IT SOLUTIONS FOR JINR TASKS ON THE “GOVORUN” SUPERCOMPUTER

D.V. Podgainy, D.V. Belaykov, A.V. Nechaevsky, O.I. Streltsova, A.V. Vorontsov, M.I. Zuev

Meshcheryakov Laboratory of Information Technologies, JINR

E-mail: podgainy@jinr.ru

The “Govorun” supercomputer is a heterogeneous computing system that contains computing architectures of different types, including graphics accelerators. The given architecture of the supercomputer allows users to choose optimal computing facilities for their tasks.

To enhance the efficiency of solving user tasks, as well as to expand the efficiency of utilizing both computing resources and data processing and storage resources, a number of special IT solutions have been implemented on the “Govorun” supercomputer. A hierarchical hyperconverged data processing and storage system with a software-defined architecture is referred to the first type of IT solutions. The implementation of this system is caused by the fact that modern supercomputers are used not only as traditional computing environments for carrying out massively parallel calculations, but also as systems for Big Data analysis and artificial intelligence tasks that arise in different scientific and applied tasks. The second type of IT solutions lies in resource orchestration, which means that computational elements (CPU cores and graphics accelerators) and data storage elements (SSDs) form independent computing and data storage fields. Due to it, the user can allocate for his task the required number and type of compute nodes (including the required number of graphics accelerators), as well as the required volume and type of data storage systems.

Keywords: high-performance platforms, data processing and storage systems, computing for high-energy physics

Dmitriy Podgainy, Dmitriy Belaykov, Andrey Nechaevsky,
Oksana Streltsova, Aleksey Vorontsov, Maksim Zuev

1. Introduction

The “Govorun” supercomputer is an integral part of the HybriLIT heterogeneous computing platform (http://hlit.jinr.ru/) of the Meshcheryakov Laboratory of Information Technologies of the Joint Institute for Nuclear Research (MLIT JINR), which also comprises the HybriLIT training and testing polygon [1, 2]. The “Govorun” supercomputer is designed for resource-intensive massively parallel calculations. It is an innovative hyperconverged software-defined system with unique properties in terms of the flexibility of customizing the user task, ensuring the most efficient use of the computing resources of the supercomputer. The “Govorun” supercomputer encompasses a GPU component, a CPU component and a hierarchical data processing and storage system. The GPU component is implemented on the basis of five NVIDIA DGX-1 servers, each of which contains eight Tesla V100 graphics accelerators. The CPU component of the supercomputer is implemented on the “RSC Tornado” high-density architecture with direct liquid cooling, which ensures a high density of compute nodes, i.e. 150 nodes per rack, and high energy efficiency about 10 GFlop/WB. The average annual PUE indicator of the system, reflecting the level of energy efficiency, is less than 1.06. The CPU and GPU components of the “Govorun” supercomputer rank 12th and 21st in the TOP50 list of supercomputers in the CIS countries respectively (http://top50.supercomputers.ru/).

The operation experience of the “Govorun” supercomputer has indicated the relevance and effectiveness of using the latest hyperconverged computing architectures that are part of it, which is reflected in 109 publications of users from July 2018 to January 2021. 11 articles were published in Q1 journals, and 15 papers were published in Q2 journals. Thus, on average, one publication accounts for 8.5 days of working on the “Govorun” supercomputer. In 2020, 65 articles were published, and on the basis of user reports, a booklet was formed and posted on the website (http://hlit.jinr.ru/users_publications/). In the first 6 months of 2021, four articles have already been published in Q1 journals, and a paper of the BM@N collaboration has been prepared for the Nature Physics journal, i.e., for three years of operation, using the resources of the “Govorun” supercomputer, two publications have been prepared for this prestigious journal.

