The Desmos supercomputer for computational materials science

Vladimir Stegailov, Nikolay Kondratyuk, Grigory Smirnov, Alexei Timofeev

Grant No. 14-50-00124
National Research University Higher School of Economics

Supercomputer Atomistic Modelling and Multi-scale Analysis

http://samma.hse.ru
Outline of the talk

• Angara network and its developer JSC NICEVT
• Building process of the Desmos supercomputer
• Benchmarking of MD calculations
• Statistical data of Desmos deployment
• Comparison with a new hybrid supercomputer in HSE
• VASP benchmarks
• A next Angara-based HPC system
ANGARA NETWORK
AND ITS DEVELOPER JSC NICEVT
JSC NICEVT:
from ES EVM computers to Angara network
Angara interconnect

Key features

- Network topology: 1D..4D-torus
- ASIC-based network card
- Up to 8 communication channels
- Remote direct memory access (RDMA)
- Multi-core CPU support
- Adaptive packet transfer
- MPI ping-pong latency: 0.85 µs
- Single hop latency: 129 ns
- Scaling: up to 32K nodes
- Power consumption: up to 20 W
- Various physical transmission media
## Angara interconnect

<table>
<thead>
<tr>
<th>Interconnect</th>
<th>Mellanox IB FDR 4x</th>
<th>Angara</th>
<th>Cray Aries</th>
<th>Mellanox IB EDR 4x</th>
<th>Intel OmniPath</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>2011</td>
<td>2013</td>
<td>2012</td>
<td>2015</td>
<td>2015</td>
</tr>
<tr>
<td>TOP500</td>
<td>13</td>
<td>–</td>
<td>5</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>Topology</td>
<td>fat tree / kD-torus</td>
<td>4D-torus</td>
<td>dragonfly</td>
<td>fat tree / kD-torus</td>
<td>fat tree</td>
</tr>
<tr>
<td>MPI latency, µs</td>
<td>1</td>
<td>0.85</td>
<td>1.3</td>
<td>0.92</td>
<td>0.9-1.0</td>
</tr>
<tr>
<td>Single hop latency, ns</td>
<td>– / 250</td>
<td>129</td>
<td>100</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>
BUILDING PROCESS OF THE SUPERCOMPUTER DESMOS
The main goals of the project

- To make a supercomputer that is effective for classical molecular dynamics and usable of ab initio MD
- To use the Angara network and get the largest possible number of nodes within a limited budget limit
- **Access to novel hardware is very beneficial for HPC training**

LAMMPS, GROMACS

VASP, ABINIT, CP2K
Intel Xeon E5-1650v3

Nvidia GeForce 1070
Desmos supercomputer

- CPU: E5-1650v3, 6 cores, 3.5 GHz
- GPU: Nvidia GTX1070
- SSD: 240 GB
- Angara NIC

1 GbE link per node
Desmos supercomputer

- **CPU**: E5-1650v3, 6c, 3.5GHz
- **SSD**: 240 GB
- **GPU**: Nvidia GTX1070
- **NIC**: Angara
- **Head node**: 1 GbE link per node
- **Gigabit Ethernet switch**
- **Storage**: FirePro S9150

...
BENCHMARKING OF MD CALCULATIONS
ApoA1 benchmark (protein in water, ~ 100 000 atoms): comparison of different supercomputers
Weak scaling of different MD models

- LJ liquid (LAMMPS)
  - 2916000 atoms
  - 50% DRAM BW

- C<sub>30</sub>H<sub>62</sub> liquid (LAMMPS)
  - 736000 atoms
  - 50% DRAM BW

- ApoA1 (GROMACS)
  - 92224 atoms
  - 50% DRAM BW

Graph shows the number of nodes vs. time per day (ns/day) for different MD models.
Best Bang for Your Buck: GPU Nodes for GROMACS Biomolecular Simulations


