THE CMS TIER1 AT JINR: FIVE YEARS OF OPERATIONS

Andrey Baginyan, Anton Balandin, Sergey Belov, Andrey Dolbilov, Alexey Golunov, Natalia Gromova, Ivan Kadochnikov, Ivan Kashunin, Vladimir Korenkov [b], Valery Mitsyn, Igor Pelevanyuk, Sergei Shmatov, Tatiana Strizh [a], Vladimir Trofimov, Nikolay Voytishin, Victor Zhiltsov

Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region, 141980, Russia

E-mail: [a] strizh@jinr.ru, [b] korenkov@jinr.ru

This article summarizes five years of operational experience of the WLCG Tier1 computer centre at the Laboratory of Information Technologies of Joint Institute for Nuclear Research, which serves the CMS experiment at LHC. In early 2013 the Tier1 prototype was deployed and its trial operation began. March, 2015 is the date of finalization of this complex and commissioning a full-scale Tier1 centre for CMS at JINR. Since its inception it was continuously adapted to the new requirements, introducing new hardware and technologies as they became available. The resources provided by the centre to the CMS experiment have increased significantly and it is on top reliability levels as compared to other Tier1 centers processing data for CMS. A special Tier1 network centre was developed to provide scalability of its network infrastructure. Additional work has been done in the recent years in hardware and services monitoring. Future modernization and increase of the Tier1 performance will provide possibilities of efficient and fast processing and reliable storage of the CMS data to cope with high luminosity and high energy of the collisions of LHC run 3.

Keywords: WLCG, Tier1, grid, CMS, distributed computing

1. Introduction

The LHC accelerator at CERN [1] started to work in November 2009. Since then hundreds of Petabytes (~50-70 PB/year) of data – raw, processed, and simulated – were gathered from all experiments. To process and analyze such an unprecedented volume of data, the largest in the world computing infrastructure was built which comprises of more than 170 computer centers in 42 countries – Worldwide LHC Computing Grid network (WLCG [2]). The WLCG is set up as a layered structure: the 14 of all the computer centers, which make the layer 1 (Tier1), receive raw data from layer 0 (Tier0) at CERN in real time to store and preprocess it and provide the centers of layer 2 (Tier2) with it for analysis. The Joint Institute for Nuclear Research (JINR [3]) is one of Tier1 sites for CMS experiment [4] at LHC [5], and one of Tier2 sites for all four experiments at LHC and other Virtual Organizations (VO) of the worldwide GRID infrastructure.

The Tier1 at JINR started as a prototype node in 2011-2013 as the Federal Target Programme Project: «Creation of the automated system of data processing for experiments at the LHC of Tier1 level and maintenance of Grid services for a distributed analysis of these data». The project assumed the deployment of distributed Tier1 center in Russia. The primary producers of experimental data at LHC are two largest experiments – CMS and ATLAS [8]. This was the reason which defined the predestination of the Russian Tier1. Two Tier1 sites were established: at the Laboratory of Information Technologies (LIT) of JINR to deal with the CMS data, and at NRC "Kurchatov Institute" [9] to support ATLAS, ALICE [10] and LHCb [11] experiments. The primary reason of assigning the CMS support to JINR is the fact that Russian institutions participating in CMS and JINR as international organization with its member states are represented in the experiment as united collaboration - "Russia and Dubna Member States" (RDMS) [12]. More than 300 scientists and specialists from 21 institutions of 8 countries participate in RDMS. This way of cooperation of many institutions of various specializations allows RDMS to be completely responsible for a number of detectors in CMS, and also to make significant contribution to the development of the basic software and computing.

The basic tasks of the Tier1 center for CMS at JINR (T1_RU_JINR in WLCG mnemonics) are:

1) to receive the experimental data from Tier-0 site in the volume determined by the WLCG agreement (WLCG MOU);
2) archiving and custodial storage of part of experimental RAW data;
3) consecutive and continuous data processing;
4) additional processing (skimming) of RAW, RECO (RECOnstructed) and AOD (Analysis Object Data) data;
5) data reprocessing with the use of new software or new calibration and alignment constants of parts of the CMS setup;
6) making available AOD data-sets;
7) serving RECO and AOD datasets to other Tier1/Tier2/Tier3 sites for their duplicated storage (replication) and physical analysis;
8) 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;
9) production of simulated data and data analysis recorded by the CMS experiment.

2. Tier1 infrastructure

Tier-1 deployment required substantial upgrade of JINR LIT data center infrastructure to grant not only enough computer power with modern software to store and process the data but also evolved and durable support facilities to provide fail-safe 24/7/365 functionality of the site.

