Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

SuperMUC Extreme Scale-Out Phase 2, lessons learned

Nicolay J. Hammer
Application Support Group LRZ

Outline

- Workshop Overview
- Some Job Statistics
- Some Results of single user groups
- Summary and Conclusions

„Extreme Scale-out“ 28 days later

Friendly-User Phase of the recently brought up SuperMUC Phase 2 (3.6 PFlop/s peak, 2.8 Pflop/s Linpack, 86016 cores)

Available: **63.4** million core-h
Used: **43.8** million core-h

41 Scientists from **14** Institutes
14 Applications running on full system

Extreme Scale-out Phase2, lessons learned

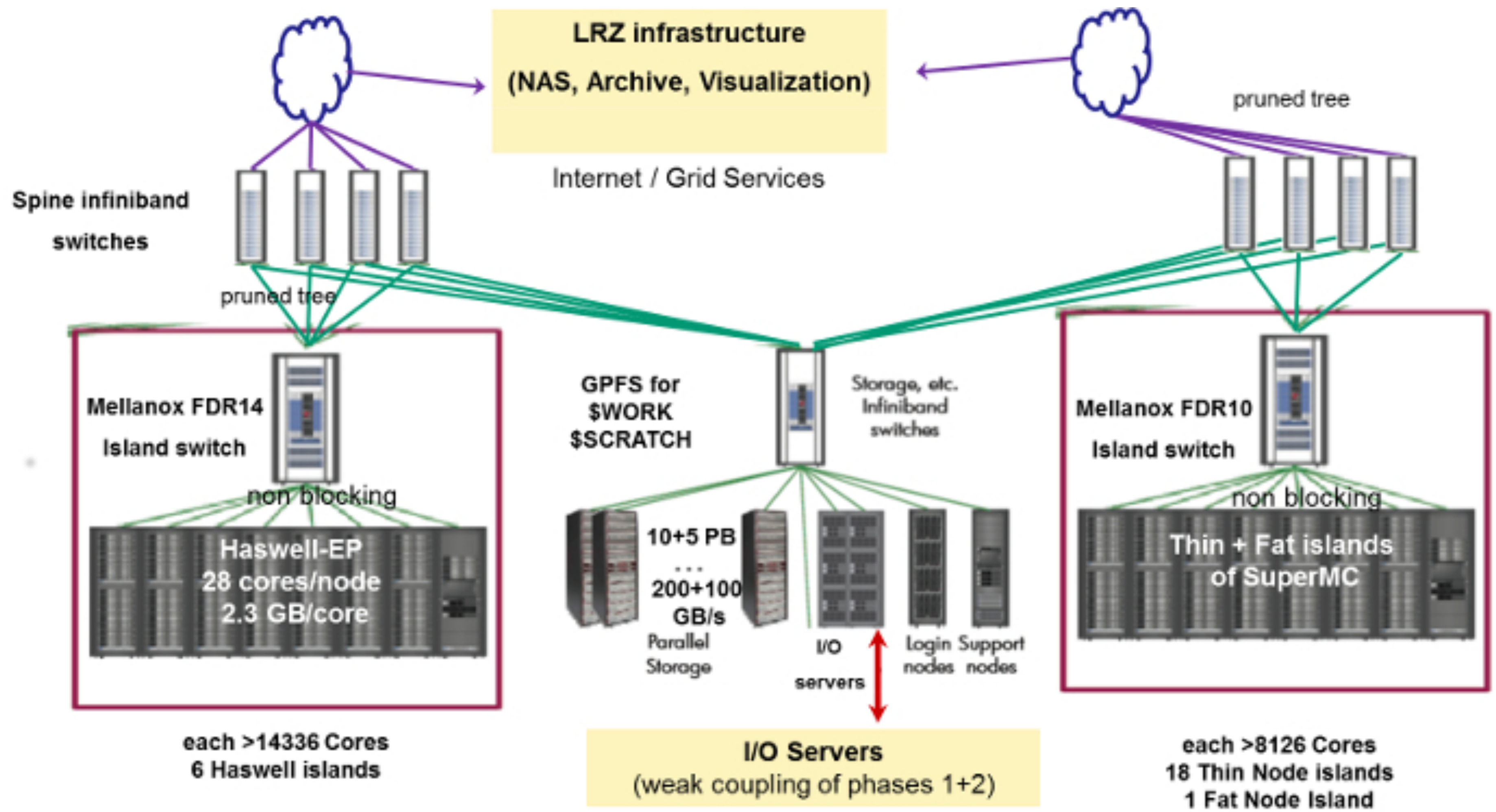
Ferdinand Jamitzky[1], Helmut Satzger[1], Nicolay Hammer[1], Momme Allalen[1], Alexander Block[1], Markus M. Müller[1], Anupam Karmakar[1], Matthias Brehm[1], Reinhold Bader[1], Luigi Iapichino[1], Antonio Ragagnin[1], Vasilios Karakasis[1], Dieter Kranzlmüller[1], Arndt Bode[1], Herbert Huber[1], Martin Kühn[2], Rui Machado[2], Daniel Grünewald[2], Philipp V. F. Edelmann[3], Friedrich K. Röpke[3], Markus Wittmann[4], Thomas Zeiser[4], Gerhard Wellein[5], Gerald Mathias[6], Magnus Schwörer[6], Konstantin Lorenzen[6], Christoph Federrath[7], Ralf Klessen[8], Karl-Ulrich Bamberg[9], Hartmut Ruhl[9], Florian Schornbaum[10], Martin Bauer[10], Anand Nikhil[11], Jiaxing Qi[11], Harald Klimach[11], Hinnerk Stüben[12], Abhishek Deshmukh[13], Tobias Falkenstein[13], Klaus Dolag[14], and Margarita Petkova[14]

[1] LRZ, Boltzmannstrasse 1, 85748 Garching b. Muenchen http://www.lrz.de
[2] CCHPC - Fraunhofer ITWM, Fraunhofer Platz 1, 67663 Kaiserslautern http://www.gpi-site.com
[3] Heidelberger Institut für Theoretische Studien, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany
[4] Erlangen Regional Computer Center (RRZE), University of Erlangen-Nürnberg, Martensstr. 1, 91058 Erlangen, Germany
[5] Department of Computer Science, University of Erlangen-Nürnberg, Germany
[6] Lehrstuhl für Biomolekulare Optik, Ludwig-Maximilians-Universität München, Oettingenstr. 67, 80538 München, Germany
[7] Research School of Astronomy and Astrophysics, The Australian National University, Canberra, ACT 2611, Australia
[8] Universität Heidelberg, Zentrum für Astronomie, Institut für Theoretische Astrophysik, Albert-Ueberle-Strasse 2, D-69120 Heidelberg, Germany
[9] Chair for Computational and Plasma Physics at the LMU, Munich
[10] Chair for System Simulation, University of Erlangen-Nürnberg, Cauerstraße 11, 91058 Erlangen, Germany
[11] Chair of Simulation Techniques & Scientific Computing, University Siegen
[12] Universität Hamburg, Zentrale Dienste, Schlüterstraße 70, 20146 Hamburg
[13] Fakultät für Maschinenwesen, Institut für Technische Verbrennung, RWTH Aachen University, Templergraben 64, 52062 Aachen
[14] Universitäts-Sternwarte München, Scheinerstr. 1, D-81679 München, Germany

Abstract. During May and June 2015 LRZ conducted a friendly user operation block operation of the their upcoming new extension of SuperMUC called Phase 2 which consists of 86,016 Intel Haswell cores distributed to 6 islands resulting in a peak performance of 3.6 PFlop/s. Selected user groups had the opportunity to use the system for 28 days of continuous operation as so called "friendly users" and run jobs up to the whole system size. This work presents results obtained during this period and the lessons learned from the operational point of view.

Keywords: Supercomputing, HPC

SuperMUC Phase 2 System



- 6 Islands (86016 cores)
- each Island has 512 nodes (14336 cores)
- each node has 2 Intel Xeon E5-2697 (Haswell) with 14 cores, 64 GB
- Interconnect Infiniband FDR14, non-blocking/pruned fat tree

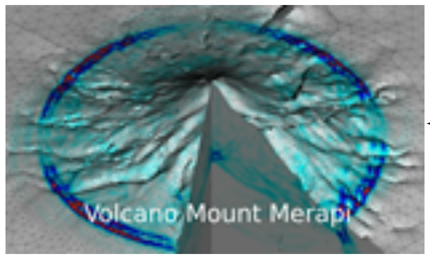
History of workshops

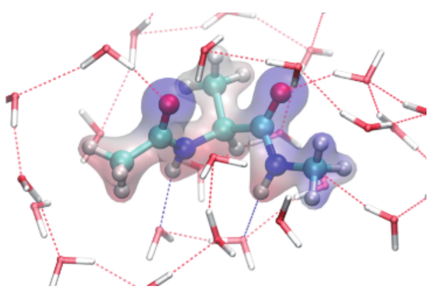
1. Workshop 2013
   12 Applications (6 on 16 Islands):
   BQCD, CIAO, Gadget3-XXL, GROMACS, LAMMPS, Nyx, Vertex3D, waLBerla, APES, SeisSol, ExaML, ICON

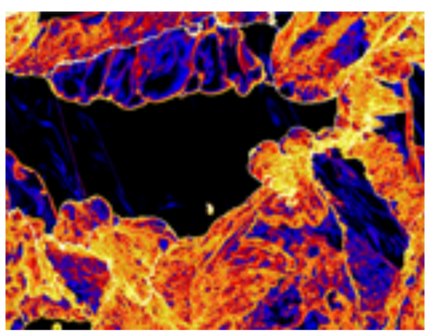
2. Workshop 2014
   10+5 Applications: Ateles, FluidNN, GASPI, nsCouette, Alya System, Seven-league, ACRONYM, PSC, BQCD, psOpen
   Production runs: GADGET, SeisSol, Vertex3D, CIAO, Intel HPCG

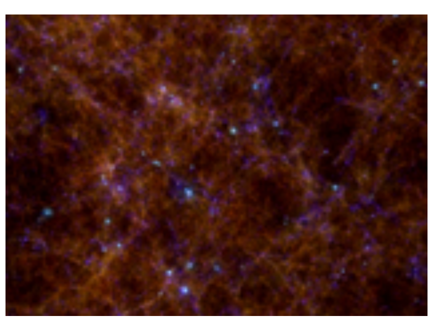
3. Workshop 2015
   14 Applications: BQCD, SeisSol, GASPI, Seven-league, ILBDC, Iphigenie, FLASH, GADGET3, PSC, waLBerla, Vertex3D, LS1-Mardyn, CIAO, Musubi

