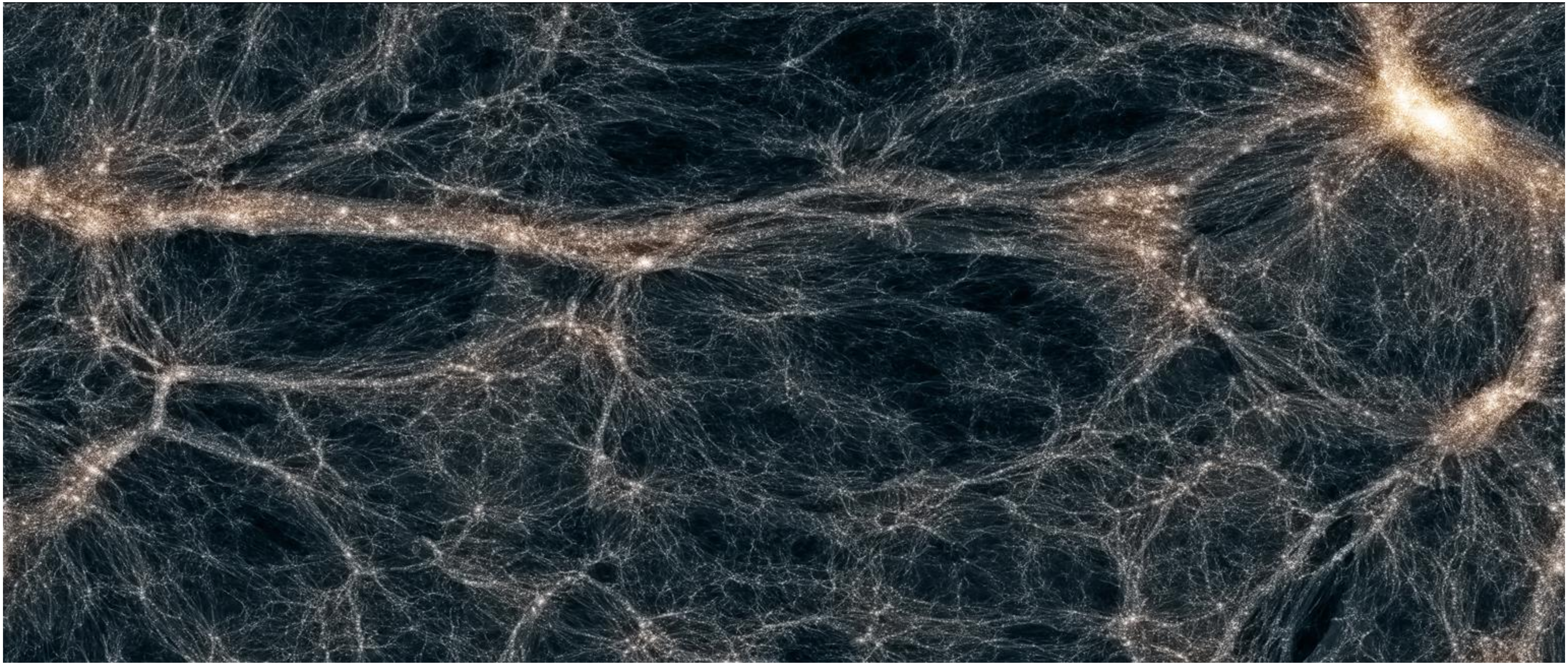
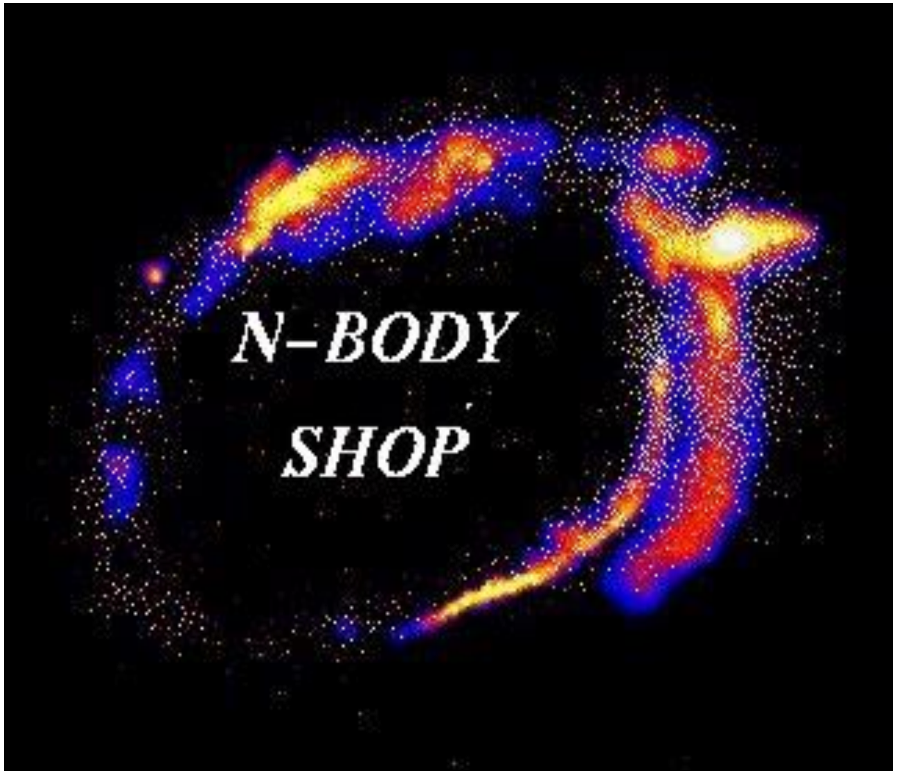
Scaling Clustered N-Body/SPH Simulations