2.1. Engineering infrastructure

Since 2014 intensive works were carried out at LIT on the engineering infrastructure upgrading. New closed-loop air conditioning system was set up. It is comprised of two UNIFLAIR ARAF 1204A chillers working in N+1 redundancy. Eight APC InRow ACRC 502 precise air conditioners of 39.56 kW cold productivity provide the required air temperature and humidity in

closed aisle of Tier1 setup. Two 300 кWA Galaxy 7000 UPS guarantee the fail-safe power for the setup. In 2018 two diesel generators were put into operation to guarantee the autonomy of Tier1 power supply.

2.2. Networking

One of the most important components of JINR Tier1 providing access to resources and the possibility to work with the Big Data is the network infrastructure. External optical telecommunication channel uses DWDM (Dense Wave Division Multiplexing) technology for data transmission. Ethernet frames are transmitted through DWDM carrier signals. The most attractive feature of DWDM technology in comparison with traditional optical technologies, in which one optical fiber is transmitted only one digital signal, is the parallel transmission of multiple digital signals over the same optical fiber. Optical equipment from Nortel Systems (Canada) was installed at three points of the JINR-Moscow data channel: at the LIT JINR central telecommunications center, at the intermediate point of the optical highway (Radishchevo settlement), at the Moscow International Exchange (Moscow Internet Exchange) Site Internet. This equipment operates with 2λ equal to 10 Gbit/s each and 1λ to 100 Gbit/s. Thus, the capacity of the optical telecommunication channel of JINR is 120 Gbit/s.

The external superimposed network LHCOPN [14] (JINR-CERN) passing through MGTS-9 in Moscow, Budapest, Amsterdam, is for communication of the Tier0 (CERN) and Tier1 (JINR) centers, while the external superimposed network LHCONE [15], passing the same route, is intended for communication of the Tier2.

The Tier1 network segment provides fail-safe operation of 160 disk servers, 25 computer blade servers, 100 Grid infrastructure support servers, and tape robot. The segment is built with Brocade equipment in which the IS–IS (Intermediate System to Intermediate System) protocol is used for network segment definition on level 2 of OSI (Open Systems Interconnection) model.

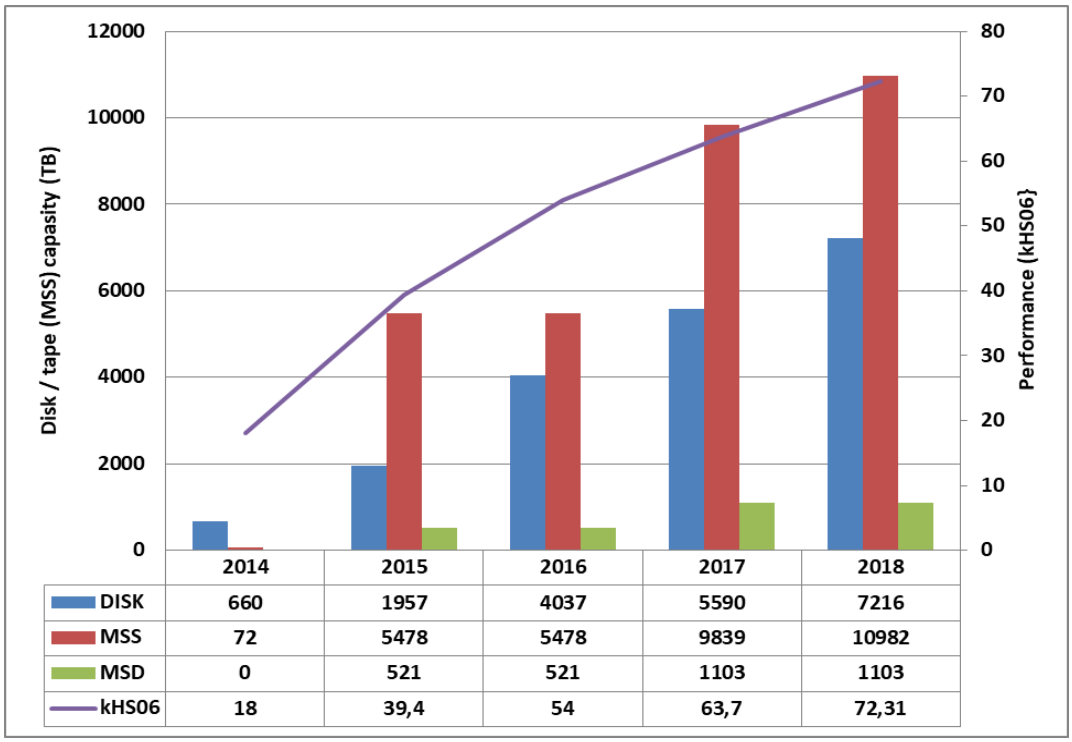
For this protocol Dijkstra's algorithm from graph theory is used. It compares and calculates the shortest path through all nodes in the network. It is constructing a shortest-path tree from the first vertex to every other vertex in the graph. On this basis was developed a modern protocol Transparent Interconnection of Lots of Links (TRILL) on the second tier of the OSI model, enabling building solutions for campus networks and data centers. Transparent Interconnection of Lots of Links (TRILL) offers many advantages. TRILL provides Layer 2 multipath and multi-hop routing.

