PROCEEDINGS OF SCIENCE

Current status of the WLCG data management system, the experience from the three years of data taking and future role of Grids for the LHC data processing

Dagmar Adamova[*]
Nuclear Physics Institute, Czech Academy of Sciences
E-mail: adamova@ujf.cas.cz

Since the start-up of the data taking in November 2009, the Large Hadron Collider (LHC) at CERN has been successfully delivering data from the proton-proton (pp), lead-lead (PbPb) and proton-lead (pPb) collisions at the energies ranging from 900 GeV to 8 TeV. The LHC experiments were continuously developing and improving their systems for the data recording and processing and as a result, these systems currently enable delivering large numbers of scientific papers within a couple of weeks after the data was recorded. The current phase of the LHC operation and data taking will end by a pPb run in February 2013, followed by the Long Shutdown 1 (LS1) intended for the LHC upgrade boosting the LHC performance to deliver luminosity of about $2 \cdot 10^{34}$ cm^2 s^{-1} and ultimate intensity of $1.7 \cdot 10^{11}$ protons per bunch with 25 ns spacing. To cope with the upcoming new conditions, upgrades of the LHC experiments will be performed, both of the detectors and the online and offline data processing systems. In this contribution, we will focus on the distributed data management infrastructure used by the LHC experiments, the Worldwide LHC Computing Grid (WLCG). We will present the current status of WLCG and the experience from the three years of the LHC data taking. While the system was successfully providing the environment for the reliable data processing and a fast delivery of scientific results, the capacity of available computing resources is becoming tight. Also, the experiments are facing the necessity to transform their software for simulation, reconstruction and analysis of data as well as for the distributed data management to comply with the new technologies including the multicore computer processors and computational clouds. The corresponding strategies adopted by the LHC experiments will be discussed.

51st International Winter Meeting on Nuclear Physics
21-25 January 2013
Bormio (Italy)

[*] Speaker.

1. Introduction

Thursday February 14th 2013 marked the end of the first operation period (Run1) of the CERN Large Hadron Collider (LHC) [1], which started in November 2009 and was enormously successful. During over 3 years, the LHC was delivering data from the proton-proton (pp), lead-lead (PbPb) and proton-lead (pPb) collisions at the energies ranging from 900 GeV to 8 TeV. In the time of writing this article, the LHC is going through the Long Shutdown 1 (LS1) intended for the LHC upgrade.

The LHC is one of the scientific projects producing very large volumes of data, of the order of tens PetaBytes (PB) per year. Therefore the development of the LHC project went along with the development of the project Worldwide LHC Computing Grid (WLCG) [2] designed to provide the environment and technical infrastructure for storage, distribution, processing and enabling access to the LHC data. As a result, national computing centers of the institutes participating in the four large experiments at the LHC: ALICE[3], ATLAS[4], CMS[5] and LHCb[6], were integrated into one world-wide system providing the required services.