Thomas Quinn



Fabio Governato

Lauren Anderson

Michael Tremmel

Ferah Munshi

Joachim Stadel

James Wadsley

PPL PARALLEL PROGRAMMING LAB

DEPT OF COMPUTER SCIENCE, UNIVERSITY OF ILLINOIS

Laxmikant Kale

Filippo Gioachin

Pritish Jetley

Celso Mendes

Amit Sharma

Lukasz Wesolowski

Gengbin Zheng

Edgar Solomonik

Cosmology at 380,000 years

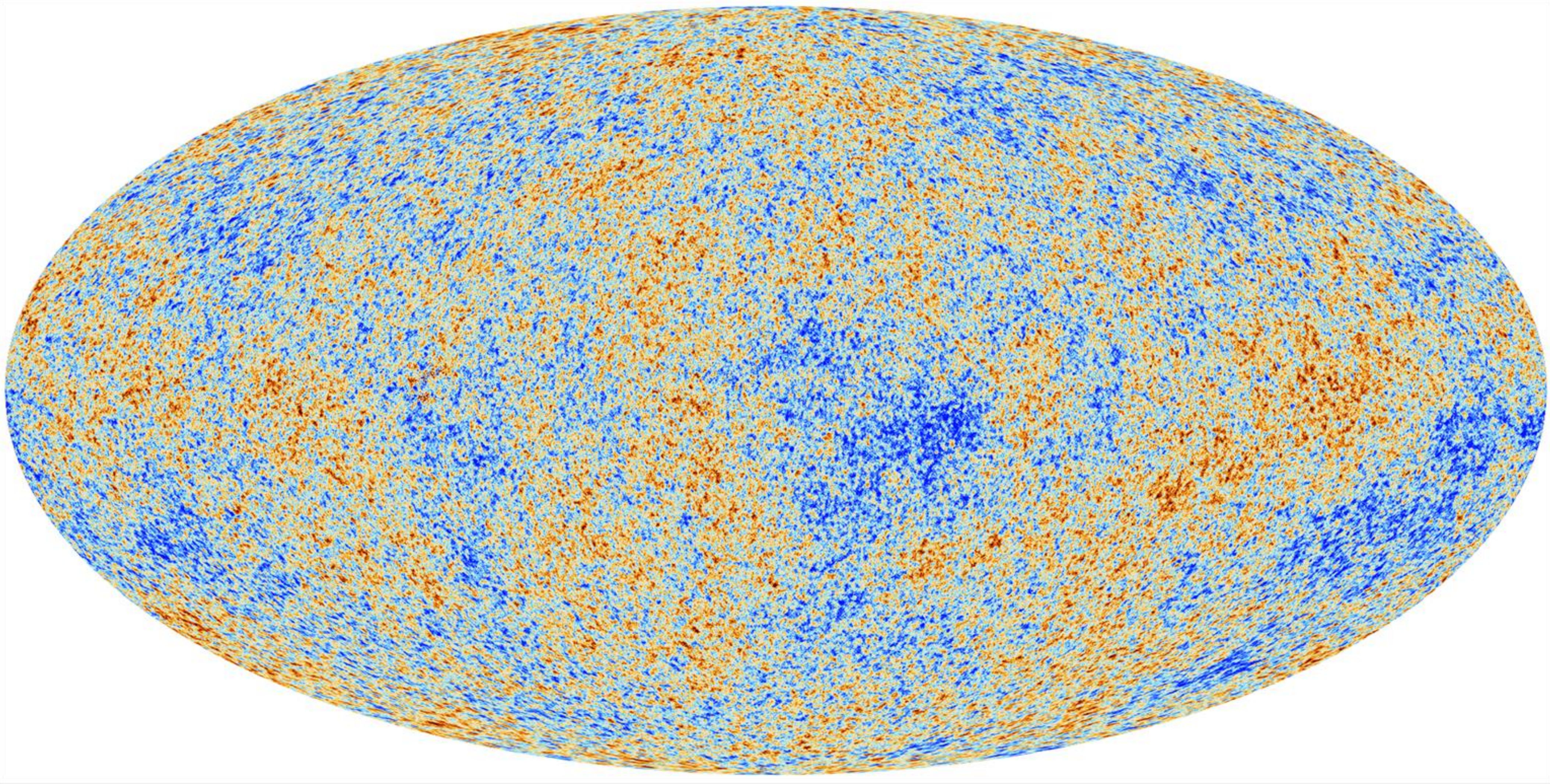
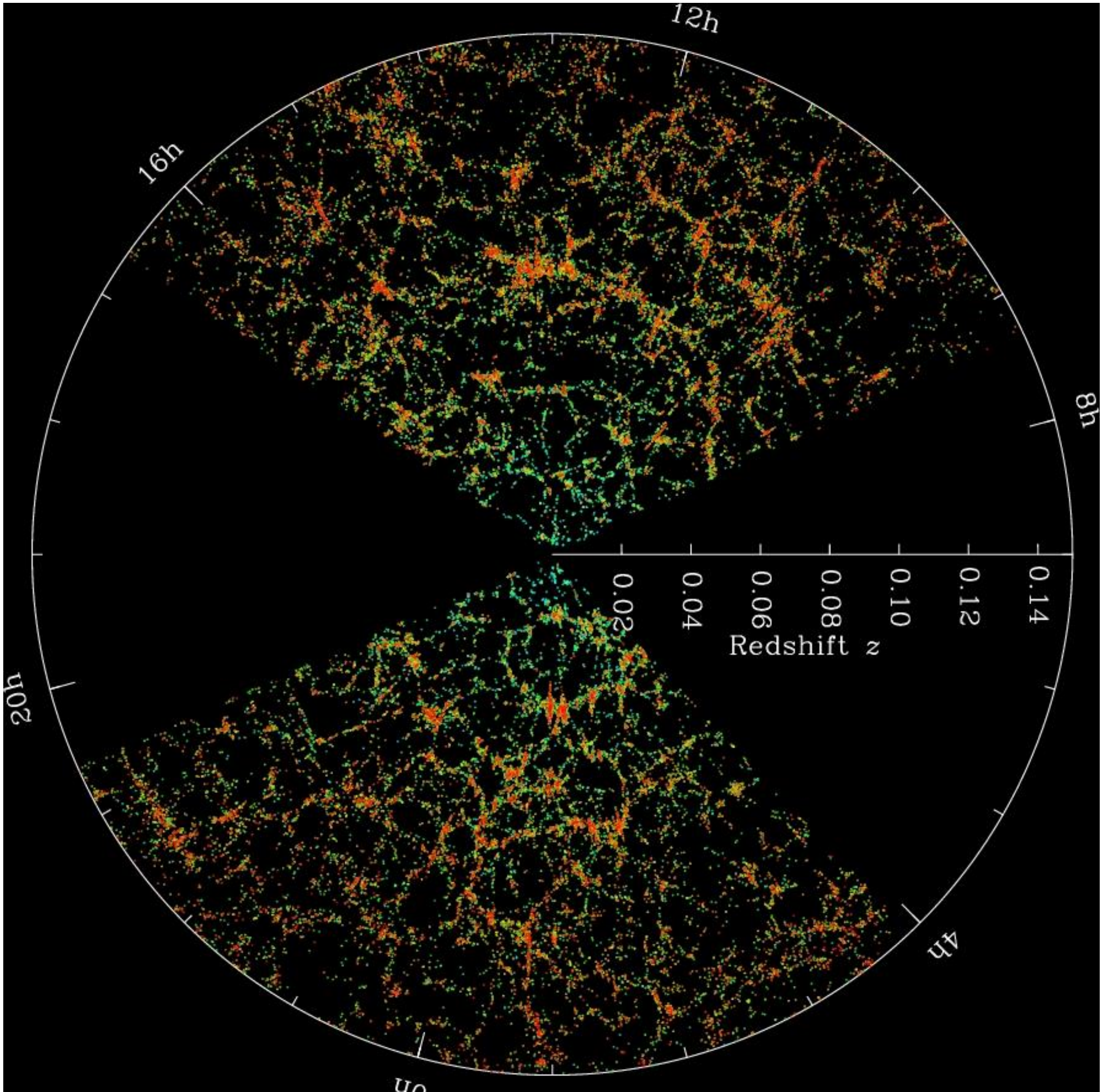
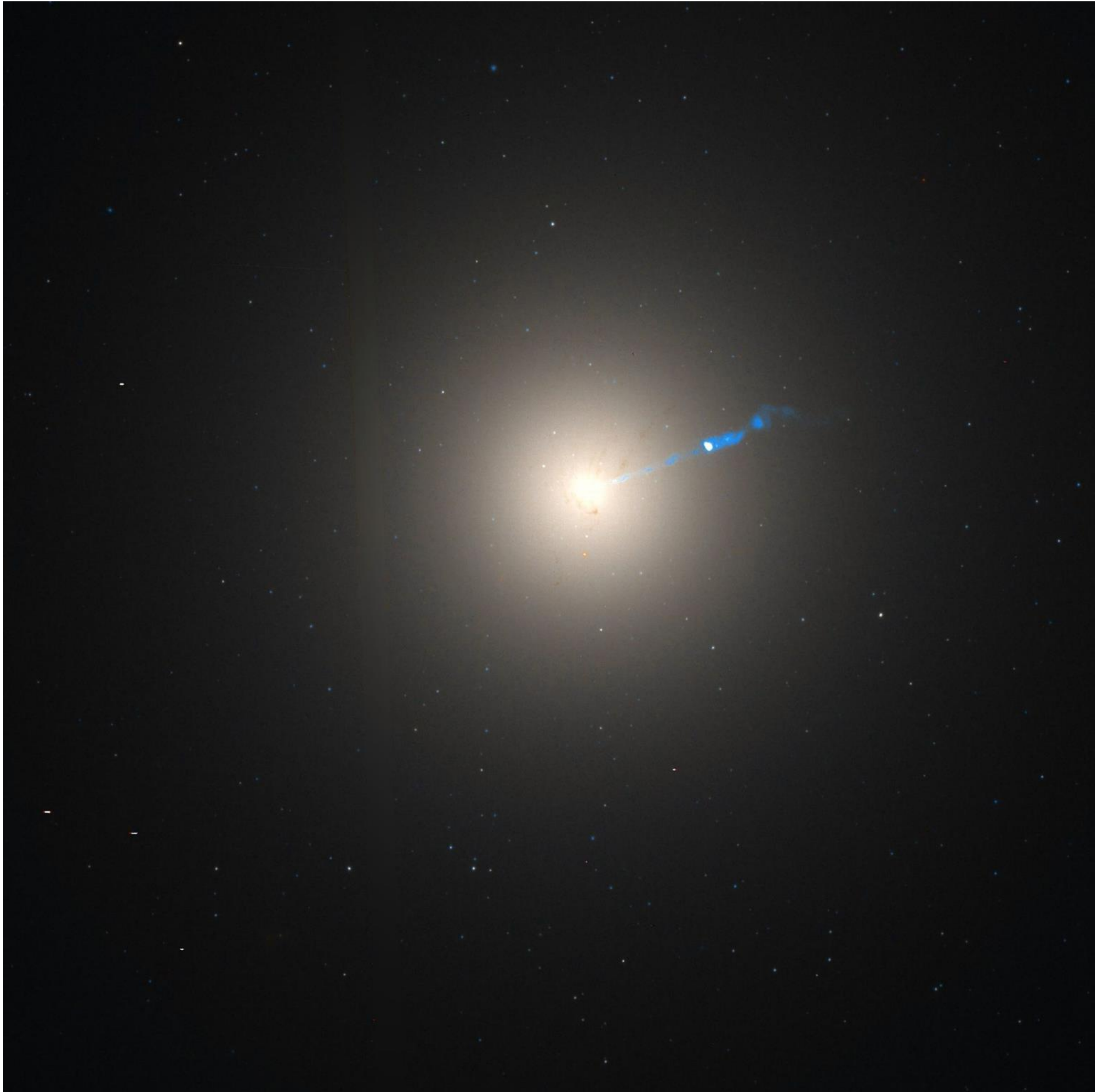