A full redundancy of links is provided for at all levels. As a result of such architecture a failure of one switch shall lead to the reduction in the total traffic capacity of the network segment only by 25%. In such case all servers will have access to the external network [16].

2.3. Hardware

The evolution of computing and storage capacity of CMS Tier1 at JINR [17-19] from 2014 to 2018 is presented on Figure 1.

Currently, the JINR Tier1 for CMS hosts 275 worker nodes (WNs) for a total amount of 4720 computing slots and a power capacity of 72,3k HS06. All the computing resources are centrally managed by a single batch system, Torque 4.2.10 (home made) and Maui 3.3.2 (home made). For batch processing support a server with the cluster resource manager system and task planner is set up. It should be noted that our Tier1 is configured to handle only 10-core pilots.



	2014	2015	2016	2017	2018
DISK	660	1957	4037	5590	7216
MSS	72	5478	5478	9839	10982
MSD	0	521	521	1103	1103
kHS06	18	39,4	54	63,7	72,31

Figure 1. The evolution of computing and storage capacity of the JINR Tier

As a Tier1 we are responsible not only for data processing but also data storage. We should support from raw data (including MC data) storage to AOD production and storage, and storage of physics data used in end user analysis. Based on those tasks the storage system must meet the following requirements:

- scalability up to hundreds PB,
- high availability (for data taking and processing),
- data-intensive processing with high I/O performance,
- data portability to Grid service.

So, the storage system is comprised of two parts: disk storage for data processing, and tape storage for long-time data keeping. JINR Tier1 operates a large storage infrastructure based typically on Supermicro and DELL disks servers and tape storage IBM TS3500. Currently there are ~ 7 PB of disks space (SE Disk only), ~1 PB disks buffer space for ~10 PB tape storage (SE Buffer + Tape).

The data can be accessed through standard protocols and interfaces, as defined by the WLCG projects (GridFTP and SRM, XRootD). dCache-3.2 software are used for disk storage system and Enstore 4.2.2 for tape robot. To support storage and data access systems, 8 physical and 14 virtual machines are installed.

Concerning Tier1 data exchange we can say that 157 sites worldwide read data from our SE Buffer+Tape, 140 sites write new file. Leaders are CERN, KIT(Germany), RAL (UK). 316 sites worldwide read data from our SE Disk only and 150 sites write new files. In 2018 data exchange to SE disk only was 36.50 PB (11.19 PB new files), to SE Buffer + Tape was 6.78 PB (3.38 PB new files, which is two times as much as in 2017. Figure 2 shows the monthly volumes of data exchange with the outer world.

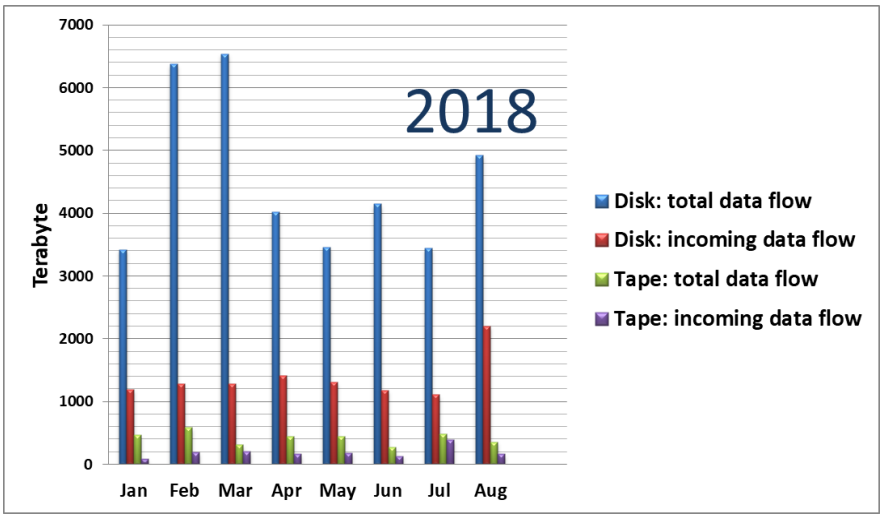


Figure 2. Monthly volumes of data exchange of Tier1 CMS with WLCG sites

The CMS PhEDEx (Physics Experiment Data Export) [20] is responsible for large-scale data transfers across the grid ensuring transfer reliability, enforcing data placement policy, and accurately reporting results and performance statistics. The system has evolved considerably since its creation in 2004, and has been used daily by CMS since then CMS PhEDEx is used for datasets transfer. Figure 3 shows the worldwide dataset transfers from T1_RU_JINR and to T1_RU_JINR since the start of operation. Information is collected from PhEDEx – CMS Data Transfers dashboard [21].