14 Applications 2015



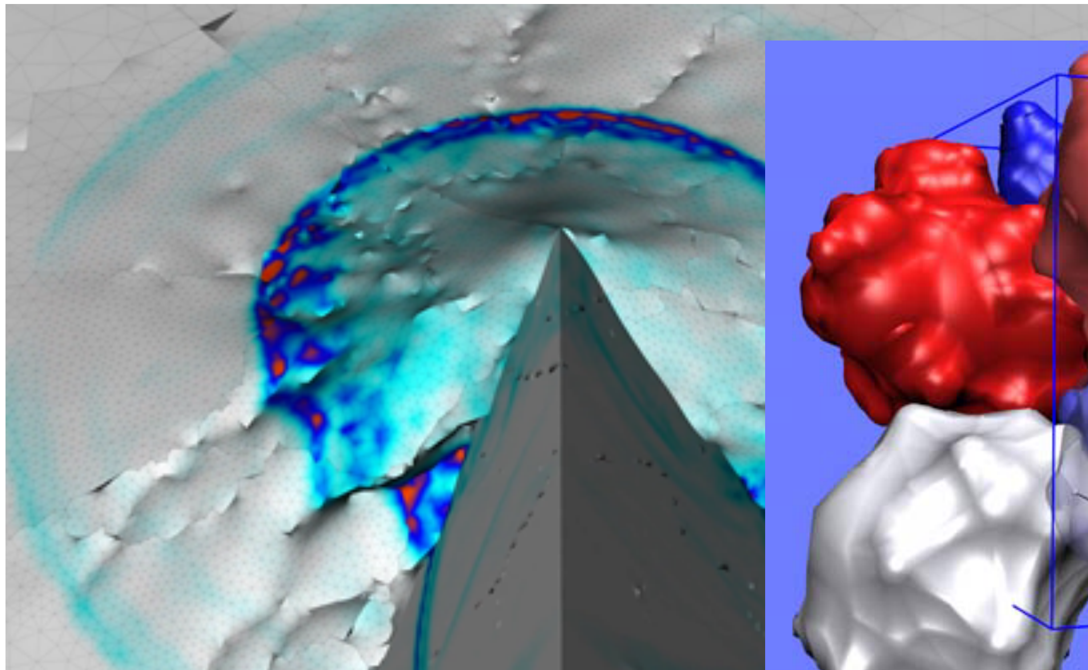


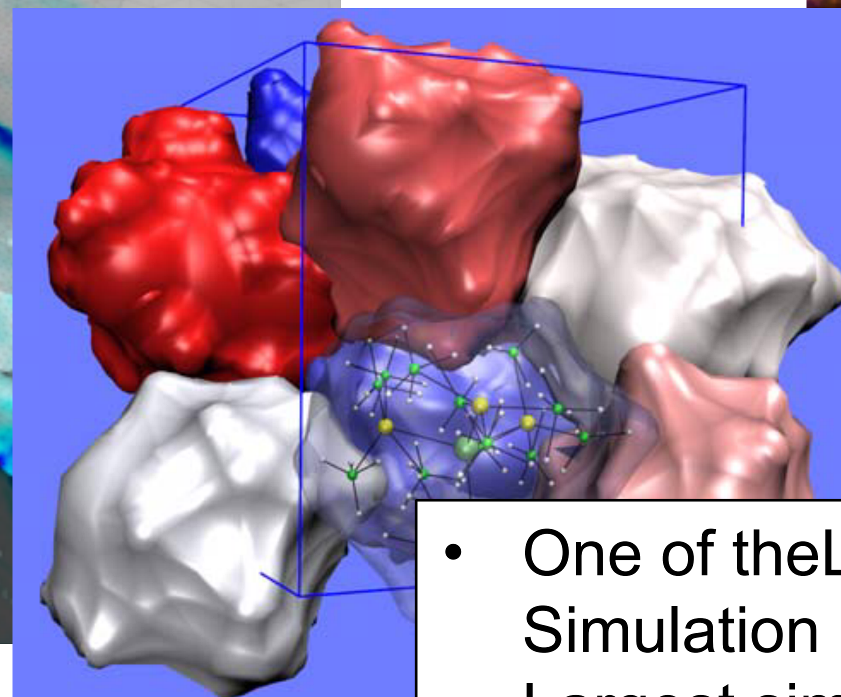


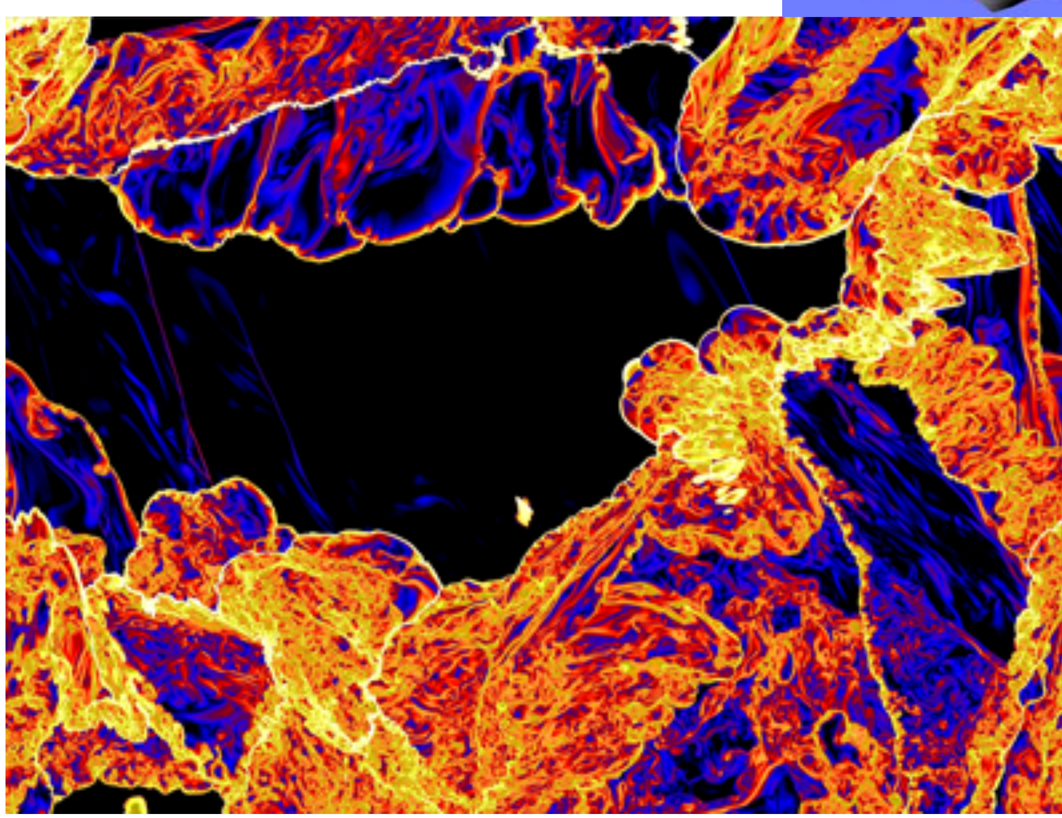


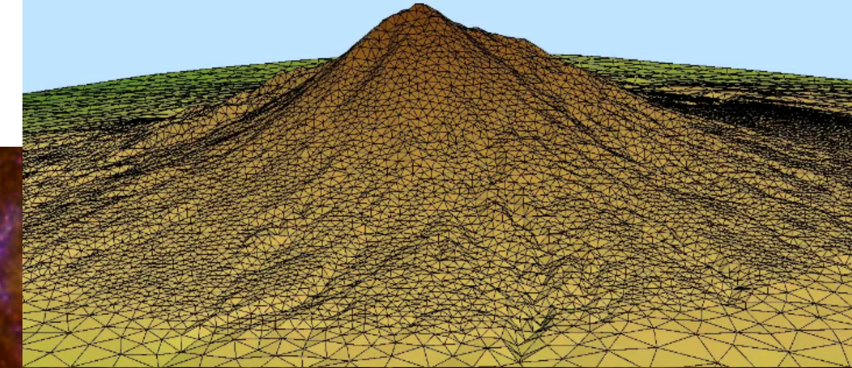
Software	Application
BQCD	Quantumchomodynamics
SeisSol	Seismology
GPI-2 / GASPI	Global Adress Space Library
Seven-League Hydro	Stellar Astropysics
ILBDC	Lattice Boltzmann
Iphigenie	Molecular Dynamics
FLASH	Astrophysics CFD
Gadget	Cosmology
PSC	Plasmaphysics
waLBerla	Lattice Boltzmann
Musubi	Lattice Boltzmann
CIAO	CFD, Combustion
Vertex3D	Stellar Astrophysics
LS1-Mardyn	Material Science

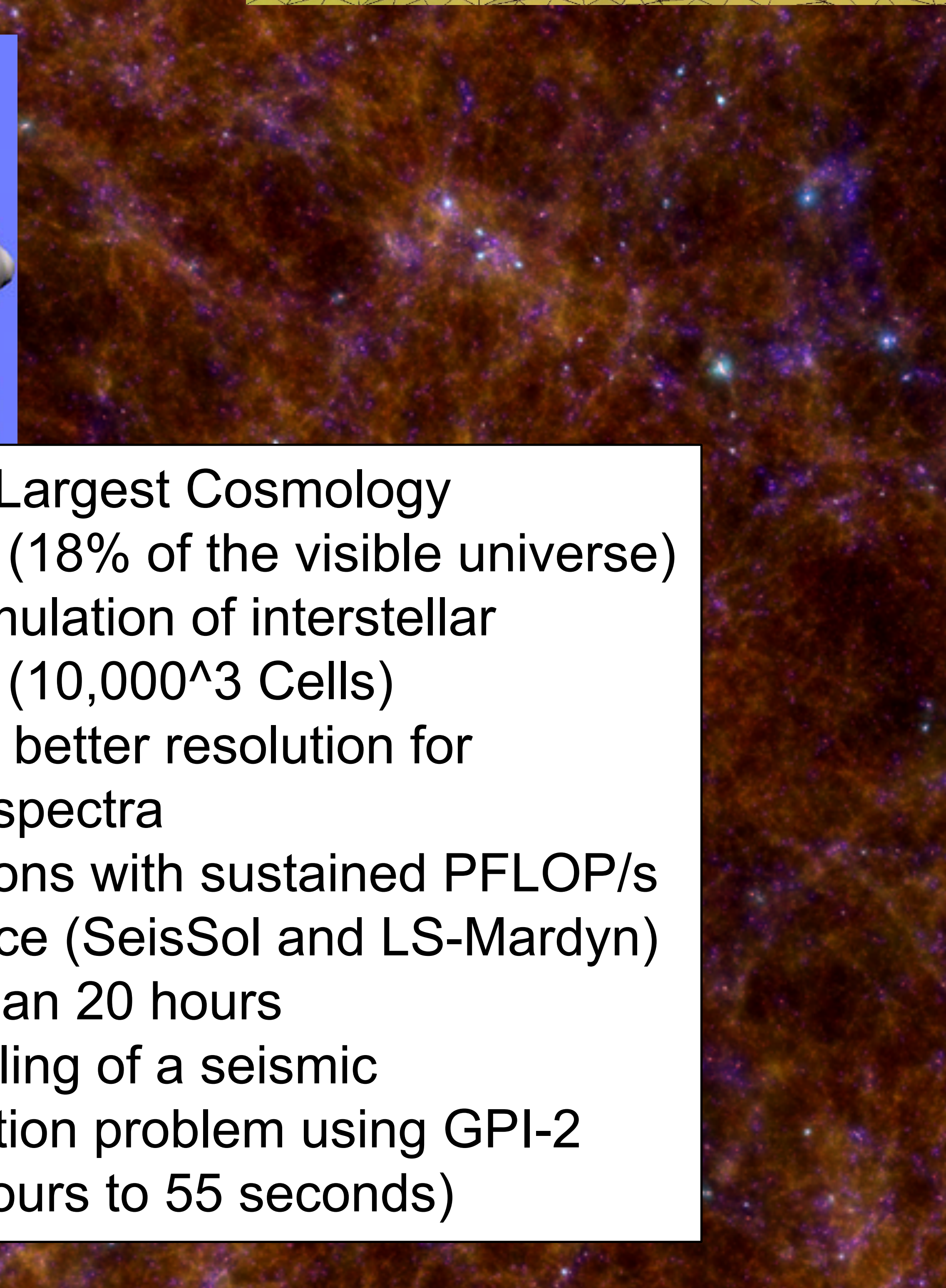
Achievments











- One of theLargest Cosmology Simulation (18% of the visible universe)
- Largest simulation of interstellar turbulence (10,000^3 Cells)
- Factor 100 better resolution for molecular spectra
- 2 Applications with sustained PFLOP/s Performance (SeisSol and LS-Mardyn) for more than 20 hours
- Strong scaling of a seismic reconstruction problem using GPI-2 (from 16 hours to 55 seconds)

Top users

TOP10 Users

TOP	User	Million Core hours
1	di72hod2	15.2
2	lu78qer5	6.4
3	di56dok2	4.7
4	lu78maw4	2.3
5	di73qeb	2.0
6	lu79hah2	1.8
7	lu24viv7	1.1
8	a2815ae	0.98
9	di98wul	0.80
10	di98bix	0.55

Total available:
63,432,000 core hours
Total used:
43,758,430 core hours
Utilisation:
68.98%

