PROCEEDINGS OF SCIENCE

Experience of the WLCG data management system from the first two years of the LHC data taking

Dagmar Adamova[1]
Nuclear Physics Institute, Czech Academy of Sciences
Rez near Prague, CZ 25068, Czech Republic
E-mail: adamova@ujf.cas.cz

The High Energy Physics is one of the research areas where the accomplishment of scientific results is inconceivable without the distributed computing. The experiments at the Large Hadron Collider (LHC) at CERN are facing the challenge to record, process and give access to tens of PetaBytes (PB) of data produced during the proton-proton and heavy ion collisions at the LHC (15PetaBytes/year of raw data alone). To accomplish this task and to enable early delivery of physics results, the LHC experiments are using the Worldwide LHC Computing Grid (WLCG) infrastructure of distributed computational and storage facilities provided by more than 140 centers. Since the first centers joined WLCG in 2002, the infrastructure had been gradually built up, upgraded and stress-tested. As a result, when the LHC started delivering beams in late 2009, the WLCG was fully ready and capable to store, process and allow the physicists to analyze this data.
The architecture of WLCG follows a hierarchical system of sites classified according to a "Tier" taxonomy. There is 1 Tier-0 (CERN), 11 Tier-1s and about 130 Tier-2s spread over 5 continents.
We will briefly summarize the experience and performance of the WLCG distributed data management system during the first two years of data taking. We will demonstrate the irreplaceable role of the WLCG Tier-2 sites as concerns providing the compute and storage resources and operation services. As an example, we will present the contribution of the WLCG Tier-2 site in Prague, Czech Republic to the overall WLCG Tier-2s operations.

50th International Winter Meeting on Nuclear Physics,
Bormio, Italy
23-27 January 2012

[1] Speaker

PoS(Bormio2012)014

1. Introduction

The start-up of the Large Hadron Collider (LHC) at CERN [1], the world's most powerful particle accelerator, in November 2009 has opened a new era in the High Energy Physics. Ever since the first collisions, the LHC has been performing unexpectedly well for a new machine, gathering data at an astonishing rate. The four large LHC experiments – ALICE [2], ATLAS [3], CMS [4] and LHCb [5], have been facing the challenge to record, process and give access to tens of PetaBytes (1PB=1 million GB) of data produced during the proton-proton and heavy ion collisions in the LHC. The production and analysis environments for the LHC experiments is provided by the distributed computing infrastructure managed and operated by a worldwide collaboration/project, the Worldwide LHC Computing Grid (WLCG) [6].
The computational Grid is the only way that the masses of data produced by the collider can be processed. The WLCG is a technological leap like the collider itself and without it the project would quickly drown in its own data. Without WLCG, the LHC would be an elaborate performance art project on the French Swiss border [7].
Since the first centers joined WLCG in 2002, the infrastructure had been gradually built up, upgraded and stress-tested. As a result, when the LHC started to deliver beams, "from the service point of view, we didn't even notice that the data had started to flow, everything worked as expected" (Ian Bird, WLCG project leader).
In this paper, we will briefly summarize the experience and performance of the WLCG distributed data management system during the first two years of data taking. One of the basic features of the WLCG workflow is a great importance of the Tier-2 [6] sites as concerns providing both the compute and storage resources and operation services. As an example, we will present one of the Tier-2 sites, the regional computing center Golias [8] in Prague, Czech Republic, and its contribution to the overall WLCG Tier-2s operations.

2. Collisions in the LHC

To arrange for collisions in the LHC, protons or lead ions are injected into the accelerator in bunches, in counter-rotating beams. In the case of the proton beams, each bunch contains 10^{11} protons. The bunch crossing rate is 40 MHz and the proton collisions rate is 10^{7} - 10^{9} Hz. However, the new phenomena looked for by the scientists, like the search for the Higgs boson [9] or measurements of the CP Violation [9], appear at a rate of ~10^{-5} Hz. So the physicists must check 10^{13} collision events to have a chance to discover a New Physics phenomenon.
Although most of the collision events are not interesting for Physics and therefore the Physics selection must be performed early on, which results in most of the collision events being thrown away, still the volume of the raw (primary, unprocessed) data produced by the LHC collisions and stored for further Physics analysis represents about 15 PetaBytes in one year. This estimate was mentioned already in the original proposals of the LCG. A little surprisingly, exactly this volume of raw data was recorded during the real collisions in 2010; in 2011 it was ~20 PB of raw data. This translates into the number of the processor cores, CPUs, needed to process this amount of data of about 200 thousands.

