EMiT 2019
Emerging Technology Conference
9th-11th April, Huddersfield, UK

The Desmos supercomputer for computational materials science

Vladimir Stegailov, Nikolay Kondratyuk, Grigory Smirnov, Alexei Timofeev

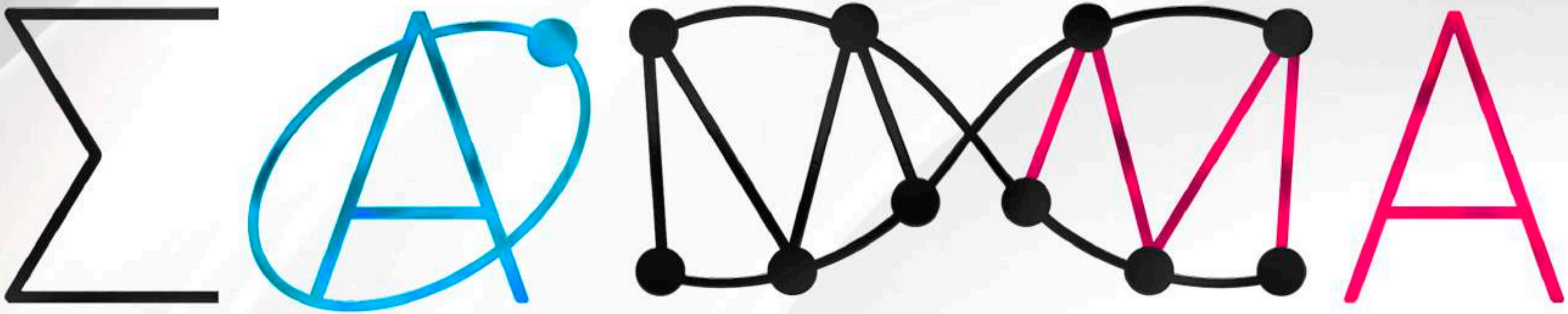
ИВТАН

MIPT
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

RSF | Russian Science Foundation
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

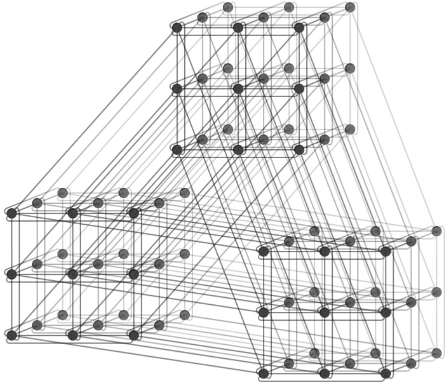
ANGARA NETWORK
AND ITS DEVELOPER JSC NICEVT

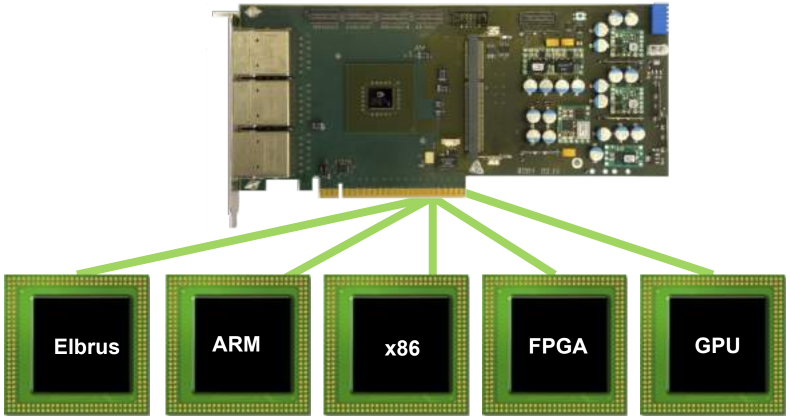
JSC NICEVT:
from ES EVM computers to Angara network

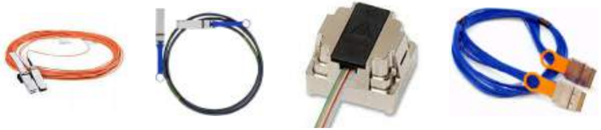
Angara interconnect

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







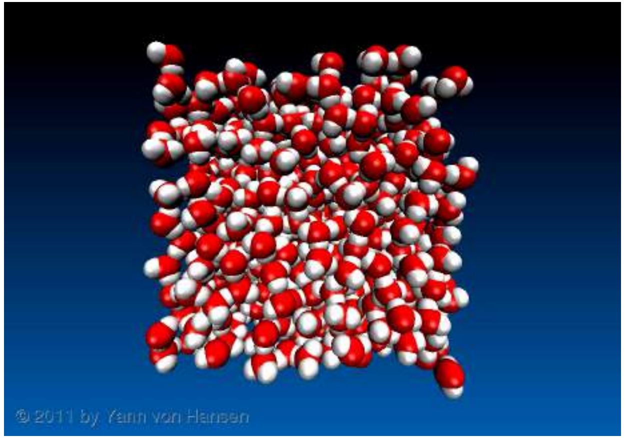
Angara interconnect

Interconnect	Mellanox IB FDR 4x	Angara	Cray Aries	Mellanox IB EDR 4x	Intel OmniPath
Year	2011	2013	2012	2015	2015
TOP500	13	–	5	20	6
Topology	fat tree / kD-torus	4D-torus	dragonfly	fat tree / kD-torus	fat tree
MPI latency, µs	1	0.85	1.3	0.92	0.9-1.0
Single hop latency, ns	– / 250	129	100	n/a	n/a

