Load Balancing and Auto-Tuning in Particle Simulations

Philipp Neumann

HPC-Status-Konferenz der Gauß-Allianz

Erlangen, October 2018

Universität Hamburg
Philipp Neumann
Thomas Ludwig

Technische Universität München
Hans-Joachim Bungartz
Nikola Tchipev
Steffen Seckler
Fabio Gratl

HLRS/ Universität Stuttgart
José Gracia
Nils Urmersbach
Christoph Niethammer

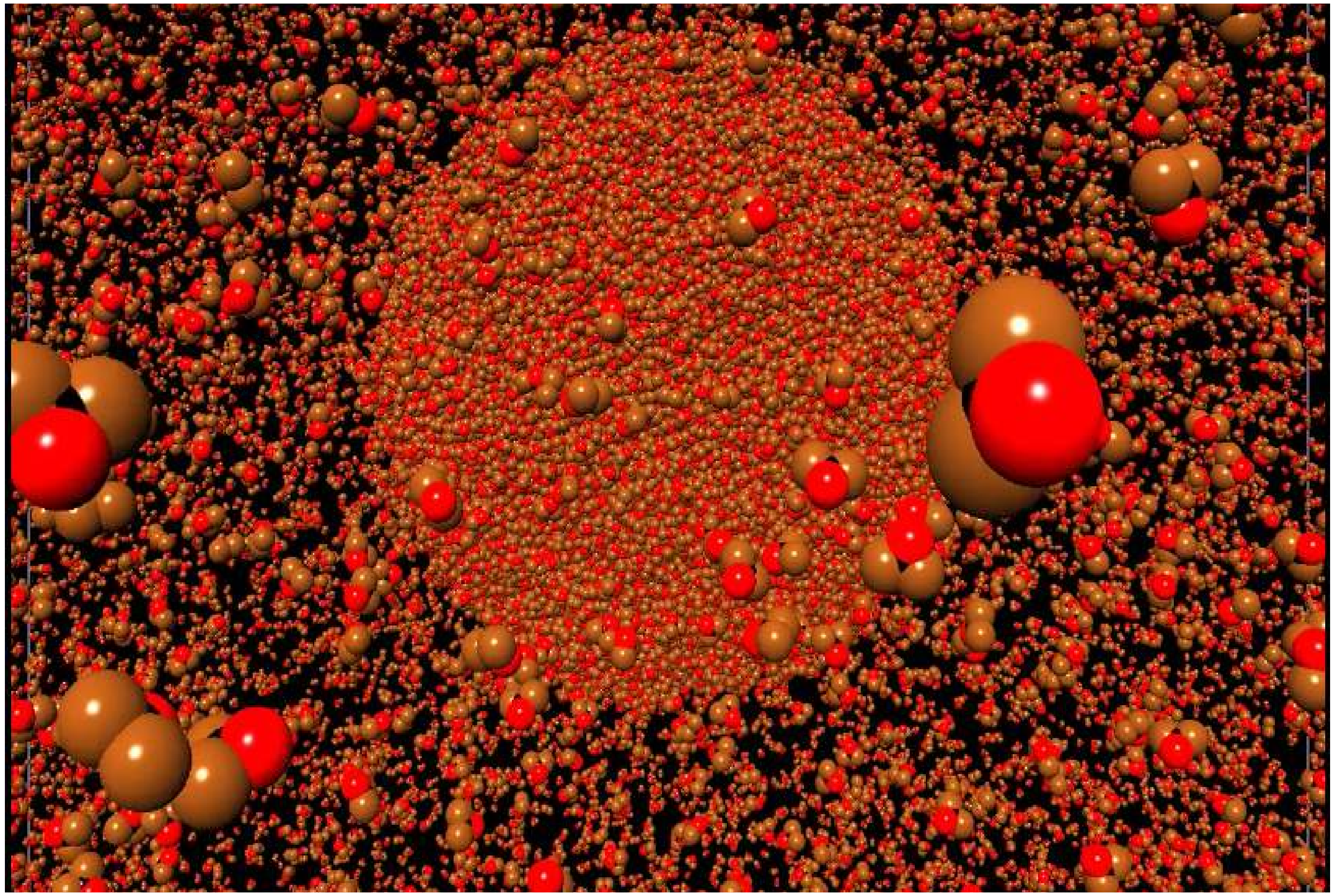
VISUS/ Universität Stuttgart
Guido Reina
Oliver Fernandes

Technische Universität Darmstadt
Felix Wolf
Sergei Shudler
Sebastian Rinke

Technische Universität Berlin
Matthias Heinen
Jadran Vrabec

Technische Universität Kaiserslautern
Hans Hasse
Kai Langenbach
Truong Vo

Sampling in MD Applications



Investigation of various thermodynamic states and properties of fluids
→ vapor-liquid systems, interfacial flows, complex fluids, …
Particularly interesting/challenging (from a computational perspective):
→ Sampling of equations of state, rare-event sampling
→ **The challenge: many inter-dependent MD runs,**
each with different compute requirements
→ Similar problem settings: Uncertainty quantification,
parameter identification, ...

Boosting Performance...

Within Particle Simulation

- **Optimal algorithms** (that is O(N) and O(N $\log^d$ N)), data structures
- Vectorization[1,2,3], Intra-node (OpenMP…)[4,8,9], Inter-node (MPI)[9]

→ tuned optimally and automatically for specific case (auto-tuning)

→ many codes with different foci: Gromacs, LAMMPS, ls1 mardyn, FDPS, NAMD,...

- Load balancing…
  - ...w.r.t. thermodynamics → vapor-liquid...
  - ...w.r.t. hardware → CPU, accelerator, …
  - Methods: triclinic cell adjustment[5], Voronoi tesselation[6], recursive bi-section/k-d trees[7], ..
- Resilience (at extreme scale) → checkpointing, ...

Between Particle Simulations

- Efficient scheduling: Various tools for particular problem settings
- Efficient scheduling requires accurate performance prediction and modeling[10]
- Resilience (at extreme scale) → checkpointing, ...

[1] Hu et al. Comput. Phys. Commun. 211:31-40, 2017
[2] Páll and Hess. Comput. Phys. Commun. 184(12):2641-2650, 2013
[3] Pennycook et al. IEEE ISPDP, pp. 1085-1097, 2013
[4] Tchipev et al. EuroPar 2015 Workshops, pp. 774-785, 2015
[5] Abraham et al. SoftwareX 1-2:19-25, 2015
[6] Fattebert et al. Comput. Phys. Commun. 183(12):2608-2615, 2012
[7] Seckler et al. HiPC 2016 proceedings, pp. 101-110, 2016
[8] Tchipev et al., Submitted, 2018
[9] Neumann et al. HLRS Review Workshop Proceedings, 2018
[10] Shudler, Vrabec, Wolf. Submitted, 2018

Overview

Auto-Tuning: The Auto-Pas Library

Efficient Scheduling: Workflow-Manager

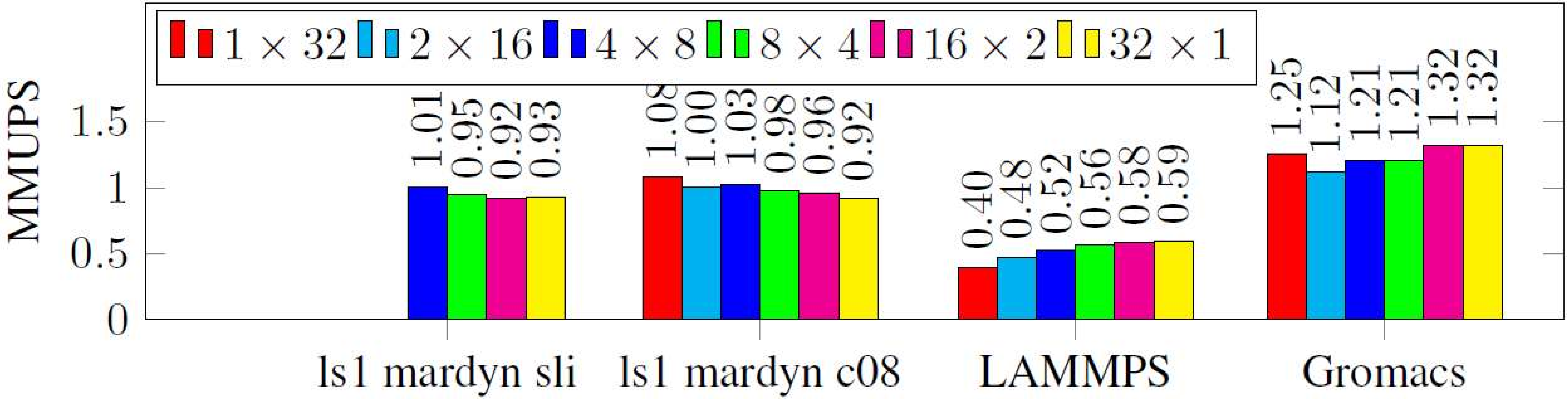
Resilience

Summary and Outlook

Towards Auto-Pas: Traversal Algorithmics

- Data structures: AoS, SoA
- Traversals:
  - Direct
  - **Linked cells**
  - **Verlet lists**
- OpenMP schemes:
  - c08
  - slicing
  - investigated: quicksched, c04

