HIGH-PERFORMANCE COMPUTING PLATFORMS FOR ORGANIZING THE EDUCATIONAL PROCESS ON THE BASIS OF THE INTERNATIONAL SCHOOL "DATA SCIENCE"

S.D. Belov[1,2 a], I.S. Kadochnikov[1,2 b], V.V. Korenkov[1,2,3 c], M.A. Matveev[1,2,3 d], D.V. Podgainy[1 e], D.I. Priakhina[1,2,3 f], R.N. Semenov[1,2 g], O.I. Streltsova[1,3 h], P.V. Zrelov[1,2 i]

[1] *Joint Institute for Nuclear Research, 6 Joliot-Curie St, Dubna, Moscow Region, 141980, Russia*

[2] *Plekhanov Russian University of Economics, Stremyanny lane, 36, Moscow, 117997, Russia*

[3] *Dubna State University, 19 Universitetskaya St, Dubna, Moscow Region, 141982, Russia*

E-mail: [a] belov@jinr.ru, [b] kadivas@jinr.ru, [c] korenkov@jinr.ru, [d] matveevma@jinr.ru, [e] podgainy@jinr.ru, [f] pryhinad@jinr.ru, [g] roman@jinr.ru, [h] strel.jinr.ru, [i] zrelov@jinr.ru

With the transition to the digital economy, the demand for specialists with advanced knowledge Mining Big Data, including using high-performance computing systems, is growing. The International School of Information Technologies "Data Science" was created on the initiative of the Joint Institute for Nuclear Research (JINR). The goal of the IT School is to train highly qualified IT specialists in the field of Data Science who will be able to use Big Data analytics to solve scientific and practical problems. In this paper, we discuss the features, structure, and properties of high-performance computing platforms that are used in educational processes and research activities of the IT School at Dubna State University and Plekhanov Russian University of Economics (PRUE). The educational program of the IT School at Dubna State University was developed taking into account the JINR personnel requirements and uses services and resources of the HybriLIT platform to help with students' practical engagement for effective learning. To solve economic and socially important problems, a universal platform that includes a high-performance heterogeneous subsystem, a cloud infrastructure, and a storage system was created at PRUE. The scientific laboratory "Cloud technologies and Big Data analytics" of PRUE elaborated a special educational program "Introduction to distributed computing and Big Data analytics" on the basis of these resources. Access to adequate tools and resources is vital for education in data science, Big Data analytics, and parallel programming.

Keywords: HPC, education, hybrid computing cluster, Big Data, analytics platform

Sergey Belov, Ivan Kadochnikov, Vladimir Korenkov, Mikhail Matveev, Dmitry Podgainy, Daria Priakhina, Roman Semenov, Oksana Streltsova, Petr Zrelov

1. Introduction

With the transition to the digital economy, the demand for specialists with advanced knowledge Mining Big Data, including using high-performance computing systems, is growing. The International School of Information Technologies "Data Science", hereinafter referred to as the IT School, was created on the initiative of the Joint Institute for Nuclear Research (JINR). Several universities participate in the School, namely, Dubna State University, Plekhanov Russian University of Economics (PRUE), Moscow State University, Saint-Petersburg State University, etc.

In this paper, we discuss the features, structure, and properties of high-performance computing platforms used in educational processes and research activities of the IT School at Dubna State University 0 and Plekhanov Russian University of Economics 0, as well as their use in educational programs.

2. Universities participating in the IT School

The goal of the IT School is to train highly qualified IT specialists in the field of Data Science who will be able to use Big Data analytics to solve scientific and practical problems. The educational program aims to develop deep knowledge in mathematical statistics, machine learning, programming, methods, and technologies of data processing and analysis, understanding business requirements and industry problems. It is expected that the participating organization will have both local programs specific to the educational institution and generic classes that will be implemented through joint events such as schools, conferences, and workshops. The events will be attended not only by the IT School participants, but also by students from JINR Member States and other countries. The international status of the IT School is predicated on the international status of JINR, the plan to invite lecturers from among prominent scientists of JINR Member States and other cooperating organizations, e.g. CERN.

Dubna State University is one the main universities of the IT School, in which the educational program "Big Data analysis" has been held since 2019. Within the school, Dubna University specializes in training IT specialists for solving high-energy and nuclear physics problems, as well as developing computing infrastructures for such scientific megaprojects as NICA, PIC, LHC, FAIR, SKA, etc. The "HybriLIT" heterogeneous platform 0, which is part of the Multifunctional Information and Computing Complex (MICC) 0 of the Laboratory of Information Technologies (LIT) 0 of JINR, plays a crucial role in the educational process.