3. LHC Computing Grid: what and why?

Data pouring out of the LHC detectors is shipped through various stages of processing (see Figure 1). The data must be moving continuously since no buffer is big enough to keep it for

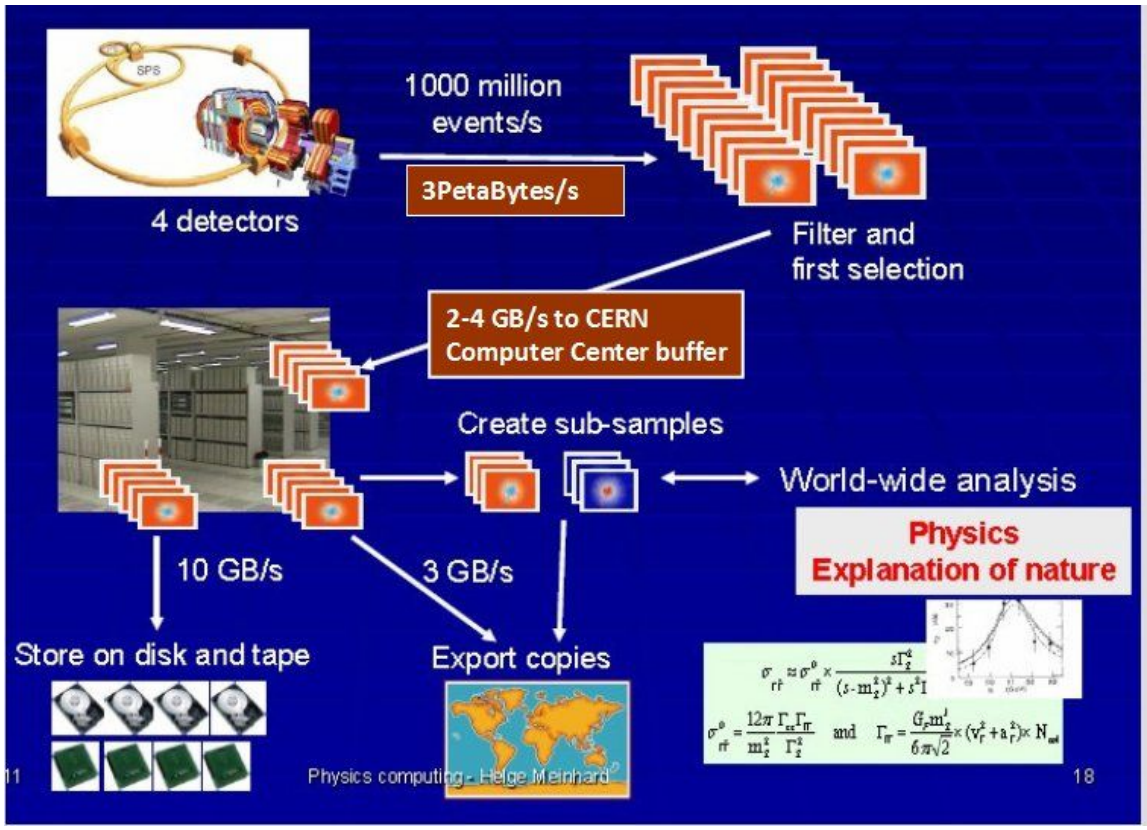

Figure 1. The LHC data flow

long. Any interruption in the data flow creates serious problems. Data must be available within a very short time. There are about 15 - 20 PB of raw data produced in one year. This raw data is processed several times. First there are multiple passes of reconstructions accompanied by Quality Assurance processing. The output files from the reconstruction passes are then further skimmed to keep only the information truly relevant for the final analysis, which in the end produces Physics results. Each cycle of the raw data production is also accompanied by a Monte Carlo simulation campaign which produces the volume of data comparable with the raw data volume itself.

It is impossible to provide the production and analysis environment for data volumes of this size within one computing center. Instead, this framework is delivered by the WLCG Collaboration in a distributed form.

