PROCEEDINGS OF SCIENCE

Computing at LHC experiments in the first year of data taking at 7 TeV

Daniele Bonacorsi*
University of Bologna, Italy
E-mail: daniele.bonacorsi@unibo.it

After many years of development, testing and validation, the computing systems of the four LHC experiments are now in operation. The Grid computing infrastructure has been heavily utilized by ALICE, ATLAS, CMS and LHCb in the first year of LHC proton-proton collisions data taking at 7 TeV - as well as in the first heavy-ion run in 2010 - with remarkable success in all major workflows. The general experience of the four experiments in building and running their computing systems is presented and discussed. Highlights will be given to performances, lessons learned and expected evolutions.

The International Symposium on Grids and Clouds and the Open Grid Forum - ISGC2011,
March 25-30, 2011
Academia Sinica, Taipei, Taiwan

*Speaker.

PoS(ISGC 2011 & OGF 31)106

1. Introduction

Computing for LHC experiments grew up together with Grids. The evolving Grid middleware and the distributed computing system achieved by previous experiments - in which most resources were located away from CERN - have been the background on which LHC experiments started to build their own computing environment. This happened through a huge collaborative effort for years, and massive cross-fertilizations among experiments and Grid developers.

The original design of distributed computing systems was based in 2001 on MONARC [1, 2, 3], which introduced the concept of a hierarchical multi-tiered computing infrastructure with defined functionalities for each Tier level. At the same time, the Worldwide LHC Computing Grid (WLCG) [4, 5] was approved by the CERN Council in 2001. WLCG organized the computing services and infrastructures for the four LHC experiments and collaborated with a number of Grid projects [6, 7, 8] in Europe and in the United States on services and interfaces needed to make all the distributed facilities function as a coherent distributed computing infrastructure based on Grid technologies.

WLCG today comprises one unique Tier-0 center at CERN, 11 Tier-1 centers and >140 Tier-2 centers on 5 continents: the combined resources are able to execute more than 1M jobs/day, having access a total of about 150k CPU cores and >50 PB of disk. Most of the LHC computing models are based on the aforementioned Tier levels, as follows. The Tier-0 (T0) facility at CERN performs prompt data reconstruction, low latency workflows for calibration and detector commissioning, and is responsible for the archival data storage and for the distribution of the data to the Tier-1 sites. The Tier-1 (T1) centers perform the data reprocessing, and are responsible for the custodial storage of real/simulated data as well as of the data serving to the Tier-2 sites. The Tier-2 (T2) centers are the primary resources for analysis for most LHC experiments, and the largest sources of simulated event production for all the experiments. Despite varying by experiment and year, CERN provides roughly 20% of the total LHC processing capacity, while T1 and T2 centers roughly provide 40% each. The computing operations programs of all the experiments have been successful in processing, storing, distributing and analyzing the data samples collected in the first year of LHC data taking at 7 TeV.

The overall experience of the four LHC experiments with their computing systems will be presented and discussed in the following. Highlights will be given on the preparatory work needed to commission the LHC computing systems prior to data taking, as well as on the transition from the preparation activities to the computing operations in a real LHC data taking environment.

2. From commissioning to data taking

The worldwide distribution of the computing Tiers and the quantity and complexity of the deployed services impose a need for a continuous monitoring of the sites availability and reliability. WLCG has been monitoring such quantities for T0, T1 and T2 sites since 2006, and verifying that they match the levels specified in Memorandums of Understanding (MoU) [9]. Looking at the reliability figures over last 5 years (see Fig. 1), an evident improvement is visible for both T1 and T2 sites. The average reliability of the best 8 T1 sites (plus CERN) was about 85% in 2006 and improved to >95% as of today. The average reliability of the top 20% (50%) T2 sites is about 100%

