CURRENT STATUS OF THE MICC: AN OVERVIEW

A. Baginyan, A. Balandin, N. Balashov, A. Dolbilov, A. Gavrish, A. Golunov, N. Gromova, I. Kashunin, V. Korenkov, N. Kutovskiy, V. Mitsyn, I. Pelevanyuk, D. Podgainy, O. Streltsova, T. Strizh[a], V. Trofimov, A. Vorontsov, N. Voytishin, M. Zuev

Meshcheryakov Laboratory of Information Technologies, Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region, 141980, Russia

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

The Multifunctional Information and Computing Complex (MICC) of the Joint Institute for Nuclear Research (JINR) runs Tier-1, which supports the NICA experiments and the LHC CMS experiment, Tier-2, which supports all LHC experiments, as well as the NICA experiments and other HEP experiments with JINR's participation, cloud computing for neutrino physics experiments (Baikal-GVD, JUNO, DANS, etc.), as well as for the JINR Member States' clouds, the "Govorun" supercomputer for all JINR programs, as well as for the NICA experiments and research in the field of machine learning and quantum computing. This activity is aimed at ensuring the further development of the network, information and computing infrastructure of JINR for the research and production activities of the Institute and its Member States on the basis of state-of-the-art information technologies in accordance with the JINR Seven-Year Plan of development for 2017-2023. This paper describes the current state of the MICC.

Keywords: MICC, NICA, WLCG, grid, Tier1, Tier2, cloud, HPC, distributed computing

Andrey Baginyan, Anton Balandin, Nikita Balashov, Andrey Dolbilov, Andrey Gavrish, Alexey Golunov, Natalia Gromova, Ivan Kashunin, Vladimir Korenkov, Nikolay Kutovskiy, Valery Mitsyn, Igor Pelevanyuk, Dmitry Podgainy, Oxana Streltsova, Tatiana Strizh, Vladimir Trofimov, Alexey Vorontsov, Nikolay Voytishin, Maxim Zuev

1. Introduction

Starting from 2017, the computing facilities of the JINR Meshcheryakov Laboratory of Information Technologies (MLIT) operate within the project “Multifunctional Information and Computing Complex” (MICC) [1]. The main aim of the project is the further development of the network, information and computing infrastructure of JINR for the research activities of the Institute and its Member States on the basis of state-of-the-art information technologies in accordance with the JINR Seven-Year Plan of development for 2017-2023.

The MICC is considered as a unique basic facility of JINR and plays a decisive role in scientific research, which requires modern computing power and storage systems. The uniqueness of the MICC is ensured by the combination of all modern information technologies from the network infrastructure with a bandwidth of 2x100 Gbit/s to 4x100 Gbit/s, the distributed data processing and storage system based on grid technologies and cloud computing, the hyperconverged liquid-cooled high performance computing infrastructure for supercomputer applications. Multifunctionality, high reliability and availability in a 24x7x365 mode for computing, scalability and high performance, a reliable data storage system, information security and an advanced software environment are the main requirements that the MICC meets. The JINR computer infrastructure includes the IT ecosystem for the NICA [2] project experiments (BM@N, MPD, SPD), which, thanks to grid technologies (DIRAC Interware [3]), embraces all computing components and storage systems; the Tier1 grid site for the CMS experiment at the LHC [4]; Tier2, which provides support for the experiments at the LHC (ATLAS, ALICE, CMS), FAIR (CBM, PANDA) and other large-scale experiments, as well as support for users of the JINR Laboratories and participating countries; the integrated cloud environment of the participating countries to support users and experiments (NICA, ALICE, BESIII, NOvA, Baikal–GVD, JUNO, etc.); the HybriLIT [5] platform with the “Govorun” supercomputer as the main resource for HPC (fig. 1).

