Review in Advance first posted online on May 10, 2011. (Changes may still occur before final publication online and in print.)

Computing for The Large Hadron Collider

Ian Bird

Information Technology Department, CERN, CH-1211 Geneva, Switzerland; email: Ian.Bird@cern.ch

Annu. Rev. Nucl. Part. Sci. 2011. 61:99–118

The *Annual Review of Nuclear and Particle Science* is online at nucl.annualreviews.org

This article's doi: 10.1146/annurev-nucl-102010-130059

0163-8998/11/1123-0099$20.00

Keywords

grid, WLCG, LCG, distributed

Abstract

Following the first full year of Large Hadron Collider (LHC) data taking, the Worldwide LHC Computing Grid (WLCG) computing environment built to support LHC data processing and analysis has been validated. In this review, I discuss the rationale for the design of a distributed system and describe how this environment was constructed and deployed through the use of grid computing technologies. I discuss the experience with large-scale testing and operation with real accelerator data, which shows that expectations have been met and sometimes exceeded. The computing system's key achievements are that (*a*) the WLCG infrastructure is distributed and makes use of all the dispersed resources, (*b*) the experiments' computing models are also distributed and can make excellent use of the infrastructure, and (*c*) the computing system has enabled physics output in a very short time. Finally, I present prospects for the future evolution of the WLCG infrastructure.

Contents

1. INTRODUCTION

The Large Hadron Collider (LHC) (see Related Resources for the Web sites of the facilities and collaborations mentioned herein) at CERN in Geneva, Switzerland, started full operation in 2010 following an initial run at the end of 2009. Its operation represents the culmination of nearly 20 years of preparation and construction. A role of the LHC is to search for new physics processes, which requires the detection of very small signals and thus the acquisition of very large statistical samples of data. The two largest experiments—ATLAS and CMS—are general-purpose detectors that use complementary detector designs. Of the two smaller experiments, LHCb is designed to make extremely precise measurements in order to understand the asymmetry between matter and antimatter in the universe, and ALICE is designed specifically to measure heavy-ion collisions (the LHC will accelerate Pb ions for part of each year) to investigate the conditions of the universe very shortly after the big bang.

Each of these detectors generates very large quantities of data in the form of discrete events, which are the detectors' recordings of the products of single proton-proton (*p-p*) or Pb-Pb collisions. The rate of *p-p* collisions within each detector is 10^9 Hz, which generates approximately 1 PB s^{-1} of data per detector [1 PB (petabyte) equals 1,000,000 GB]. Not all of these data are of interest for further analysis, so dedicated electronics and a large farm of processors located close to the detector make real-time decisions about which data should be retained. For ATLAS and CMS, this selection process reduces the event rate to a few hundred hertz. The resulting data rate of a few hundred megabytes per second is sent to the CERN Computer Center for archiving. LHCb has a much higher rate (~2 kHz) of interesting events, but the events are small and the resulting data-archiving rate is approximately 50 MB s^{-1}. The events that occur at ALICE during heavy-ion running, however, are extremely large, and the resulting data rate is approximately 1.25 GB s^{-1} (equivalent to creating a DVD full of data every 3 s). This article discusses the system of archiving, distributing, processing, and analyzing data built for the LHC. Including the data generated by the necessary physics and detector simulations, the LHC computing system must manage approximately 15 PB of new data each year. Initially, it was assumed that some 200,000 central processing units (CPUs) would be required, together with some 40 PB of disk space.

Although the main computing challenge for LHC is simply the huge volume of data and the numerous files that must be processed, several other considerations have been important in the

design of the computing environment. In addition to the accelerator data, numerous simulations must be performed, parts of which are computationally very intensive. The size of the LHC collaborations must also be taken into account, as they range between ~800 people for ALICE to ~2,000 each for ATLAS and CMS. Therefore, a great many people must be able to access and analyze the data, and the design of the computing system must allow physicists to access their data from their home institutes all over the world.

To summarize, the basic requirements for the computing environment are (*a*) to be able to manage very large data volumes at very high data rates; (*b*) to provide the requisite raw computing capacity—both CPU and storage; (*c*) to allow thousands of users to access the data; and (*d*) to provide long-term data archiving in a robust way. In addition, the environment must be able to manage organized data processing for real and simulated data, as well as for so-called chaotic user analysis. In data processing, a few production managers for each experiment run a standard set of programs on all of the data sets, whereas in user analysis, the need for specific data sets is unpredictable and depends on the needs of the various physics-analysis teams and the details of the analysis to be undertaken—which evolve as the analysis progresses.

Early on, it became evident that, for various reasons, placing all of the computing and storage power at CERN to satisfy each of the above-described needs would not be possible. First, the infrastructure of the CERN computing facilities could not scale to the required level without significant investments in a power and cooling infrastructure many times larger than what was available at the time. Second, the CERN member states have historically encouraged roughly two-thirds of the computing needs of an experiment to be located outside of CERN. However, the physics institutes and national laboratories participating in LHC experiments do, of course, have local computing facilities, often of significant scale; given the expectation of improvements in wide-area networking, it seemed reasonable that a distributed system able to couple these resources could achieve the scale needed. There are sociological advantages in such a distributed system: Investment is made locally and benefits the institute in several ways, namely by providing training for students and providing the ability to leverage the resources for other local uses.

The first model of a potential distributed computing system for LHC, created in 1999, was the so-called MONARC model (1). **Figure 1** shows the essential components of the model, namely the tiers of computing centers and the outward flow of data as they are processed. This model is discussed in greater detail below.

At the same time the MONARC model was being conceptualized, the computer science community was developing grid computing (2). In 2000, at the International Conference on Computing in High-Energy and Nuclear Physics (3), the two ideas came together, and it was realized that grid computing could provide the technical implementation of the LHC distributed computing environment. Certain fundamental concepts in grid computing addressed several of the problems that we were trying to solve for LHC, specifically (*a*) integration of computing resources across multiple administrative domains and (*b*) construction of an authentication and authorization infrastructure to allow users to access resources that may be distributed around the world.

Following this conference, a series of development projects were launched, in both Europe and the United States, to create prototypes and to test the basic ideas of grid computing and whether it would fit the needs of the LHC. All these projects were based on the reference software (i.e., grid middleware) provided by the Globus project. In Europe, the EU-Datagrid project aimed to create a prototype of a set of middleware that would enable the LHC applications to use the basic grid system; in the United States, the PPDG (Particle Physics Data Grid) project pursued similar goals. Thus, by the end of the year 2000, the new paradigm of grid computing had the potential to implement a globally distributed computing environment for the LHC.