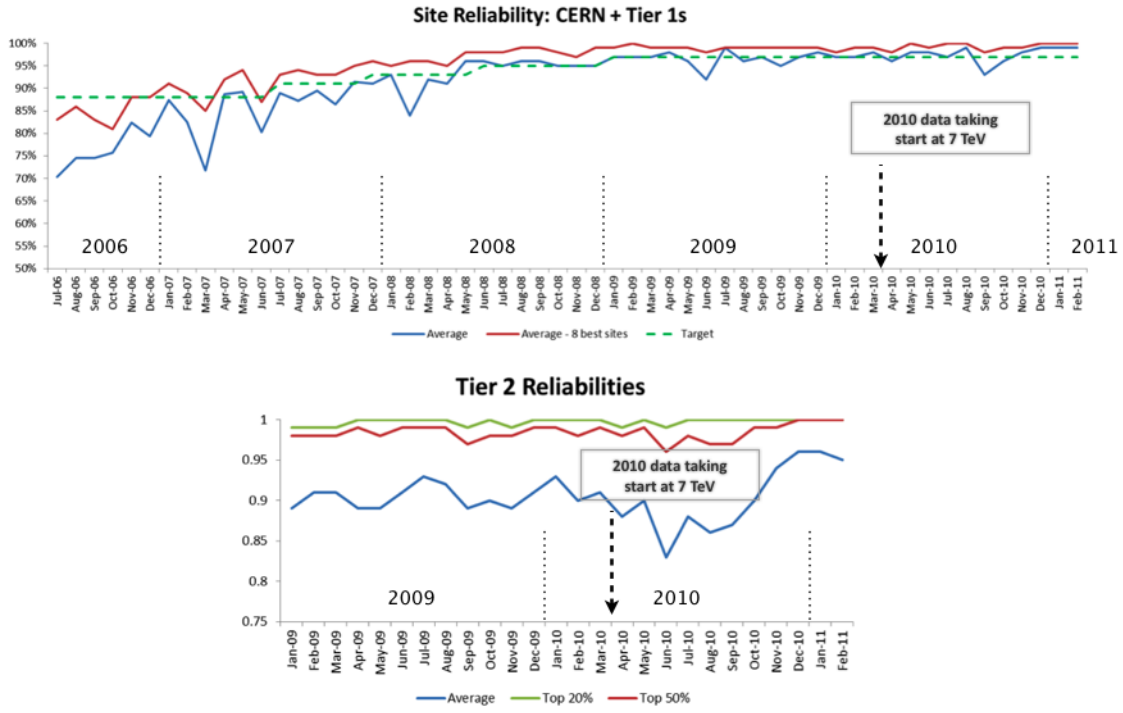

Figure 1: Evolution of WLCG Tiers reliability over time.

(>95%) since early 2009, respectively. More recently, experiment-specific workflows were added to allow a more detailed view of the site performances and hence conclude on their "readiness" to support the activities of the experiments. The continuous monitoring of the site readiness has shown that sites have improved constantly. The number of sites ready for LHC operations stabilized already before the start of the LHC program at 7 TeV, thus yielding a large collection of reliable resources to the LHC physics community even in the first year of data taking.

The smoothness of LHC Computing has to be credited also to a long series of ad-hoc computing exercises at increasing scale since 2004, involving both the experiment communities and the WLCG. In general, their goal was to exercise the computing systems and determine which components were functioning well when operated at the expected scale, and which elements instead needed some more design or development. Despite not all the exercises were equally successful in terms of validating each of the involved components, all of them were extremely beneficial to direct future development plans and commissioning efforts. The community started in 2004 with "*Data Challenges*", i.e. experiment-specific, independent tests in which the full chain of workflows as from the computing models were tested on the Grid. Then, WLCG organized and coordinated a series of four "*Service Challenges*" (SC1, SC2, SC3, SC4) to demonstrate service aspects e.g. sustained data transfers, effective scaling of job submission systems, Grid interoperability, support structures and security systems [10]. Since 2007, given the fact that the experiments had been running Monte Carlo simulations on Grid since years already, and entered in cosmics data taking mode also, the focus moved to the real and continuous production use of the services over several years. From Data/Service challenges on specific topics, the computing exercises became "*Readi-*

ness/Scale challenges" interested to exercise all aspects of the overall service at the same time. This phase culminated in two major worldwide computing exercises: the WLCG Common Computing Readiness Challenge (CCRC) in 2008 [11] and the Scale Test for the Experiment Program (STEP) in 2009 (see e.g. [12] for CMS). They emphasized the simultaneity of many tests and the overlap among experiments to check possible interferences, hence they were crucial as final exercises before the start of 7 TeV data taking.

