EPJ Web of Conferences **214**, 03009 (2019) https://doi.org/10.1051/epjconf/201921403009
CHEP 2018

The JINR distributed computing environment

Vladimir Korenkov, *Andrei* Dolbilov, *Valeri* Mitsyn, *Ivan* Kashunin, *Nikolay* Kutovskiy, *Dmitry* Podgainy, *Oksana* Streltsova, *Tatiana* Strizh*, *Vladimir* Trofimov, and *Peter* Zrelov

Joint Institute for Nuclear Research, Laboratory of Information Technologies, 141980 Dubna, Moscow reg., Russia

Abstract. Computing in the field of high energy physics requires usage of heterogeneous computing resources and IT, such as grid, high performance computing, cloud computing and big data analytics for data processing and analysis. The core of the distributed computing environment at the Joint Institute for Nuclear Research is the Multifunctional Information and Computing Complex. It includes Tier1 for CMS experiment, Tier2 site for all LHC experiments and other grid non-LHC VOs, such as BIOMED, COMPASS, NICA/MPD, NOvA, STAR and BESIII, as well as cloud and HPC infrastructures. A brief status overview of each component is presented. Particular attention is given to the development of distributed computations performed in collaboration with CERN, BNL, FNAL, FAIR, China, and JINR Member States. One of the directions for the cloud infrastructure is the development of integration methods of various cloud resources of the JINR Member State organizations in order to perform common tasks, and also to distribute a load across integrated resources. We performed cloud resources integration of scientific centers in Armenia, Azerbaijan, Belarus, Kazakhstan and Russia. Extension of the HPC component will be carried through a specialized infrastructure for HPC engineering that is being created at MICC, which makes use of the contact liquid cooling technology implemented by the Russian company JSC "RSC Technologies". Current plans are to further develop MICC as a center for scientific computing within the multidisciplinary research environment of JINR and JINR Member States, and mainly for the NICA mega-science project.

1 Introduction

The Joint Institute for Nuclear Research (JINR) [1] is an international intergovernmental organization and is developing as a large multidisciplinary international scientific center incorporating basic research in the field of modern nuclear physics, development and application of cutting edge technologies, and university education in the relevant fields of knowledge. Presently JINR has 18 Member States and six associate members. JINR possesses an information-computational complex. The uninterrupted functioning of all its elements at the acceptable level is mandatory for the fulfillment of the JINR scientific

* Corresponding author: strizh@jinr.ru

research programmes. Support of this fully functional infrastructure is the major task of the Laboratory of Information Technologies (LIT). The future development of the basic computer infrastructure of the JINR is necessary for the success of the ambitious targets of the research carried out at JINR and collaborating organizations both in frames of the JINR research programmes, in particular the NICA (Nuclotron-based Ion Collider fAcility) mega-project [2], and within cooperation with the leading scientific and research centers (CERN, FAIR, BNL, etc.). In the last few years, in the context of various works on organization of the computing for NICA, commissioning of the Tier1 center for the CMS experiment [3], as a part of the Worldwide LHC Computing Grid infrastructure (WLCG) [4], implementations of a cloud computing structure and of a cluster for hybrid computations, the information-computing environment of JINR evolved in a set of stand-alone structures that have a common engineering and networking infrastructures. Multifunctional Information and Computing Complex (MICC) [5] of JINR currently has the following basic components:

- the Central Information and Computing Complex (CICC) of JINR with in-house build compute and mass storage elements,
- Tier2 for all experiments at the Large Hadron Collider (LHC) and other virtual organizations (VOs) in the grid environment [6],
- Tier1 for CMS experiment,
- heterogeneous platform HybriLIT for High Performance Computing (HPC) [7],
- the cloud infrastructure [8].

MICC JINR resources are used for storing, analyzing, and simulating data in the fields of particle physics, nuclear physics and condensed matter physics. The grid center resources of the MICC JINR are a part of the WLCG infrastructure, developed for the LHC experiments.