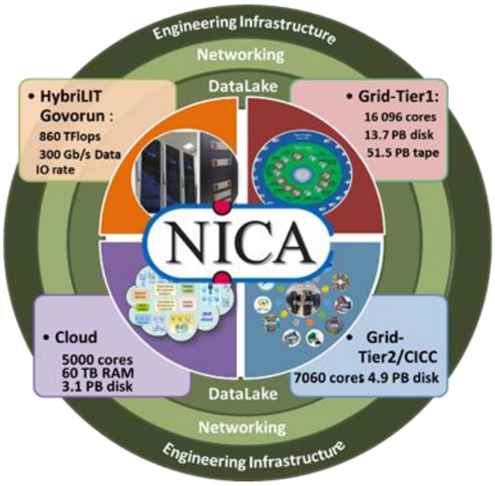

Figure 1. Diagram of the MICC structure

2. Engineering infrastructure

The MICC computing facilities at MLIT JINR are hosted in a single computing hall of 900 m2 of floor-space on the 2nd floor of the MLIT building. It was built in the late 1970s for hosting mainframe computers. After a number of refurbishments throughout the years, it currently consists of eight separate IT equipment modules (fig. 2) with 2 MW power and slightly different features:

- Module 1 and module 2: 22.55 m2 of floor-space each, 33 server racks and 20 kW power per rack;
- Tier1 module: 29.33 m2 of floor-space, 16 server racks and 35 kW power per rack;
- Tape library space: 13 m2 of floor-space, two installations IBM TS3500 and IBM TS4500 with a total capacity of 51 PB;
- “Govorun” supercomputer: 1.97 m2 of floor-space, 4 racks and 100 kW power per rack;
- Small module that hosts critical services of a standard business computing type (administrative systems and databases, etc.);
- Module 4: 36.12 m2 of floor-space, 20 server racks and 35 kW power per rack;
- Network equipment module that hosts the main network services for the MICC, JINR local and wide area networking.

The six modules mainly host the “physics processing” of different experiments and use different technologies for computing such as grid, cloud and HPC.

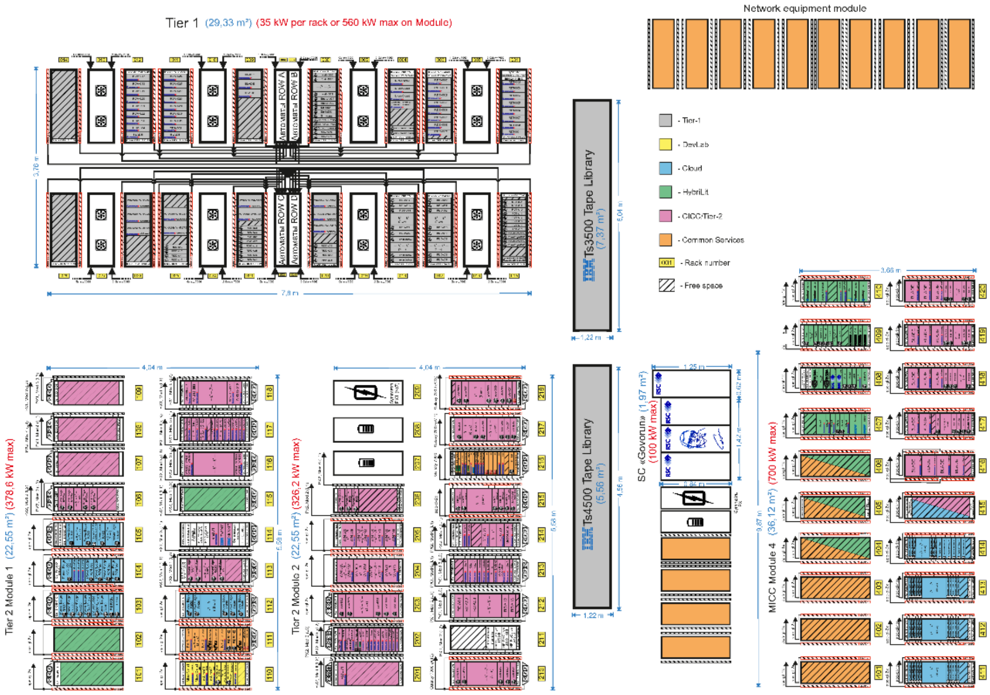
Tier 1 (29,33 m²) (35 kW per rack or 560 kW max on Module)
Автоматы ROW A
Автоматы ROW B
Автоматы ROW C
Автоматы ROW D
3,76 m
7,8 m
Network equipment module
IBM Ts3500 Tape Library (7,37 m²)
6,04 m
1,22 m
- Tier-1
- DevLab
- Cloud
- HybriLit
- CICC/Tier-2
- Common Services
001 - Rack number
- Free space
3,66 m
Tier 2 Module 1 (22,55 m²) (378,6 kW max)
4,04 m
Tier 2 Module 2 (22,55 m²) (326,2 kW max)
4,04 m
IBM Ts4500 Tape Library (6,58 m²)
1,22 m
SC «Govorun» (1,97 m²) (100 kW max)
1,25 m
MICC Module 4 (36,12 m²) (700 kW max)