During the first year of LHC data taking at 7 TeV the community faced unique challenges because of the exponential luminosity increase. During the first several months, a "good" weekend could double or triple the entire dataset, and actually at the very beginning the same was true even for a single "good" fill. It was important to maintain stability in the operations of the computing systems under such rapidly changing conditions, given the fact that a relatively large fraction of the total dataset could have been lost by a significant failure or outage of a single computing component for a single fill. Stably throughout 2010, all LHC experiments have been able to utilize the computing facilities to perform the prompt reconstruction of the events coming out of the detectors and archive them at the T0 as a first archival copy. From CERN the data has been successfully exported to distributed T1 facilities to manage a second archival copy and where enough computing resources for reprocessing were available. The derived data interesting for physics analyses were produced and transferred to T2 centers. Depending on the models, Monte Carlo simulation samples were generated at several Tier levels and uploaded to T1 centers for permanent storage. Some of the major workflows and their performances in 2010 will be briefly discussed in the following paragraphs.

3. Network and data transfers

Substantial commissioning efforts were performed by all four LHC experiments to be able to efficiently transfer data between all the computing Tier levels. The network infrastructure worked very well for the needs of LHC experiments in the recent years and in the first year of data taking at 7 TeV. In particular, the crucial route from CERN to the T1 centers is served by the dedicated LHC-OPN, on which all T0-T1 links have reached a 10 Gbps capacity and back-up channels exist to handle possible temporary service interruptions. Additionally, the LHC-OPN is also serving well the needs of data replication among the T1 sites. In total, the LHC-OPN available bandwidth including all the connections adds to roughly 120 Gbps. Rates of 70 Gbps have been observed across the network during peak periods of synchronization of recently reprocessed data across T1 sites: despite this was driven by reconstructed data replication across T1's by ATLAS, spikes of about 40 Gbps are still reasonably routine under normal experiment activity, and the LHC-OPN has behaved well under such load.

The CERN outbound traffic showed high performance and reliability throughout all data taking in 2010. From historical views of network traffic plots (see Fig. 2) it is clearly visible that the T0-T1 traffic load in 2010 had already been experienced several times in the past over the infrastructure, e.g. during the CCRC'08 and STEP'09 challenges. This demonstrates how the final full-scale service challenges were representative of the first year of LHC running, and the importance of such collaborative efforts done by all LHC experiments together with WLCG during past years.