Image courtesy ESA/Planck

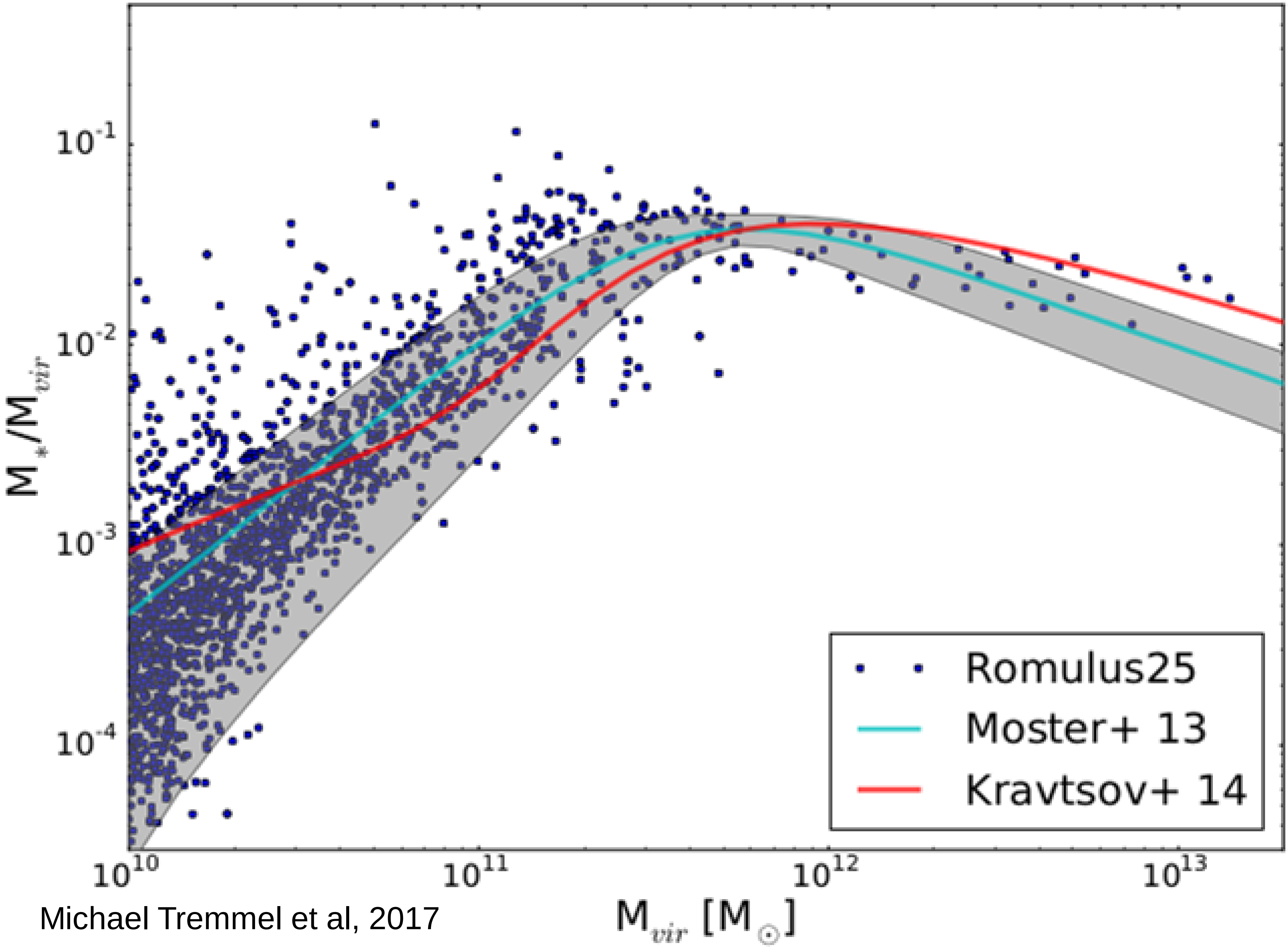

12h
16h
8h
20h
4h
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Redshift z

... is not so simple



Computational Cosmology

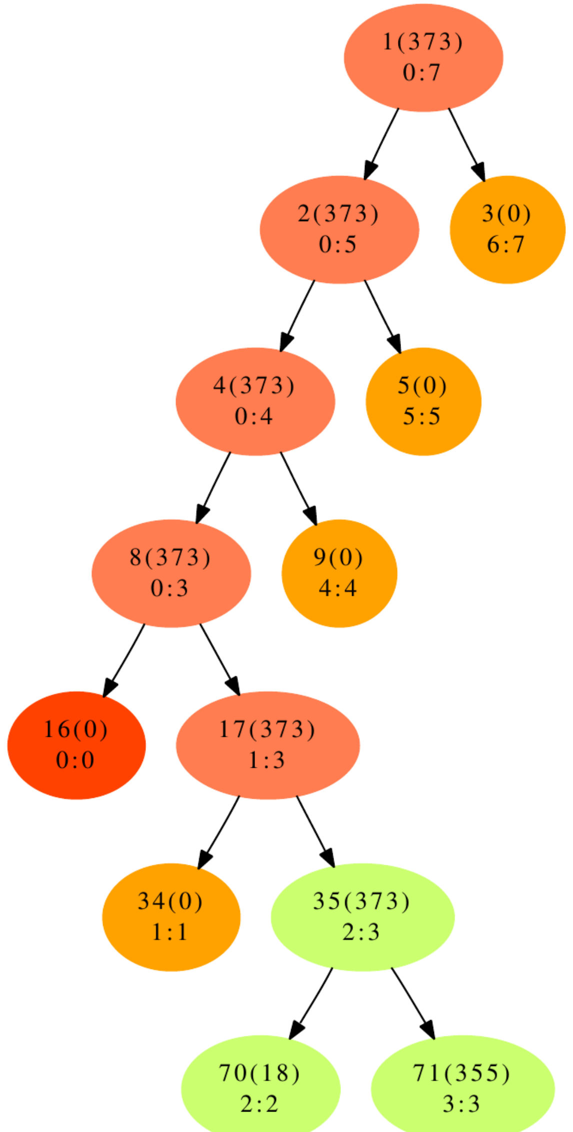
- CMB has fluctuations of 1e-5
- Galaxies are overdense by 1e7
- It happens (mostly) through **Gravitational Collapse**
- Making testable predictions from a cosmological hypothesis requires
  - Non-linear, dynamic calculation
  - e.g. **Computer simulation**