The concept of a computational Grid which would provide the needed computing and storage resources to handle the data produced by the LHC experiments was conceived and approved by the CERN Council 10 years ago. The WLCG has integrated all the resources provided by individual sites into one system using the latest technologies in Grid middleware [10] and networking. The result is an amazingly complex infrastructure featuring about 250 thousands of CPUs and 150 PB of disks provided by more than 140 computing centers spread over 5 continents (see Figure 2.).

The WLCG has a hierarchical structure based on the data processing paradigm. The individual centers are ranked as Tier-0, Tier-1 or Tier-2 according to their hardware resources and level of services. Tier-0 center is CERN and by now there is just one Tier-0 in WLCG. Then there are 11 centers of the level Tier-1, which are very large computing centers providing in addition to

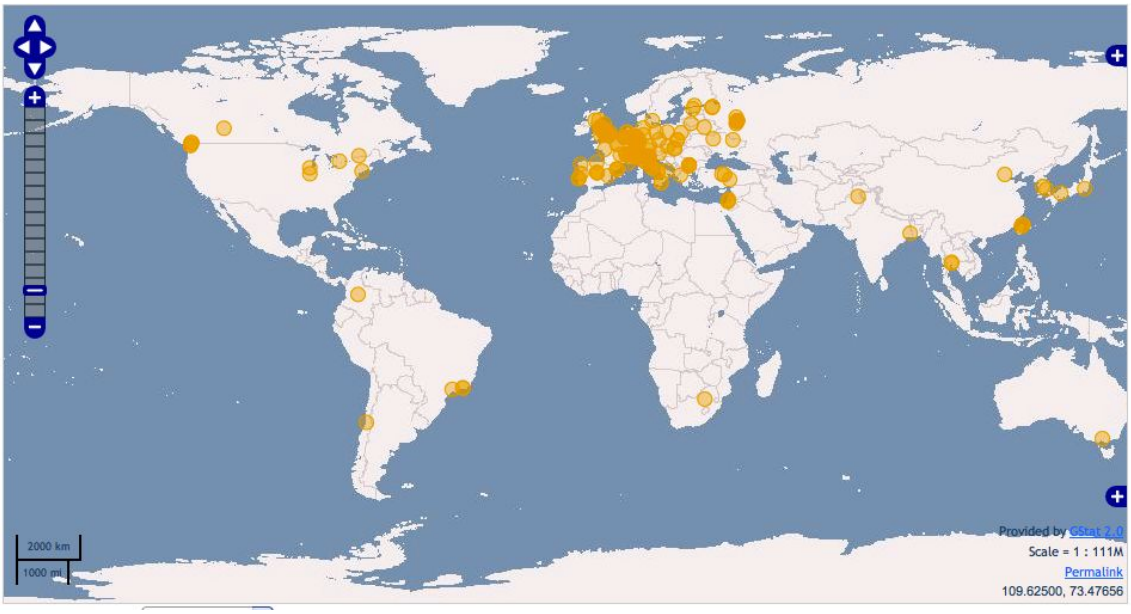

Figure 2. Map of the WLCG sites

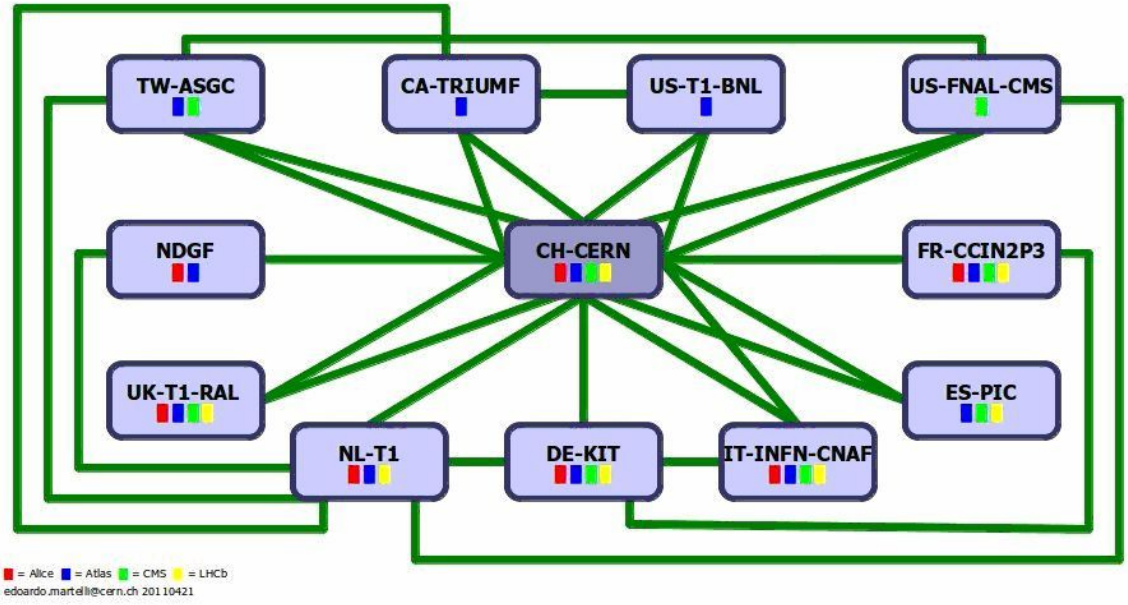

Figure 3. LHC Optical Private Network (OPN)

thousands of CPUs and PBs of disks also tape storage and 24/7 service support. Tier-1s are connected to CERN with dedicated 10Gbit/s optical links composing a backbone called LHC Optical Private Network (LHC OPN) [11], see Figure 3. The Tier-0 - Tier-1 links are in addition either duplicated and/or there are backup routes of 10Gb/s between Tier-1s themselves, which make the LHC OPN an extremely reliable system.
Then there are over 130 Tier-2 centers which are regional computing sites smaller than Tier-1s, not necessarily providing the tape storage. The connectivity of Tier-2s is provided by the national NRENs [12] and in the frame of a system in development called LHCONE (LHC Open Network Environment) [13]. 1Gb/s links capacity is usually a standard.

The raw data recorded by the LHC experiments is shipped at first to the CERN Computing Center (see Figure 1), through dedicated fiber links. At CERN, the data is archived in the CERN tape system CASTOR2 [14] and is pushed through the first level of processing - the first pass of reconstruction. The raw data is also replicated to the Tier-1 centers, so there always are 2 copies of the raw data files.
