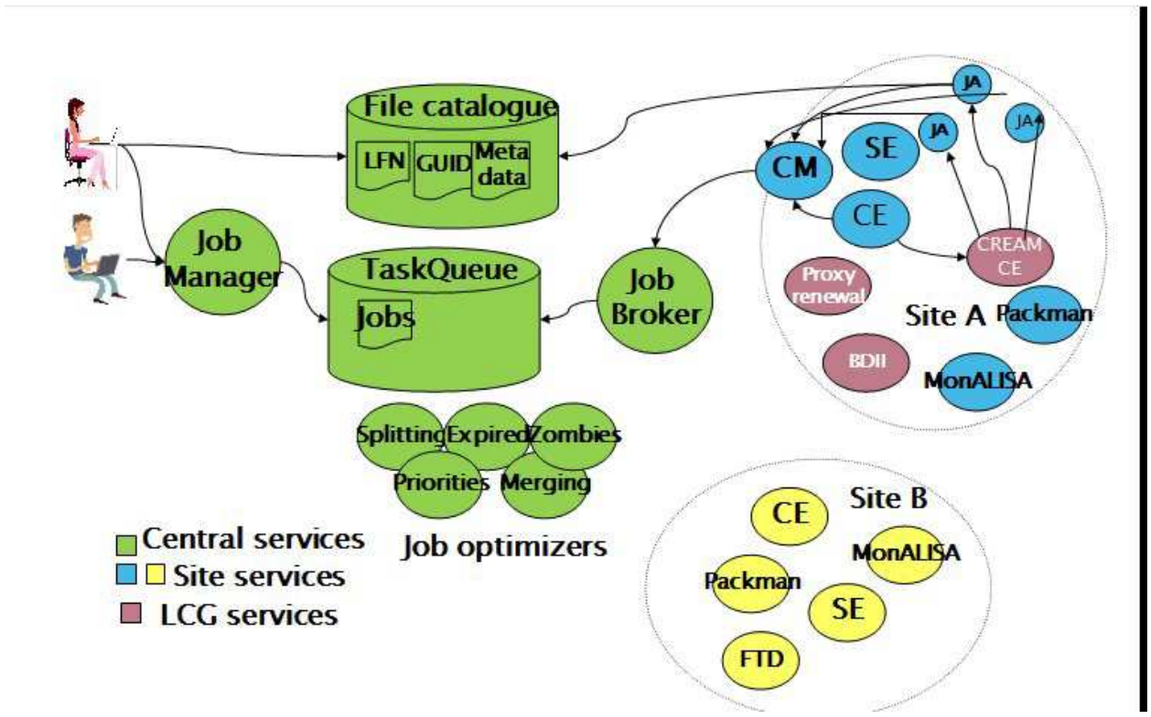


Fig. 16. AliEn + WLCG services

In the pull architecture, local agents (pilot status-checking test jobs) running at individual sites ask for real jobs after having checked the local conditions and found them appropriate for the processing of the job. Thus, AliEn only deals with the requests of local pilot jobs, so-called Job Agents (JA), to assign appropriate real jobs. The descriptions of jobs in the form of ClassAds are managed by the central Task Queue.

Each site runs several AliEn services: CE, ClusterMonitor, Package Manager (PackMan) and a MonALISA client. The AliEn CE automatically generates Job Agents and submits them to the local batch system. The ClusterMonitor manages the connections between the site and central services, so there is only one connection from each site. The AliEn CE can also be submit to the CREAM-CE, ARC, OSG or even the WMS, and delegate the communication with the local batch system to such a service. Schemas of the job submission procedure in AliEn are shown in Figures 16 and 17.

5.3 Jobs

When a job is submitted by a user, its description in the form of a ClassAd is kept in the central TQ where it waits for a suitable Job Agent for execution. There are several Job optimizers that can rearrange the priorities of the jobs based on the user quotas. These optimizers can also split jobs, or even suggest data transfers so it would be more likely that some Job Agent picks up the job. After it has been submitted, a job gets through several stages [62]. The information about running processes is kept also in the AliEn FC. Each job is given a unique id and a corresponding directory where it can register its output. The JAs provide a job-wrapper, a standard environment allowing a virtualization of resources. The whole job submission and

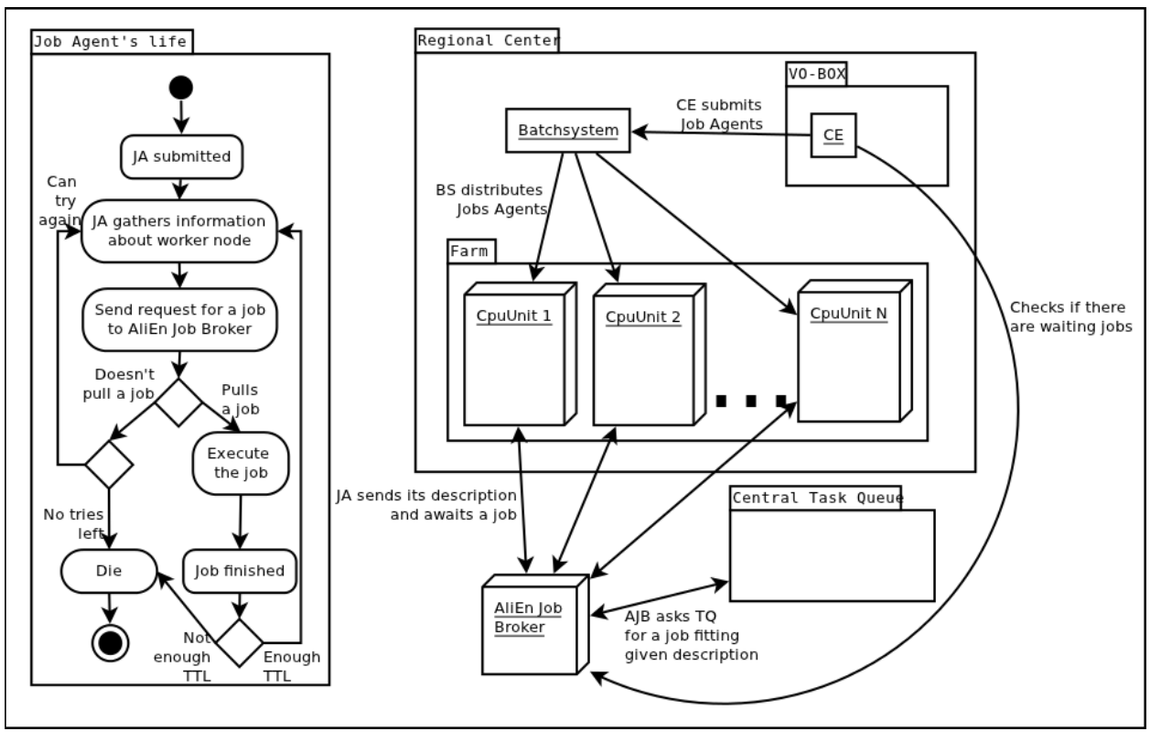


Fig. 17. The Job Agent model in AliEn: the JA does five attempts to pull a job before it dies.

processing chain is extensively monitored so a user can any time get the information on the status of his/her jobs.

5.4 Site services

As mentioned before, there are several AliEn services running at each ALICE site: CE, ClusterMonitor, PackMan and MonALISA. These services are running on a dedicated machine, so-called VOBOX described in section 3.