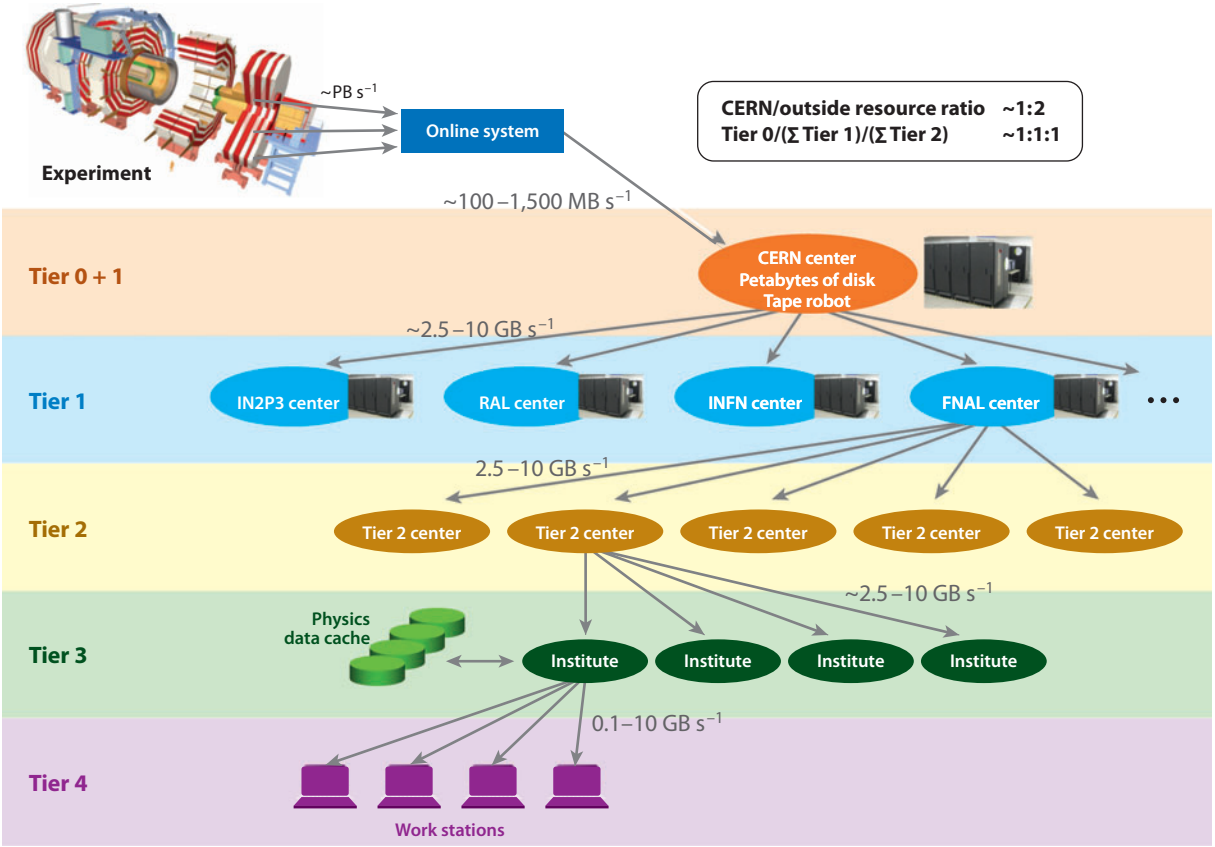

Figure 1

The MONARC model of LHC computing. Data flow outward from Tier 0 to the distributed Tier 1 sites and then to Tier 2. MONARC is the original concept for distributed LHC computing in 1999.

2. DEVELOPMENT AND EVOLUTION OF THE WLCG INFRASTRUCTURE

2.1. The LCG Project and the WLCG Collaboration

Although the projects mentioned above were pioneering the use of grid technology for physics, building a real production environment for use by thousands of physicists required a significant effort to create the underlying infrastructure, including, for instance, user, operational, and software support. Also, physicists needed to gain basic experience in actually managing such a novel computing environment. In 2001, the LHC Computing Grid (LCG) project (4) was proposed and approved to address these issues.

LCG was conceived in two phases. The first phase, from 2002 to 2005, involved developing and creating prototypes for various aspects of the distributed computing environment and learning how to integrate these components into the production environments of the participating computer centers. The second phase, from 2005 to 2008, involved setting up the distributed computing environment and commissioning services in preparation for the start-up and initial running of the LHC. During the second phase of the project, the framework for the long-term management of the

infrastructure was organized; this framework is known as the Worldwide LHC Computing Grid (WLCG) Collaboration. This collaboration is managed through a memorandum of understanding (MoU) (5) between CERN and the various national funding agencies that provide resources and support for the participating computing centers. Currently, there are 49 signatories to the MoU, representing 34 countries. The collaboration includes the Tier 0 site at CERN, 11 Tier 1 sites distributed around the world, and 130 Tier 2 sites organized into 66 management "federations" (the tiers are described more fully in Section 2.2). These sites, together with the four LHC experiments, are the formal members of the WLCG Collaboration. In addition, there are numerous other sites that are physically equivalent to the Tier 2 sites but that have not (yet) signed the MoU, along with many Tier 3 sites that are outside the scope of the MoU.

One of the principles embodied in the MoU is that a site, be it Tier 1 or Tier 2, that provides resources for an experiment does so for all members of the experiment. A Tier 3 site, however, may be national or local and available only to members of the experiments at that particular site.

In addition to functioning as an agreement of collaboration, the MoU defines the levels of service to be provided by the sites in each tier. It also describes the mechanism with which to define the level of resources—computing, storage (disk and tape), and networking—to be provided by each site. The collaboration reports twice a year to the funding agencies on the actual use of the resources and undergoes regular scientific review by the LHCC (6), which is the main scientific review body for the LHC experiments.

All the members of the collaboration support and use the WLCG infrastructure, which consists of significantly more than the software that enables distributed computing. The following sections summarize the elements of that infrastructure and the software architecture that implements the grid.

2.2. The WLCG Infrastructure

The main elements of the infrastructure are the tiers of computer centers. Their roles and responsibilities are described below.

The Tier 0 center at CERN is responsible for accepting the data streaming from the experiments at the full data rates, and it must ensure that these data are immediately archived to tape for safekeeping. The policy for ensuring that raw data are reliably stored stipulates that all the data be stored at CERN and that a second copy be distributed among the Tier 1 centers. Thus, two full copies of the raw data are kept. Ensuring distribution to the Tier 1 centers is also the responsibility of Tier 0. For this purpose, dedicated 10 GB s^{-1} network connections to each Tier 1 have been established. Should a Tier 1 site go off-line, to ensure that data distribution can catch up once the site is back online, the bandwidth of data distribution out of CERN has been scaled to sustain at least twice the nominal data-taking rate. This rate is some 1.3 GB s^{-1}—equivalent to a DVD of data every 3 or 4 s. In fact, the network must be able to support even more than this because, in addition to the raw data, processed data must also be distributed out of CERN. This network is illustrated in **Figure 2**.

Tier 0 is responsible for ensuring that the data on tape are usable and accessible for the long term. These responsibilities are shared by the Tier 1 centers, which is one reason that Tier 1 sites are generally large centers that have experience with large-scale data management.

Tier 0 also provides a large computing cluster; as of the end of 2010, some 50,000 CPU cores are used for tasks that must take place close to the experiments. These tasks include detector calibration and alignment, which must occur before large-scale processing of the data can be started. Tier 0 is also used for first-pass reconstruction of the data. The results of such processing

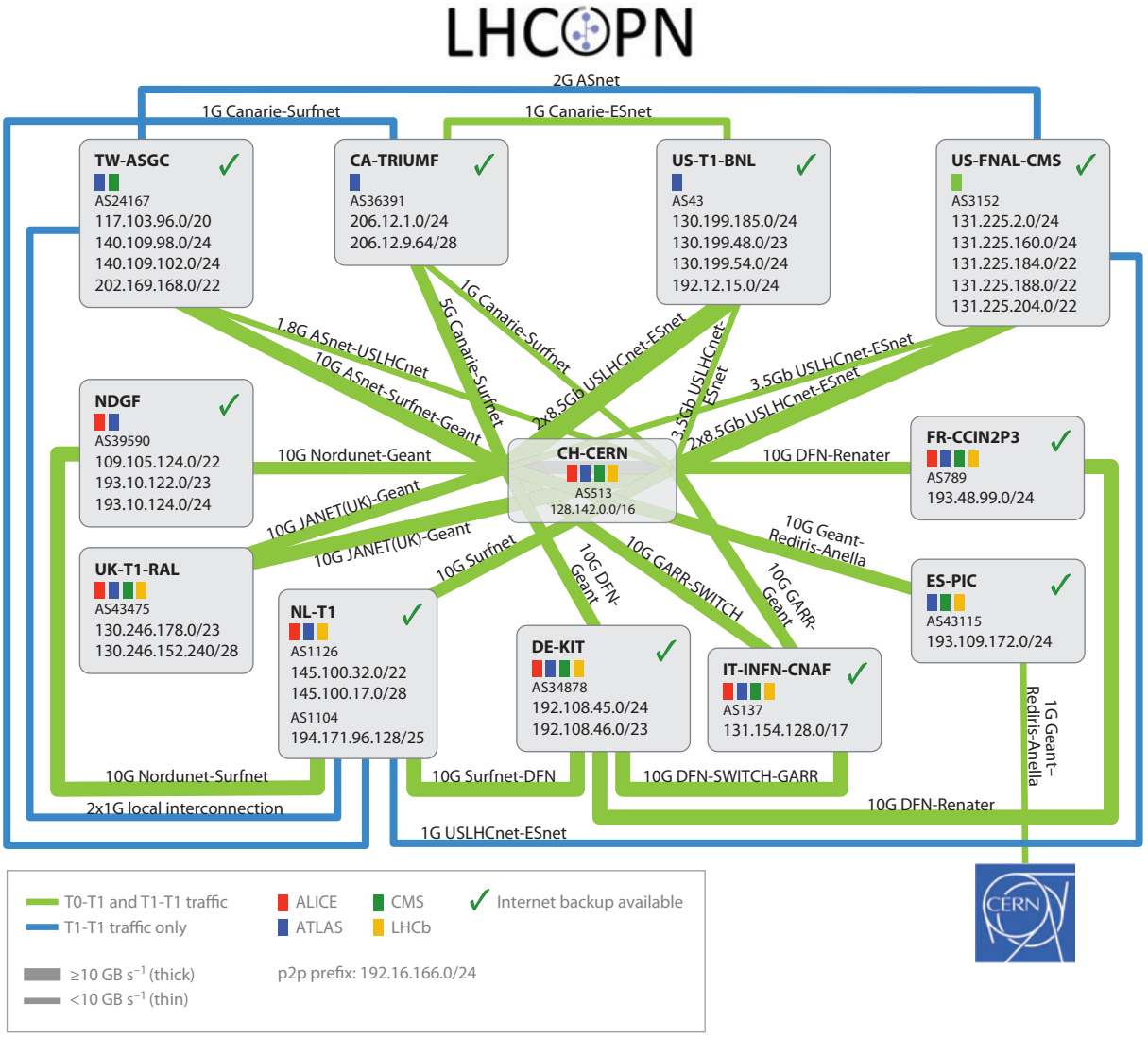

Figure 2

The LHC Optical Private Network (LHCOPN). Dedicated "dark-fiber" connections between CERN and each of the Tier 1 sites were implemented to ensure the necessary data rates. These connections are supplemented with secondary connections to ensure reliability of the network. Connectivity between Tier 1 and Tier 2 sites is provided by national and international academic and research network infrastructures.

are also archived at CERN and distributed to the Tier 1 sites. Part of the Tier 0 cluster is also designated as a data-intensive analysis facility.