At the same time, the number of supercomputer users is growing. After commissioning in July 2018, the total number of users was 46 (41 from JINR and 5 from other organizations, including organizations of the Russian Federation and JINR’s Member States), in 2019, the number of users increased and amounted to 133 (93 from JINR and 40 from other organizations, including organizations of the Russian Federation and JINR’s Member States), and in 2020, it reached 161 (95 from JINR and 66 from other organizations, including organizations of the Russian Federation and JINR’s Member States).It is noteworthy that access to the resources of the “Govorun” supercomputer is provided only to those users who are directly involved in the implementation of the JINR Topical Plan.

It should be pointed out that the range of tasks solved by the “Govorun” supercomputer is constantly expanding, and their specificity requires not only scaling supercomputer resources, but also introducing new IT solutions, which are not characteristic of traditional HPC systems. An example of such a task is computing being created for the NICA megascience project, for which the “Govorun” supercomputer, due to the flexibility of its architecture, is a key resource for testing and creating IT solutions. In particular, for this task, a hierarchical data processing and storage system has been developed and implemented together with RSC Group on the “Govorun” supercomputer. Let us single out several fundamental features related to the computing being created for the NICA project, which are also characteristic of other high-energy physics projects. First of all, it is necessary to work with large amounts of data, while at different stages of event reconstruction and simulation workflows, there is a need for different access speeds, for example, the access speed is not an important factor for long-term storage tasks, but it is essential for reconstruction tasks. In addition, a large amount of RAM is required for a number of tasks, which leads to the need to implement specialized nodes with large memory in the supercomputer architecture. Thus, methodologically, to ensure all work processes for tasks of the NICA megaproject, a system that combines both computing architectures of different types and a developed hierarchical data processing and storage system is being created on the

"Govorun" supercomputer. The approach to the implementation of new IT solutions is schematically presented in Fig. 1.

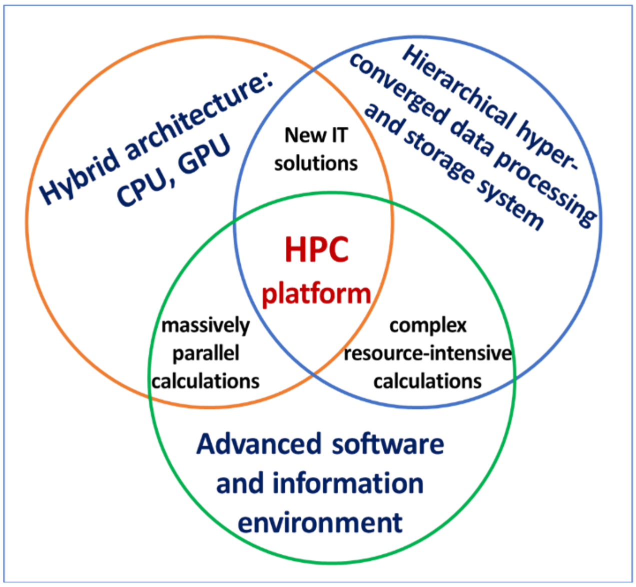
Hybrid architecture: CPU, GPU
Hierarchical hyper-converged data processing and storage system
New IT solutions
HPC platform
massively parallel calculations
complex resource-intensive calculations
Advanced software and information environment

Figure 1. Venn diagram for approaches to the development of the "Govorun" supercomputer.

2. Architecture of the "Govorun" supercomputer

The supercomputer includes computing modules, network infrastructure modules, the RSC BasIS software and an infrastructure module (https://rscgroup.ru/). Computing modules have CPU and GPU components. The GPU component is based on five NVIDIA DGX-1 servers, each of which comprises eight NVIDIA Tesla V100 graphics accelerators. The CPU component is implemented on 21 RSC Tornado compute nodes containing Intel® Xeon Phi™ processors and 88 RSC Tornado nodes each containing two Intel® Xeon® Platinum 8268 processors and two Intel® SSDs DC P4511 (NVMe, M.2), with a capacity of 2TB each.