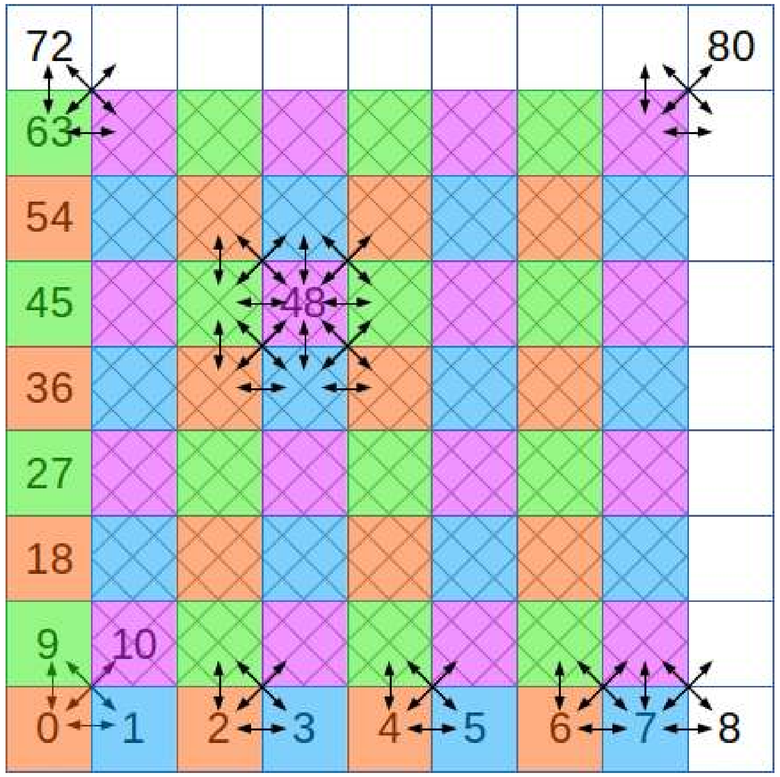
ls1 mardyn (Linked cells)
vs. LAMMPS (Verlet lists)
vs. Gromacs (Verlet lists)

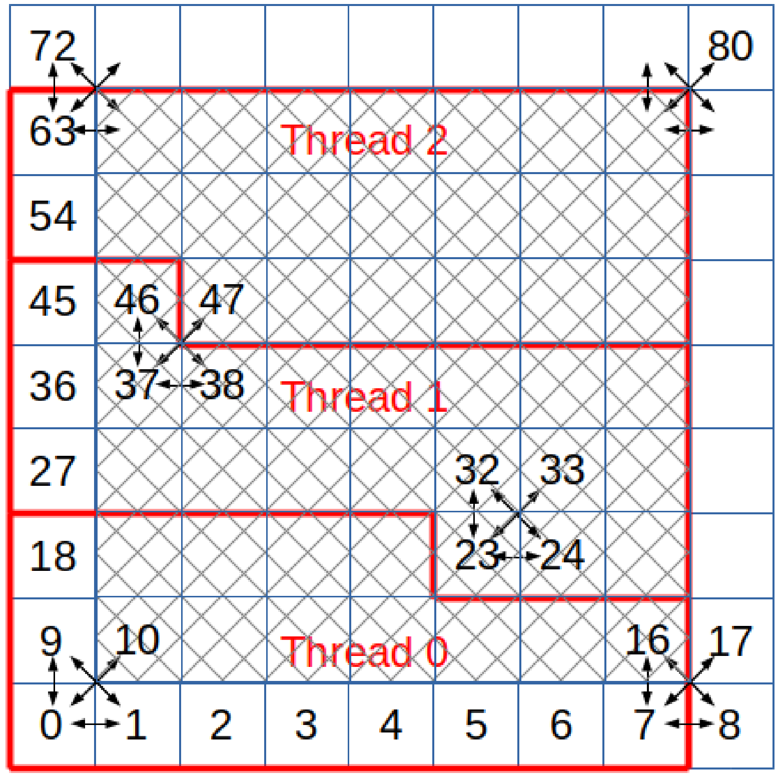


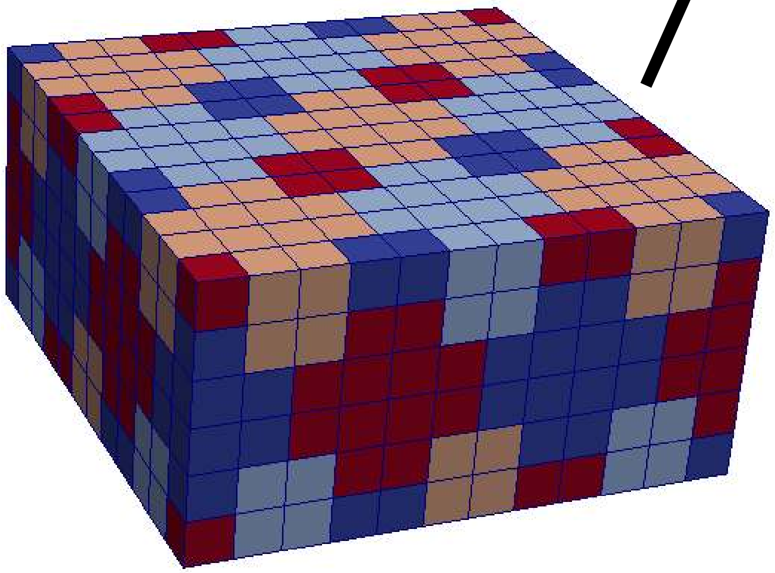
Tchipev, Seckler, …, Bungartz, Neumann. TweTriS: Twenty Trillion-atom Simulation. Submitted, 2018

Towards Auto-Pas: OpenMP

- Data structures: AoS, SoA
- Traversals:
  - Direct
  - Linked cells
  - Verlet lists
- OpenMP schemes:
  - **c08**
  - **slicing**
  - investigated: quicksched, **c04**

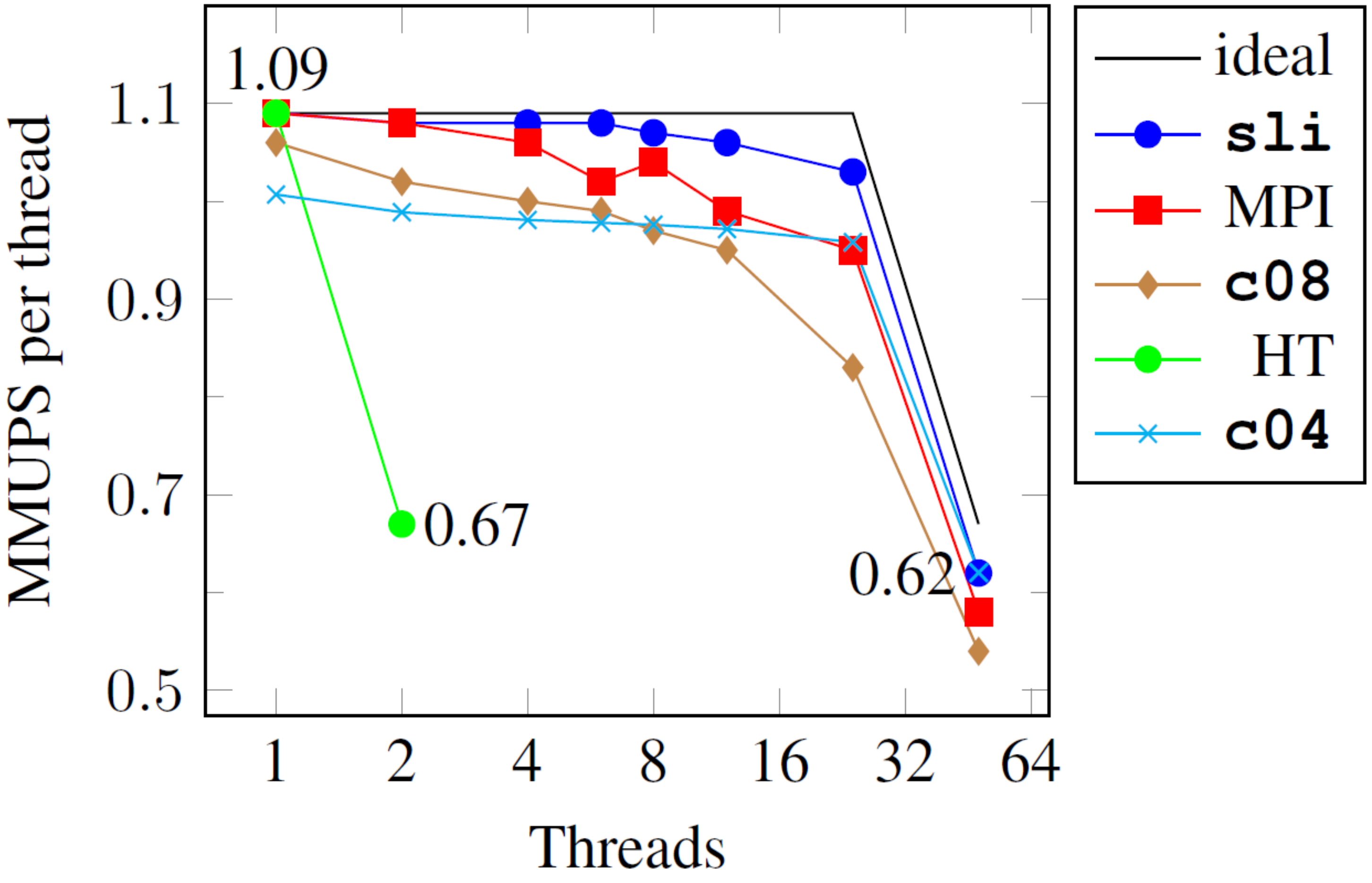






Tchipev, Seckler, …, Bungartz, Neumann. TweTriS: Twenty Trillion-atom Simulation. Submitted, 2018