Finally, Tier 0 is responsible for the coordination and overall operation of the grid. The levels of service defined for Tier 0 in the MoU are relatively high. For example, during accelerator running, main services such as raw data recording must be available 99% of the time. These parameters, together with those that define the service levels for the other tiers, can be found in appendix 3 of the MoU (5).

Of the 11 Tier 1 sites, several support many or all of the experiments. There is one site in Canada (ATLAS), two in the United States (one for CMS and one for ATLAS), one in the Asia-Pacific region (for CMS and ATLAS), and seven in Europe; four of those seven support all four experiments, and the other three support three of the experiments.

The Tier 1 centers must accept an agreed share of the raw and reconstructed data from Tier 0 and must provide for the long-term archiving of those data as well as for the storage of simulated data. The Tier 1 sites also operate computing clusters that are used for (re)processing of the raw data with improved calibrations. Typically, they provide data-intensive analysis facilities that are used for large-scale organized analysis.

The Tier 1 sites support a number of Tier 2 centers. These Tier 2 sites are generally in the same country as the Tier 1 sites, but there are numerous exceptions, particularly in Europe, where several countries have Tier 2 sites but no Tier 1 sites. The assignment of Tier 2 sites to Tier 1 sites may be different for each experiment. In addition to providing data to the Tier 2 sites, the Tier 1 sites provide support in case of operational problems. The Tier 1 centers also run a number of grid services whose levels are defined within the MoU.

The Tier 2 centers are essentially analysis facilities that are provided for the use of an entire experiment. They also provide computing power for the Monte Carlo simulations needed by the experiments. The results of those simulations are sent to Tier 1 for archiving and processing.

The organization of the Tier 2 sites varies from country to country. In some cases, as in the United States, a Tier 2 site supports only one experiment. In Europe, it is common for a Tier 2 site to support several or all of the experiments. The organization depends on the host institution and its membership in the experiments.

Tier 2 centers provide computing and disk storage resources but are not required to archive data; they generally run only the grid services that are needed to provide access to their resources. Thus, Tier 2 centers are simpler to manage than Tier 1 sites and are often run by a very small team. The service levels for Tier 2 sites are defined in the MoU.

Although Tier 3 sites are not described in the MoU, they are important as clients to the Tier 2 sites, which must be able to distribute data sets to Tier 3 for end-user analysis. Tier 3 sites vary in scale enormously, ranging from a few CPUs to large national analysis facilities.

The network between the various tiers is critical to the success of this infrastructure. Whereas the connection between Tier 0 and Tier 1 is based on dedicated and redundant links, those between the Tier 1 and Tier 2 sites rely on national and international academic and research networks.

Figure 3 shows that several layers of support and management make up a large part of the overall infrastructure. These components include processes and tools to support daily operation, problem solving, and metrics reporting. An important tool is the site availability monitor (SAM), which measures and improves grid-site reliability and availability. The SAM measures the availability of a grid site by running grid jobs at the site to test elements of the overall service. The results of the tests are used both to send alarms to the site to report possible service problems and as a metric to report the site's overall availability and reliability.

Tier 0 provides ~20% of the total computing resources, and Tier 1 and Tier 2 sites contribute ~40% each. Currently, there are approximately 250,000 CPU cores, along with 100 PB of disk space provided by WLCG.

Management of identity within the WLCG and similar grid infrastructures is based on the use of X.509 certificates (7). Access to resources at grid sites is controlled by a two-part process: (*a*) A user presents a certificate that authenticates his or her identity, and (*b*) the user is then recognized as a member of a valid virtual organization (VO) that is allowed access to the grid-site resources by the resource provider.

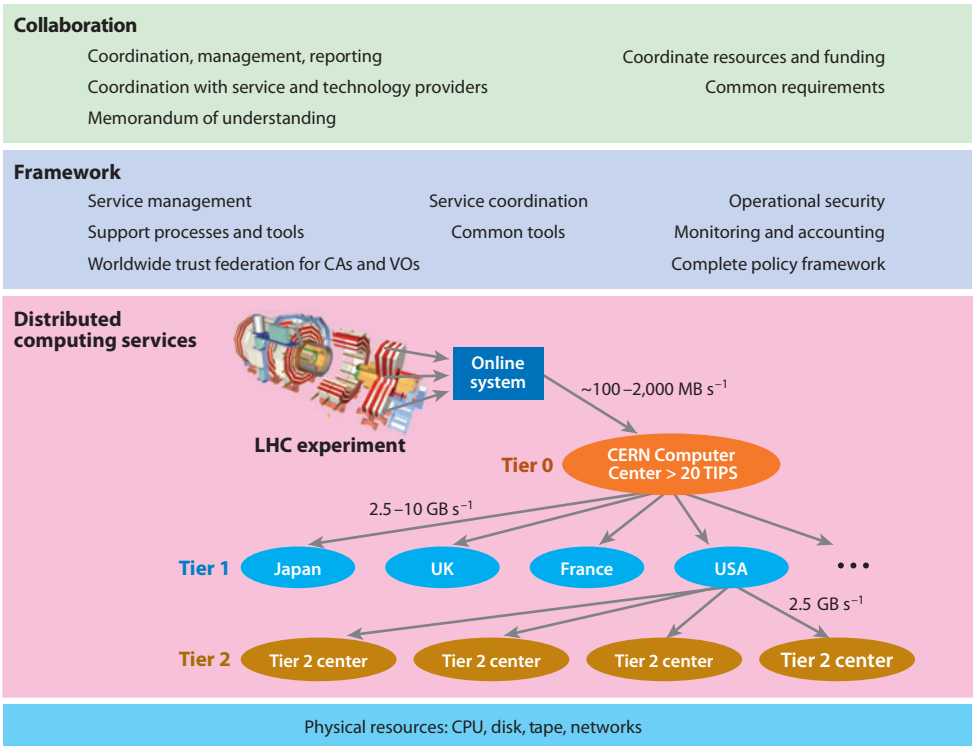

Figure 3

Elements of the WLCG infrastructure. In addition to the physical resources and the middleware layer, the support and management structures are important for deploying and managing the service. Abbreviations: CA, certification authority; VO, virtual organization.

Networks of trust have been set up to provide the authentication of users through the issuance of certificates and through membership in VOs. In both cases, the foundations of the trust relationships are in legal entities—usually the employing organizations of the users.

Having comprehensive policies in place is essential to deploying the infrastructure across so many different countries, and it allows the grid to work within the requirements of privacy and security. These policies have been developed over many years, in consultation with other grid projects, to ensure that the same policies can be used everywhere to the greatest extent possible.

2.3. The WLCG Grid Implementation

The technical implementation of the grid is based on an underlying set of grid middleware (8) and some higher-level services. These services are summarized below and in **Figure 4**.

2.3.1. Storage services. A storage element provides a set of storage services at a given site. Its standard interfaces are based on the storage resource manager, which provides the same interface regardless of the underlying storage system. Various implementations are available; those at Tier 1 sites have a tape archive behind the disk cache. Storage elements also provide a GridFTP service to move data into and out of the storage, as well as a mechanism to provide local data transfer to and from the disk cache.

