DOI: 10.14529/jsfi230208

The High Performance Interconnect Architecture for Supercomputers

Alexey S. Simonov[1,2]***, Alexander S. Semenov***[1]***, Andrey N. Shcherbak***[1]***, Ivan A. Zhabin***[1]

In this paper, we introduce the design of an advanced high-performance interconnect architecture for supercomputers. In the first part of the paper, we consider the first generation high-performance Angara interconnect (Angara G1). The Angara interconnect is based on the router ASIC, which supports a 4D torus topology, a deterministic and an adaptive routing, and has the hardware support of the RDMA technology. The interface with a processor unit is PCI Express. The Angara G1 interconnect has an extremely low communication latency of 850 ns using the MPI library, as well as a link bandwidth of 75 Gbps. In the paper, we present the scalability performance results of the considered application problems on the supercomputers equipped with the Angara G1 interconnect. In the second part of the paper, using research results and experience we present the architecture of the advanced interconnect for supercomputers (G2). The G2 architecture supports 6D torus topology, the advanced deterministic and zone adaptive routing algorithms, and a low-level interconnect operations including acknowledgments and notifications. G2 includes support for exceptions, performance counters, and SR-IOV virtualization. A G2 hardware is planned in the form factor of a 32-port switch with the QSFP-DD connectors and a two-port low profile PCI Express adapter. The switches can be combined to 4D torus topology. We show the performance evaluation of an experimental FPGA prototype, which confirm the possibility of implementing the proposed advanced high performance interconnect architecture.

Keywords: interconnect, high performance computing, supercomputer, Angara.

Introduction

An interconnection network (interconnect) is a critical supercomputer component to achieve high scalability and performance on different applications. Modern supercomputer applications include mathematical physics, business analytics and machine learning problems.

Analysis of world experience in designing custom interconnection networks for high-end supercomputers, primarily the IBM BlueGene L/P/Q series [6, 8, 11] and Cray SeaStar/Gemini [1, 5, 7], simulation results [20] allowed to design the principles of operation of the first generation of the high-performance Angara interconnect (Angara G1) [21]. Series-produced Angara G1 hardware was presented in 2013. The Angara G1 interconnect is a direct network, it supports topologies from 1D-mesh to 4D-torus and provides the possibility of building a supercomputer with a size of up to 32K nodes.

The operation experience of existing high-performance interconnect solutions allows to develop architecture features and design the principles of operation of an advanced high-performance interconnection network for supercomputers.

The paper is structured as follows. Section 1 describes the architecture and the performance results of the Angara G1 interconnect. Section 2 presents the architecture and features of the advanced high-performance interconnect for supercomputers, and provides the performance evaluation obtained on a FPGA prototype of the proposed interconnect architecture. Conclusion summarizes the paper and points directions for further work.

[1] Federal State Unitary Enterprise “Russian Federal Nuclear Center-Zababakhin All – Russia Research Institute of Technical Physics”, Snezhinsk, Russia

[2] Federal State Budgetary Educational Institution of Higher Education Moscow Aviation Institute (National Research University), Moscow, Russia

1. The First Generation Angara Interconnect

The choice of the interconnect topology determines the range of applicable routing algorithms and methods for solving the main problems of communication networks: deadlocks, starvation and livelocks. As a research result, the following decisions were made on the Angara G1 interconnect architecture:

- torus topology;
- the deterministic directional order routing algorithm with a fixed order of directions [10], a direction bit (dirbit) routing rule does not allow both positive and negative directions of a dimension in a route. There is a possibility of the first and last steps of a route in an arbitrary positive and negative directions for bypassing failed nodes and links [18];
- the bubble routing method [2, 16] is used to avoid deadlocks in a ring (movement without changing a direction). The direction order routing [9] garantees no deadlocks in torus directions. The first and last steps can violate the direction order rule, the routing table generation algorithm provides deadlock freedom [12]. The adaptive routing has the ability by a timeout to switch to a nonblocking deterministic virtual channel in case of potential deadlock;
- a minimal full-adaptive routing algorithm with assignment of possible directions;
- virtual subnets for deterministic, adaptive and broadcast routing based on virtual channels;
- a virtual cut-through flow control method;
- the used routing algorithms are minimal, i.e., each step of a packet route reduces the distance to a destination node. Minimal routing provides livelock freedom;
- fair arbitrage algorithms allow to avoid starvation;
- the PCI Express interface with CPU;
- a network adapter at the hardware level supports remote direct memory access (RDMA) write, read, and atomic operations;
- hardware and software support for the main functions of the SHMEM and MPI libraries;
- each node has a dedicated memory region available for remote access from other nodes to support, OpenSHMEM and partitioned global address space (PGAS);
- router implementation in the form of custom ASIC.