Figure 2. Layout of the MICC hall equipment

All power feeds are UPS backed up with an approximate autonomy of 10-15 minutes. In addition, there are two diesel generator backups for critical services. The majority of racks are nowadays equipped with intelligent (switched and metered) power distribution units (PDU), which enables a fine-grained monitoring of power consumption.

The Tier1 module and module 4 are air-cooled with in-row racks arranged between server racks for cold-air containments. Modules 1 and 2 are air-cooled, and the cold air is blown through large ducts underneath the false floor, where it diffuses into cold aisles through perforated floor tiles. The “Govorun” supercomputer is fully “hot” water-cooled, which allows for a power density of 100 kW per rack and PUE = 1.06.

All engineering equipment that provides both the guaranteed energy supply to the MICC and the cooling system is located on the first and basement floors of the building. Only chillers and diesel generators are located on the territory adjacent to the MLIT building.

3. Networking

One of the most important components of the MICC, providing access to the resources and the possibility to work with experimental data processing and computing, is the network infrastructure. In the frames of these works, it is necessary to ensure the reliable and fault-tolerant operation of all network components of the infrastructure: external telecommunication channels, the JINR backbone network with a multisite cluster network and the local MICC network.

At the moment, external telecommunication channels are presented by the JINR-Moscow optical link with a 3x100 Gbit/s capacity, the direct JINR-CERN 100 Gbit/s link for LHCOPN, connecting all WLCG Tier1 centers with the CERN Tier0 and Tier1 centers with each other and the JINR-Amsterdam 100 Gbit/s link for the LHCOPN, LHCONE, GEANT networks, direct channels up to 100 Gbit/s for communication using RU-VRF technology with the collaboration of RUHEP

(Gatchina, NRC Kurchatov Institute, Protvino) and with the RUNNet, RASnet networks. IPv6 routing for the Tier1 and Tier2 sites is implemented.

The local area network (LAN) is presented by the JINR LAN 2x100 Gbit/s backbone and the distributed MultiSite Cluster Network between the DLNP and VBLHEP sites (4x100 Gbit/s).

The internal MICC network has the Tier1 segment built on the Brocade factory with a throughput of 80 Gbit/s. The EOS data storage system, Tier2, cloud environment, and “Govorun” supercomputer segments are built on the Dell and Cisco equipment. Ports up to 10 Gbit/s and 100 Gbit/s are used to connect server components on access level switches in the MICC network core, built on Cisco Nexus 9504 and Nexus 9336C switches with an N x 100 Gbit/s port bandwidth.