The AliEn site CE is usually associated with the local batch system. It is periodically submitting testing pilot jobs (Job Agents) to the local WLCG CE or an appropriate external Resource Broker or WMS. The role of the Job Agents is to verify the local hardware and software capacities at the site. After the usual matchmaking procedure, the JA is sent, through the site CE, into the local batch queue and then to a local Worker Node (WN). After its startup, the JA performs its task and in the case of a positive checkup, the JA requests a "real" job from the central Task Queue via the AliEn Job Broker, or dies otherwise. The PackMan automates the process of installation, upgrades, configuration and removal of the ALICE software packages from the shared software area on the site. It also advertises known/installed packages. The packages are installed on demand, when requested by a Job Agent running on a Worker Node or during the central software deployment over the Grid sites. If a package is not already installed the PackMan would install it along with its dependencies and return a string with commands that client has to execute to configure the package and all its dependencies. The PackMan manages the local disk cache and cleans it, when it needs more space to install newer packages. The Cluster Monitor handles communication with the AliEn Job Broker and gives configuration to JAs. It gets “heartbeats” from the JAs. If it gets no heartbeats from a JA, the existing job will get into the ZOMBIE status (after 1.5 hours) and then it will expire (after 3 hours).

5.5 Monitoring

Since the AliEn Workload Management does not depend directly on sophisticated monitoring, no special monitoring tools were developed in AliEn. As the monitoring solution, ALICE has adopted and further developed the Java-based MonALISA framework [44] mentioned already in the previous section. The MonALISA system is designed as an ensemble of autonomous multithreaded, self-describing agent-based subsystems which are registered as dynamic services, and together can collect and process large amounts of information.

The collected monitoring information is published via Web Service for use by AliEn Optimizers or for visualization purposes. An extension of the network simulation code which is a part of MonALISA can provide a tool for optimization and understanding of the performance of the AliEn Grid system.

5.6 Storage

Experience with the performance of different types of storage managers shows that the most advanced storage solution is the native XRootD manager [37] described in section 3. It has been demonstrated that with all other parameters being equal (protocol access speed and security) the native XRootD storage clusters exhibit substantially higher stability and availability. The ALICE distributed system of native XRootD clusters is orchestrated by the

global redirector which allows interacting with the complete storage pool as a unique storage. All storages are on WAN (Wide Area Network).

5.7 AliEn Shell - aliensh

To complete the brief description of AliEn, we mention the client called AliEn shell. It provides a UNIX-shell-like environment with an extensive set of commands which can be used to access AliEn Grid computing resources and the AliEn virtual file system. There are three categories of commands: informative and convenience commands, File Catalogue and Data Management commands and TaskQueue/Job Management commands. The AliEn shell has been created about 4 years ago and become a popular tool among the users for job handling and monitoring.

5.8 Concluding remarks

AliEn is a high-level middleware adopted by the ALICE experiment, which has been used and validated in massive Monte Carlo events production since 2001, in end-user analysis since 2005 and during the real data management and processing since 2007. Its capabilities comply with the requirements of the ALICE computing model. In addition to modules needed to build a fully functional Grid, AliEn provides interfaces to other Grid implementations enabling the true Grid interoperability. The AliEn development will be ongoing in the coming years following the architectural path chosen at the start and more modules and functionalities are envisaged to be delivered.

The Grid (AliEn/gLite/other) services are many and quite complex. Nonetheless, they are working together, allowing to manage thousands of CPUs and PBs of various storage types. The ALICE choice of single Grid Catalogue, single Task Queue with internal prioritization and a single storage access protocol (xrootd) has been beneficial from user and Grid management viewpoint.

6. WLCG and ALICE performance during the 2010/2011 LHC data taking

In this section, we will discuss the experience and performance of the WLCG in general and the ALICE Grid project in particular during the real LHC data taking both during the proton and the lead ion beam periods.

6.1 LHC performance

The LHC delivered the first pp collisions in the end of 2009, and the stable operations startup was in March 2010. Since then, the machine has been working amazingly well compared to other related facilities. Already in 2009, the machine beaten the world record in the beam energy and other records have followed. In 2010, the delivered integrated luminosity was 18.1 pb^{-1} and already during the first months of operation in 2011, the delivered luminosity was 265 pb^{-1}. This is about a quarter of the complete target luminosity for 2010 and 2011 [63], which is supposed to be sufficient to get the answer concerning the existence of the Higgs boson. Also, as mentioned in section 1, the machine has beaten the records concerning the stored energy and also the beam intensity.

The target number of bunches per a beam of protons stated for 2011 was reached already in the middle of the year: 1380 bunches. The final target luminosity of 10^{34} cm^{-2}s^{-1} is being approached rapidly, at present it is about 10^{33} cm^{-2}s^{-1} [64].

6.2 WLCG performance

The performance of the data handling by the WLCG has also been surprisingly good. This wrapped up several years to the LHC startup, when the WLCG and the experiments themselves were regularly performing a number of stress-tests of their distributed computing infrastructure and were gradually upgrading the systems using new technologies. As a result, when the data started to flow from the detectors, the performance of the distributed data handling machinery was quite astounding. All aspects of the Grid have really worked for the LHC and have enabled the delivery of Physics results incredibly quickly.

During 2010, 15 PetaBytes of data were written to tapes at the CERN Tier-0 reaching the level expected from the original estimates for the fully-tuned LHC operations, a nominal data taking year. As mentioned in section 2, in average 2 PB of data per a month were written to tapes at Tier-0 with the exception of the heavy ions period when this number got about doubled and a world record was reached with 225 TB written to tapes within one day, see Figure 18. The CERN Tier-0 moved altogether above 1 PB of data per day.

As mentioned in section 2, the mass storage system at the CERN Tier-0 supported data rates at an average over the year of 2.5 GB/s IN with peaks up to 11 GB/s, and data was served at an average rate of $\sim$ 7 GB/s with peaks up to 25 GB/s.

The data processing went on without basic show-stoppers. The workload management system was able to get about 1 million of jobs running per day, see Figure 19, and this load is gradually going up. This translates into significant amounts of computer time. Towards the end of 2010 there was a situation when all of the available job slots at Tier-1s and Tier-2s were often fully occupied. This has been showing up also during 2011, so the WLCG collaboration has already now fully used all the available computing resources. During 2010, the WLCG delivered about 100 CPU-millennia.

As also briefly mentioned in section 2, the WLCG was very successful concerning the number of individuals really using the grid to perform their analysis. At the start of the project, there was a concern that end users will be discouraged from using the grid due to complexity of its structure and services. But thanks to the effort of the WLCG and experiments themselves a reasonably simple access interfaces were developed and the number of end users reached up to 800 in the large experiments.

The distribution of the delivered CPU power between sites has been basically according to the original design, but the Tier-2s provided more than the expected 40%: it was in fact 50% or more of the overall delivery, see section 2. Member countries pledged different amounts of resources according to their capacities and have been delivering accordingly. So the concept of collaborative Grid resource sharing really works and enables institutes worldwide to share data, and provide resources to the common goals of the collaboration.