The Tier-1 centers hold the permanent replicas of raw data and are used for further raw data re-processing. This multiple-stage data re-processing applies methods developed to detect interesting events through the processing algorithms, as well as improvements in detector calibration, which are in continuous evolution and development. Also, the scheduled analyses, as well as some of the end user analyses, are performed at Tier-1s.
Tier-2 centers are supposed to host simulation campaigns (Monte Carlo simulations of the collision events) and end-user analysis.
The number of end users regularly running their analysis jobs on the WLCG infrastructure is quite large: it varies from about 300 to 800 people depending on the experiment. The fact that so many people are using the Grid for their analysis is considered a great success of WLCG: both WLCG and the experiments themselves worked hard to develop interfaces to the Grid which hide from the users the tremendous complexity of the system.

4. WLCG current status and performance during the real data taking

When the data from the first LHC collisions flooded the WLCG, it has been after years of a continuous development and massive stress-testing campaigns. The infrastructure was able to handle the data as expected.
During 2010/2011 data taking, the CERN Tier-0 tape system CASTOR2 was recording approximately 2PB of data per month for proton-proton and 4 PB per month for heavy ion collisions. Currently, there is about 50 PB of stored data at CERN from the LHC. Up to 600 MB/s of data was recorded at CERN by the ATLAS and CMS experiments during proton-proton collisions and 4 GB/s of data was recorded by the ALICE and CMS experiments during heavy ion collisions. In total, CERN Tier-0 accepts data at average of 2.6 GBytes/s with peaks up to 11 GB/s and serves data at average rate of 7 GB/s with peaks up to 25 GB/s (Figure 4). The LHC OPN is able to move data with the global rate up to 100 Gbit/s.