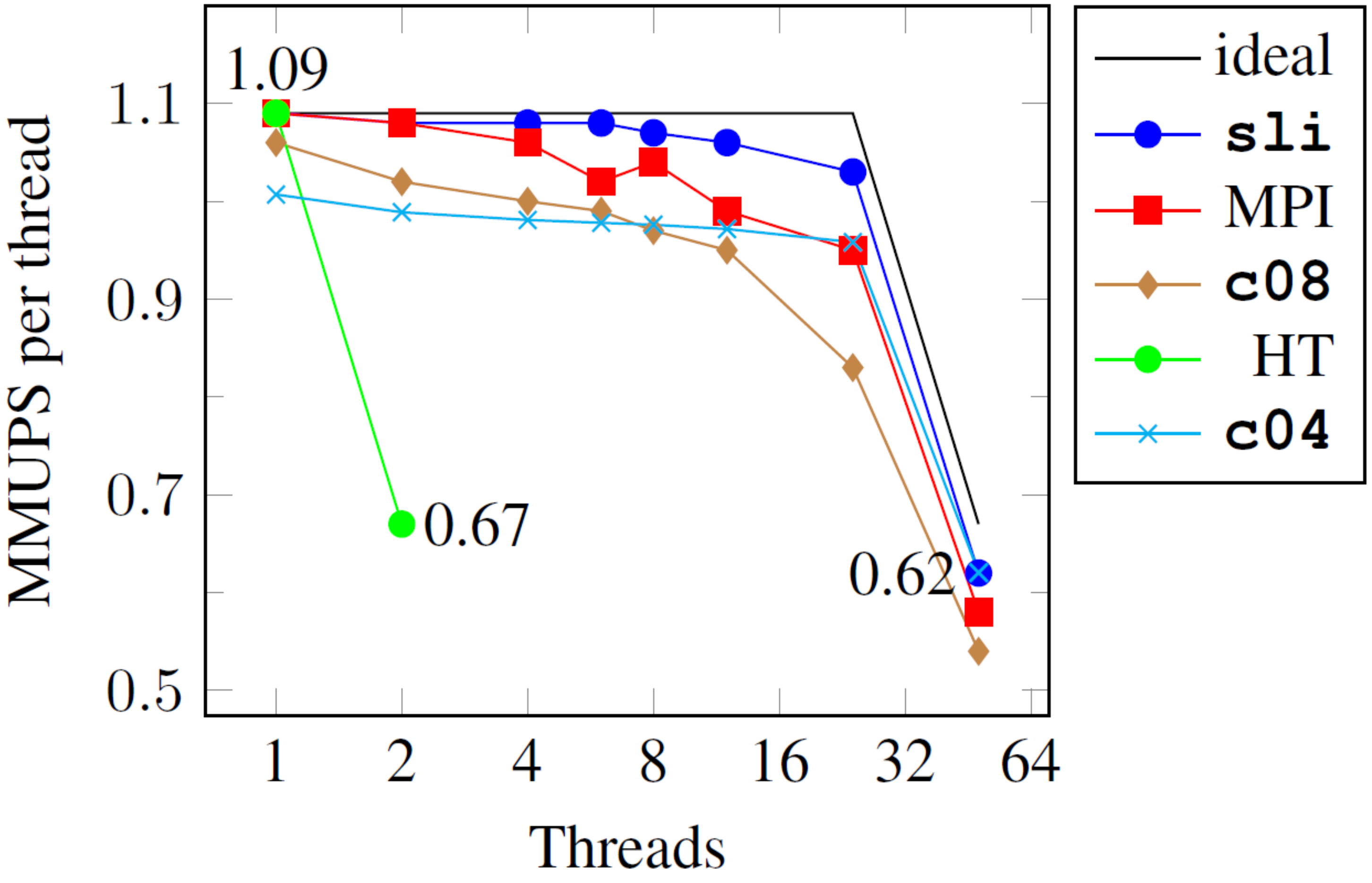
OpenMP Schemes: Performance

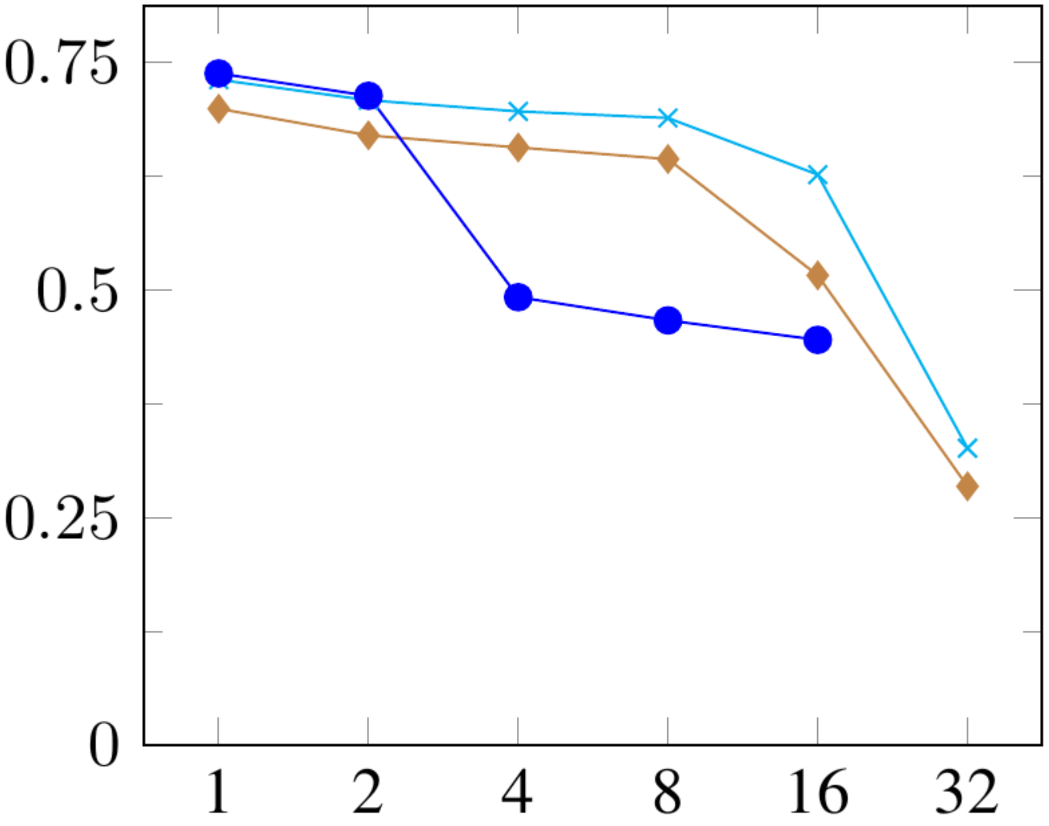


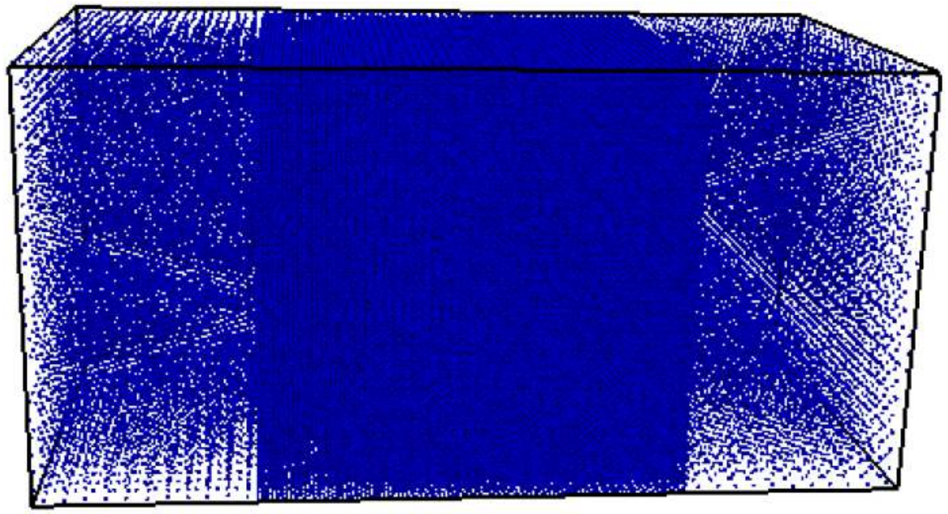
- OpenMP parallelization, exploiting linked cell structure
- Schemes: c08, c04, sli
- Good scalability observed for all schemes
- Overall performance dependent on actual scenario

Tchipev, Seckler, …, Bungartz, Neumann. TweTriS: Twenty Trillion-atom Simulation. Submitted, 2018

