Parallel Computing: Technology Trends
I. Foster et al. (Eds.)

doi:10.3233/APC200046

The Architecture of Heterogeneous Petascale HPC RIVR

Miran ULBIN[a,1] and Zoran REN[a]
[a] *Faculty of Mechanical Engineering, University of Maribor, Maribor, Slovenia*

Abstract. The EU has launched EuroHPC Joint Undertaking initiative plan to build an exascale HPC by 2025. A petascale HPC will be built in Slovenia in a concerted effort by 2020. The aim is to establish a national HPC system by own design with low maintenance and power consumption costs of the system. The HPC architecture will be unique, built from the off-the-shelf state-of-the-art components, and will operate using the open-source system software. A small HPC prototype system of about 200 TFLOP/s computing capability will be built in the first phase to test various computing nodes and components, which will be later integrated into a full-scale supercomputer with approx. 2 PFLOP/s. The throughput of Infiniband and Ethernet interconnect solutions will be of particular interest. The presentation is first focused on the architecture of HPC prototype consisting of 82 heterogeneous nodes based on double AMD Epyc and Intel Xeon SCL processors in combination with GPU nodes, with the discussion of possible variations of interconnect configurations. The network configuration of full-scale HPC with 600 AMD Epyc nodes, GPU nodes and large hard drive storage with a connection to HPC prototype will be discussed next. Possibilities of open source software for operating, provisioning and maintaining system, as well as flexibility and security of several options for user access, will be given in conclusion.

Keywords. HPC architecture, petascale, heterogenous CPU-GPU, open-source

1. Introduction

Roadmaps are clear for building exascale HPC in China by 2020 [1] and the USA by 2021 [2], while the EU has the plan to build exascale HPC by 2025 [3]. Slovenia has joined to Declaration Cooperation framework on High-Performance Computing in 2017 with an obligation to build integrated high-performance computing infrastructure, which will enable a competitive level of research and industry.

There is some computer infrastructure in Slovenia, which could be classified as high-performance computing. Some systems had begun in the late 20th century as research projects building computer grids and HPC systems, but most systems were small scale. Largest HPC in Slovenia today is performing with speed about 43 TFLOPS. Therefore, it was decided to build an HPC in the scope of petascale. As this is considerably more than any other system in Slovenia, it will be established as a national HPC system.

[1] Corresponding Author, Miran Ulbin, Faculty of Mechanical Engineering, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia; E-mail: miran.ulbin@um.si.

There are two ways to build an HPC system. The easiest way is to buy an existing system from vendors like IBM, HP, Cray, NVIDIA, etc. Beside of higher initial and maintenance cost, such a system is also more rigid than custom build system. Various researchers, from the very different research field, will use national HPC system. The system should also provide a large number of services for the research community. One major requirement is the provision of a national repository of open access publications and research data. HPC vendors usually compete with benchmark test, which will prove that the system efficiently solves customer problems. With a diversity of research areas from massive parallel simulations to artificial intelligence problems or big data analysis, it is impossible to identify a simple benchmark test.

Because of that, HPC-RIVR was designed to be custom build heterogeneous [4] and very flexible to cover various research areas. Besides that, it should include large storage for research data. Another aspect is the maintenance cost of HPC-RIVR system. Power consumption is the main cause of HPC running cost and one of the major requirement for equipment was power efficiency. Another aspect was compatibility with the majority of software and efficiency of various hardware solutions. After comparing different solutions using benchmarks [5], the initial design was drawn. Some unknown remains, so it was decided to build HPC prototype first, to test some hardware and software configurations.

The system software is also not vendor based and only open-source software will be used. This requires the development of custom-based procedures and scripts for provisioning and maintenance of system software. Beside standard batch submission of jobs, user-friendly interfaces for HPC usage are developed [6]. HPC prototype is dedicated to development and testing system software, user interfaces and special configurations, while HPC-RIVR is designed as a production system where tested changes will be applied when needed.

2. HPC prototype

HPC prototype was designed to test various configurations and setup. Size of HPC prototype is about 10% of the size of the complete HPC system. It is built with a flexible design with equipment installed in a container presented if figure 1. Container with a length of 6,5 m and width and height of 2,9 m is equipped with all support system needed for HPC computer. Several racks are installed in the container, as can be seen from figure 1. Four racks are dedicated to HPC servers and two racks include a power supply with UPS and supporting systems for fire alarm and remote surveillance. The redundant mechanical and electrical cooling system is installed in the container so that there is a warm zone on one side of racks and cool zone on the other side.