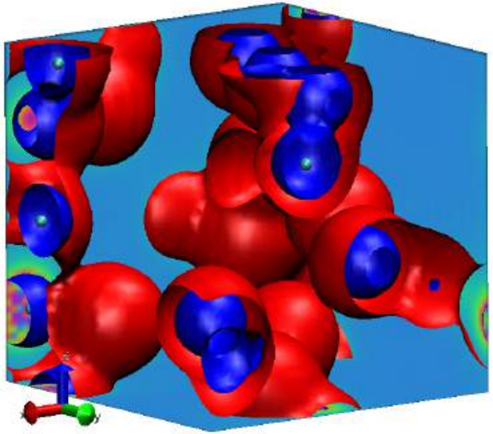
BUILDING PROCESS OF THE SUPERCOMPUTER DESMOS

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**



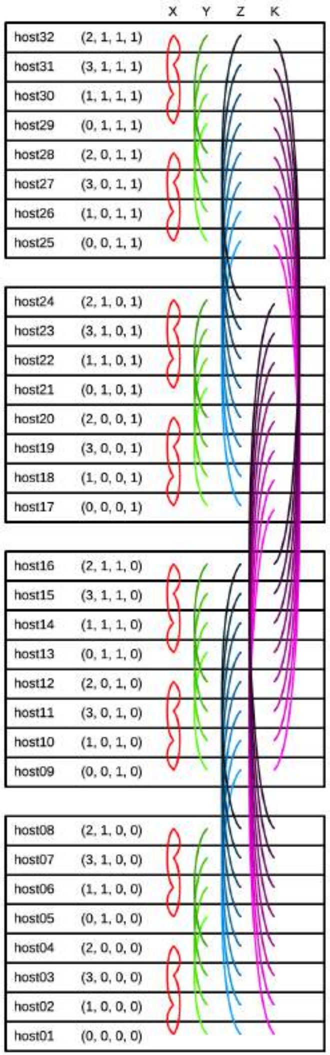
LAMMPS, GROMACS



VASP, ABINIT, CP2K

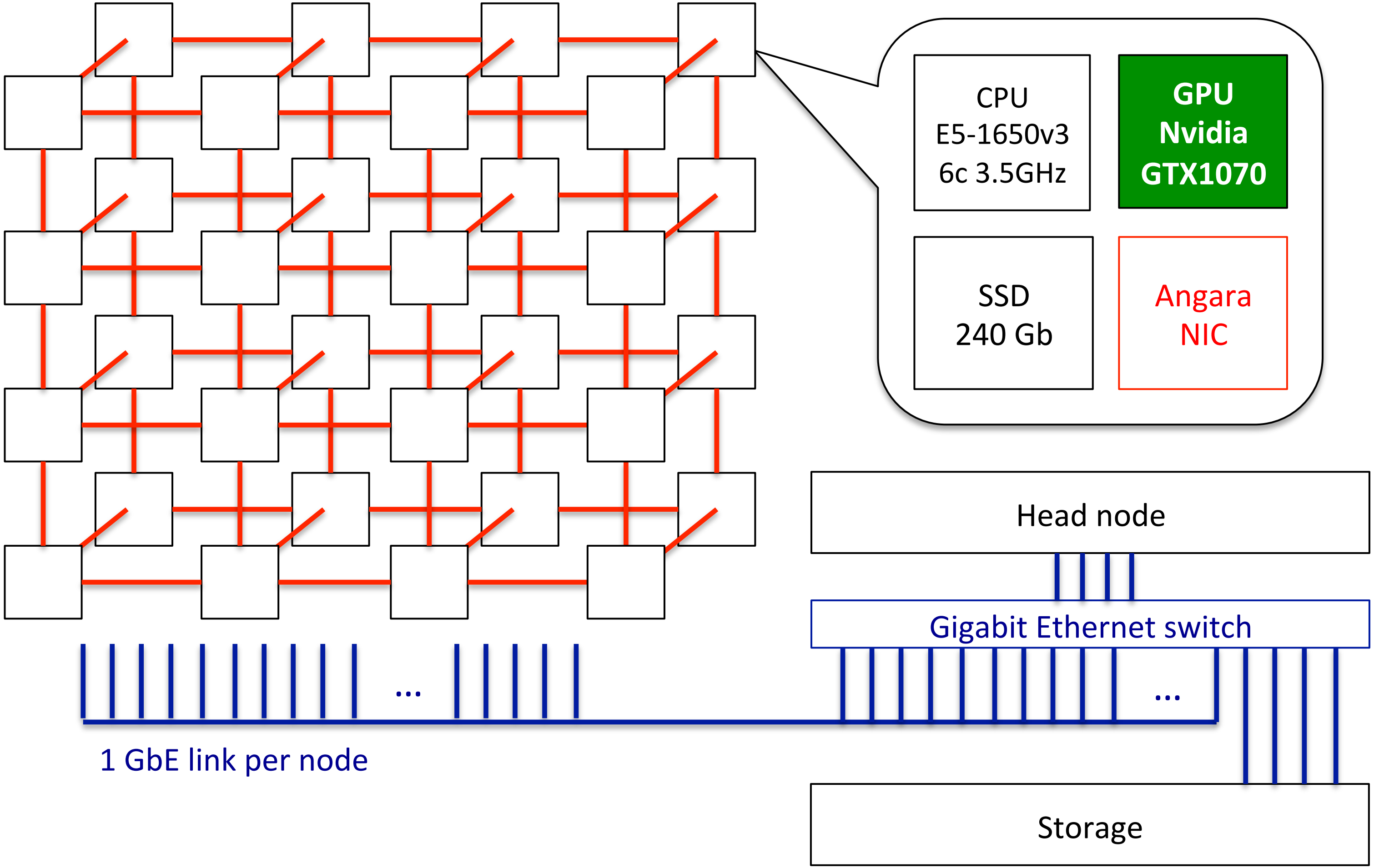
Intel Xeon E5-1650v3

Nvidia GeForce 1070

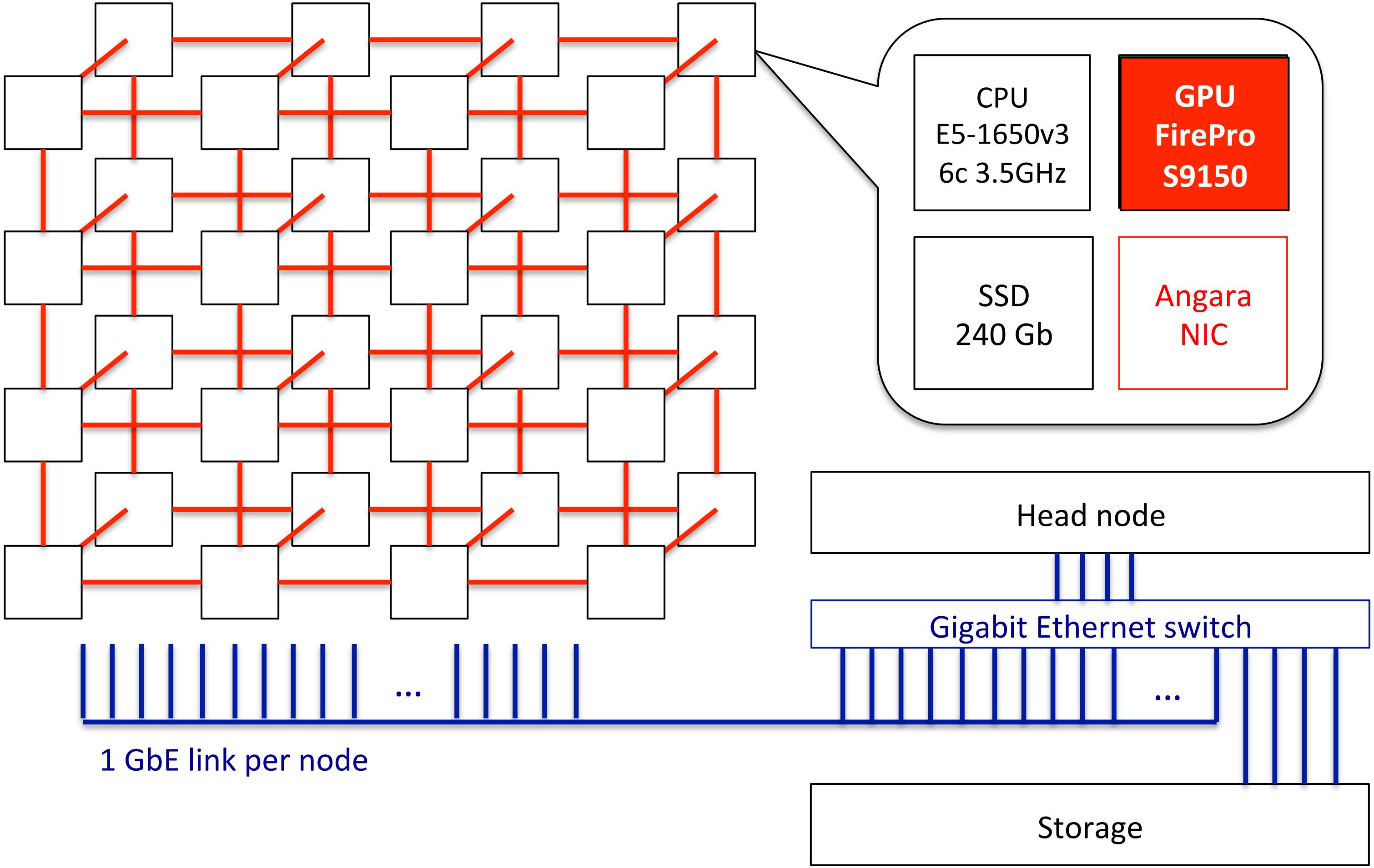


host32	(2, 1, 1, 1)
host31	(3, 1, 1, 1)
host30	(1, 1, 1, 1)
host29	(0, 1, 1, 1)
host28	(2, 0, 1, 1)
host27	(3, 0, 1, 1)
host26	(1, 0, 1, 1)
host25	(0, 0, 1, 1)
host24	(2, 1, 0, 1)
host23	(3, 1, 0, 1)
host22	(1, 1, 0, 1)
host21	(0, 1, 0, 1)
host20	(2, 0, 0, 1)
host19	(3, 0, 0, 1)
host18	(1, 0, 0, 1)
host17	(0, 0, 0, 1)
host16	(2, 1, 1, 0)
host15	(3, 1, 1, 0)
host14	(1, 1, 1, 0)
host13	(0, 1, 1, 0)
host12	(2, 0, 1, 0)
host11	(3, 0, 1, 0)
host10	(1, 0, 1, 0)
host09	(0, 0, 1, 0)
host08	(2, 1, 0, 0)
host07	(3, 1, 0, 0)
host06	(1, 1, 0, 0)
host05	(0, 1, 0, 0)
host04	(2, 0, 0, 0)
host03	(3, 0, 0, 0)
host02	(1, 0, 0, 0)
host01	(0, 0, 0, 0)