Job statistics

Extreme Scale-Out Phase 2

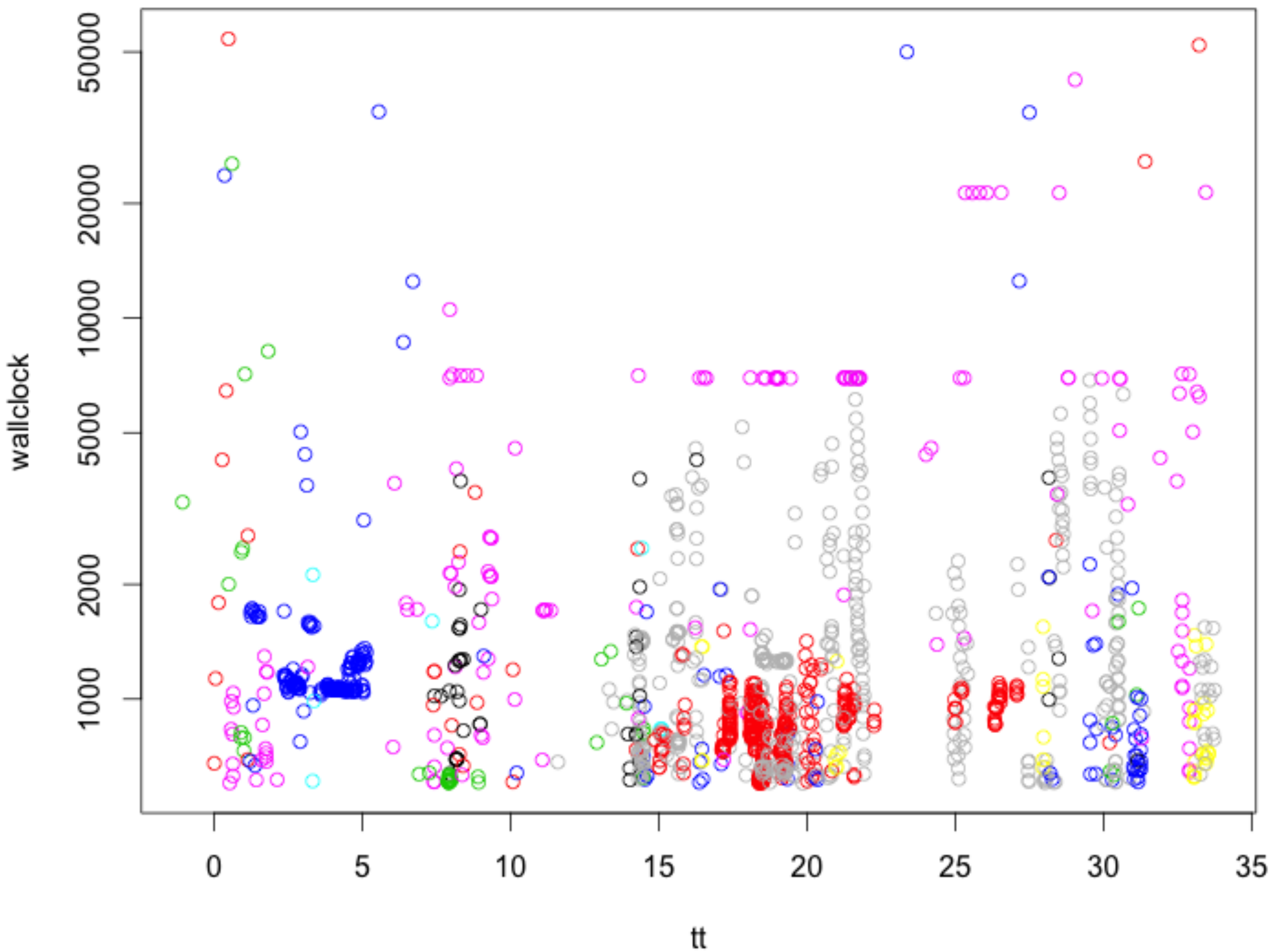
12.5.2015 – 12.6.2015
30 (28) days

Nightly Operation: general queue max 3 islands
Daytime Operation: special queue max 6 islands (full system, dedicated)

6751 jobs (> 1 min)
2054 jobs (>10 min)

general	special	test	tmp1
3600	1567	1575	9

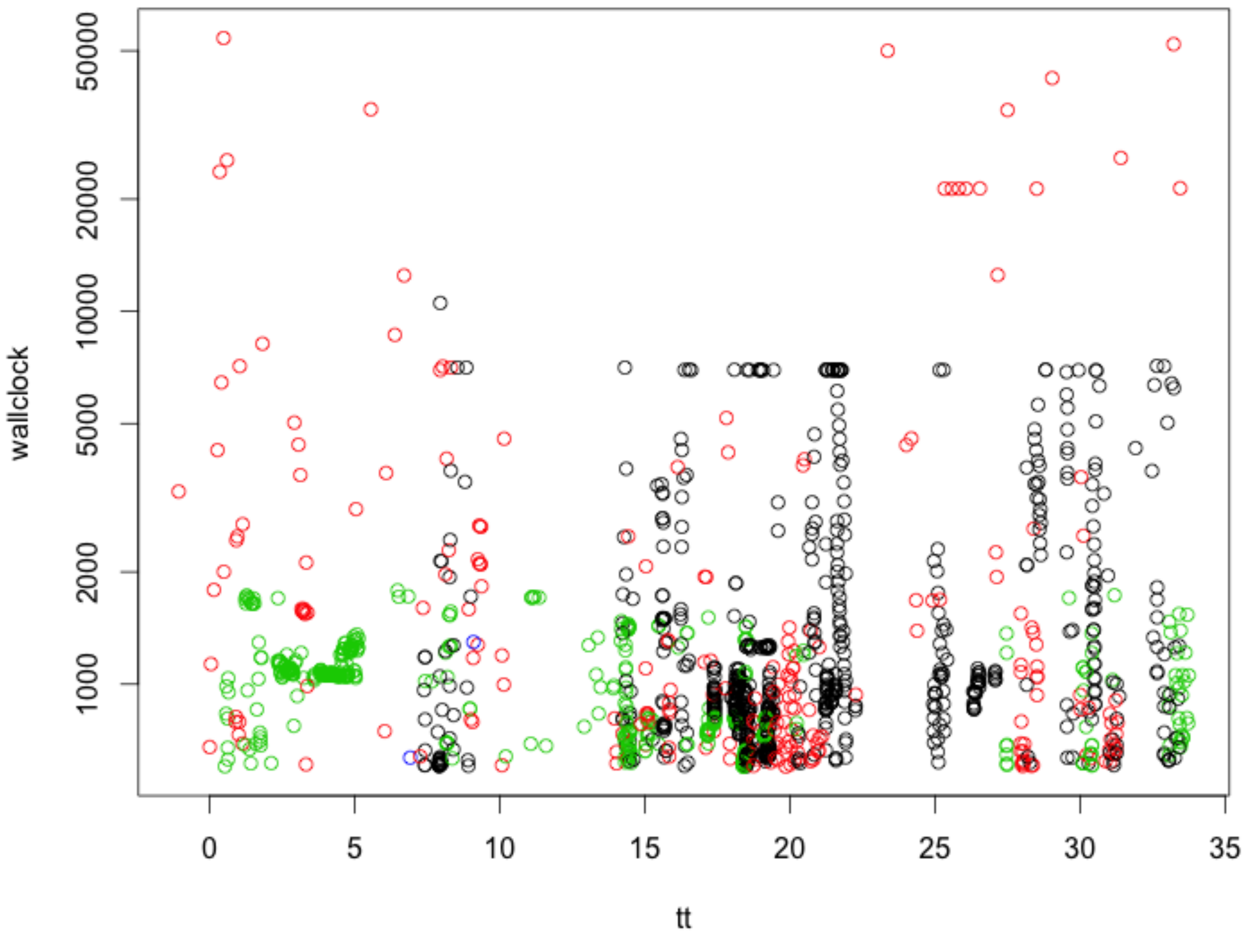
Timeline May 12th – June 15th (jobs > 10 min)



Jobs by users:

- dedicated slots
- weekends
- production
- scaling/tests

Timeline May 12th – June 15th (jobs > 10 min)

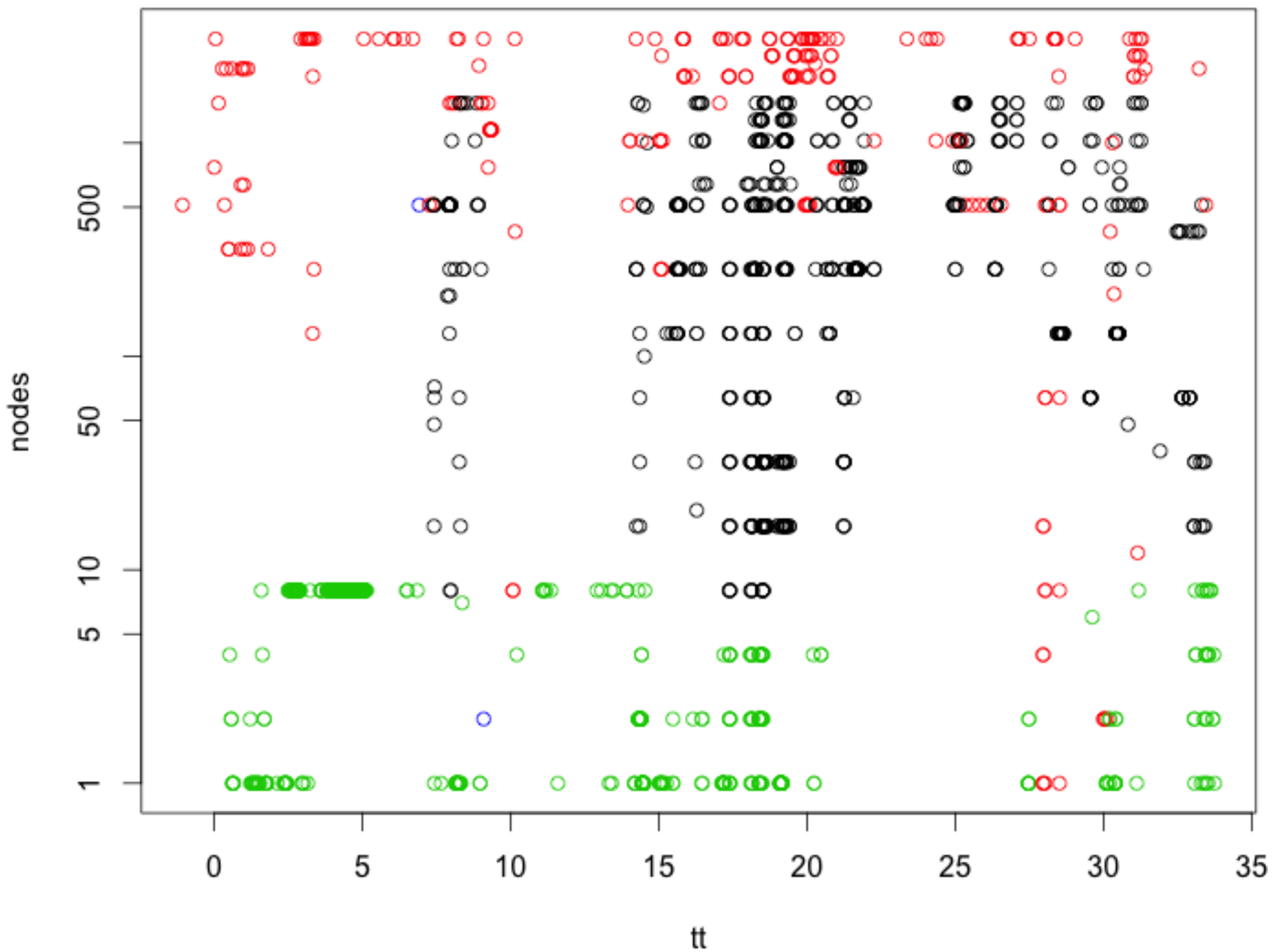


Queues:

- Special (6 Islands, daytime)
- Test (max 30min)
- General (max 2 hours)

- jobs in Special/General
  - production runs
  - scaling/test run
- series of scaling runs (script)

Timeline May 12th – June 15th (jobs > 10 min)