Network connection of HPC prototype is realized with optical connectors providing 10 Gb/s connections to the internet, which will be later replaced by 100 Gb/s connection to HPC-RIVR system. Interconnect is built with two Mellanox 3800 Ethernet 100Gb/s switches each with 64 ports. As every HPC component is at a short distance to the switch, DAC copper cables using QSFP28 connector are used for the connection. Switches also enable connection of 10GBase-LR SFP+ module using an adapter which converts 40 Gb/s speed to 10 Gb/s and enables connection to existing 10 Gb/s switches. Three additional Ethernet Quanta T148-LY4R 1 Gb/s switches are used for network management connection.

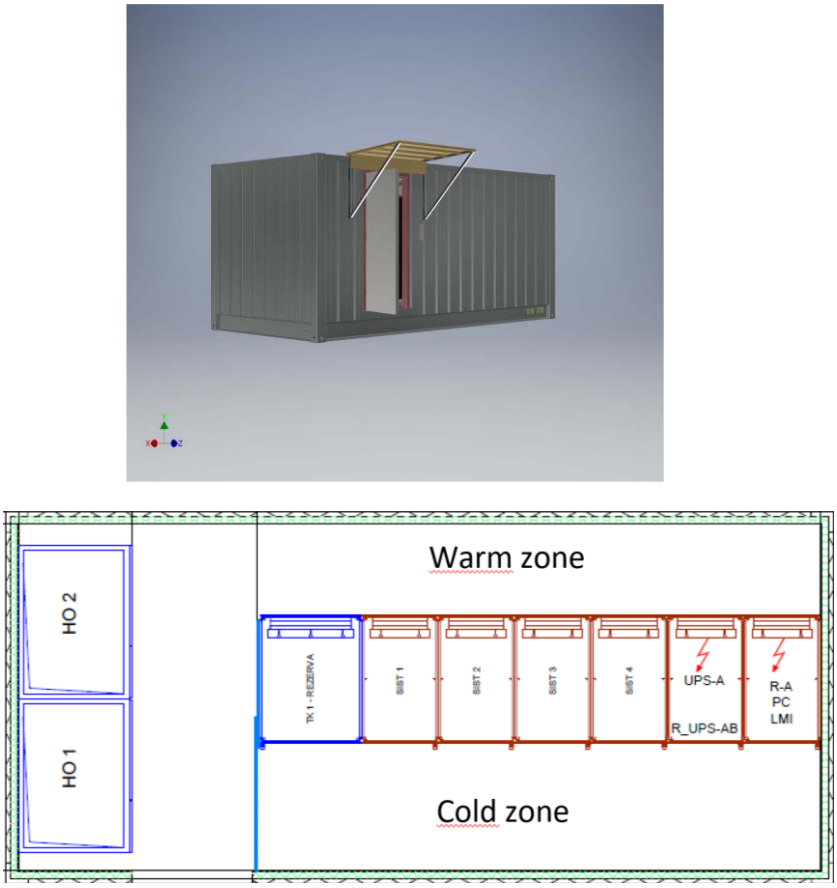

Figure 1. Container for HPC prototype.

Part of the HPC prototype is interconnected using Mellanox 8700 HDR100 100 Gb/s Infiniband switch. The first draft of the system considered EDR Infiniband switch with only 36 ports therefore, there were not enough ports for the whole system. With the new switch, it is possible to connect all nodes to Infiniband as each port can connect to two HDR100 interface cards. Also, previously planned interfaces in form of single port ConnectX-5 VPI network card were replaced by new ConnectX-6 interface cards which also have two modes of operation, one for 100 Gb/s Ethernet and one for HDR100 Infiniband. The distance of Ethernet connection between switch and component is rather small, so copper cables with QSFP56 HDR connectors were used.

The main purpose of the Infiniband switch is the comparison of performance between Ethernet and Infiniband interconnect. Although there are some results of Infiniband versus Ethernet comparisons [7], it is strongly dependent on packet size. In the report [7] the speed is almost identical for small packets while there is a huge difference for larger packets. Therefore, our purpose is to test network speed using typical applications utilizing either Infiniband or Ethernet interconnect. Results of comparison will influence the architecture of HPC-RIVR regarding interconnecting and it is possible that this will enable more cost-effective architecture of a network for our petascale HPC-RIVR.