OpenMP Schemes: Performance



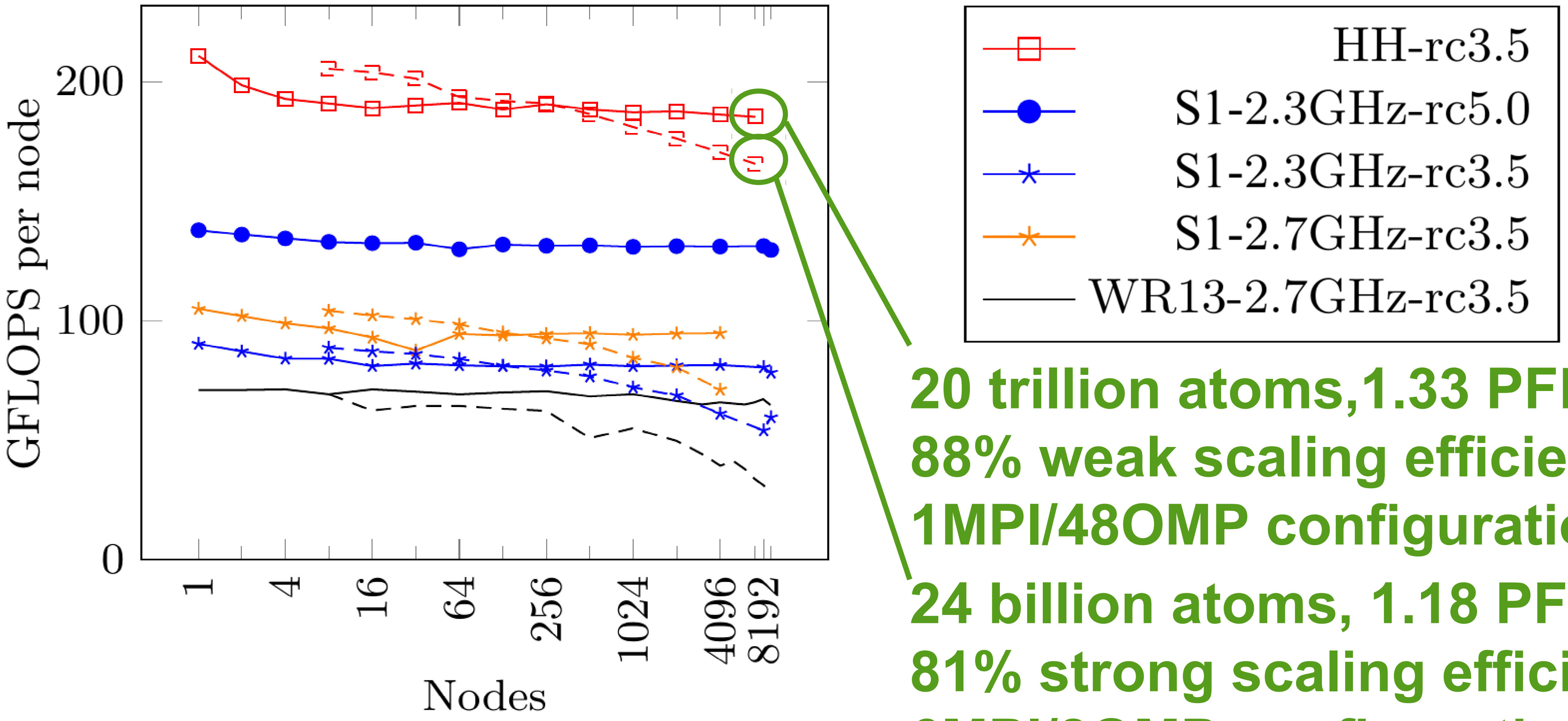




- OpenMP parallelization, exploiting linked cell structure
- Schemes: c08, c04, sli
- Good scalability observed for all schemes
- **Overall performance dependent on actual scenario**

Tchipev, Seckler, …, Bungartz, Neumann. TweTriS: Twenty Trillion-atom Simulation. Submitted, 2018

From Node-Level to Multi-Node



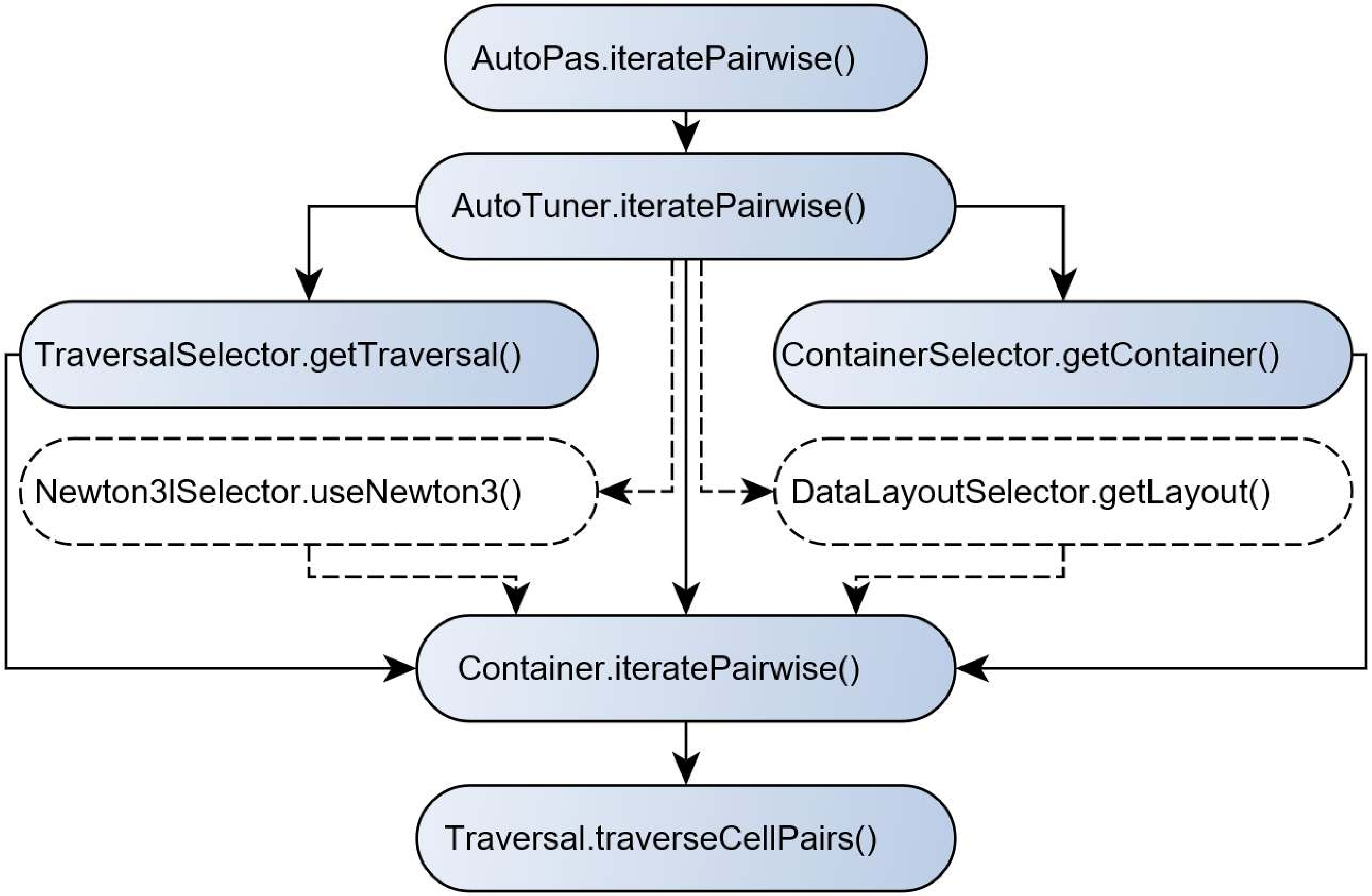
20 trillion atoms,1.33 PFLOPS, 88% weak scaling efficiency 1MPI/48OMP configuration

24 billion atoms, 1.18 PFLOPS, 81% strong scaling efficiency 6MPI/8OMP configuration

- MPI parallelization via domain decomposition
- Nonblocking MPI-3 collectives
- Scalability studies on SuperMUC (LRZ) and Hazel Hen (HLRS)

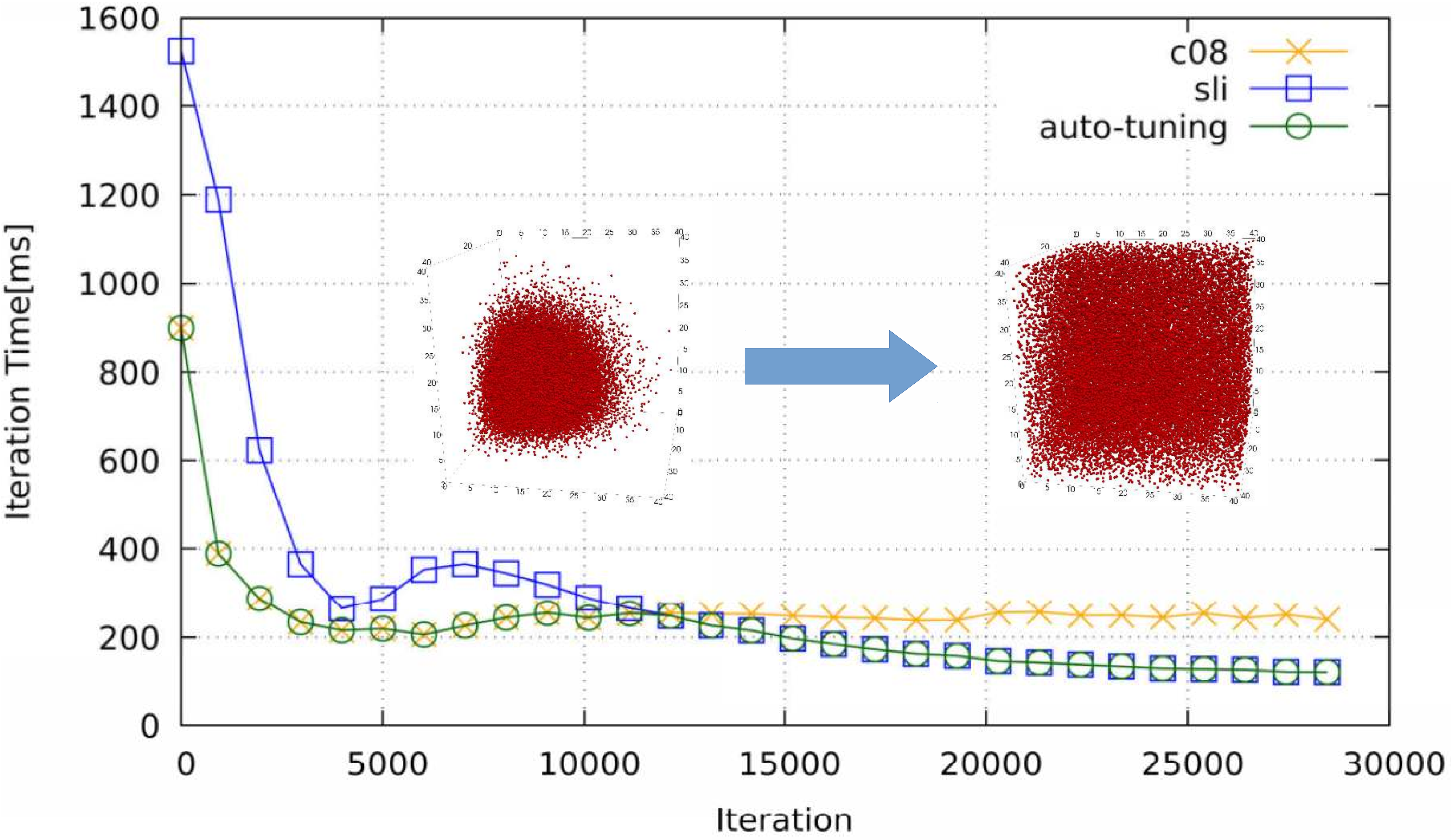
Tchipev, Seckler, …, Bungartz, Neumann. TweTriS: Twenty Trillion-atom Simulation. Submitted, 2018