On the basis of custom designed router ASIC the following options of the Angara G1 interconnect hardware were released:

- a full-height full-length PCI Express adapter, which provides the ability to combine directly up to 32 thousand computing nodes of a supercomputer into a topology up to 4D torus topology $8 \times 16 \times 16 \times 16$;
- a 24-port switch for installation in a 19-inch rack and a low profile PCI Express adapter for installation in supercomputer nodes. This option provides the ability to combine up to 2048 computing nodes with 2D torus topology of switches.

1.1. Experimental Results and Performance Evaluation

The first computing system equipped with the Angara G1 interconnect is the Angara-C1 36-node computing cluster [3], designed for hardware and system software development of the interconnect.

During the development, different options for combining computing nodes into a torus topology were tested, system software was debugged, including the Slurm job scheduling system, a plugin for the Zabbix monitoring system. Figure 1 shows the Angara G1 system software stack.

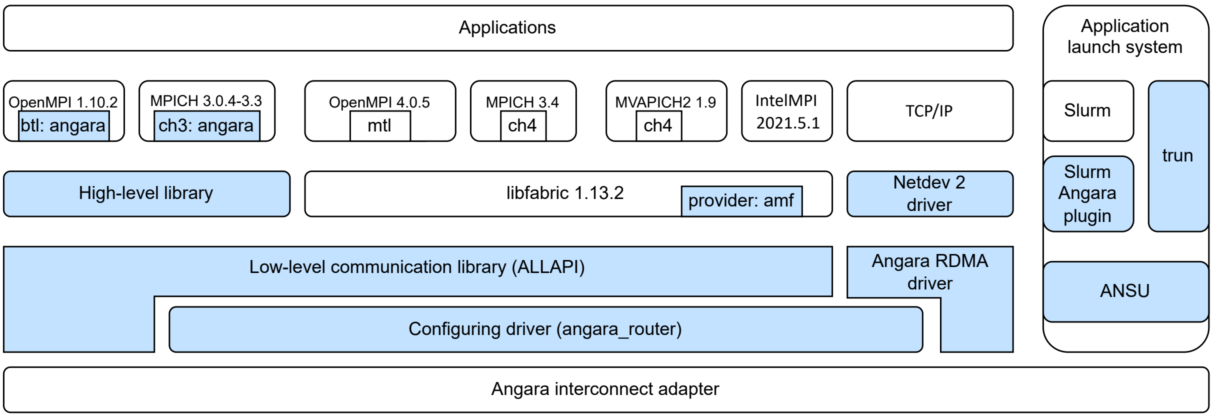


Figure 1. The Angara G1 system software stack

A supercomputer series of difference performance based on the Angara G1 interconnect were produced, including the Desmos [22] and Fisher [19, 23] supercomputers installed at the JIHT RAS.