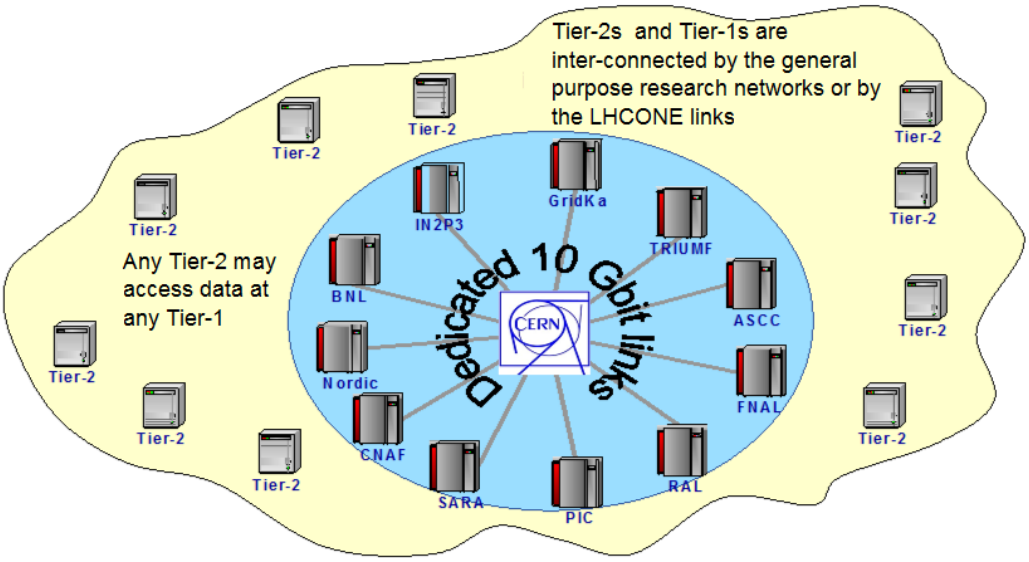
In this article, we will present the current status of WLCG and the experience from the three years of the LHC data taking (Run1). After almost 10 years of development, preparations and stress testing, the WLCG infrastructure worked perfectly during the LHC operations, although the volumes of produced data were in excess of the expected 15 PB per year. However in 2012/2013, the computing resources became tight and urgent became the problem of losing some Physics results due to the shortage of computing resources. We will present the strategy followed by the WLCG to meet the requests of the LHC experiments for the time after LS1 (Run2) and beyond, when the volumes of produced data will get larger due to the boosted performance of the LHC.

2. WLCG: the structure

WLCG represents a federation of heterogeneous national computing sites, enabled by fast networks, high throughput computing (HTC) technologies and a middleware layer. It provides a single sign-on and delegation of access rights through common interfaces for basic services. The integrated computing sites are classified according their size and level of services as centers of the level Tier-0, Tier-1 or Tier-2 (figure 1).

WLCG is the world's largest computing grid. It is based on two main grids: the European Grid Infrastructure [7] in Europe, and Open Science Grid [8] in the US, but has many associated regional and national grids (such as TWGrid [9] in Taiwan and EU-IndiaGrid [10]), which supports grid infrastructures across Europe and Asia.

The primary/raw data from the LHC experiments/detectors is first processed online by the Trigger (HLT) and Data Acquisition (DAQ) systems and then migrated to the CERN Data Center which is the Tier-0 of the WLCG. CERN is responsible for the safe keeping of the raw data and performs the first pass of the raw data reconstruction. It also distributes the raw data and the reconstructed output to Tier-1s. In the time of writing this article, an extension of the CERN Tier-0 is under construction at the Wigner Institute in Budapest, Hungary.



Tier-2s and Tier-1s are inter-connected by the general purpose research networks or by the LHCONE links

Any Tier-2 may access data at any Tier-1

Figure 1: Tier-0, Tier-1, Tier-2 layout and interconnectivity. Tier-0 is CERN and 11 Tier-1s are represented inside the turquoise circle. The LHCONE network project is described in [11].

The raw data is replicated at one of 11 Tier-1 centers which keep a proportional share of raw, reconstructed and simulated data. They usually perform large-scale reprocessing and Monte Carlo simulation campaigns.

And finally there are about 140 Tier-2 sites designed mainly for Monte Carlo simulations and end user ("chaotic") analysis. Tier-2 centers are very important: they provide about half of the total computing and storage resources aggregated within WLCG.

3. WLCG: the current status

The distribution of computing sites integrated in the WLCG over the world is shown on the map in figure 2 (see also [12]).

In the time of writing this article, the single Tier-0 of the system is the CERN Data Center, with the Budapest extension under construction, as already mentioned. The center will be connected to CERN by two links of 100 Gigabit(Gb)/s throughput and this way the new Tier-0 will show up as a remote part of the CERN Data Center.