Auto-Pas: Concept


AutoPas.iteratePairwise()
AutoTuner.iteratePairwise()
TraversalSelector.getTraversal()
ContainerSelector.getContainer()
Newton3lSelector.useNewton3()
DataLayoutSelector.getLayout()
Container.iteratePairwise()
Traversal.traverseCellPairs()

- User defines: particle class + pairwise force functors
- Auto-Pas provides: data structures, traversals, OpenMP schemes

Auto-Pas: Example



- Drastic load imbalance changes over execution time
- Periodic evaluation of runtime
  → Selection of fastest OpenMP scheme at runtime

Overview

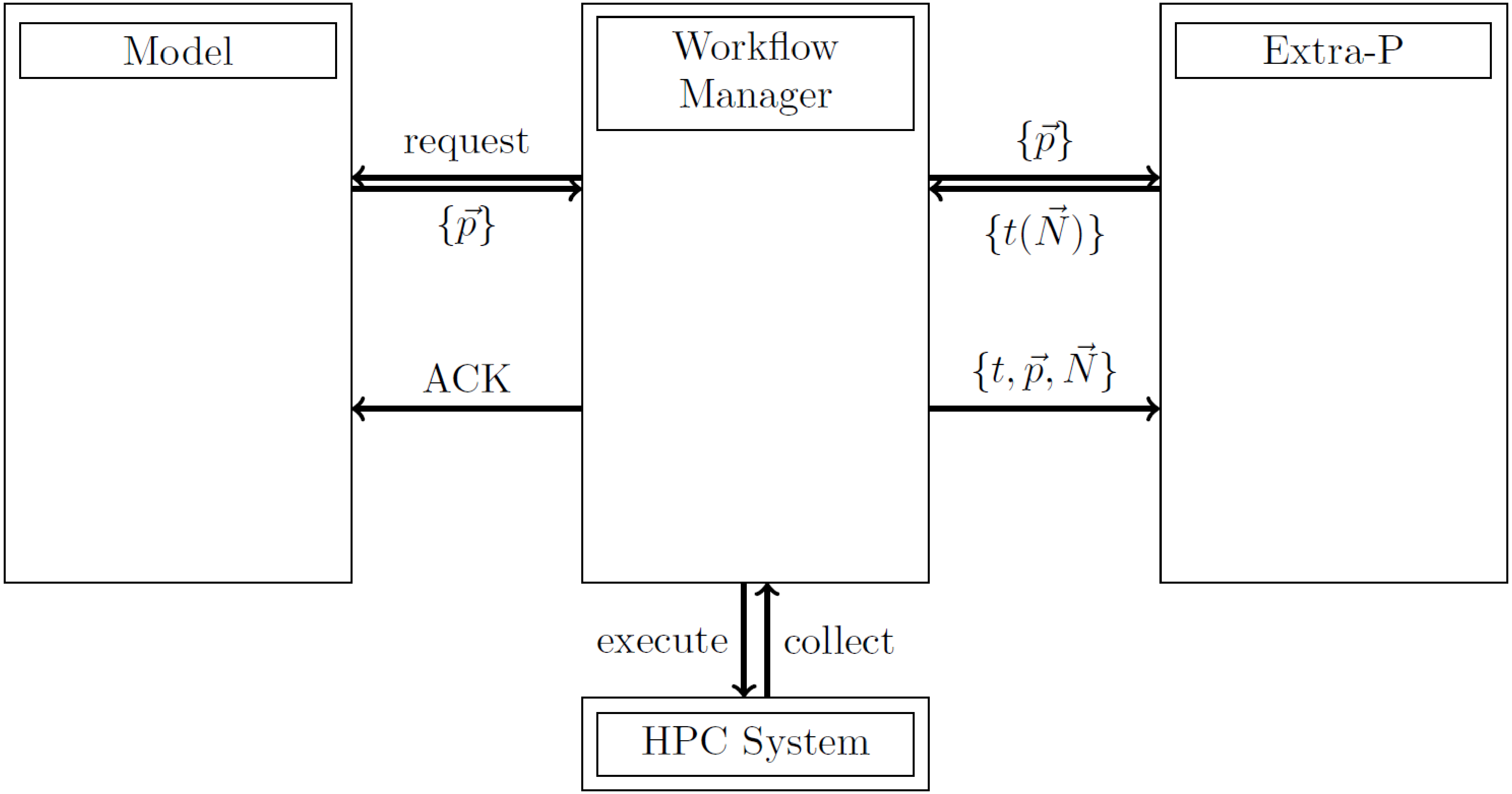
Auto-Tuning: The Auto-Pas Library

Efficient Scheduling: Workflow-Manager

Resilience

Summary and Outlook

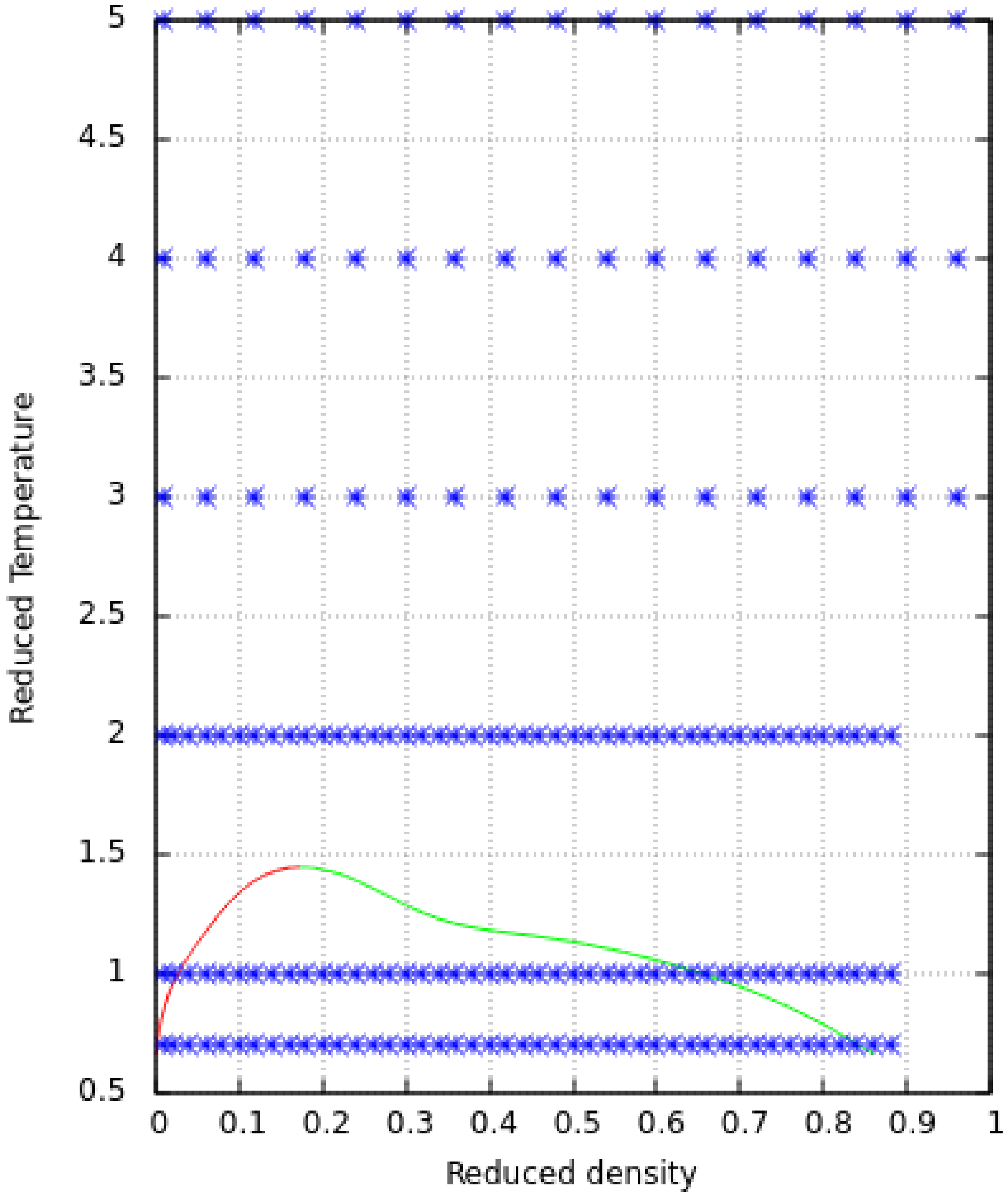
Workflow-Manager



- Python-based implementation
- API
  - get_task() → delivers new task or “end”, if no tasks are available
  - deploy(task,N,mpi) → prepares task for execution based on MPI parameters
  - record_result(task) → called after task is finished

Workflow-Manager in Action: Equation of state (EOS) Fitting and Vapor-Liquid-Equilibrium (VLE) Envelope