2 Networking

One of the most important components of JINR and MICC providing access to resources and the possibility to work with the big data is a network infrastructure that uses one lambda (single frequency) of 100 Gbps, as the main communication channel, and two lambdas (two frequencies) of 10 Gbps each. The external network of JINR includes an external overlay network LHCOPN (Large Hadron Collider Optical Private Network) [9] (JINR – CERN) passing through MGTS-9 in Moscow, Budapest and Amsterdam to link the centers of Tier0/Tier1 and our Tier1, and another external overlay network LHCONE (LHC Open Network Environment) [10] of the same route which is intended for Tier2 center at JINR; direct communication links based on EN-VRF technology with the collaboration of research centers RUHEP (Gatchina, NRC Kurchatov Institute, Protvino) and with networks RUNNet, RASnet. The IPv6 routing for Tier1 and Tier2 centers was implemented. There is also a backup communication channel with a bandwidth of 20 Gbps.

3 Central information and computing complex

The central information and computing complex at LIT allows carrying out computations, including parallel ones, outside the grid-environment. This is required by NOvA [11], BESIII [12], NICA/MPD and other experiments, as well as by the local users of the JINR laboratories. The JINR and the grid users have access to all computer facilities through a unified batch processing system. dCache [13] and XRootD [14] storage systems ensure work with data both for local JINR users and for the WLCG users and collaborations. All the storage systems are built with the help of hardware data protection mechanism based on RAID6 and a software mechanism RAIDZ2 which is not inferior to the reliability of the

RAID6 hardware. We start a collaboration within the WLCG Datalake R&D project [15] and a prototype is based on EOS [16] storage system. As the next step, deployment of the common EOS based data storage for MICC components is already in progress. Its space is 3,740TB and usable space is 1,870TB (for we keep two replicas of each file on the EOS based storage).

4 Grid infrastructure

We have two WLCG sites at the JINR LIT: Tier1 (T1_RU_JINR) provides one of 7 dedicated real-time computing facilities for experimental data obtained from the CMS detector at CERN, and Tier2 (JINR-LCG2) provides one of ~170 dedicated real-time computing facilities for experimental data obtained from the ATLAS, ALICE, CMS, LHCb detectors, and other VOs. We participate in some other WLCG activities such as software development for ATLAS, ALICE, CMS; WLCG Dashboard (WLCG Google Earth monitoring, Global data transfer monitoring; local and global monitoring of Tier3 centers; XRootD monitoring, etc.); NoSQL data storage integration (Hadoop, ElasticSearch, etc.); integration of Grid, Clouds and HPC; PanDA WMS Development (Pilot2, Harvester) [17]; COMPASS Production System [18]; GENSER & MCDB [19], etc.

4.1 Tier1 for CMS

Data processing system at JINR Tier1 for CMS consists of 275 64-bit machines: 2 x CPU, 6-10 core/CPU, which is in total 4,720 core for the batch system. All servers are Supermicro blades.

The storage system supports disks and long-term data storage on tapes. One of the dCache instances is used with disk servers for online storage and fast access to data. The second dCache system includes disk servers and a tape robot. The total usable capacity of disk-only servers is 7.2 PB (Supermicro and Dell), total disk buffer space of Mass Storage System (MSS) is 1.1 PB, and tape robot IBM TS3500 is 9 PB. The Torque 4.2.10 / Maui 3. 3.2 (custom build) software is used as a resource manager and task scheduler respectively. PhEDEx [20] is used for our site as a data-placement management tool. The dCache-3.2 software is used for disk storage system and Enstore 4.2.2 for tape robot.

The standard program stack of the WLCG is deployed for data processing: 2 x CREAM, 4 x ARGUS, BDII top, BDII site, APEL parsers, APEL publisher, EMI-UI, 220 x EMI-WN + gLExec-wn, 4 x FTS3, LFC, WMS, L&B, glite-proxyrenewal.