HPC prototype is presented in figure 2, where it is shown that some nodes are connected only to Ethernet switch, while others are connected to Ethernet and Infiniband switch as discussed above. During tests, this configuration could change and SSD storage servers might be connected to Infiniband switch for testing purposes.

There are three computers of figure 2, which are designated with the name SuperMicro Server. These servers have the role of a head node and general-purpose servers with different services like Slurm, Web servers, etc. Each SuperMicro Server consists of two AMD Epyc 16C/32T 7301 2.2G 64M processors on board with 256 GB DDR4-2666 LRDIMM ECC, with two 480 GB SSD SATA drive configured in RAID 1 and with two 100 Gb/s Eth/IB Mellanox ConnectX-6 VPI network card. Each ConnectX-6 VPI network card could be connected to 100 Gb/s Ethernet switch or HDR Infiniband switch.

There are also three computers in figure 2, with the name SuperMicro Storage. These are storage servers with one AMD EPYC 24C/48T 7401P 2.0G 64M processor on board, with 256 GB DDR4-2666 LRDIMM ECC, with two 480 GB SSD SATA drive configured in RAID 1 and with two 100 Gb/s Eth/IB Mellanox ConnectX-6 VPI network card. Although there is a connection to Ethernet switch in figure 2, network cards can also connect to Infiniband switch. The better configuration will be considered for future use. Each storage servers contains 24 1,92 TB SSD drives beside two system drives. Storage servers are organized using CEPH [8], which also will provide testing for implementation on petascale HPC. The total size of SSD storage is 138 TB, which gives 69 TB of risky storage size and 46 TB of safe storage size [9].

Node SuperMicro GPU in figure 2 is a compute node with NVIDIA graphical accelerator boards. Each node consists of two Intel Gold SKL-SP 6128 6C/12T 3.4G processors, 256 GB DDR4-2666 LRDIMM ECC RAM, with two 480 GB SSD SATA drive configured in RAID 1, two 100 Gb/s Eth/IB Mellanox ConnectX-6 VPI network cards and four NVIDIA TESLA V100 32G PCI-E x16 boards. Each NVIDIA TESLA V100 board have 5120 cores enabling 7 double-precision TFLOPS. Usually, NVIDIA graphics boards are used with combination with INTEL processors and usage of AMD processors is not well proven with this combination. GPU nodes are connected to HDR100 Infiniband and to 100 Gb/s Ethernet and there is also plan to compare computational speed using one or another network.

Ethernet and Infiniband compute nodes in figure 2 are AMD processor-based. Because the first draft of HPC prototype was planned with one EDR Infiniband switch, there were only 36 ports available for connecting Infiniband nodes. Change to HDR Infiniband switch occurred at last moment, therefore only 28 nodes were equipped with two interface cards and other nodes will be equipped with additional interface cards in case it will be shown that HDR100 Infiniband offers considerable benefits over 100 Gb/s Ethernet.

Each compute node consist of two AMD EPYC 32C/64T 7501 2.0G 64M processors, 512 GB DDR4-2666 LRDIMM ECC RAM and two 960 GB SSD SATA drive configured in RAID 1. There are 28 nodes which have two 100 Gb/s Eth/IB Mellanox ConnectX-6 VPI network cards and are connected to HDR100 Infiniband and 100Gb/s Ethernet. Other 48 nodes have only one ConnectX-6 VPI network card and are connected only to 100 Gb/s Ethernet. Nodes are built into 2U chassis, which can contain four nodes. This enables compact build in half of the space required for 1U/1N chassis.