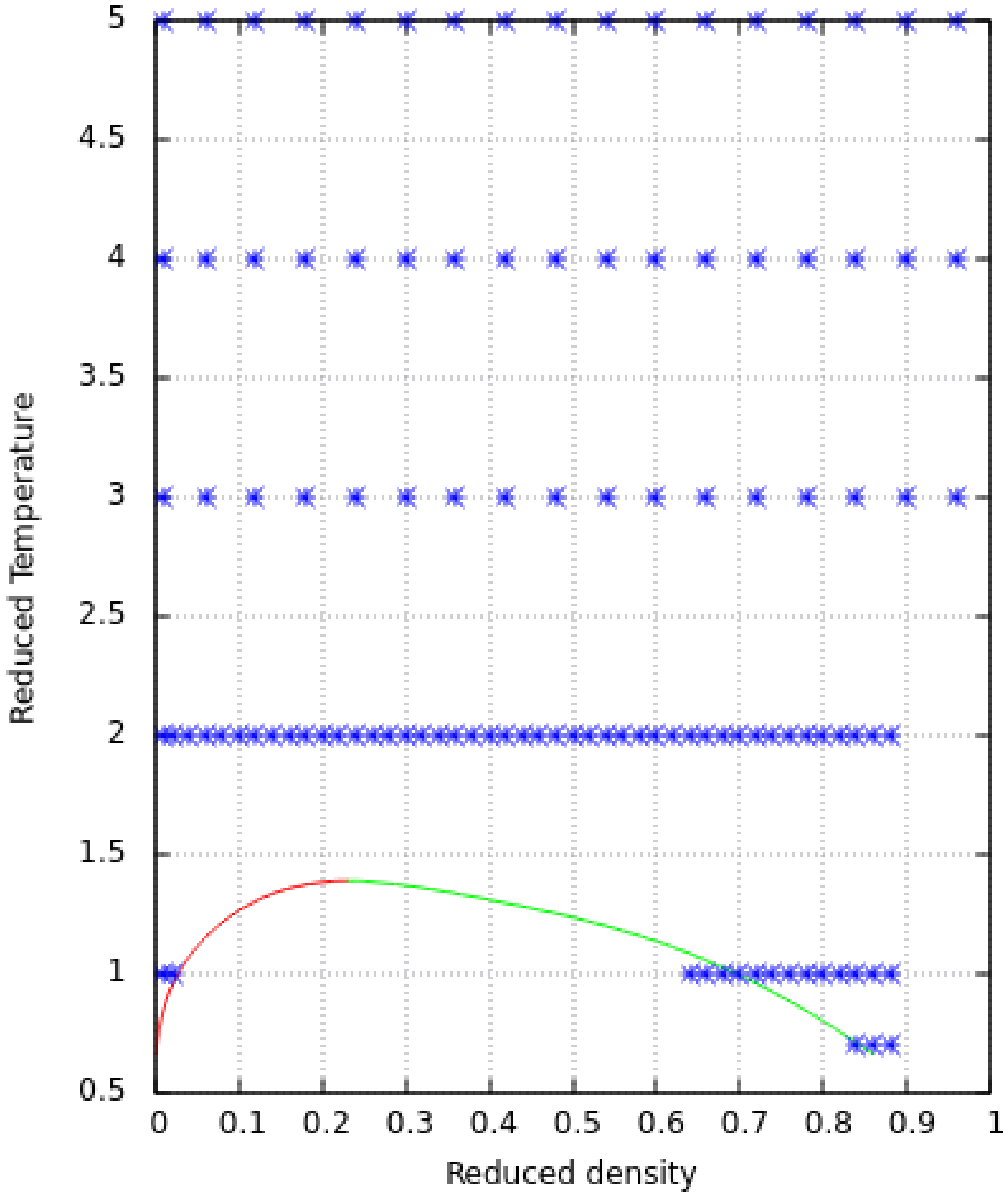
1. **Select two or three high temperature isotherms.**
   Fit the EoS to the simulation data along these isotherms.
   Calculate a preliminary VLE envelope from it.
2. Remove state points if they are in the vapor--liquid two phase region.
   Add more state points that are in the homogeneous fluid region.
3. Fit the EoS to the extended set of state points, and calculate an updated VLE envelope.
4. Repeat steps (1) to (3) until the VLE envelope does not change significantly any more.



Rutkai, Vrabec. J Chem. Eng. Data 60(10):2895-2905, 2015

Workflow-Manager in Action: Equation of state (EOS) Fitting and Vapor-Liquid-Equilibrium (VLE) Envelope

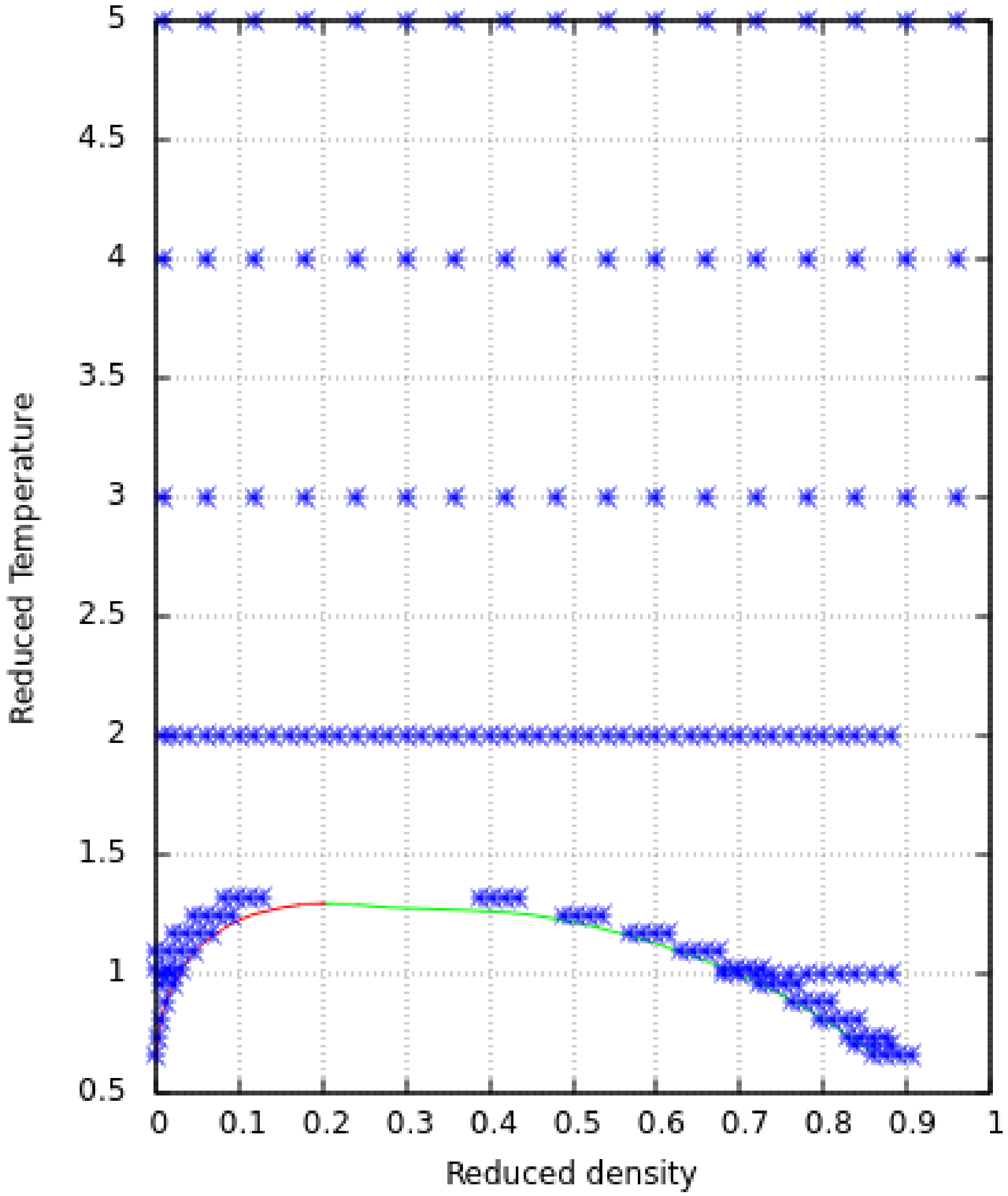
1. Select two or three high temperature isotherms.
   Fit the EoS to the simulation data along these isotherms.
   Calculate a preliminary VLE envelope from it.
2. **Remove state points if they are in the vapor--liquid two phase region.**
   Add more state points that are in the homogeneous fluid region.
3. Fit the EoS to the extended set of state points, and calculate an updated VLE envelope.
4. Repeat steps (1) to (3) until the VLE envelope does not change significantly any more.



Rutkai, Vrabec. J Chem. Eng. Data 60(10):2895-2905, 2015

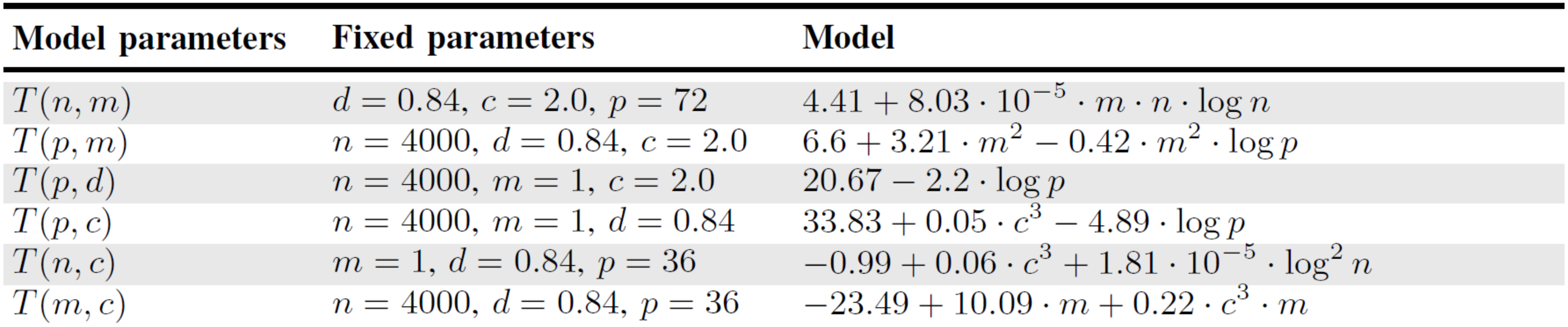
Workflow-Manager in Action: Equation of state (EOS) Fitting and Vapor-Liquid-Equilibrium (VLE) Envelope

1. Select two or three high temperature isotherms.
   Fit the EoS to the simulation data along these isotherms.
   Calculate a preliminary VLE envelope from it.
2. Remove state points if they are in the vapor--liquid two phase region.
   Add more state points that are in the homogeneous fluid region.
3. **Fit the EoS to the extended set of state points, and calculate an updated VLE envelope.**
4. Repeat steps (1) to (3) until the VLE envelope does not change significantly any more.