In 2014, the Laboratory of cloud technologies and Big Data analytics was created at Plekhanov University 0. Lectures on Big Data technologies have been given since the creation of the laboratory, and the educational program "Introduction to distributed computing and Big Data analytics" was elaborated on their basis. In September 2018, an inter-faculty group "Data Science" was created with the goal of teaching Big Data and machine learning methods with the perspective application to economic and social problems. For practical classes and student research, the resources of the universal high-performance platform at PRUE are used.

2. "HybriLIT" heterogeneous platform

Heterogeneous resources at JINR function as part of the JINR MICC 0 and consist of the "HybriLIT" education and testing polygon and the Govorun supercomputer, which share a unified software and information environment 0. The supercomputer is used for problems requiring massively parallel computing in various fields of nuclear physics and high-energy physics. The structure of the HybriLIT platform is shown in Figure 1.

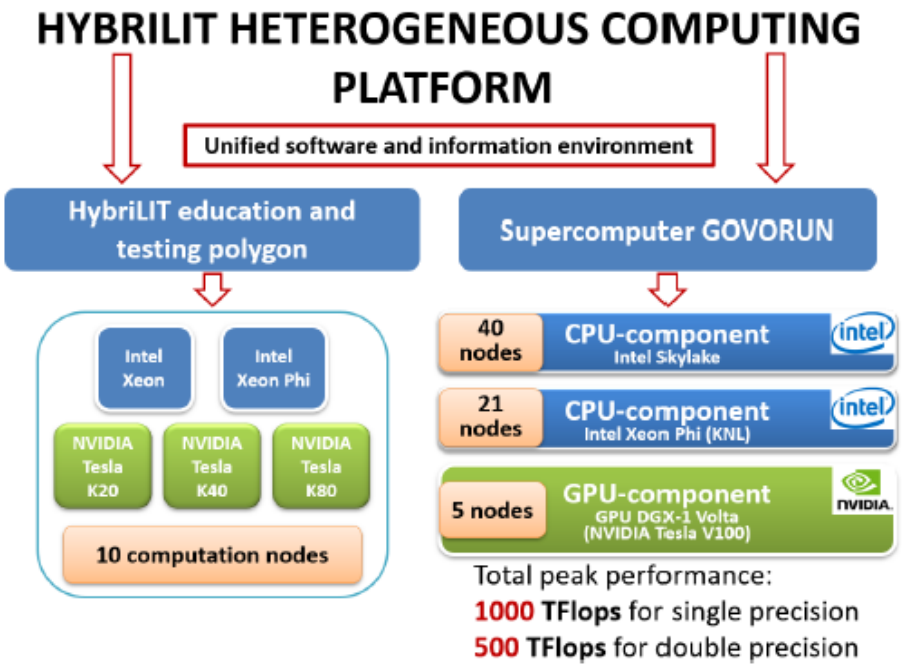

Figure 1. Structure of the HybriLIT platform

Training courses on parallel programming and hybrid computing technologies are held using the "HybriLIT" infrastructure. An ecosystem of machine learning/deep learning (ML/DL) and data analysis was deployed on "HybriLIT" for the research and development of ML/DL algorithms, mathematical models, and resource-intensive computing on CPUs and graphics accelerators 0. The platform has two components: a) component for resource-intensive massively parallel tasks of neural network learning on NVIDIA GPUs and b) model and algorithm development ecosystem based on JupyterHub 0, which is a multi-user platform for Jupyter Notebooks being an interactive web-based programming environment for different languages, including Python.

At present, the platform includes 10 computing nodes, containing NVIDIA graphics accelerators (Tesla K20, K40, K80), and Intel Xeon co-processors. The total computing performance of the cluster reaches 142 Tflops.

For effective utilization, a software and information environment, including a website, an Indico service, a GitLab service 0, etc., was deployed and is being actively developed.

2.1 Use of the "HybriLIT" platform in the educational process of the IT School at Dubna University

The educational program of the IT School at Dubna State University was developed taking into account the JINR personnel requirements. The program includes the study of such subjects as "Mathematical Basis and Tools for Data Analysis", "Technologies and Platforms for Distributed and Parallel Computing", "Big Data Analytics" 0.

Within the subject of "Technologies and Platforms for Distributed and Parallel Computing" students study OpenMP, MPI, CUDA 0 and OpenCL 0 parallel programming technologies. In addition, hybrid computing is considered, such as the combined use of MPI and OpenMP, or MPI and CUDA within one problem. For organizing practical lessons, the "HybriLIT" platform is used, since it contains both computing resources and programming libraries to use these technologies.

"Mathematical Basis and Tools for Data Analysis" aims to familiarize students with computational and statistical methods and tools for solving a wide range of data analysis problems, as well as develop practical programming skills in languages mainly used for large-scale data analysis, such as Python, R, Scala, and C++. To facilitate the educational process, Python data analysis,

machine learning, and deep learning libraries (namely, *NumPy*, *SciPy*, *matplotlib*, *scikit-learn*, *pandas*, *TensorFlow*) were installed on the platform. The lesson results are prepared in Jupyter Notebooks with Markdown markup and can be exported to PDF for review.