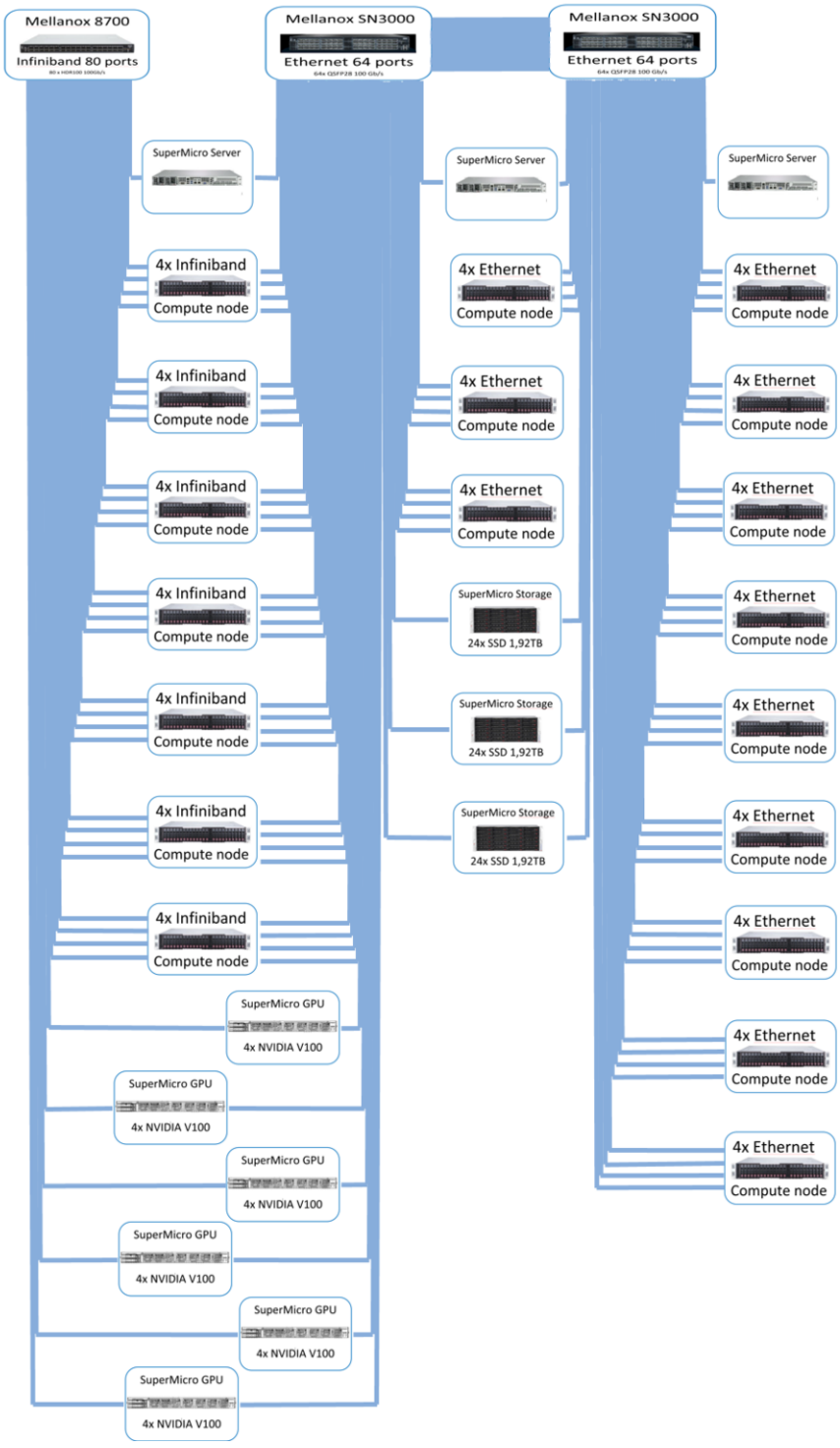

Figure 2. HPC prototype.

Power consumption is maximum 2,2 kW per chassis and total maximum power consumption for AMD compute nodes is 41.8 kW. GPU compute nodes maximum total power consumption is 12 kW, storage server consumes 6 kW, servers 1,5 kW and switches 3 kW. Maximum total power consumption without cooling is about 64,3 kW.

System software for HPC prototype is based on open source solutions. Head node operating system is CentOS [10] while compute nodes will use the most efficient UNIX flavour available. Provisioning is based on Foreman [11] with a combination of Puppet [12] customization. There will be several configurations tested on HPC prototype with either batch usage using Slurm [13] and Nordungrid ARC [14] and more direct access via web interfaces and virtualization. Most of the work will be based on running applications built-in containers like Singularity [15].

HPC prototype system was installed and is in the testing phase since July 2019. Container with cooling where cabinets with HPC servers and nodes were stacked is shown in figures 3 and 4.