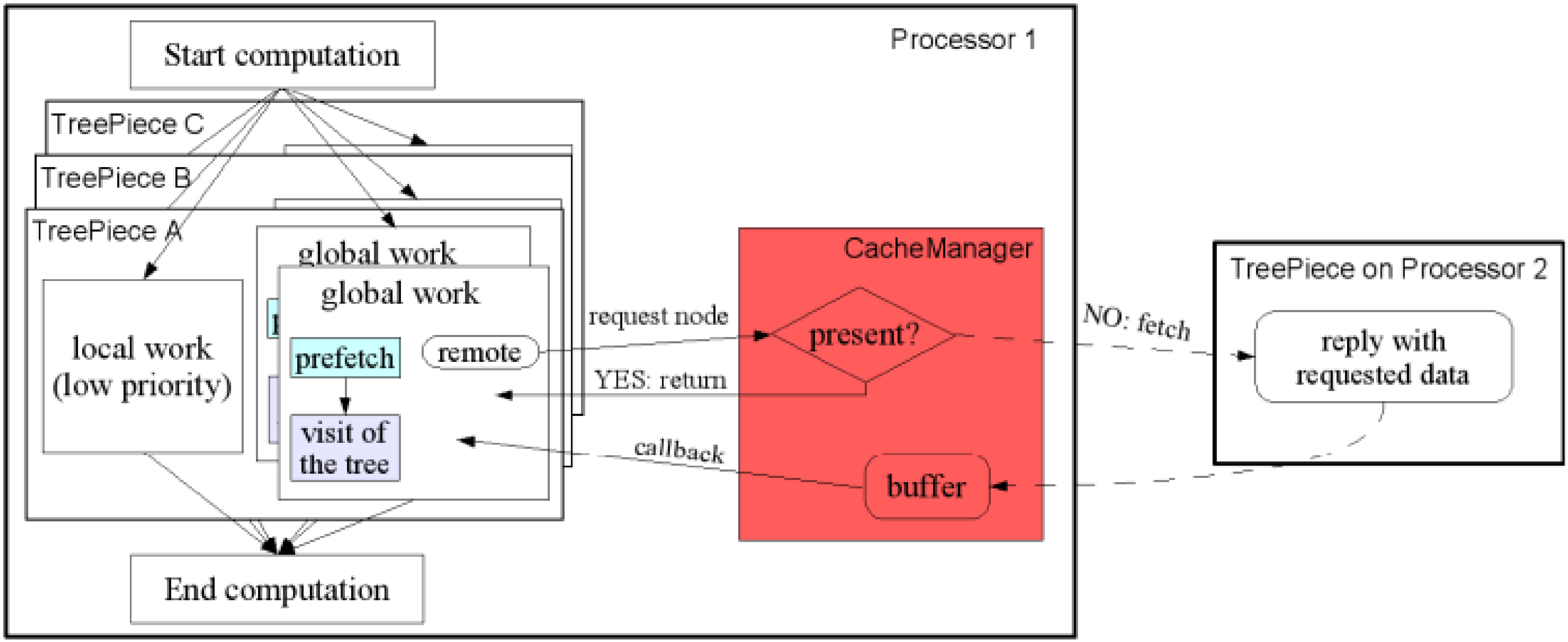
Michael Tremmel et al, 2017

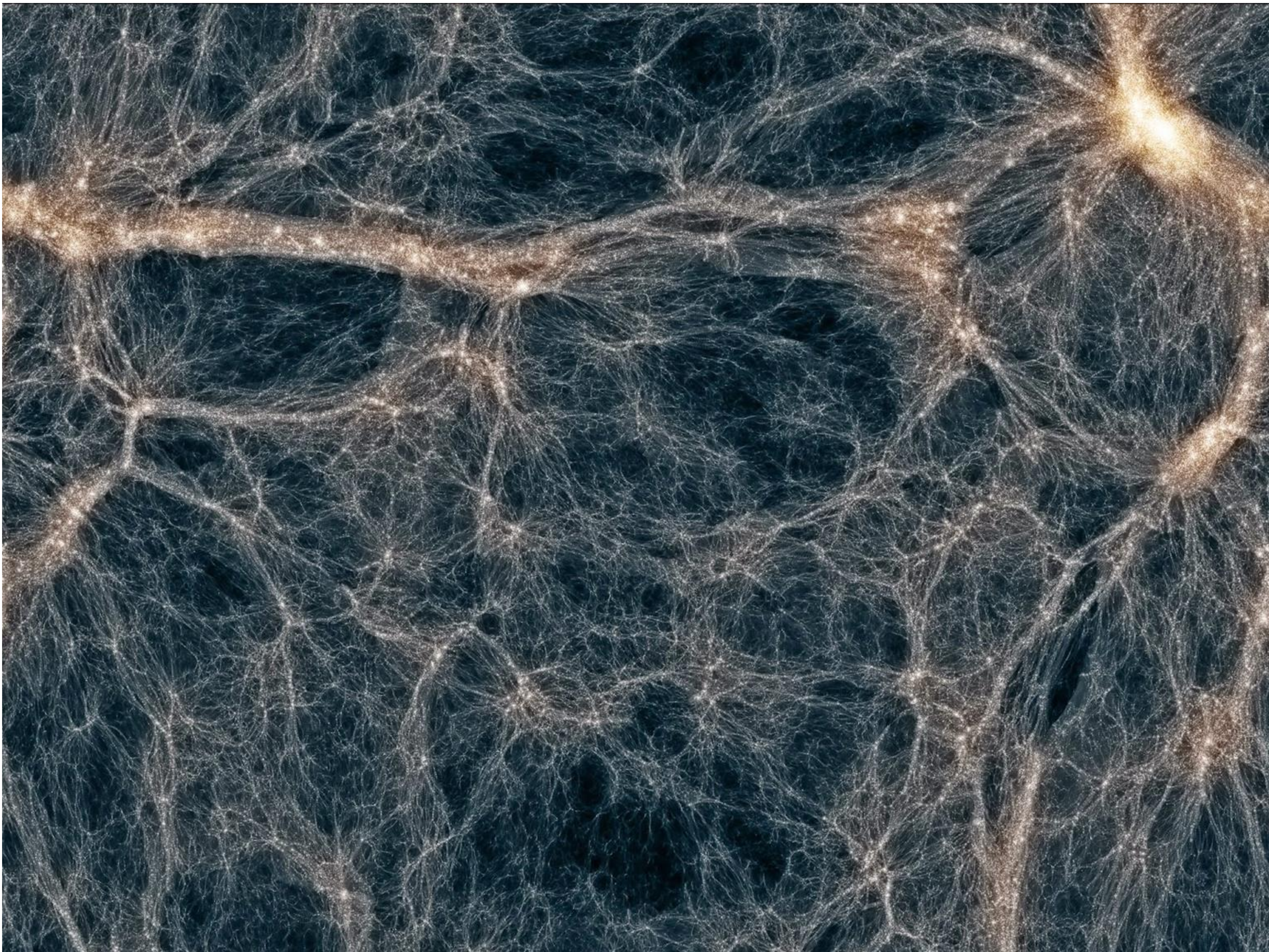
TreePiece: basic data structure

- A “vertical slice” of the tree, all the way to the root.
- Nodes are either:
  - Internal
  - External
  - Boundary (shared)