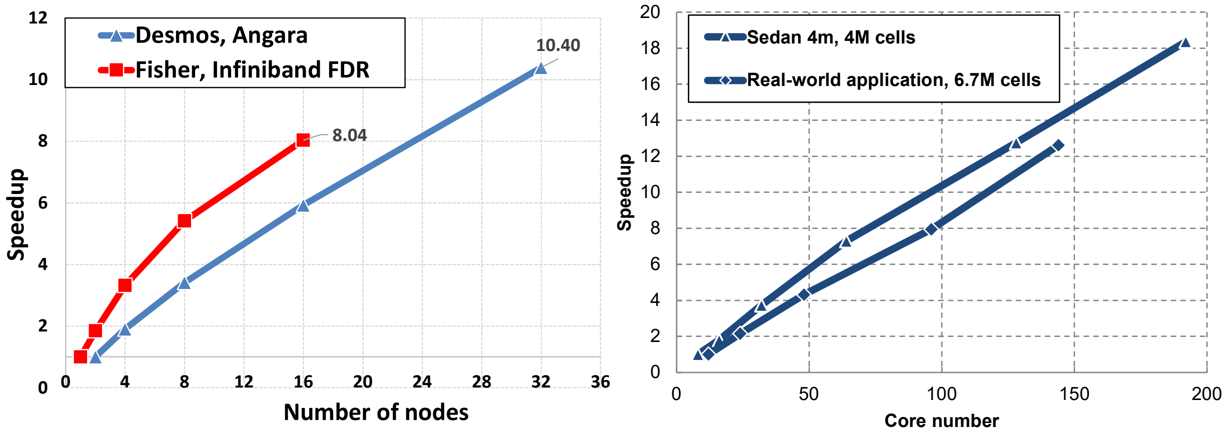
Several years of operation and maintenance of supercomputers based on the the Angara G1 interconnect made it possible to identify both the advantages and disadvantages of Angara G1. The advantages include:

- extremely low communication latency, for example on Desmos the obtained latency is 850 ns using the MPI library on the osu_latency benchmark. In comparison with other commercially available communication networks, on Angara G1 the latency is stable with a network load increase and with an increase of the supercomputer node number;
- 75 Gbit/s link bandwidth, which does not change with increasing distance between supercomputer nodes. This feature is ensured both by the chosen topology and by a good balance of the stages of the entire Angara G1 pipeline.

The mentioned advantages made it possible to obtain good performance scalability of benchmarks and applications on the supercomputers based on the Angara G1 interconnect [4, 13–15, 17, 22–24]. Figure 2 presents performance scalability obtained on the Desmos, Fisher and Angara C1 supercomputers on quantum mechanical calculation software called VASP and a fluid simulation software.

The main disadvantages of the Angara G1 interconnect are the following:

- network connectivity degradation in case of computing node and communication cable failure, associated with the limited capabilities of the routing algorithms for bypassing the failed network resources;
- supercomputer performance degradation in the situation of interconnect congestions, associated with the features of the deterministic and adaptive routing algorithms;
- insufficient hardware support of fault tolerance in the router ASIC;
- insufficient hardware support of modern multi-core processors;
- insufficient hardware support for GPU integration;
- lack of hardware support of the TCP/IP protocol stack;
- lack of hardware support of virtualization;



(a) Quantum mechanical calculations VASP, Desmos

(b) A fluid simulation software, Angara-C1

Figure 2. Performance result scaling of the Angara based supercomputers on the simulation application softwares

- insufficient hardware support for supercomputer monitoring and control systems, which is especially important for high-end large supercomputers;
- large physical dimensions of the full-height full-length PCI Express Angara G1 adapter;
- limitation on the number of supercomputer nodes when using the Angara G1 24-port switch and the low profile PCI Express adapter.

These disadvantages were taken into account when developing an advanced high-performance interconnect architecture.

2. Advanced Interconnect Architecture

An advanced G2 high-performance interconnect is designed to combine computing nodes of supercomputers up to a subexascale performance level. During its development, technical risks were minimized, the advantages and disadvantages of the Angara G1 interconnect were taken into account. In addition to the goal of ensuring high performance and high scalability of supercomputers built using the developed interconnect, the goal was to extend the product segments of the interconnect. The intended product segments are not only supercomputers, but also storage systems, business analytic processing systems and data centers.

To ensure the scalability of application performance and support new product areas, it was necessary to:

- use more flexible routing algorithms;
- have low communication latency;
- increase the message rate;
- extend the interconnect functional capabilities and features.