Desmos supercomputer

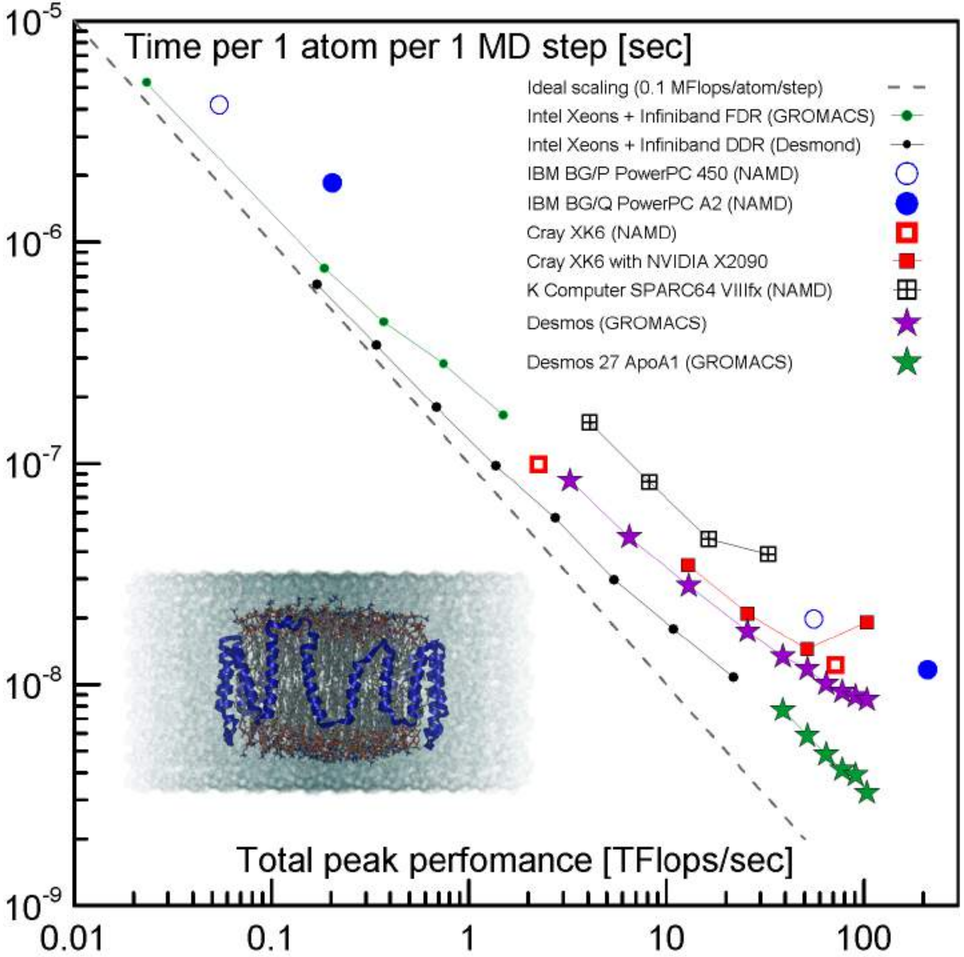


Desmos supercomputer

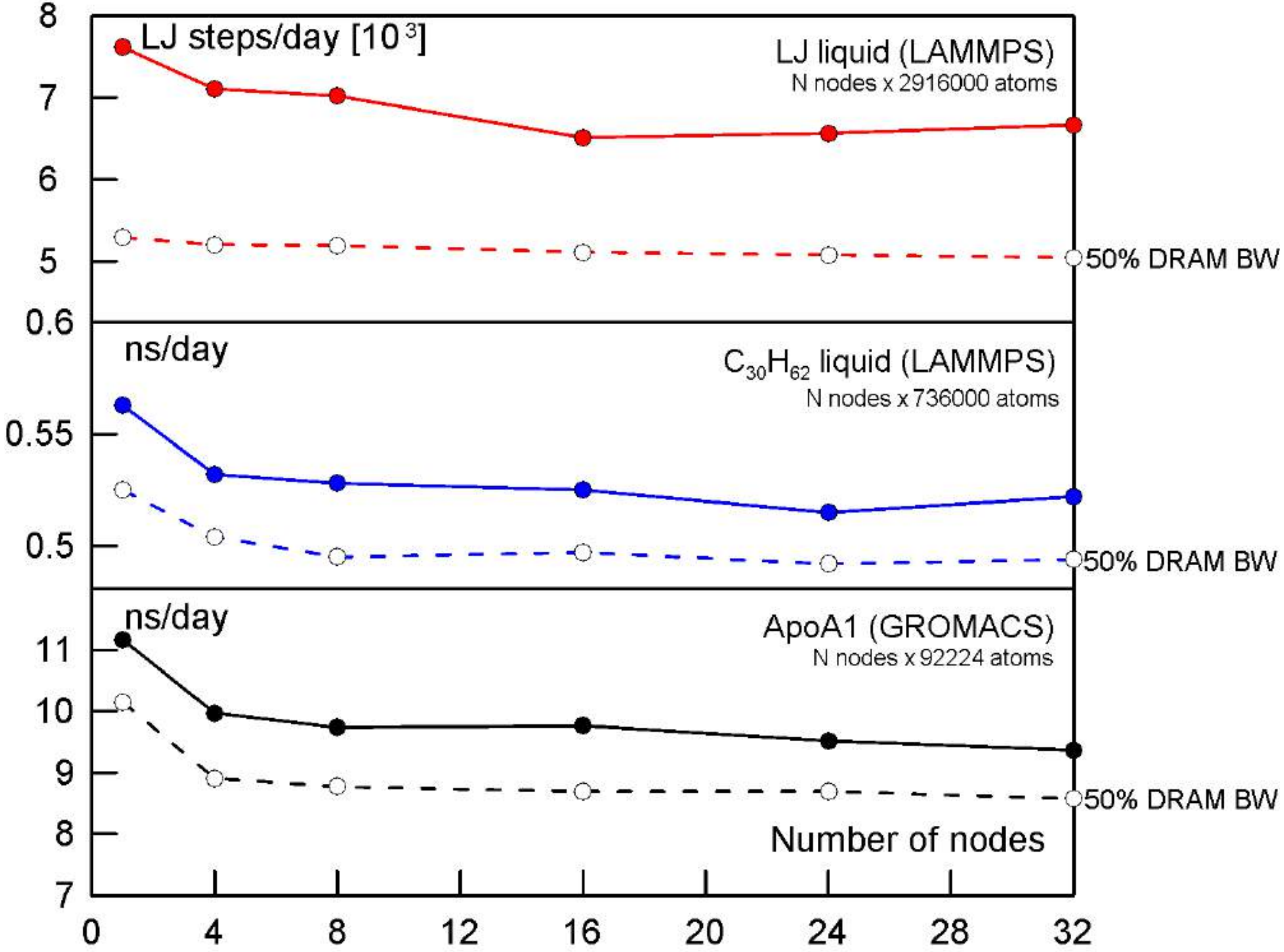


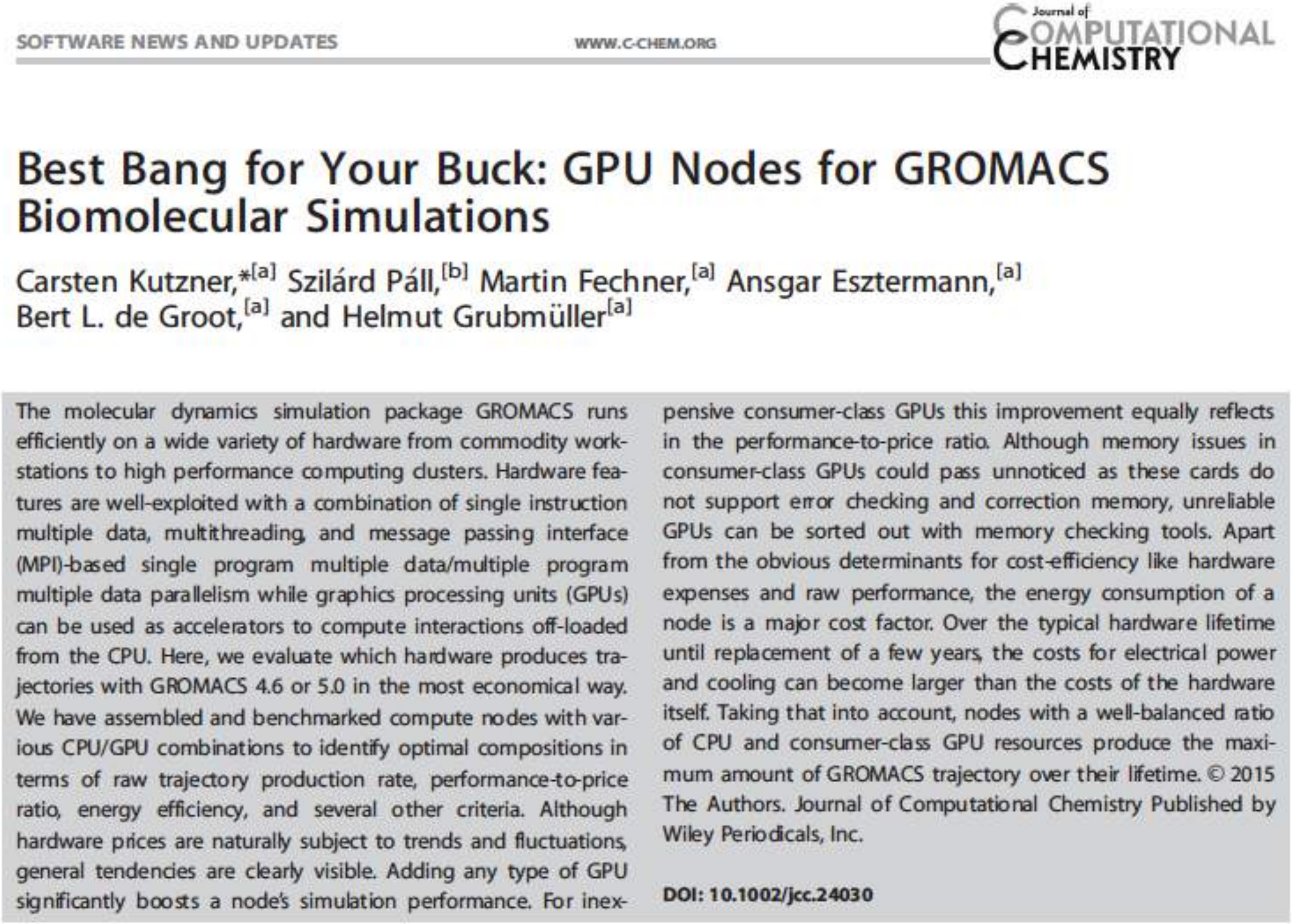
BENCHMARKING OF MD CALCULATIONS

ApoA1 benchmark (protein in water, ~ 100 000 atoms): comparison of different supercomputers



Weak scaling of different MD models





Best Bang for Your Buck: GPU Nodes for GROMACS Biomolecular Simulations

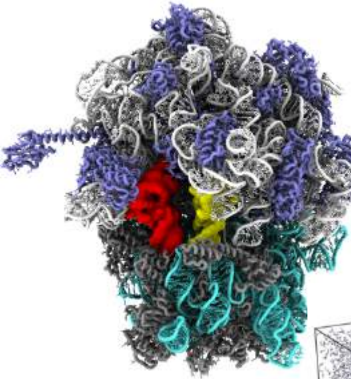
Carsten Kutzner,*[a] Szilárd Páll,[b] Martin Fechner,[a] Ansgar Esztermann,[a] Bert L. de Groot,[a] and Helmut Grubmüller[a]

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.