Rutkai, Vrabec. J Chem. Eng. Data 60(10):2895-2905, 2015

Feeding the Manager with Performance Data: Performance Modeling for Molecular Dynamics



Model parameters	Fixed parameters	Model
$T(n,m)$	$d=0.84,\ c=2.0,\ p=72$	$4.41+8.03\cdot 10^{-5}\cdot m\cdot n\cdot\log n$
$T(p,m)$	$n=4000,\ d=0.84,\ c=2.0$	$6.6+3.21\cdot m^2-0.42\cdot m^2\cdot\log p$
$T(p,d)$	$n=4000,\ m=1,\ c=2.0$	$20.67-2.2\cdot\log p$
$T(p,c)$	$n=4000,\ m=1,\ d=0.84$	$33.83+0.05\cdot c^3-4.89\cdot\log p$
$T(n,c)$	$m=1,\ d=0.84,\ p=36$	$-0.99+0.06\cdot c^3+1.81\cdot 10^{-5}\cdot\log^2 n$
$T(m,c)$	$n=4000,\ d=0.84,\ p=36$	$-23.49+10.09\cdot m+0.22\cdot c^3\cdot m$

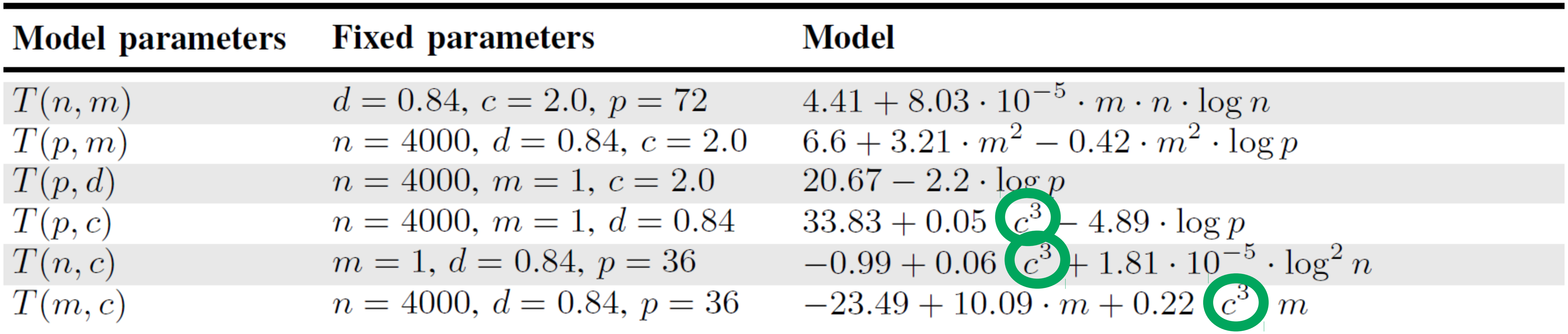
- Extra-P to model performance in ms2
- Multi-parameter model, exploiting performance model normal form

$$f(r_1, r_2, \ldots, r_q) = \sum_{k=1}^{n} c_k \cdot \prod_{l=1}^{q} r_l^{i_{k_l}} \cdot \log^{j_{k_l}}(r_l)$$

- Considered quantities:
  - number of molecules (n)
  - number of interaction sites (m)
  - density (d)
  - cut-off radius (c)
  - number of MPI processes (p)

Shudler, Vrabec, Wolf. Submitted 2018

Feeding the Manager with Performance Data: Performance Modeling for Molecular Dynamics



Model parameters	Fixed parameters	Model
$T(n,m)$	$d = 0.84,\ c = 2.0,\ p = 72$	$4.41 + 8.03 \cdot 10^{-5} \cdot m \cdot n \cdot \log n$
$T(p,m)$	$n = 4000,\ d = 0.84,\ c = 2.0$	$6.6 + 3.21 \cdot m^2 - 0.42 \cdot m^2 \cdot \log p$
$T(p,d)$	$n = 4000,\ m = 1,\ c = 2.0$	$20.67 - 2.2 \cdot \log p$
$T(p,c)$	$n = 4000,\ m = 1,\ d = 0.84$	$33.83 + 0.05\ c^3 - 4.89 \cdot \log p$
$T(n,c)$	$m = 1,\ d = 0.84,\ p = 36$	$-0.99 + 0.06\ c^3 + 1.81 \cdot 10^{-5} \cdot \log^2 n$
$T(m,c)$	$n = 4000,\ d = 0.84,\ p = 36$	$-23.49 + 10.09 \cdot m + 0.22\ c^3\ m$

cubic dependence on cut-off radius

- Extra-P to model performance in ms2
- Multi-parameter model, exploiting performance model normal form

$$f(r_1, r_2, \ldots, r_q) = \sum_{k=1}^{n} c_k \cdot \prod_{l=1}^{q} r_l^{i_{k_l}} \cdot \log^{j_{k_l}}(r_l)$$

- Considered quantities:
  - number of molecules (n)
  - number of interaction sites (m)
  - density (d)
  - cut-off radius (c)
  - number of MPI processes (p)