The purpose of the "Big Data Analytics" subject is to introduce students to modern tools and technologies of data analysis, such as Hadoop and Apache Spark, as part of the high-performance computing platform. In addition, participants get acquainted with version control and community development tools of GitLab, which is part of the "HybriLIT" platform, as well as cloud infrastructure technologies and Unix-like operating systems for constructing and managing computing platforms.

HybriLIT platform services (Indico, GitLab, etc.) form the basis of the most practical engagement of students for effective learning. This allows teaching the latest technologies and IT solutions that are not yet taught at most universities.

3. High-performance computing platform of Plekhanov University

For solving economic and socially important problems, as well as for IT education, a universal platform that includes a high-performance heterogeneous subsystem, a cloud infrastructure, and a storage system, was built at Plekhanov University.

The cloud infrastructure is based on the Open Nebula 5.6 cloud platform with 1 control node, 3 nodes in the Ceph storage cluster used for image and data storage, and 4 virtualization host nodes. The HPC infrastructure was deployed in October 2018 and consists of the high-performance Dell PowerEdge C4140 server containing 2 Intel Xeon Gold 6130 processors, 4 NVIDIA Tesla V100 graphics accelerators and 256GB of RAM. For internal storage, it has 480GB of SSD. The estimated HPC power is 60 single-precision Tflops and 30 double-precision Tflops. The components of the PRUE HPC system are shown in Figure 2.

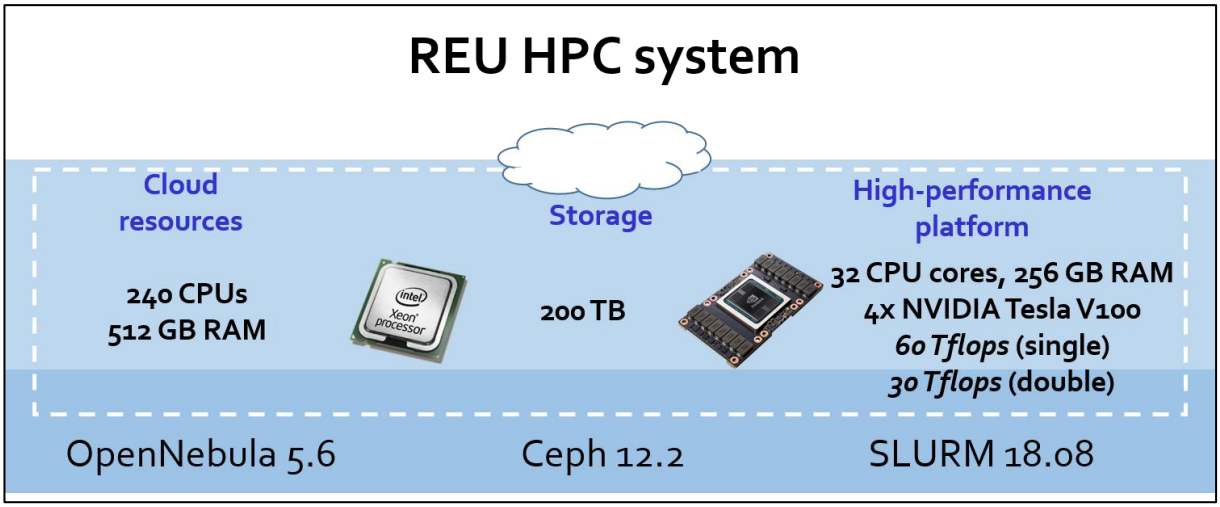

Figure 2. High-performance system of Plekhanov University

The system software includes the Centos 7.5 operating system and Intel MPI 0, OpenMPI 0 and NVIDIA CUDA parallel programming frameworks. C/C++ and Fortran compilers and a Python development environment are deployed on the high-performance platform with the most relevant parallel computing modules: *braid*, *chainer*, *cntk*, *keras*, *numpy*, *pandas*, *scipy*, *tensorflow-gpu*, *torch* [10]. Database and *Ceph* storage interfaces are provided. The list of available software is constantly expanding according to user requirements. Docker containerization is used to allow users to run a custom docker container and thus support research or computing in any provided software environment 0. Distributed resource access is ensured via a SLURM workload manager, which uses job queues to manage user tasks 0. A user can select the most efficient computing resource for his job

to run on GPUs or CPUs. By combining CPU and GPU resources on one heterogeneous server or cluster, one can meet the needs of more developers and data scientists.