We have experienced several difficulties while using hardware RAID6 on FST pools. It provides only one-way data consistency check **(**VD health status and PD presence). There are no S.M.A.R.T data collection available, no info about physical drive (PD) temperature/health. Pre-fail condition HDD's may suddenly faults while rebuilding is already in progress and more or equal 3 faulted HDD's in one pool makes data unavailable.

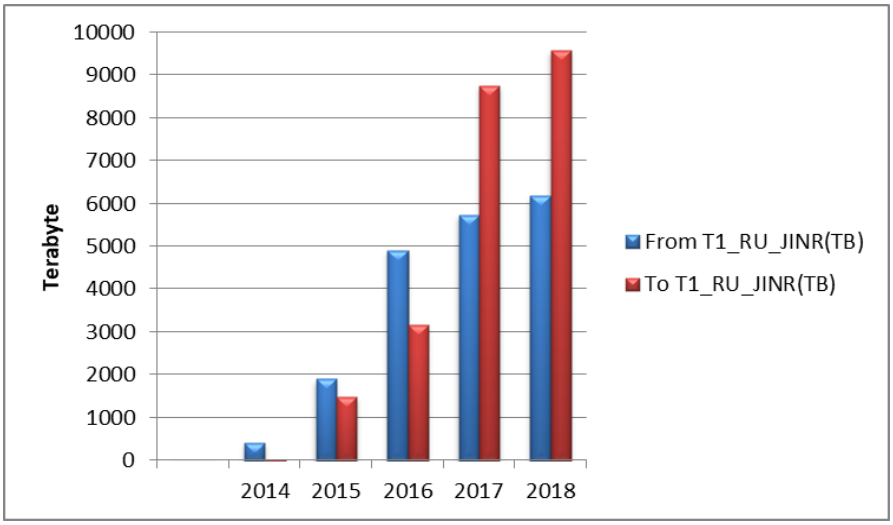


Figure 3. PhEDEx data transfers from and to T1_RU_JINR by years

With transition to ZFS RAIDZ2 we have two-way data consistency check. In HBA mode S.M.A.R.T of each installed HDD is available, and we can collect several parameters: reallocated sectors count, pending sectors count, HDD temperature, read uncorrected errors count, write uncorrected errors count, ECC correction algorithm invocations count, and overall health status. ZFS provides information about pool and devices: "State" to indicate the current health of the pool, "Status" to describe what is wrong with the pool, "Scrub" to identify the current status of a scrub operation, and "Errors" to identify known data errors or the absence of them. Thus, we can use both of that information to find faulting PDs and replace them in advance.

Rebuild process (resilvering) for 1 or 2 HDD's in ZFS is almost the same as for RAID6. In our configuration we scheduled every week pool-scrub to check data consistency. Resilvering operation also takes about 1 week for 6TB/8TB HDD's. If ZFS automatically marks another one HDD as fault, while resilvering is incomplete, it looks like faulted HDD is usually still present in pool, but in read only state. In such a case we can successfully finish previous operation before starting another one, and finally lose only several files instead of tens TB of data.

For Tier1 operation, WLCG grid environment support services are required. Some WLCG services have been installed on physical machines, the other – on virtual ones. The WLCG services are installed with software EMI-3 for compatibility with software grid-environment of WLCG. Currently, 21 services are installed. The services provide the entire infrastructure of remote work with the grid, namely:

- user and virtual organizations (VO) authorization,
- task run from VO remote services,
- the WLCG information system,
- different algorithms of remote testing and verification of the service environment on local resources.

3. Tier1 hardware and services monitoring

For a robust performance of the complex it is necessary to monitor the state of all nodes and services – from the supply system to the robotized tape library. We developed special monitoring system for our Tier1 [22]. The monitored data are collected from the wide range of hardware and software related to Tier1: cooling systems, temperature sensors, uninterruptable power supplies (UPS), computing servers, disk servers, managing services, L2 and L3 switches/routers, and tape robot. More than 850 nodes are under observation, ~8000 checks were performed in real time, and ~100 scripts were written to support this system based on Icinga2 [23].