Additionally, the "Govorun" supercomputer has the RSC Storage-on-Demand system, which is a single centrally managed system and has several levels of data storage, namely, very hot data, hot data and warm data. The very hot data storage system is built on top of four RSC Tornado TDN511S blade servers. Each server has 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 for 4.2 TB of very hot data per server. The hot and warm data storage system consists of a static storage system with the Lustre parallel file system, created on the basis of 14 RSC Tornado TDN511S blade servers, and a dynamic RSC Storage-on-Demand system on top of 84 RSC Tornado TDN511 blade servers with support for the Lustre parallel file system and the DAOS distributed object storage system. Low-latency solid state drives Intel® Optane™ SSD DC P4801X 375GB M.2 Series are used to quickly access metadata of the Lustre file system. Intel® SSDs DC P4511 (NVMe, M.2) are used to store Lustre hot data.

The network infrastructure module encompasses a communication and transport network, a control and monitoring network and a task control network. NVIDIA DGX-1 servers are interconnected by the communication and transport network based on InfiniBand 100 Gbps technology, and this component communicates with the CPU module via Intel OmniPath 100 Gbps. 3.5. The communication and transport network of the CPU module 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 HP 2530-48 switches. The task control network connects all compute nodes and the control node into a single Gigabit Ethernet network. The network is built using HPE Aruba 2530 48G switches.

An equally important part of the “Govorun” supercomputer architecture is the RSC BasIS supercomputer control software. RSC BasIS uses the CentOS Linux version 7.8 operating system on all compute nodes (CN) and performs the following functions:

- monitors compute nodes with the emergency shutdown functionality in the case of detecting critical malfunctions (such as CN overheating);
- collects indicators of the functioning of communication and transport network components;
- collects performance indicators of compute nodes, i.e., load of processors and RAM;
- stores monitored indicators with the ability to view statistics for a given time interval (at least one year);
- collects the readings of the integral indicator of the state of CNs and displays them on the geometric form of the calculator rack;
- displays the status of the leak detection system by moisture control sensors on compute nodes and displays it on the geometric form of the calculator rack;
- displays the efficiency of using the allocated resources via the SLURM scheduler to the cluster user for a specific task as an indicator of the average load of CPUs allocated by the user (%);
- displays the availability of CNs on the computing network and the control network on the geometric form of the calculator rack.

The supercomputer control software includes:

- system for managing user profiles and environments;
- node software control system, including installation and updating of the operating system and application software packages;
- SLURM scheduler;
- software for secure remote access to the supercomputer;
- tools for parallel administration and supercomputer control.

In addition, RSC BasIS provides the management of the RSC Storage-on-Demand system, which ensures the following:

- configuration of logical instances of storage systems using drives installed inside compute nodes designed to perform user tasks;
- configuration of the parameters and hierarchy of the storage system instance;
- monitoring of the key parameters of the storage system;
- tasks are automatically launched on nodes that provide drives for use by other CNs via the task control system;
- management of groups of CNs acting as clients for storage systems;
- automatic mounting and unmounting of storage systems to a group of CN clients after making changes to the configuration;
- graphical interface for creating static storage systems on demand with the ability to verify the scheme of a future storage system;
- manual and automatic replacement of drives in degraded raid arrays;
- ability to connect drives over a network with support for RDMA and Ethernet using the NVMe-over-Fabrics and NVMe-over-TCP protocols respectively.

The infrastructure module comprises subsystems of refrigeration, power supply and an automated remote control system. The refrigeration subsystem provides the absorption, removal from compute nodes and dissipation of thermal energy in the atmosphere. Thermal energy dissipation is ensured by a liquid cooler, a dry closed cooling tower. To transfer thermal energy from nodes to the liquid cooler, circulation pumps and heat exchange units, assembled in a rack form factor, i.e., a pumping unit, are used. All computing racks are connected to a single refrigeration system. The refrigeration system is created on the basis of collectors of internal and external circuits, which allows connecting new liquid coolers, pumping units and computing racks without stopping the system. Pumping units are installed with N+1 redundancy. The refrigeration subsystem provides cooling of compute nodes with a total electrical power of up to 200 kW. The power supply subsystem performs the distribution and accounting of power supply. Computing racks, network infrastructure cabinets, refrigeration subsystems and the automated remote control system are connected to the power supply subsystem. The automated remote control system manages the operation parameters of the refrigeration subsystem. It also monitors the current states of subsystems and provides the emergency shutdown of hardware.

