Journal of Physics: Conference Series **513** (2014) 062041 doi:10.1088/1742-6596/513/6/062041

LHC Grid Computing in Russia: present and future

A. Berezhnaya[⋆], A. Dolbilov[+], V. Ilyin[⋆†], V. Korenkov[+], Y. Lazin[⋆], I. Lyalin[⋆], V. Mitsyn[+], E. Ryabinkin[⋆], S. Shmatov[+], T. Strizh[+], E. Tikhonenko[+], I. Tkachenko[⋆], V. Trofimov[+], V. Velikhov[⋆], V. Zhiltsov[+]

[⋆]National Research Centre "Kurchatov Institute", Kurchatov square, 1, 123182, Moscow, Russia
[+]Joint Institute for Nuclear Research, Joliot-Curie 6, 141980, Dubna, Moscow region, Russia
[†]Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Leninskiye gory, 1 (2), GSP-1, 119234, Moscow, Russia

Abstract. The review of the distributed grid computing infrastructure for LHC experiments in Russia is given. The emphasis is placed on the Tier-1 site construction at the National Research Centre "Kurchatov Institute" (Moscow) and the Joint Institute for Nuclear Research (Dubna).

1. Introduction

In accordance with the protocol between CERN, Russia and the Joint Institute for Nuclear Research (JINR) on participation in LCG Project approved in 2003 and Memorandum of Understanding (MoU) on Worldwide LHC Computing Grid (WLCG) signed in October 2007. Russia and Joint Institute for Nuclear Research bear responsibility for nine Tier-2 centers. This Tier-2 infrastructure currently fully satisfies the WLCG computing requirements and provides proper support of the LHC experiments' Data Processing and Analysis Tasks.

In March 2011 the proposal to create Russian LCG Tier-1 center as an integral part of the central data handling service of the LHC Experiments in Russia was expressed in the official letter by Minister of Science and Education of Russia to CERN Director General.

2. Current status of Tier-2 activities

Currently there are 3 major Tier-2 centres in JINR, Kurchatov Institute and Protvino and a number of smaller Tier-2 centers in ITEP, PNPI, SINP MSU, Troitsk, SPbSU and MEPHI. Together they contribute more than 3% of the cumulative CPU time used for processing data coming from LHC.

Apart from Tier-2 centers, there is central regional operations team that is physically spread between multiple institution and which provides central monitoring, resource coorddination, user/admin support and security operations, including running national certification authority for research and education. There is also an unofficial, but rather powerful and healthy community of system administrators and experiment support persons from Tier-2 who exchange knowledge about Grid operations, problems and their solutions. Tier-2 community in Russia is currently mostly static in the sense of new institutions joining it: there are no new actors in the field since a couple years.

3. Tier-1 construction and operations

In 2011 The Federal Target Programme Project "Creation of the automated system of data processing for experiments at the Large Hadron Collider of Tier-1 level and maintenance of Grid services for distributed analysis of this data" [1] was approved for the period 2011 – 2013.

The Project is aimed at the creation of a Tier-1 computer-based system in Russia and JINR for the processing of experimental data received from LHC and provisioning of Grid services for a subsequent analysis of this data at the distributed centers of the LHC computing Grid. It is agreed that the National Research Centre "Kurchatov Institute" (Moscow) is responsible primarily for support of ALICE, ATLAS, and LHCb experiments while the JINR (Dubna) provides Tier-1 services for the CMS experiment.

The master construction plan consists of two phases: prototype construction in the middle of 2013 and fully functional Tier-1 at the end of 2014.

4. Tier-1 in JINR

To be aligned with the WLCG and LHC Experiments requirements JINR has to provide a support of a number of the main Tier-1 services for the CMS experiment. Tier-1 in JINR will provide computing facilities of around 10% of the total existing CMS Tier-1 resources (excluding CERN) for 2013, after that resources will be further increased to catch up with the Tier-1 pledges for 2014. The network bandwidth as part of LHCOPN for Tier-0 – Tier-1 and Tier-1 – Tier-1 connections is about 2 Gbps for 2012 and will be increased consistently up to 20 Gbps in 2014. JINR link to public network with a bandwidth of 20 Gbps will be used to connect the Tier-1 with all other Tier-2/Tier-3 centres.