The internal network of the “Govorun” supercomputer consists of three main parts: a communication and transport network, a control and monitoring network and a task control network. The communication and transport network uses Intel OmniPath 100 Gbit/s technology. The network is built on a "thick tree" topology based on 48-port Intel OmniPath Edge 100 Series switches with full liquid cooling. The control and monitoring network enables the unification of all compute nodes and the control node into a single Fast Ethernet network. This network is built using Fast Ethernet switches HP 2530-48. The job control network connects all compute nodes and the control node into a single Gigabit Ethernet network. The job control network is built using HPE Aruba 2530 48G switches.

4. Grid infrastructure

The first level grid resource center (T1_RU_JINR) is now used to process and store data for the CMS experiment and to perform simulations for the NICA project. At present, there are 16,096 cores with a total performance of 253,135.18 HEP-SPEC06 and an average of 15.73 HEP-SPEC06 per core. The software and compilers used are CentOS Scientific Linux release 7.9, gcc (GCC) 4.4.7, C ++(g ++ (GCC) 4.4.7), GNU Fortran (GCC) 4.4.7, dCache-5.2 for data storage, Enstore 6.3 for tape libraries and FTS.

From the end of 2020, the entire JINR grid sites processing resources were migrated from CREAM-CE and Torgue-Maui to the ARC-SE [6] Compute Element and the SLURM [7] Batch System (adapted to kerberos and AFS). It involved the migration of more than 20,000 CPU cores, as well as new policies for both the CE and the Batch System. To support NICA computing FairSoft, FairRoot and MPDroot were installed.

In terms of performance (number of processed events, jobs per site, etc.) JINR Tier1 ranks second among the WLCG (Worldwide LHC Computing Grid) [8] Tier1 centers (FNAL, JINR, CCIN2P, KIT, CNAF, RAL, PIC) for the CMS experiment, and about 22% of the sum CPU work was performed.

One of the main functions of Tier1 level centers is to provide data exchange with all global sites that run CMS jobs. Since the beginning of the year, 8.5 PB of data from more than 180 grid sites has been transferred to Tier1, and more than 10 PB of data has been downloaded.

The JINR Tier2 site is the most productive in the Russian RDIG Consortium (Russian Data Intensive Grid) [9]. Over 61% of the total CPU time in the RDIG is used for computing on our Tier2. There are ~7,700 cores with a total performance of 121,076.99 HEP-SPEC06. The Tier2 resources are used by all four LHC experiments, as well as by the NICA experiments, ILC, NOvA, BES, JUNO, etc. The software and the batch system are the same as for Tier1.

5. Cloud computing

The task of JINR cloud computing [10] is twofold. Firstly, it is necessary to provide JINR users with cloud services for conducting research, and secondly, to ensure the functioning and expansion of the distributed computing environment using the cloud resources of the JINR Member States.

At the moment, there are 200+ physical servers, 5,000+ non-hyperthreaded CPU cores, 60+ TB of RAM and the KVM hypervisor only. The software used is a custom opennebula collector for the prometheus TSDB, prometheus alertmanager and grafana dashboard cloud.

Ceph-based storages are used for:

- general purpose for VMs (RADOS block devices, cephfs, and object storage) – 1.1 PiB of raw disk space with 3x replicas;
- dedicated for the NOvA experiment – 1.5 PiB of raw disk space with 3x replicas;
- SSD-based for VMs with intensive disk I/O – 419 TiB of raw disk space with 3x replicas.

All three ceph-based storages are monitored by the prometheus plugin, prometheus alertmanager and grafana dashboards.

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 distribute a peak load across them, the DIRAC Interware was put into operation [11]. It enables the integration of the JINR cloud with the partner organizations via DIRAC. The geography of organizations that share part of their resources using the distributed cloud infrastructure is illustrated in Figure 3.

