JINR Tier-1 Center*

N.S. Astakhov[1], S.D. Belov[1], P.V. Dmitrienko[1], A.G. Dolbilov[1], I.N. Gorbunov[2], V.V. Korenkov[1], V.V. Mitsyn[1], S.V. Shmatov[2], T.A. Strizh[1], E.A. Tikhonenko[1], V.V. Trofimov[1], V.E. Zhiltsov[1]

e-mail: korenkov@jinr.ru, [1]Laboratory of Information Technologies, Joint Institute for Nuclear Research, Dubna

[2]Veksler and Baldin Laboratory of High Energy Physics, Joint Institute for Nuclear Research, Dubna

Abstract

JINR collaboration with the NRC "Kurchatov Institute" in 2011-2013 on the design of an automated system for the LHC data processing at the Tier-1 level and provision of the grid-services for distributed data analysis is of great importance for the development of both the JINR informational-computational infrastructure as a whole and its grid-segment.

Today the CMS Russia and Dubna Member States (RDMS) computing infrastructure allows a proper support of the CMS data processing and analysis tasks.

This report is devoted to a new element of RDMS Computing - the CMS Tier-1 site which is under construction in JINR. The plans foresee:

- organization of LHC OPN network,
- in 2012 - beginning of 2014 a prototype needs to be designed and commissioned,
- in 2013-2015 - preparation of the infrastructure (cooling, power system, network, etc.) for the full-featured Tier-1.

The goal – CMS Tier-1 in Dubna starts in a full scope in the WLCG for the end of 2015.

In accordance with the protocol between CERN, Russia and JINR on participation in LCG Project approved in 2003 and Memorandum of Understanding (MoU) on Worldwide LHC Computing Grid (WLCG) [1] signed in October of 2007, Russia and the Joint Institute for Nuclear Research (JINR) bier responsibility for seven Tier-2 sites (CMS support). Here and now JINR computing Tier-2 infrastructure fully satisfies the WLCG Computing Requirements and provides a proper support for the LHC experiments data processing and analysis tasks.

In March of 2011, the proposal to create the WLCG Tier-1 site as an integral part of the central data handling service of the LHC Experiments in Russia was expressed in an official letter by the RF Minister of Science and Education Andrey Fursenko to CERN Director General Rolf-Dieter Heuer. In pursuance to achieve principal provisions of this proposal discussed on Russia-CERN meeting in October of 2011, CMS Computing Meeting in April of 2012, CMS Workshop on "Perspectives on Physics and on CMS at Very High Luminosity, HL-LHC" in May of 2012, JINR has agreed to accept responsibility of creation of a Tier-1 site to serve the LHC CMS experiment.

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 a distributed analysis of these data" was approved on the period of 2011-2013 with the budget amounted to about 8.5 MCHF. The Project is aimed to creation of a Tier-1 computer-based system in NRC KI and JINR for processing experimental data received from LHC and provision of grid services for a subsequent analysis of the data at the distributed centers of the LHC computing grid. It is shared that the National Research Center "Kurchatov Institute" is responsible for support of ALICE, ATLAS, and LHC-B experiments, while the JINR provides Tier-1 services for the CMS experiment.

The master execution plan of the Tier-1 development (or construction) consists of three phases (2012-2014). The first phase was the construction of the prototype by the end of 2012. The next one is implementation of full Tier-1 functionality, which has to be completed in 2014 (Phase I). Phase II foresees the upgrade of Tier-1 resources in 2015 .

Since the start-up in line with the WLCG and LHC Experiments requirements, the JINR has to provide a support of a number of the main Tier-1 services for the CMS experiment. The Tier-1 site in JINR will provide computing facilities of 10% of the total existing CMS Tier-1 resources (excluding CERN) for 2015. 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 10 Gbps in 2014. The 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 sites.

In agreement with the CMS Computing model [1, 2], the JINR Tier-1 site will provide

*The work is 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).

acceptance of an agreed share of raw data and Monte Carlo data and provision of access to the stored data by other CMS Tier-2/Tier-3 sites of the WLCG, will serve FTS-channels for Russian and DMS 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 on 28 September 2012.