As per CMS computing model [2, 3], JINR Tier-1 centre will store an agreed share of raw data and Monte Carlo data and will allow access to this data to other WLCG CMS Tier-2/Tier-3 centers, will serve FTS channels for Russian and JINR Tier-2 storage elements including monitoring of data transfers.

The corresponding presentation of detailed working plan was given and adopted on the WLCG Overview Board at 28^{th} of September 2012.

The main functions of the Tier-1 centre are as follows:

- reception of experimental data from Tier-0 in the volume determined by the WLCG agreement (WLCG MOU);
- archival and protected storage of part of experimental RAW data;
- performance of consecutive and continuous data processing;
- selection of events and formation of a data set by predetermined criteria (skimming);
- data reprocessing with the help of new software or new constants of calibrations of detector systems and spatial smoothing of parts of the CMS installation;
- creation of AOD data;
- distribution of RECO and AOD to other Tier-1/Tier-2/Tier-3 centers for the duplicated storage (replication) and physical analysis;
- remodeling with the help of new software and calibration constants and constants of spatial smoothing of parts of the CMS installation, protected storage of the simulated events.

Implementations for functions of the Tier-1 center are provided by various services of the computing model of CMS collaboration with a high level of functionality and reliability [4, 5].

Services of CMS centers of the Tier-1 level are subdivided into the system ones (inaccessible to users) and the services accessible to users (user services).

A first stage of the Tier-1 center prototype was designed in 2012. The modules consist of:

- Worker node (WN). 100 64-bit machines: 2xCPU (Xeon X5675 @3.07GHz, 6 cores per processor); 48 GB RAM; 2x1000 GB SATA-II; 2x1GbE. Total: 1200 core/slots for batch processing.
- Storage system (SE) (dCache). Disk Only:
  - 7 disk servers: 65906GB h/w RAID6 (24x3 TB SATA-III); 2x1GbE; 48GB RAM.
  - 1 disk server: 2x17974GB h/w RAID6 (2x8x3 TB SATA-III); 2x1GbE; 48GB RAM.
  - 3 head node machines: 2xCPU (Xeon X5650 @2.67GHz; 48GB RAM; 500GB SATA-II; 2x1GbE.
  - 8 KVM (Kernel-based Virtual Machine) nodes for access protocols support.
- Mass Storage System:
  - 2 disk servers: 65906GB h/w RAID6 (24x3 TB SATA-III); 2x1GbE; 48GB RAM.
  - 1 tape robot: IBM TS3200, 24xLTO5; 4xUltrium5 FC8; 72TB.
  - 3 head node machines: 2xCPU (Xeon X5650 @2.67GHz; 48GB RAM; 500GB SATA-II; 2x1GbE.
  - 6 KVM machines for access protocols support.
- WLCG services: 17 virtual machines (KVM).
- Infrastructure servers: 17 machines: 2xCPU (Xeon X5650 @2.67GHz; 48GB RAM; 500GB SATA-II; 2x1GbE.
- Software:
  - OS: Scientific Linux release 6/x86_64 and Scientific Linux release 5/x86_64 for PhEDEx.
  - WLCG services: EMI-2 SL6 x86_64, EMI-3 SL6 x86_64 Argus server, and WLCG SL5 x86_64 VOBOX for PhEDEx.
  - Batch system: Torque 4.1.4 and Maui 3.3.2 (heavily-tuned in-house builds).
  - Storage system: dCache-2.2.11 (dcache.org).

4.1. Ongoing and future work

Currently this prototype system was fully validated by CMS and passes all functional tests, has certified transfer channels and is used for real CMS jobs.

Once Tier-1 in JINR will be fully functional and well-tested, we will start to work with Tier-2 centers from Russia and Dubna Member States (RDMS) who support CMS experiment; currently RDMS is a very important part of worldwide CMS, so supporting these centres on the Tier-1 level is our priority.