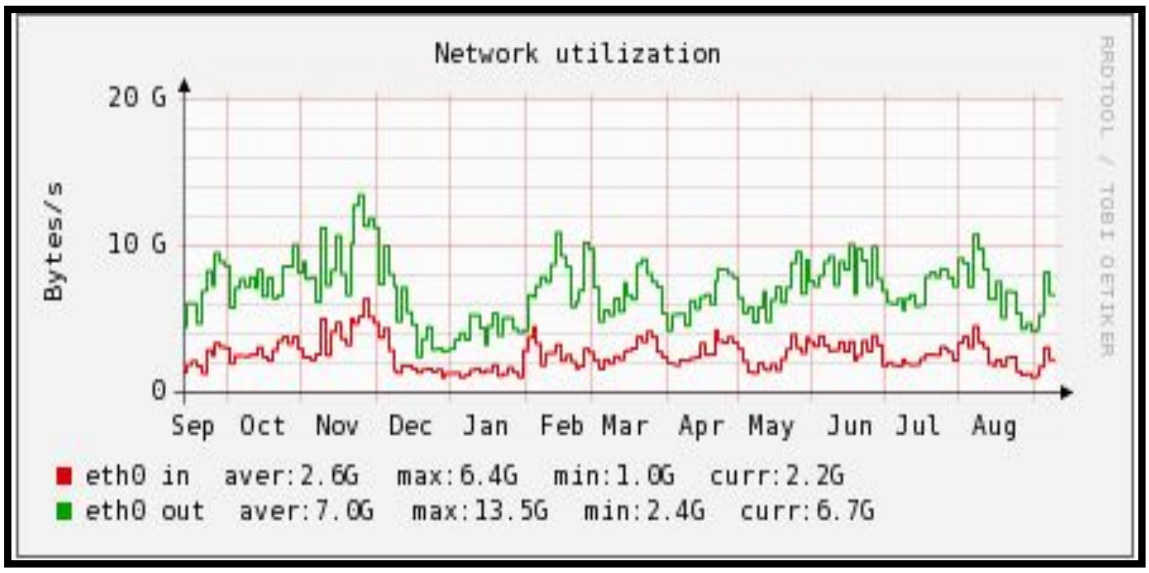

Figure 4. CERN Tier-0 Disk Servers (GB/s), 2010/2011.

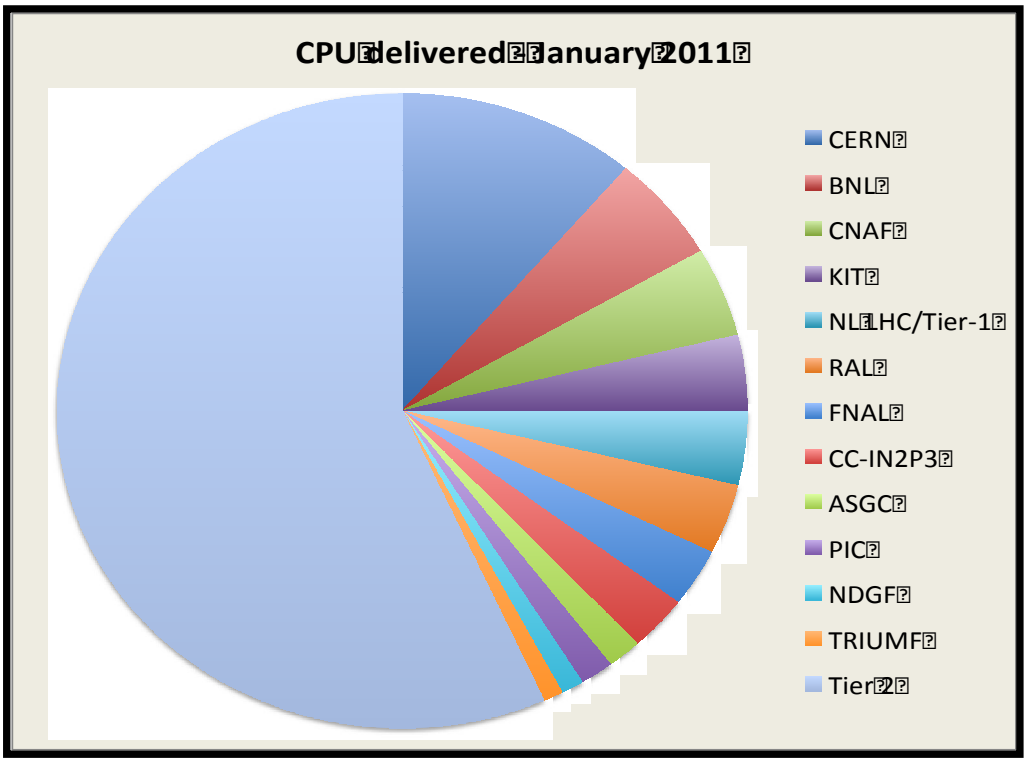

Figure 5. CPU resources in WLCG, January 2011. More than 50% was delivered by Tier-2s.