Figure 1 shows the contribution of the world Tier1 centers in the processing of the CMS

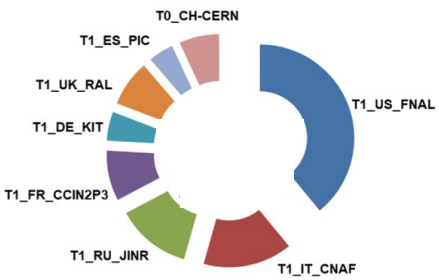

Fig. 1. Number of events processed for good jobs in percentage from 2015-04-01 to 2018-05-31 at all WLCG Tier1 sites for CMS experiment.

experimental data (in a percentage of millions of processed events) for the period from 2015-04-01 to 2018-05-31. The Tier1 center for CMS in JINR demonstrates stable operation throughout the entire period after its launch in full operation in 2015. T1_RU_JINR site ranks third in the world in terms of performance.

4.2 Tier2

The Tier2 at JINR allows data processing for all four LHC experiments (ALICE, ATLAS, CMS, LHCb) as well as supports a number of VOs that are not included in the LHC (BESIII, BIOMED, COMPASS, FUSION, MPD, NOvA, STAR, etc.). Computational resources of the Tier2 center consist of 4,128 cores (typically SuperMicroBlade, SuperMicroTwin2, Dell FX). Data storage system is using dCache and XRootD software. One of the dCache instance is used by the CMS and ATLAS. The second dCache is used by the JINR users and user groups (as well as for the NICA tasks). Besides, this instances is used to store data of several third-party experiments (BIOMED, BESIII, FUSION). One XROOTD installation is used by ALICE. The total capacity of the storage system is 2,929 TB.

WLCG grid-environment is supported by special servers installed for the operation of the VOs. Part of the WLCG services is deployed on physical machines, some of them - on virtual ones. Currently, 23 WLCG services are in operation. They provide access to entire resources installed in the Tier2 at JINR for remote work with the Grid. There are five settings of user interface (UI) for running jobs in a distributed grid-environment. The OSG HTCondor computing element is integrated into the Tier2 center infrastructure. It provides a way for VO STAR to process data using our Tier2 resources.

The main users of the JINR grid resources are virtual organizations of all LHC experiments. During 2018, the Tier2 site performed 2,448,777 jobs, and normalized CPU time was 127,105,363 HS06 hours.

5 Cloud infrastructure

The JINR cloud infrastructure is built on the basis of the OpenNebula software [21]. The main purposes are to increase the efficiency of hardware utilization and to improve IT-services management. Two types of virtualization are used: OpenVZ containers [22] - CTs (Linux only) and KVM virtual machines [23] - VMs (any OS) are used. Ceph [24] block device is used as storage back-end for KVM VM images [25]. Web-based graphical user interface (web-GUI) and command line one are provided for user interactions with the JINR cloud. A JINR single sign-on (SSO) service is used for authentication in the cloud web-GUI. RSA/DSA keys or Kerberos credentials are used for VM/CT access via ssh protocol.

Cloud resources are used as VMs & CTs for JINR users and as computational resources for Baikal-GVD [26], BESIII, Daya Bay [27], JUNO [28], NOvA experiments, also as testbeds for development and R&D in IT. The new fault-tolerant architecture of the JINR cloud is designed on the basis of the Raft algorithm implemented in the new version of OpenNebula.

One of the most important trends in cloud technologies is the development of methods for integrating various cloud infrastructures. In order to join the cloud resources of partner organizations from the JINR Member States to solve common tasks, as well as to balance a peak load across them, a special driver was developed [29]. It allows integration of the JINR cloud with partner ones that are either built on OpenNebula or support the Open Cloud Computing Interface. The geography of organizations which share a part of their resources via distributed cloud infrastructure is presented in Figure 2. A new challenge is integration

JINR Member State organizations clouds and the Supercomputer into unified distributed computational environment.