3.1 Scientific applications of the PRUE HPC system

The HPC system at Plekhanov University is widely used by university scientific laboratories. For example, the Research Laboratory "Monetary System Studies and Financial Market Analysis" conducts research on the applicability of the Echo state of neural networks for forecasting (forecast of the trajectory of derivatives of stock prices by the ESN neural network) and studies of the option exchange segment for the most liquid assets of RTS and Si stock exchanges. The Research Laboratory "Applied Modeling" uses the HPS system for studies in the field of logistics (optimization of air transport loading), machine learning (training a neural network for detecting an aircraft in satellite images), factor-frame modeling (analysis of the mutual influence of factors in complex socio-economic models with feedback).

3.2 Educational program for Big Data training at Plekhanov University

In September 2018 at Plekhanov Russian University of Economics, an inter-faculty group "Data Science" was created in order to study Big Data and machine learning methods for the application in sociology and economics. The scientific laboratory "Cloud technologies and Big Data analytics" of PRUE elaborated a special educational program "Introduction to distributed computing and Big Data analytics". In addition to Big Data, the program covers modern programming methods, distributed computing, basic data structures and algorithms, database theory and practice, etc. Practical lessons and student research are performed using resources of the high-performance platform.

4. Conclusion

Access to adequate tools and resources is vital for education in data science, Big Data analytics, and parallel programming. Creating a heterogeneous platform that combines hardware, software and necessary services provides an effective way to facilitate students' training. The "HybriLIT" platform at JINR and the high-performance platform at PRUE met the requirements of the educational process of groups within the International School of Information Technologies "Data Science". The experience gained provides a blueprint for deploying educational computing platforms in the future.

5. Acknowledgments

The study was carried out at the expense of the Russian Science Foundation grant (project No. 19-71-30008).

References

[1] Dubna State University, n.d. Available at: https://int.uni-dubna.ru/ (accessed 18.11.2019)

[2] Plekhanov Russian University of Economics, n.d. Available at: https://www.rea.ru/en/Pages/default.aspx (accessed 18.11.2019).

[3] Supercomputer "Govorun" – Heterogeneous cluster | LIT/JINR, n.d. Available at: http://hlit.jinr.ru/en/ (accessed 15.11.2019).

[4] MICC - INFRASTRUCTURE, n.d. Available at: https://micc.jinr.ru/?id=29 (accessed 15.11.2019).

[5] lit.jinr.ru | LIT, n.d. Available at: http://lit.jinr.ru/en (accessed 11.18.19).

[6] Scientific Laboratory «Cloud technologies and Big Data analytics», n.d. Available at: https://www.rea.ru/ru/org/managements/unitscires/Laboratorija-Oblachnykh-tekhnologijj-i-analitiki-Bolshikh-dannykh/Pages/lotiabd.aspx (accessed 15.11.2019) (in Russian)

[7] Adam Gh., Bashashin M., Belyakov D., Kirakosyan M., Matveev M., Podgainy D., Sapozhnikova T., Streltsova O., Torosyan Sh., Vala M., Valova L., Vorontsov A., Zaikina T., Zemlyanaya E., Zuev M.. 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

[8] Hardware and software environment – Supercomputer "Govorun," n.d. URL http://hlit.jinr.ru/en/for_users_eng/hardware-and-software-environment_eng/ (accessed 18.11.2019).

[9] Ecosystem for tasks of machine learning, deep learning and data analysis – Supercomputer "Govorun", n.d. Available at: http://hlit.jinr.ru/en/ecosystem-for-tasks-of-machine-learning-deep-learning-and-data-analysis/ (accessed 18.11.2019).

[10] JupyterHub, n.d. Available at: https://jhub.jinr.ru/hub/login (accessed 15.11.2019).

[11] Projects · Dashboard, n.d. GitLab. Available at: https://gitlab-hybrilit.jinr.ru/ (accessed 15.11.2019).

[12] Korenkov V.V., Podgainy D.V., Streltsova O.I. Educational program on HPC technologies on the basis of the HybriLIT heterogeneous cluster (LIT JINR) // Modern Information Technology and IT-education. 2017. V. 13, no. 4, pp. 141-146 (in Russian)

[13] CUDA Toolkit, 2013. NVIDIA Developer. Available at: https://developer.nvidia.com/cuda-toolkit (accessed 18.11.2019).

[14] OpenCL - The open standard for parallel programming of heterogeneous systems, 2013. The Khronos Group. Available at: https://www.khronos.org/opencl/ (accessed 18.11.2019).

[15] Intel® MPI Library, n.d. Available at: https://software.intel.com/en-us/mpi-library (accessed 18.11.2019).

[16] Open MPI: Open Source High Performance Computing, n.d. Available at: https://www.open-mpi.org/ (accessed 18.11.2019).

[17] Enterprise Container Platform, n.d. Docker. Available at: https://www.docker.com/ (accessed 18.11.2019).

[18] Slurm Workload Manager - Overview, n.d. Available at: https://slurm.schedmd.com/overview.html (accessed 18.11.2019).