In accordance with the CMS computing model [1, 2], a Tier-1 site is intended for long-term data archiving and processing of "raw" data coming from the detectors of experimental setup and for their preparation for subsequent analysis at the Tier-2 sites. In the computing model of CMS at the Tier-1 there are two main types of stream job processing: data re-reconstruction and simulated data re-digitization/re-reconstruction. During the data re-reconstruction, Tier-1 sites process RAW data with the use of newer software and/or in view of updated calibration and alignment constants of detector systems as well as more exact information about the setup state during data taking. The output data are recorded in RECO (RECOnstructed data) and in AOD (Analysis Object Data) formats. RECO are the data containing the values of parameters of physical objects (tracks, interaction vertices, jets, electrons, muons, photons, etc.) as well as clusters and hits reconstructed from the RAW data with the usage of various algorithms. They are an output data stream from Tier-0 for redistribution over various Tier-1 sites. The volume of one event is about 0.5 MB. These data can be used for the analysis, but they are inconvenient due to their large size. AOD represents a selective set of information from the RECO data. AOD events contain the parameters of high-level physics objects, plus sufficient additional information to allow kinematic refitting. AOD has a considerably smaller size, as compared to RECO (0.12 MB per one event), and is used for reconstruction of the final topology of a physical event and a subsequent analysis.

In addition, by selected criteria Tier-1 provides selection of events from the reconstructed data (skimming). A similar selection can be carried out from "raw" (RAW) data or elsewhere from already reconstructed RECO data. These events pass through processing and accumulation similar to that during a re-reconstruction; then they are recorded in files of RECO format or in incorporated format (RAW-RECO).

In the Tier-1 sites, like in the case of experimental data, there is a reprocessing of simulation data with the use of newer versions of software and/or in view of the updated calibration and alignment constants of detector systems. The input data of GEN-SIM-RAW type are exposed to re-digitization (part of GEN-SIM data) for receiving updated versions of the simulated data such as RAW which are further reconstructed again (on output GEN-SIM-RECO and/or AODSIM data).

Thus, the main functions of the Tier-1 site are as follows:

- Receiving of experimental data from a Tier-0 site in the volume determined by the WLCG agreement (WLCG MOU);
- Archiving and custodial storage of part of experimental RAW data;
- Consecutive and continuous data processing;
- Additional processing (skimming) of RAW, RECO and AOD data;
- Data reprocessing with the use of new software or new calibration and alignment constants of parts of the CMS setup;
- Making available AOD data-sets;
- Serving RECO and AOD datasets to other Tier-1/Tier-2/Tier-3 sites for their duplicated storage (replication) and physical analysis;
- Running production reprocessing with the use of new software and new calibration and alignment constants of parts of the CMS setup, protected storage of the simulated events;
- Production of simulated data and data analysis recorded by the CMS experiment.

The implementation of the Tier-1 site functions is provided by various services of the computing model of CMS collaboration with a high level of functionality and reliability.

Tier-1s will provide the following set of user-visible services:

- Data Archiving Service;
- Disk Storage Services;
- Data Access Services;
- Reconstruction Services;
- Analysis Services;
- User Services (Tier-2 type user services).

Also a Tier-1 site must provide some specialized system-level services:

- Mass storage system;
- Site security;
- Prioritization and accounting;
- Database Services.

Figure 1 presents the JINR Tier-1 CMS infrastructure scheme.