6.2.1 Network performance

The key basis for building up a distributed system is the data transfer infrastructure. The network which the WLCG operates today is much in advance of what was anticipated in

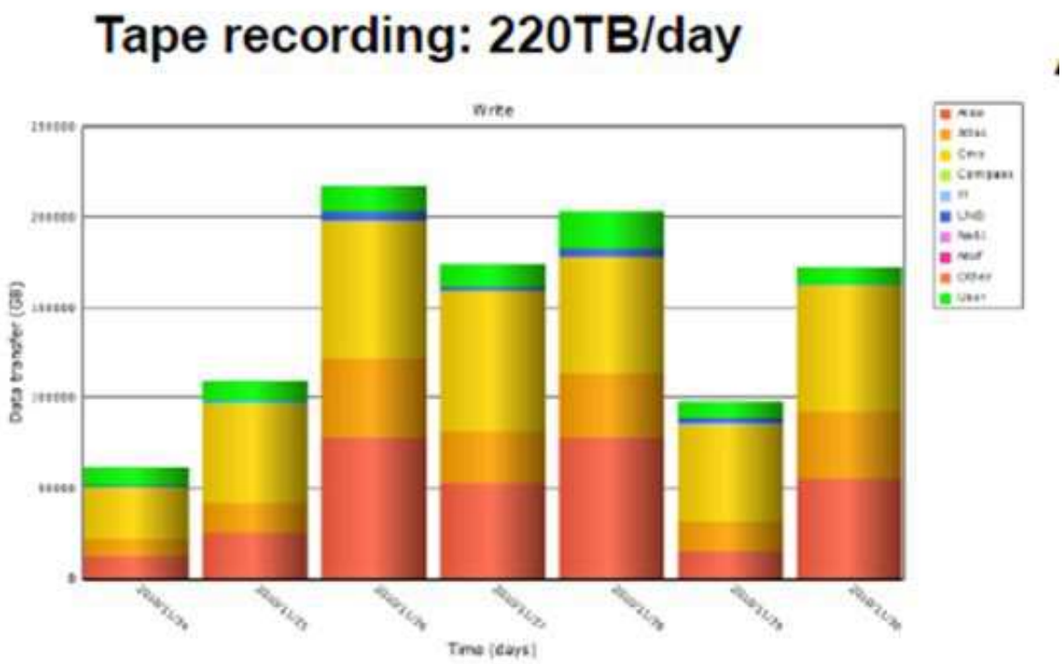


Fig. 18. A record in data tape recording: over 220 TB/day.

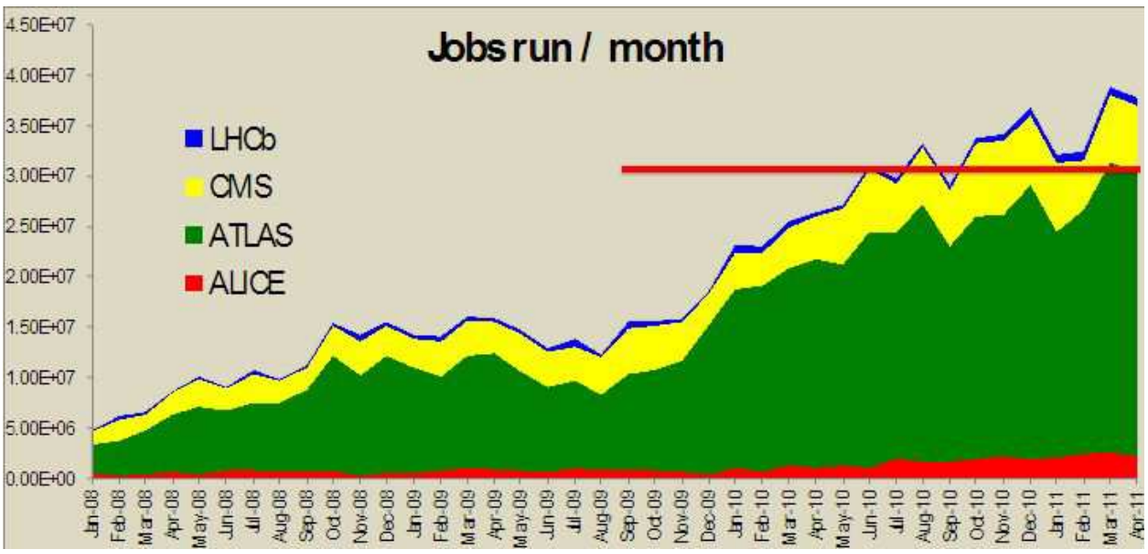


Fig. 19. WLCG job profile

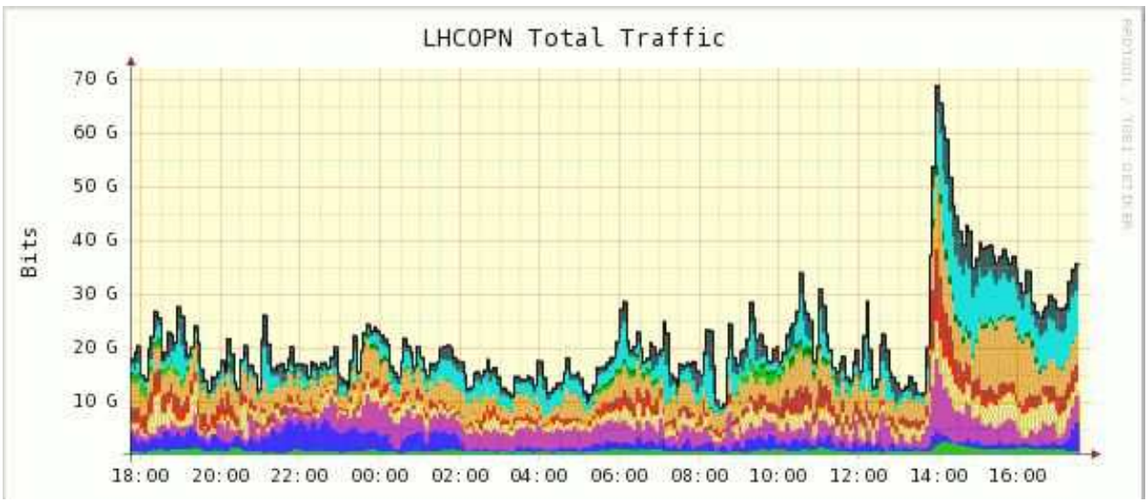


Fig. 20. WLCG OPN traffic in 2010 with a peak of 70 Gbit/s

the time of writing the WLCG TDR. The OPN, see also section 2, started as dedicated fiber links from CERN to each of the Tier-1s with the throughput 10 Gbit/s. Today, there is a full redundancy in this network with the original links doubled and with back-up links between Tier-1s themselves. The OPN is a complicated system with many different layers of hardware and software and getting it into the current shape was a difficult task, which evidently paid-off.

The original concerns about the possible network unreliability and insufficiency were not realized. The network infrastructure relying on the OPN and the complementary GEANT, US-LHCNet and all the R&E national network infrastructures, extensively monitored and continuously checked with the test transfer jobs, has never been a problem in the data transfer except for occasional glitches. The originally estimated sustained transfer rate of 1.3 GB/s from Tier-0 to Tier-1s was reached without problems and exceeded and reached up to 5 GB/s. Within the OPN, a peak of 70 Gb/s was supported without any problem during a re-processing campaign of one of the LHC experiments, see Figure 20.

6.2.2 Concluding remarks - WLCG

The experience from the first year and a half of the LHC data taking implies that the WLCG has built a truly working grid infrastructure. The LHC experiments have their own distributed models and have used the WLCG infrastructure to deliver Physics results within weeks after the data recording which has never been achieved before. The fact that a significant numbers of people are doing analysis on the Grid, that all the resources are being used up to the limits and the scientific papers are produced with an unprecedented speed is proving an evident success of the WLCG mission.

6.3 ALICE performance

To conclude this section, we will briefly summarize the experience and performance of the ALICE experiment. ALICE started extremely successfully the processing of the LHC data in 2009: the data collected during the first collisions delivered by the LHC on November 23rd

(2009) got processed and analyzed so fast that within one week the article with the results from the first collisions was accepted for publication as the first ever scientific paper with the Physics results from the LHC collisions [65].

6.3.1 Jobs

During the data taking in 2010, ALICE collected 2.3 PB of raw data, which represented about 1.2 million of files with the average file size of 1.9 GB. The data processing chain has been performing without basic problems. The Monte Carlo simulation jobs together with the raw data reconstruction and organized analysis (altogether the organized production) represented almost 7 millions of successfully completed jobs, which translates into 0.3 jobs/second. The chaotic (end user) analysis made for 9 millions of successfully completed jobs, which represents 0.4 jobs/s, consuming approximately 10% of the total ALICE CPU resources (the chaotic analysis jobs are in general shorter than the organized processing jobs). In total, there were almost 16 millions of successfully done jobs, which translates to 1 job/s and 90 thousands jobs/day. The complimentary number of jobs which started running on the Grid but finished with an error was in excess of this.