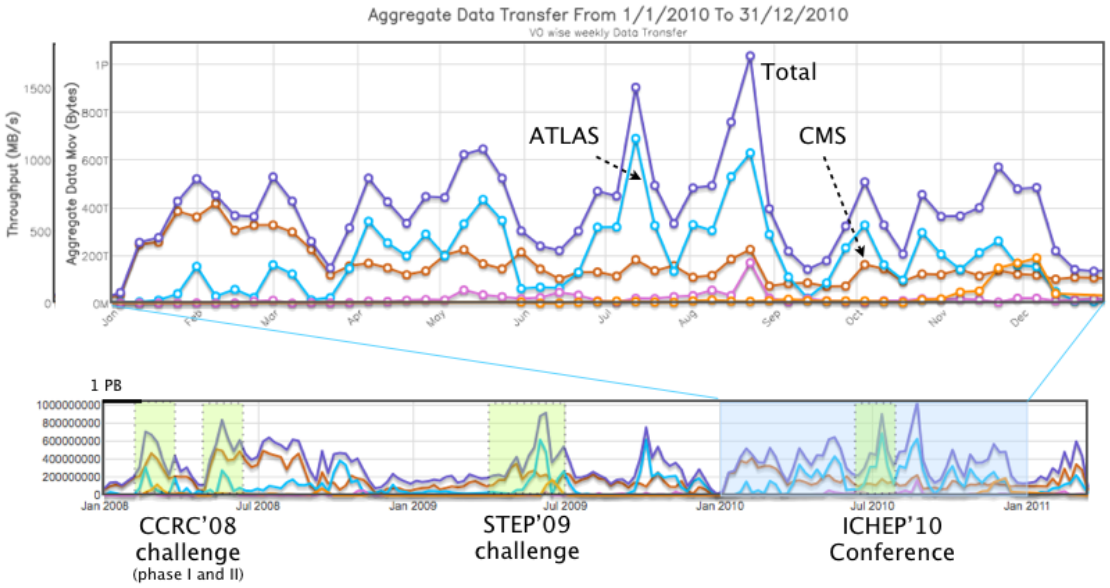

Figure 2: Aggregate volume of CERN-outbound data traffic, in 2010 (top) and since 2008 (bottom).

The ATLAS experiment measured the highest volume of RAW data being distributed from CERN to the T1 sites, and accounted for about 60% of the total in that route. ATLAS experienced a constant data traffic among Tiers also, daily averaging at about 2.3 GB/s per day, with peaks up to about 7 GB/s (corresponding to large volumes of data which need to be distributed after reprocessing campaigns). The overall ATLAS traffic is instead composed of many activities, e.g. T0 exports (including calibration streams), Monte Carlo data transfer in regional clouds, data consolidation (extra-clouds traffic), user subscriptions, and more.

The T0-T1 and T1-T1 traffic is indeed only a part of the overall data transfer load measured in 2010. Once data has been transferred to T1 sites, in most LHC experiments the data is processed to extract derived data that needs to be moved to the T2 level for physics analyses. In doing so, the networking becomes potentially challenging since the data serving is heavier than the data ingesting at the T1 level. WLCG comprises more T2 sites than T1 sites by roughly a factor of 5, thus yielding a T1-T2 matrix with many more possible destinations than sources. Additionally, while the transfer rate from CERN to T1's is controlled by the trigger rate of each experiment, the rate from T1's to T2's is driven by the evolving interest of the analysts in different datasets and data formats, and it is hence partially unpredictable. In terms of data distribution down to the T2 level, for example, ALICE and CMS have computing models foreseeing the distribution of derived data to a high number of T2 centers for analysis. While ALICE adopted a model based on xrootd [13] for data distribution and access, CMS is currently using on-demand data replication techniques on a full mesh of computing centers to minimize the arrival latency of complete data samples. In particular, the CMS experiment conducted a massive commissioning of their data transfer system, which is now in continuous production-mode of operation since 2004. CMS improved by ad-hoc challenges of increasing complexity and by regular computing commissioning activities, and today can sustain up to >200 TB/day of production transfers on the overall topology. The transfers have

been commissioned on the full mesh, i.e. a T2 can get data from any T1, and transfers among T2's (and with Tier-3 centers also) are possible and widely exploited. The transfer rates achieved within a region are as high as several hundreds MB/s; between regions - including transatlantic - about 100 MB/s is becoming quite common.

4. Data reprocessing

Once landed at the T1 level, and safe on custodial storage, LHC data gets reprocessed as needed, to apply new calibrations, to profit of improved software, to derive new data formats. The reprocessing step involves a large number of (usually centrally managed) job submissions exploiting the Grid compute elements and worker nodes, and successful interacting with the Grid storage elements: this workflow has been repeatedly and successfully demonstrated by all the experiments.