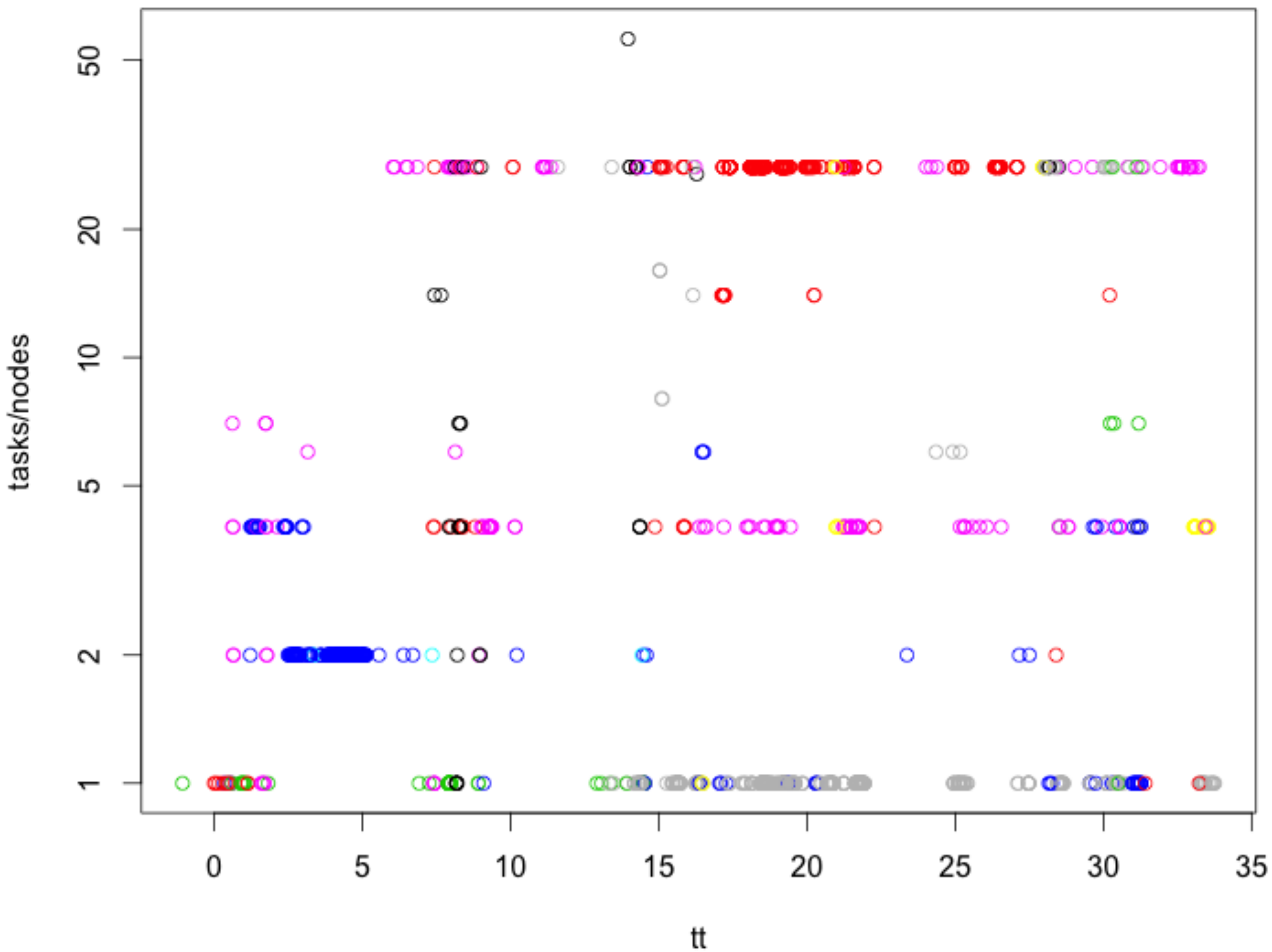


Queues:

- Special (6 Islands, daytime)
- Test (max 8 nodes)
- General (max 3 Islands)

- jobs in Special/General
  - weekends
  - “back filling”
- series of scaling runs (script)

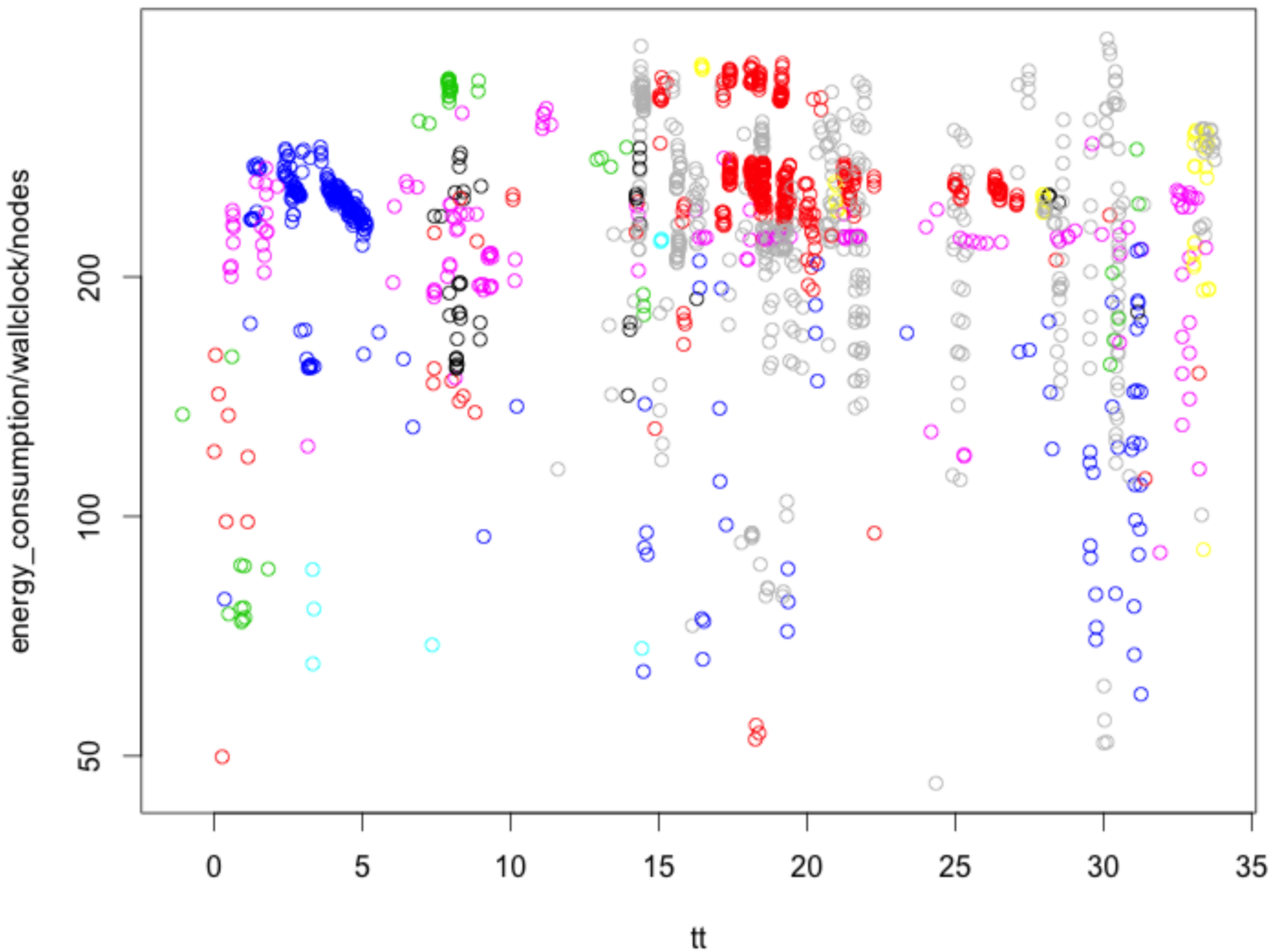
Timeline Tasks/Node (jobs > 10 min)



Jobs by users

- pure MPI
  - task/node = 28
- hybrid MPI+OpenMP
  - task/node = 1,2,4
- hyper threading
- scaling task per node

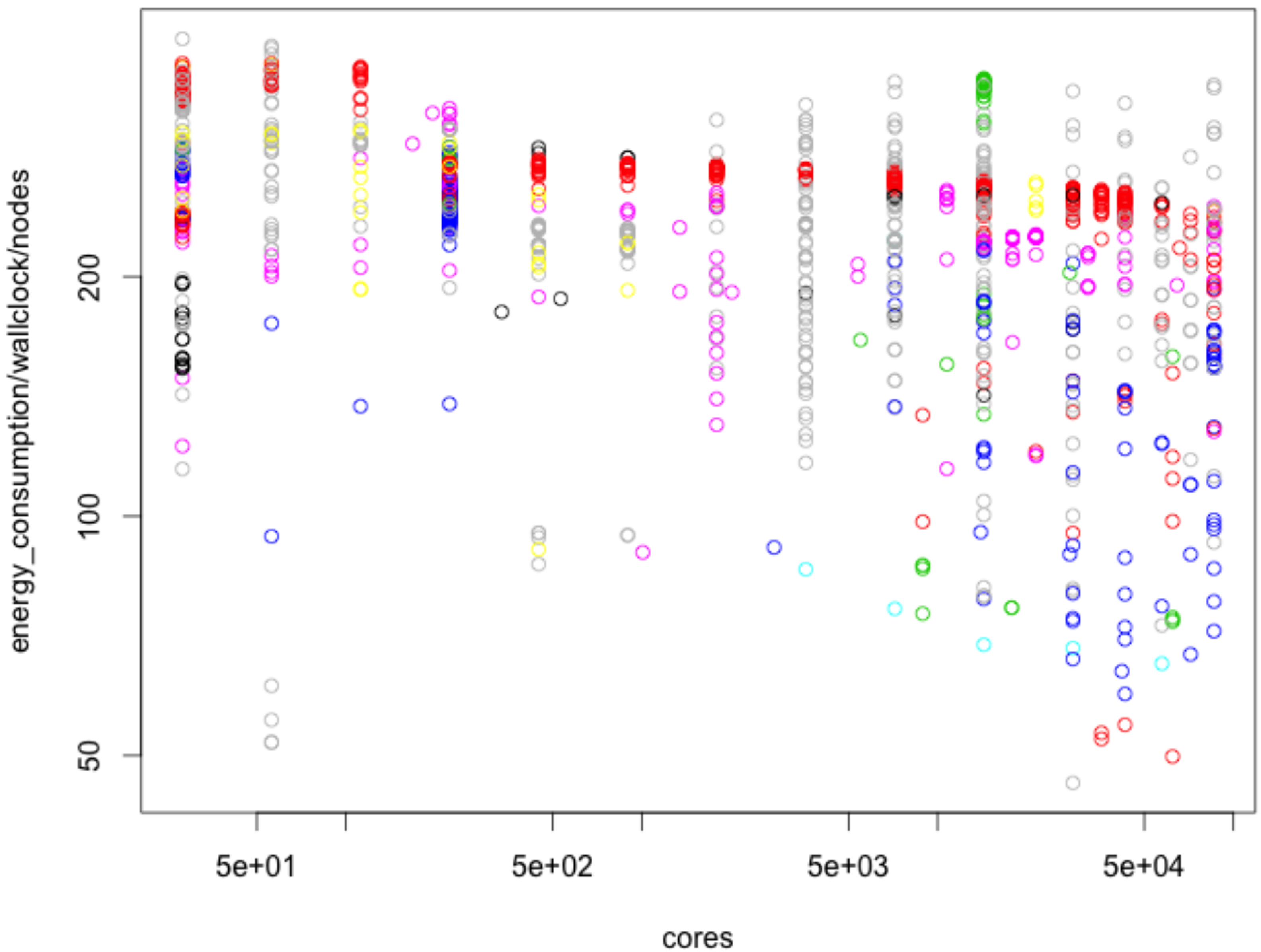
Power (Watts/node) vs. Time



Jobs by users

- max. ~300 W/node
- performance optimization
- type of algorithms
- scaling clock speed