Shudler, Vrabec, Wolf. Submitted 2018

Overview

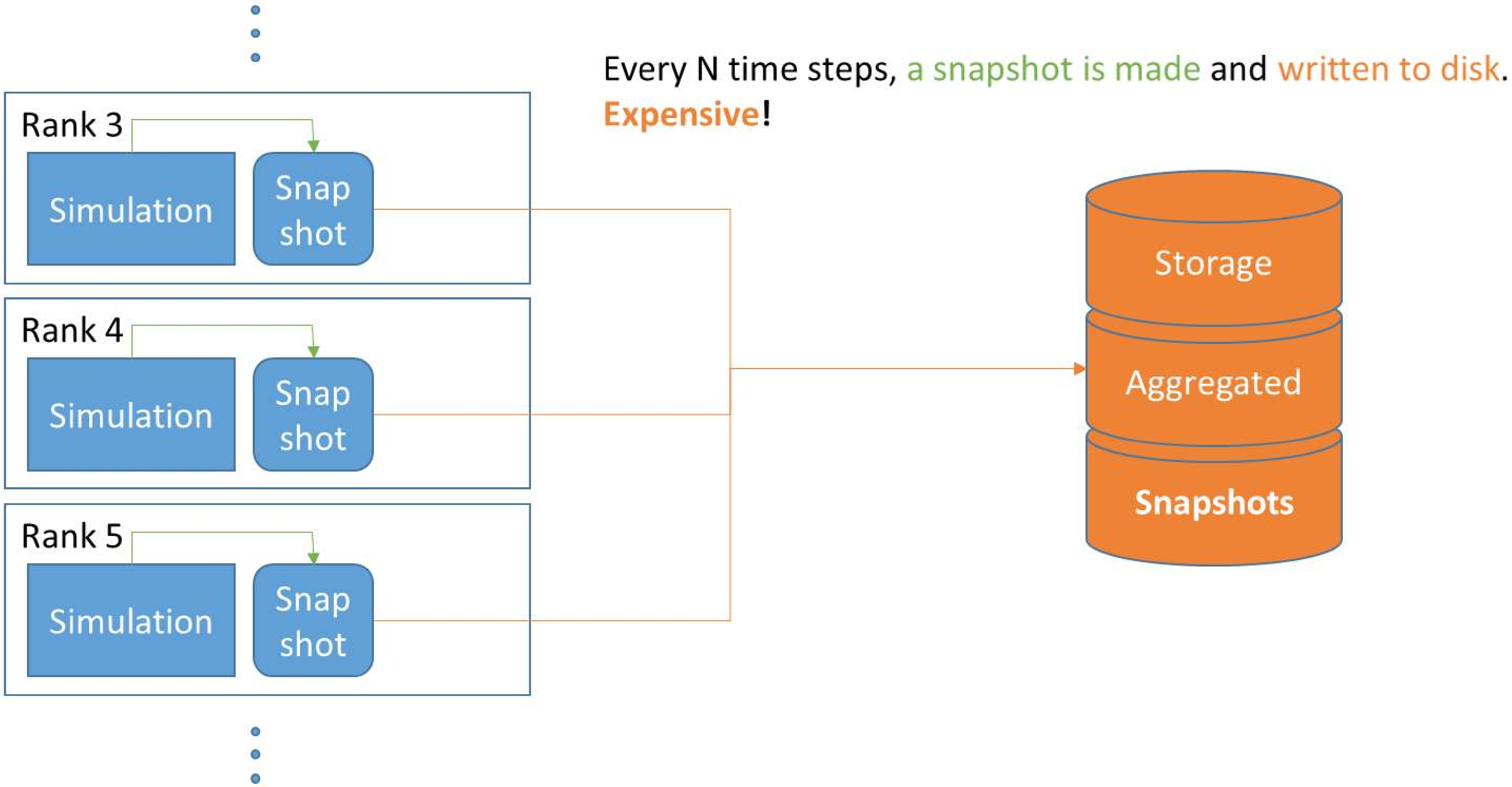
Auto-Tuning: The Auto-Pas Library

Efficient Scheduling: Workflow-Manager

Resilience

Summary and Outlook

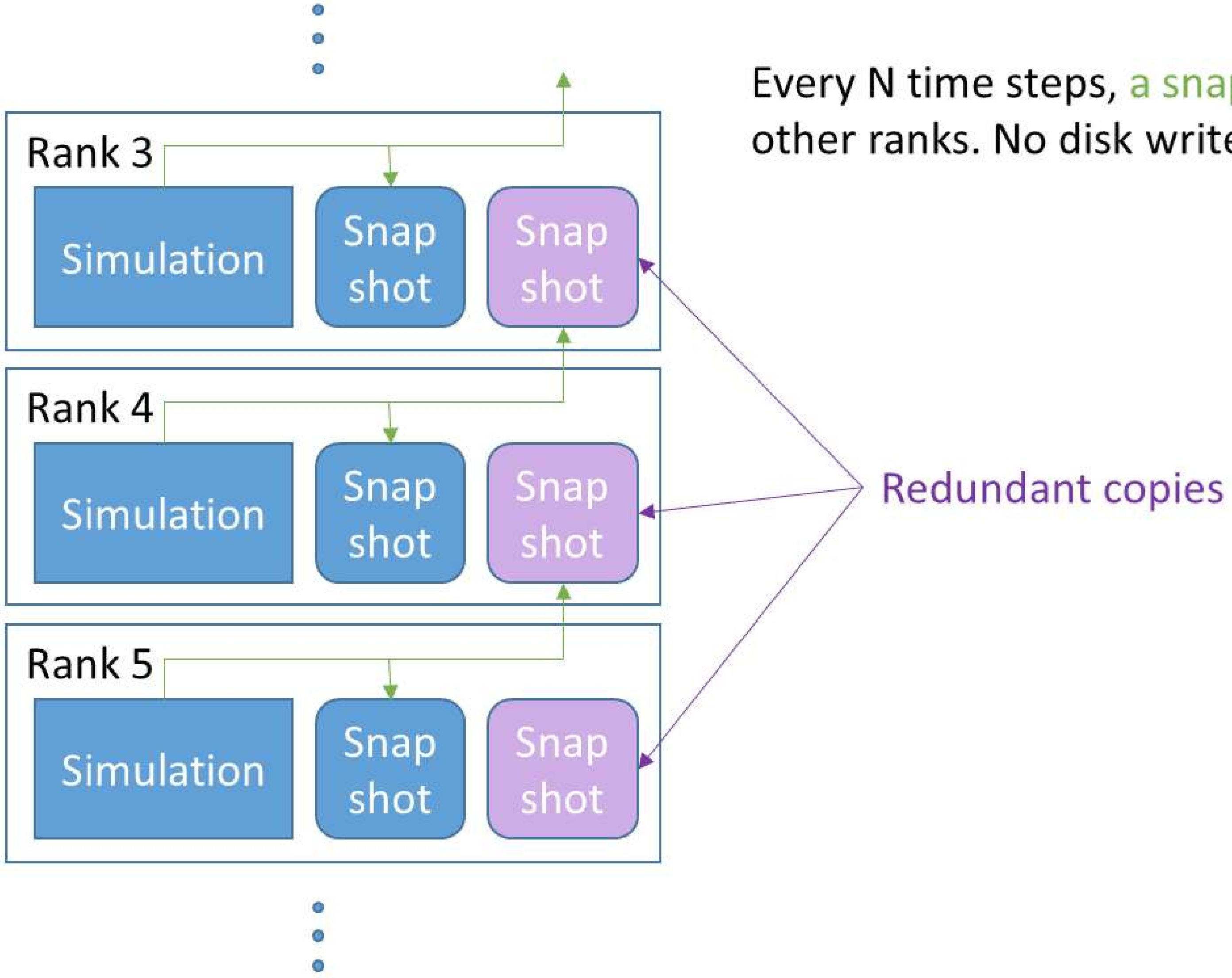
Resilience: Evaluation in ls1 mardyn



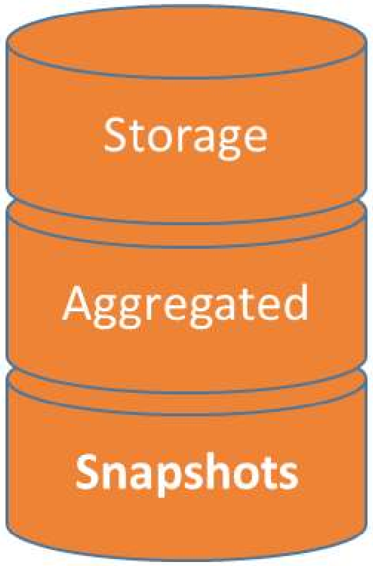
Every N time steps, a snapshot is made and written to disk. **Expensive!**

- Development of particle storage format to support both efficient checkpointing/restart and visualization
- Checkpointing
- **In-memory approaches to resilience (work-in-progress)**

Resilience: Evaluation in ls1 mardyn

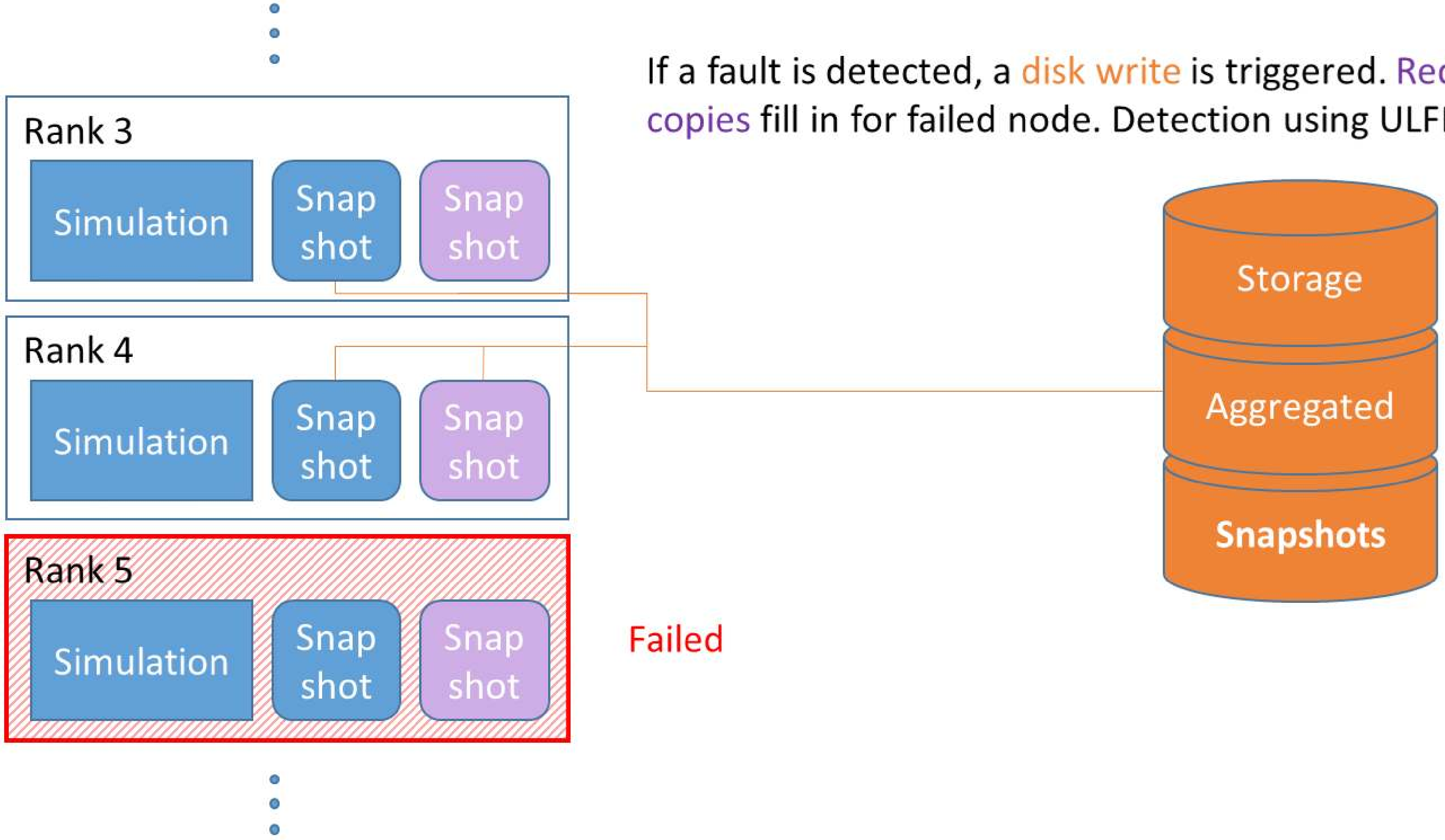


Every N time steps, a snapshot is made and distributed to other ranks. No disk writes occur.



- Development of particle storage format to support both efficient checkpointing/restart and visualization
- Checkpointing
- **In-memory approaches to resilience (work-in-progress)**

Resilience: Evaluation in ls1 mardyn



If a fault is detected, a disk write is triggered. Redundant copies fill in for failed node. Detection using ULFM is WIP.

- Development of particle storage format to support both efficient checkpointing/restart and visualization
- Checkpointing
- **In-memory approaches to resilience (work-in-progress)**

Summary and Outlook

- **Auto-Pas**
  → all ingredients available for auto-tuning
  → work in progress: actual tuning, different scenarios, etc.

- **Workflow Manager**
  → prototype and interfaces available
  → show case: EOS fitting
  → Extra-P for multi-parameter performance modeling
  → how about high-dimensional parameter spaces?

- **Resilience**
  → checkpointing, in-memory approach (wip)
  → compression of particle data (wip)

- **Outlook (project year 3): component integration**

- **Acknowledgements**
  - BMBF project TaLPas, www.talpas.de , 01IH16008
  - GCS large-scale project *Extreme-Scale MD Simulation of Droplet Coalescence*

SPONSORED BY THE

Federal Ministry of Education and Research

GCS

Gauss Centre for Supercomputing

- **Thank you for your attention!**