The running jobs profile got in peaks to 30 thousands of concurrently running jobs (see Figure 21) with more than 50% of the CPU resources delivered by the Tier-2 centers. About 60% of the total number of jobs represented the end user analysis (see Figure 22). In general, the user analysis already in 2010 was a resounding success, with almost 380 people actively using the Grid. Since the chaotic analysis brings sometimes problems concerning not completely perfect code resulting, e.g., in a high memory consumption (cf. section 4), ALICE was running a mixture of the organized production and end user jobs at all its sites, and this scenario was working well.

6.3.2 Storage-2010

The distributed storage system endured and was supporting an enormous load. During 2010/2011, 25.15 PB of data (raw, ESDs, AODs, Monte Carlo productions) was written to xrootd Storage Elements with the speed maximum of 621.1 MB/s. 59.97 PB of data was read from the xrootd Storage Elements, with the speed maximum of 1.285 GB/s, see Figure 23.

6.3.3 Data taking 2011

Constantly upgrading and extending its hardware resources and updating the grid software, ALICE continued a successful LHC data handling campaign in 2011. By September, the total volume of the collected raw data was almost 1.7 PB with the first reconstruction pass completed. The year 2011 was marked by massive user analysis on the Grid. In May, the most important conference in the Heavy Ion Physics community, Quark Matter 2011 (QM2011) [66], took place and was preceded by an enormous end user analysis campaign. In average, 6 thousands end-user jobs were running all the time, which represents almost 30% of the CPU resources officially dedicated to ALICE (The number of running jobs is higher than that most of the time due to use of opportunistic resources). During the week before the QM2011, there was a peak with 20 thousands of concurrently running end-user jobs, see Figures 23,24. The number of active Grid users reached 411.

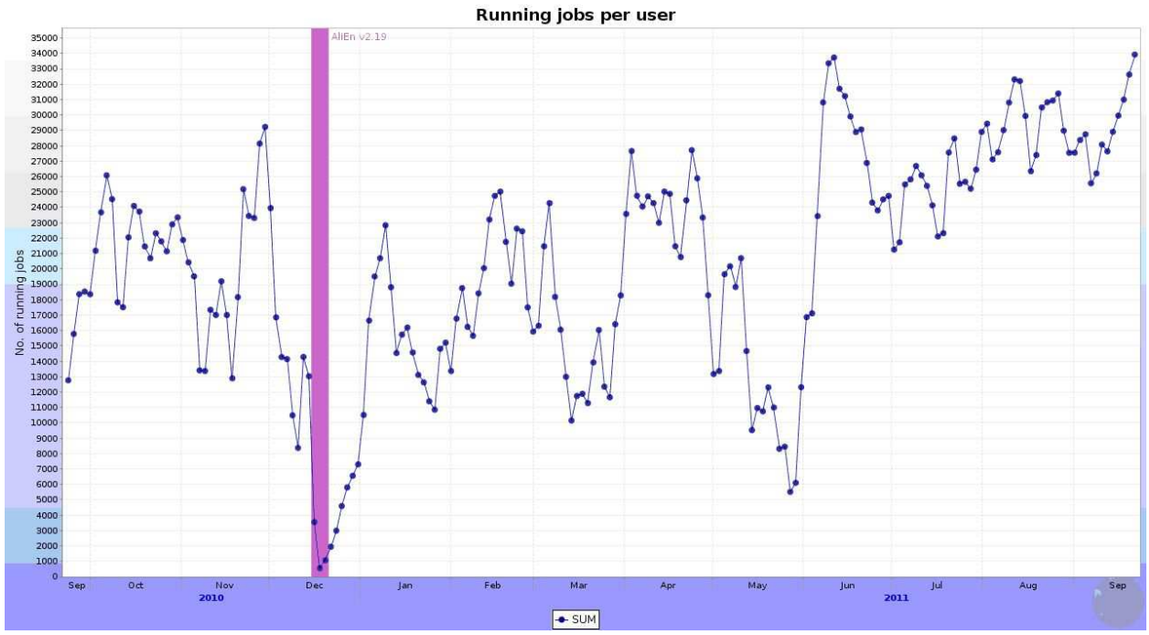


Fig. 21. ALICE running jobs profile 2010/2011.