The amount of data delivered by LHC in the first year of data-taking at 7 TeV was not as large as expected for a nominal data taking year, so it was possible for all experiment to perform reprocessing campaigns - even over the complete dataset - more quickly and frequently than originally planned. ALICE reconstructed heavy-ion data in November-December 2010 using up to about 8k computing slots. ATLAS performed four reprocessing rounds in 2010, the last one in November 2010 for the full 2010 data/MC sample, peaking at about 16k concurrent reprocessing jobs at T1 sites. CMS went through a dozen of reprocessing passes in 2010, with different fraction of the overall data taken, submitting up to about 6k concurrent jobs at the T1 level, with each facility taking a commensurate share of the RAW data. LHCb worked in a continuous reprocessing mode, reconstructing all existing data when changes were worth it, with a major reprocessing round done in August 2010. Interestingly, the experiments determined for the first time their reprocessing profile during data-taking. CMS demonstrated to be able to reprocess all the p-p data taken in 2010 ($>$1.5G evts) within about 10 days (including resubmission tails) on 7 T1 sites. ATLAS demonstrated to be able to reprocess all data (from the raw data to several derived data types) within within about 7 days on 10 T1 sites, plus some more days to digest the tails.

5. Monte Carlo production

The simulated event production accounts for a large fraction of the global Grid usage. It is a organized and scheduled processing activity, as well as one of the earliest Grid applications, and thus resulted in being very successful on the Grid. Several factors, like realistic simulations, or different pile-up scenarios, will make this a more interesting problem in the future. Simulation production continued in the background all the time since the early utilization of Grid by the LHC experiments, despite with fluctuations caused by a range of causes, including release cycles, sites downtimes, etc. Depending on the experiment, it is done mainly on the T1/T2 level.

ALICE performs Monte Carlo production on all T1/T2 sites: in 2010, ALICE ran on average about 12k simultaneous jobs on the infrastructure (peaking at about 27k) with simulation alone accounting for about 9k. The ATLAS processing is managed centrally on all Tiers resources, organized in regional clouds, with a relatively constant load of up to about 60k ATLAS simulation

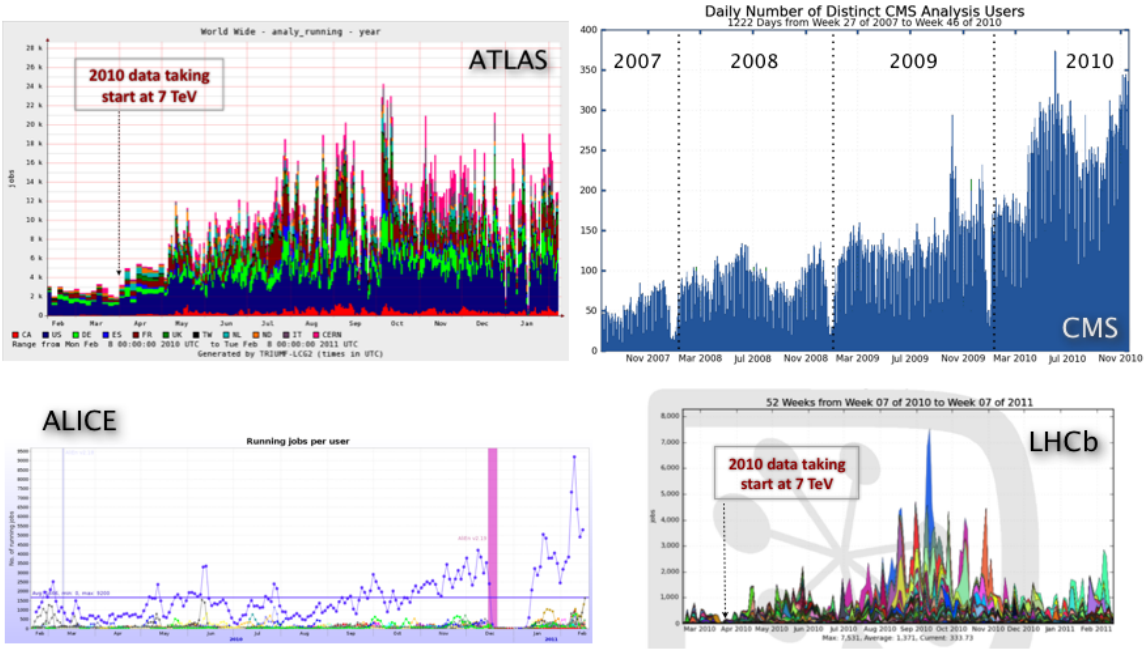

Figure 3: Distributed analysis at LHC (see text for further explanations).