3. IT solutions: hierarchical hyperconverged data processing and storage system with a software-defined architecture and resource orchestration

At present, in the development of supercomputer technologies, there is a tendency to use supercomputers not only for massively parallel calculations, but also for working with Big Data [4]. The latter circumstance is related to the fact that the volume of Big Data is increasing exponentially from day to day. If we talk about the sources of Big Data in science, then the Large Hadron Collider at CERN, which generates about hundreds of petabytes of data per year, can be called the main one [5]. SKA, i.e., a project to create a radio telescope with an area of one square kilometer, the expected flow of raw data from which will be 1 exabyte per year, can be singled out of the upcoming experiments [6]. At the same time, when considering the issue of working with Big Data, two directions can be distinguished, the first one is associated with the use of tools for manipulating data (MapReduce algorithms, Hadoop libraries, etc.) [7], and the second one is related to the development of software and hardware solutions, which enables to effectively solve tasks on data manipulation. The NICA accelerator complex is currently under construction at JINR, and its experiments will generate tens of petabytes of data per year. The accounting of this trend and the creation of computing for the NICA project have defined the development vector for the “Govorun” supercomputer, which plays the role of a research polygon for the elaboration of software-hardware and IT solutions for the NICA project. To work with Big Data, a hierarchical hyperconverged data processing and storage system with a software-defined architecture has been implemented on the “Govorun” supercomputer. According to the speed of accessing data, the system is divided into layers that are available for the user’s choice. Each layer of the developed data storage system can be used both independently and as part of data processing workflows. The “Govorun” supercomputer ranks 31st in the IO500 list (https://io500.org/) with a bandwidth of over 35 GiB/s and a speed of accessing metadata of over 230 kIOP/s.

An equally important property of the “Govorun” supercomputer is the hyperconvergence of compute nodes, which allows orchestrating computing resources and data storage elements, as well as creating computing systems on demand using the RSC BasIS software. The notion “orchestration” means the software disintegration of a compute node, i.e., the separation of compute nodes and data storage elements (SSDs) with their subsequent integration in accordance with the requirements of user tasks. Thus, computing elements (CPU cores and graphics accelerators) and data storage elements (SSDs) form independent fields. Due to orchestration, the user can allocate for his task the required number and type of compute nodes (including the required number of graphics accelerators), the required volume and type of data storage systems. After the task is completed, compute nodes and storage elements are returned to their corresponding fields and are ready for the next use. This feature allows one to effectively solve user tasks of different types, to enhance the level of confidentiality of working with data and avoid system errors that occur when crossing the resources for different user tasks.

The use of these IT solutions makes it possible to formulate the concept of the development of the “Govorun” supercomputer as the implementation of mapping the main characteristics of Big Data V^3 (*Volume*, large amounts of data for processing and storage; *Velocity*, the need for high-speed data processing, *Variety,* data of different types) on the hardware and software characteristics of the supercomputer H^3 (*Heterogeneity, Hierarchy, Hyperconvegence*). Heterogeneity, i.e., the presence of different types of calculators, namely, central processors, both classical and with a large number of processor cores (Intel Xeon Phi), graphics accelerators, allows choosing one or another type of calculator for efficient work with a given type of data. Hierarchy consists of two components, i.e., the use of file systems with different read/write speeds, from ultra-fast (DAOS, Lustre) to distributed, intended for long-term storage (EOS, Tapes). Hyperconvegence stands for the ability of the system to scale to the required capacity, namely, the same compute nodes can participate in data processing and be used for storage systems.