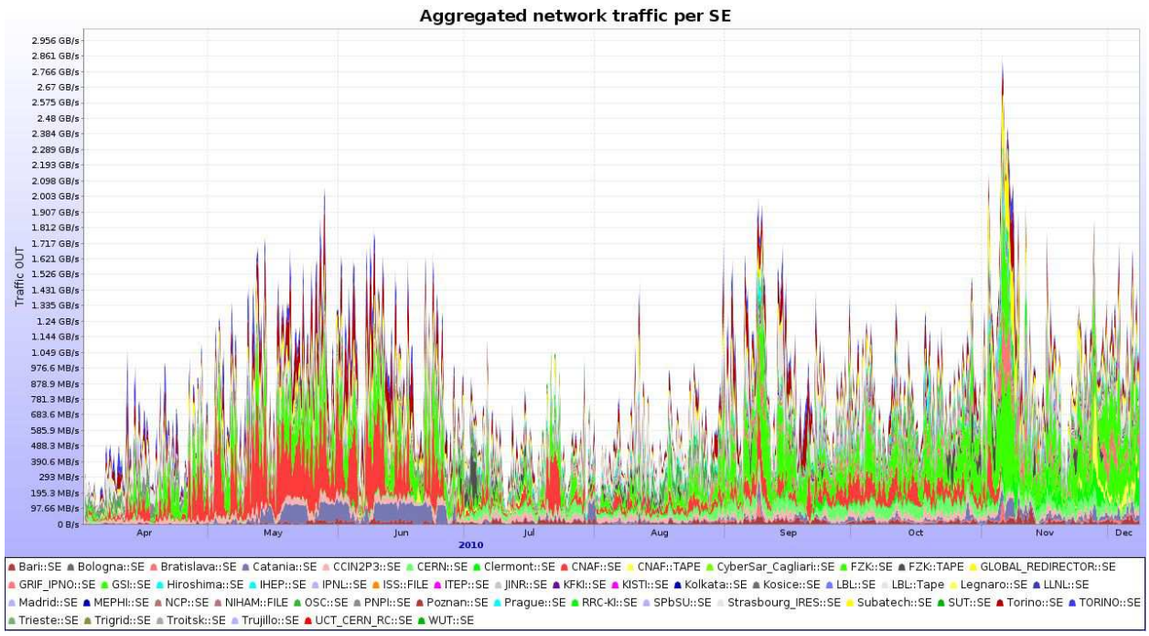


Fig. 22. Network traffic OUT by analysis jobs - 2010

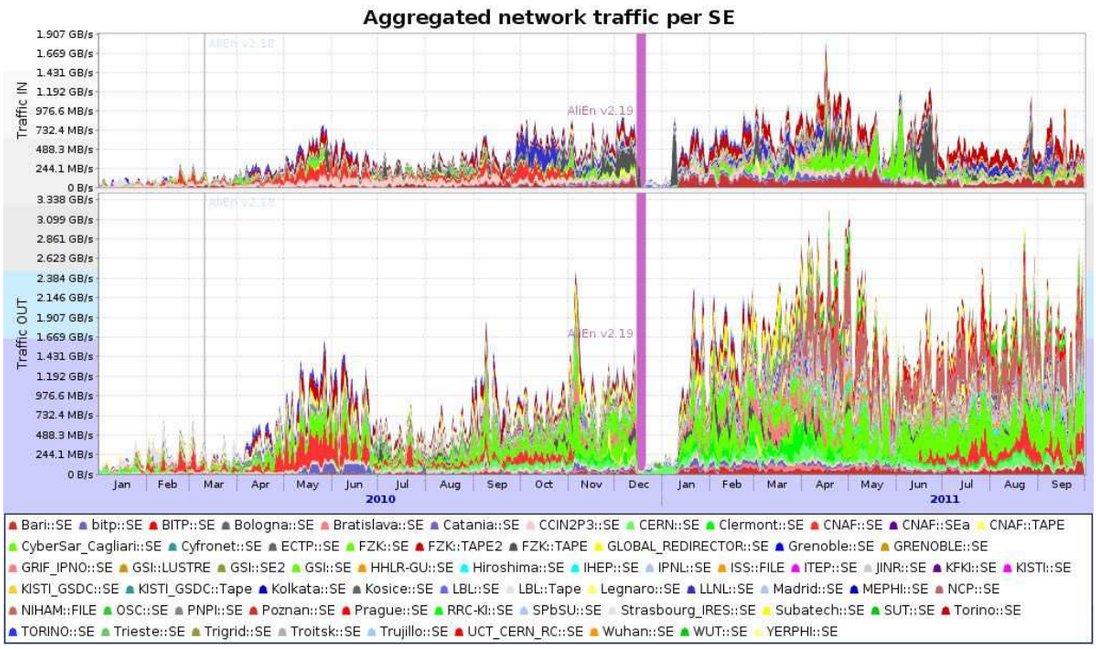


Fig. 23. Total network traffic at the ALICE Storage Elements - 2010/2011

Fig. 24. End-user jobs profile - 2011

In total, the ALICE sites were running in average 21 thousands of jobs with peaks up to 35 thousands (Figure 21). The resources ratio remained 50% delivered by Tier-0 and Tier-1s to 50% delivered by Tier-2s. Altogether, 69 sites were active in the operations. The sites' availability and operability kept very stable throughout the year. The gLite (now EMI) middleware, see section 3, is mature and only a few changes are necessary.

In the beginning of the 2011 campaign, there was a concern that the storage would be saturated. In fact the storage infrastructure was performing basically without problems supporting the enormous load from the end-user analysis and was getting ready for the Pb-Pb operations. The network situation, as was already mentioned for the WLCG in general, has been excellent and allowed for the operation scenario where the hierarchical tiered structure got blurred, the sites of all levels were well interconnected and running a similar mixture of jobs. As a result, the ALICE Grid in a sense has been working as a cloud.

6.4 Concluding remarks - ALICE

In general, similar to the overall characteristics of the WLCG performance also the ALICE operations and data handling campaigns were notably successful right from the beginning of the LHC startup, making the Grid infrastructure operational and supporting a fast delivery of Physics results. By September 2011, ALICE has published 15 scientific papers with the results from the LHC collisions and more is on the way. Two papers [67,68] were marked as "Suggested reading" by the Physical Review Letters editors and the later was also selected for the "Viewpoint in Physics" by Physical Review Letters.

The full list of the ALICE papers published during 2009-2011 can be found on [69]. One of the Pb-Pb collision events recorded by ALICE is shown on Figure 25.

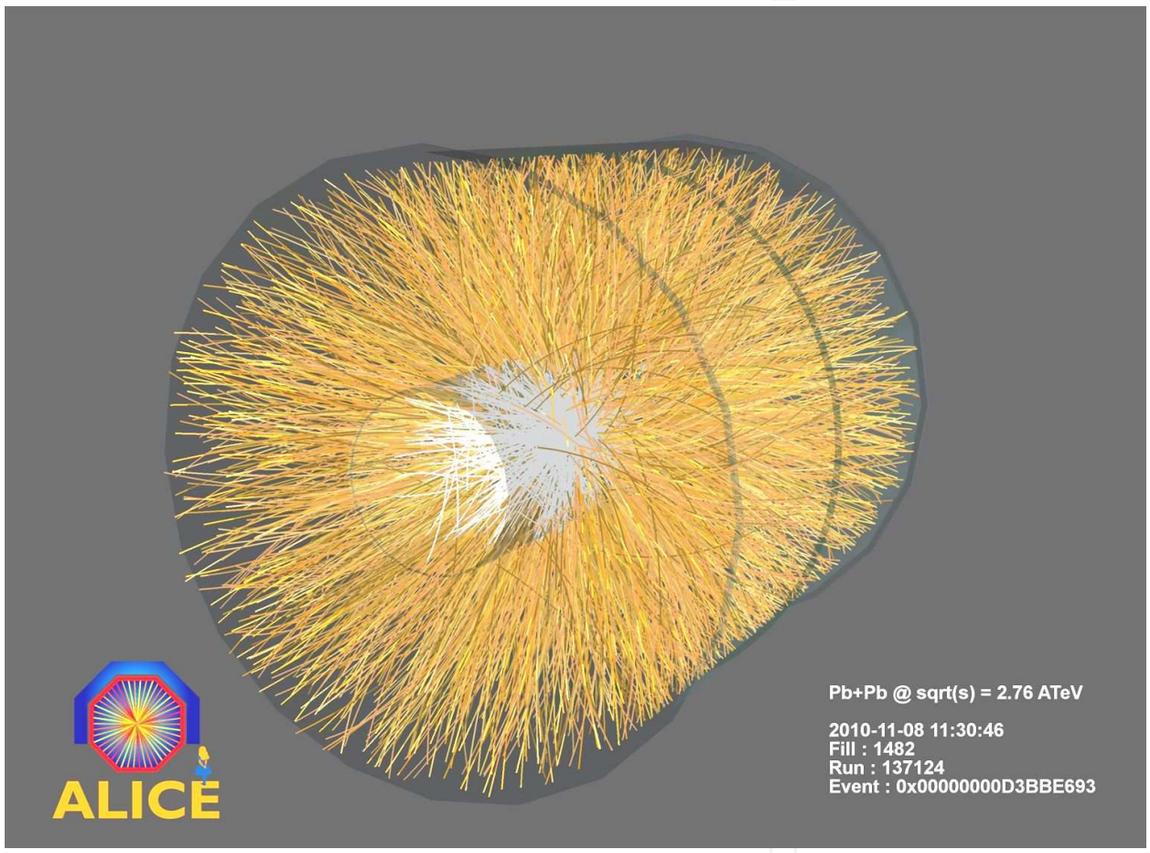


Fig. 25. Pb-Pb collision event recorded by ALICE

7. Summary and outlook

This chapter is meant to be a short overview of the facts concerning the Grid computing for HEP experiments, in particular for the experiments at the CERN LHC. The experience gained during the LHC operations in 2009-2011 has proven that for this community, the existence of a well performing distributed computing is necessary for the achievement and fast delivery of scientific results. The existing WLCG infrastructure turned up to be able to support the data production and processing thus fulfilling its first-plan mission. It has been and will be

continuously developing into the future absorbing and giving rise to new technologies, like the advances in networking, storage systems and inter-operability between Grids and Clouds [70,71].

7.1 Data management

Managing the real data taking and processing in 2009-2011 provided basic experience and a starting point for new developments. The excellent performance of the network which was by far not anticipated in the time of writing the WLCG (C)TDR shifted the original concept of computing models based on hierarchical architecture to a more symmetrical mesh-like scenario. In the original design, the jobs are sent to sites holding the required data sets and there are multiple copies of data spread over the system due to anticipation that network will be unreliable or insufficient. It turned out that some data sets were placed on sites and never touched.