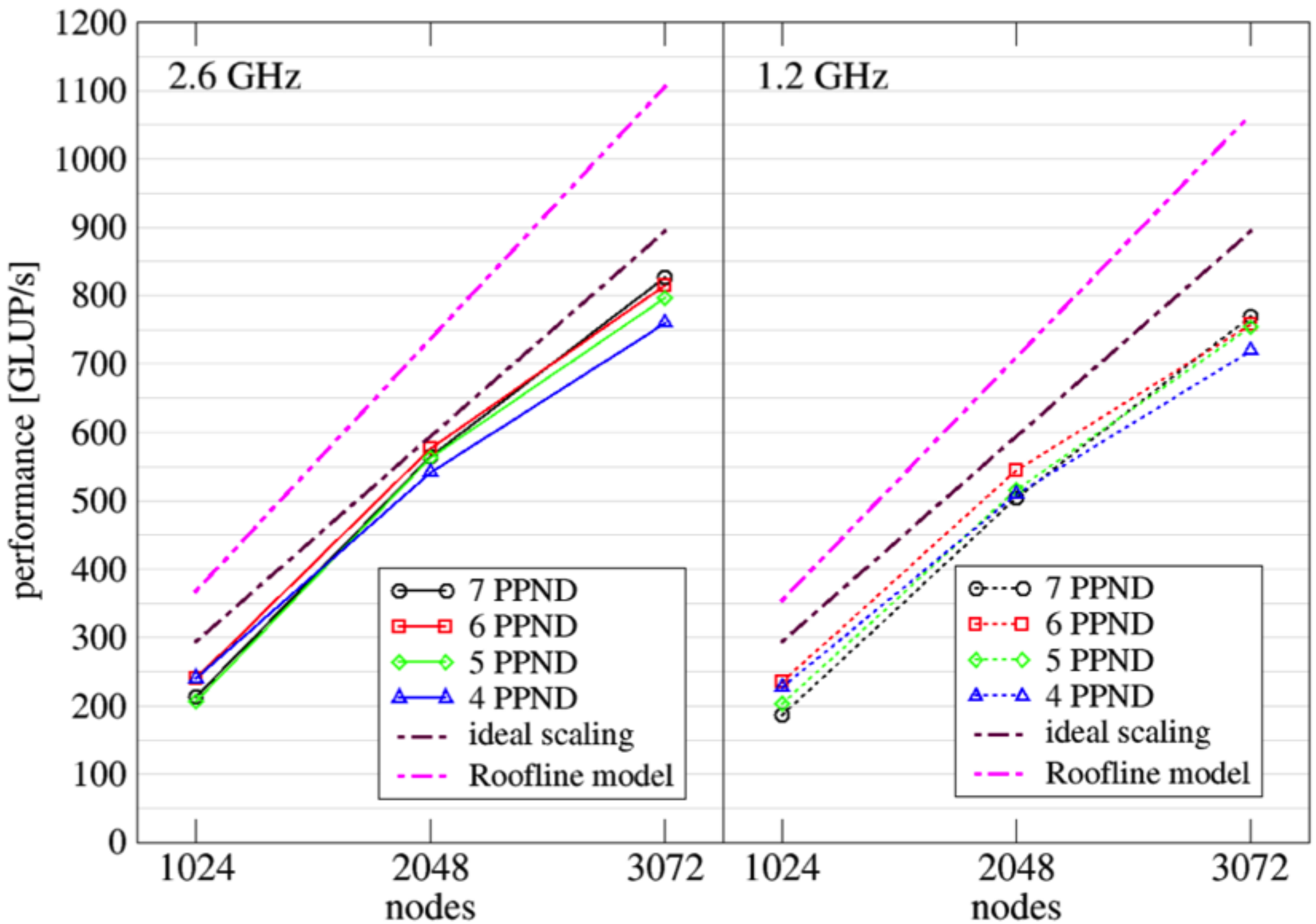
Power (Watts/node) vs. #Cores



Jobs by users

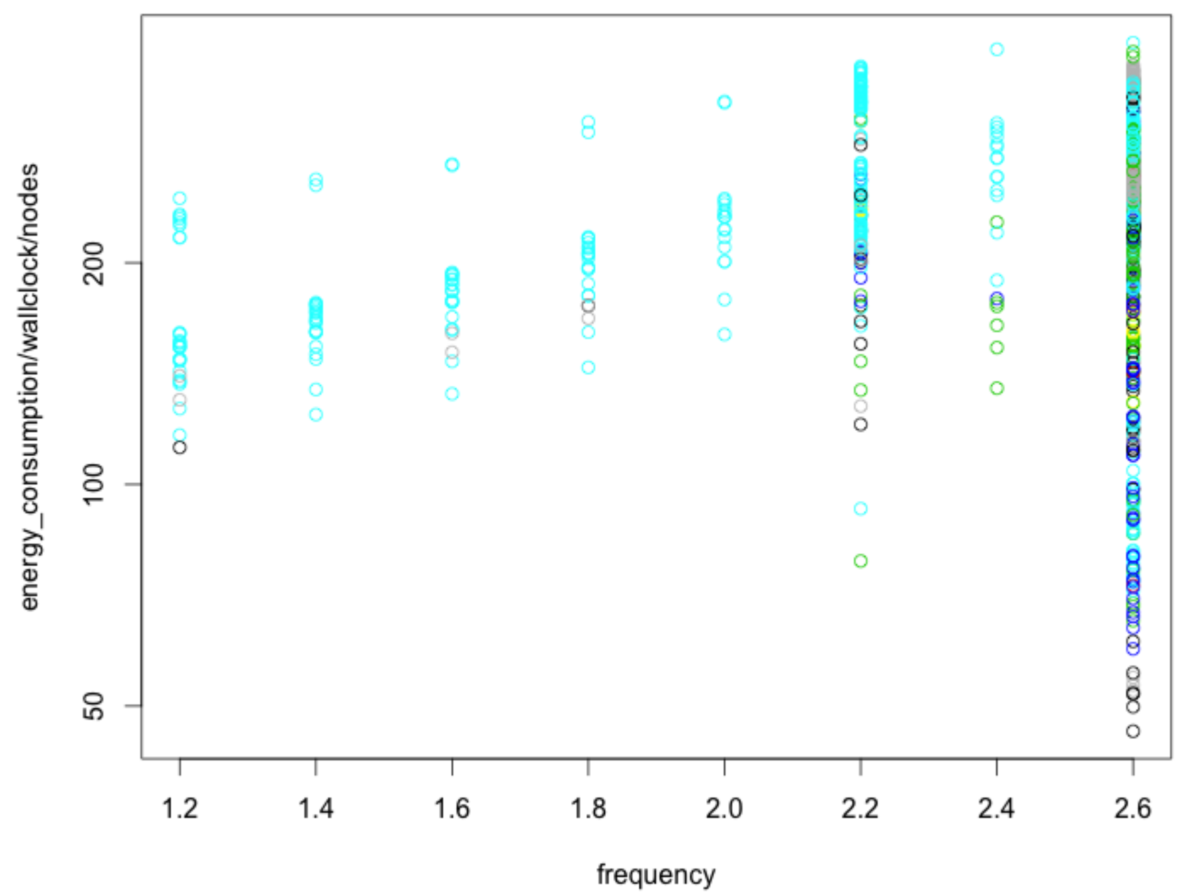
- power/node vs.
  - number of nodes
  - clock speed
  - job setup
- job failures
  - idle nodes
  - hanging jobs

ILBDC (RRZE, FAU Erlangen)



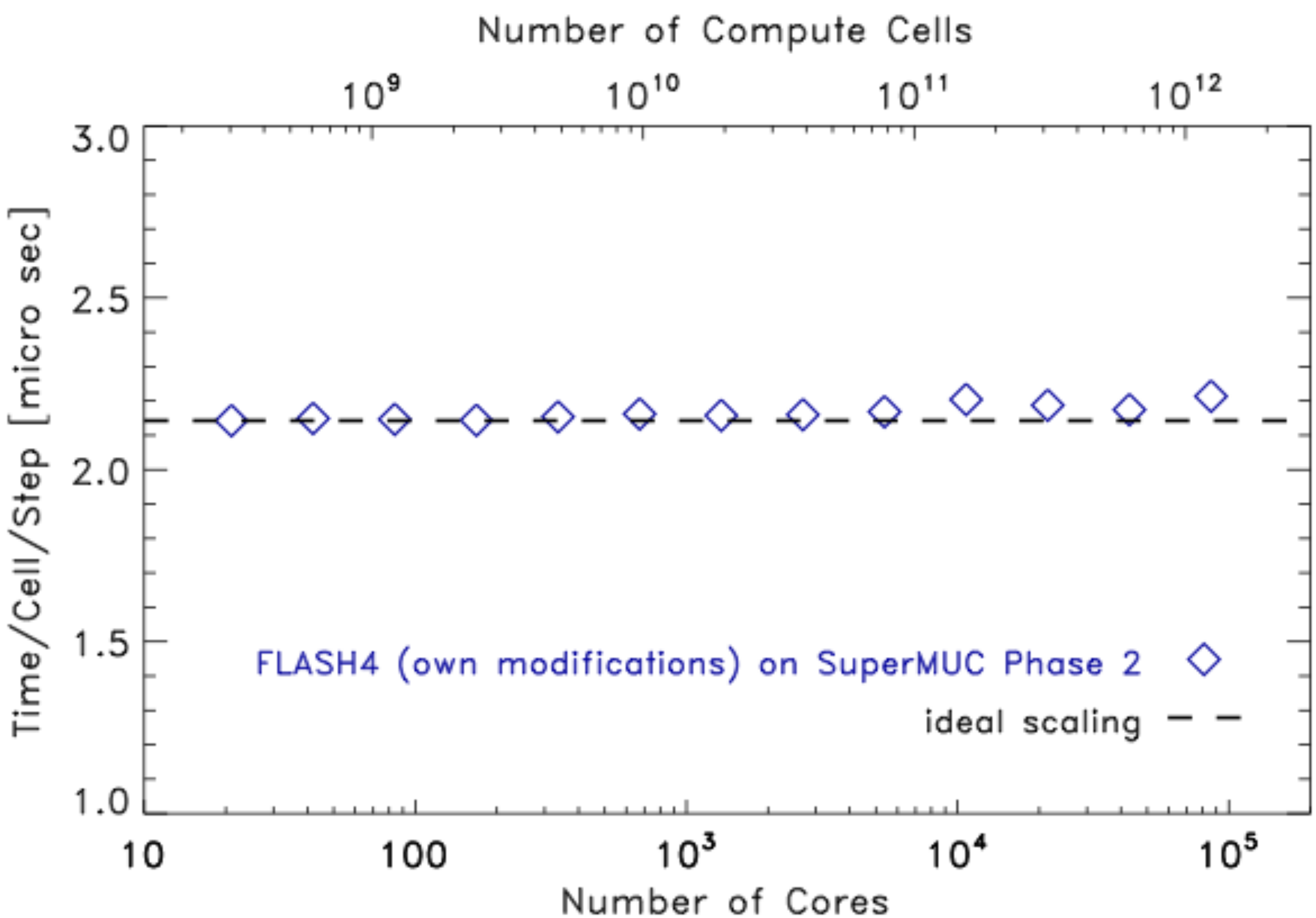
D3Q19-TRT Lattice Boltzmann solver

- strong scaling case
- scaling 1.2GHz - 2.6GHz
- 4 -7 tasks / NUMA domain



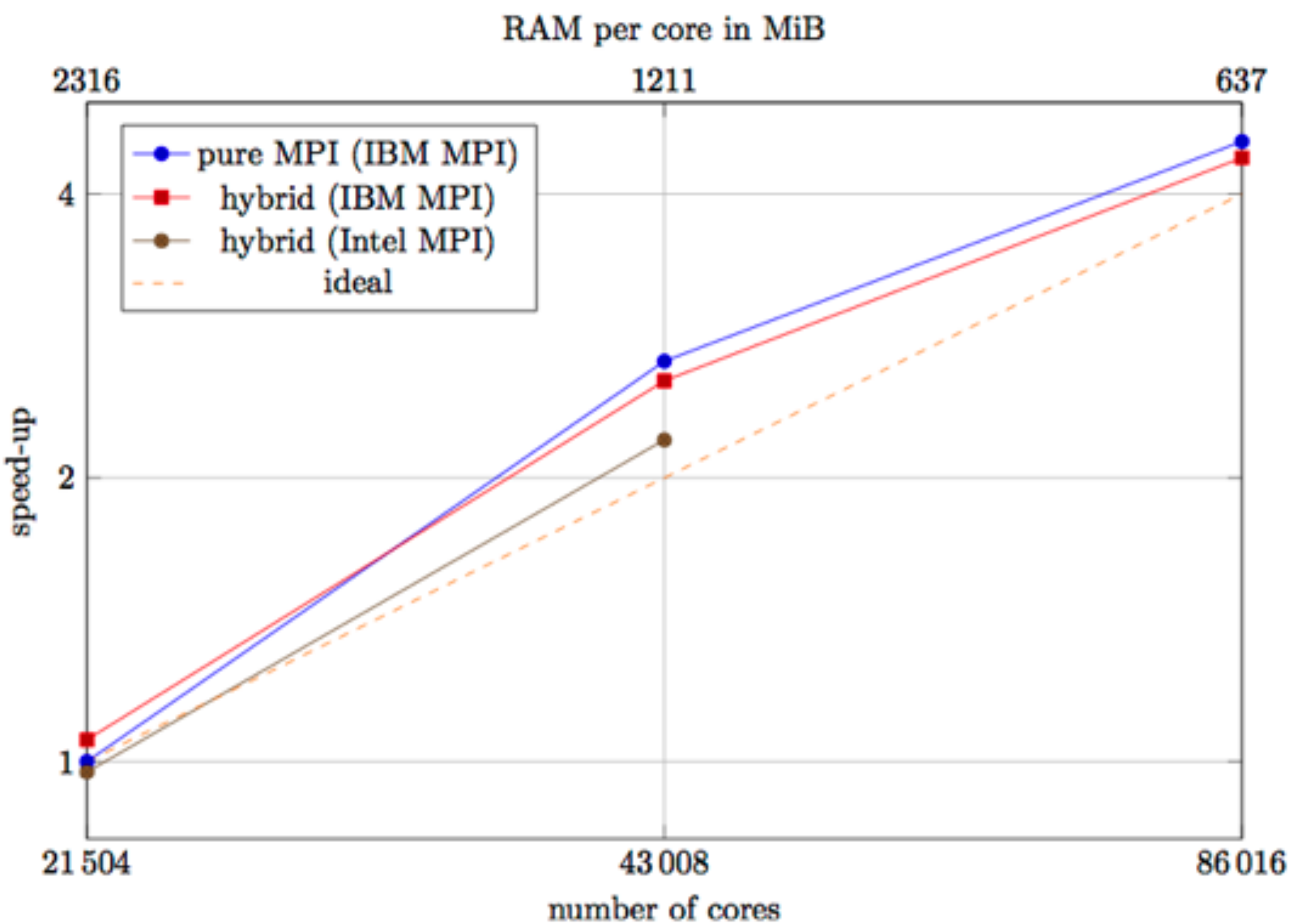
- 87% of STREAM mem. BW
- 93% performance @ 1.2GHz
- clock speed in-sensitive
- energy efficient computing