There are 11 external Tier-1 centers: Amsterdam/NIKHEF-SARA (the Netherlands), Bologna/CNAF (Italy), NDGF (Nordic European countries), UK-RAL (United Kingdom), Lyon/CCIN2P3 (France), Barcelona/PIC (Spain), DE-FZK (Germany), FNAL (USA), BNL (USA), TRIUMF (Canada), Taipei/ASGC (Taiwan). Then there are 3 centers in preparation to become additional official Tier-1s: KISTI-GSDC in South Korea, JINR and RRC-KI in Russia. The system is completed with 68 Tier 2 federations representing about 140 computing sites. All the WLCG participants are bound by the Memorandum of Understanding signed by 54 signatories, representing 36 countries.

WLCG by the numbers:

- More than 8000 physicists use it

	Design	2010	2011	2012
Beam energy [TeV]	7	3.5	3.5	4
Bunches/Beam	2835	368	1380	1380
Proton/Bunch [10^{11}]	1.15	1.3	1.5	1.5
Peak Luminosity [10^{32} cm^{-2} s^{-1}]	100	2	30	~ 80
Integrated Luminosity [fb^{-1}]	100/year	0.036	6	~ 25
Pile-up	23	~ 1	20	> 30

Table 1: LHC performance during Run1

- Over 300000 available processor cores (CPU)
- Over 170 PB of disk space
- 10 Gb/s optical fiber links connect CERN to each Tier-1

4. WLCG: the performance during Run1

The LHC running conditions over the Run1 period (see Table 1) put higher demands on the data processing systems than originally anticipated. The continuous upgrade of delivered luminosities combined with 50 ns bunch spacing resulted in the pile-up (number of pp interaction events per bunch crossing) in excess of 30 which was out of warranty of the experiments' online systems and produced more than 15 PB of LHC data in one year, as originally anticipated. In 2011, about 22 PB of LHC data were recorded and in 2012 close to 30 PB. The events taken during 2011/2012 were very complex, as displayed for CMS and ALICE in figures 3 and 4.

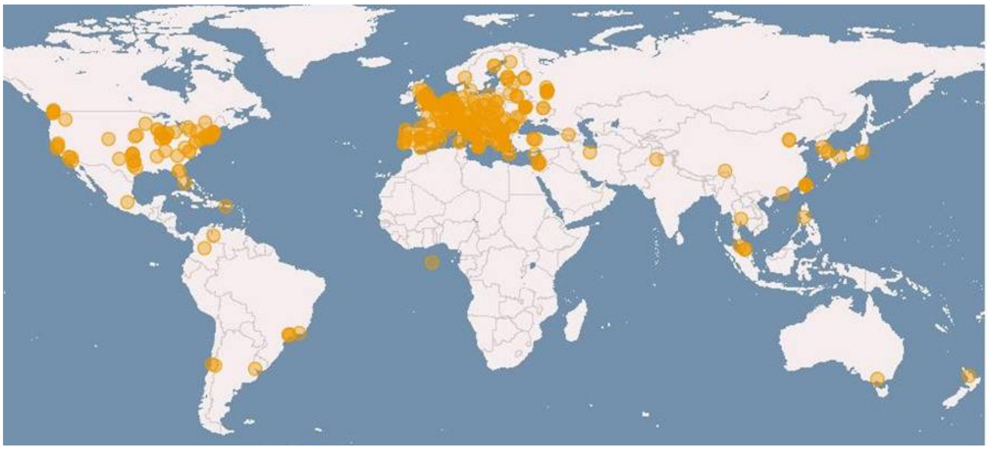


Figure 2: Geographical distribution of WLCG sites.