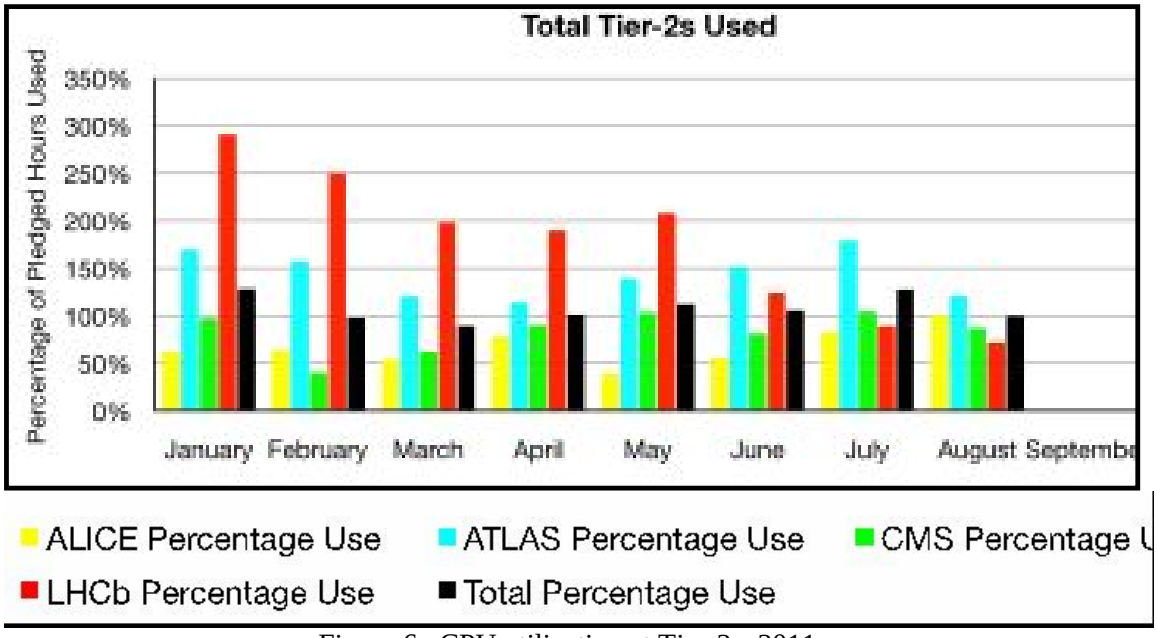

Figure 6. CPU utilization at Tier-2s, 2011.

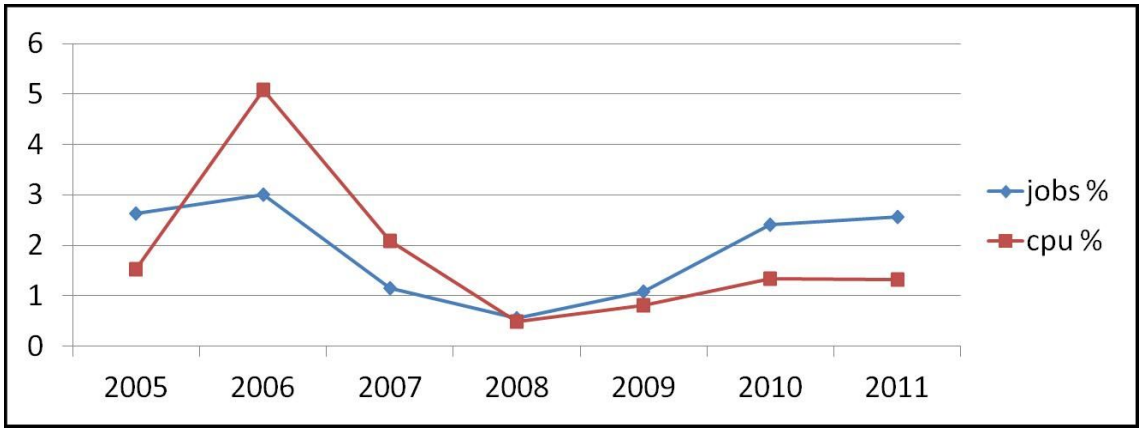

Figure 7. Contribution of the Prague Tier-2 site to the overall WLCG Tier-2s operations.