Based on the existing excellent network reliability and growing throughput, the data models start to change along a dynamical scenario. This includes sending data to a site just before a job requires it, or reading files remotely over the network, use remote (WAN) I/O to the running processes. Certainly, fetching over the network one needed data file from a given data set which can contain hundreds of files is more effective than a massive data sets deployment and will spare storage resources and bring less network load.

The evolution of the data management strategies is ongoing. It goes towards caching of data rather than strict planned placement. As mentioned, the preferences go to fetching a file over the network when a job needs it and to a kind of intelligent data pre-placement. The remote access to data (either by caching on demand and/or by remote file access) should be implemented.

7.2 Network

To improve the performance of the WLCG-operated network infrastructure, the topology of LHC Open Network Environment (LHCONE [24]) is being developed and built. This should be complementary to the existing OPN infrastructure providing the inter-connectivity between Tier-2s and Tier-1s and between Tier-2s themselves without putting an additional load on the existing NREN infrastructures. As we learned during the last years, the network is extremely important and better connected countries do better.

7.3 Resources

During the 2010 data taking the available resources were sufficient to cover the needs of experiments, but during 2011 the computing slots as well as the storage capacities at sites started to be full. Since the experience clearly shows that delivery of the Physics results is limited by resources, the experiments are facing a necessity of more efficient usage of existing resources. There are task forces studying the possibility of using the next generations computing and storage technologies. There is for instance a question of using multicore processors which might go into the high performance computing market while WLCG prefers usage of commodity hardware.

7.4 Operations

Another important issue is sustainability and support availability for the WLCG operations. The middleware used today for the WLCG operations is considerably complex with many services unnecessarily replicated in many places (like, e.g., databases) mainly due to original worries concerning network. The new conception is to gradually search for more standard solutions instead of often highly specialized middleware packages maintained and developed by WLCG.

7.5 Clouds and virtualization

Among the new technologies, the Clouds is the right buzzword now and the virtualization of resources comes along. The virtualization of WLCG sites started prior to the first LHC collisions and has gone quite far. It helps improving system management, provision of services on demand, can make use of resources more effective and efficient. Virtualization also enables to make use of industrial and commercial solutions.

But, no matter what the current technologies advertise, the LHC community will always use a Grid because the scientists need to collaborate and share resources. No matter what technologies are used underneath the Grid, the collaborative sharing of resources and the network of trust and all the security infrastructure developed on the way of building the WLCG is of enormous value, not only to WLCG community but to e-science in general. It allows people to collaborate across the infrastructures.

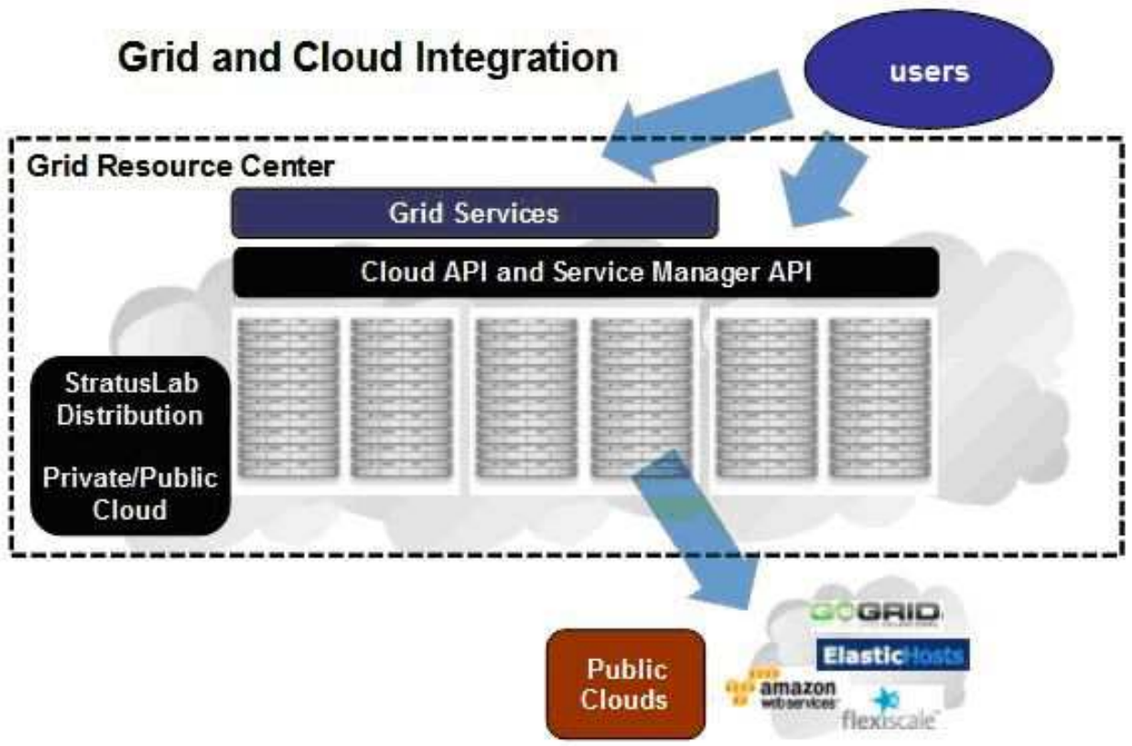


Fig. 26. Schema of StratusLab IaaS Cloud interoperability with a Grid

The basic operations like distributed data management, the high data throughput and the remote job submission can probably be more cloud-like. There is a great interest among people to use commercial Clouds resources, especially when the experiments see their resources becoming full.

So, can we use Amazon or Google to do processing of data from LHC? The point is, one cannot be sure what the level of services will be and what the IN/OUT bandwidth will be. This can in principle be negotiated with these companies and may bring some level of agreement. That in principle is doable.

7.6 Grids and Clouds

As argued in [70], "Cloud Computing not only overlaps with Grid Computing, it is indeed evolved out of Grid Computing and relies on Grid Computing as its backbone and infrastructure support. The evolution has been a result of a shift in focus from an infrastructure that delivers storage and compute resources (in the case of Grids) to one that is economy based ....". Both the Grids and the Clouds communities are facing the same problems like the need to operate large facilities and to develop methods by which users/consumers discover, request and use resources provided by the centralized facilities.

There exist a number of projects looking into and developing Cloud to Grid interfaces with the idea that Grid and Cloud Computing serve different use cases and can work together improving Distributed Computing Infrastructures (see, e.g., [71]). Also CERN is involved in this activity together with other international laboratories in Europe.

With the WLCG resources becoming used up to their limits, using commercial Clouds to process the LHC data is a strategy that should be assessed. Several of the LHC experiments have done tests whether they can use commercial Clouds. But today, the cost is rather high. Also, there are issues like whether academic data can be shipped through academic networks to a commercial provider or how to make sure what happens to this data.

Nevertheless the strategy towards deployment over the WLCG resources Cloud interfaces, managed with high level of virtualization, is under evaluation. Some level of collaboration with industry would provide the understanding how to deploy this properly and what would be the cost. The Cloud and Grid interfaces can be deployed in parallel or on top of each other. This development might also give a way to evolve into a more standardized infrastructure and allow to make a transparent use of commercial Clouds.

A testbed of such an architecture is the CERN LXCloud [72] pilot cluster. Implementation at CERN allows to present a Cloud interface or to access other public or commercial Clouds. This is happening with no change to any of the existing Grid services. Another interesting example is the development of a comprehensive OpenSource IaaS (Infrastructure as a Service) Cloud distribution within the StratusLab project [71], see Figure 26. Anyone can take the code and deploy it on his site and have IaaS Cloud running on his site. The project is focused on deploying Grid services on top of this Cloud, 1) to be a service to existing European Grid infrastructures and to enable these people to use Cloud-operated resources and 2) because the developers consider the Grid services very complex and making sure they run safely on this Cloud should guarantee that also other applications will run without problems.