FLASH (Uni Heidelberg)



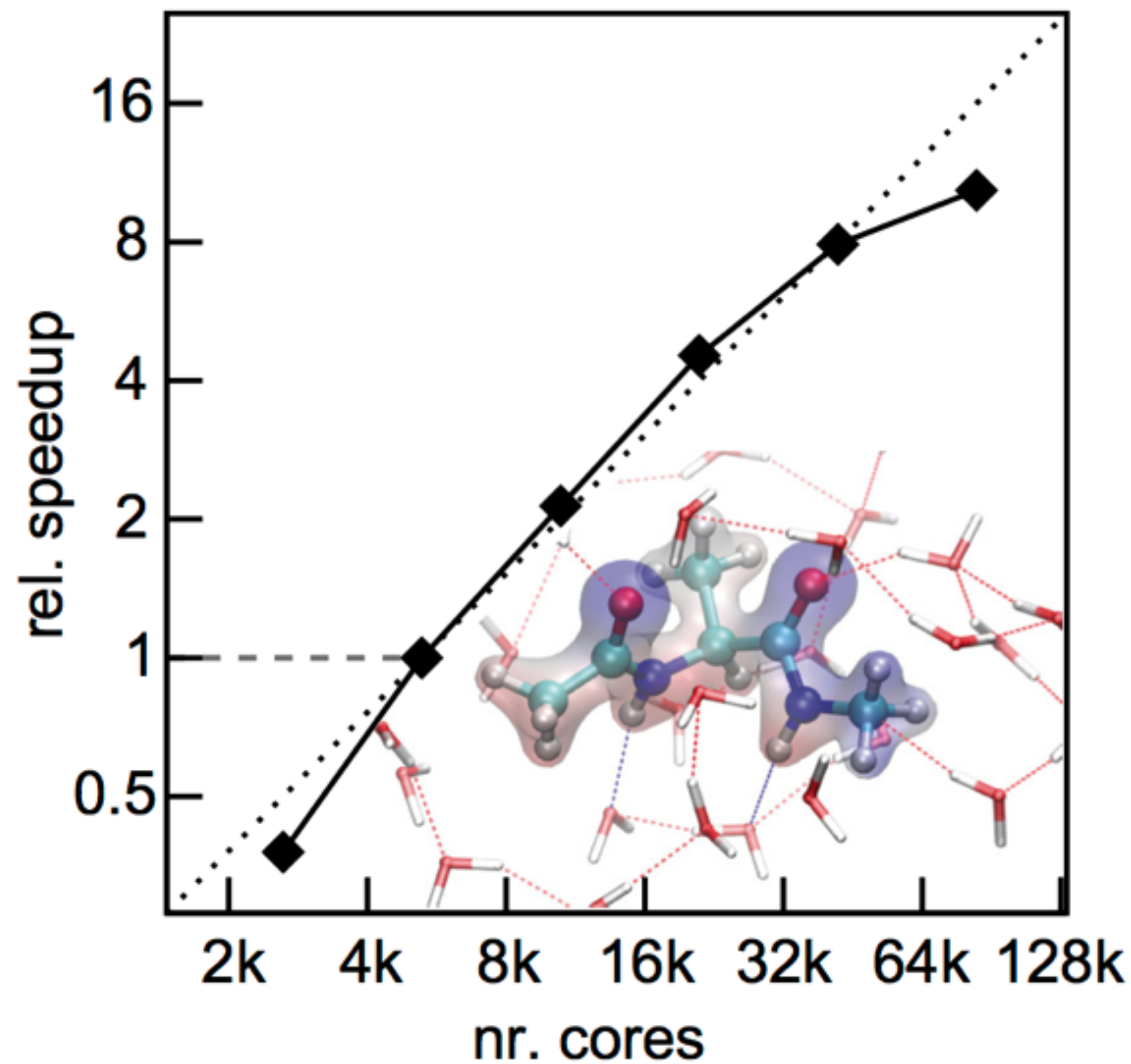
- 10,000^3 grid
- 155 TB memory
- 19 TB output/chckp.
- prep. simulation
- resolve sonic length of astrophysical turbulence

Seven-League Hydro (Uni Würzburg)



- 3D stellar evolution
  - •low Mach number
  - •implicit t-steps (overcome CFL)
  - •reduced art. dissipation
- pure MPI
- MPI+OpenMP
- 2016^3 grid
- 50 TB memory

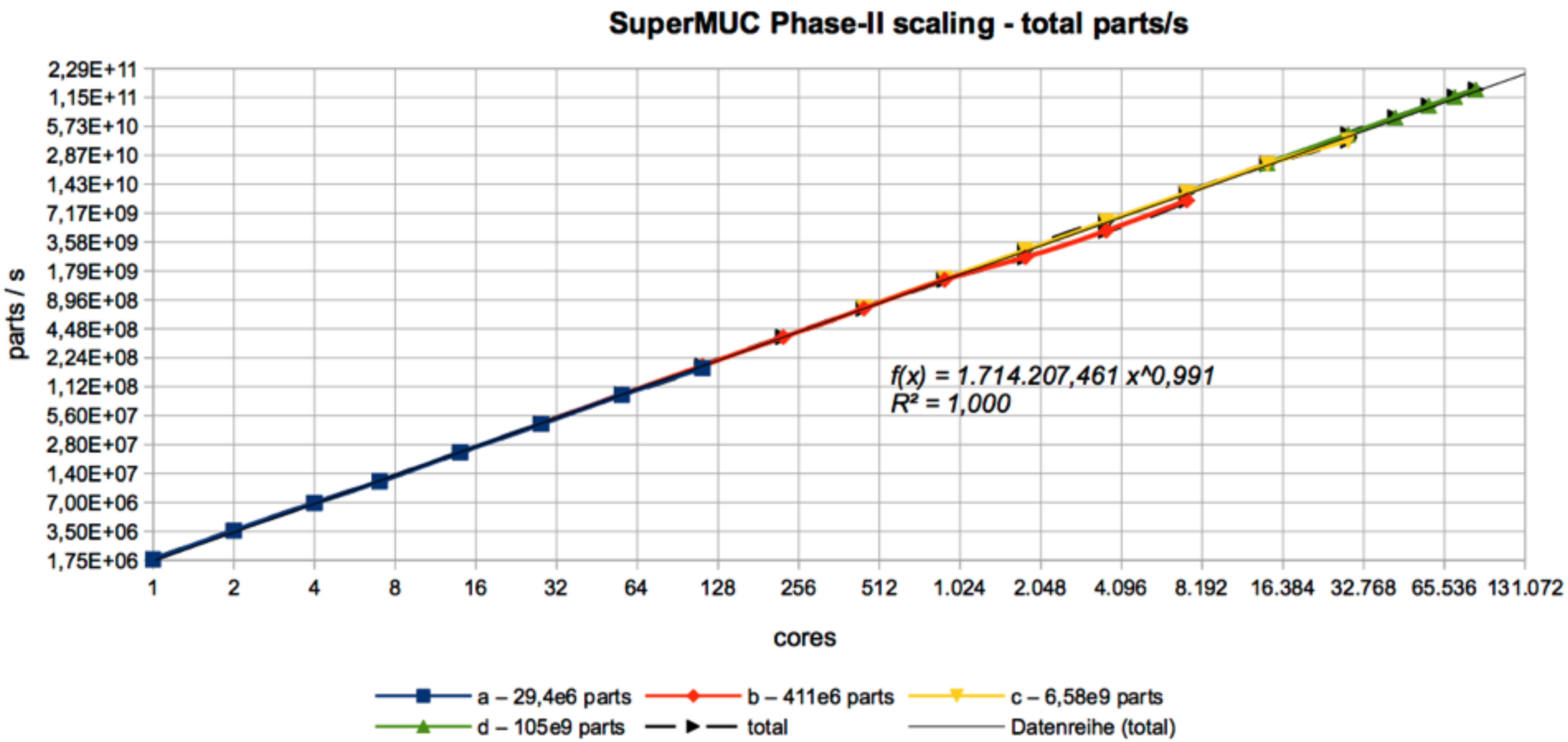
Iphigenie (LMU München)



MD package *IPHIGENIE / CPMD*

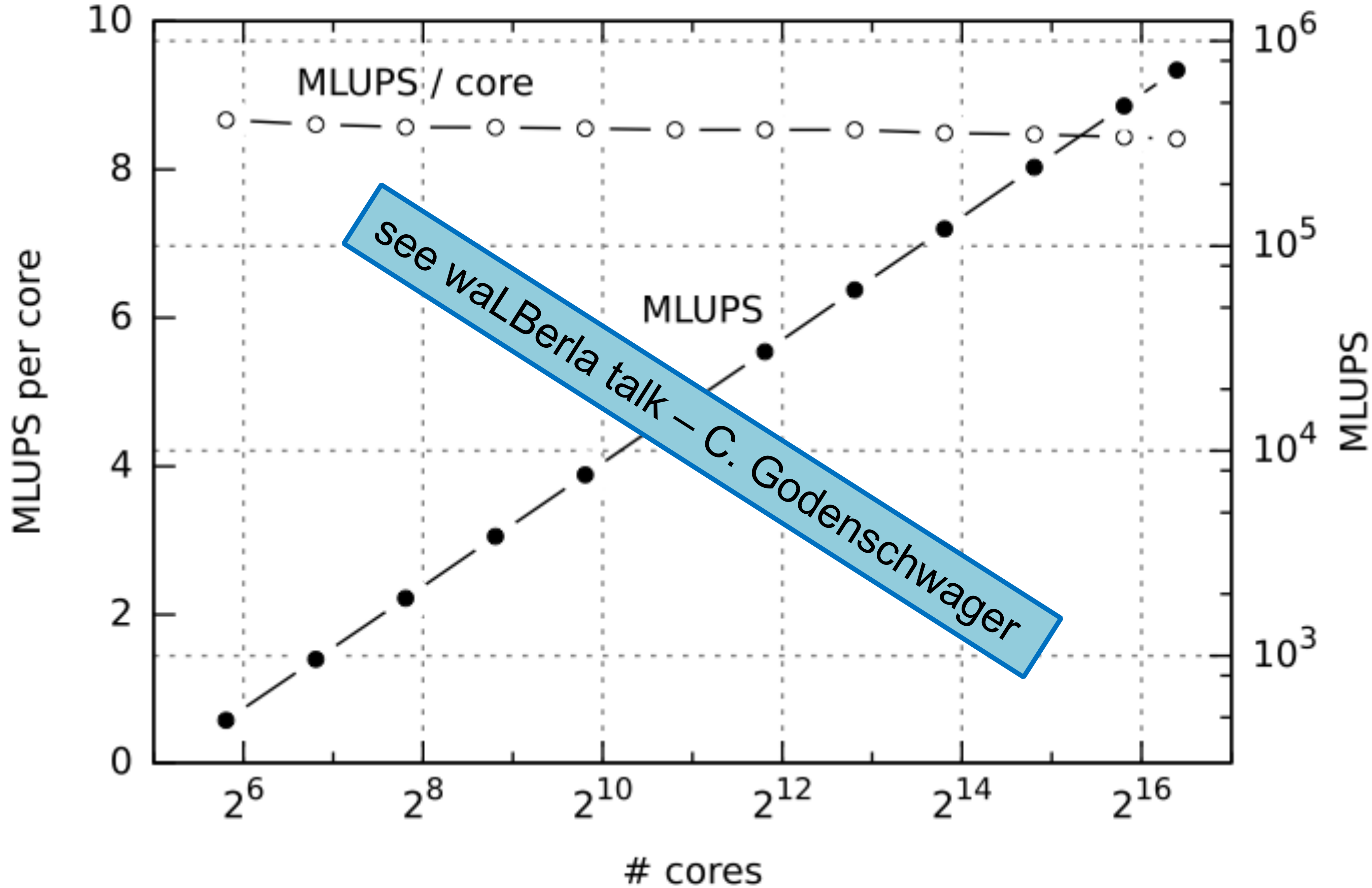
- Coupling molecular mechanics (*IPHIGENIE*) with quantum-mechanical DFT (*CPMD*)
- biomoleculs in native solvent environment, e.g. Alanine Dipeptide (MD) in polarized water(MM)
- Overcome scaling problems of MD
- 4 MPI tasks with 7 OpenMP threads / node
- 128 replicas
- scaling reference @ 5376 cores