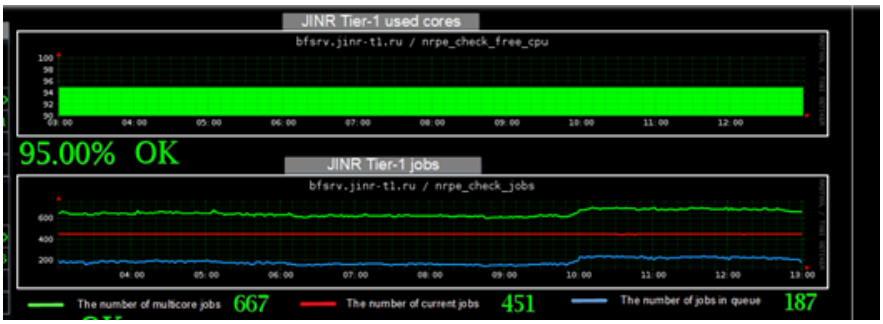


Figure 4. Screenshot from Tier1 monitoring dashboard: used cores (top) and jobs (bottom)

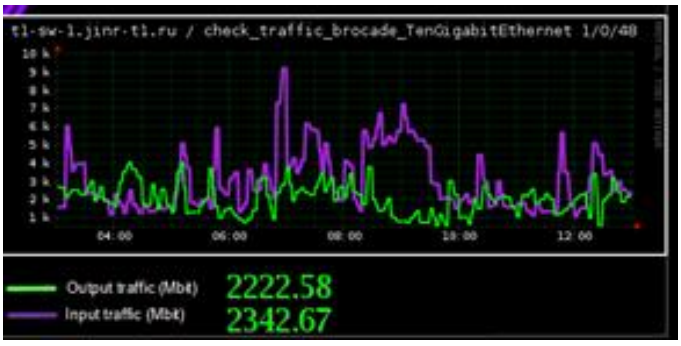


Figure 5. Screenshot from Tier1 monitoring dashboard: output (green) and input (magenta) traffic

The created monitoring system allows one not only to control the state of computing nodes and disk servers, but also to obtain information about the number of executed and waiting in the queue jobs, the Tier1 CPU load (Figure 4), local network load and incoming and outgoing data streams (Figure 5).

Apart from hardware metrics, service metrics are scattered among many internal and external systems. This information relates to data transfers, data storage, and data processing. In order to keep track of the services administrator should regularly check several dozens of web pages on common experiment and WLCG dashboards. Interpretation of data is more complex. To provide a single source of aggregated monitoring information and to perform basic analysis of data, and provide status of the system the special service monitoring system was developed [24]. Now we can collect data from four types of sources: JSON data, HTML data, Databases and command line. This collected data information is relevant only for our Tier1, stored in the database and shown on the web-page.

Other little, but helpful issue is connected to PhEDEx system, which was designed to operate mostly automatically. But sometimes, due to different reasons it requires intervention to fix errors manually. Source of information about errors is a corresponding PhEDEx webpage. Every error is a big form with source/destination site, time of assigned/start/done, PFNs to/from, transfer/detail/validate logs. In order to simplify operation *python* script was written to list important errors and provide relevant information about them. This script allows one to detect next types or errors: nsf - no such file or directory, csmm – checksum mismatch, smm – size mismatch, and uto – user timeout over. Finally the list of files in error state is produced.

4. Tier1 performance results

Since commissioning, the JINR Tier1 has steadily increased its productivity and maintained the level of availability and reliability required by the CMS experiment. The Figure 6 shows JINR Tier1 (blue) monthly reliability results compared to the average Tier1 reliabilities (orange) as well as with the WLCG target for site reliability (yellow area), which is set to 97% since 2009, according to the WLCG MoU.