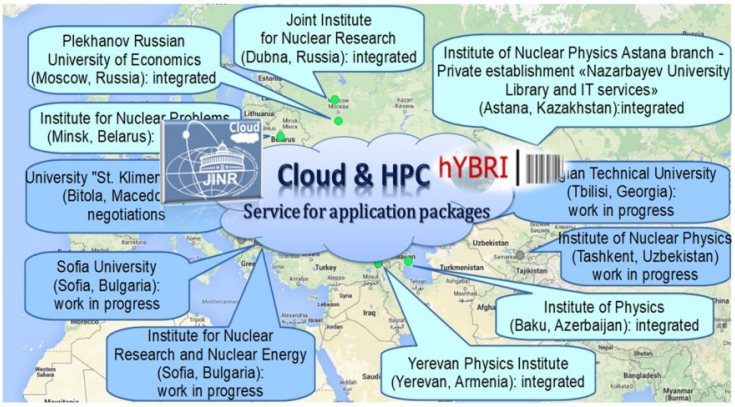

Fig. 2. Map with the distributed cloud infrastructure participants**.**

6 HPC – HybriLIT platform

In 2014, HybriLIT heterogeneous computing cluster was put into operation at the MICC, containing at that time the latest computing architectures: not only multicore CPUs, but also computing accelerators - Intel Xeon Phi coprocessors and GPU by NVIDIA. The chosen heterogeneous architecture of the cluster is related to the need to provide computations to a wide range of problems that JINR scientific groups face with, and requiring massive parallel calculations. The cluster meets the following requirements: the ability to dynamically expand the cluster by adding new computing nodes; provide the ability to synchronously update or change the software on existing and commissioned computing nodes; quick installation of nodes and quick recovery of cluster nodes after failures and reboots. The specialized software and information environment are developed for the effective use of the cluster. It includes a monitoring system, services for working with users, services for joint application development. The “GOVORUN” supercomputer [30] has become a natural development of the heterogeneous cluster HybriLIT. The total performance is increased by 10 times (figure 3).

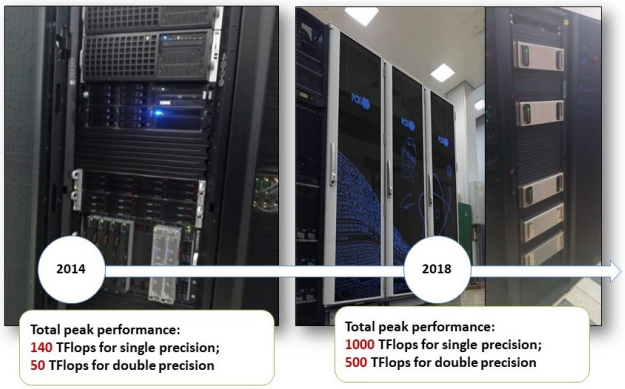

Fig. 3. From HybriLIT cluster to a HybriLIT platform, the performance increased by 10 times.

It is important to note that the basic idea is preserved - heterogeneity of the computing platforms: the supercomputer contains both graphics accelerators by NVIDIA, and specialized Intel Xeon Phi 7290 (KNL) and Intel Xeon Gold 6154 (Skylake) processors. The supercomputer is included in a single software environment allowing users to perform calculations on all parts of the new platform without migrating their data, settings, etc. Thus, the HybriLIT heterogeneous platform is developed: it consists of training and test polygon (previously the cluster HybriLIT) and supercomputer "GOVORUN". One of the main features of the CPU - components of the supercomputer is its implementation on the basis of a "hot water" liquid cooling technology by RSC [31]. This solution has a high energy efficiency (PUE = 1.03), and high-density architecture of servers in the rack with direct liquid cooling providing a packing density of more than 150 servers per rack.

Currently, the supercomputer is used for theoretical research in the field of quantum chromodynamics, molecular dynamics, and it is planned to use the supercomputer to simulate computing for the NICA megaproject.

7 Monitoring

All major components of MICC rely on the network and computing infrastructure. It is important to monitor all of the components on three levels: hardware, network, and service. Currently, there are many monitoring systems applied for MICC. These systems are used by different groups and are built on different technologies.

