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[13], 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
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. The interconnect is based on the Angara NIC that supports 3D and 4D torus network topology. We describe the corresponding ASIC structure and the software stack (shmem and MPI including). Firstly, this cluster is aimed at cost-effective classical molecular dynamics calculations. We present strong and weak scaling benchmarks for GROMACS and LAMMPS. Secondly, the cluster serves as a test bed for the Angara interconnect and verifies its ability to unite large MPP systems and to speed-up effectively MPI-based applications.

1 Introduction

2 Related work

2.1 Scalability of classical MD on supercomputers

3 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 [], direction ordered routing [] and initial and final hops [] 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.

3.1 Hardware

3.2 Software stack (MPI)

Including analytic estimates that show no need for extra MPI tuning below 256 nodes

4 Cluster "Desmos"

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.

Fig. 3. The scheme of the links between 32 cluster nodes (4x2x2x2).

5 MD benchmarks

ApoA1 model (100000 atoms) with GROMACS comparison with BlueGene, Cray and other large machines.

Large LJ system with LAMMPS on 32 nodes influence on special decomposition mapping on torus topology.

Comparison ns/day vs hardware cost.

6 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).

Fig. 4. ApoA1 benchmark. Timestep per 1 atom per 1 MD step vs R_{peak}. The comparison of different systems.

Fig. 5. Time vs cost. Comparison with the published results for GROMACS.