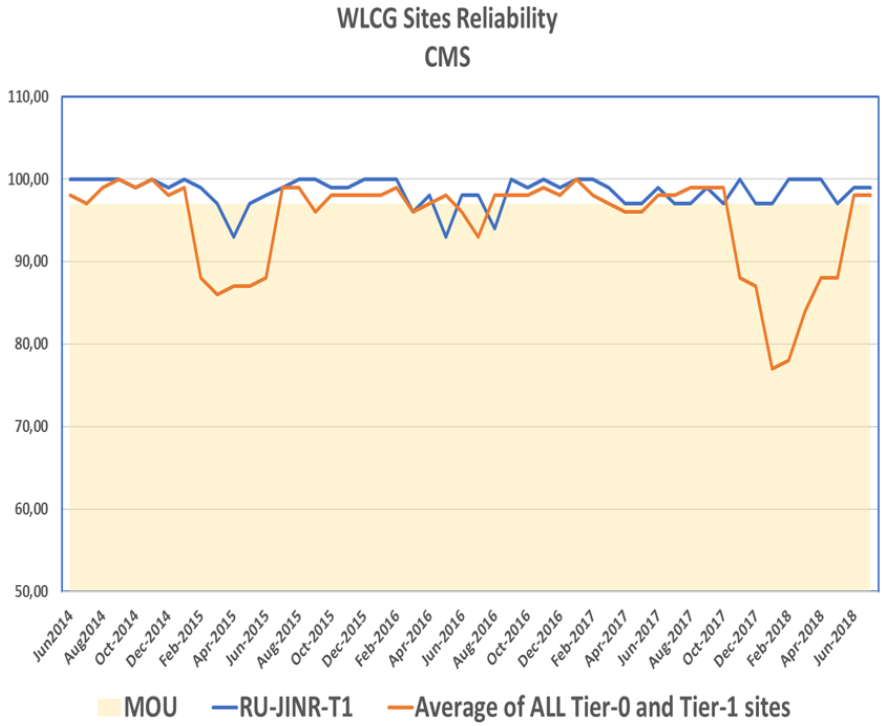


Figure 6. WLCG Tier1 sites reliability results from 2014 to 2018

The contribution of the Tier1 center in JINR to the processing of CMS experiment data for the last 5 years is shown in Figure 7 as a percentage of the total number of jobs and processed events by all Tier1 sites. The percentage of jobs has been increasing during the year by year and reached about 13% of total CMS jobs and about 20% of good events processed. In total, since the beginning of operation, 27 722 950 jobs have been completed and 308 560 399 956 events processed on our site.

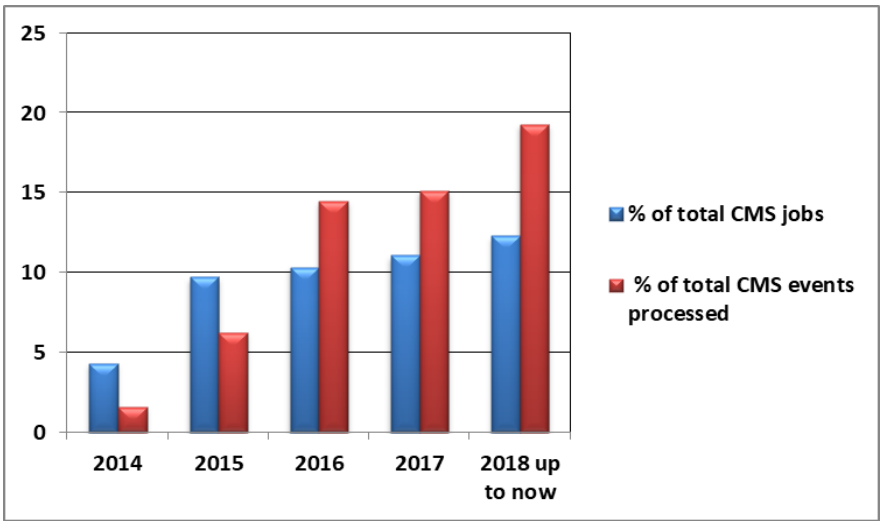


Figure7. Percentage of total CMS jobs and events processed by JINR Tier1

Organized data processing activities are carried on at the Tier1 centres by the Data Operations team. These include reprocessing of raw data, production of AOD, analyzing, large scale physics skims and processing. During this years our Tier1 processed 281 555 814 018 events by analysis jobs, 10 042 089 961 events were reprocessed, 5 569 434 071 events were processed through production activity, etc. Figure 8 shows the distribution of completed jobs by activities.

Figure 9 shows the number of CMS events processed in Million events at our Tier1 in August 2018 and figure 10 shows the number of jobs processed in the same period of time [25].