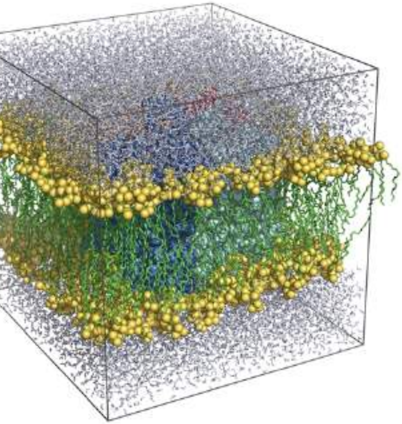
DOI: 10.1002/jcc.24030

Comparison with the IRUS17 supercomputer

2 x Intel Xeon E5-2699 v4 per node
with Intel Omni-Path

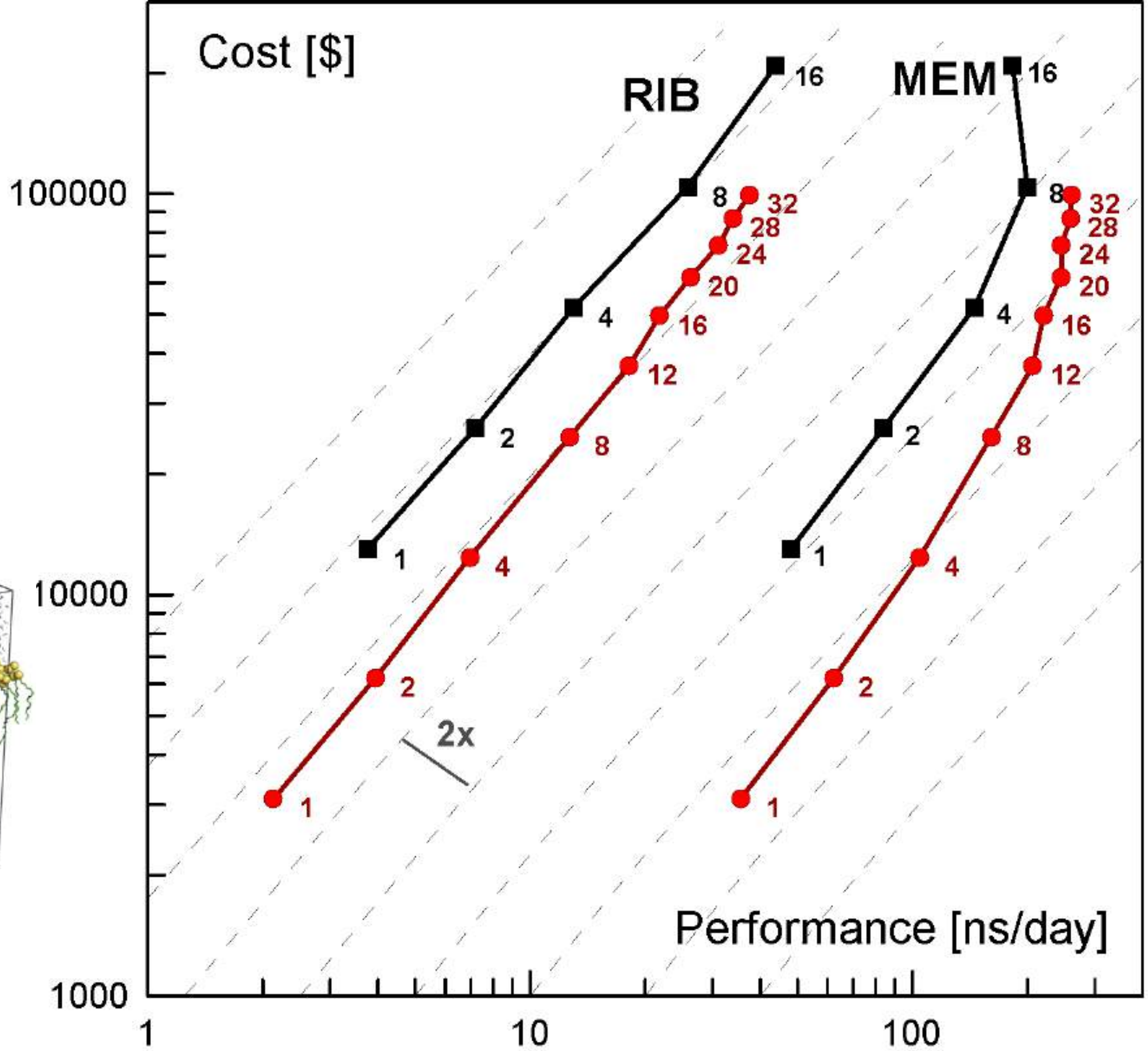


RIB:
2136412
atoms



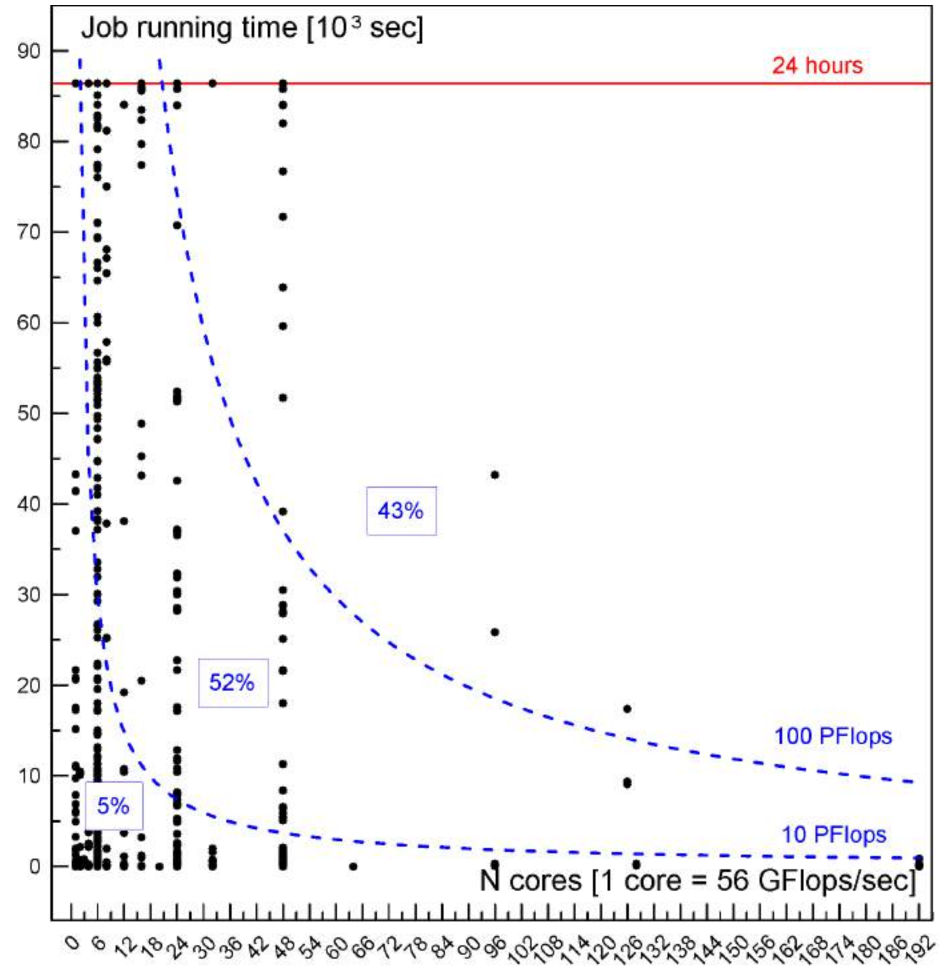
MEM:
81743
atoms

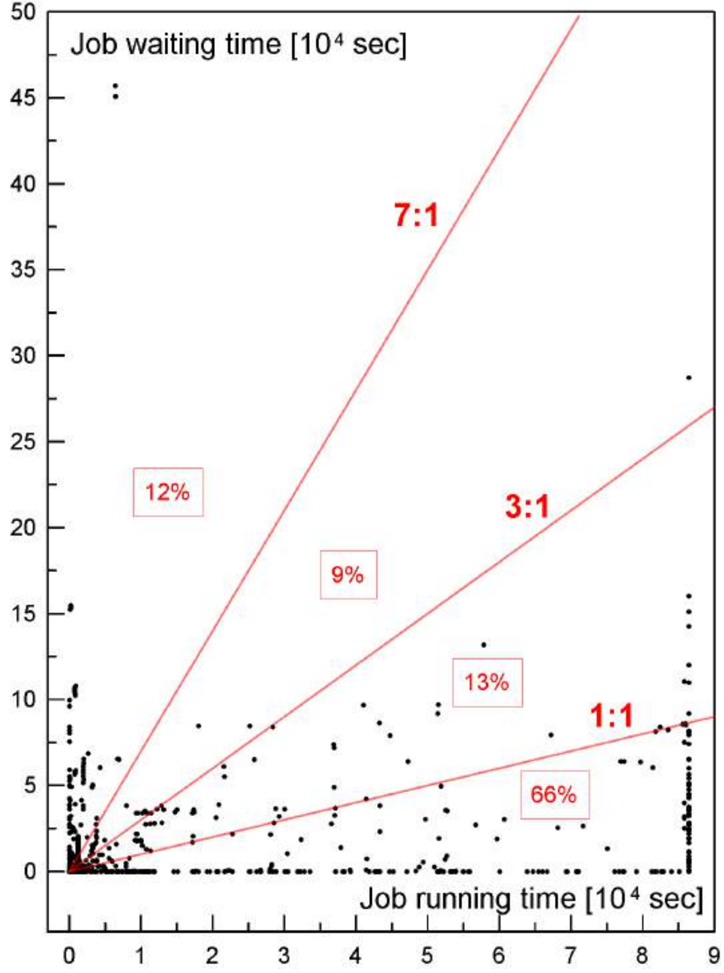
Prices from thinkmate.com
(November 2017)

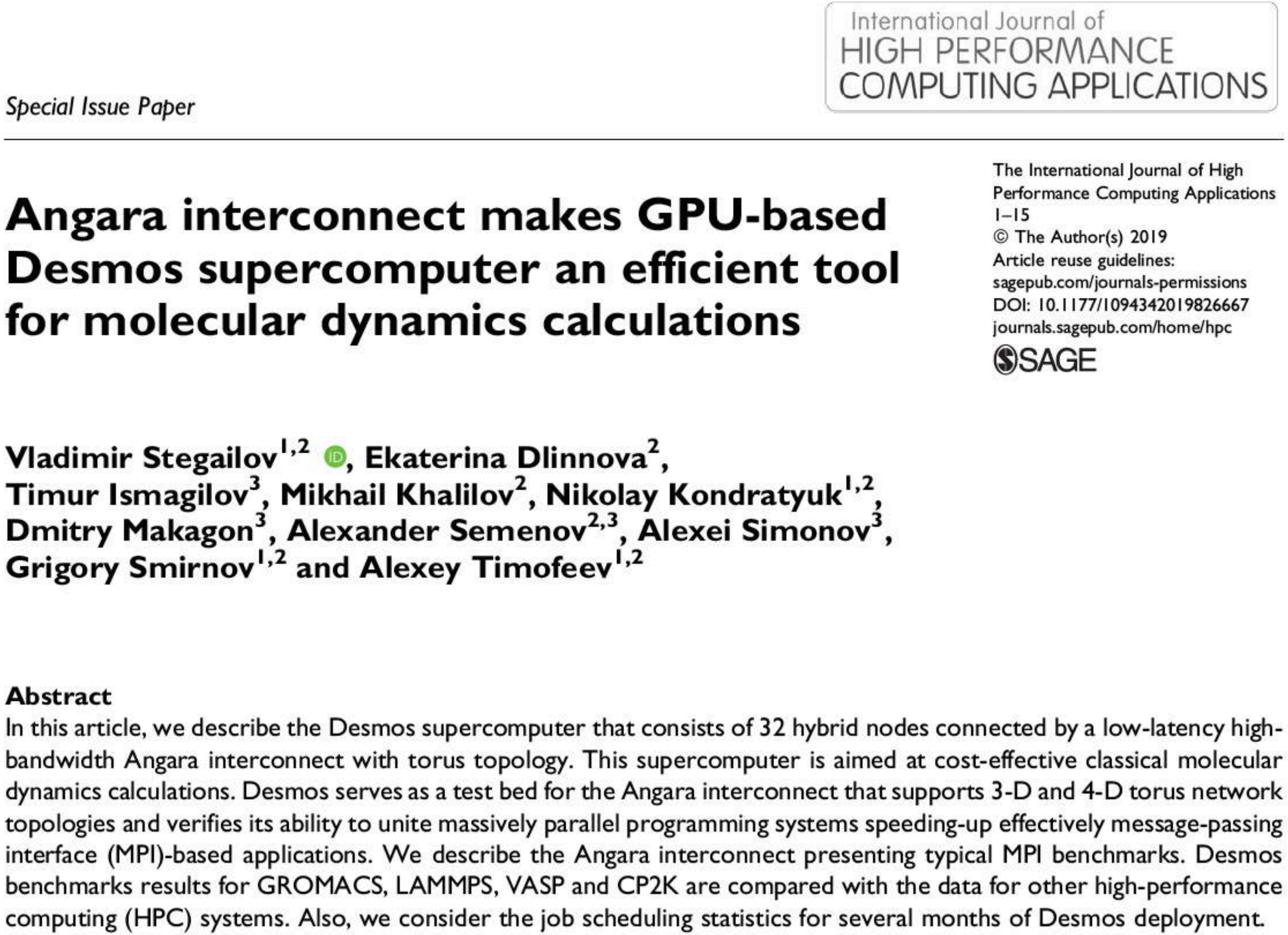


STATISTICAL DATA
OF DESMOS DEPLOYMENT

SlurmDB data for September-November 2017







International Journal of
HIGH PERFORMANCE
COMPUTING APPLICATIONS

Special Issue Paper

The International Journal of High
Performance Computing Applications
1–15
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1094342019826667
journals.sagepub.com/home/hpc
SAGE

Angara interconnect makes GPU-based Desmos supercomputer an efficient tool for molecular dynamics calculations

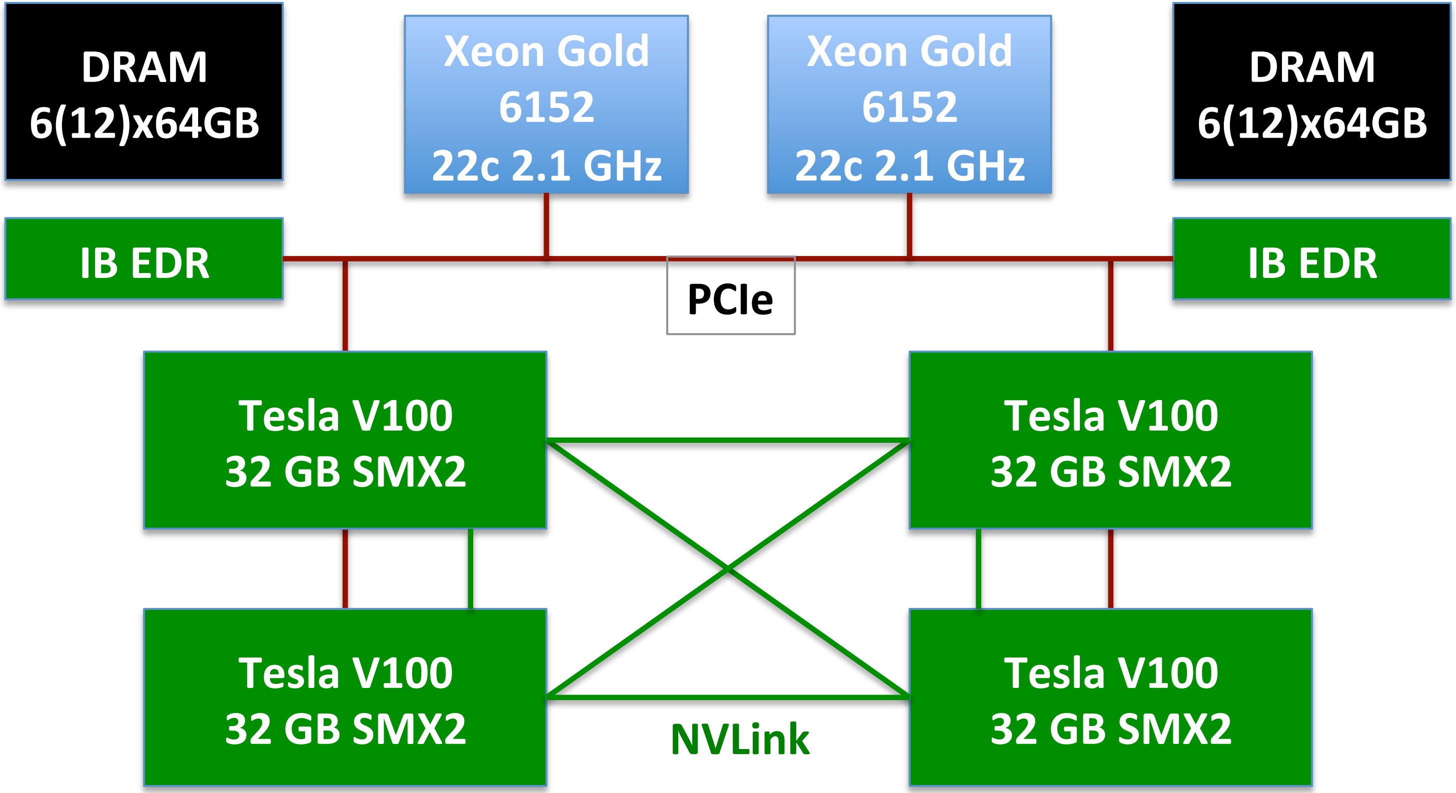
Vladimir Stegailov[1,2], Ekaterina Dlinnova[2],
Timur Ismagilov[3], Mikhail Khalilov[2], Nikolay Kondratyuk[1,2],
Dmitry Makagon[3], Alexander Semenov[2,3], Alexei Simonov[3],
Grigory Smirnov[1,2] and Alexey Timofeev[1,2]

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 speeding-up effectively message-passing interface (MPI)-based applications. We describe the Angara interconnect presenting typical MPI 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.