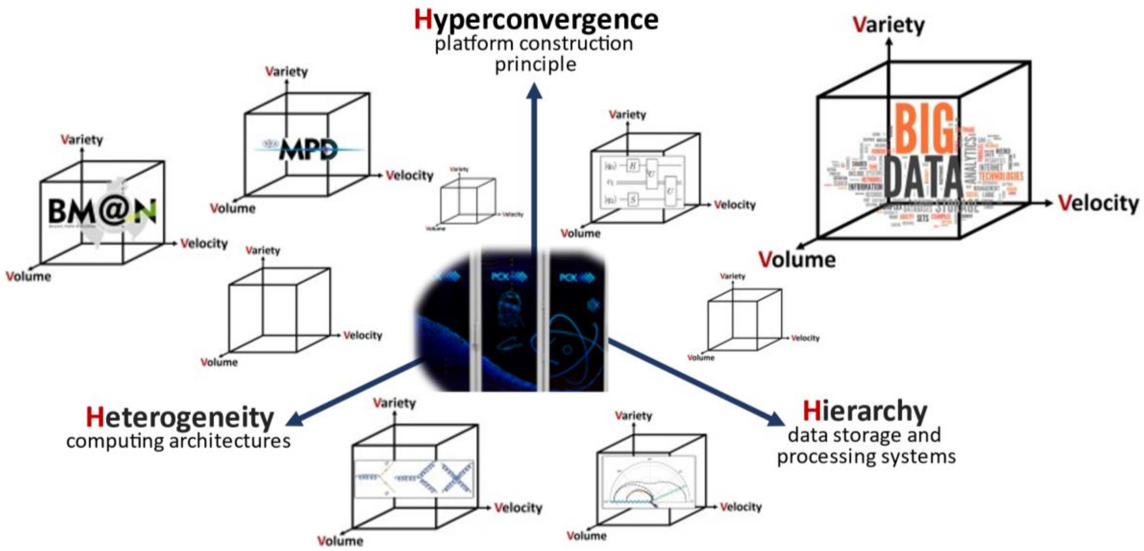

Figure 2. Concept of ensuring work with Big Data on the “Govorun” supercomputer: $V^3 \rightarrow H^3$.

4. Conclusion

The implementation of the above technologies on the “Govorun” supercomputer has made it possible to perform a number of complex resource-intensive calculations in the field of lattice quantum chromodynamics to study the properties of hadronic matter at high energy density and baryon charge and in the presence of supramaximal electromagnetic fields, to qualitatively increase the efficiency of modeling the dynamics of collisions of relativistic heavy ions, to speed up the process of event generation and reconstruction for conducting experiments within the NICA megaproject implementation, to carry out computations of the radiation safety of JINR’s experimental facilities, to significantly accelerate studies in the field of radiation biology and other applied tasks solved at JINR under international scientific cooperation.

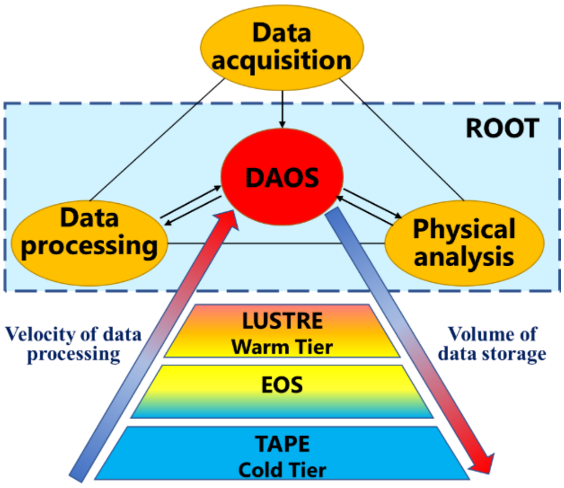

Figure 3. Intended usage of DAOS for the NICA project.

It is noteworthy that the “Govorun” supercomputer is an actively developing platform that combines state-of-the-art computing architectures and IT solutions. At present, one of such directions of development is research related to working with Big Data using the latest DAOS (Distributed Asynchronous Object Storage) technology [8]. For this purpose, a polygon of eight nodes with DAOS has been deployed on the “Govorun” supercomputer, it has demonstrated a high read/write speed and ranks 16th in the “10 node challenge” nomination in the current edition of the IO500 list (https://io500.org/list/isc21/ten). Great prospects for the use of this technology are associated with its application to the NICA project at all stages of its work, from experimental data acquisition to final physical analysis (Fig. 3).