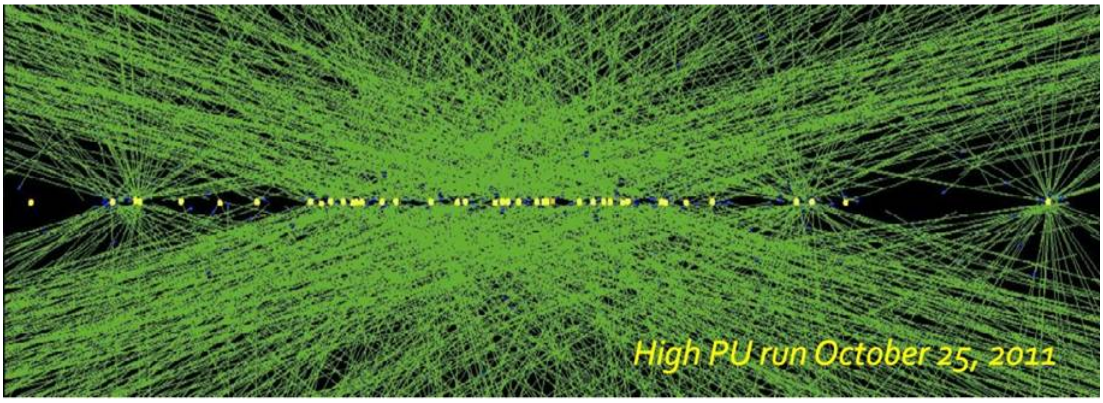


Figure 3: Events pile-up reconstructed by CMS in October 2011.

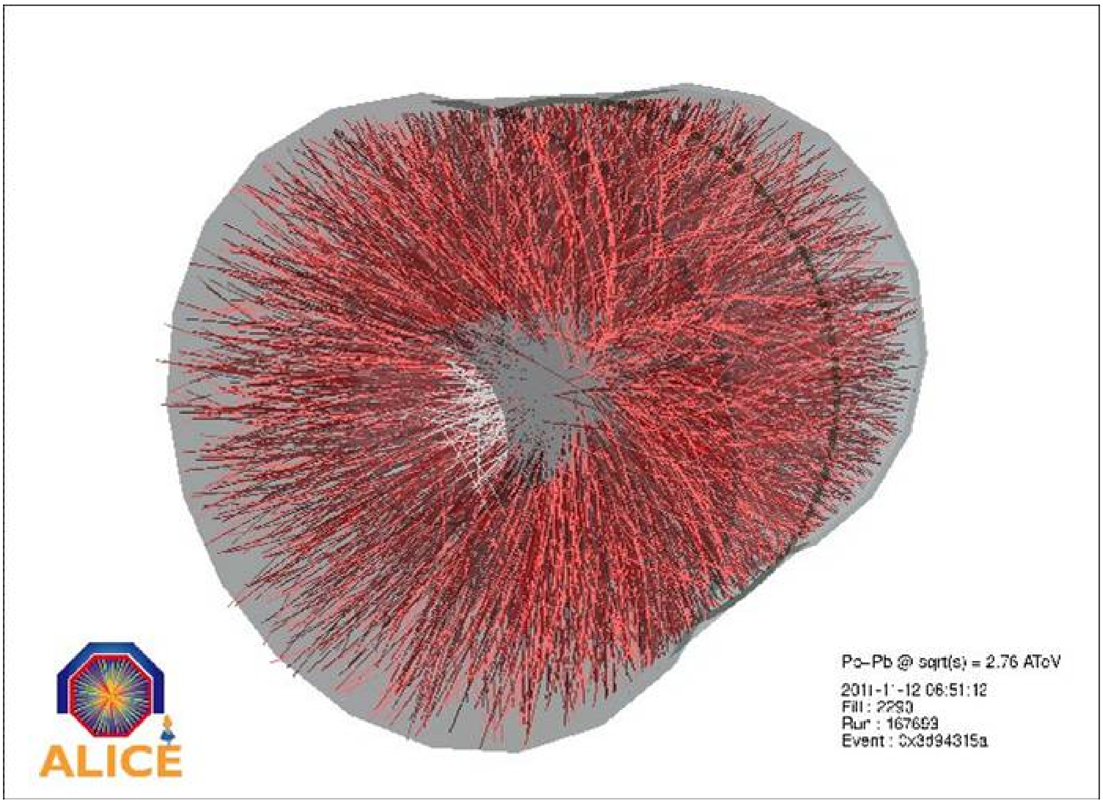


Figure 4: Pb-Pb collision event recorded by ALICE in November 2011.

WLCG resources and services were able to handle all this data without problems (a recognized success), although a part of 2012 data was "parked" and will be processed during the LS1. Some plots showing the performance of different WLCG services during Run1 are shown in figures 5–10. A detailed information on CASTOR, the CERN Advanced Storage manager, is given in [13].