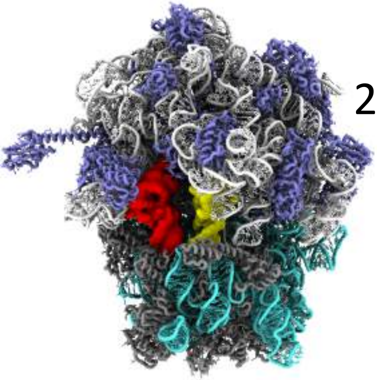
COMPARISON WITH A NEW HYBRID SUPERCOMPUTER IN HSE

The new supercomputer in Higher School of Economics: a hybrid system based on 26 DELL C4140 nodes

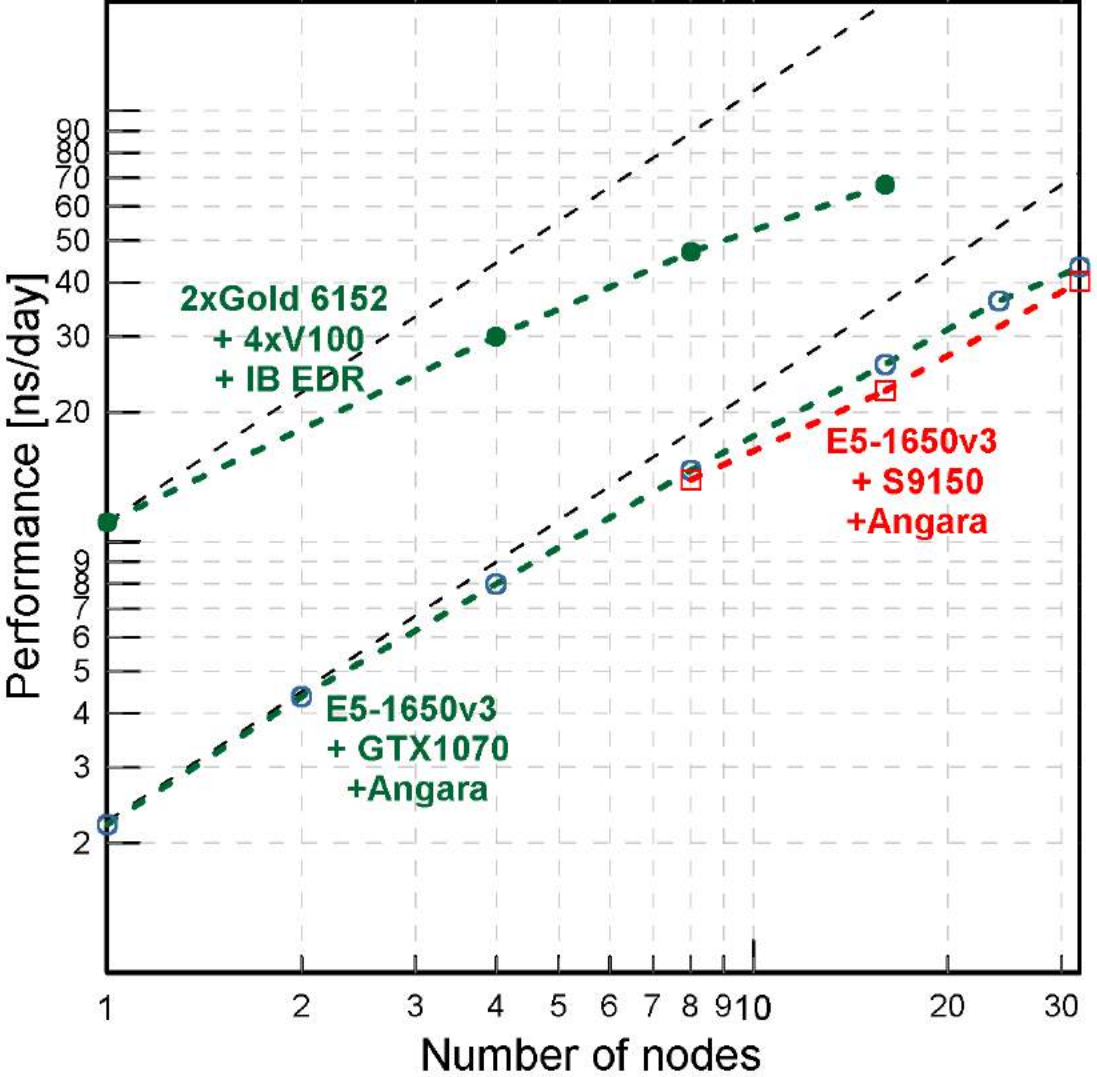


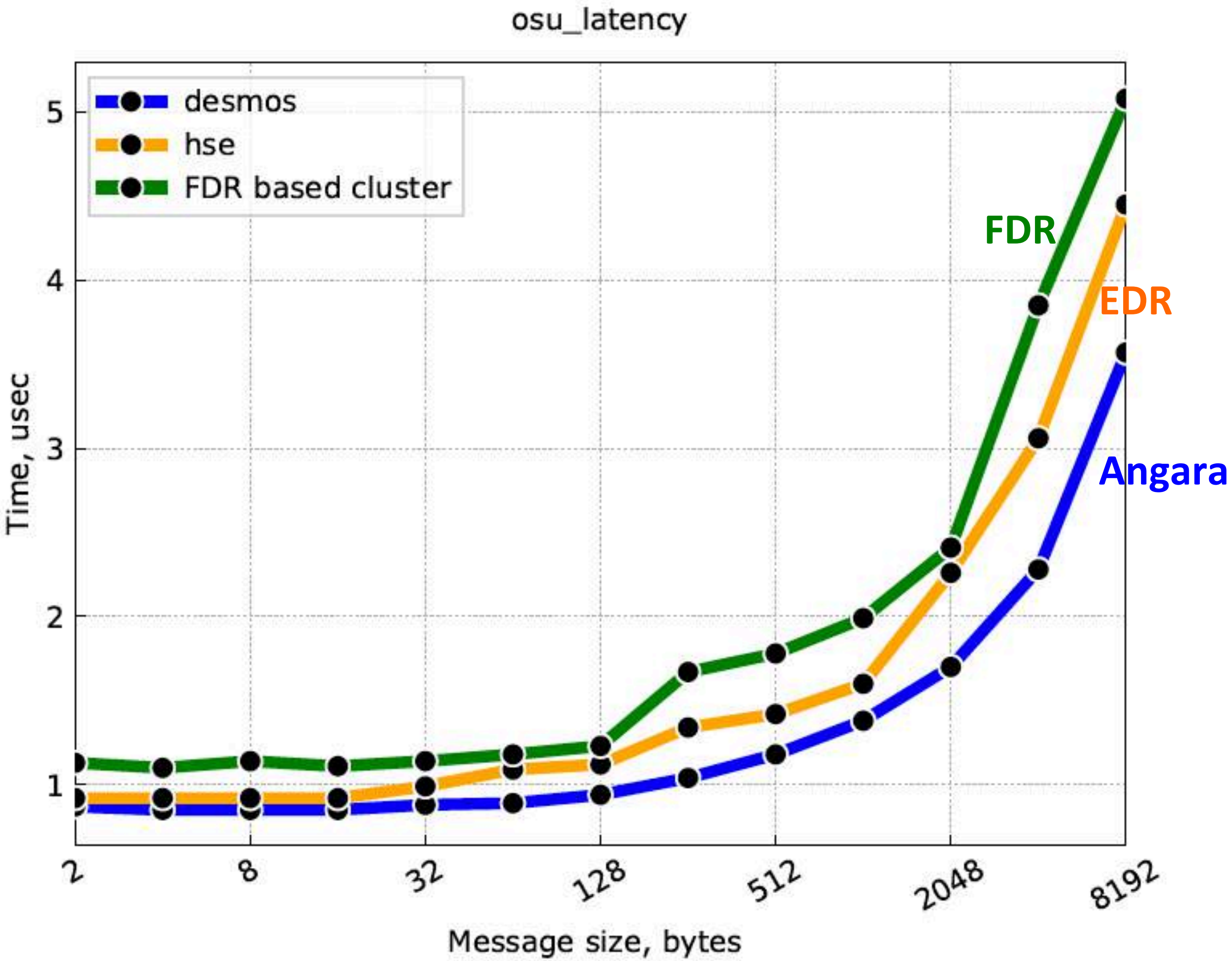
Comparison Desmos
with the new HSE supercomputer