2.1. Interconnect Architecture

The essential differences between the G2 interconnect architecture and the existing solution are the following:

- flexible options for combining available topologies from 1D mesh to 6D torus, topologies with shifted connections are also supported;

- an advanced deterministic delta routing algorithm that provides more route options when bypassing failed or congested network sections;
- an advanced zone-adaptive routing algorithm;
- an extended low-level network operation set, including acknowledgements, notifications and synchronization operations;
- multihost hardware support, that allows to divide the PCI Express interconnect adapter interface into 2 or 4 independent interfaces and connecting different computing nodes via them;
- hardware support for modern multicore processors, 128 independent injection pipelines are implemented;
- hardware methods to solve the starvation network problem;
- advanced fault tolerance support, as well as monitoring and control systems for supercomputers, including support for exceptions, traps and performance counters;
- hardware support for SR-IOV virtualization, up to 15 virtual functions are supported in each PCI Express endpoint;
- hardware support for hot swap of failed interconnect hardware.

The most important difference between the G2 interconnect and the existing solution is 6D torus topology. This architecture feature follows the global trend of high-radix communication networks to increase the connectivity of each node or switch.

The advanced delta routing algorithm implements the dimension order routing rule. In the dirbit routing algorithm implemented in the Angara G1 interconnect a packet header stores not only a target node address, but also a bit that specify a torus direction, in which a packet will move. In contrast to the dirbit routing algorithm, in the delta routing algorithm a packet header for each torus coordinate stores only the difference between the current and a target node, which allowed to reduce bit number in the packet header. Also, the delta routing algorithm is more flexible and allows you to transmit a packet both in a positive and in a negative torus direction, and by a large number of steps, which allows to significantly extend the possibilities of bypassing failed supercomputer nodes or communication cables. In addition, the delta routing algorithm allows to build shifted connections in a torus topology, which in some cases can reduce the network diameter and improve communication latency.

The second important difference is the possibility of a zone-adaptive routing. In the Angara G1 interconnect, when a packet moves using adaptive routing, in case of, for example, congestion or failure, the packet waits too long for a connection, the packet is transferred to the deterministic subnet, and it is impossible to return the packet back to movement with the adaptive routing. The zone-adaptive routing allows to specify in which directions a packet can be transmitted using an adaptive algorithm, and in which directions using the deterministic algorithm. Thus, the packet movement using the deterministic algorithm will be used only in hyperplanes in which there are problems, for example, failed computing nodes or communication channels.

The nomenclature of low-level G2 operations includes simple put, get, and atomic operations (including returning a value) in a memory of a remote node. In addition to the mentioned operations there are write to a segment operations, moreover three types of segments are available: number of incoming packets, aggregating segments and circular segments. Also there are operations with acknowledgements and notifications, including interrupts. The use of these types of operations allows to optimize the implementation of low-level software of the MPI and TCP/IP libraries, reduces the load on CPU and significantly improves the application perfor-

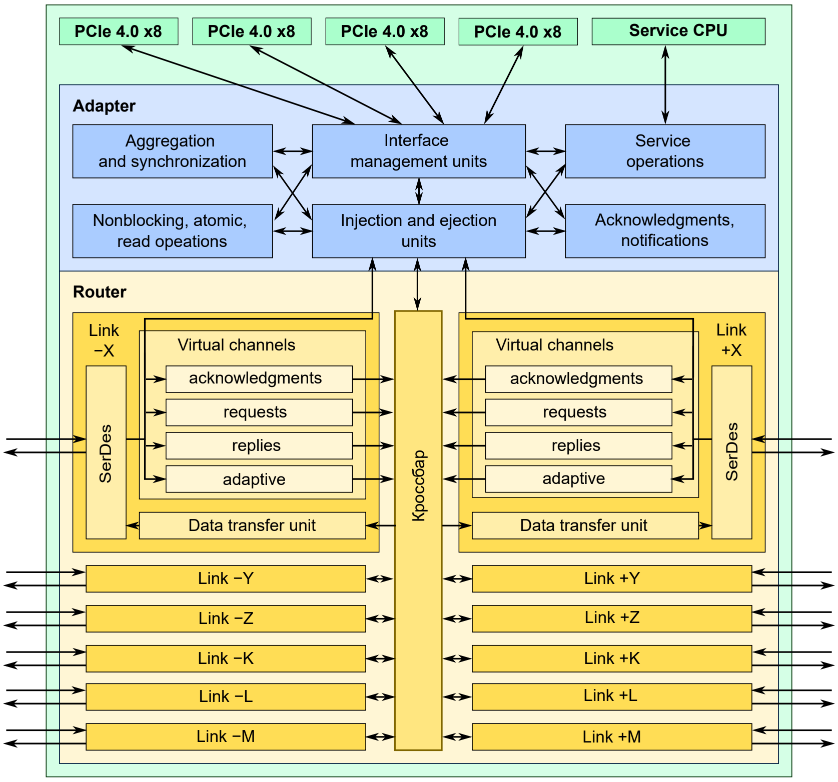


Figure 3. The G2 interconnect architecture

mance and scalability. The G2 interconnect supports the technology of direct memory access to the remote node (RDMA), which significantly affects the obtained performance not only in the high-performance computing systems, but also in storage and Big Data processing systems.

Figure 3 presents the G2 interconnect architecture, which consists of two parts: a router and an adapter. The router is intended for routing and transmission of packets between nodes, includes four virtual channels in each link, including: three deterministic virtual channels for request, response and confirmations subnets and an adaptive subnet. Deterministic virtual channels implement the deterministic delta routing algorithm with dimension-order routing and an order of dimensions can be set at system startup.

The adapter performs the injection of packets into the network and the processing of packets ejected from the network, as well as a set of service functions that ensure the minimization of traffic over the network and through the interface to the node processor. The adapter includes:

- interface control units, which support PCI Express protocols and implement virtual functions;
- injection and ejection units, that form packets for sending to the network and parse the headers of packets that came from the network;
- a service unit, that processes packets going to and from the service processor;
- aggregation and synchronization units, as well as confirmation and notification units, which support the execution of different network operations;
- non-blocking, atomic and read operation units, which support RDMA technology.

The G2 interconnect hardware is planned to be released in the form factor of a 32-port switch for installation in a standard 19-inch rack and a low profile PCI Express adapter for installation in supercomputer nodes. This option provides the possibility of combining up to several thousand computing nodes by connecting switches in a topology from 1D to 4D torus.

Each switch will include 32 ports with QSFP-DD connectors, and the low profile adapter will include 2 ports with QSFP-DD connectors and support multihost operation of several PCI Express interfaces.

2.2. Prototype Performance Evaluation

Development of the underlying architectural solutions of G2 was on a simulation model, as well as on a cluster with prototype network adapters based on the Virtex7 FPGA, each computing node includes 2x Intel Xeon CPU E5-2630 v3 processors @ 2.40 GHz, 32 GB of memory.

A prototype network adapter is made in the form factor of a full-length full-height PCI Express expansion card. Due to the technical limitations of the FPGA, PCI Express 2.0 x8 was used as an interface with CPU, there are four independent injection packet pipelines and 8 QSFP connectors were used for communication with other adapters.

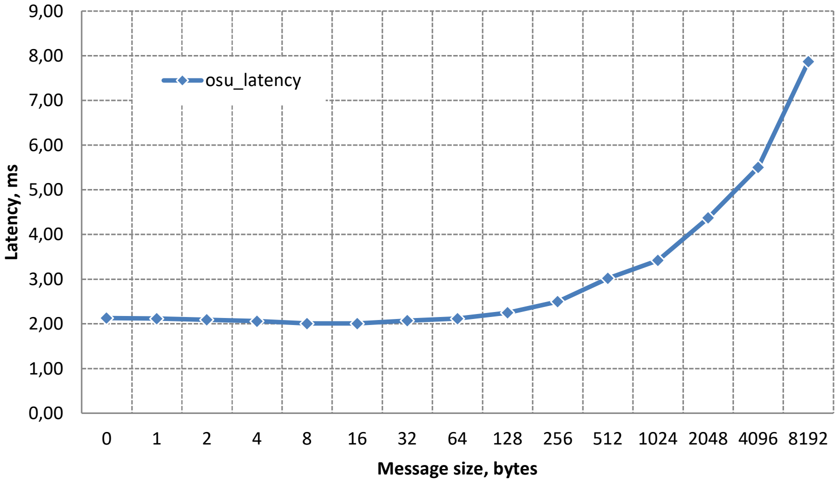


Figure 4. The osu_latency results obtained on two nodes of the M4 prototyping cluster

The frequency of the prototype adapter is 167 MHz, we obtained 2.01 us communication latency between two neighboring nodes on the osu_latency test using the MPI library. Figure 4 presents osu_latency results for different message sizes. The latency for each additional hop is no more than 0.5 us. The 30 Gbit/s bandwidth of the prototype adapter is limited by the performance of packet injection into the network, the overhead of the G2 interconnect data transmission protocol when transmitting large messages is less than 9% of the peak bandwidth of the communication channel.

Conclusion

In this paper, we introduce the main features of the advanced high-performance interconnect architecture for supercomputers, including topology, adaptive and deterministic routing algorithms, remote direct memory access technology.

We have considered the first generation high-performance Angara interconnect (Angara G1). Angara G1 has an extremely low communication latency of 850 ns, as well as a link bandwidth of 75 Gbps. However, it has insufficient hardware support for GPU integration, TCP/IP protocol stack, virtualization, for supercomputer monitoring and control systems, which is especially important for high-end large supercomputers.

Based on the performance evaluation and operation of the high-performance computing systems built with the Angara G1 interconnect solution, we present the main architectural features for the advanced interconnect for supercomputers. The proposed G2 architecture supports 6D torus topology, the advanced deterministic delta routing and zone adaptive routing algorithms, and extended low-level interconnect operations. G2 includes support for exceptions, performance counters, and the SR-IOV virtualization technology. The G2 hardware is planned in the form factor of a 32-port switch with the QSFP-DD connectors and a two-port low profile PCI Express adapter. We showed the performance evaluation of the experimental FPGA prototype. In future works, we plan to present more detailed performance evaluation of the FPGA prototype.

Acknowledgements

This work is supported by the Ministry of Industry and Trade of the Russian Federation under the complex project, which is performed by FSUE "VNIITF named after Academ. E. I. Zababakhin".

References

1. Abts, D.: The Cray XT4 and Seastar 3D torus interconnect (2011)

2. Adiga, N.R., Blumrich, M.A., Chen, D., *et al.*: Blue Gene/L torus interconnection network. IBM Journal of Research and Development 49(2.3), 265–276 (2005). https://doi.org/10.1147/rd.492.0265

3. Agarkov, A., Ismagilov, T., Makagon, D., *et al.*: Performance evaluation of the Angara interconnect. In: Proc. Int. Conf. on Russian Supercomputing Days, Moscow, Russia. pp. 626–639 (2016). https://russianscdays.org/files/pdf16/626.pdf, accessed: 2023-05-15 (in Russian)

4. Akimov, V., Silaev, D., Aksenov, A., *et al.*: FlowVision scalability on supercomputers with Angara interconnect. Lobachevskii Journal of Mathematics 39, 1159–1169 (2018). https://doi.org/10.1134/S1995080218090081

5. Alam, S.R., Kuehn, J.A., Barrett, R.F., *et al.*: Cray XT4: an early evaluation for petascale scientific simulation. In: Proceedings of the 2007 ACM/IEEE Conference on Supercomputing. pp. 1–12 (2007). https://doi.org/10.1145/1362622.1362675

6. Almasi, G., Asaad, S., Bellofatto, R.E., *et al.*: Overview of the IBM Blue Gene/P project. IBM Journal of Research and Development 52(1-2), 199–220 (2008). https://doi.org/10.1147/rd.521.0199

7. Alverson, R., Roweth, D., Kaplan, L.: The Gemini system interconnect. In: 2010 18th IEEE Symposium on High Performance Interconnects. pp. 83–87. IEEE (2010). https://doi.org/10.1109/HOTI.2010.23

8. Chen, D., Eisley, N., Heidelberger, P., *et al.*: The IBM Blue Gene/Q interconnection fabric. IEEE Micro 32(1), 32–43 (2011). https://doi.org/10.1109/MM.2011.96

9. Dally, W.J., Seitz, C.L.: The torus routing chip. Distributed computing 1(4), 187–196 (1986). https://doi.org/10.1007/BF01660031

10. Duato, J., Yalamanchili, S., Ni, L.: Interconnection networks. Morgan Kaufmann (2003)

11. Gara, A., Blumrich, M.A., Chen, D., *et al.*: Overview of the Blue Gene/L system architecture. IBM Journal of research and development 49(2.3), 195–212 (2005). https://doi.org/10.1147/rd.492.0195

12. Mukosey, A., Simonov, A., Semenov, A.: Extended routing table generation algorithm for the Angara interconnect. In: Russian Supercomputing Days. pp. 573–583. Springer (2019). https://doi.org/10.1007/978-3-030-36592-9_47

13. Nikolskiy, V., Pavlov, D., Stegailov, V.: State-of-the-art molecular dynamics packages for GPU computations: Performance, scalability and limitations. In: Russian Supercomputing Days. pp. 342–355. Springer (2022). https://doi.org/10.1007/978-3-031-22941-1_25

14. Ostroumova, G., Orekhov, N., Stegailov, V.: Reactive molecular-dynamics study of onion-like carbon nanoparticle formation. Diamond and Related Materials 94, 14–20 (2019). https://doi.org/10.1016/j.diamond.2019.01.019

15. Polyakov, S., Podryga, V., Puzyrkov, D.: High performance computing in multiscale problems of gas dynamics. Lobachevskii Journal of Mathematics 39(9), 1239–1250 (2018). https://doi.org/10.1134/S1995080218090160

16. Puente, V., Izu, C., Beivide, R., *et al.*: The adaptive bubble router. Journal of Parallel and Distributed Computing 61(9), 1180–1208 (2001). https://doi.org/10.1006/jpdc.2001.1746

17. Pugachev, L., Umarov, I., Popov, V., *et al.*: PIConGPU on Desmos supercomputer: GPU acceleration, scalability and storage bottleneck. In: Russian Supercomputing Days. pp. 290–302. Springer (2022). https://doi.org/10.1007/978-3-031-22941-1_21

18. Scott, S.L., *et al.*: The Cray T3E network: adaptive routing in a high performance 3D torus (1996)

19. Shamsutdinov, A., Khalilov, M., Ismagilov, T., *et al.*: Performance of supercomputers based on Angara interconnect and novel AMD CPUs/GPUs. In: International Conference on Mathematical Modeling and Supercomputer Technologies. pp. 401–416. Springer (2020). https://doi.org/10.1007/978-3-030-78759-2_33

20. Simonov, A.: Simulation model of high-speed Angara communication network with kd-tor topology. Trudy MAI (109), 22–22 (2019). https://doi.org/10.34759/trd-2019-109-22, (in Russian)

21. Simonov, A., Makagon, D., Zhabin, I., *et al.*: The first generation of Angara high-speed interconnect. Science Technologies 15(1), 21–28 (2014) (in Russian)

22. Stegailov, V., Dlinnova, E., Ismagilov, T., *et al.*: Angara interconnect makes GPU-based Desmos supercomputer an efficient tool for molecular dynamics calculations. The International Journal of High Performance Computing Applications 33(3) (2019). https://doi.org/10.1177/1094342019826667

23. Stegailov, V., Smirnov, G., Vecher, V.: VASP hits the memory wall: Processors efficiency comparison. Concurrency and Computation: Practice and Experience, p. e5136 (2019). https://doi.org/10.1002/cpe.5136

24. Tolstykh, M., Goyman, G., Fadeev, R., Shashkin, V.: Structure and algorithms of SLAV atmosphere model parallel program complex. Lobachevskii Journal of Mathematics 39(4), 587–595 (2018). https://doi.org/10.1134/S1995080218040145