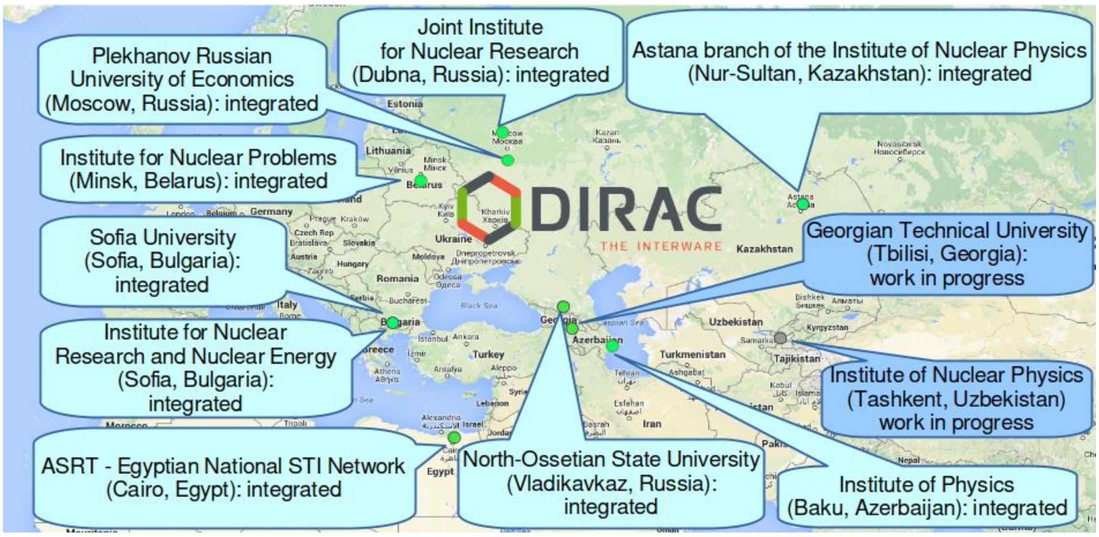
Joint Institute for Nuclear Research (Dubna, Russia): integrated
Plekhanov Russian University of Economics (Moscow, Russia): integrated
Astana branch of the Institute of Nuclear Physics (Nur-Sultan, Kazakhstan): integrated
Institute for Nuclear Problems (Minsk, Belarus): integrated
DIRAC
THE INTERWARE
Georgian Technical University (Tbilisi, Georgia): work in progress
Sofia University (Sofia, Bulgaria): integrated
Institute for Nuclear Research and Nuclear Energy (Sofia, Bulgaria): integrated
Institute of Nuclear Physics (Tashkent, Uzbekistan) work in progress
ASRT - Egyptian National STI Network (Cairo, Egypt): integrated
North-Ossetian State University (Vladikavkaz, Russia): integrated
Institute of Physics (Baku, Azerbaijan): integrated

Figure 3. JINR Member States' cloud environment: participants

The main users of cloud computing are the Baikal-GVD, BESIII, DayaBay, JUNO, NOvA, DUNE experiments. At the end of 2020, Baikal-GVD started using clouds via DIRAC for Monte-Carlo simulation.

The cloud resources of both the Institute and the organizations of its participating countries, which are free from the main activity of scientific computing, are successfully involved in research on COVID-2019 within the Folding@Home platform.

6. HybriLIT platform and "Govorun" supercomputer

The unified software and information environment of HybriLIT (fig. 4) can be used as an educational platform and as a testing polygon, which is aimed at exploring the possibilities of novel computing architectures, IT solutions, developing and debugging their applications, furthermore, carrying out calculations on the supercomputer, which enables the efficient use of the supercomputer resources.

In the HybriLIT environment, the latest versions of over 20 software packages, in particular, GSL, FairSoft, FairRoot, PyROOT with add-ons for BmnRoot and MpdRoot, SMASH, Valgrind, ABINIT, Wien2k, Amber, AmberTools, DIRAC, ELPA, FLUKA, LAMMPS, FreeSurfer, FSL, MRIConvert, GROMACS, FORM, SMILEI COMSOL, Maple, Mathematica, etc., were implemented and are supported at the request of user groups.