Also Tier-1 team from JINR will provide operational support for regional centers and users, including consultations on deployment of computing centres of various scale, help with generic and Grid-specific problems, security support and dissemination of best practices in the field of Grid computing and general system administration.

5. Tier-1 in Kurchatov Institute

Tier-1 centre in Kurchatov Institute is built to support three LHC virtual organizations, ALICE, ATLAS and LHCb. Resources are scaled to be on the 10% level of the aggregate capacity of all other existing Tier-1 centres, at least for period 2014 – 2016.

During the prototype phase, backed up by the govermental project, we had built the prototype for our Tier-1 that includes all WLCG and VO-specific services for the supported experiments:

- computing field that consists of batch system that is based on Torque and Maui, backed up by the EMI-based worker nodes and supplemented by CREAM CE services that are used as the gateways for the job submission and output retrieval;
- disk-based storage system that has dual personality: ATLAS and LHCb instances run dCache and ALICE instance runs EOS;

- tape-based storage system; currently only dCache has tape backend based on Fermilab's Enstore;
- various site services like site BDII, top-level BDII, APEL box and VO-BOX machines;
- infrastructure services like CVMFS Squid farm, NAT farm, DNS/firewall services, logging nodes, Nagios monitoring, NFS;
- installation and configuration system that is currently based on Puppet.

As was outlined in our technological review [6], service-wise we're concentrating on providing highly-reliable services, so all service instances that have possibility to be clustered are clustered:

- BDII nodes are running uCARP-based master/slave clusters;
- Squid nodes are running CARP-based master/master clusters;
- CREAM CEs are just provided in many instances that serve all supported VOs and we're relying on experiment's ability for running round-robin job submission;
- PostgreSQL for Enstore and dCache runs master/slave clusters comprised from two nodes and shared external storage that runs dual-controller master/master stack for giving away SAS connections to the disk arrays.

Just after we had finished the prototype phase, the built Tier-1 was used for ATLAS to reprocess their historical data from the 2011's 2.76 TeV run: our resources crunched 54 TB of data in roughly 37000 tasks and job efficiency was 98%, that we and ATLAS consider to be a major success. The said reprocessing was the part of validation of our Tier-1 for ATLAS in the tape-less mode: it was successfully passed and since then our resources are running ATLAS production and being prepared to be validated as the full Tier-1 with tape-based storage system.

We also did a major step in improving our team practices on collaborative work and monitoring, though this area is still very much open to the future progress as our resources and team will grow, so this is also one of the most demanding problems.

5.1. Ongoing and future work

Currently our team is rolling out new computing and storage resources for our Tier-1 and is moving to the new 10 GE-based internal fabric that will scale up to 15 Tbit of aggregate bandwitdh.

We're working on improving of reliability for our resources. One of the main concerns is the Torque/Maui batch system: while Torque 4.x made some progress in employing modern algorithms to handle computing fields sized to thousands of nodes, it still carries legacy of OpenPBS and its older predcessors. Slurm is our next system of choice and we're currently working on software components (BLAH and accounting parts) that will allow to use it for WLCG middleware stack.

Another connected problem is our current usage of NFS for staging input and output files for Torque. Having non-clustered NFS, we already faced some issues when server's hardware was failed, but instead of shooting at clustered NFS, we're currently working on NFS-less stagein/stageout system that will be HTTP-based and will allow to scale with the number of employed CREAM CE nodes.

Support for this system will be integrated with Torque and Slurm will be armed with simple stagein/stageout, the feature that its currently completely lack, being targeted at supercomputer systems that almost in every case imply the existence of shared filesystem. Older WLCG and LHC VO software stacks also wanted shared filesystem area for software installation, but with the advent of CVMFS [7] this requirement is gone, so we are going to leverage much simpler model of transferring data between CE and worker nodes, since it has only "write-once, then read" semantics without distributed append functionality and read/write concurrency that is one the main problems for the distributed file systems.