7.7 Physics results achieved by the LHC experiments

Before we bring the final concluding remarks for our chapter, we will briefly summarize the Physics results delivered by the LHC experiments by the time of writing this document.

In addition to many specific new results describing different Physics phenomena in the energy regime never explored before, there have been new findings concerning some of the major issues addressed by the LHC research.

- ATLAS and CMS experiments have been delivering results concerning the energy regions excluding the mass of the Higgs boson. The latest results on the topic of Higgs boson searches exclude a wide region of Higgs boson masses: ATLAS excludes Higgs boson masses above 145 GeV, and out to 466 GeV (apart from a couple of points in-between, which are however excluded by CMS studies). For some of the latest references see [73-75].
- To contribute new results on the topic of the dominance of matter over antimatter in the present Universe, the LHCb experiment has been pursuing studies of phenomena demonstrating the so-called CP-symmetry violation. Violation of this symmetry plays an important role in the attempts of Cosmology to explain the dominance of matter over antimatter in our world. The latest LHCb results concerning the demonstration of the existence of the CP-violation can be found, e.g., in [76].
- The study of properties of the Quark Gluon Plasma (QGP), the phase of matter which existed in a fraction of a second after the Big Bang, is the mission of the ALICE experiment. During the lead-lead collisions at the LHC energies, the individual collisions can be seen as "little Big Bangs". The matter produced in these collisions is under extreme conditions: the energy density corresponds to a situation when 15 protons are squeezed into the volume of one proton and the temperature reaches more than 200000 times the temperature in the core of the Sun. ALICE has confirmed the previous findings of the STAR experiment at the Relativistic Heavy Ion Collider (RHIC) at Brookhaven that this QGP behaves like an ideal liquid [68] even at the LHC energies.

7.8 Concluding remarks

As we already stressed, the WLCG performance during the LHC data taking in 2009-2011 was excellent and the basic mission of the WLCG has been fulfilled: the data taking and processing is ongoing without major show-stoppers, hundreds of people are using the Grid to perform their analysis and unique scientific results are delivered within weeks after the data was recorded. In addition, the experience gained during this data taking .stress test. launched new strategies to be followed on the way of the future WLCG development. There are fundamental issues like the approaching lack of WLCG resources and the expansion of new technologies like the Cloud computing. In the time of writing this chapter it looks like we will see in the future some combination of Grid and Cloud technologies will be adopted to operate the distributed computing infrastructures used by the HEP experiments.

8. Acknowledgements

We would like to thank Jiri Adam for a critical reading of the manuscript and for a great help with the LaTeX matters. The work was supported by the MSMT CR contracts No. 1P04LA211 and LC 07048.

9. References

[1] D.H. Perkins: Introduction to High Energy Physics, Cambridge University Press, 4th edition (2000), ISBN-13: 978-0521621960.
Procedings of 35th International Conference of High Energy Physics, July 22-28, 2010, Paris, France, Proceedings of Science (PoS) electronic Journal: ICHEP 2010

[2] STAR Collaboration: Experimental and theoretical challenges in the search for the quark gluon plasma: The STAR Collaboration's critical assessment of the evidence from RHIC collisions, Nucl. Phys. A757 (2005) 102-183.

[3] CERN - the European Organization for Nuclear Research; http://public.web.cern.ch/public/

[4] The Large Hadron Collider at CERN; http://lhc.web.cern.ch/lhc/; http://public.web.cern.ch/public/en/LHC/LHC-en.html

[5] ALICE Collaboration: http://aliceinfo.cern.ch/Public/Welcome.html

[6] ATLAS Collaboration: http://atlas.ch/

[7] CMS Collaboration: http://cms.web.cern.ch/

[8] LHCb Collaboration: http://lhcb-public.web.cern.ch/lhcb-public/

[9] TOTEM Experiment: http://cern.ch/totem-experiment

[10] LHCf Experiment: http://cdsweb.cern.ch/record/887108/files/lhcc-2005-032.pdf

[11] W.N. Cottingham and D.A. Greenwood: An Introduction to the Standard Model of Particle Physics, Cambridge University Press, 2nd edition (2007), ISBN-13: 978-0521852494

[12] Worldwide LHC Computing Grid: http://public.web.cern.ch/public/en/lhc/Computing-en.html

[13] LHC Computing Grid: Technical Design Report, http://lcg.web.cern.ch/LCG/tdr/

[14] I. Foster and C. Kesselman. The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, 1999;
I. Foster et al: The Anatomy of the Grid: Enabling Scalable Virtual Organizations, International Journal of High Performance Computing Applications Vol.15(2001), p.200.

[15] WLCG Memorandum of Understanding, http://lcg.web.cern.ch/lcg/mou.htm

[16] EGI - The European Grid Initiative; http://web.eu-egi.eu/

[17] OSG - The Open Science Grid, http://www.opensciencegrid.org/; https://osg-ress-1.fnal.gov:8443/ReSS/ReSS-prd-History.html

[18] I. Legrand et al: MONARC Simulation Framework, ACAT'04, Tsukuba, Japan,2004; http://monarc.cacr.caltech.edu:8081/www_monarc/monarc.htm

[19] CERN Advanced Storage Manager: http://castor.web.cern.ch/castor/

[20] I. Bird: LHC Computing: After the first year with data, TERENA Networking Conference (TNC2011), Prague, 2011, https://tnc2011.terena.org/web/media/archive/7A

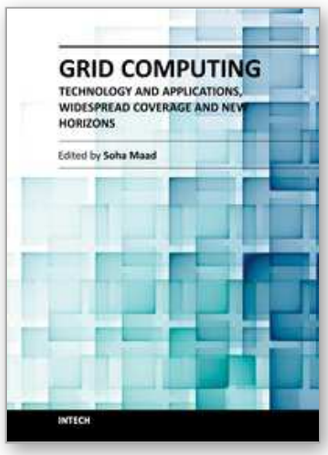
[21] LHCOPN - The Large Hadron Collider Optical Private Network, https://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome

[22] The GEANT Project, http://archive.geant.net/

[23] USLHCNet: High speed TransAtlantic network for the LHC community, http://lhcnet.caltech.edu/
[24] LHCONE-LHC Open Network Environment: http://lhcone.net/
[25] Virtual Organisation: http://technical.eu-egee.org/index.php?id=147
[26] The European Middleware Initiative: http://www.eu-emi.eu/home
[27] The Globus Toolkit: http://www-unix.globus.org/toolkit/
[28] OMII - The Open Middleware Infrastructure Institute: http://www.omii.ac.uk/
[29] The Virtual Data Toolkit, http://vdt.cs.wisc.edu/
[30] P. Saiz et al., AliEn-ALICE environment on the GRID, Nucl. Instrum. Meth. A502 (2003) 437; http://alien2.cern.ch/
[31] Computing Element: http://glite.cern.ch/lcg-CE/
[32] Workload Management System: http://glite.cern.ch/glite-WMS/
[33] The CREAM (Computing Resource Execution And Managemen) Service, http://glite.cern.ch/glite-CREAM/
[34] Storage Element: http://glite.cern.ch/glite-SE_dpm_mysql/
[35] dCache: http://www.dcache.org/
[36] The Disk Pool Manager: https://www.gridpp.ac.uk/wiki/Disk_Pool_Manager; https://twiki.cern.ch/twiki/bin/view/LCG/DataManagementTop
[37] XRootD: http://project-arda-dev.web.cern.ch/project-arda-dev/xrootd/site/index.html
[38] SLAC (Stanford Linear Accelerator Center): http://slac.stanford.edu/
[39] INFN - The National Institute of Nuclear Physics: http://www.infn.it/indexen.php
[40] R. Brun and F. Rademakers, ROOT: An object oriented data analysis framework, Nucl. Instrum. Meth. A389 (1997) 81; http://root.cern.ch
[41] VOMS-Virtual Organization Membership Service: http://glite.web.cern.ch/glite/packages/R3.1/deployment/glite-VOMS_mysql/glite-VOMS_mysql.asp
[42] The VO-box: http://glite.cern.ch/glite-VOBOX/
[43] ALICE Experiment Computing TDR: http://aliceinfo.cern.ch/Collaboration/Documents/TDR/Computing.html
[44] Monitoring Agents using a Large Integrated Services Architecture: http://monalisa.cern.ch/monalisa.html; C. Grigoras et al., Automated agents for management and control of the ALICE Computing Grid, Proceedings of the 17th Int. Conf. CHEP 2009, Prague, March 21-27, 2009, J. Phys.: Conf. Ser. 219, 062050.
[45] ALICE raw data production cycles: http://alimonitor.cern.ch/production/raw.jsp
[46] AliRoot: http://aliceinfo.cern.ch/Offline/AliRoot/Manual.html
[47] GEANT3-Detector Description and Simulation Tool: http://wwwasd.web.cern.ch/wwwasd/geant/
[48] FLUKA-Particle Physics MonteCarlo Simulation package: http://www.fluka.org/fluka.php