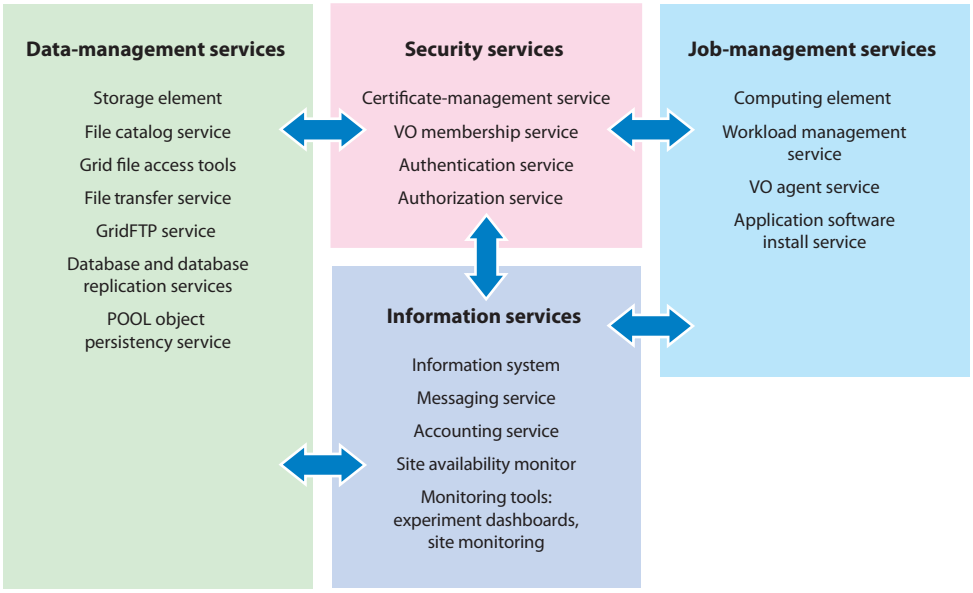

Figure 4

Grid middleware components are shown in terms of their functionality. Abbreviation: VO, virtual organization.

2.3.2. File transfer services. The underlying protocol for file transfer is GridFTP. A file transfer service provides error recovery, retry on failure, and a mechanism to share bandwidth between experiments on a specific network connection. The service enables third-party copies between remote sites. The file transfer service is run at Tier 0 and at each of the Tier 1 sites; other sites act as clients.

2.3.3. Computing services. The compute element provides the interface to the computing facilities at a site, providing job submission and query functions, the interface to the accounting system, authentication and authorization mechanisms based on a user's certificate, and VO membership and roles.

2.3.4. Workload management. Workload management is provided by a general service that provides brokering capabilities to locate resources on the grid that satisfy the needs of a particular job. During the development of the infrastructure, this functionality has generally moved into the experiments' own work-flow systems, as the necessary integration with experiment information can be better managed.

2.3.5. File catalog and database services. Database services are run at the Tier 0 and Tier 1 sites. These services include file catalogs that show which data files are located at each storage element, conditions databases that describe the alignment and calibration parameters of the detectors as input to the reconstruction-and-analysis software, and databases for experiment-defined metadata. The conditions databases are regularly updated replicas of a master at Tier 0, whereas the file catalogs are explicitly updated with information about files at a particular location. The metadata of each experiment is usually in a central database.

2.3.6. Virtual organization agents. A mechanism is provided to permit experiment-specific services to be run in a standard way at grid sites. The experiment takes responsibility for the management of the service, although the site manages the underlying hardware.

2.3.7. Information service. An information service provides a lookup point for information published by each of the service instances, describing its configuration and details of, for example, the storage and computing resources.

2.3.8. Application software. The experiment application software is regularly updated and must be made available at each site. A suite of standard utilities are also provided.

2.3.9. Interoperability. The WLCG grid infrastructure relies on large national and international grid projects for some of the underlying tools and support services. The major science grid infrastructures are the OpenScience Grid (OSG) in the United States and the European Grid Infrastructure (EGI) in Europe. The latter is the successor to the EGEE and NorduGrid projects, and it coordinates the various national grid initiatives. WLCG works closely with these infrastructures to ensure interoperability, even though they may use different base middleware and tools.

3. DEVELOPMENT AND EVOLUTION OF THE EXPERIMENTS' COMPUTING MODELS

All the experiments' computing models were based on the MONARC model discussed above. However, each experiment based its implementation on different key choices, which has resulted in quite distinct models. These models have also evolved after having been tested at scale. Each of the experiments has developed a software layer that integrates its applications with the distributed computing environment.

The computing model used by ATLAS (**Figure 5**) (9) is the closest in concept to the MONARC model. Raw data from the detector are sent to Tier 0 at ~320 MB s^{-1}; these data are then archived

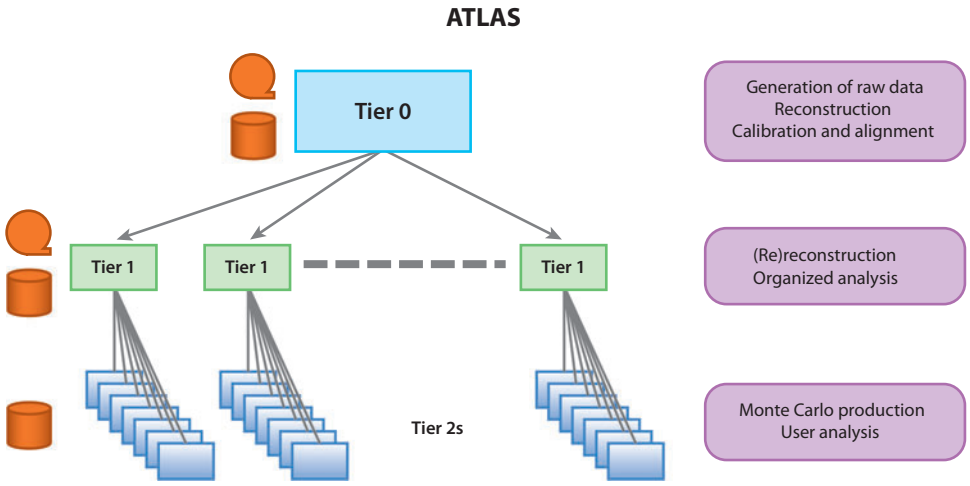

Figure 5

The ATLAS computing model.

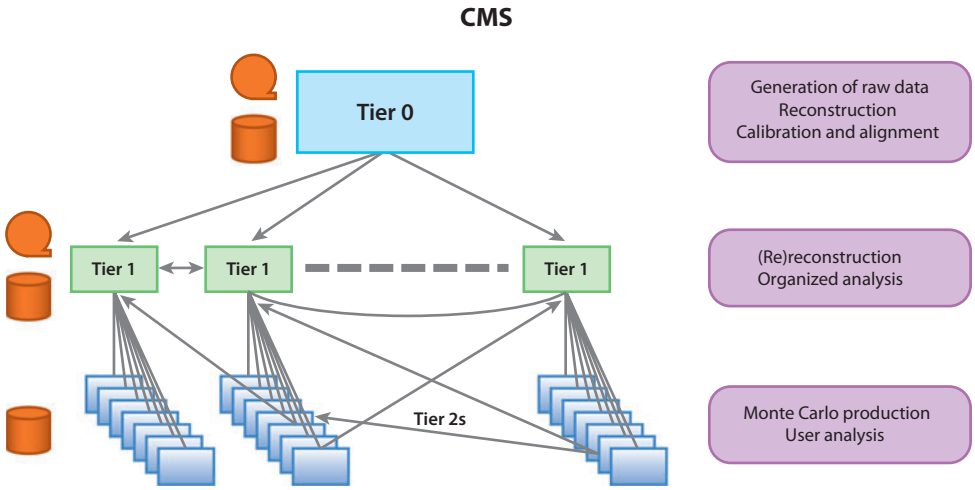

Figure 6

The CMS computing model.

to tape and cached on disk, and a second copy is distributed among the 10 ATLAS Tier 1 sites. Calibration and alignment are run at Tier 0, and a small sample of raw data is immediately reconstructed for quality checks. Once the calibration and alignment are reasonably accurate, the first-pass reconstruction is performed on the raw data, and the results [event summary data (ESD) and a reduced format, AOD] are archived at Tier 0 and distributed to Tier 1. Two copies of the ESD are distributed so that each Tier 1 site has a primary copy of its share and a replica of the share at another Tier 1.

The Tier 1 sites are responsible for reprocessing their shares of the raw data once better calibrations are available, as well as for making their ESD and AOD available to the collaboration. They also host simulated data produced at their associated Tier 2 centers. The AOD samples at each Tier 1 site are replicated and shared among the associated Tier 2 sites.

The Tier 2 sites are responsible for hosting their share of the AOD data sets and for providing analysis facilities as well as resources for producing an agreed amount of simulated data. Physics analysis work is sent to a site that hosts the relevant data sets.