The molecular dynamics simulation package GROMACS runs efficiently on a wide variety of hardware from commodity workstations to high performance computing clusters. Hardware features are well-exploited with a combination of single instruction multiple data, multithreading, and message passing interface (MPI)-based single program multiple data/multiple program multiple data parallelism while graphics processing units (GPUs) can be used as accelerators to compute interactions off-loaded from the CPU. Here, we evaluate which hardware produces trajectories with GROMACS 4.6 or 5.0 in the most economical way. We have assembled and benchmarked compute nodes with various CPU/GPU combinations to identify optimal compositions in terms of raw trajectory production rate, performance-to-price ratio, energy efficiency, and several other criteria. Although hardware prices are naturally subject to trends and fluctuations, general tendencies are clearly visible. Adding any type of GPU significantly boosts a node’s simulation performance. For inexpensive consumer-class GPUs this improvement equally reflects in the performance-to-price ratio. Although memory issues in consumer-class GPUs could pass unnoticed as these cards do not support error checking and correction memory, unreliable GPUs can be sorted out with memory checking tools. Apart from the obvious determinants for cost-efficiency like hardware expenses and raw performance, the energy consumption of a node is a major cost factor. Over the typical hardware lifetime until replacement of a few years, the costs for electrical power and cooling can become larger than the costs of the hardware itself. Taking that into account, nodes with a well-balanced ratio of CPU and consumer-class GPU resources produce the maximum amount of GROMACS trajectory over their lifetime. © 2015 The Authors. Journal of Computational Chemistry Published by Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24030
Comparison with the IRUS17 supercomputer
2 x Intel Xeon E5-2699 v4 per node with Intel Omni-Path

RIB: 2136412 atoms

MEM: 81743 atoms

Prices from thinkmate.com (November 2017)
STATISTICAL DATA OF DESMOS DEPLOYMENT
SlurmDB data for September-November 2017

Job running time \([10^3 \text{ sec}]\)

Job waiting time \([10^4 \text{ sec}]\)
Angara interconnect makes GPU-based Desmos supercomputer an efficient tool for molecular dynamics calculations

Vladimir Stegailov\textsuperscript{1,2}, Ekaterina Dlinova\textsuperscript{2}, Timur Ismagilov\textsuperscript{3}, Mikhail Khalilov\textsuperscript{2}, Nikolay Kondratyuk\textsuperscript{1,2}, Dmitry Makagon\textsuperscript{3}, Alexander Semenov\textsuperscript{2,3}, Alexei Simonov\textsuperscript{3}, Grigory Smirnov\textsuperscript{1,2} and Alexey Timofeev\textsuperscript{1,2}

Abstract
In this article, we describe the Desmos supercomputer that consists of 32 hybrid nodes connected by a low-latency high-bandwidth Angara interconnect with torus topology. This supercomputer is aimed at cost-effective classical molecular dynamics calculations. Desmos serves as a test bed for the Angara interconnect that supports 3-D and 4-D torus network topologies and verifies its ability to unite massively parallel programming systems (MPS) based on the HPC systems. We describe the Angara interconnect presenting typical MPS benchmarks. Desmos benchmarks results for GROMACS, LAMMPS, VASP and CP2K are compared with the data for other high performance computing (HPC) systems. Also, we consider the job scheduling statistics for several months of Desmos deployment.
COMPARISON WITH A NEW HYBRID SUPERCOMPUTER IN HSE
The new supercomputer in Higher School of Economics: a hybrid system based on 26 DELL C4140 nodes
Comparison Desmos with the new HSE supercomputer

GROMACS
RIB: 2136412 atoms
VASP BENCHMARKS
SPECIAL ISSUE PAPER

VASP hits the memory wall: Processors efficiency comparison

Vladimir Stegailov\textsuperscript{1,2}  Grigory Smirnov\textsuperscript{1,2}  Vyacheslav Vecher\textsuperscript{1,2}

\textsuperscript{1}Joint Institute for High Temperatures of the Russian Academy of Sciences, Moscow, Russia
\textsuperscript{2}Moscow Institute of Physics and Technology (State University), Dolgoprudy, Russia