The monitoring system for Tier1, Tier2 and CICC farm at JINR that is based on Nagios3.5 [32] and is currently moved to Icinga2 [33], provides real-time information about: work nodes, disk servers, tape robot, network equipment, UPS, and cooling system. It also can be used for creating network maps and network equipment load maps, for drawing state tables and different plots. The system allows one to observe the whole computing complex state and send system alerts to users via e-mail, sms, etc. in a real time mode. About 1,000 nodes are under monitoring and more than 8,000 checks in real time are performed. Monitoring results are displayed on the web-page in the form of a dashboard.

The monitoring for JINR cloud consists of several parts. The first part is based on Nagios and it monitors the state of every physical host, which VMs are running on. The second part is responsible for virtual machines and hosts usage monitoring. A special monitoring system is developed to collect performance metrics related to virtual machines and hypervisors: CPU, RAM and network usage. Such data is collected by Icinga2 agents installed on every host with hypervisor. Grafana is used for visualization.

A special monitoring system of heterogeneous cluster HybriLIT is developed. A sensor written in C++ is responsible for collecting metrics on servers and sending them to a message queue. Data from the message queue go to MySQL and clients if they are connected to the web-interface. Part of the monitoring data for users is taken from the Slurm database. The biggest advantage of this monitoring is that it provides present time information about HybriLIT cluster which allows using it as one of the debugging tools.

All monitoring systems are mostly independent despite the fact that some systems collect the same monitoring data. Their independence makes it difficult to see the overall effectiveness and bottlenecks of MICC because the data are scattered among many systems. The role of the multi-level monitoring system for MICC is to unite the existing systems and solve the problem of providing high level information about the whole computing complex and its services [34].

8 Conclusion

The further development of the JINR distributed environment is aimed at creating a technological frame which enables scientific research mainly for the NICA mega-science project at JINR to be conducted in unified information and computing environment incorporating a lot of technological solutions, concepts and practices. Such an environment has to combine supercomputer (heterogeneous), grid- and cloud-complexes and systems with the aim at providing the best approaches for the solution of different kinds of scientific and applied tasks. Essential requirements for this environment are scalability, interoperability and adaptability to new technical solutions. Transition to distributed experimental data processing and storage based on grid and DataLake technologies is a necessary condition for the successful participation of the physicists of JINR and JINR Member State institutes in the NICA and LHC experiments.

References

1. Joint Institute for Nuclear Research web-portal: http://www.jinr.ru/
2. NICA (Nuclotron-based Ion Collider fAcility web-portal: http://nica.jinr.ru/
3. N.S. Astakhov, A.S. Baginyan, S.D. Belov, *et al.*, Phys. Part. Nuclei Lett. **13**, 714, https://doi.org/10.1134/S1547477116050046 (2016)
4. The Worldwide LHC Computing Grid (WLCG) web-portal: http://wlcg.web.cern.ch/
5. The JINR Multifunctional Information and Computing Complex web-portal: https://micc.jinr.ru/
6. N.S. Astakhov, A.S. Baginyan, A.I. Balandin, *et al.*, *Proc. of the XXVI International Symposium on Nuclear Electronics & Computing (NEC'2017)*, CEUR-WS.org/ Vol-2023/68-74-paper-10.pdf (2017)
7. Gh. Adam, V.V. Korenkov, D.V. Podgainy, *et al.*, *Proc. of the XXVI International Symposium on Nuclear Electronics & Computing (NEC'2017)*, CEUR-WS.org/ Vol-2023/351-356-paper-57.pdf (2017)
8. A.V. Baranov, N.A. Balashov, N.A. Kutovskiy, R.N. Semenov, Phys. Part. Nuclei Lett., **13**, 672 (2016)
9. LHCOPN (Large Hadron Collider Optical Private Network) web-portal: http://lhcopn.web.cern.ch/lhcopn
10. LHCONE (LHC Open Network Environment) web-portal: http://lhcone.web.cern.ch/
11. D.S. Ayres, G.R. Drake, M.C. Goodman, *et al.*, NOvA Technical Design Report, DOI: 10.2172/935497 (2007)
12. M. Ablikim, Z.H. An, J.Z. Bai, *et al.*, Nucl. Instrum. Methods Phys. Res. A, **614**, 3, 345-399, DOI: 10.1016/j.nima.2009.12.050 (2010)
13. M. Ernst, P. Fuhrmann, M. Gasthuber, *et al.*, *Proceedings of CHEP 2001*, p. 757, China: Science Press (2001)
14. XRootD web-portal: http://xrootd.org/
15. WLCGDatalakes: https://twiki.cern.ch/twiki/bin/view/WLCGDatalakes/WebHome
16. EOS Open Storage web-portal: http://eos.web.cern.ch/
17. F.H. Barreiro Megino, K. De, A. Klimentov, *et al.*, J. Phys.: Conf. Ser., **898**, 052002, DOI https://doi.org/10.1088/1742-6596/898/5/052002 (2017)
18. A.Sh. Petrosyan, *Proc. of the XXVI International Symposium on Nuclear Electronics & Computing (NEC'2017)*, CEUR-WS.org/ Vol-2023/234-238-paper-37.pdf (2017)
19. A. Karneyeu, M. Kirsanov, D. Konstantinov, *et al.*, J. Phys.: Conf. Ser., **331,** 032025, doi:10.1088/1742-6596/331/3/032025 (2011)