CMS has a somewhat different model (**Figure 6**) (10) that, although based on the original MONARC model, is less hierarchical in its implementation. CMS allows data to be moved between any Tier 1 and Tier 2 sites. Raw data are received at Tier 0 from CMS at a rate of approximately 250 MB s^{-1}, and Tier 0 is used in the same way as for ATLAS. Raw and processed data are distributed to the seven CMS Tier 1 centers.

Although the functions of CMS's Tier 1 and Tier 2 sites are very similar to those of ATLAS, the way data are distributed is driven by policies that can be changed to adapt to different priorities. Although the raw data are shared among the Tier 1 centers to ensure a full second copy, the placement of ESD and AOD is driven by specific needs, and a given Tier 1 site may hold the only copy of a given data set, which thus must be made accessible from any other CMS site.

Although a CMS Tier 2 site is associated with a specific Tier 1 center that runs services and provides support for it, the Tier 2 site may nevertheless request data from any Tier 1 center (or from another Tier 2 site). Thus, the CMS model has potentially many more data connections than the ATLAS model and can be visualized more as a mesh of connections than as a hierarchy.

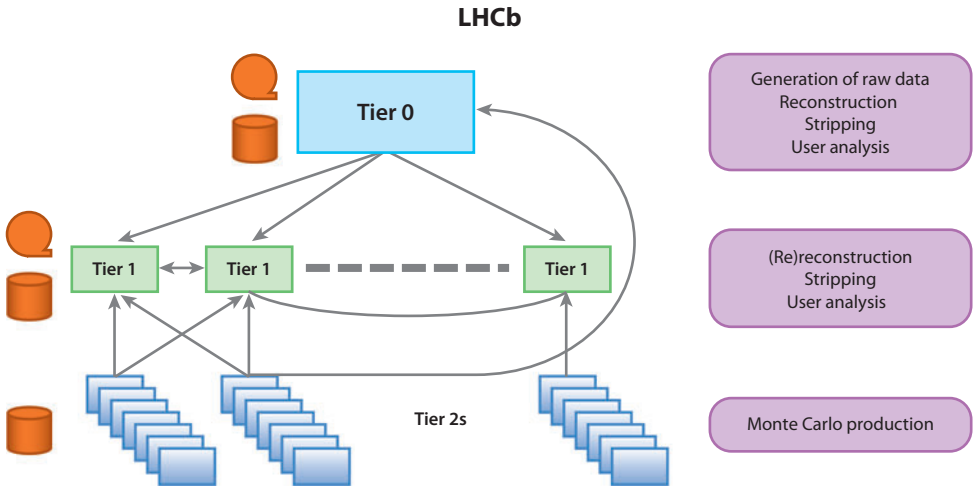

Figure 7

The LHCb computing model.

The LHCb computing model (**Figure 7**) (11) differs from those of CMS and ATLAS in terms of data rates and analysis. The LHCb trigger rate is very high, but the event size is small, so LHCb sends raw data to Tier 0 at a rate of approximately 50 MB s^{-1}. LHCb's use of Tier 0 and its distribution of data to its six Tier 1 sites are very similar to the procedures followed by ATLAS and CMS. The results of the reconstruction, known as DST in LHCb, are also distributed to the Tier 1 centers. Subsequent reconstruction passes use Tier 1 resources.

The Tier 1 centers (which, for LHCb, include the CERN facility) also provide resources for selecting events according to various physics criteria. This process, known as stripping, is data-transfer intensive and must be run on the full DST samples to produce stripped DSTs that can be used for analysis. In contrast to ATLAS and CMS, LHCb analysis work takes place at the Tier 1 centers using the stripped DSTs available there. These DSTs are shared among the Tier 1 sites.

The LHCb Tier 2 sites are used only for producing simulated data. These data are sent to the associated Tier 1 sites for archiving and processing.

The ALICE experiment's model (**Figure 8**) (12) differs from those used by the other three experiments. Its main feature is that there is no significant distinction between the Tier 1 and Tier 2 sites, apart from the data-archiving responsibilities of the Tier 1 sites.

During heavy-ion running, ALICE sends data to Tier 0 at a very high rate: ~1.25 GB s^{-1}. These data are archived at Tier 0, but not all the data can be immediately sent to the six Tier 1 sites because these sites' aggregate data-transfer capacity is insufficient for this rate. Rather, the data are sent during the period following the heavy-ion run. To insure against data loss, two copies of the raw data are kept at Tier 0 until a second copy has been made at the Tier 1 sites.

The Tier 1 centers provide reconstruction to the ESD level, whereas subsequent data reduction to AOD, data analysis, and simulation work are shared across all Tier 1 and Tier 2 sites and the CERN analysis facility. The ALICE model allows data transfer between any two sites. In this model, the only real distinctions between the tiers are in the quality of service and the archiving capabilities of the Tier 1 sites.

As noted above, each of the experiments has built a layer of software infrastructure to implement these models, in addition to the basic set of grid middleware services described in Section 2. All these models arose from fundamental concerns that wide-area network capacity would not be

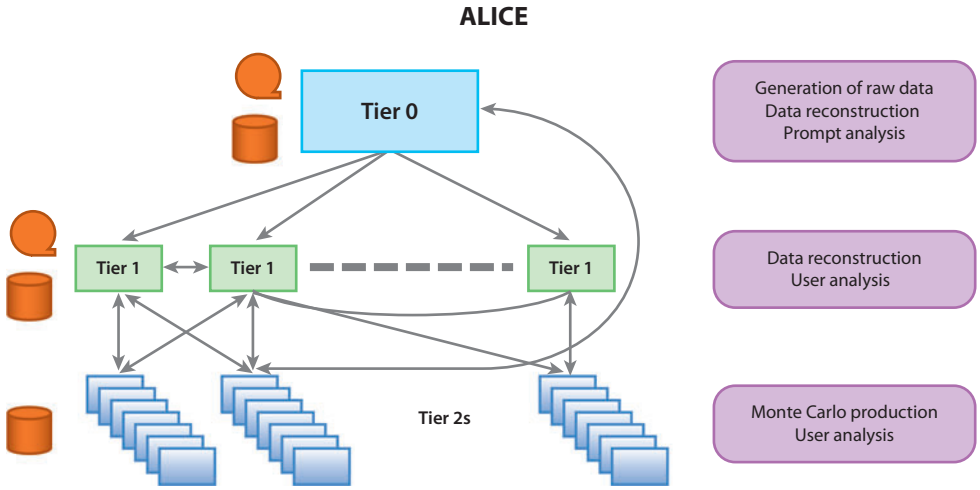

Figure 8

The ALICE computing model.

sufficient for the experiments' needs and that the network would not be reliable enough. Many of the models' services have been replicated to ensure reliability. Between the time the models were conceived and the start of the LHC, the network capabilities exceeded all expectations in terms of both capacity and reliability. The latter is ensured by secondary connections among Tier 0 and all the Tier 1 sites (**Figure 2**).

Following the first year of data taking at the LHC, it became clear that the network could be used far more optimally. The computing models are evolving to reduce the amount of unnecessary data movement and are allowing remote access to data under certain circumstances. The strict hierarchy of the original MONARC model will be relaxed, and concepts of models such as that used by ALICE and the mesh structure of CMS will become more developed and prevalent.

4. WLCG RESULTS IN TESTING AND WITH DATA

As of early 2011, the first year of LHC data taking has completed, and the performance of the WLCG infrastructure and the experiments' computing models has exceeded expectations. Many physics results have already been published, which would not have been possible without the significant program of testing and challenges that had been run over the six years prior to the start-up of the LHC. Since 2003, when the initial prototype grid system was put in place with only a handful of grid sites, the system has been used for the experiments' production of simulated data, and since 2004 an extensive program of data and service challenges has been employed. **Figure 9** illustrates the main features of this program.

Early testing took the form of limited-scope data challenges organized by the experiments along with the CERN team. An example of such a data challenge is the effort to build up the rate of data transfer into Tier 0 for ALICE, which had the most demanding data-acquisition rate. The full chain of simulation production and processing was first implemented in the grid infrastructure in 2004.