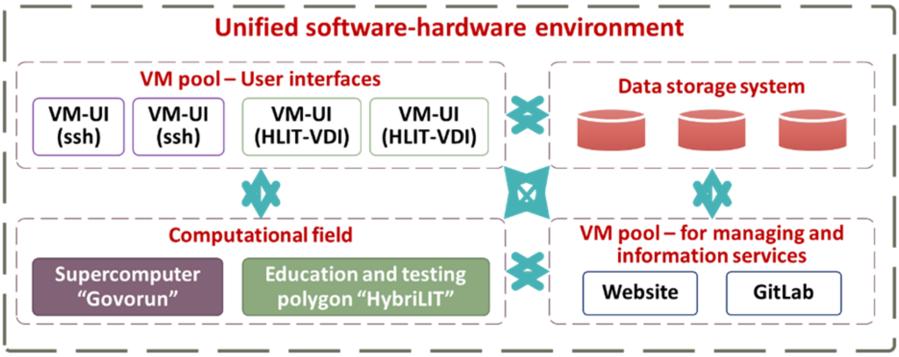
Unified software-hardware environment
VM pool – User interfaces
VM-UI (ssh)
VM-UI (ssh)
VM-UI (HLIT-VDI)
VM-UI (HLIT-VDI)
Data storage system
Computational field
Supercomputer "Govorun"
Education and testing polygon "HybriLIT"
VM pool – for managing and information services
Website
GitLab

Figure 4. Unified software and information environment of the HybriLIT platform

The main HPC resource of the MICC is the "Govorun" supercomputer. The second stage of its upgrade was in 2019, and now the full peak performance is 1.7 PFlops for single precision and 860 TFlops for double precision. The 288 TB storage has an I/O speed of >300 GB/s. Like most supercomputers "Govorun" has:

- GPU component: 5 NVIDIA DGX-1 servers with 8 NVIDIA Tesla V100 GPUs (NVIDIA Volta), 40,960 cores CUDA on one NVIDIA, NVLink 2.0 wire (a bandwidth of up to 300 Gbit/s);
- CPU component: 21 RSC Tornado nodes based on Intel® Xeon Phi™ (Intel® Xeon Phi™ 7290 processors (72 cores), Intel® Server Board S7200AP, Intel® SSD DC S3520 (SATA, M.2), 96GB DDR4 2400 GHz RAM, Intel® Omni-Path 100 Gbit/s adapter), 88 RSC Tornado nodes based on Intel® Xeon® Scalable gen 2 (Intel® Xeon® Platinum 8268 processors (24 cores), Intel® Server Board S2600BP, Intel® SSD DC S4510 (SATA, M.2), 2x Intel® SSD DC P4511 (NVMe, M.2) 2TB, 192GB DDR4 2933 GHz RAM, Intel® Omni-Path 100 Gb/s adapter;
- 14 storage modules of the fast scalable parallel file system (Lustre, EOS, etc.) based on Intel® SSD DC P4511 (NVMe, M.2) 2TB with a total capacity of 288 TB;
- additionally "Govorun" has 4 special nodes with 12 high-speed, low-latency solid state drives Intel® Optane™ SSD DC P4801X 375GB M.2 Series with Intel® Memory Drive Technology (IMDT), which allows getting 4.2 TB for very hot data per server.

The CPU component of "Govorun" is a hyperconverged software-defined system and ranks 16th (DAOS-10 node) in the current edition of the IO500 list (July 2021). Now the "Govorun" system has unique properties for the flexibility of customizing the user's job, ensuring the most efficient use of the computing resources of the supercomputer.

The resources of the "Govorun" supercomputer are used by scientific groups from all the Laboratories of the Institute within 25 themes of the JINR Topical Plan for solving a wide range of tasks in the field of theoretical physics, as well as for the modeling and processing of experimental data.

First of all, the resources are used to study the properties of quantum chromodynamics (QCD) and Dirac semimetals in a tight-binding mode under extreme external conditions using lattice modeling. The given study entails the inversion of large matrices, which is performed on video cards (GPU), as well as massive parallel CPU calculations, to implement the quantum Monte-Carlo method.