jobs running on the Grid in the second half of 2010. CMS initially performed Monte Carlo production on 50% of T2 resources (plus some opportunistic T3 sites), and only recently expanded to T1 sites as well; only the Tiers that pass the CMS Site Readiness criteria are used, and in 2010 a total of about 3.6 billion of RAW simulated events were produced on the Grid, peaking at >500M evts/month. LHCb processing activities in 2010 consisted of simulation 50%, analysis 29%, reconstruction 21%; simulation is done mainly at T2 level, with more than 100 T2 sites stably used. Also, evolutions in the derived data formats used for physics analyses in several LHC experiments started to impact the Monte Carlo production strategies: e.g. ATLAS produced more simulated ESD's since December 09 to match real data analysis, and CMS is quickly moving away from RECO's and basing most of physics analyses on AOD's.

6. Distributed analysis

While the simulated event production is a scheduled processing activity, the physics analysis is largely unpredictable and chaotic. In this perspective, the ability to transition the complex analysis workflows of all experiments to a common and worldwide distributed infrastructure based on Grid technologies is one of the most remarkable achievements of the experiments in collaboration with the Grid projects.

The LHC Computing activities in the analysis sector are based on considerable investments by each experiment in developing specific tools following a common key paradigm: to shield the user from the structure and complexity of the underlying Grid(s). Each developed framework implement in different ways some instance of this same concept, as can be seen in the design features of *pAthena* and *Ganga* for ATLAS, *Alien* for ALICE, *CRAB* for CMS, *Ganga* and *Dirac* for LHCb. Of

course, custom development in the individual experiment systems is needed to support experiment specific applications and concepts in their data management sector, but the basic functionalities are quite similar, and all such tools succeeded in the task of making a complicated chain of communication look as much as possible like a batch queue submission. In fact, all such tools manage the creation, the submission and the tracking of jobs and the mechanisms to return the results to the users. The success of the Grid analysis systems depends on the capability to properly deal with possible failures of each of the single steps in the chain, e.g. local environment packaging (to make it available at the remote location), choice of site(s) with the desired datasets or resources, submission of individual task sections through grid interfaces, arrival on a batch farm and user/VO authentication on Grid sites, capability to source a proper local environment, discovery functionalities and local data file or remote file opens. Additionally, since the largest fraction of analysis computing at LHC is at the T2 level, a special effort is spent by most experiments in the data placement optimizations and in data transfer operations. As an example, CMS launched an ad-hoc program in 2010 to commission all the T2-T2 transfer links in a full mesh model, at a rate of up to 30 links commissioned per day, about 7 links/day over the first 6 months of data taking. This resulted in a large utilization of such links for production transfers (several hundreds of TB/month transferred among T2's already in the second half of 2010), and the exploitation of such flexibility in the transfer operations allowed to ultimately reduce the latency seen by the analysis end-users. For most experiments there are still wide margins of improvement in the distributed analysis sector, at several levels like efficiency of completion, CPU efficiency, user experience, job status tracking, monitoring and accounting, debugging and troubleshooting models, etc. Nevertheless, the LHC experiments successfully performed analysis on Grid over the LHC data taken in 2010.

It is impressive to observe the level of adoption of the analysis frameworks of the experiments (see Fig. 3). Some detailed figures are given below per each experiment. The collaborations are of course much larger than the number of submitters of analysis jobs, but it's remarkable to note that the vast majority of active analysis users are performing their own analyses successfully using Grid resources, and that the analysis submissions are a sizeable fraction of the total jobs on the Grid.