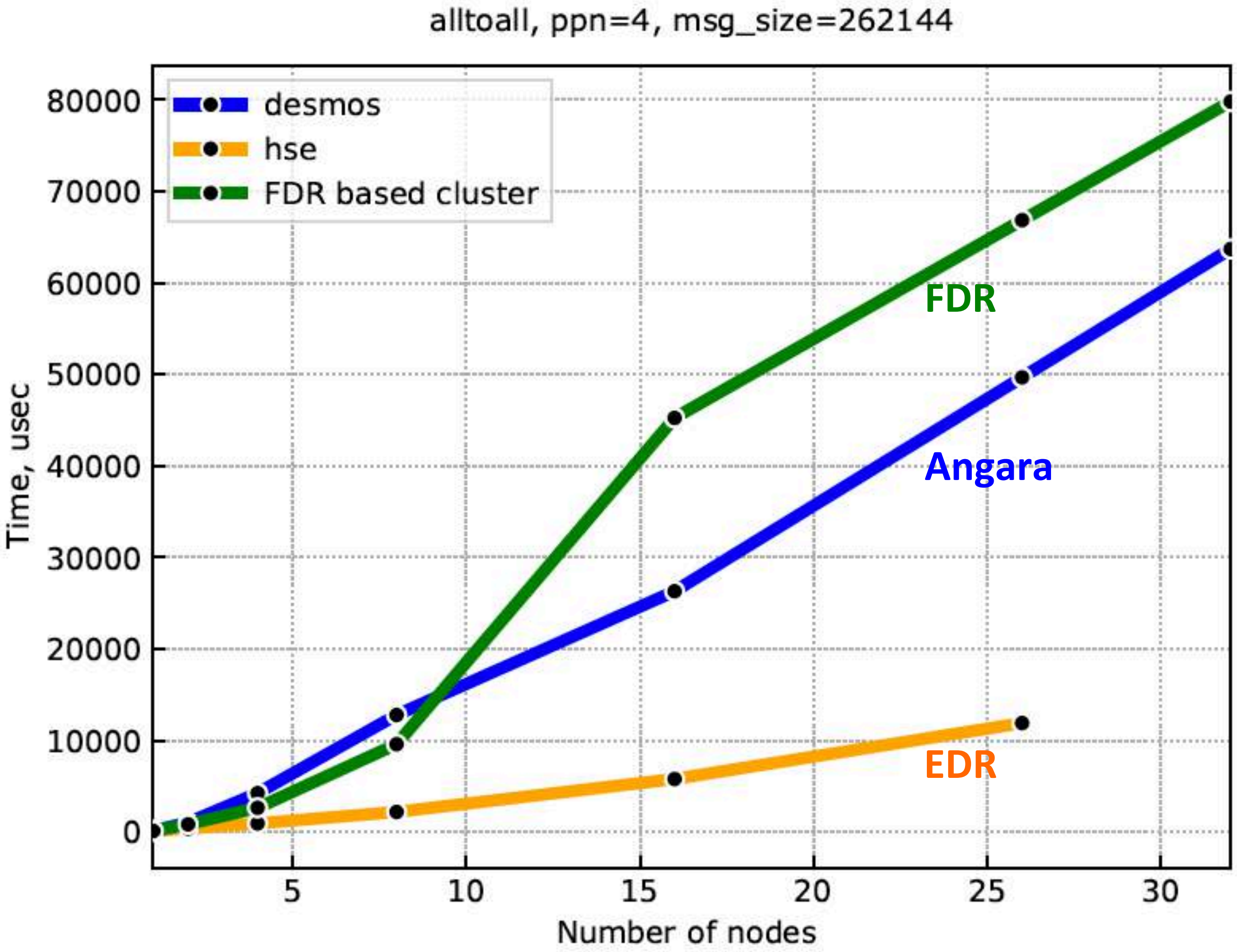
GROMACS

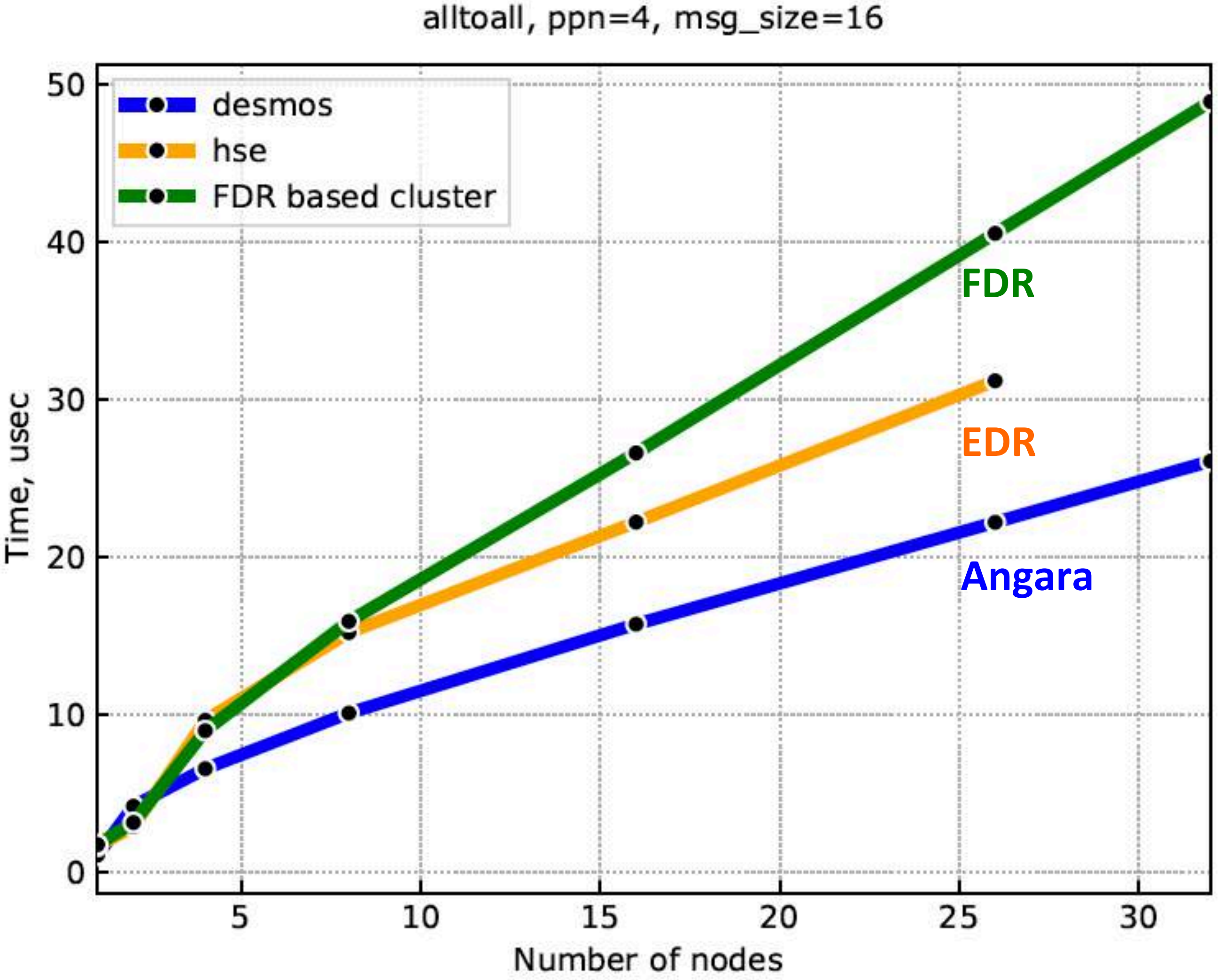


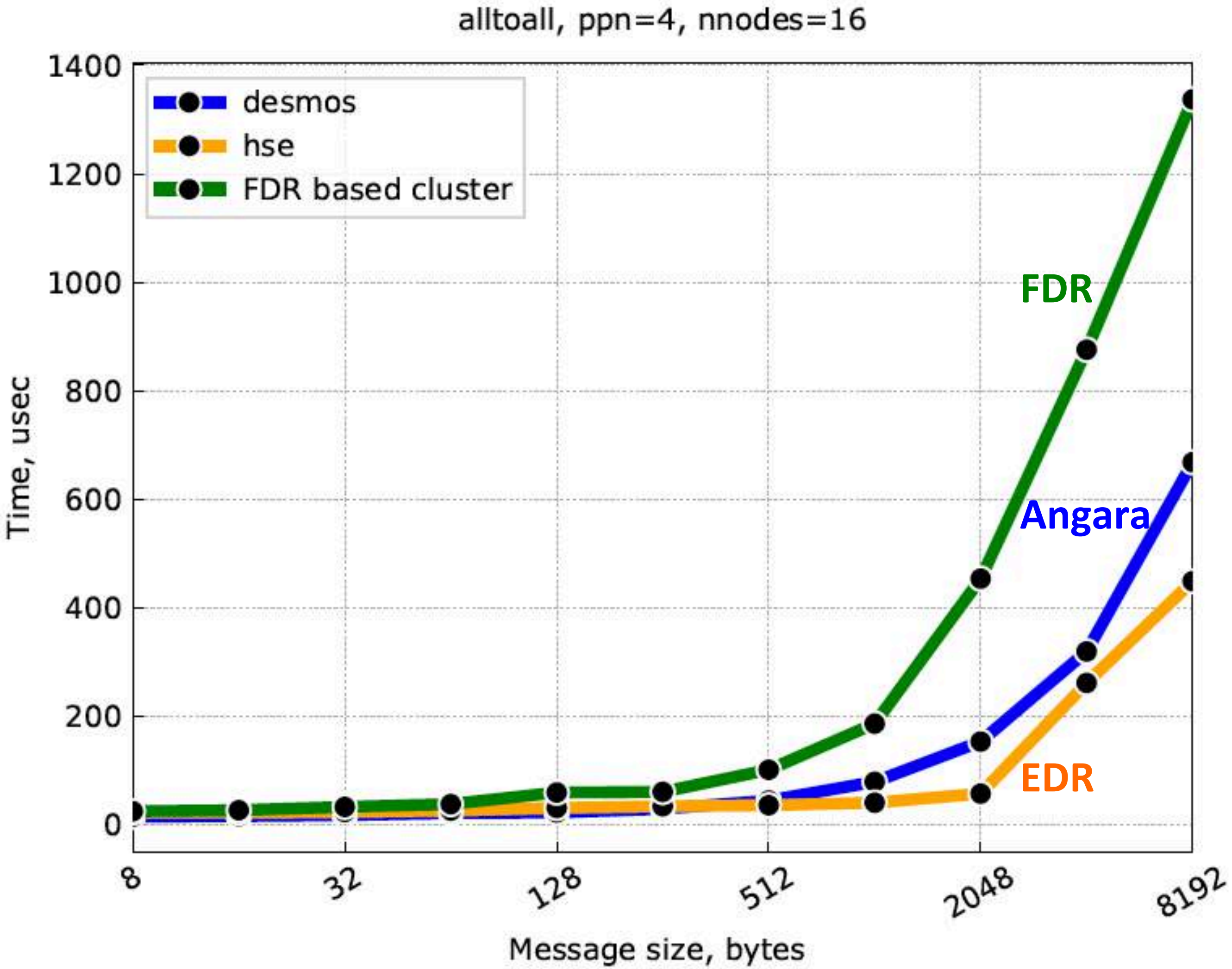
RIB:
2136412
atoms

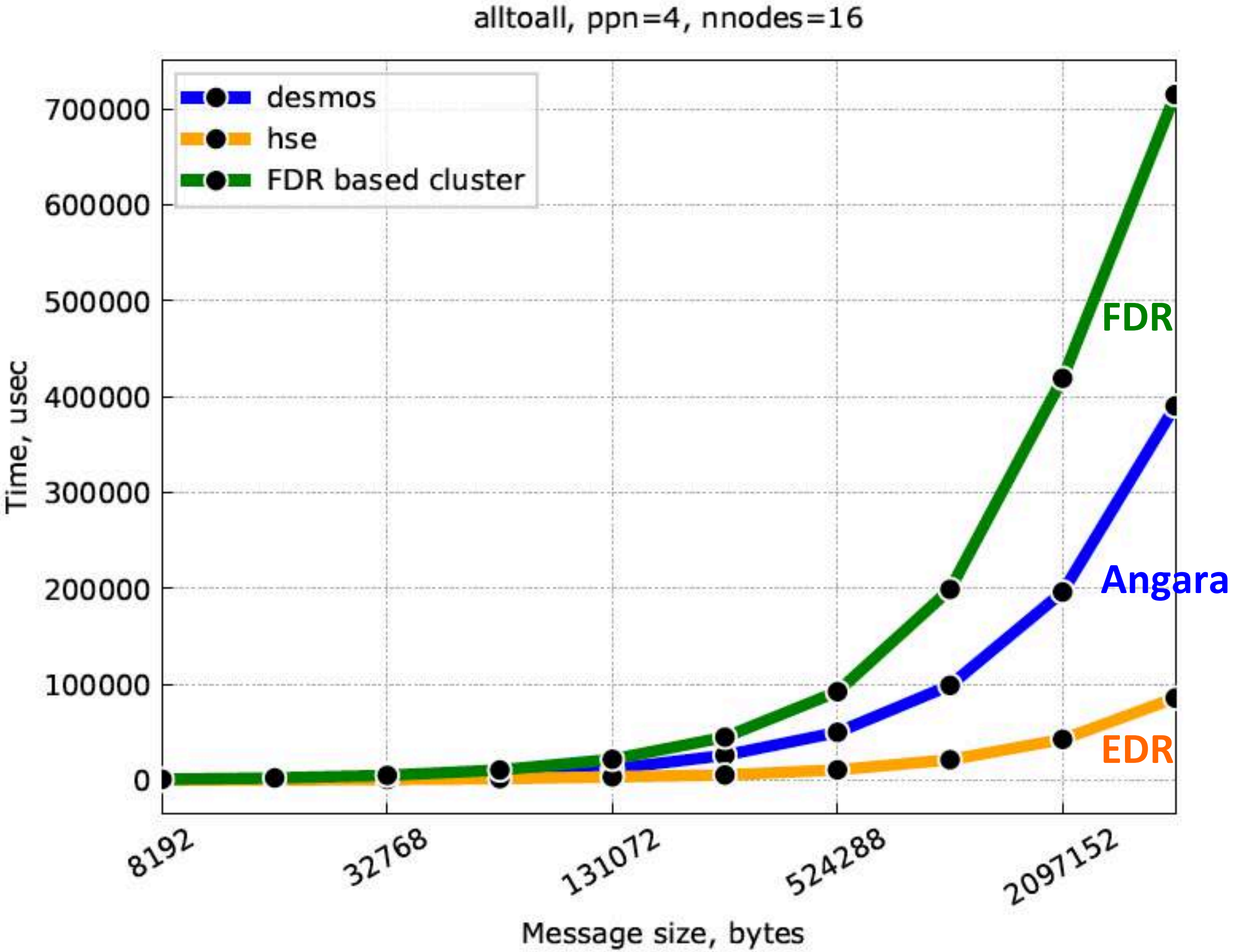



osu_latency
desmos
hse
FDR based cluster
FDR
EDR
Angara
Time, usec
1
2
3
4
5
2
8
32
128
512
2048
8192
Message size, bytes

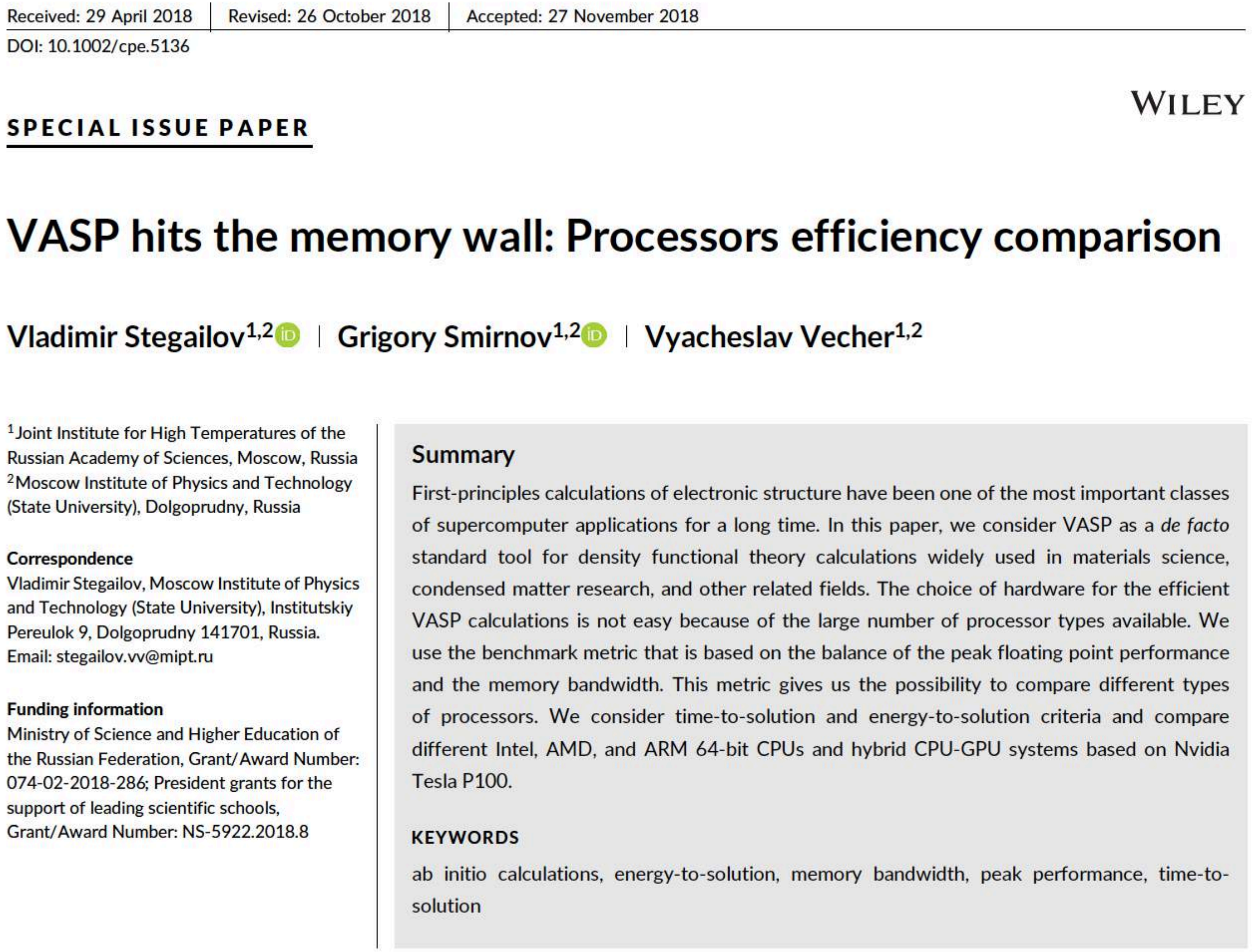

alltoall, ppn=4, msg_size=262144
desmos
hse
FDR based cluster
FDR
Angara
EDR
Time, usec
0
10000
20000
30000
40000
50000
60000
70000
80000
5
10
15
20
25
30
Number of nodes


alltoall, ppn=4, msg_size=16
desmos
hse
FDR based cluster
FDR
EDR
Angara
Time, usec
Number of nodes
0
10
20
30
40
50
5
10
15
20
25
30


alltoall, ppn=4, nnodes=16
desmos
hse
FDR based cluster
FDR
Angara
EDR
Time, usec
1400
1200
1000
800
600
400
200
0
8
32
128
512
2048
8192
Message size, bytes


alltoall, ppn=4, nnodes=16
desmos
hse
FDR based cluster
Time, usec
700000
600000
500000
400000
300000
200000
100000
0
8192
32768
131072
524288
2097152
Message size, bytes
FDR
Angara
EDR

VASP BENCHMARKS



Received: 29 April 2018 | Revised: 26 October 2018 | Accepted: 27 November 2018

DOI: 10.1002/cpe.5136

WILEY

SPECIAL ISSUE PAPER

VASP hits the memory wall: Processors efficiency comparison

Vladimir Stegailov[1,2] | Grigory Smirnov[1,2] | Vyacheslav Vecher[1,2]

[1]Joint Institute for High Temperatures of the Russian Academy of Sciences, Moscow, Russia
[2]Moscow Institute of Physics and Technology (State University), Dolgoprudny, Russia

Correspondence
Vladimir Stegailov, Moscow Institute of Physics and Technology (State University), Institutskiy Pereulok 9, Dolgoprudny 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