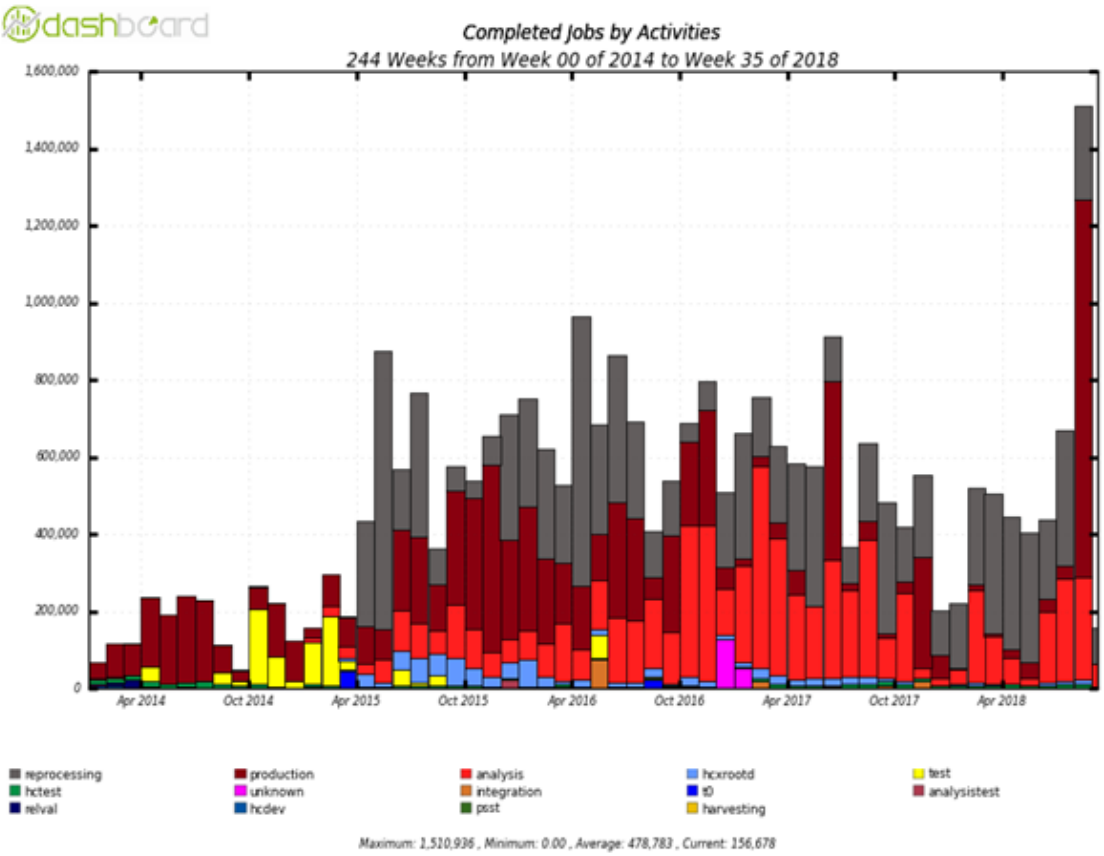


Figure 8. Completed jobs by activities at the JINR Tier1

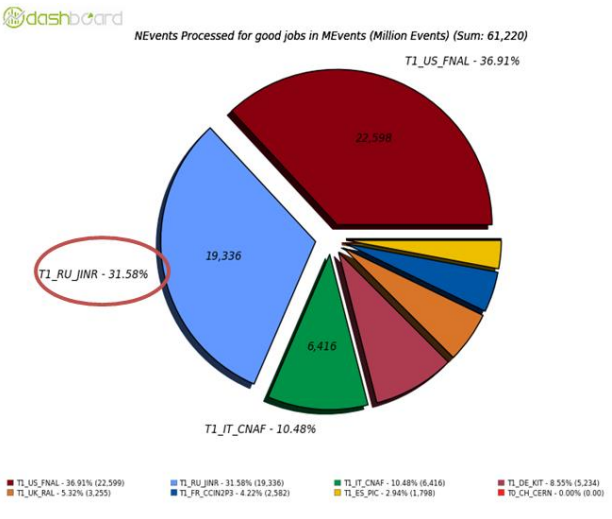


Figure 9. Number of CMS events processed at the JINR Tier1 in August 2018

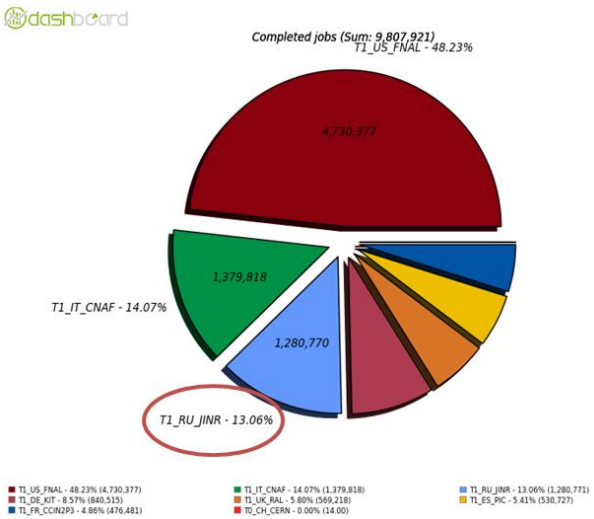


Figure 10. Number of CMS jobs completed at the JINR Tier1 in August 2018

5. Conclusion

During the last five years the JINR Tier1 centre for CMS experiment at CERN progressively increased its scale and improved its reliability. It creates conditions for physicists from JINR and its JINR Member States, RDMS-CMS collaboration for a full-scale participation in processing and analysis of data of the CMS experiment on the Large Hadron Collider.

The invaluable experience of launching the Tier1 center will be used for creating a system of storage and data processing of megaproject NICA and other scale projects of the JINR-participating countries.