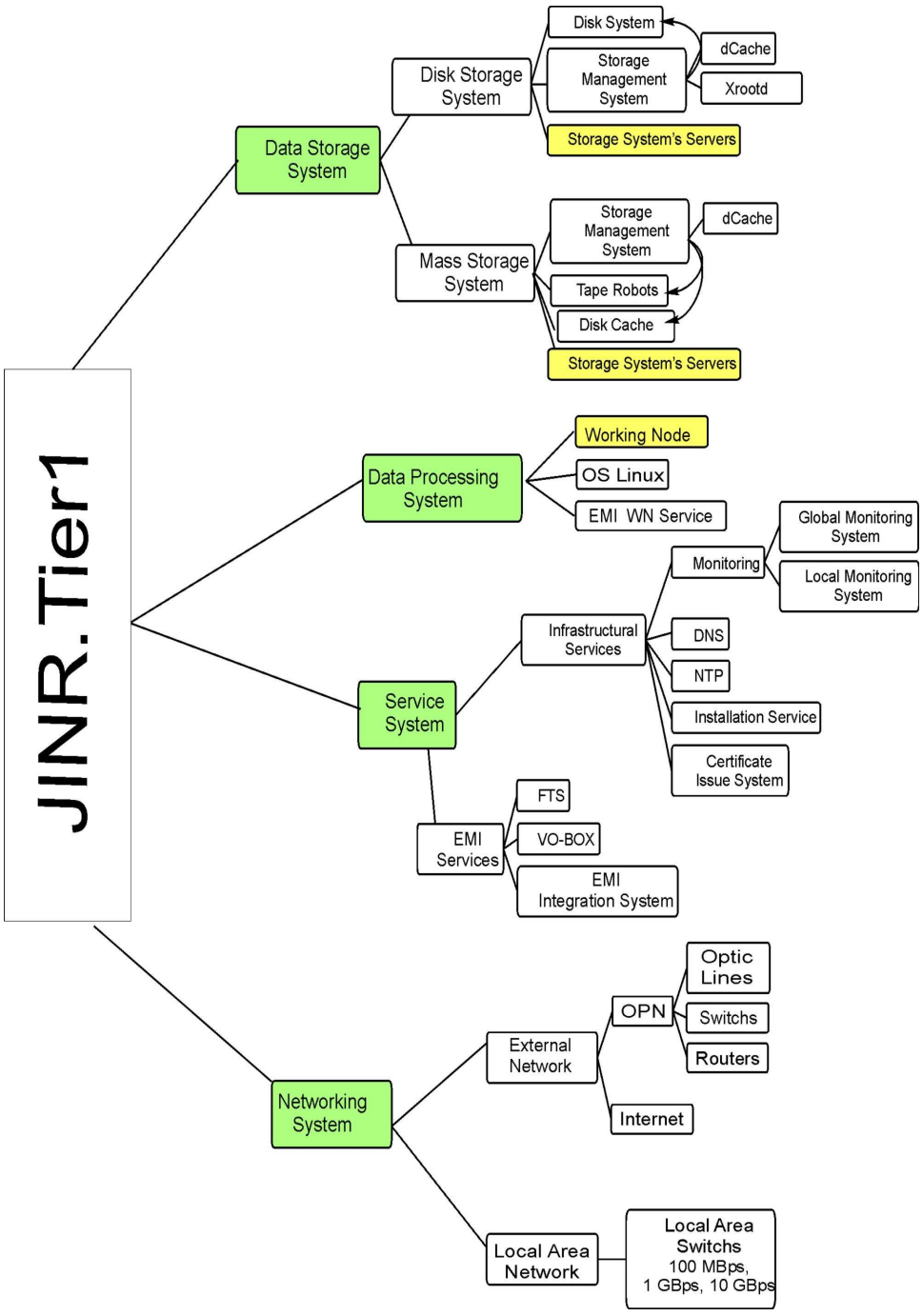

Figure 1. JINR Tier-1 infrastructure scheme

Data storage subsystem is developed for experimental data storage from Tier-0 site (CERN) of LHC experiment, for simulated data that come from the Tier-2 sites, and also for output data of jobs on data processing that are carried out in the CMS sites. According to the CMS data model, the size of files with data should not exceed 10 GB. Therefore, the data storage system should provide not less than 10 GB of disk-space on computation process. Also a system should provide RAM not less than 2 GB per process.

Computing system *(Computing Elements)* is developed for data processing from the LHC experiment. This subsystem includes a number of services that are a part of grid middleware for providing access of grid jobs to local resources of Tier-1 site (LRMS - Local Resource Management System). Usually, the system provides access to a number of job's queues at the computing node inside the grid (calling the Worker Nodes (WN)). The number of WN should correspond to the expected system workload. The jobs of common users are not

allowed to be processed on the resources of Tier-1 sites; the access to the queues is limited in accordance with the subsystem of access roles (t1 access roles).

Data transfer subsystem *(FTS)* between sites of various levels. This service can also be realized with the help of the grid middleware tools. Data transfer is carried out by means of constructing virtual channels with the possibility of indication of transfer sequences and priorities. According to the CMS computational model, all transfers between various Tier-1 sites and between associated Tier-2 sites and Tier-1 ones are carried out by means of FTS services.

Management of data transfer and data storage *(CMS VOBOX)* on the level of CMS datasets and data flows is carried out by means of PhEDEx project tools which include:

- Transfer management database - TMDB.
- Transfer agents managing files transfer between sites, data migration on local storages, checksum of transferred data.
- Managing agents providing files allocation according to the site's subscription for the data.
- Local agents providing files processing after their reception or before their transfer: files' match-merge, files' registration in catalogues, provision of information about files in the files management database.
- Transfer monitoring and presentation of results by means of a web-interface.

Load distribution system and interface between grid and local computing queues provide information interchange and commands transfer between various devices and subsystems of a developing system, between the system and related systems, and, also, between Tier-2 sites WLCG in Russian and in the world.

CMS Tier-1 network infrastructure in JINR includes LHC OPN subsystem for organization of dedicated channels of data transfer that connects Tier-1 and Tier-0 sites with the Tier-1 local network infrastructure. LHC OPN transfer capacity between Tier-0 - Tier-1 and Tier-0 - Tier-1 equaled to 2 Gb/s at the end of 2012 and will be increased to 10 Gb/s in 2014. JINR is also connected with academic networks with transfer capacity of 2x10 Gb/s, that provides connection of Tier-1 JINR with Tier-2/Tier-3 sites.

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

1. **Worker node (WN)**

100 64-bit machines: 2xCPU (Xeon X5675 @ 3.07GHz, 6 cores per processor); 48GB RAM; 2x1000GB SATA-II; 2x1GbE. Total: 1200 core/slots for batch.