Figure 3. HPC prototype container.

Figure 4. HPC cabinets with servers and nodes.

3. Petascale HPC-RIVR

While HPC prototype is assembled and different hardware and software solutions are tested, complete HPC-RIVR is still in design. Reason for that is partly in the fact that results of testing will influence the final design and partly because technology is currently changing. There are new processors with more cores presented and will be in production this year. Infiniband speed will double this year and products with new PCI 4.0 bus will emerge. Therefore, the design of HPC-RIVR is somehow outlined, but it is very flexible to adapt to new technology and findings of tests on HPC prototype. Therefore, the final version is still under consideration and might be influenced by the vendor's proposals.

Rough design of HPC-RIVR consist of Ethernet network and Infiniband interconnect of head nodes, servers, compute nodes with an x86 processor, compute node with GPUs, SSD storage servers and hard disk storage servers. Network scheme of the design is shown in figure 5.

Ethernet is designed with redundant switches which should have large buffers to allow high throughput as it is expected that a large amount of data will be transferred to and from HPC. The consequence of that is more latency, but traffic between compute nodes should flow on Infiniband with low latency. If testing on HPC prototype proves that 100 Gb/s Ethernet switches with low latency can compete with Infiniband, the design will change accordingly and the cost of interconnect will be much lower. Research results from literature shows, that latency of Infiniband is still much lower than Ethernet, hence, a dual network is required. Additionally, there is the third network for management, which is served by 1 Gb/s switches and is not shown in figure 5.

The speed of the Ethernet network with presented configuration can be much lower at compute nodes and hard disk storage servers. It is designed with speeds of 25 Gb/s resulting in a lower cost of Ethernet network. Hard disk storage servers are connected with redundant connections, which on one hand double the speed and on the other hand makes the network more reliable.

Infiniband interconnect is based on HDR 200 Gb/s, but 200 Gb/s speed will be used only between switches. Nodes will use HDR 100 Gb/s interfaces and for HPC-RIVR 24 HDR 200 Gb/s switches are required for level 1 and 10 HDR 200 Gb/s switches for level 2 switch with the 3-1 blocking scheme. The total number of nodes connected to Infiniband is about 700, which requires 700 cables or optical fibres, while the connection between level 1 and level 2 switch requires 240 cables.

There are 30 general-purpose servers, which will be used as head nodes or dedicated nodes for services like scheduling, maintenance or user interfaces. The configuration of such server is identical as SuperMicro servers in HPC prototype, which is described above.

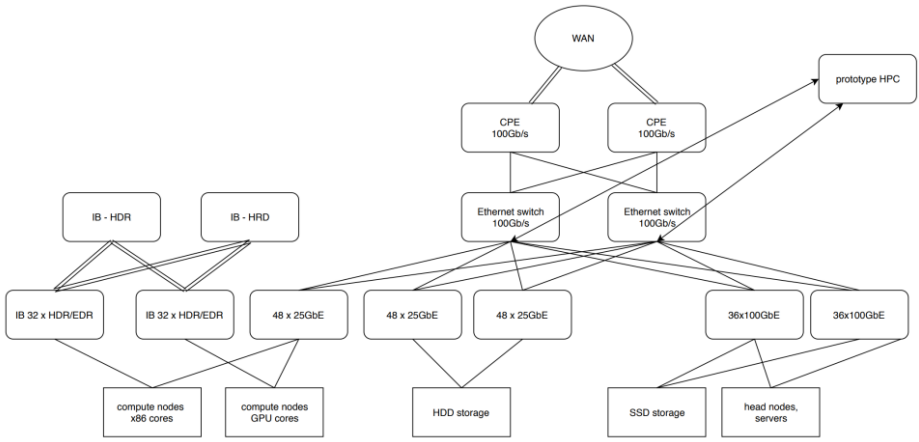

Figure 5. HPC-RIVR design.

In figure 5 HDD storage and SSD storage is presented in the bottom row. SSD storage is runtime storage for program and user data. SSD storage is designed with 22 servers each containing 24 SSD drives each with 1,92 TB space. The server is identical to SuperMicro storage in HPC prototype. Total SSD storage is therefore about 1 PB of raw space and will be configured using CEPH, which gives 483 TB of safe space. Similarly, HDD servers are designed with 20 servers where each is containing 90 hard disks each with 12 TB space. Storage servers will be configured using CEPH, with total raw space of 21.6 PB with 10.26 PB of safe space with two replicas.