Correspondence
Vladimir Stegailov, Moscow Institute of Physics and Technology (State University), Institutsky Pereulok 9, Dolgoprudy 141701, Russia.
Email: stegailov.vv@mipt.ru

Funding Information
Ministry of Science and Higher Education of the Russian Federation, Grant/Award Number: 074-02-2018-286; President grants for the support of leading scientific schools, Grant/Award Number: NS-5922.2018.8

Summary
First-principles calculations of electronic structure have been one of the most important classes of supercomputer applications for a long time. In this paper, we consider VASP as a de facto standard tool for density functional theory calculations widely used in materials science, condensed matter research, and other related fields. The choice of hardware for the efficient VASP calculations is not easy because of the large number of processor types available. We use the benchmark metric that is based on the balance of the peak floating point performance and the memory bandwidth. This metric gives us the possibility to compare different types of processors. We consider time-to-solution and energy-to-solution criteria and compare different Intel, AMD, and ARM 64-bit CPUs and hybrid CPU-GPU systems based on Nvidia Tesla P100.

KEYWORDS
ab initio calculations, energy-to-solution, memory bandwidth, peak performance, time-to-solution
VASP: GaAs benchmark with 80 atoms
<table>
<thead>
<tr>
<th>PE type</th>
<th>LREAL=TRUE τ&lt;sub&gt;iter&lt;/sub&gt; (sec)</th>
<th>Options</th>
<th>LREAL=FALSE τ&lt;sub&gt;iter&lt;/sub&gt; (sec)</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 x Xeon E5-1650v3</td>
<td>36</td>
<td>6 MPI</td>
<td>66</td>
<td>2 MPI x 3 OpenMP</td>
</tr>
<tr>
<td>2 x Xeon E5-2660v4</td>
<td>25</td>
<td>14 MPI</td>
<td>66</td>
<td>4 MPI x 3 OpenMP</td>
</tr>
<tr>
<td>2 x Xeon E5-2698v4</td>
<td>24</td>
<td>20 MPI</td>
<td>59</td>
<td>4 MPI x 4 OpenMP</td>
</tr>
<tr>
<td>2 x Epyc 7301</td>
<td>20</td>
<td>16 MPI</td>
<td>39</td>
<td>8 MPI x 2 OpenMP</td>
</tr>
<tr>
<td>2 x Epyc 7551</td>
<td>13</td>
<td>32 MPI</td>
<td>33</td>
<td>8 MPI x 4 OpenMP</td>
</tr>
<tr>
<td>1 x Tesla P100 (JUPITER)</td>
<td>12</td>
<td>NSIM=8</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>2 x Tesla P100 (DGX-1)</td>
<td>7.5</td>
<td>NSIM=8</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>4 x Tesla P100 (DGX-1)</td>
<td>5.5</td>
<td>NSIM=8</td>
<td>n/a</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 2: The selected best time-to-solution values (NCORE=1, KPAR=4). The number of MPI processes per socket is given together with the number of OpenMP threads per one MPI process.
VASP: GaAs benchmark with 80 atoms
VASP: GaAs benchmark with 80 atoms

E5-1650v3 + Angara
2x E5-2698v4 + OmniPath
P8 + P100 + EDR
2x Epyc 7301 + FDR
2x Gold 6152 + EDR

Lower means better balanced
A NEXT ANGARA-BASED HPC SYSTEM
In March 2019 a next HPC system in JIHT RAS has been installed based on the Angara switch and one-port low-profile PCIe cards.
Conclusions

• The scaling tests for the classical MD and electronic structure calculations show the high efficiency of the MPI-exchanges over the Angara network.

• GPU-accelerated classical MD with Gromacs runs faster and is more cost-effective on supercomputers similar to Desmos than on wide-spread supercomputers based on expensive Intel Xeon multi-core CPUs.

• The job accounting statistic of the Desmos supercomputer has been reviewed. Two methods of quantitative efficiency monitoring have been proposed.

• All-to-all performance of the Angara network confirms its efficiency for electronic structure codes. The detailed analysis of modern CPUs efficiency for typical VASP model calculations has helped to select the best option for the next system.