References

[1] CERN European Organization for Nuclear Research, URL https://home.cern

[2] WLCG Worldwide LHC Computing Grid, URL http://wlcg.web.cern.ch/

[3] JINR Joint Institute for Nuclear Research, URL http://www.jinr.ru/

[4] CMS Compact Muon Solenoid experiment, URL https://cms.cern/detector

[5] LHC – The Large Hadron Collider, URL http://lhc.web.cern.ch

[6] N.S. Astakhov, A.S. Baginyan, A.I. Balandin, S.D. Belov, A.G. Dolbilov, A.O. Golunov, N.I. Gromova, I.S. Kadochnikov, I.A. Kashunin, V.V. Korenkov, V.V. Mitsyn, I.S. Pelevanyuk, S.V. Shmatov, T.A. Strizh, V.V. Trofimov, N.N. Voitishin, V.E. Zhiltsov, Proc. of the XXVI International Symposium on Nuclear Electronics & Computing (NEC'2017), CEUR-WS.org/ Vol-2023/68-74-paper-10.pdf, (2017)

[7] 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, J.Phys.Conf.Ser. 513, 062041, (2014)

[8] ATLAS Experiment, URL http://atlas.cern/

[9] National Research Center "Kurchatov Institute", URL http://www.nrcki.ru/

[10] ALICE A Large Ion Collider Experiment, URL http://aliceinfo.cern.ch/Public/Welcome.html

[11] LHCb Large Hadron Collider beauty experiment, URL http://lhcb-public.web.cern.ch/lhcb-public/

[12] RDMS: Russia and Dubna Member States CMS Collaboration, URL http://rdms-cms.jinr.ru/

[13] CMS: The Computing Project Technical Design Report, URL http://cdsweb.cern.ch/ record/838359/ files/lhcc-2005-023.pdf

[14] The Large Hadron Collider Optical Private Network URL http://lhcopn.web.cern.ch/lhcopn/

[15] The Large Hadron Collider Open Network Environment URL http://lhcone.web.cern.ch/

[16] S. Baginyan, A. Dolbilov, V. Korenkov, Network for datacenter Tier 1 at JINR for experiment CMS (LHC), T-Comm Vol.10, No.1, 25-19 (2016)

[17] N.S. Astakhov, S.D. Belov, P.V. Dmitrienko, A.G. Dolbilov, I.N. Gorbunov, V.V. Korenkov, V.V. Mitsyn, S.V. Shmatov, T.A. Strizh, E.A. Tikhonenko, V.V. Trofimov, V.E. Zhiltsov, CMS Tier-1 at JINR, Proc. NEC'2013, p.5 (2014)

[18] N.S. Astakhov, A.S. Baginyan, S.D. Belov, A.G. Dolbilov, A.O. Golunov, I.N. Gorbunov, N.I. Gromova, I.S. Kadochnikov, I.A. Kashunin, V.V. Korenkov, V.V. Mitsyn, I.S. Pelevanyuk, S.V. Shmatov, T.A. Strizh, E.A. Tikhonenko, V.V. Trofimov, N.N. Voitishin, V.E. Zhiltsov, JINR Tier-1 centre for the CMS experiment at LHC, Particles and Nuclei, Letters, v.13, 5(203).1103-1107 (2016)

[19] N.S. Astakhov, A.S. Baginyan, S.D. Belov, A.G. Dolbilov, A.O. Golunov, I.N. Gorbunov, N.I. romova, I.S. Kadochnikov, I.A. Kashunin, V.V. Korenkov, V.V. Mitsyn, I.S. Pelevanyuk, S.V. Shmatov, T.A. Strizh, E.A. Tikhonenko, V.V. Trofimov, N.N. Voitishin, V.E. Zhiltsov, Tier-1 CMS at JINR: Status and Perspective. CEUR Workshop Proceedings V1787, P.1-14 (2016)

[20] R. Egeland, T. Wildish and S. Metson. Data transfer infrastructure for CMS data taking, PoS(ACAT08)033 pdf (2008)

[21] PhEDEx – CMS Data Transfers, URL https://cmsweb.cern.ch/phedex/prod/Info::Main

[22] I.A. Kashunin, A.G. Dolbilov, A.O. Golunov, V.V. Korenkov, V.V. Mitsyn, T.A. Strizh, Proc. of the 7th International Conference Distributed Computing and Grid-technologies in Science and Education, CEUR-ws.org/Vol-1787/256-263-paper-43.pdf, (2016)

[23] Icinga, URL https://icinga.com/

[24] I.S. Kadochnikov, I.S. Pelevanyuk JINR Tier1 service monitoring system: Ideas and Design, Proc. of the 7th International Conference Distributed Computing and Grid-technologies in Science and Education, http://ceur-ws.org/Vol-1787/275-278-paper-46.pdf, (2016)

[25] CMS Dashboard URL http://dashb-cms-jobsmry.cern.ch/dashboard