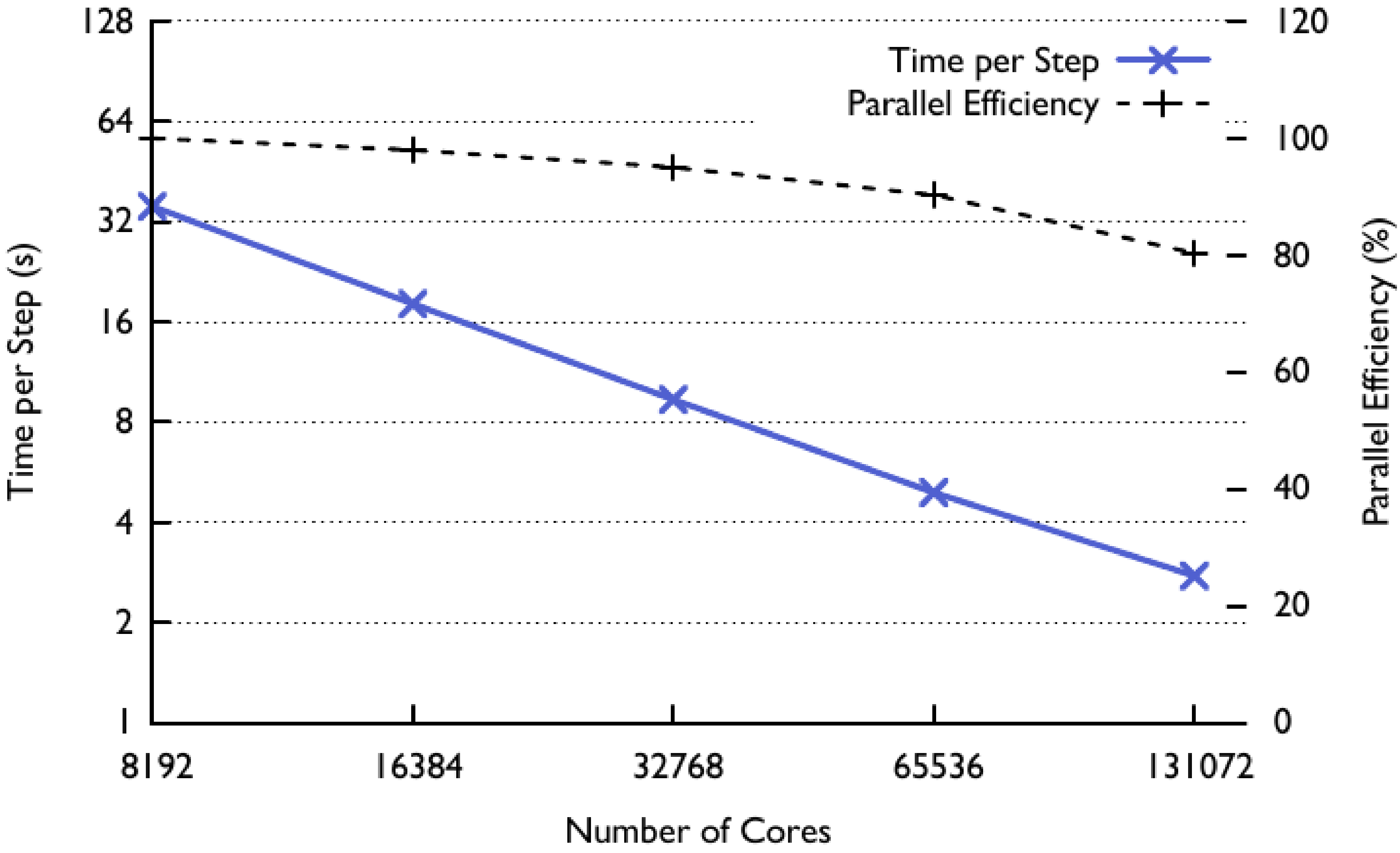


Overall treewalk structure

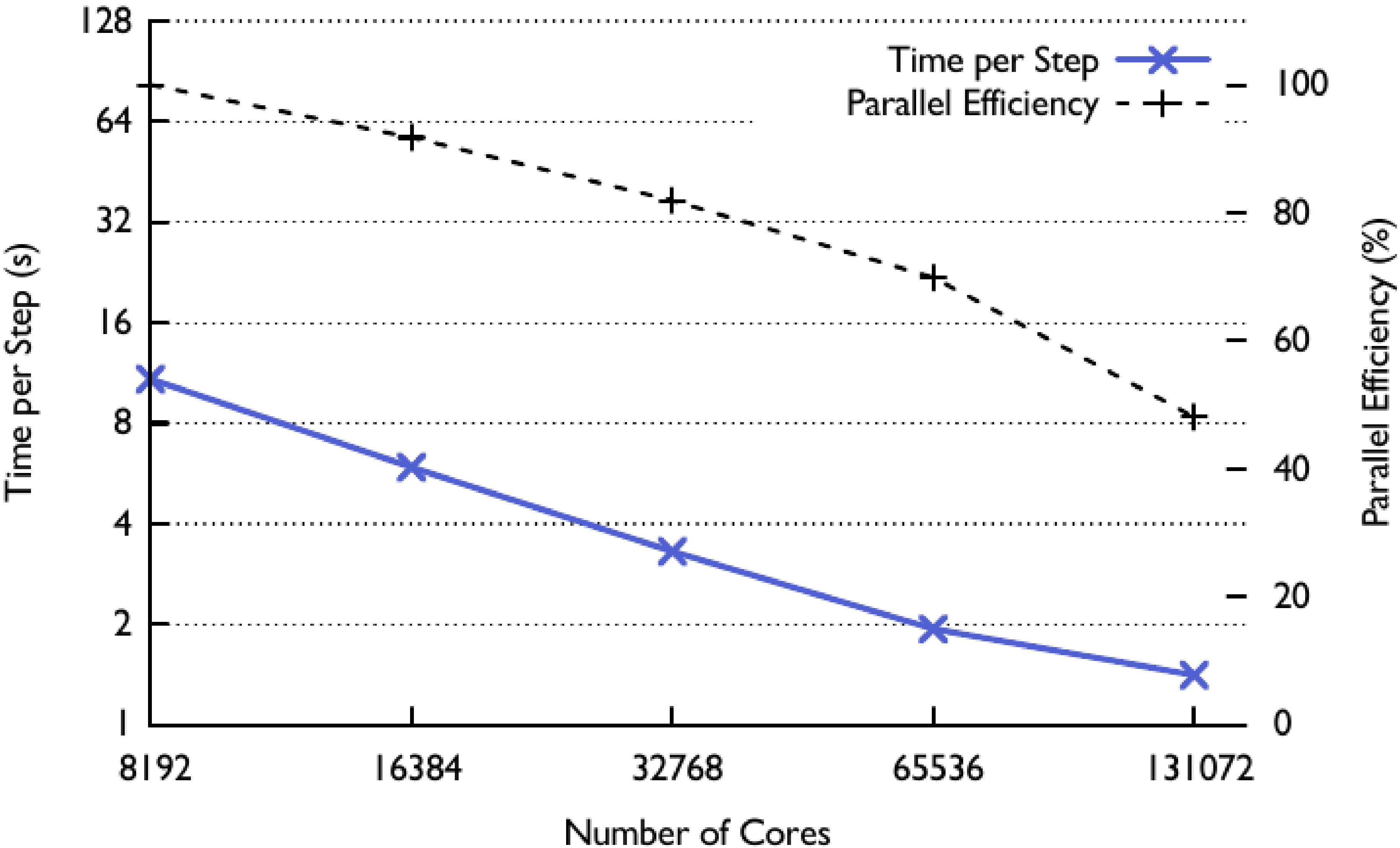




Speedups for 2 billion clustered particles



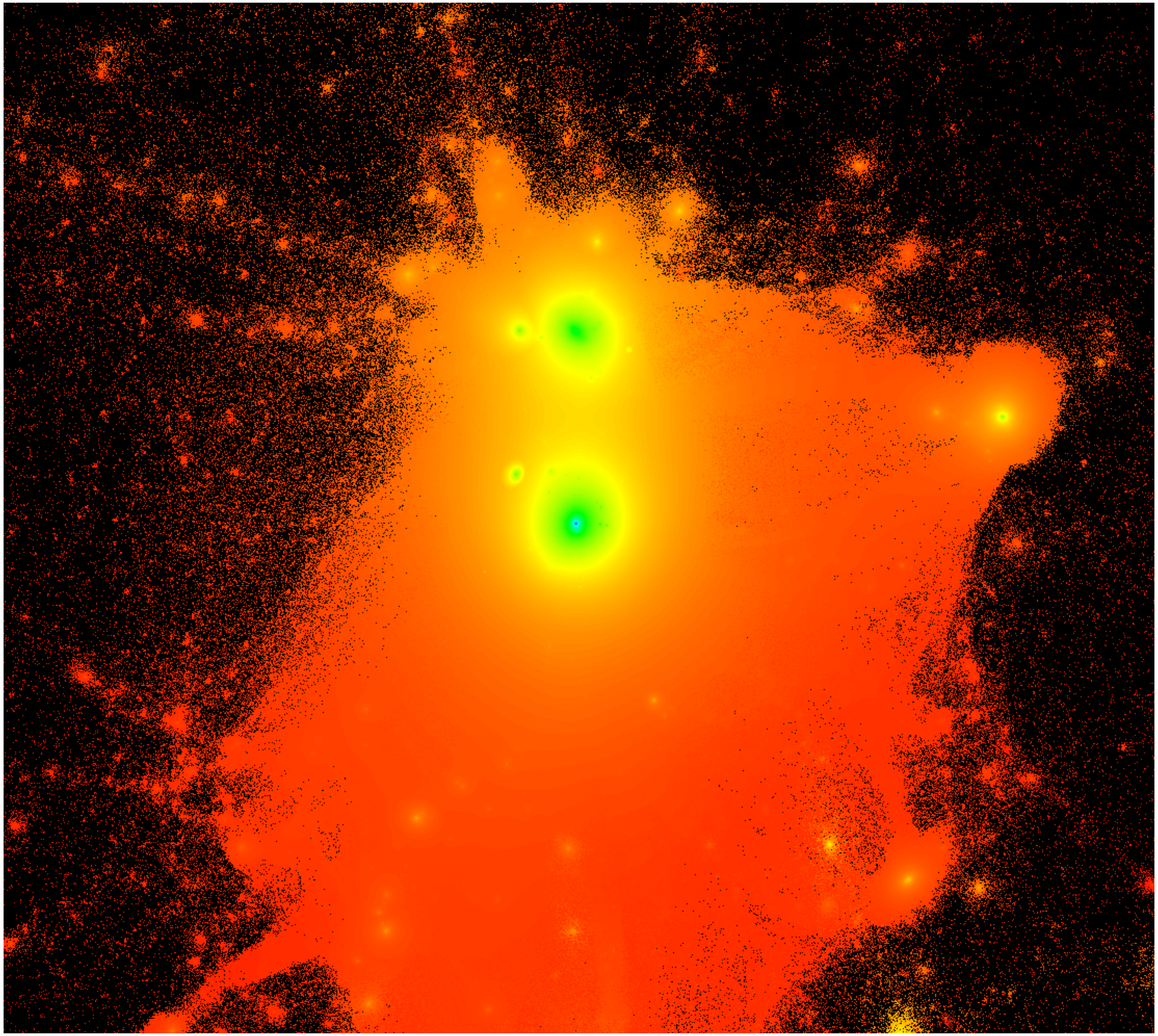
Multistep Speedup



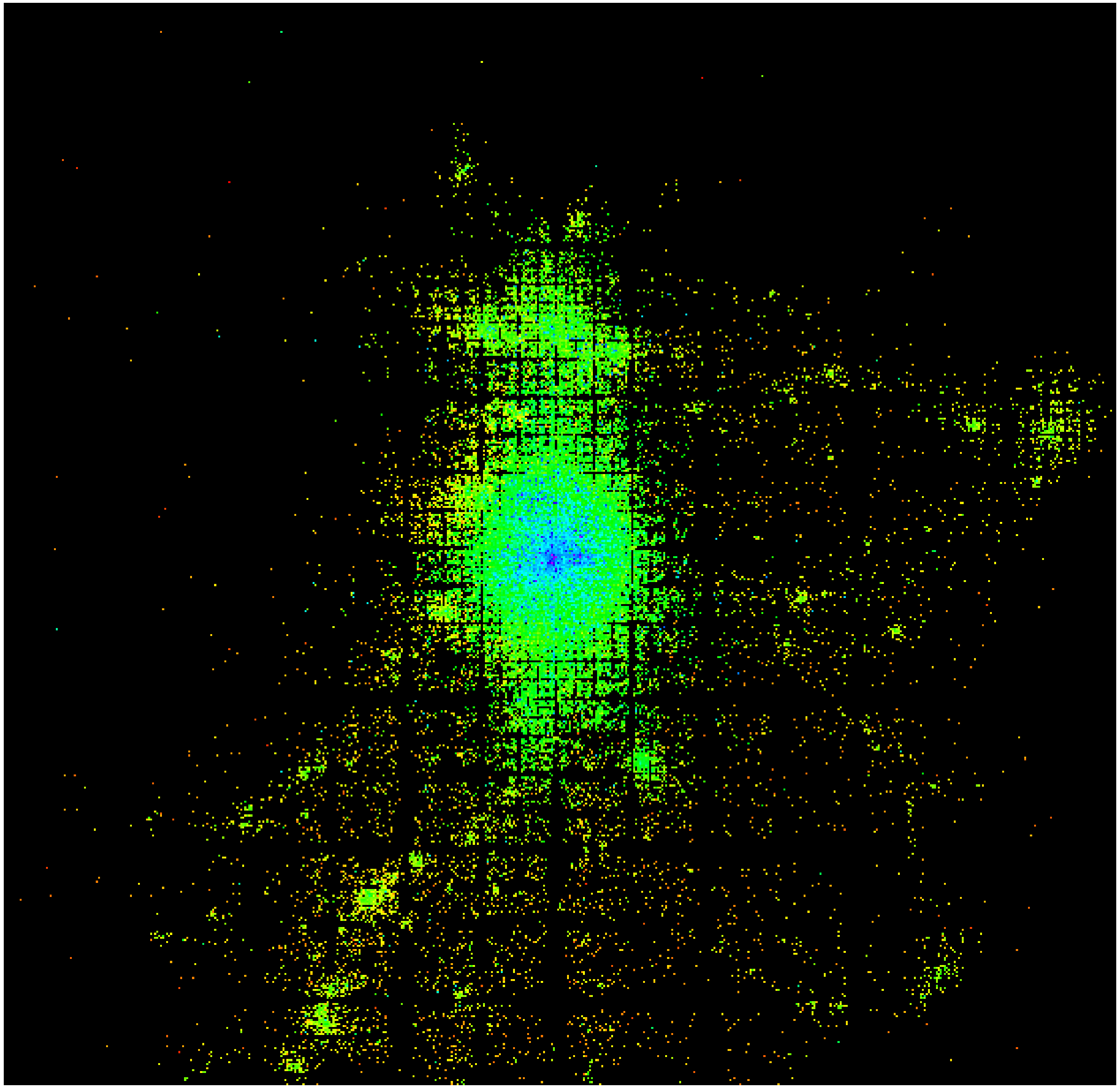
Clustered/Multistepping Challenges

- Load/particle imbalance
- Communication imbalance
- Rapid switching between phases
  - Gravity, Star formation, SMBH mergers
- Fixed costs:
  - Domain Decomposition
  - Load balancing
  - Tree build