Other hot topics of resource usage are storing, processing and analyzing data within the NICA megascience project [12]. To speed up the simulation and reconstruction of events for the NICA MPD experiment, a hierarchical structure of working with data was implemented on the "Govorun" supercomputer. Events of the MPD experiment are simulated and reconstructed on the ultrafast data storage system under the Lustre file system management with a subsequent transfer to semi-cold storages and to the tape library for long-term storage. About 2 million events were generated for the

MPD experiment using the hierarchical structure of working with data. The acceleration of calculations on the upgraded supercomputer in comparison with the previous configuration was 1.45 times.

The HybriLIT platform is widely used for investigations based on machine and deep learning technologies. An example is the use of computer vision algorithms to accelerate experimental data processing by automating the morphological classification of neural cells, track interpretation and reconstruction algorithms, etc.

Research on the development of applied quantum algorithms using a quantum simulator started within the project “Superheavy nuclei and atoms: the limits of the masses of nuclei and the boundaries of D.V. Mendeleev’s Periodic Table”. The pie charts below demonstrate “Govorun” resources usage by groups of users.

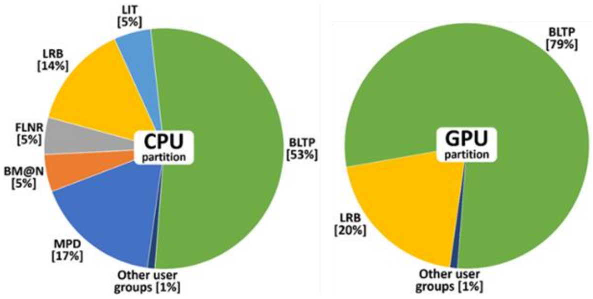

Figure 5. Distribution of the “Govorun” resources by user groups: NICA experiments, BM@N and MPD; BLTP – Bogoliubov Laboratory of Theoretical Physics, LRB – Laboratory of Radiation Biology, FLNR – Flerov Laboratory of Nuclear Reactions, LIT – Meshcheryakov Laboratory of Information Technologies

7. Data storage

The MICC data storage operates on two systems, namely, IBM TS3500 and IBM TS4500 tape libraries with a total capacity of 51 PB used for long-term and archival data storage, and the EOS [13] disk storage ~ 7 PB intended for storing and accessing large amounts of information, including distributed collective data generation, raw data storage, data conversion and analysis. EOS is common for Tier1, Tier2, Cloud and HybriLIT platform users.

8. Monitoring and accounting

The developed integrated monitoring system of the MICC allows receiving information from various components of the computing complex: engineering infrastructure, network, compute nodes, batch systems, data storage elements, grid services, which guarantees a high level of reliability of the entire MICC. The Litmon monitoring system is modular and distributed [14]. The role of the multi-level monitoring system for the MICC is to combine the existing systems and solve the problem of providing high-level information about the computing complex and its services. At present, an accounting system for the JINR Tier1 and Tier2 grid sites [15] was developed on top of Grafana dashboards and integrated into the Litmon monitoring system.

9. Conclusion

The development of the JINR distributed environment is aimed at creating a technological frame that enables scientific research at JINR to be conducted in a unified information and computing environment, incorporating a multitude of technological solutions, concepts and practices. Such an environment has to combine supercomputer (software-defined hyperconverged server solutions), grid technologies, cloud computing and systems to provide the best approaches for the solution of different kinds of scientific and applied tasks. The essential requirements for this environment are scalability, interoperability and adaptability to new technical solutions. The transition to distributed experimental data processing and storage based on modern technologies is a necessary condition for the successful participation of physicists of JINR and the JINR Member States’ institutes in the NICA project at JINR, as well as in other worldwide experiments and applied studies performed in collaboration with JINR scientists. It should be mentioned that HPC is also needed for theoretical investigations.