ALICE experienced about 1.7k concurrent user jobs in 2010 on average, for a total of >9M user jobs completing over last 12 months, and about 200 distinct users on average (and increasing). An interesting analysis train model is also adopted, in which instead of only standard user jobs (CPU efficiency lower than simulations or reconstruction, variable job duration, many failures, far-from-perfect code, chaotic job submissions) - "analysis trains" are preferred (optimized I/O - read once and do many tasks, streamlined code as much as possible, managed and scheduled submissions). ATLAS experiences an evident increase in the analysis load after the start of 2010 data taking, with a roughly stable load since then, recording peaks at about 20k concurrent user jobs, with dips due to holidays - as well as peaks before major conferences - clearly visible. ATLAS measured more than 1000 active users per month submitting analysis jobs on the Grid. CMS measured a constant increase in the number of distributed analysis users, up to 2010 figures: about 300-350 distinct daily users, up to >500 users per week during peaks, >800 individuals per month submitting analysis jobs on the Grid. The CMS analysis is done at the T2 level only, and the resources utilization was measured to be as high as up to about 12k jobs slot used per week. CMS records approximately about 100k Grid submissions per day from analysis, which is roughly what

is predicted as from the computing model. In the LHCb distributed analysis there is no a-priori assignment of site, the share being done by availability of resources and data. Only about 2% of the analysis is performed at the T2 level (toy MC, private small simulations, etc), the vast majority is done at CERN and at the T1 sites. LHCb recorded up to about 320 unique analysis users on the Grid.

7. Conclusions

The overall experience of the four LHC experiments with their Grid Computing systems in the first year of LHC data taking at 7 TeV was a success. After a long period of dedicated tests and commissioning efforts, the computing operations programs of all the experiments have been successful in processing, storing, distributing and analyzing the data samples collected at the LHC. A close, constant and fruitful collaboration with WLCG allowed to achieve the challenging goal of deploying a distributed computing infrastructure in operations while still commissioning the LHC detectors. Grid Computing is serving well the LHC physics program, and the user activity level and enthusiasm are high. The data volume collected so far has been relatively small with respect to the original planning: a more resource-constrained environment is expected in 2011 and beyond.

Acknowledgements

The author would like to credit and thank the LHC accelerator division for such a fruitful 2010. Thanks to the LHC Computing teams (managers, team coordinators, developers and operators) for their hard work and collaboration; the Grid developers, the entire WLCG community and all the site admins at the distributed Tiers for their committed, competent and constant work.

References

[1] M. Aderholz et al., "*Models of Networked Analysis at Regional Centres for LHC Experiments (MONARC), Phase 2 Report*", CERN/LCB 2000-001 (2000)

[2] MONARC: *http://monarc.web.cern.ch/MONARC/*

[3] S. Bethke et al., "*Report of the Steering Group of the LHC Computing Review*", CERN/LHCC 2001-004 (2001)

[4] J. D. Shiers, "*The Worldwide LHC Computing Grid (worldwide LCG)*", Computer Physics Communications 177 (2007) 219–223, CERN, Switzerland

[5] WLCG: `http://lcg.web.cern.ch/lcg/`

[6] EGEE: `www.eu-egee.org/`; EGI: `www.egi.eu/`

[7] OSG: `www.opensciencegrid.org/`

[8] NORDUGRID: `http://www.nordugrid.org/`

[9] "*Memorandum of Understanding for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid*", CERN-C-RRB-2005-01/Rev. 20 April 2009

[10] D. Bonacorsi, "*WLCG Service Challenges and Tiered architecture in the LHC era*", IFAE, Pavia, 2006

[11] J.D. Shiers et al., "*The (WLCG) Common Computing Readiness Challenge(s) - CCRC'08*", contribution N29-2, Grid Computing session - Nuclear Science Symposium, IEEE (Dresden), October 2008

[12] D.Bonacorsi and O. Gutsche, "*CMS from STEP'09 to Data Taking: CMS Computing experiences from the WLCG STEP'09 challenge to the First Data Taking of the LHC era*", ISGC'10, Taipei, Taiwan, 2010

[13] Xrootd: `http://xrootd.slac.stanford.edu`