Also in 2004, the LCG project proposed a series of service challenges that were intended to demonstrate and test various aspects of the grid service. Initially, these challenges focused on achieving the necessary scale and reliability of data transfers; later, they expanded to the testing of data management, scaling of job workloads, and testing of key support processes including response

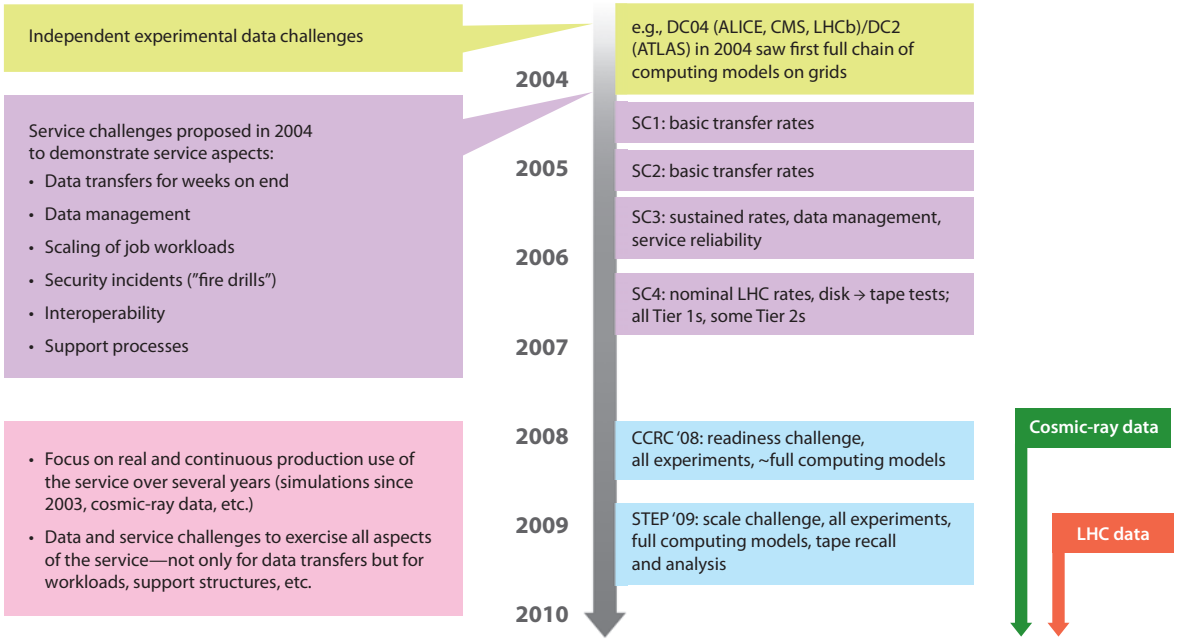

Figure 9

The time line of the program of testing and service challenges and the roles of the WLCG infrastructure.

to simulated security incidents. Results of these challenges and the growth of the infrastructure have been reported in many international conferences and symposia; refer to the Related Resources (7, 8).

In parallel with these organized periods of service challenges, the grid infrastructure was gradually ramped up in terms of resources in preparation for data taking. Because the infrastructure has been the de facto production environment for all of the computing needs of the experiments since approximately 2004, it has been in full-scale use for many years.

In addition to these organized tests, between late 2008 and late 2009, when the LHC restarted operations, significant amounts of cosmic-ray data were acquired and processed by the experiments. These data allowed full testing in real conditions of the experiments' entire computing models, from data acquisition to analysis, and thereby validated the entire system. Often, by increasing the data-acquisition rates this testing could be performed at rates close to that anticipated for LHC running. The testing also allowed preliminary calibrations and alignment of the detectors in preparation for the actual LHC start-up. The testing program culminated in two major readiness challenges—one in 2008 (CCRC'08) and one in 2009 (STEP'09)—that demonstrated WLCG's preparation for the start of data taking.

During 2010—the first full year of LHC data taking—the data rates achieved across the grid were much higher than those achieved during testing. The peak rates of data transfer out of CERN were greater than 5 GB s^{-1}; in contrast, the nominal rate is approximately 1.5 GB s^{-1}. **Figure 10** shows the rates achieved during STEP'09 and during the 2010 run.

Grid-wide data movement is also important. For example, ATLAS reports an overall data-transfer rate of up to 10 GB s^{-1} globally; this rate includes data movement to the Tier 2 sites for analysis. The expected grid-wide job workload during nominal data-taking years was estimated to be approximately 1 million jobs per day. This workload has consistently increased over the past

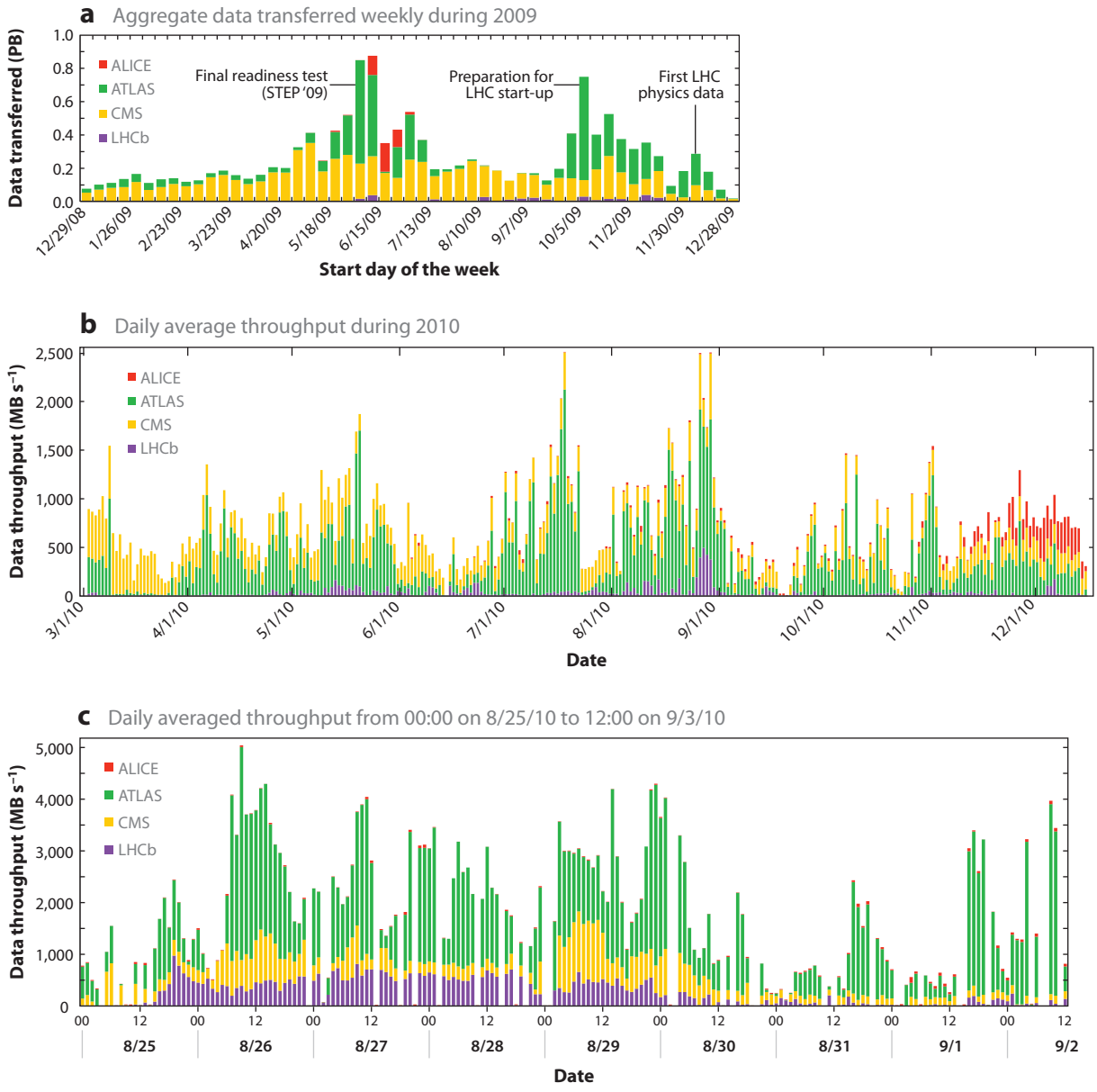

Figure 10

Data rates achieved during (*a*) STEP'09 and (*b*) the year 2010. (*c*) An expansion of several days in 2010 shows peak transfer rates. The aggregate data rate out of CERN to the Tier 1 sites is shown. The nominal rate of 1.3 GB s^{-1} was often significantly exceeded.

several years (**Figure 11**) and is probably significantly higher than indicated, as the number of jobs plotted in the figure is based on those jobs submitted through the grid interfaces. Currently, such submissions represent the so-called pilot jobs of the experiments—they function as containers into which the actual workloads are pulled during execution, and several tasks can be run in a single pilot job.