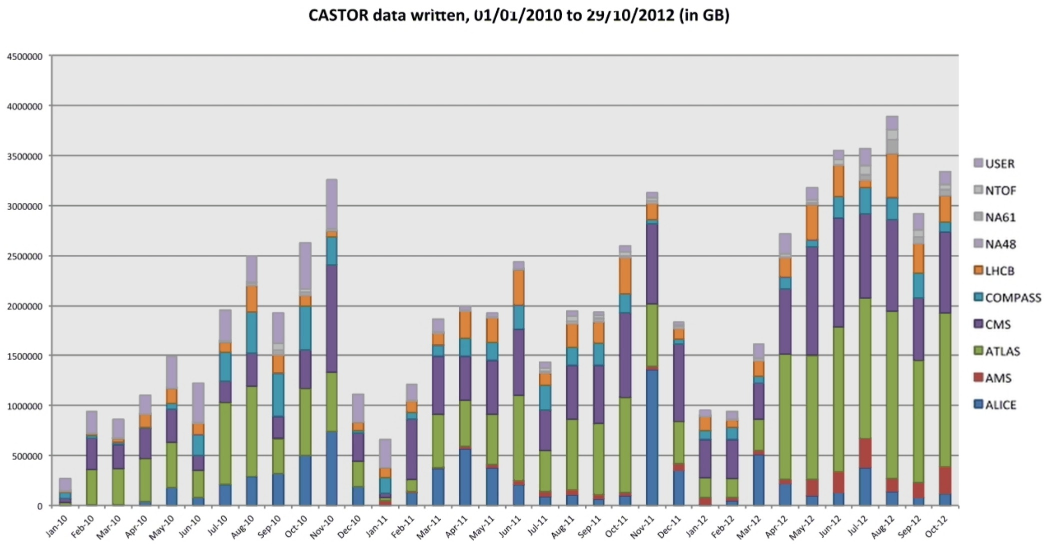


Figure 5: Data written to CERN CASTOR during 2010-2012: In 2012 written $\sim$ 30 PB (LHC data); Close to 3.5 PB written per month with the LHC running.

5. Conditions during Run2 and beyond

The upcoming energy and luminosity upgrade of the LHC (to luminosity $2 \cdot 10^{34}$ cm^{-2} s^{-1} and bunch intensity of $1.7 \cdot 10^{11}$ p/bunch with 25 ns spacing) will produce yet higher pile-up and more complex events. The upgrade is necessary to clarify aspects concerning the nature of the Higgs boson and of electro-weak symmetry breaking, as well as to perform high-precision measurements in heavy-ion collisions:

- is the Higgs fundamental or composite?
- how many doublets? singlets? charged Higgs bosons?

It is needed to measure, as accurately as possible:

- Higgs couplings to fermions, gauge bosons and self-couplings
- Rare decay modes, possible Flavor Changing Neutral Current
- WW scattering at high energy and gauge boson self-couplings
- Charm and Beauty production in heavy-ion collisions

These measurements need huge volume of data to be analyzed. For instance, for a Higgs decay measurement $H \rightarrow Z\,\gamma$ with 3.5 σ one needs 600 fb^{-1} and with 11 σ one needs 6000 fb^{-1}.

6. WLCG and LHC experiments strategy for the future

The ever growing volumes of data anticipated after the LHC upgrade require changes in the computing models of the LHC experiments, adoption of new hardware and software technologies and upscale of the existing WLCG resources.