[49] Pythia6: http://projects.hepforge.org/pythia6/;
Pythia8: http://home.thep.lu.se/~torbjorn/pythiaaux/present.html
[50] Xin-Nian Wang and Miklos Gyulassy: HIJING: A Monte Carlo model for multiple jet production in pp, pA and AA collisions,
Phys. Rev. D44 (1991), 3501;
http://www-nsdth.lbl.gov/~xnwang/hijing/
[51] ALICE Offline policy:
http://aliceinfo.cern.ch/Offline/General-Information/Offline-Policy.html
[52] Monitoring of Analysis trains in ALICE:
http://alimonitor.cern.ch/prod/
[53] ALICE simulation framework:
http://aliceinfo.cern.ch/Offline/Activities/Simulation/index.html
[54] ALICE MC simulation cycles:
http://alimonitor.cern.ch/job_details.jsp
[55] ALICE Computing sites:
http://pcalimonitor.cern.ch:8889/reports/;
ALICE Distributed Storage:
http://pcalimonitor.cern.ch/stats?page=SE/table
[56] Yves Schutz: Computing resources 2011-2013, ALICE Computing Board Sept. 1st, 2011:
http://indico.cern.ch/materialDisplay.py?contribId=3&materialId=2&confId=153622;
https://twiki.cern.ch/twiki/bin/view/FIOgroup/TsiBenchHEPSPEC
[57] gLite-Lightweight Middleware for Grid Computing,
http://glite.cern.ch/
[58] ARC-The Advanced Resource Connector middleware,
http://www.nordugrid.org/arc/about-arc.html
[59] The AliEn Shell-aliensh, AliEn User Interfaces:
http://project-arda-dev.web.cern.ch/project-arda-dev/alice/apiservice/guide/guide-1.0.html#_Toc156731986;
ALICE Grid Analysis:
http://project-arda-dev.web.cern.ch/project-arda-dev/alice/apiservice/AA-UserGuide-0.0m.pdf
[60] The PANDA Experiment: http://www-panda.gsi.de/
[61] The CBM Experiment: http://www-cbm.gsi.de/
[62] Job statuses in AliEn:
http://pcalimonitor.cern.ch/show?page=jobStatus.html
[63] LHC Design Report: http://lhc.web.cern.ch/lhc/LHC-DesignReport.html
[64] LHC Performance and Statistics:
https://lhc-statistics.web.cern.ch/LHC-Statistics/
[65] The ALICE Collaboration: First proton-proton collisions at the LHC as observed with the ALICE detector: measurement of the charged-particle pseudorapidity density at $\sqrt{s} = 900$ GeV, Eur. Phys. J. C65 (2010), 111-125.
[66] Quark Matter 2011: http://qm2011.in2p3.fr/
[67] The ALICE Collaboration: Charged-Particle Multiplicity Density at Midrapidity in Central Pb-Pb Collisions at $\sqrt{s_{NN}} = 2.76$ TeV, Phys. Rev. Lett. 105 (2010), 252301.

[68] The ALICE Collaboration: Elliptic Flow of Charged Particles in Pb-Pb Collisions at $\sqrt{s_{NN}} = 2.76$ TeV, Phys. Rev. Lett. 105 (2010), 252302.
[69] Physics Publications of the ALICE Collaboration in Refereed Journals, http://aliceinfo.cern.ch/Documents/generalpublications
[70] I. Foster et al: Cloud Computing and Grid Computing 360-Degree Compared, Proc. of the Grid Computing Environments Workshop, 2008. GCE '08, Austin, Texas, http://arxiv.org/ftp/arxiv/papers/0901/0901.0131.pdf
[71] C. Loomis: StratusLab: Enhancing Grid Infrastructures with Cloud and Virtualization Technologies, TERENA Networking Conference (TNC2011),Prague, 2011, https://tnc2011.terena.org/web/media/archive/11C
[72] LXCloud: https://twiki.cern.ch/twiki/bin/view/FIOgroup/LxCloud
[73] The ATLAS Collaboration: Search for the Higgs boson in the $H \to WW \to lvjj$ decay channel in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector; arXiv:1109.3615v1, Sep. 2011.
[74] The ATLAS Collaboration: Search for a Standard Model Higgs boson in the $H \to ZZ \to llvv$ decay channel with the ATLAS detector; arXiv:1109.3357v1, Sep. 2011.
[75] The CMS Collaboration: New CMS Higgs Search Results for the Lepton Photon 2011 Conference; http://cms.web.cern.ch/news/new-cms-higgs-search-results-lepton-photon-2011-conference
[76] The LHCb Collaboration: A search for time-integrated CP violation in $D^0 \to h^+h^-$ decays and a measurement of the D0 production asymmetry; http://cdsweb.cern.ch/record/1349500/files/LHCb-CONF-2011-023.pdf
The LHCb Collaboration: Measurement of the CP Violation Parameter A_Γ in Two-Body Charm Decays; http://cdsweb.cern.ch/record/1370107/files/LHCb-CONF-2011-046.pdf



Grid Computing - Technology and Applications, Widespread Coverage and New Horizons

Edited by Dr. Soha Maad

ISBN 978-953-51-0604-3
Hard cover, 354 pages
Publisher InTech
Published online 16, May, 2012
Published in print edition May, 2012

Grid research, rooted in distributed and high performance computing, started in mid-to-late 1990s. Soon afterwards, national and international research and development authorities realized the importance of the Grid and gave it a primary position on their research and development agenda. The Grid evolved from tackling data and compute-intensive problems, to addressing global-scale scientific projects, connecting businesses across the supply chain, and becoming a World Wide Grid integrated in our daily routine activities. This book tells the story of great potential, continued strength, and widespread international penetration of Grid computing. It overviews latest advances in the field and traces the evolution of selected Grid applications. The book highlights the international widespread coverage and unveils the future potential of the Grid.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Dagmar Adamová and Pablo Saiz (2012). Grid Computing in High Energy Physics Experiments, Grid Computing - Technology and Applications, Widespread Coverage and New Horizons, Dr. Soha Maad (Ed.), ISBN: 978-953-51-0604-3, InTech, Available from: http://www.intechopen.com/books/grid-computing-technology-and-applications-widespread-coverage-and-new-horizons/grid-computing-in-high-energy-physics-experiments

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen