Early evaluation of the hybrid cluster with torus interconnect aimed at cost-effective molecular-dynamics simulations

Vladimir V. Stegailov[1], Alexander Agarkov[2], Sergey Biryukov[2], Timur Ismagilov[2], Nikolai Kondratyuk[134], Evgeny Kushtanov[2], Dmitry Makagon[2], Anatoly Mukosey[2], Alexander Semenov[2], Alexey Simonov[2], Vyacheslav Vecher[13]

[1] Joint Institute for High Temperatures of RAS, Moscow, Russia
[2] NICEVT, Moscow, Russia
[3] Moscow Institute of Physics and Technology, Dolgoprudny, Russia
[4] National Research University Higher School of Economics, Moscow, Russia
v.stegailov@hse.ru

Abstract. In this paper, we describe the Desmos cluster that consists of 32 hybrid nodes connected by a low-latency high-bandwidth torus interconnect. This cluster is aimed at cost-effective classical molecular dynamics calculations. We present strong scaling benchmarks for GROMACS and compare the results with other HPC systems. Moreover, the cluster serves as a test bed for the Angara interconnect and verifies its ability to unite large MPP systems speeding-up effectively MPI-based applications. The interconnect is based on the Angara NIC that supports 3D and 4D torus network topologies. We describe the interconnect presenting typical MPI benchmarks.

1 Introduction

2 Related work

2.1 Scalability of classical MD on supercomputers

A lot of work on redesigning software for HPC with GPUs and testing new hardware is done by this moment. MD codes are also become rewritten for running on CPU+coprocessor machines which allow to investigate real size and timescales in MD models. Authors of [1] show how to double the performance of running MD biomolecular model in NAMD [2] with GPUs on Jaguar Cray XK6 machine with Gemini interconnect. In [3] it is described how to to achieve an excellent efficiency using cluster with GPUs by using suitable memory access patterns and mechanisms like CUDA streams and profiling tools. The errors propagation in GPU systems is studied in [4].

The Mont-Blanc project is aimed to developing of the supercomputer based on ARM cores [5]. The main advantages of this project are flexibility in designing an application-specific system-on-chip, in turn providing the possibility in balancing performance, energy-efficiency, and cost.

The topology of connections between supercomputer nodes also influences on the scaling of the current simulation. The comparison of torus and fat tree topologies is done in [6] based on SIESTA electronic structure code [7] for *ab initio* MD on six large-scale supercomputers. Torus topologies are showed to demonstrate a better scalability to large system sizes than those implementing fat tree topologies. Torus topology is implemented in 512-node Anton 2 MD supercomputer which allows to achieve simulation rates of multiple microseconds of physical time per day for systems with millions of atoms [8].

High MD code optimization for concrete supercomputer can provide better large-scaling performance than use of wide spread MD packages such as NAMD [2], GROMACS [9], CHARMM [10], LAMMPS and AMBER [11] which can be installed almost on any machine. In [12], authors demonstrate novel code-level and algorithmic improvements to Sandia's miniMD benchmark [13] and show that the usage of Intel Xeon Phi coprocessor provides up to 2x performance increase. The optimizations of MD code ls1mardyn are done for Intel Sandy Bridge processor including vectorization and shared-memory parallelization which allow to simulate liquid of multi-trillion Lennard-Jones particles on 146016 Cores of SuperMUC. Special MD code is developed in [14] called MODYLAS for machines like K-computer with torus topology. MODYLAS allows to investigate 100 ns-long MD calculation of 10 million-atom systems within 3 days (5 ms/step) by using machines as K-computer. The optimized for Intel Xeon Phi vectorization scheme is developed and applied to speed up LAMMPS calculations in [15]. The accuracy of MD with single precision is also studied.

3 Cluster "Desmos"

Table 1: Desmos cluster system configuration.

Server	Supermicro SuperServer 1018GR-T
Processor	Xeon E5-1650 v3 (6 cores, 3.0 GHz)
GPU	Nvidia GeForce GTX 1070 (1920C, 8 GB GDDR5)
Memory	DDR4 8 GB
Number of nodes	32
Interconnect	Angara 4D-torus $4 \times 2 \times 2 \times 2$
Operating system	SLES 11 SP4
Compiler	Intel Parallel Studio XE 2017
MPI	Angara MPI (based on MPICH 3.0.4)

3.1 GPU Testing

The Nvidia GTX 1070 cards have no memory errors correction (ECC) in contrast to the professional accelerators. For this reason it was necessary to make sure that there is no hardware memory errors in each GPU.

Testing of each GPU was performed using MemtestG80 [16]. Testing was performed on the available memory of the accelerator (8 GB) and wasted at least 4 hours (depending on the number of test iterations). Testing time was chosen on the basis of the results obtained in the paper [17] and restrictions imposed by the number of cards and the total time of the test the entire batch. There are no errors detected during the testing process.

3.2 Angara Interconnect

Angara interconnect is a Russian-designed communication network with torus topology. Inteconnect chip was developed by JSC NICEVT and manufactured by TSMC with 65 nm process. The chip supports deadlock-free adaptive routing based on bubble flow control [18], direction ordered routing [19], [20] and initial and final hops [20] for fault tolerance.

Each node has a dedicated memory region available for remote access (read, write, atomic operations) from other nodes to support the OpenSHMEM and PGAS languages. Multiple parallel programming models are supported, including MPI, OpenMP, OpenSHMEM.

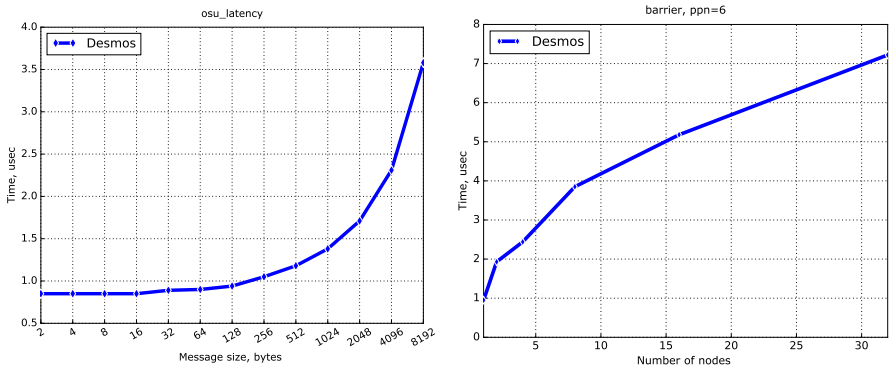
The network adapter is a PCI Express extension card that is connected to the adjacent nodes by up to 6 cables (or up to 8 with an extension card). The following topologies are supported: ring, 2D, 3D and 4D torus (or grid).

Fig. 1: The scheme of the “Angara” chip.

Description
Early HPL performance
Energy consumption

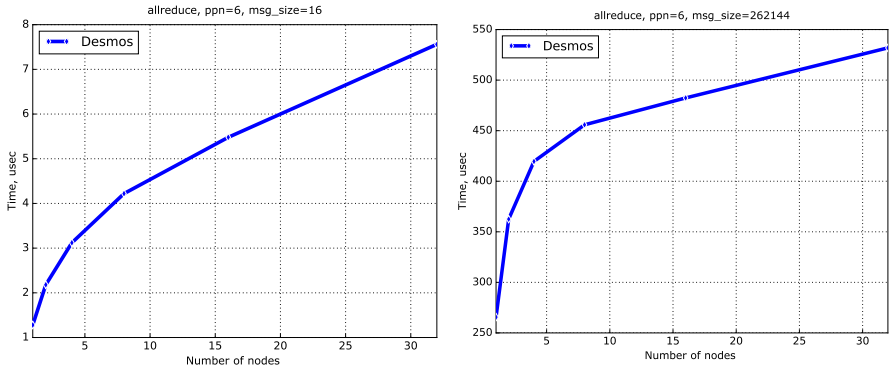
Fig. 2: The photo of one “Desmos” node and the photo of the rear side of the rack with cabling.

4 Experimental Results



(a) (b)

Fig. 3: (a) Communication latency between two adjacent Desmos nodes (OSU Micro-Benchmarks 5.3.2). (b) Intel MPI Benchmarks 2017 MPI_Barrier, processes per node (ppn) = 6.



(a) (b)

Fig. 4: FIXME: REMOVE?? Intel MPI Benchmarks 2017 MPI_Allreduce, ppn = 6 (a) message size = 16 bytes (b) message size = 256 KB.

4.1 ApoA1 benchmark

The Apolipoprotein A1 in water (∼100k atoms) system (ApoA1) is used to compare Desmos cluster with other supercomputers. The performances of different clusters based on ApoA1 test are shown in Fig. 5 in terms of seconds per 1 atom for 1 MD step and declared peak performance (3.236 TFlops per node consist of 6 core Xeon performance which is 0.336 TFlops and GPU performance which is estimated as 2.9 TFlops). The dotted line shows ideal scalability with performance 0.1 MFlops/atom/step.

Table 2: Info about different clusters.

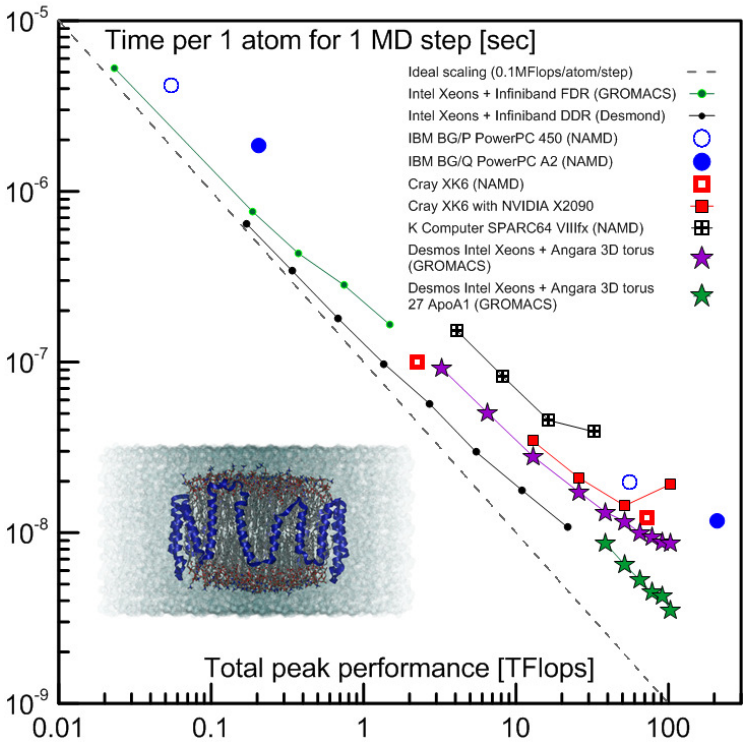


Fig. 5: ApoA1 benchmark. Timestep per 1 atom per 1 MD step vs R_{peak}. The comparison of different systems.

Despite different architectures and MD software, this data can be used to analyze the applicability of some supercomputer systems for biomolecular MD system. Black and green lines with dotes show the performance of CPU cluster with Intel Xeon nodes that consist of 28 cores [21]. New Desmos machine (purple filled stars) with GROMACS shows better scalability then CRAY XK6 with NVIDIA X2090 [1] (red filled squares) and K Computer SPARC64 Vlllfx with NAMD [?]. It also outgoes benchmarks ran on IBM BlueGene/P and BlueGene/Q [22] (blue open and filled points correspondingly).

Green stars are 27 times replicated ApoA1 benchmarks on Desmos machine. This result shows the continuous scaling with larger number atoms per node. It indicates the possibility for adding new nodes to Desmos without lack of performance.

4.2 MEM and RIB benchmarks

The work [23] gives a good guidelines for achieving the best performance for a minimal price in 2015. Authors compare different configurations of supercomputers using two wide spread biological benchmarks (are available at [24]): membrane channel protein embedded in a lipid bilayer surrounded by water (MEM, ∼100k atoms) and ions and bacterial ribosome in water with ions (RIB, ∼2M atoms). GROMACS package is used for all tests. The longer simulations than in [23] are performed to obtain the best performance. Other benchmark options are kept the same.

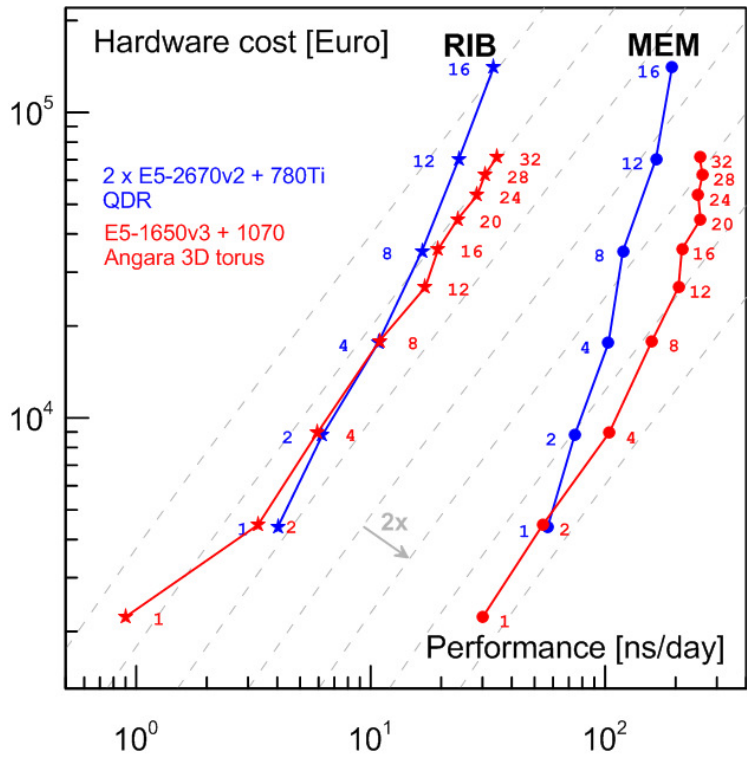


Fig. 6: The cost of hardware vs achieved performance for MEM and RIB benchmarks. Dashed lines show ideal scaling. Comparison with the published results for GROMACS in [23].

We compare the results obtained on Desmos cluster with the best choice of [23]: the nodes consist of 2 sockets Xeon E5-2670v2 with 780Ti and connected via InfiniBand QDR. The costs of hardware in euros are displayed on Y axis and the performance in ns/day is shown on X axis in Fig. 6. The numbers show the amount of nodes used. Grey lines indicate ideal scaling.

Desmos (red color) shows better weak scaling for MEM benchmark than system configuration provided by [23] (blue color). The saturation is achieved after 20 nodes which corresponds to the small amount of atoms per node. In the case of RIB benchmark, Desmos demonstrates ideal scaling after 16 nodes which shows that the productivity can be increased with more number of nodes.

5 Conclusions

The work of the JIHT team (N.K., V.S. and V.V.) was supported by the grant No. 14-50-00124 of the Russian Science Foundation (this work includes the development of the cluster node architecture, the preliminary benchmarks, the purchase of the "Desmos" cluster). The NICEVT team developed the "Angara" interconnect and the corresponding low-level software stack, built and tuned the "Desmos" cluster).

References

1. Yanhua Sun et al. Optimizing fine-grained communication in a biomolecular simulation application on Cray XK6. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2012.
2. James C. Phillips et al. Scalable molecular dynamics with namd. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.
3. Colloquium: Large scale simulations on GPU clusters. *The European Physical Journal B*, 88(6):158, 2015.
4. Guanpeng Li and Karthik Pattabiraman. Understanding Error Propagation in GPGPU Applications. (November), 2016.
5. Nikola Rajovic et al. The Mont-Blanc prototype : An Alternative Approach for HPC Systems. *SC16 Supercomputing*, (November):444–455, 2016.
6. Fabiano Corsetti. Performance analysis of electronic structure codes on HPC systems: A case study of SIESTA. *PLoS ONE*, 9(4), 2014.
7. M. Soler Jos et al. The siesta method for ab initio order- n materials simulation. *Journal of Physics: Condensed Matter*, 14(11):2745, 2002.
8. David E. Shaw et al. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2015-Janua(January):41–53, 2014.
9. Sander Pronk et al. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845, 2013.
10. B. R. Brooks et al. Charmm: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009.
11. Romelia Salomon-Ferrer et al. An overview of the amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(2):198–210, 2013.
12. Simon J. Pennycook et al. Exploring SIMD for molecular dynamics, using Intel Xeon processors and Intel Xeon Phi coprocessors. *Proceedings - IEEE 27th International Parallel and Distributed Processing Symposium, IPDPS 2013*, pages 1085–1097, 2013.

13. M. A. Heroux et al. Improving performance via mini- applications. *Sandia National Laboratories, Albuquerque, NM, Tech. Rep. SAND2009-5574, 2009.*
14. Noriyuki Yoshii et al. MODYLAS: A highly parallelized general-purpose molecular dynamics simulation program. *International Journal of Quantum Chemistry*, 115(5):342–348, 2015.
15. Markus Höhnerbach, Ahmed E. Ismail, and Paolo Bientinesi. The Vectorization of the Tersoff Multi-Body Potential: An Exercise in Performance Portability. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis(SC '16)*, page Article No.7, 2016.
16. Imran S. Haque and Vijay S. Pande. Hard data on soft errors: A large-scale assessment of real-world error rates in gpgpu. In *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, CCGRID '10, pages 691–696, Washington, DC, USA, 2010. IEEE Computer Society.
17. Carsten Kutzner, Szilárd Páll, Martin Fechner, Ansgar Esztermann, Bert L. de Groot, and Helmut Grubmuller. Best bang for your buck: Gpu nodes for gromacs biomolecular simulations. In *Journal of Computational Chemistry*, 2015.
18. V. Puente, R. Beivide, J. A. Gregorio, J. M. Prellezo, J. Duato, and C. Izu. Adaptive bubble router: A design to improve performance in torus networks. In *Proceedings of the 1999 International Conference on Parallel Processing*, ICPP '99, pages 58–, Washington, DC, USA, 1999. IEEE Computer Society.
19. N. R. Adiga, M. A. Blumrich, D. Chen, P. Coteus, A. Gara, M. E. Giampapa, P. Heidelberger, S. Singh, B. D. Steinmacher-Burow, T. Takken, M. Tsao, and P. Vranas. Blue gene/l torus interconnection network. *IBM J. Res. Dev.*, 49(2):265–276, March 2005.
20. Steven L. Scott and et al. The cray t3e network: Adaptive routing in a high performance 3d torus, 1996.
21. G. S. Smirnov and V. V. Stegailov. Efficiency of classical molecular dynamics algorithms on supercomputers. *Mathematical Models and Computer Simulations*, 8(6):734–743, 2016.
22. Sameer Kumar, Yanhua Sun, and Laximant V. Kale. Acceleration of an asynchronous message driven programming paradigm on ibm blue gene/q. In *Proceedings of the 2013 IEEE 27th International Symposium on Parallel and Distributed Processing*, IPDPS '13, pages 689–699, Washington, DC, USA, 2013. IEEE Computer Society.
23. Carsten Kutzner et al. Best bang for your buck: Gpu nodes for gromacs biomolecular simulations. *Journal of Computational Chemistry*, 36(26):1990–2008, 2015.
24. `http://www.mpibpc.mpg.de/grubmueller/kutzner/publications`.