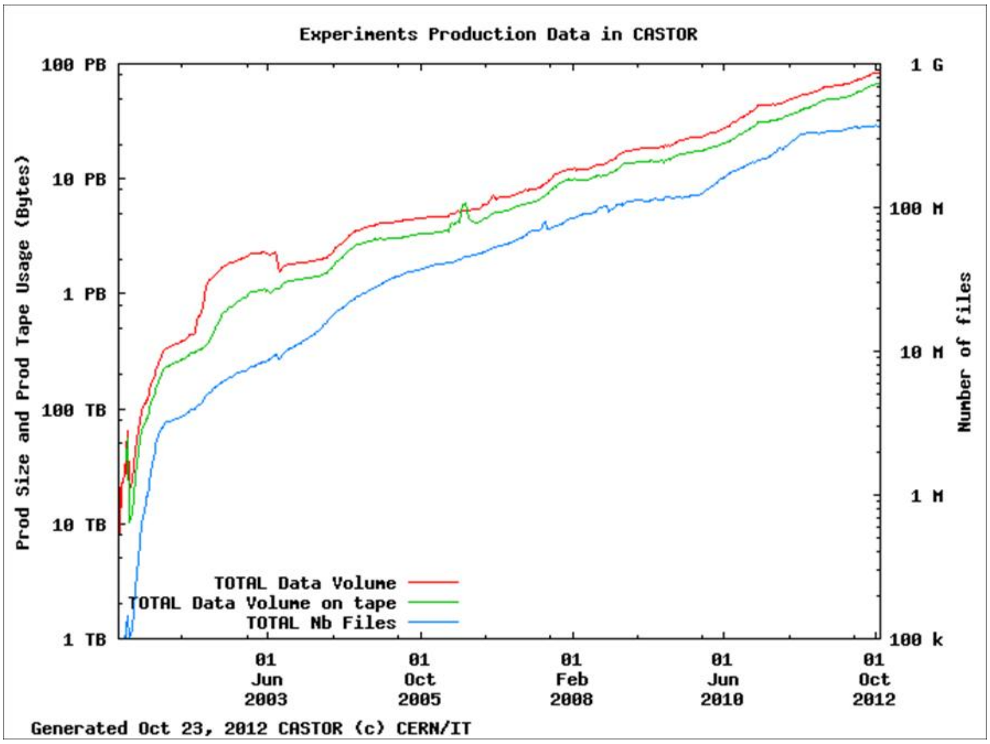


Figure 6: CERN CASTOR: close to 100 PB archive in October 2012. Increases by a rate ∼ 1 PB/week with the LHC on.

One of the new features in the data handling strategy is a maximum possible online data size reduction. Detectors' throughput in the new conditions can reach up to Terabytes/s and this must get reduced in the online DAQ/HLT farms by data compression including reconstruction (tracking) so that the output rate to the on-site storage be reduced to less than 100 GB/s and the transfer rate to the CERN Data Center to about 15 GB/s. The same computing farms (same hardware) will be used for the online and offline data processing: during the technical stops, the DAQ/HLT farms will operate as Tier-1 centers [14].

The choice of optimal hardware architecture to be used after the LS1 and beyond is still uncertain: CPU, GPU or FPGA? High performance or low power? Multicore or singlecore processors? The acquisition, ownership, development and maintenance costs must also be assessed. Most likely outcome will be a hybrid solution.

In any case, the computing models must elaborate strategies to adapt to new technologies and architectures, to improve efficiency etc. Unlike in the original computing models, the new ones emphasize the need to share as many common solutions as possible.

Under investigation and testing is also the possibility to find additional resources using public and private computing clouds [15]. The key technology – virtualization – is already widely used at the WLCG sites. But new standard interfaces to the existing clouds are necessary. The costs of using commercial clouds' services for the processing of LHC data are currently too high. Still, there are ongoing tests of using clouds as a replacement for the Computing Element (CE, [16]), for a batch-less interaction with the computing resources [17].