20. R. Egeland, T. Wildish, S. Metson, *Proc. of XII Advanced Computing and Analysis Techniques in Physics Research,* PoS(ACAT08)033, doi:10.22323/1.070.0033 (2009)
21. R. Moreno-Vozmediano, R.S. Montero, I.M. Llorente, IEEE Computer, **45**, 65-72 (2012)
22. K. Kolyshkin, Virtualization Comes in More than One Flavor, *Virtualization Magazine* (Ulitzer, Inc., 2007)
23. Kernel-based Virtual Machine web-portal, https://www.linux-kvm.org
24. Ceph storage web-portal, https://ceph.com
25. N. Balashov, A. Baranov, S. Belov, *et al.*, *Proc. of the XXVI International Symposium on Nuclear Electronics & Computing (NEC'2017)*, CEUR-WS.org/Vol-2023/88-91-paper-13.pdf (2017)
26. A.D. Avrorin, A.V. Avrorin, V.M. Aynutdinov, *et al.*, EPJ Web Conf. **136**, 04007, https://doi.org/10.1051/epjconf/201713604007 (2017)
27. Jun Cao, Kam-Biu Luk, Nucl. Phys. B, **908**, 62-73, https://doi.org/10.1016/j.nuclphysb.2016.04.034 (2016)
28. G. Ranucci and JUNO Collaboration, J. Phys.: Conf. Ser., **888,** 012022, https://doi.org/10.1088/1742-6596/888/1/012022 (2017)
29. A.V. Baranov, V.V. Korenkov, V.V. Yurchenko, *et al.*, Computer Research and Modeling, **8**, 3, 583-590 (2016)
30. SUPERCOMPUTER "GOVORUN", http://hlit.jinr.ru/en/about_govorun_eng/
31. RSC Group. http://www.rscgroup.ru/en/company
32. I.A. Kashunin, A.G. Dolbilov, A.O. Golunov, *et al.*, *Proc. of the 7th International Conference Distributed Computing and Grid-technologies in Science and Education (2016)*, CEUR-ws.org/Vol-1787/256-263-paper-43.pdf (2016)
33. Icinga2: https://icinga.com/products/icinga-2/
34. A.S. Baginyan, N.A. Balashov, A.V. Baranov, *et al.*, *Proc. of the XXVI International Symposium on Nuclear Electronics & Computing (NEC'2017)*, CEUR-ws.org/Vol-2023/226-233-paper-36.pdf (2017)