PSC (LMU München)

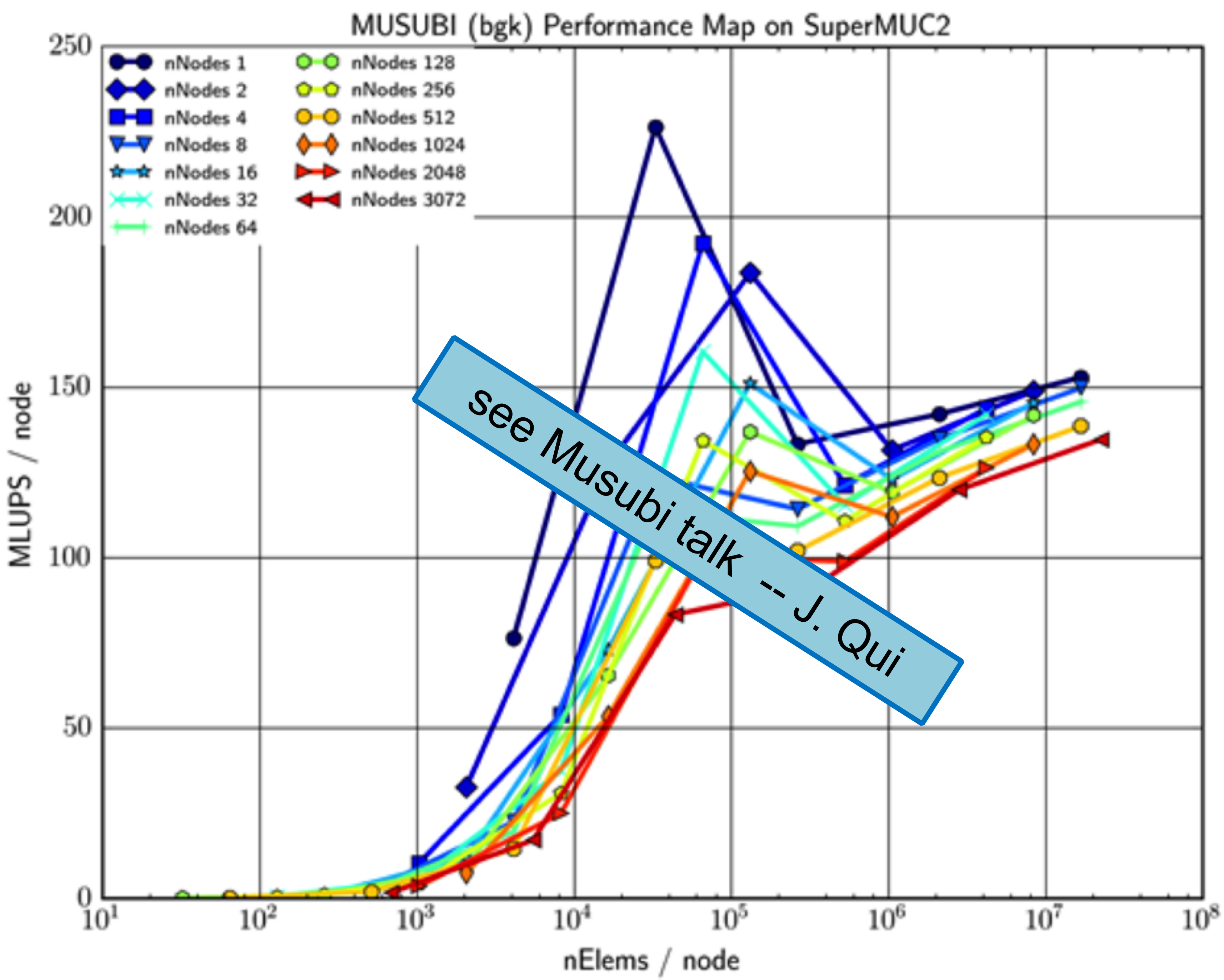


- Plasma Simulation Code (PSC) is a Particle-In-Cell code for solving the ext. Maxwell-Vlasov and Maxwell-Vlasov-Boltzmann equation
- set of strong scaling tests
- improvements of task-local I/O scheme

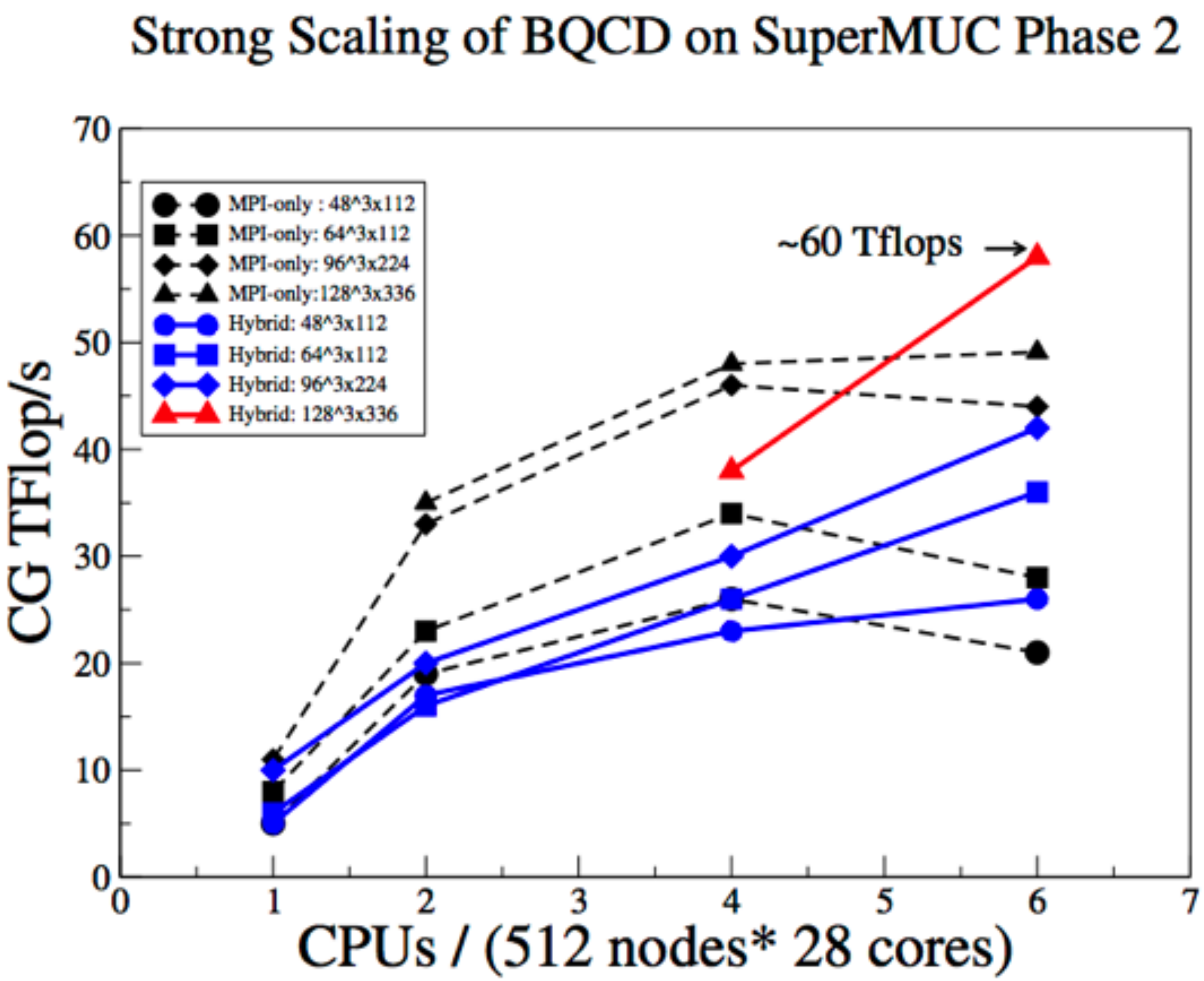
waLBerla (FAU Erlangen)



Musubi (Uni Siegen)



BQCD (LRZ,FU Berlin)



- Berlin Quantum Chromodynamics package (BQCD)
- performance of internal CG solver (95% of execution time)
- setup needs to fit the fabric
- pure MPI: super-linear/linear scaling up to 2/4 islands
- beyond 3 islands hybrid MPI+OpenMP favorable

CIAO (RTWH Aachen)