During last few years, the renovation of the engineering infrastructure (power supply and cooling systems) was in progress. We gradually modernized the local area and MICC networks, and configured the 4x100 Gbps multisite cluster network for the IT ecosystem of the NICA project. We are constantly enhancing the performance of the JINR Tier1 and Tier2 grid sites in accordance with the requirements of the experiments at the LHC, as well as increasing the data storage capacity on the HDD up to 7.35 PB and the tape robot capacity up to 51 PB. The “Govorun” supercomputer currently has a total performance of 0.86 Pflops, and its performance is projected to expand from year to year.

The MICC is a dynamically evolving IT platform that responds to the rapidly developing IT world. The promising directions of modern information technologies are Artificial Intelligence and Robotics, as well as Quantum Technologies and Big Data Analytics. The development of the scientific IT-ecosystem will depend on novel technologies for acquiring, analyzing and sharing data. Thus, such a system must be very flexible and open to new computing methods such as quantum, cognitive calculations, machine learning methods and data mining, as well as to any developments of new algorithmic bases.

References

[1] A.G. Dolbilov, I.A. Kashunin, V.V. Korenkov et al. Multifunctional Information and Computing Complex of JINR: Status and Perspectives. CEUR Workshop Proceedings (CEUR-WS.org), 2019. V. 2507. P. 16 – 22. Available at: http://ceur-ws.org/Vol-2507/16-22-paper-3.pdf

[2] NICA (Nuclotron-based Ion Collider fAcility): http://nica.jinr.ru/

[3] DIRAC Interware: https://dirac.readthedocs.io/en/latest/index.html

[4] A.S. Baginyan, A.I. Balandin, A.G. Dolbilov et al. Grid at JINR. CEUR Workshop Proc. 2019. V. 2507. P. 321 – 325. Available at: http://ceur-ws.org/Vol-2507/321-325-paper-58.pdf

[5] HybriLIT Platform: https://micc.jinr.ru/?id=30

[6] ARC Compute Element (CE): https://www.nordugrid.org/arc/ce/

[7] SLURM. Available at: https://slurm.schedmd.com/documentation.html (accessed 15.07.2021).

[8] The Worldwide LHC Computing Grid (WLCG): http://wlcg.web.cern.ch/LCG

[9] A. Soldatov, V. Korenkov, V. Ilyin, Russian Segment of the LCG Global Infrastructure, Open Systems, N1, 2003, in Russian

[10] N.A. Balashov, A.V. Baranov, N.A. Kutovskiy et al. Present status and main directions of the JINR cloud development. CEUR Workshop Proc. 2019. V. 2507. P. 185 – 189. Available at: http://ceur-ws.org/Vol-2507/185-189-paper-32.pdf

[11] N.A. Balashov, R.I. Kuchumov, N.A. Kutovskiy et al. Cloud integration within the DIRAC Interware. CEUR Workshop Proc. 2019. V. 2507. P. 256 – 260. Available at: http://ceur-ws.org/Vol-2507/256-260-paper-45.pdf

[12] D.V. Belyakov, A.G. Dolbilov, A.N. Moshkin et al. Using the “Govorun” supercomputer for the NICA megaproject. CEUR Workshop Proc. 2019. V. 2507. P. 16 – 22. Available at: http://ceur-ws.org/Vol-2507/16-22-paper-3.pdf

[13] EOS Open Storage: http://eos.web.cern.ch/

[14] I. Kashunin, V. Mitsyn, V. Trofimov, A. Dolbilov. Integration of the Cluster Monitoring System Based on Icinga2 at JINR LIT MICC. PEPAN Letters, v. 17, No 3(228), P. 345–352. Available at: http://www1.jinr.ru/Pepan_letters/panl_2020_3/14_kashunin.pdf

[15] I.A. Kashunin, V.V. Mitsyn, T.A. Strizh. JINR WLCG Tier1 & Tier2/CICC accounting system. Ibid.