Another major work is building the unified monitoring system for VO-specific and site-specific metrics: since we are supporting 3 different VOs that have many monitoring endpoints and we need to be aware of the status of all tests and be able to correlate failures with the site-level sensors, unified dashboard is the tool that is really needed for this task. Being working with many monitoring systems during both Tier-2 and Tier-1 activities, we had laid out requirements for this system and are currently building its architecture and the first prototype.

Other work item is adding support for LHCb and ALICE; this work is progressing just at the time of writing this paper and we expect these two VOs to be able to use our resources in the coming weeks.

The future piece of work is integrating existing Tier-2 centres in Russia with our Tier-1 according to the practices of each supported VO. We are currently working on this through making our VO support persons to become experts in the respective fields; though the work on "connecting" Tier-2 centres to our resources will be done only after our Tier-1 centre will show good results in stand-alone mode: almost any VO needs Tier-1 centers to be really reliable to allow for co-operation with Tier-2 centres, since the latter depend on the former in many respects, especially for VOs that still running some variants of an old hierarchical model from MONARC.

6. Conclusions

Russian teams from Dubna and Kurchatov Institute are progressing in their activities in building Tier-1 centre for LHC experiments. Both institutions show first promising results from their prototypes, although, of course, there is much future work to carry on.

We expect that fully-fledged Tier-1 centres in both institutions will be ready for the second LHC run that is seen to happen somewhere in the beginning of 2015.

Acknowledgements

The work in JINR and KI was partly carried out within the federal program "Research and development on the priority directions of the development of the scientific-technological complex in Russia for 2007 – 2013" (contract № 07.524.12.4008).

The work in KI was partly carried out within thematic plan for research and development programme of NRC "Kurchatov Institute" for 2013 – 2014.

References

[1] National federal program "Research and development on the priority directions of the development of the scientific-technological complex in Russia for 2007 – 2013, http://www.fasi.gov.ru/fcp/compl/797/

[2] LHC Computing Grid Technical Design Report. CERN-LHCC-2005-024, 2005; Worldwide LHC Computing Grid (WLCG), http://lcg.web.cern.ch/LCG/public/default.htm

[3] C. Grandi, D. Stickland, L. Taylor, CMS NOTE 2004-031 (2004), CERN LHCC 2004-035/G-083; CMS Computing Technical Design Report, CERN-LHCC-2005-023 and CMS TDR 7, 20 June 2005

[4] V.V. Korenkov, N.S. Astakhov, S.D. Belov, A.G. Dolbilov, V.E. Zhiltsov, V.V. Mitsyn, T.A. Strizh, E.A. Tikhonenko, V.V. Trofimov, S.B. Shmatov. Creation at JINR of the data processing automated system of the TIER-1 level of the experiment CMS LHC. // Proceedings of the 5th Inter. Conf. "Distributed Computing and Grid-technologies in Science and Education", ISBN-5-9530-0345-2, Dubna, 2012.

[5] N.S. Astakhov, S.D. Belov, I.N. Gorbunov, P.V. Dmitrienko, A.G. Dolbilov, V.E. Zhiltsov, V.V. Korenkov, V.V. Mitsyn, T.A. Strizh, E.A. Tikhonenko, V.V. Trofimov, S.V. Shmatov. Tier-1-level computing system of data processing for the CMS experiment at the Large Hardon Collider. 15 p., Information Technologies and Computation Systems, № 3, accept., 2013.

[6] A.Y. Berezhnaya, V.E. Velikhov, V.A. Ilyin, R.N. Kolchin, Y.A. Lazin, I.N. Lyalin, E.A. Ryabinkin, I.A. Tkachenko, F.K. Checherov. Technological problems and solutions for the process of building Tier-1 resource center using ATLAS as an example of LHC's virtual organization. // Proceedings of the 5th Inter. Conf. "Distributed Computing and Grid-technologies in Science and Education", ISBN-5-9530-0345-2, Dubna, 2012

[7] J. Blomer, P. Buncic, I. Charalempidis, A. Harutyunyan, D. Larsen, R. Meusel. 2012 Status and future perspectives of CernVM-FS *J. Phys.: Conf. Ser.* **396** 052013