The daily jobs load represents about 150 thousands of CPUs in a continuous use, and well above 1.5 million of completed jobs per day. The sharing of the job load is displayed on Figure 5, showing the jobs statistics in January 2011. It is evident that more than 50% of the computing power is delivered by the Tier-2s. In general, the standard contribution of Tier-2s to the WLCG computing and storage resources is more than 50%, which makes Tier-2 centers an indispensable part of the WLCG infrastructure. A Tier-2 level site is "affordable" for almost every country participating in the LHC research. This brings a substantial sociological importance to Tier-2 centers: the scientists of all countries can participate regardless the size of the delivered resources.

5. An example of a WLCG Tier-2 site: Golias in Prague, Czech Republic

As a rule, WLCG Tier-2 sites are regional computing centers which provide services and resources not only to WLCG but also to other scientific projects. Typically, some kind of a sharing of resources is set and the LHC experiments/WLCG, being the most resource hungry of all hosted projects, are very good in using the opportunistic resources. This shows up in an excess of resources delivered by Tier-2s to WLCG in comparison what was officially pledged (see Figure 6).

Let us shortly describe a rather typical WLCG Tier-2 site, the regional computing center Golias [8] in Prague, Czech Republic. The site started with a couple of rack servers in 2002 and was built up and growing since then. In 2008 it became one of the signatories of the WLCG Memorandum of Understanding. According to what was mentioned above, the center provides services to the LHC experiments ALICE and ATLAS and then to a number of other projects like D0 [15] at Fermilab or STAR [16] at Brookhaven National Laboratory. Currently, the site provides 2 PetaBytes on disk servers (DPM [17], XRootD [18], NFSv3) and about 3500 cores on WorkerNodes. The site uses extensively virtual machines [19] and prepares tests of interfaces to computational clouds [20].

For a WLCG site to deliver pledged services and resources, the fundamental condition is a good connectivity. The Prague site has 1 or 10 Gbit/s end-to-end connections to local and foreign collaborating institutions and 1 Gb/s to European network system GEANT2 [21].

In Figure 7, the Prague site ATLAS+ALICE contributions to resources provided by WLCG Tier-2s are shown. The lapse in 2008 was caused by problems with funding.