At the same time, it is expected that the use of DAOS for the NICA project will enable to save and read multi-dimensional data structures of TB scale in a single address space, to create a multi-user presentation layer for analyzing physical results and easily integrate DAOS with other hot/warm storages. The latter circumstance is especially important due to the fact that a distributed, heterogeneous computing environment based on the DIRAC system has recently been developed for the MPD experiment to launch event generation and reconstruction tasks [9]. Meanwhile, in the process of calculating tasks, data is written to ultra-fast file storages and gradually transferred to slower ones, up to distributed or tape storages [10]. In addition, the DAOS technology looks promising in applying it to other types of tasks related to Big Data. These are primarily ML/DL tasks, including computer vision tasks, as well as the actively developing field of quantum computing.

5. Acknowledgments

The studies in this direction were supported by the RFBR special grant (“Megascience – NICA”), No. 18-02-40101.

References

[1] Gh. Adam, M. Bashashin, D. Belyakov, M. Kirakosyan, M. Matveev, D. Podgainy, T. Sapozhnikova, O. Streltsova, Sh. Torosyan, M. Vala, L. Valova, A. Vorontsov, T. Zaikina, E. Zemlyanaya, M. Zuev. IT-ecosystem of the HybriLIT heterogeneous platform for high-performance computing and training of IT-specialists. Selected Papers of the 8th International Conference «Distributed Computing and Grid-technologies in Science and Education» (GRID 2018), Dubna, Russia, September 10-14, 2018, CEUR-WS.org/Vol-2267″.

[2] Dmitry Belyakov, Andrey Nechaevskiy, Igor Pelevanuk, Dmitry Podgainy, Alexey Stadnik, Oksana Streltsova, Aleksey Vorontsov, Maxim Zuev, “Govorun” Supercomputer for JINR Tasks, CEUR Workshop Proceedings, 2020, 2772, pp. 1-12.

[3] Albrecht, J., Alves, A. A., Amadio, G., Andronico, G., Anh-Ky, N., Aphecetche, L., et al. (2019). A Roadmap for HEP Software and Computing R&D for the 2020s. Comput. Softw. big Sci. 3 (1), 7. doi:10.1007/s41781-018-0018-8.

[4] Semin A. DAOS: Data storage system for HPC/BigData/AI applications in the era of exascale computing, “Storage News”, № 2 (74), 2019 (In Russian).

[5] https://home.cern/news/news/computing/cern-data-storage-gets-ready-run-3.

[6] https://www.spiedigitallibrary.org/journals/Journal-of-Astronomical-Telescopes-Instruments-and-Systems/volume-8/issue-1/011004/Toward-a-Spanish-SKA-Regional-Centre-fully-engaged-with-open/10.1117/1.JATIS.8.1.011004.full.

[7] CMS BIG DATA PROJECT: https://cms-big-data.github.io/.

[8] https://docs.daos.io/.

[9] N. Kutovskiy, V. Mitsyn, A. Moshkin, I. Pelevanyuk, D. Podgayny, O. Rogachevsky, B. Shchinov, V. Trofimov, A. Tsaregorodtsev, Integration of distributed heterogeneous computing resources for the MPD experiment with DIRAC Interware, PEPAN, v. 52, № 4, pp. 999-1005, 2021.

[10] A.A. Moshkin, I.S. Pelevanyuk, D.V. Podgainy, O.V. Rogachevsky, O.I. Streltsova, M.I. Zuev, Approaches, services, and monitoring in a distributed heterogeneous computing environment for the MPD experiment, in Proceedings of the International Conference. September 27–28, 2021, Moscow, Russia / Ed. by Vl. Voevodin. – Moscow: MAKS Press, 2021. DOI: https://doi.org/10.29003/m2454.RussianSCDays2021.