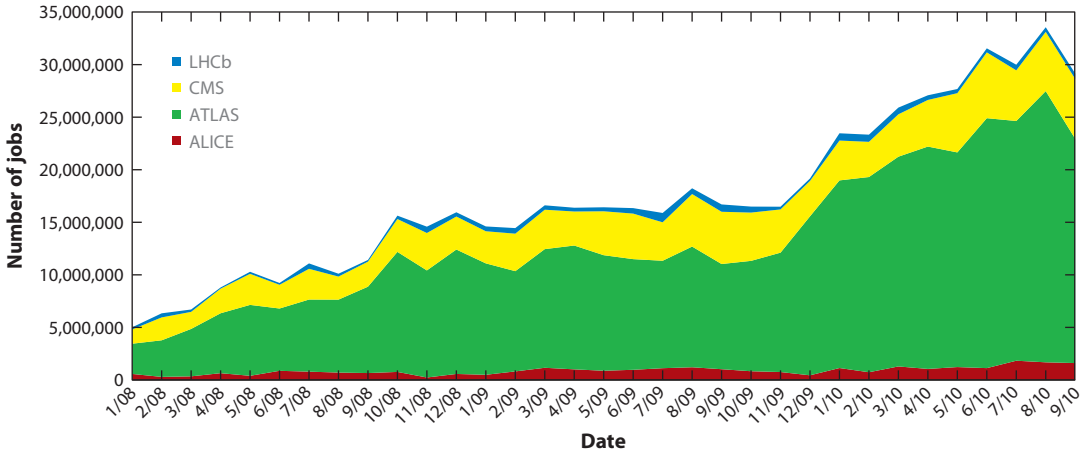

Figure 11

Evolution of grid job workloads. The number of jobs run per month, summed over the full WLCG infrastructure, is shown. More than 1 million jobs per day are currently run.

One of the successes of the grid is that work is distributed across the entire system. **Figure 12*a*** shows an example of the distribution of CPU time among the Tier 0, Tier 1, and Tier 2 centers, and **Figure 12*b*** illustrates the distribution among the Tier 2 centers by country.

Notably, the Tier 2 centers provide more than 50% of the overall computing resources, and the use of CPU time in each country is close to the capacity the country has pledged. The grid really does enable all institutes, both large and small, to contribute to the overall effort. **Figure 12** also shows that the early concern that users would attempt to use CERN to perform all analyses was unfounded; that analysis is actually performed at the Tier 2 centers.

Similarly, the number of users now using the grid system is significant. ATLAS and CMS each report between 800 and 1,000 users per month, and LHCb and ALICE each report between 200 and 250 users. These figures are very encouraging and are significantly higher than the number of users who tested analysis during the testing campaigns.

The reliability of the services provided by all of the sites has been continually monitored since 2006. The improvement of service reliability during this time has been slow but continuous. The main causes of unreliability at the larger sites are problems with the large mass-storage systems, which Tier 2 sites do not have.

Significant effort has been invested in improving the overall reliability of the site services. These levels are now acceptable, but further improvements will be necessary to ensure that the operational support effort can be sustained over the long term. Some failures, be they due to power or cooling failures at a site or to hardware problems, will be unavoidable. In summary, the grid's performance during the first year of LHC running has been impressive and has enabled very rapid production of physics results.

5. EVOLUTION OF HIGH-ENERGY PHYSICS COMPUTING

The LHC experience has been leveraged to help build other major science grid infrastructures, such as EGEE and EGI in Europe and OSG in the United States. For both EGEE and OSG, WLCG has been the main inspiration, although many other application communities have contributed. A wide range of sciences can employ these infrastructures; they range from astroparticle

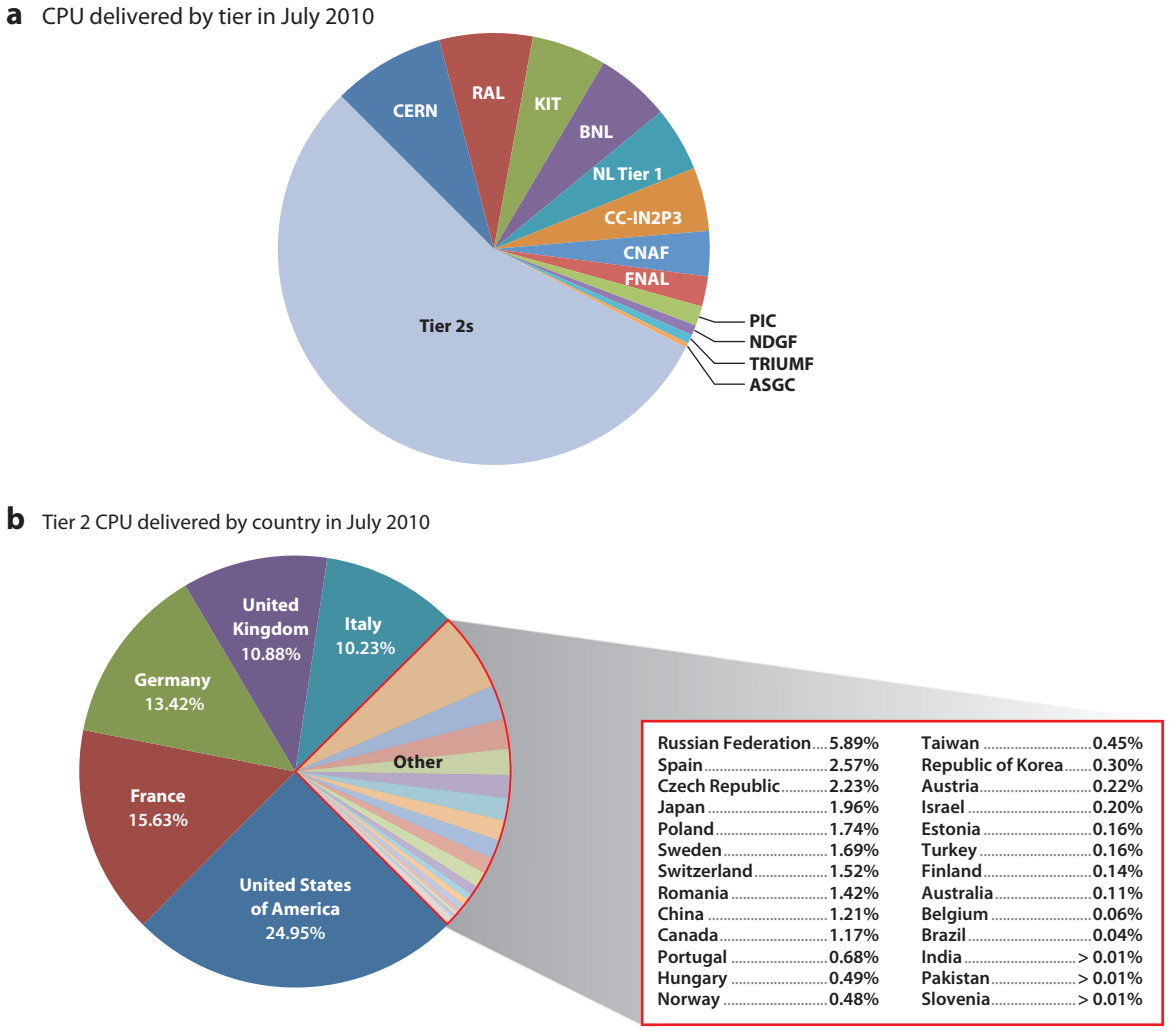

Figure 12

Distribution of central processing unit (CPU) time between tiers in July 2010. (*a*) The partition of CPU time delivered by Tier 0, the individual Tier 1 sites, and the sum of the Tier 2 sites. Tier 2 sites deliver more than 50% of the total. (*b*) The distribution of CPU time across the Tier 2 sites in each country. This distribution corresponds well with the amount promised by each country.

physics, whose needs are similar to those of high-energy physics, to biomedicine, which has very different requirements.

Although the first year of LHC data taking has been an excellent experience, there are many areas in which there are lessons to be learned or in which there are causes for concern. Such challenges (as those noted at the end of Sections 3 and 4) will drive the evolution of the WLCG infrastructure in the future and influence high-energy physics computing, given that other experiments are either using or intending to use similar distributed computing infrastructures.

Some early ideas about grid functionality were more complex than what was actually required by the experiments; the converse is also true. Complex work-flow management is a good example: When an experiment has complete information about types of data and their location, it can

straightforwardly direct work to the appropriate data, whereas attempting to implement such a work flow in a generic way leads to complex solutions. As noted above, concern that networks would be unreliable led to deployment of services at several sites, which often added to the complexity.