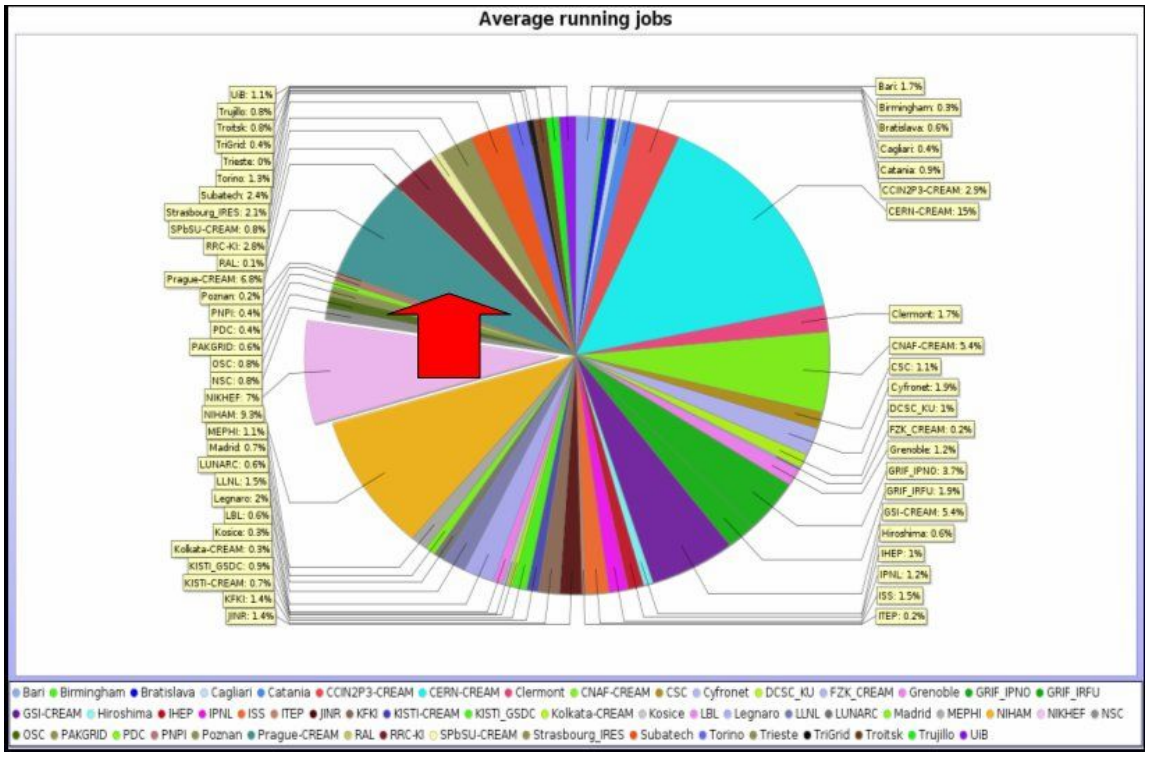

Figure 8. ALICE - data processing January 2011 : Prague contribution 6.8%
(the red arrow points to the section showing the Prague share)

A nice example of using the opportunistic resources is shown on Figure 8 with the ALICE jobs statistics for January 2011. The Prague site contribution makes 6.8%, although the officially pledged computing capacity is less than 2% of the ALICE total resources.

6. Concluding remarks

The scale at which the LHC experiments produce data makes the LHC science consume tremendous computing and storage resources which are delivered within the infrastructure of the Worldwide LHC Computing Grid. After 10 years of continuous building, testing, integration of new resources and application of newly emerged technologies, WLCG absorbed and handled the data from the LHC collisions as expected and without problems.
In the beginning, the LHC Computing Grid was a rather unique and pilot project, which has driven the development of multi-science Grids in Europe and in the US. Many different communities have benefitted from the advent of Grid infrastructures. Nowadays, the LHC Physics is by far not the only community doing the large scale computing. Other communities include e.g. Astronomy, Astrophysics, Civil Protection, Earth Sciences or Geophysics.
WLCG performance during the first two years of LHC running has been impressive and has enabled very rapid production of Physics results. The time scale at which the scientific papers are produced is unprecedented: it takes a few weeks since the raw data is recorded to scientific papers with Physics results to be submitted to journals.
Large numbers of people actively using the Grid for their analysis (up to ~800 per experiment) represent another very significant success of the WLCG project. WLCG has created a truly collaborative environment where the integrated distributed resources can be used by scientists

from all around the globe based on a global single-sign-on schema (use same credentials everywhere).
After 10 years of accumulation of resources, only the first year of data taking - 2010 was without worries about the resources becoming fully used. Already in 2011 it was clear that the current rate of resources increase can not catch up with the rising demands of the experiments. In the future, the experiments and WLCG itself must achieve a better utilization of existing resources which means e.g. adjusting the software to fully exploit the new multi-core processors, which will require fundamental changes towards parallelization of the code. Another inevitable challenges include the use of virtualization, which is already ongoing on majority of sites, and the use of (commercial or institutional) compute clouds to supplement WLCG own resources.

Acknowledgements

The work was supported by the MSMT CR contracts No. 1P04LA211 and LC 07048.

References

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

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

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

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

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

[6] Worldwide LHC Computing Grid: http://lcg.web.cern.ch/LCG/tdr/.htm

[7] Geoff Brumfiel: *High-energy physics: Down the petabyte highway*, *Nature* **469** (2011) 282-283.

[8] Regional Computing Center for Particle Physics, http://www.particle.cz/farm/public.aspx

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

[10] see e.g. The European Middleware Initiative: http://www.eu-emi.eu/home

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

[12] National research and education networks - see e.g. TERENA, http://www.terena.org/

[13] LHCONE-LHC Open Network Environment: http://lhcone.net/

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

[15] The D0 Experiment: http://www-d0.fnal.gov/

[16] The STAR Experiment: http://www.star.bnl.gov/

[17] The Disk Pool Manager: https://www.gridpp.ac.uk/wiki/Disk_Pool_Manager

[18] XRootD: http://project-arda-dev.web.cern.ch/project-arda-dev/xrootd/site/index.html

[19] Virtual Machines: http://en.wikipedia.org/wiki/Virtual_machine

[20] 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

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