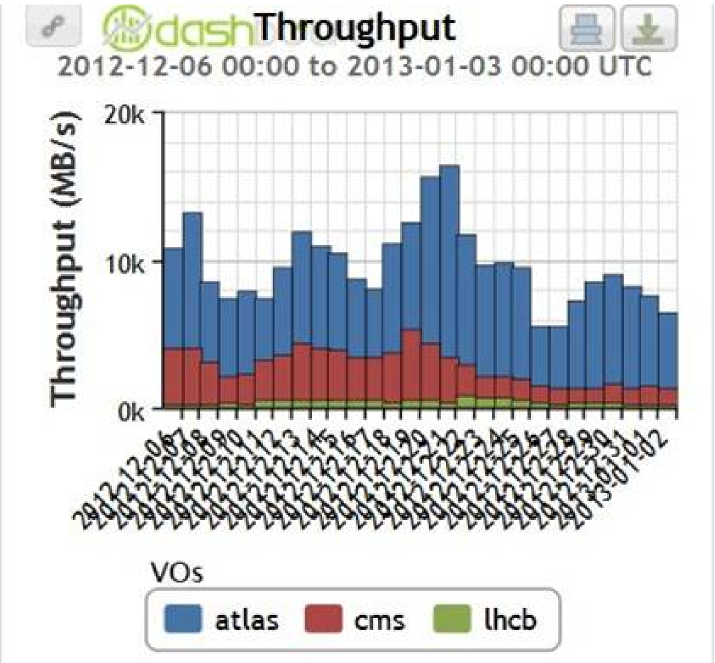


Figure 7: Data export from CERN to the outside WLCG centers in 2012: global rate > 15 GB/s.

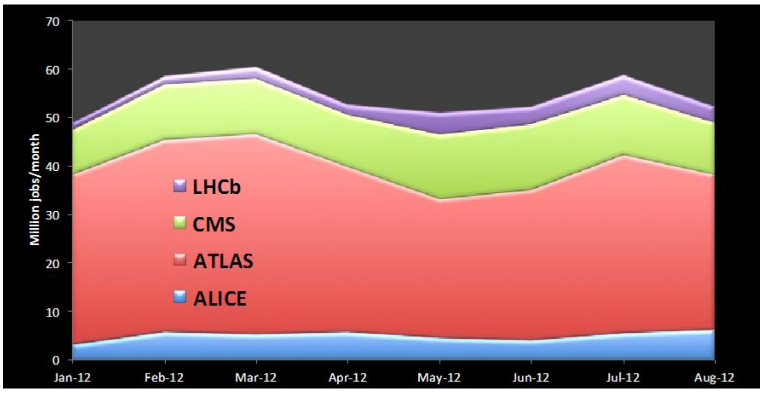


Figure 8: CPU workload in 2012: 2 million jobs/day (vertical scale in million jobs/month).

7. Summary and conclusions

After about 10 years of building, development and stress-testing, the WLCG operations and performance during the real data taking were a recognized success. Although the collider produced significantly more data than originally anticipated, the WLCG infrastructure was able to handle all this data without problems and enabled fast delivery of scientific results. Thus, the number of CERN EP preprints with LHC results submitted during 2012 reached 352.

However, towards the end of Run1 the WLCG resources were exploited up to the edge. It is evident that new strategies in handling the LHC data must be adopted because the simple accumulation of new additional hardware resources can not continue forever.

As we mentioned, the LHC experiments are designing updated computing models to be able

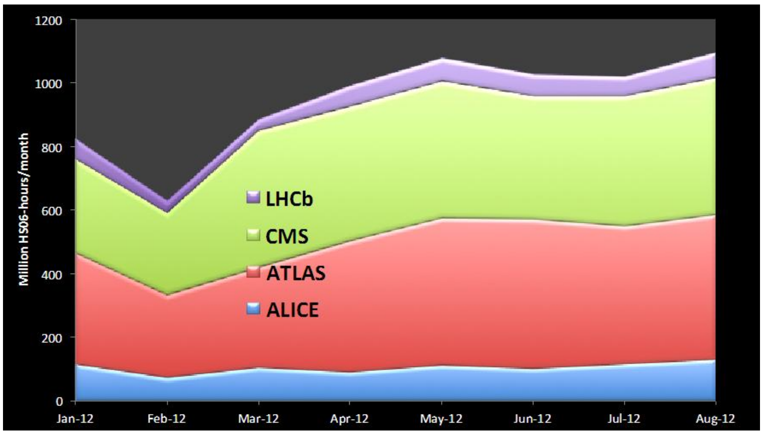


Figure 9: CPU workload in 2012: 10^9 HepSpec06-hours/month (vertical scale in million HS06-hours/month). Corresponds to > 200 thousands CPUs in continuous use. The unit of CPU performance HepSpec06 is defined in [18].