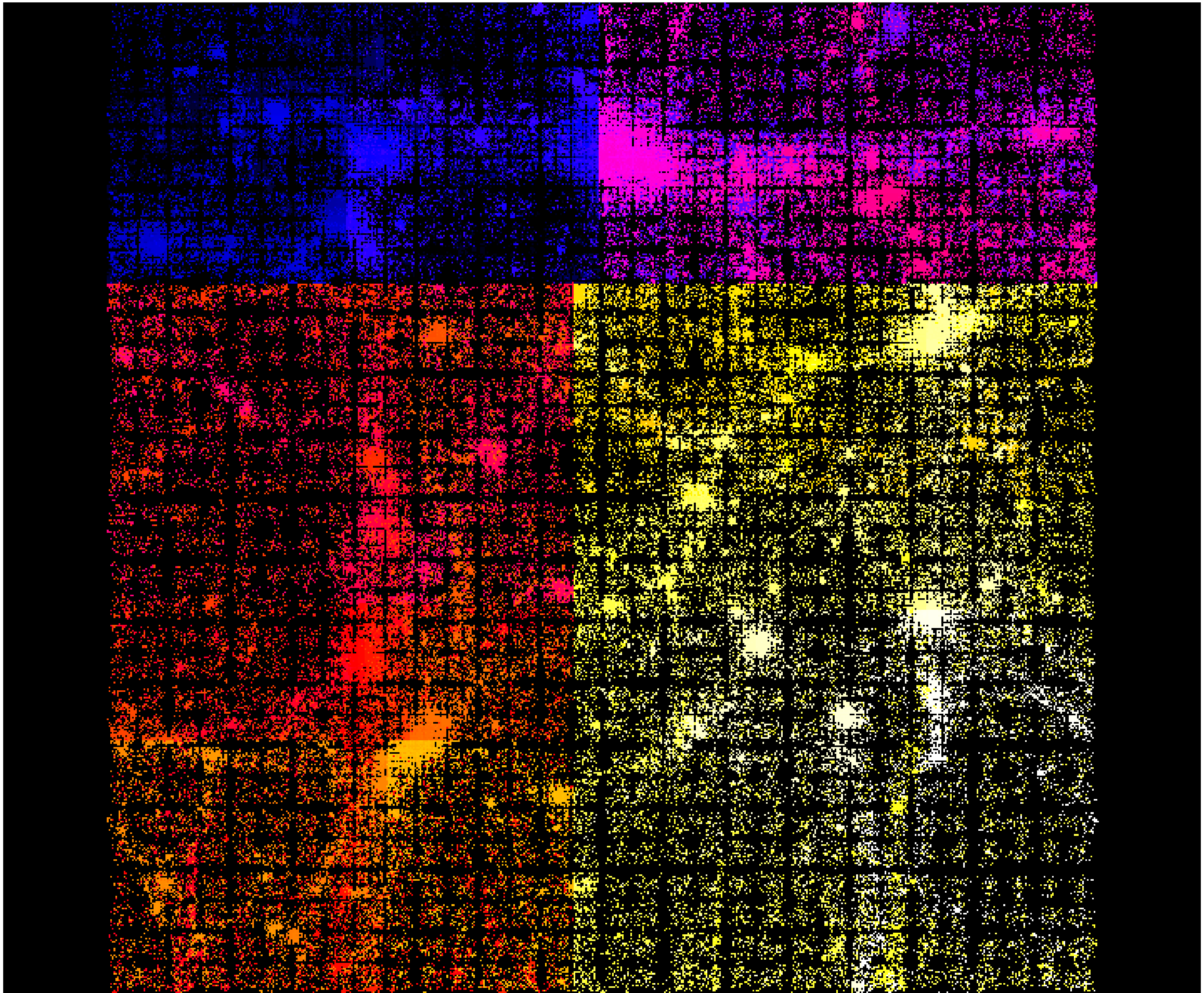
Zoomed Cluster simulation



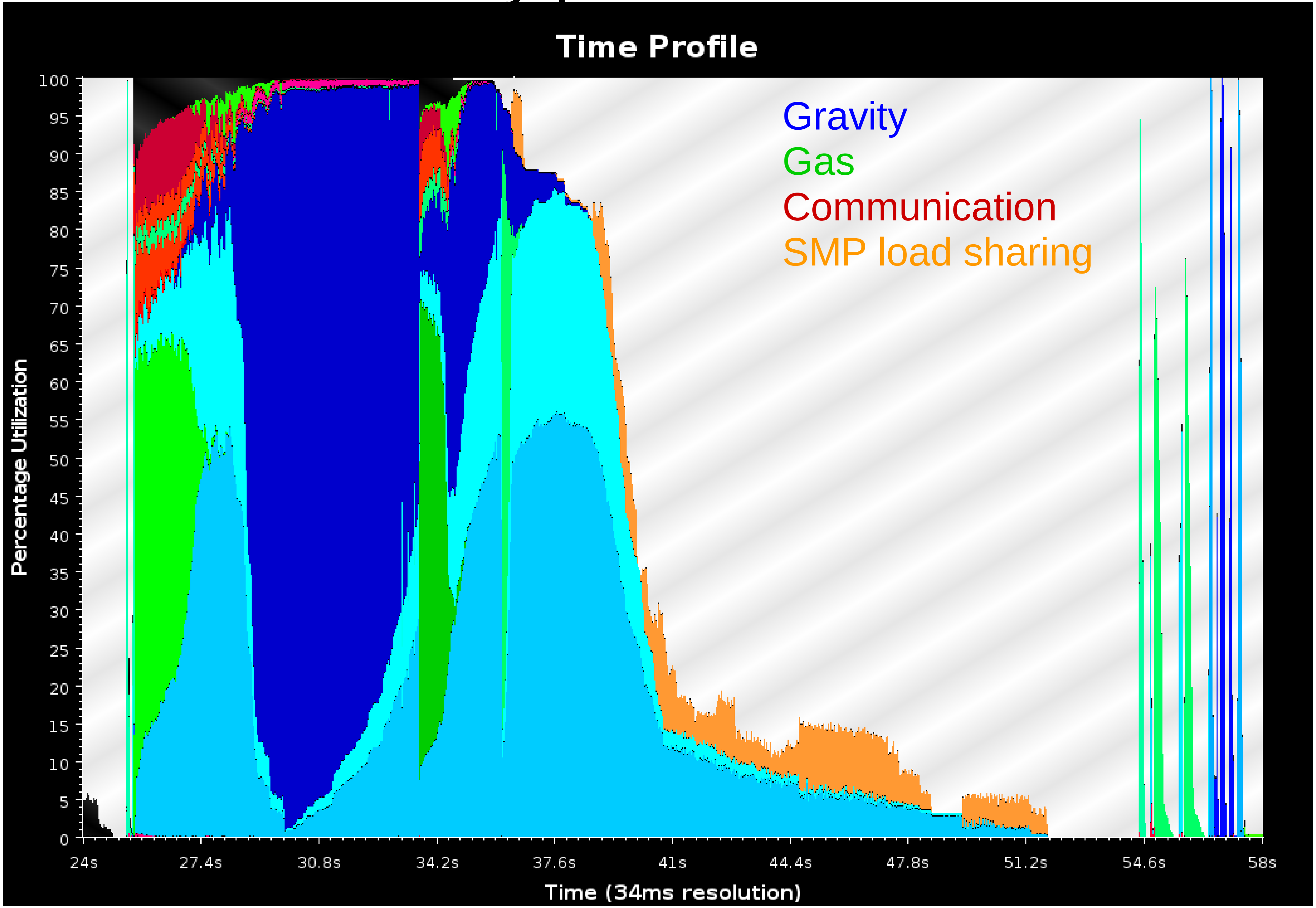
Load distribution



ORB Load Balancing

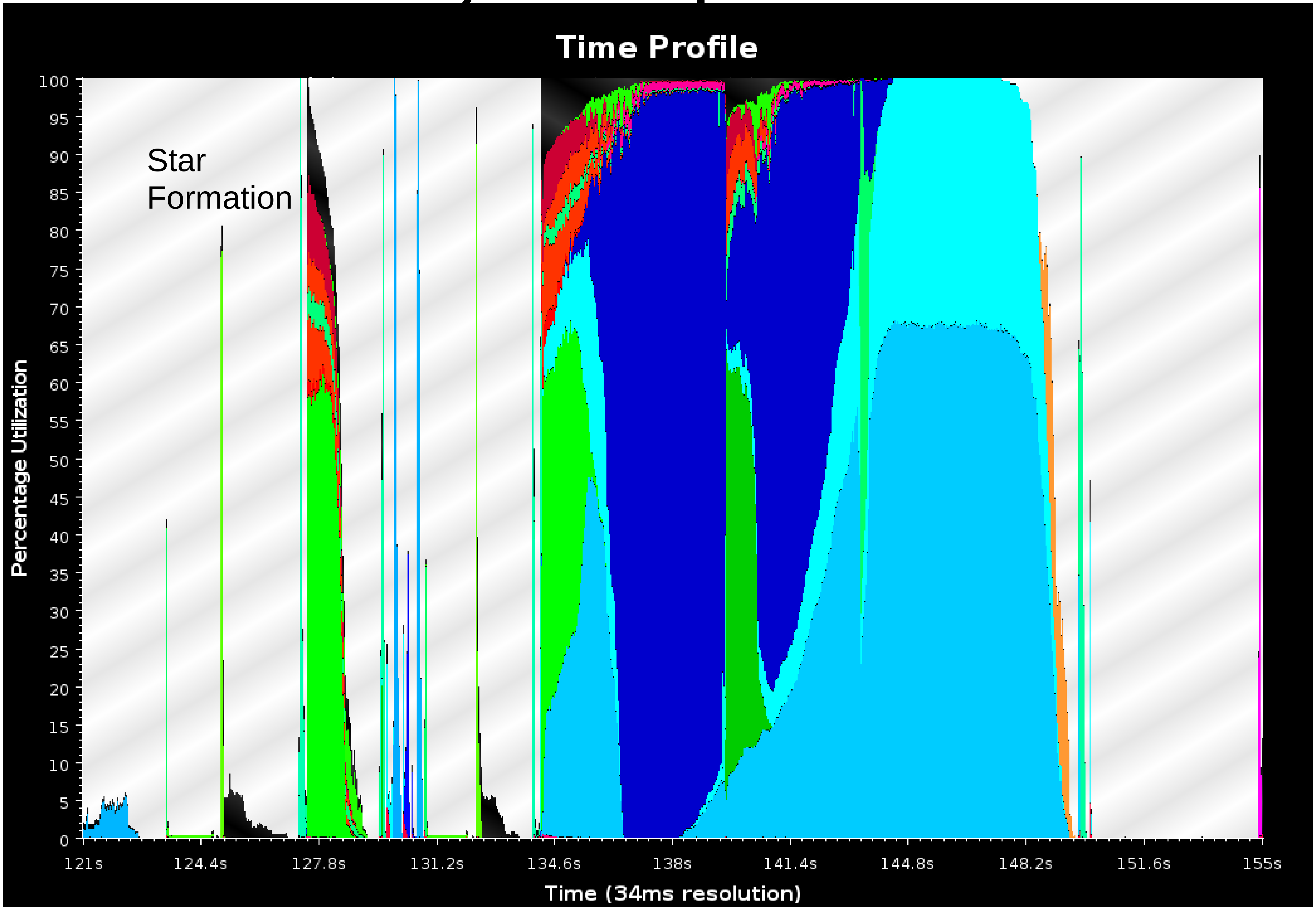


LB by particle count



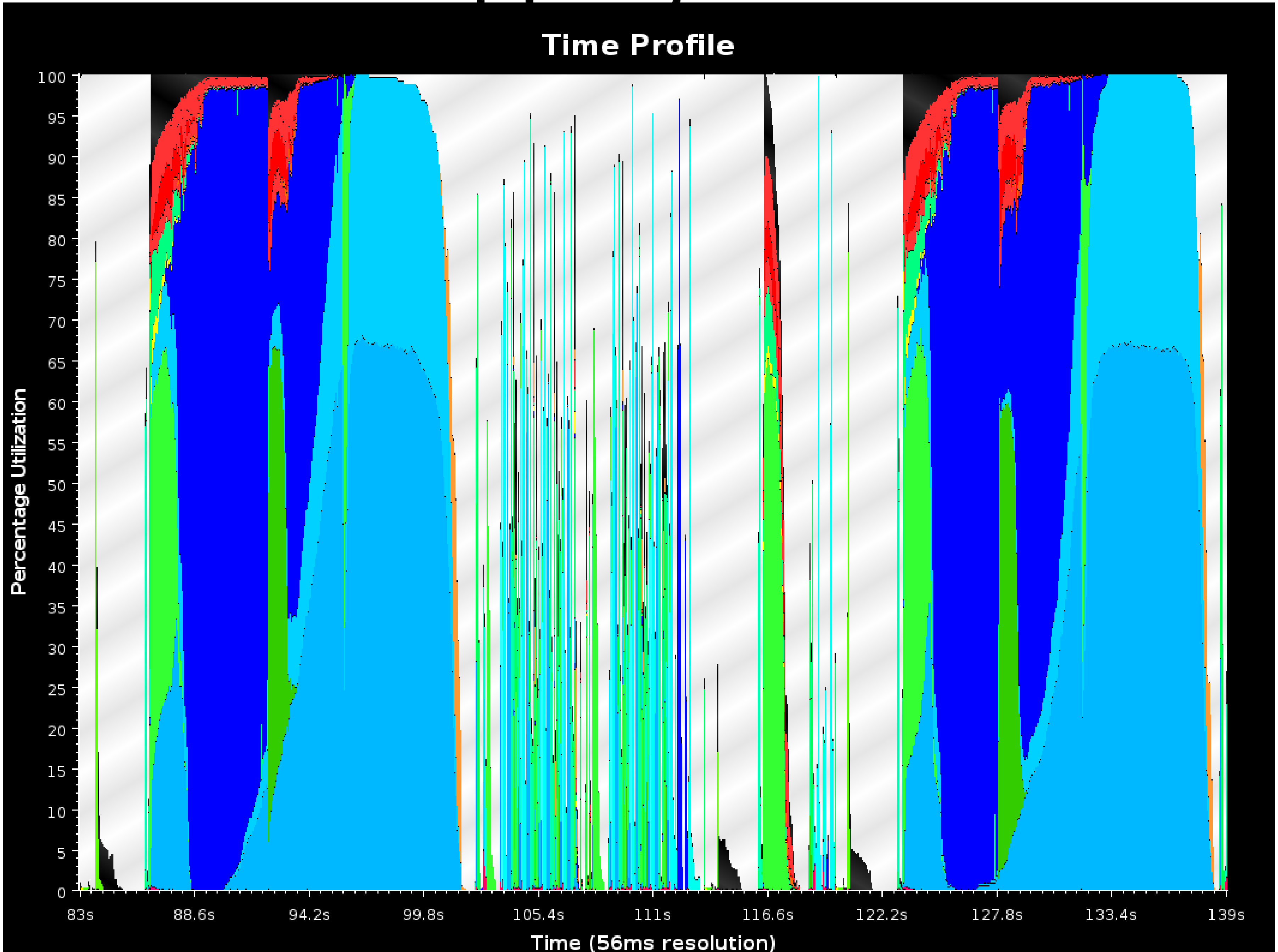