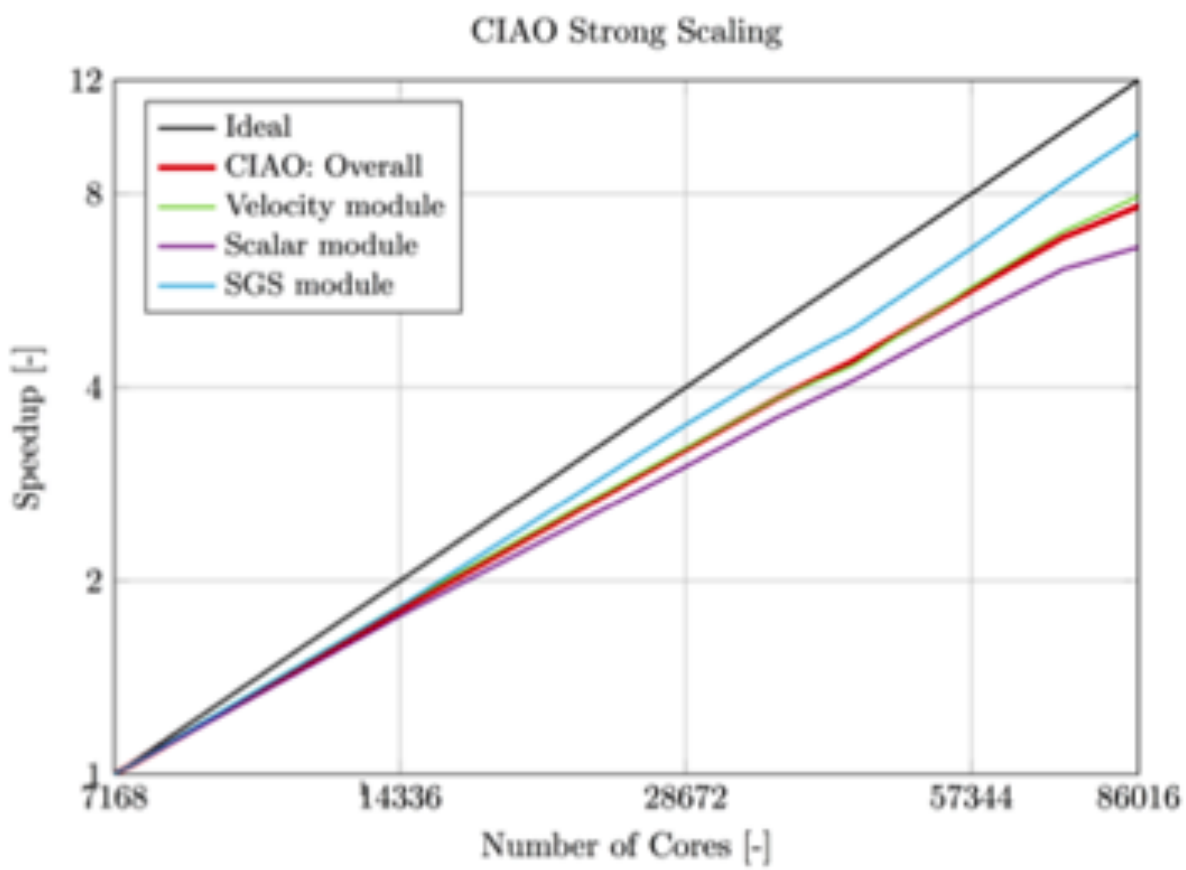
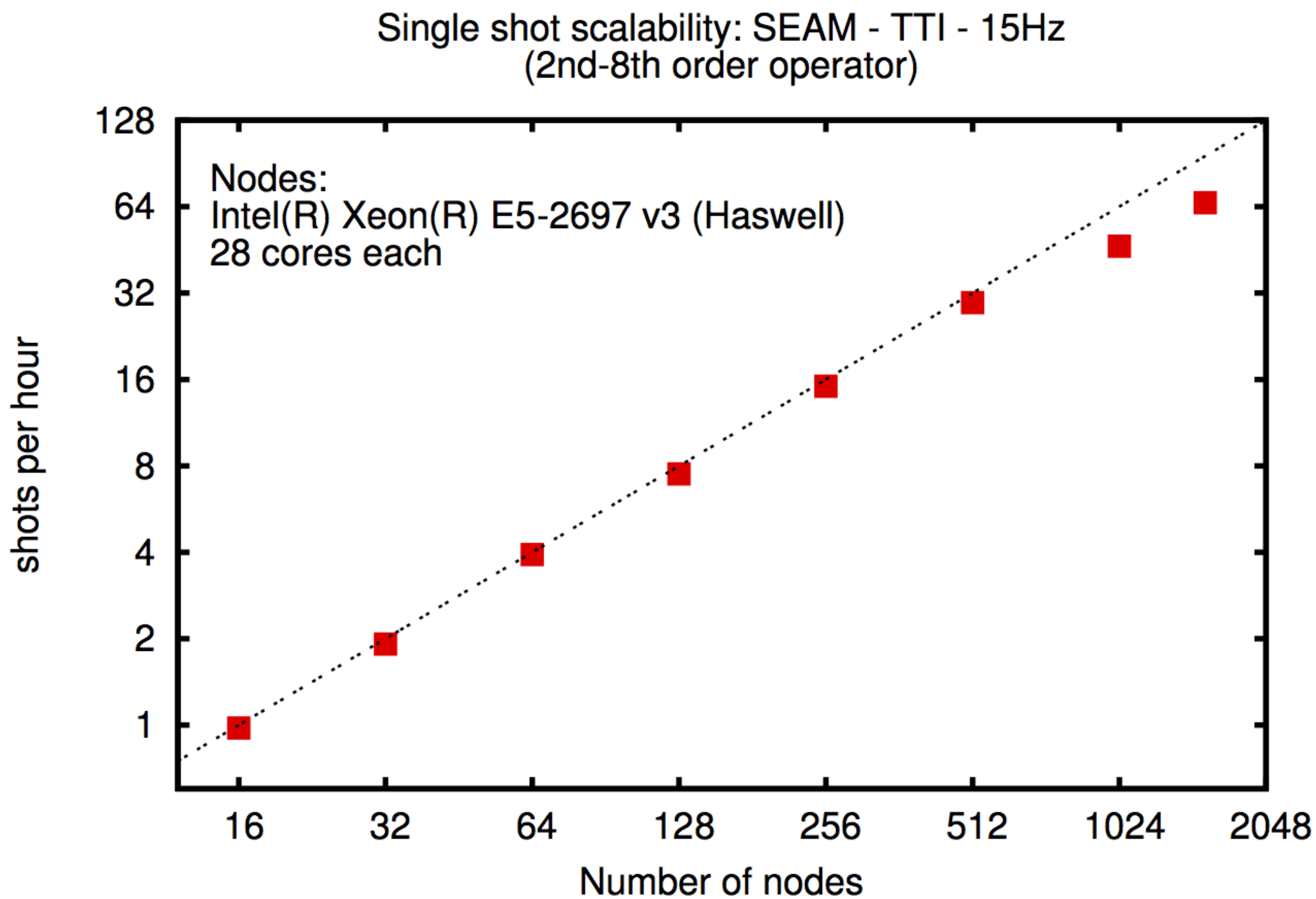


Figure 1: CIAO Strong Scaling

CIAO was completely developed at the Institute for Combustion Technology (RWTH) and Sogang University.

- pure MPI parallelization
- strong scaling: large eddy simulation
- compressible Navier-Stokes solver
- 5 stage expl. Runge Kutta
- high order WENO scheme

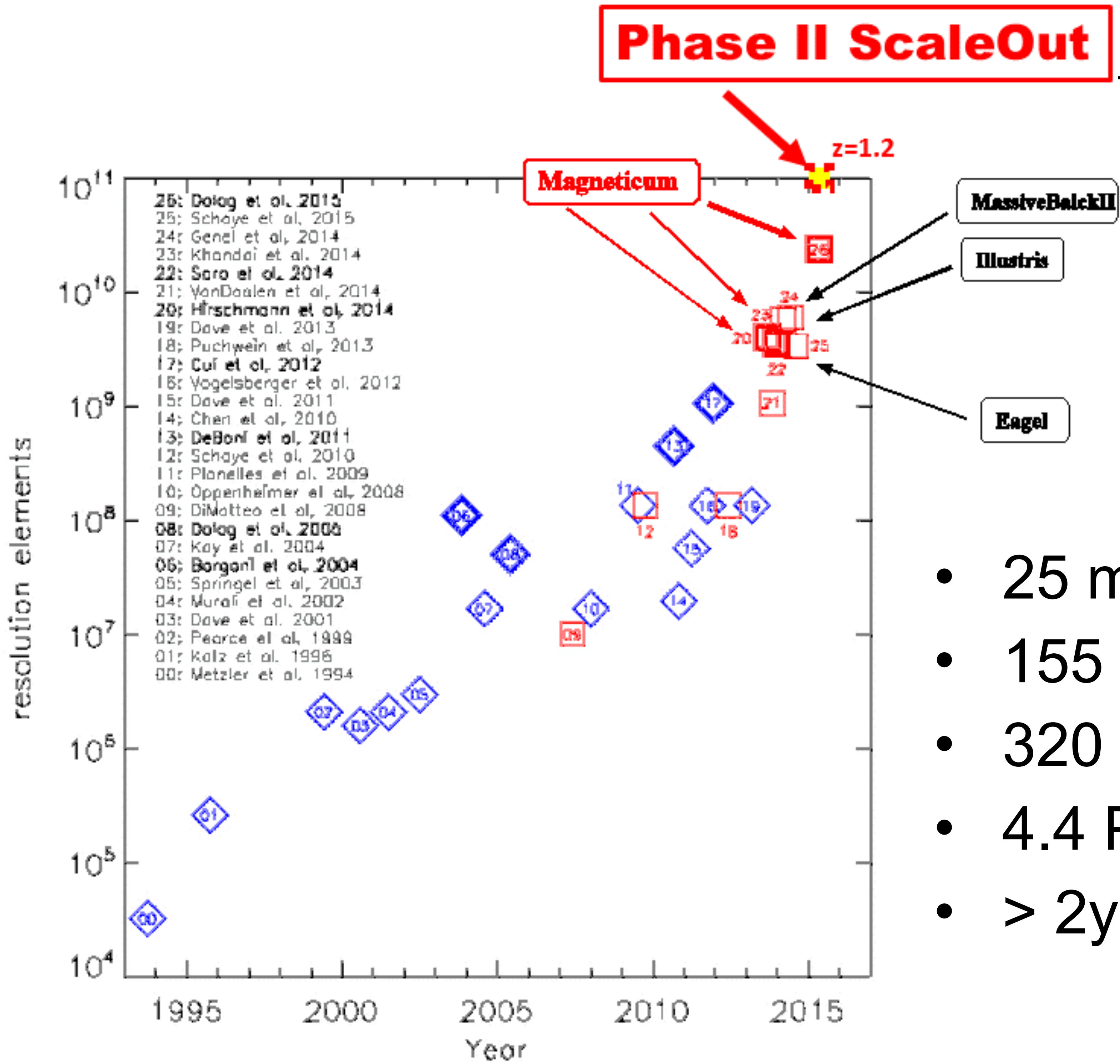
GASPI/GPI-2 (Fraunhofer ITWM)



- GPI 2.0 show case
- Framework for Reverse Time Migration (FRTM)
- seismic imaging, e.g. oil / gas
- one-side, truly async. comm.
- parallel efficiency
  - 94% @ 512 nodes
  - 75% @ 1024 nodes
  - 70% @ 1536 nodes
- ~ 210 Tflop/s

Problem execution time reduced in strong scaling from 16h to 55s

GADGET (USM/C2PAP/LRZ)



***Magneticum* - Box0:**

Largest cosmology simulation of its kind

- 1.8×10^{10} particles
- (2.7Gpc/h)^3
- => ~18% visible universe

- 25 million CPU-h
- 155 TB memory
- 320 TB scientific data
- 4.4 PB written / 2.6 PB read
- > 2yr preparation

Summary & Conclusions

- finding hardware bugs
  - unusual usage pattern
  - only one application a time
  - a final accept. stress test
  - experienced users
- MPI/OpenMP
  - stack size per node gets significant (performance vs. memory)
  - OpenMP/MPI hybrid parallelization is favorable
  - efficient parallel IO gets more and more important
- preparation is everything
  - scaling tests on lower level
  - OpenMP/MPI balance
  - selection of input cases
  - checkpoints / restart files
  - plan I/O strategy (MPI-IO, p-HDF5, p-netCDF)
  - risk management
    - Plan B input case
    - debug test cases
    - debug strategies
    - expect the unexpected ;-)

Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

Thanks for your Attention

Special Thanks

- Special Thanks to the ADMINS at LRZ
- Thanks to IBM operation support
- Thanks to our users for good collaboration

Projects

1. BQCD / Quantum Physics (M. Allalen)
2. SeisSol / Geophysics, Seismics(S. Rettenberger)
3. GPI-2/GASPI / Toolkit for HPC (M. Kühn)
4. Seven-League Hydro / Computational Fluid Dynamics CFD (P. Edelmann)
5. ILBDC / Lattice Boltzmann CFD (M. Wittmann)
6. Iphigenie / Molecular Dynamics (G. Mathias)
7. FLASH / CFD (L. Iapichino and C. Federrath)
8. GADGET / Cosmological Dynamics (K. Dolag)
9. PSC / Plasma Physics (K. Bamberg)
10. waLBerla / Lattice Boltzmann CFD (F. Schornbaum)
11. Musubi / Lattice Boltzmann CFD (A. Nikhil, J. Qi, H. Klimach, S. Roller)
12. Vertex3D / Stellar Astrophysik (T. Melson, A. Marek)
13. LS1-Mardyn / Material Science (N. Tchipev)
14. CIAO / Combustion Simulation (A. Deshmukh, T. Falkenstein)

Paper Abstract

- Preparation is everything
- Finding Heisenbugs is difficult
- MPI reaches at its limits
- hybrid is the way to go
- I/O libraries are more important than ever

The extreme scale-out workshop at LRZ again showed that preparation of a simulation campaign is crucial for the success of the project. This preparation has to address scaling tests, choice of OpenMP/MPI balance, interval for check- point and restart files, good preparation of input files , I/O strategy, and risk management. Under these conditions it was possible to use a brand new system like SuperMUC Phase 2 directly after installation and obtain scientific results from the start.

A big advantage of the extreme scale-out workshop was that only one code was running at a time and this code was filling up the whole system. Thus hardware bugs were much easier to detect and resolve. One especially hard to find bug was a combination of two timeouts and a hardware problem. During normal user operation this error would have been close to impossible to detect because of the low probability of two errors occurring simultaneously for smaller jobs.

MPI is at its limits. The stack size of the MPI stack is growing on each node and for a system of almost 100,000 cores it occupies a significant amount of memory. The startup time can exceed the range of minutes and become a significant part of the overall run time. One way to overcome this bottleneck is the use of hybrid OpenMP/MPI programming models. However, this implies very deep system knowledge on the user side, since process pinning and the choice of the OpenMP/MPI balance has to be evaluated and decided by the user. Furthermore, I/O strategies have to be developed and tested before the complete system can be used. In the future I/O libraries which can mediate this task become more and more important.

Even for hybrid openMP/MPI Set-ups with a single MPI-task per node, problems arise due to internal limit of the MPI send/receive buffer. This limit is caused by the Integer*4 Byte implementation of the MPI index values. Such problems can be overcome by using application internal buffering.