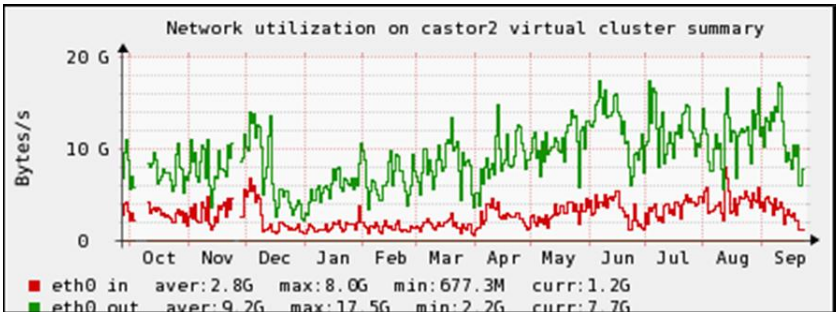


Figure 10: Data rates in CERN CASTOR in 2011/2012: input-max > 3.5 GB/s and output-max > 15 GB/s.

to deliver Physics results using the existing hardware and software resources. As also mentioned, it will be necessary to adapt the existing experiments' software and WLCG middleware to the new technologies like e.g. multicores. Also, the online raw data compression is very important.

And evidently, it would be unwise not to try to find a way how to use the ever growing resources of commercial computing clouds. An example of the initiative in this direction is the European project Helix Nebula - the Science Cloud: Big science teams up with big business [16]. This project will support the massive IT requirements of European scientists and will enable the development and exploitation of a cloud computing infrastructure, initially based on the needs of European IT-intense scientific research organizations. It will also allow the inclusion of other stakeholders' needs (governments, businesses and citizens). The participating scientific project from the High Energy Physics community is the CERN ATLAS Collaboration. Currently, this project is in the phase 'Proof of Concept'.

The work on all the mentioned strategies and tasks was ongoing already during Run1 but started fully with the beginning of LS1. The outcome should partly be completed towards the start of Run2 and finalized towards the end of LS2. It is essential to guarantee/ensure that the Physics potential of the LHC will not be limited by the shortage or inability to fully utilize the available computing resources.

8. Acknowledgements

This work was supported by the MSMT CR contracts No. LG 13031, CZ.1.07/2.3.00/20.0207 and LC 07048.

References

[1] The Large Hadron Collider at CERN; `http://lhc.web.cern.ch/lhc/`

[2] Worldwide LHC Computing Grid: `http://wlcg.web.cern.ch/`

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

[4] ATLAS Collaboration: `http://atlas.ch/`

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

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

[7] European Grid Infrastructure: `http://www.egi.eu/`

[8] Open Science Grid: `https://www.opensciencegrid.org/bin/view`

[9] TWGrid: `http://www.twgrid.org/en/index.php`

[10] EU-IndiaGrid: `http://www.euindiagrid.eu/`

[11] LHCONE: `http://lhcone.net/`

[12] `http://gstat2.grid.sinica.edu.tw/gstat/geo`

[13] CASTOR: `http://castor.web.cern.ch/`

[14] P. Vande Vyvre: *ALICE Online upgrade*; `https://indico.cern.ch/getFile.py/access?contribId=59&sessionId=1&resId=1&materialId=slides&confId=209628`

[15] 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`

[16] Computing Element, `http://glite.cern.ch/lcg-CE/`

[17] M. Jouvin: *Cloud pre-GDB Summary*; `http://indico.cern.ch/getFile.py/access?contribId=2&resId=0&materialId=slides&confId=197801`

[18] HepSpec06: `https://wiki.egi.eu/wiki/FAQ_HEP_SPEC06`

[19] Helix Nebula: Big science teams up with big business; `http://helix-nebula.eu/`