2. **Storage system (SE) (dCache).**

Disk Only:

- 7 disk servers: 65906GB h/w RAID6 (24x3000GB SATAIII); 2x1GbE; 48GB RAM
- 1 disk server: 2x17974GB h/w RAID6 (2x8x3000GB SATAIII); 2x1GbE; 48GB RAM
- 3 head node machines: 2xCPU (Xeon X5650 @ 2.67GHz; 48GB RAM; 500GB SATA-II; 2x1GbE.
- 8 KVM (Kernel-based Virtual Machine) for access protocols support

Mass Storage System:

- 2 disk servers: 65906GB h/w RAID6 (24x3000GB SATAIII); 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

3. **WLCG services:** 17 virtual machines (KVM).
4. **Infrastructure servers:** 17 machines: 2xCPU (Xeon X5650 @ 2.67GHz; 48GB RAM; 500GB SATA-II; 2x1GbE.
5. **Software:**

- OS: Scientific Linux release 6 x86_64 and Scientific Linux release 5 x86_64 for Phedex.
- WLCG services: EMI2 sl6 x86_64, EMI3 sl6 x86_64 Argus server, and WLCG sl5 x86_64 VOBOX for Phedex.
- Batch system: Torque 4.1.4 (home made) and Maui 3.3.2 (home made)

6. **Storage system**: dCache-2.2/2.6 (dcache.org)

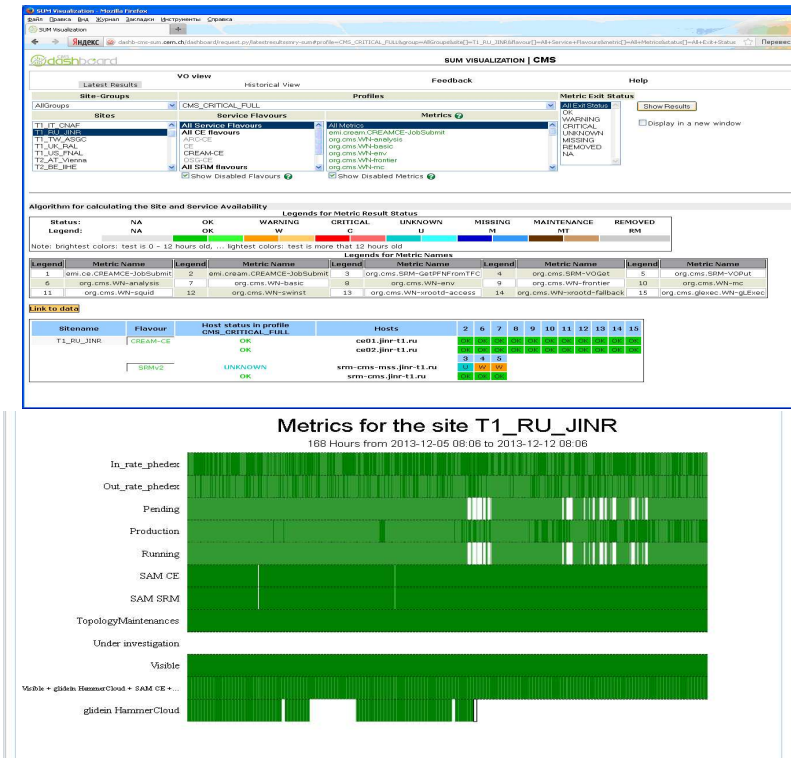

Figure 2: Metrics for the JINR Tier-1 site

In order to improve the site reliability for production activities, CMS defined metrics to determine whether a site is ready for production. The metrics are based on the result of common WLCG tests and CMS specific tests. Failing the metrics for more than 2 days over the last five causes the site to be considered not ready for production. A consecutive period of 2 days with all the metrics satisfied must be achieved before being considered ready again.

The following parameters are used to estimate the readiness of the site:

- time per day during which all necessary tests have successfully accomplished;
- percentage of successfully running data analysis jobs per day;
- a number of connections with other sites for data transfer.

Test results are summarized in the status summary table at Site Status Board (SSB) monitoring at the CMS Dashboard: (http://dashb-ssb.cern.ch/ dashboard/ request.py/siteviewhome) [5].

Fig. 2 present test results of a prototype of the JINR Tier-1 site.

References

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

[2] 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.

[3] 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 5^{th} Inter. Conf. "Distributed Computing and Grid-technologies in Science and Education", ISBN-5-9530-0345-2, Dubna, pp. 254-265, 2012(in Russian).

[4] 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, The 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.

[5] R. Rocha et al., Experiment Dashboard for Monitoring of the Computing Activities of the LHC Experiments on the Grid, Grid Computing. Nuclear Science Symposium, IEEE (Dresden), October 2008; http://dashboard.cern.ch/cms/