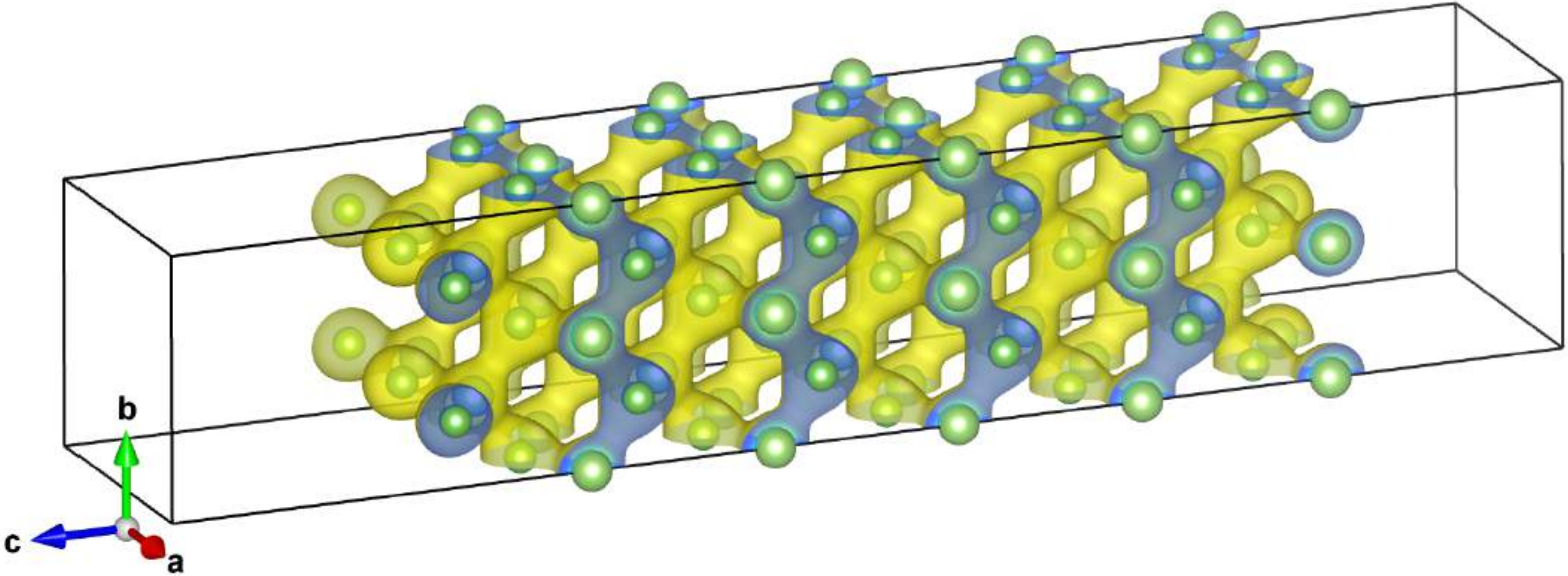
Summary

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



VASP: GaAs benchmark with 80 atoms

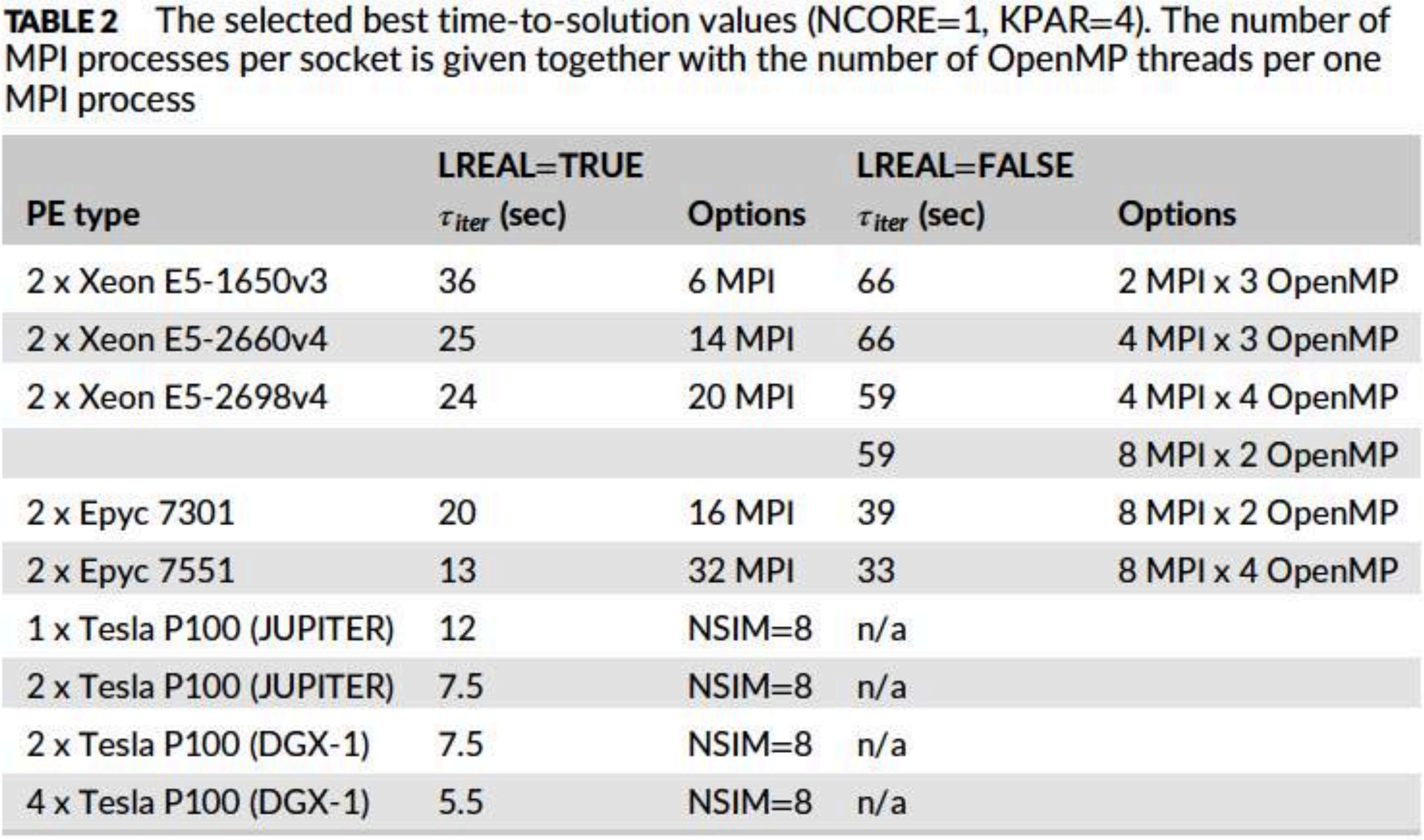
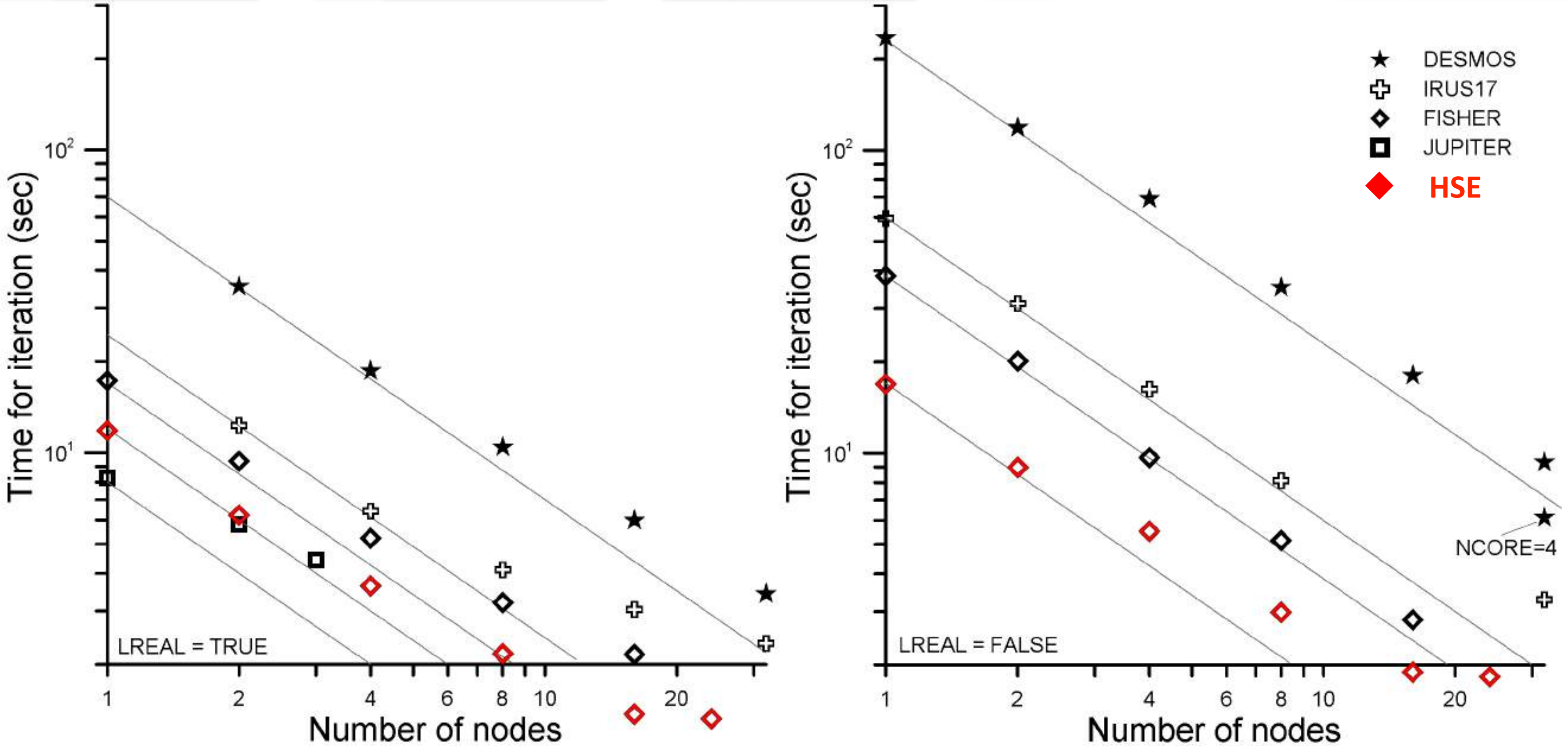


TABLE 2 The selected best time-to-solution values (NCORE=1, KPAR=4). The number of MPI processes per socket is given together with the number of OpenMP threads per one MPI process

	LREAL=TRUE		**LREAL=FALSE**	
PE type	τ_{iter} **(sec)**	**Options**	τ_{iter} **(sec)**	**Options**
2 x Xeon E5-1650v3	36	6 MPI	66	2 MPI x 3 OpenMP
2 x Xeon E5-2660v4	25	14 MPI	66	4 MPI x 3 OpenMP
2 x Xeon E5-2698v4	24	20 MPI	59	4 MPI x 4 OpenMP
			59	8 MPI x 2 OpenMP
2 x Epyc 7301	20	16 MPI	39	8 MPI x 2 OpenMP
2 x Epyc 7551	13	32 MPI	33	8 MPI x 4 OpenMP
1 x Tesla P100 (JUPITER)	12	NSIM=8	n/a	
2 x Tesla P100 (JUPITER)	7.5	NSIM=8	n/a	
2 x Tesla P100 (DGX-1)	7.5	NSIM=8	n/a	
4 x Tesla P100 (DGX-1)	5.5	NSIM=8	n/a	

VASP: GaAs benchmark with 80 atoms

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



VASP: GaAs benchmark with 80 atoms

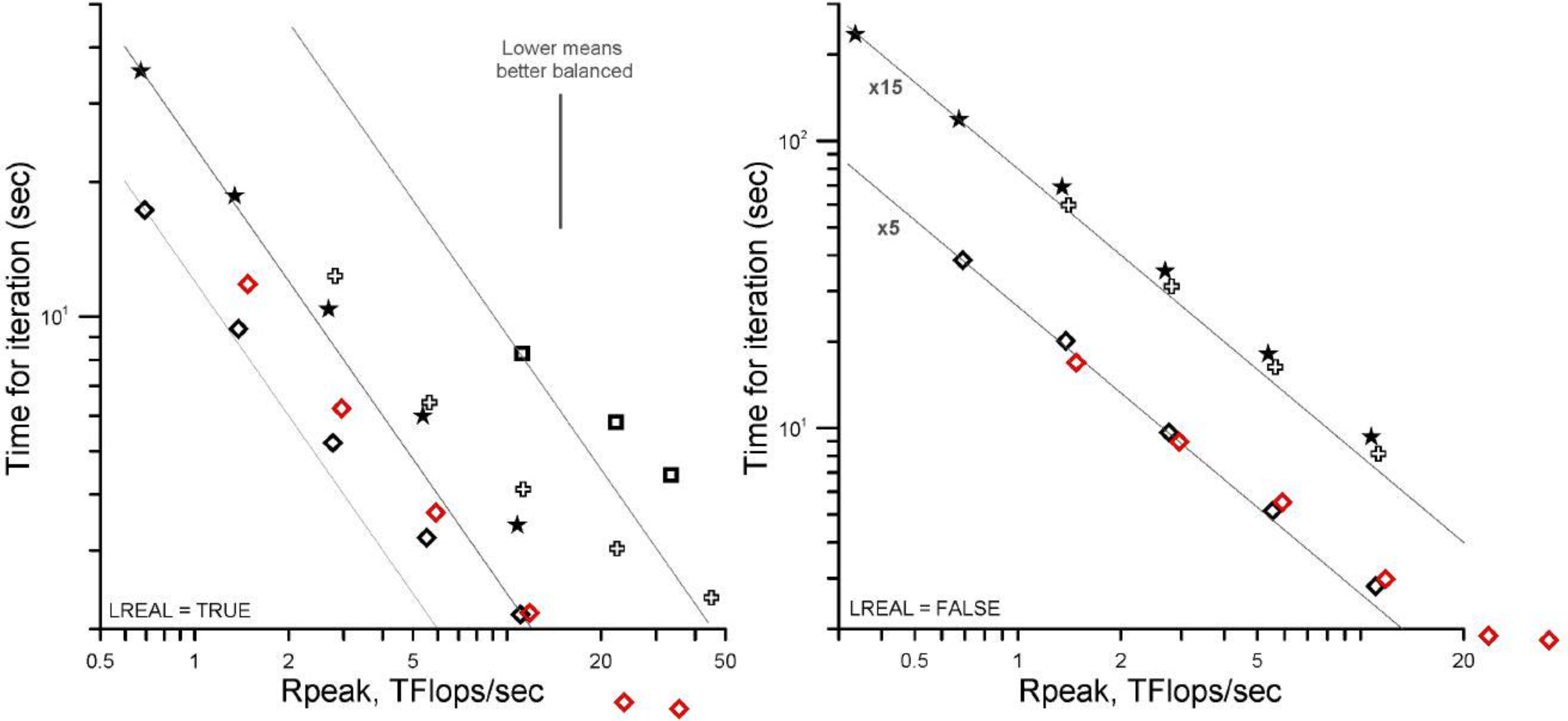
★ E5-1650v3 + Angara

✚ 2x E5-2698v4 + OmniPath

□ P8 + P100 + EDR

◆ 2x Epyc 7301 + FDR

◆ 2x Gold 6152 + EDR



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.