GPU compute nodes are identical as in HPC prototype with four NVIDIA V100 boards. There is a total of 30 GPU servers in the design, with a total of 120 NVIDIA V100 boards. Compute nodes with x86 cores are same as Infiniband nodes in HPC prototype in the current design of HPC RIVR. As new processor with 64 cores emerges now, compute nodes might be different and there will be fewer nodes and required network connections.

The expected power consumption of HPC-RIVR is about 600 kW and an additional 200 kW for cooling. Power supplies and cooling equipment is currently under construction and dedicated space is prepared for installation in the next months. In figure 6 preparation for building HPC-RIVR is shown.

Figure 6. The equipment and dedicated space for HPC-RIVR.

4. Conclusions

Designing petascale HPC from scratch is a difficult task, comparing with the selection of the final product offered by established vendors. The final design has required an initial design of HPC prototype, where some results and solutions are expected. HPC prototype will also enable a more realistic estimate of our goal of achieving PFLOPS operation. Theoretically, designed hardware should result in value above 1 PFLOPS, but the final number will be calculated with the LINPACK [16] test.
As hardware is near the final design and system software will be finalized in HPC prototype, the majority of the work is still ahead. HPC-RIVR will be used by a variety of researchers with specific needs and requirements. Teams of researchers are currently building different applications for these needs and creating web portals and special user interfaces, which will enable user-friendly use of HPC. Findings during building HPC-RIVR and various software solutions could be used on the way to European exascale HPC.

Acknowledgements
The authors would like to thank the Ministry of Education, Science and Sport of the Republic of Slovenia and to the European Union – European Structural and Investment Fund for financial support.

References

[1] China Navigating The Homegrown Waters For Exascale (21.03.2019). Retrieved from: https://www.nextplatform.com/2018/11/15/china-navigating-the-homegrown-waters-for-exascale/ .
[2] The Roadmap Ahead For Exascale HPC In The US (21.03.2019). Retrieved from: https://www.nextplatform.com/2018/03/06/roadmap-ahead-exascale-hpc-us/ .
[3] R. Ammendola et al., 'Large Scale Low Power Computing System - Status of Network Design in ExaNeSt and EuroExa Projects', arXiv e-prints, p. arXiv:1804.03893, Apr. 2018.
[4] S. Lee, H. Seo, H. Kwon, and H. Yoon, 'Hybrid approach of parallel implementation on CPU–GPU for high-speed ECDSA verification', The Journal of Supercomputing, Jan. 2019.
[5] Kutzner C., Páll S., Fechner M., Esztermann A., L. de Groot B., Grubmüller H., ' More Bang for Your Buck: Improved use of GPU Nodes for GROMACS 2018', arXiv:1903.05918 [cs.DC], 2019.
[6] N. Fareghzadeh, M. A. Seyyedi, and M. Mohsenzadeh, 'Toward holistic performance management in clouds: taxonomy, challenges and opportunities', The Journal of Supercomputing, vol. 75, no. 1, pp. 272–313, Jan. 2019.
[7] Erickson, K., Kachelmeier, L., Van Wig, F.. 2016. Comparison of High Performance Network Options: EDR InfiniBand versus 100Gb RDMA Capable Ethernet; Poster; accepted Supercomputing Conference, SC'16 (21.03.2019). Retrieved from: https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-16-26126 .
[8] CEPH (21.03.2019). Retrieved from: https://ceph.com/ .
[9] CEPH storage calculator (21.03.2019). Retrieved from: http://florian.ca/ceph-calculator/ .
[10] CentOS (21.03.2019). Retrieved from: https://www.centos.org/ .
[11] Foreman (21.03.2019). Retrieved from: https://www.theforeman.org/ .
[12] Puppet (21.03.2019). Retrieved from: https://puppet.com/ .
[13] Slurm (21.03.2019). Retrieved from: https://slurm.schedmd.com/ .
[14] Nordugrid ARC (21.03.2019). Retrieved from: http://www.nordugrid.org/ .
[15] Singularity (21.03.2019). Retrieved from: https://www.sylabs.io/ .
[16] HPL - A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers (21.03.2019). Retrieved from: http://www.netlib.org/benchmark/hpl/ .