Also, there is a certain level of site unreliability that may be difficult to reduce, particularly for power and cooling failures. The computing models generally do not take into account the periodic unavailability of sites, although sites do go off-line either in an organized way for maintenance or when something fails. Tier 1 sites are particularly critical for the experiments. Therefore, computing models must evolve to take periods of unavailability into account.

Another area in which evolution of the models and tools is needed, based on experience in the past few years, concerns the availability of data sets. A complete data set may consist of many thousands of files, and often some small number of those files either is not present at a site or is temporarily unavailable. Such instances of unavailability are considered failures of processing. However, given the reliability and bandwidth available in the network layer, in most cases it should be possible to find the data remotely and to be able to access them (either remotely or by copying the data locally). Although doing so may lead to some inefficiency in CPU use, from the user's point of view, it represents greater reliability.

During the first year of the LHC's operation, all the data were copied to the tiers, as had been proposed in the computing models, but not all the data were accessed, and some were accessed very infrequently. Thus, it seems reasonable to implement a more dynamic model of data placement, particularly at the Tier 2 sites, either according to the popularity of the data or by request. Such a model, in conjunction with organized data placement at Tier 1 sites, would probably improve both the efficiency of the disk caches and the use of the networks, thereby reducing unnecessary data movement and freeing up disk cache for data that will be used.

Along with such natural evolution in the models, there are issues of long-term sustainability that must be addressed:

1. Use of multicore and other CPU types. Typically, in high-energy physics data analysis, a process is run on a single processor core—with a single data file as input and a single file as output. This model has worked well for many years and has allowed work to scale to many processors in parallel simply by running more jobs. This process is not suitable for the multicore CPUs that are now becoming common, because the needs for memory scale with the number of processes and the cost is prohibitive. Thus, efforts are under way to allow applications to effectively share memory between cores, or to make the applications truly multithreaded. Other strategies to better use multicore processors are also being investigated, as is the possibility of using graphics-processing units in certain circumstances.
2. Replacement of certain components of grid middleware with more standard software. In certain areas, this has already happened—for example, open-source monitoring systems are used to reimplement the SAM service, and open-source (or commercial) messaging services provide a mechanism for interprocess communication. Such replacements will help to provide long-term support for the software underlying the infrastructure.
3. Use of virtualization. Virtualization, already in use at many grid sites, is a good way to eliminate interdependence among the operating system, the grid middleware, and the application software. This dependence has been a significant problem over the past few years because it hinders upgrades to all parts of the software. Virtualization will probably play an increasingly important role in the management of the various grid sites, and eventually individual sites may well be run as fully virtualized infrastructures, although concerns about data-transfer overheads and how virtual machines fit with ideas of how to best make use of multicore processors must be addressed. Nevertheless, the concept of hybrid clouds may, over the very long term, allow for the implementation of a grid infrastructure that uses technology

that is in widespread use outside the grid community. There is also some interest in utilizing public clouds for certain types of work as a supplement to the WLCG resources. Use of a virtualized infrastructure with cloud interfaces would simplify such extensions.

In the world outside particle physics, grid computing has been replaced by cloud computing, and the WLCG infrastructure must be able to evolve to accommodate new mainstream technologies while maintaining the elements that make the grid concept so useful, namely (*a*) the collaborative aspects of worldwide authentication and authorization mechanisms and (*b*) the ability to bring together many dispersed resources in different administrative domains as an integrated whole. These features are crucial and must be maintained regardless of the underlying technology.

6. CONCLUSIONS AND OUTLOOK

The WLCG Collaboration has brought together some 150 computer centers from around the world for use by the LHC experiments as a solid production and analysis environment. The first year of LHC data taking has validated the infrastructure through the very rapid production of physics output. Several lessons have been learned from the years of testing and commissioning as well as from the early experience with real data. Although these lessons are important for guiding the evolution of the infrastructure, no major problems have been uncovered. The performance of the system in terms of scale and data throughput has matched and often exceeded expectations.

The experiments' computing models will evolve over the coming years to more effectively use the distributed infrastructure. Importantly, the infrastructure itself can evolve to make use of new technologies while retaining the key elements of the distributed system and of the support and management structures. It is likely that in the near future, WLCG will utilize other computing paradigms, such as cloud computing or citizen cyberscience for some cases, to supplement the existing infrastructure. It is also likely that new collaborators will join and add resources to the existing infrastructure.

For several reasons, the experiments' computing models over the past 10 years have diverged somewhat. The use of such a large distributed service had no precedent, so different experiments had different ideas of how to proceed. It is becoming clear that certain concepts are common to all the experiments, which are focusing on using the infrastructure in the most effective possible way. Thus, we anticipate increasing commonality between the models, which will also be reflected in better prospects for solid support in the long term. The LHC computing environment will outlive the accelerator itself, but it will evolve along with technology and is likely to become very different over the next few years.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Aderholz M, et al. Models of networked analysis at regional centres. MONARC phase 2 rep. *CERN-LCB-2000-001* (2000)
2. Foster I, Kesselman C. *The Grid: Blueprint for a New Computing Infrastructure*. Amsterdam: Elsevier (2004)
3. Mazzucato M, Michelotto M, eds. *Comput. Phys. Commun.* 140(1/2) (2001)
4. CERN. Text of council resolution. *CERN/2369/Rev.* **http://cern.ch/LCG/PEB/Documents/c-e-2379Rev.final.doc** (2001)

5. WLCG Collab. Memorandum of understanding. *CERN-C-RRB2005-01*. **http://lcg.web.cern.ch/LCG/mou.htm** (2009)
6. LHC Exp. Comm. Mandate. **http://committees.web.cern.ch/committees/lhcc/welcome.html** (2010)
7. Int. Telecommun. Union. **http://www.itu.int/rec/T-REC-X.509/en** (2008)
8. Bird I, ed. LCG Baseline Services Group report. **http://cern.ch/LCG/PEB/BS/BSReport-v0.7.pdf** (2005)
9. ATLAS Comput. Tech. des. rep. *ATLAS-TDR-017*, *CERN-LHCC-2005-022*. **http://cdsweb.cern.ch/record/837738/files/lhcc-2005-022.pdf** (2005)
10. CMS Comput. Tech. des. rep. *CMS-TDR-007*, *CERN-LHCC-2005-023*. **http://cdsweb.cern.ch/record/838359/files/lhcc-2005-023.pdf** (2005)
11. LHCb Comput. Tech. des. rep. *LHCb-TDR-007*, *CERN-LHCC-2005-019*. **http://lhcb.web.cern.ch/lhcb/TDR/LHCb-comp-050613+authors.pdf** (2005)
12. ALICE Comput. Tech. des. rep. *ALICE-TDR-012*, *CERN-LHCC-2005-018*. **http://aliceinfo.cern.ch/static/Documents/TDR/Computing/Chap0/chap0.pdf** (2005)

RELATED RESOURCES

1. LHC home page. **http://public.web.cern.ch/public/en/LHC/LHC-en.html**
2. CERN home page. **http://cern.ch**
3. ATLAS Experiment home page. **http://atlas.ch**
4. CMS Experiment home page. **http://cms.web.cern.ch/cms/index.html**
5. LHCb Experiment home page. **http://lhcb-public.web.cern.ch/lhcb-public**
6. ALICE Experiment home page. **http://aliceinfo.cern.ch/Public/Welcome.html**
7. Globus Alliance home page. **http://www.globus.org**
8. DataGrid Project home page. **http:// http://eu-datagrid.web.cern.ch/eu-datagrid**
9. Particle Physics Data Grid home page. **http://www.ppdg.net**
10. Storage Resource Management Working Group home page. **http://sdm.lbl.gov/srm-wg**
11. Grid File Transfer Protocol documentation. **http://globus.org/toolkit/docs/3.2/gridftp**
12. Open Science Grid home page. **http://www.opensciencegrid.org**
13. European Grid Initiative home page. **http://www.egi.eu**
14. Enabling Grids for E-Science home page. **http://www.eu-egee.org**
15. Nordugrid home page. **http://www.nordugrid.org**
16. Computing in High Energy and Nuclear Physics (CHEP) conference series:
    a. CHEP 2006. **http://www.tifr.res.in/~chep06/**
    b. CHEP 2007. **http://www.chep2007.com/**
    c. CHEP 2009. **http://www.particle.cz/conferences/chep2009/**
    d. CHEP 2010. **http://event.twgrid.org/chep2010/**
17. International Symposium on Grid Computing (ISGC) series:
    a. ISGC 2008. **http://event.twgrid.org/isgc2008/**
    b. ISGC 2010. **http://event.twgrid.org/isgc2010/**