29.4 seconds

LB by Compute time

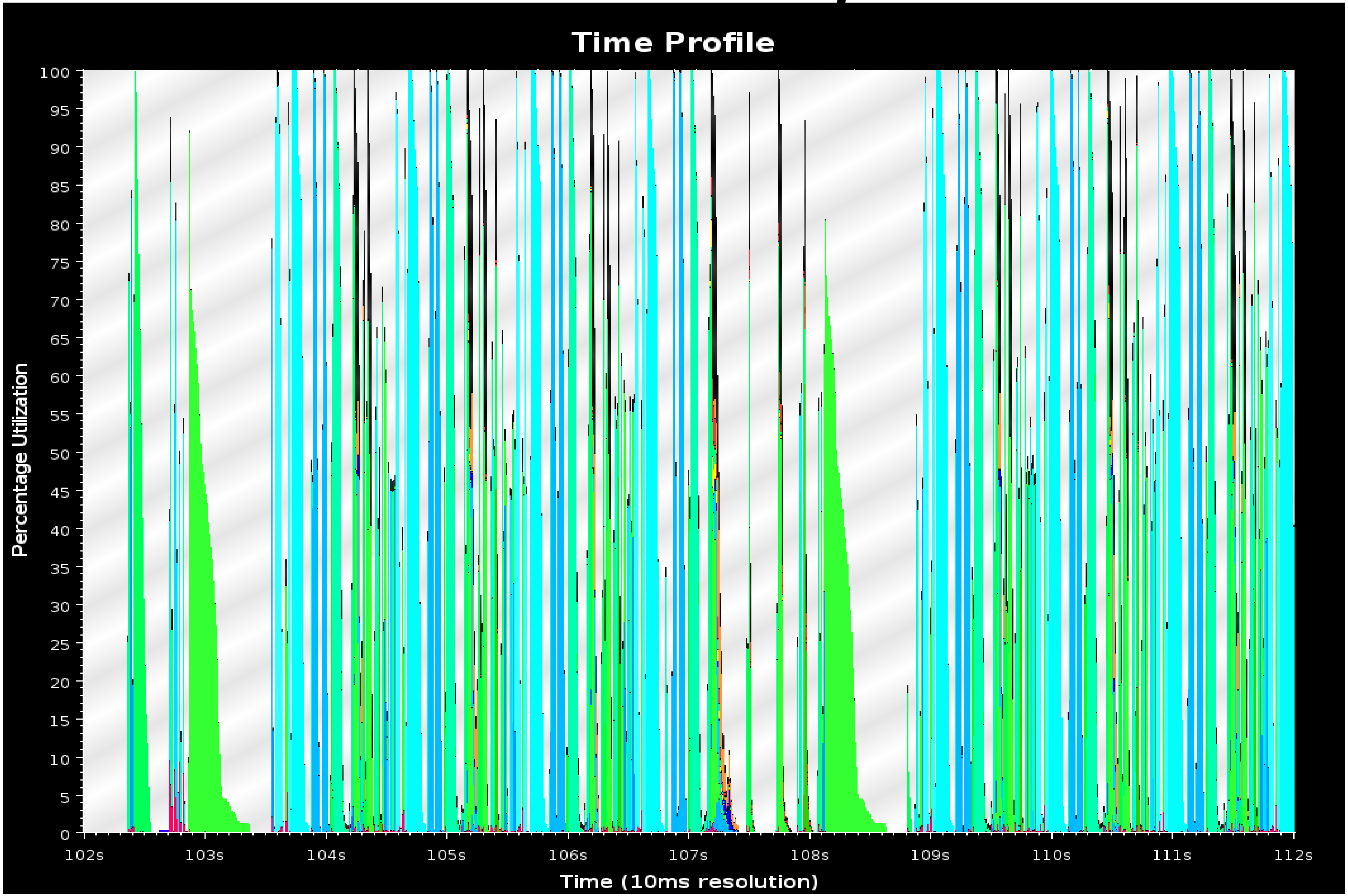


15.8 seconds

Multistepping Utilization



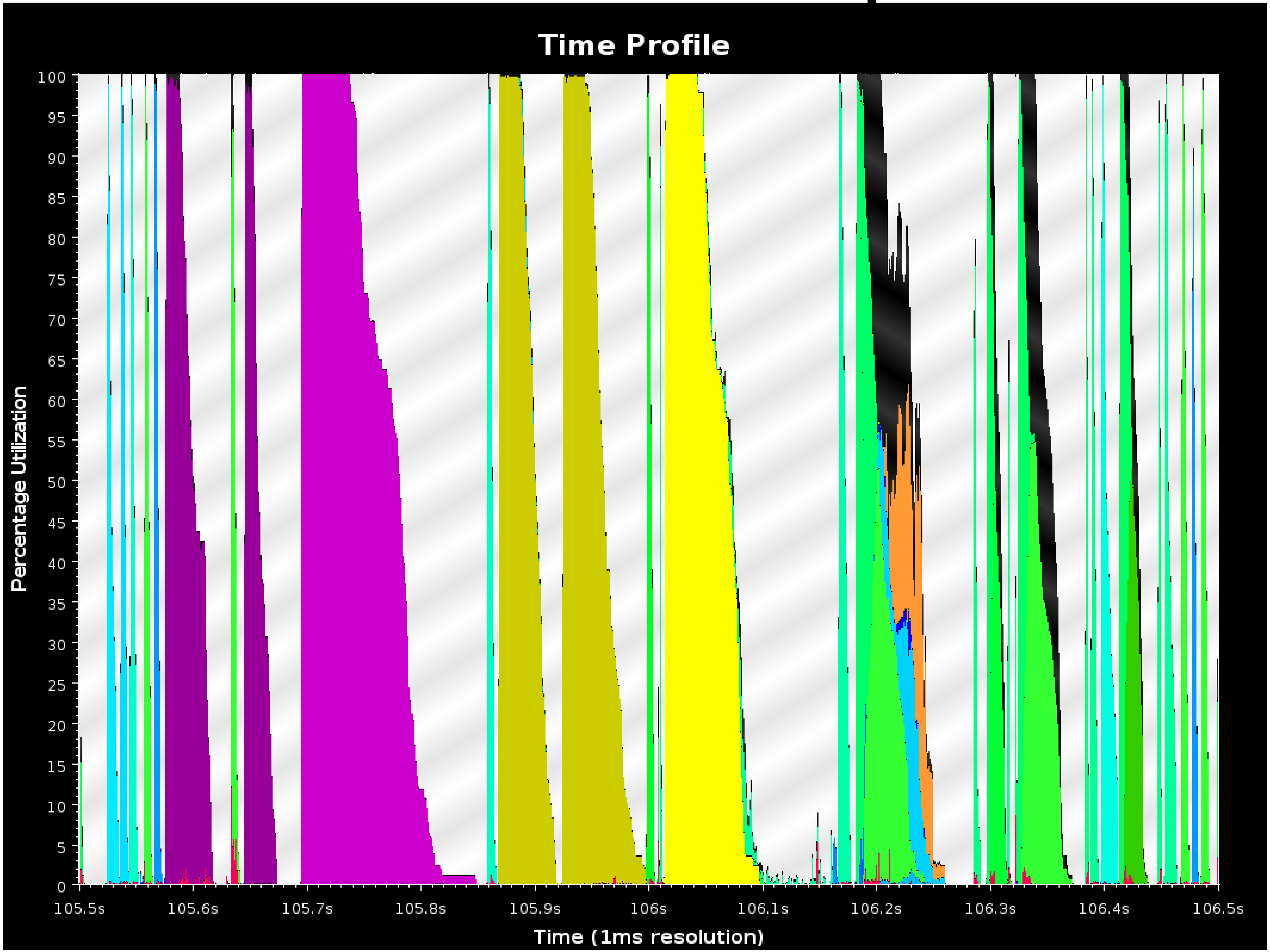
Small rungs:



Energy

Energy

Smallest step



Total interval: 1 second

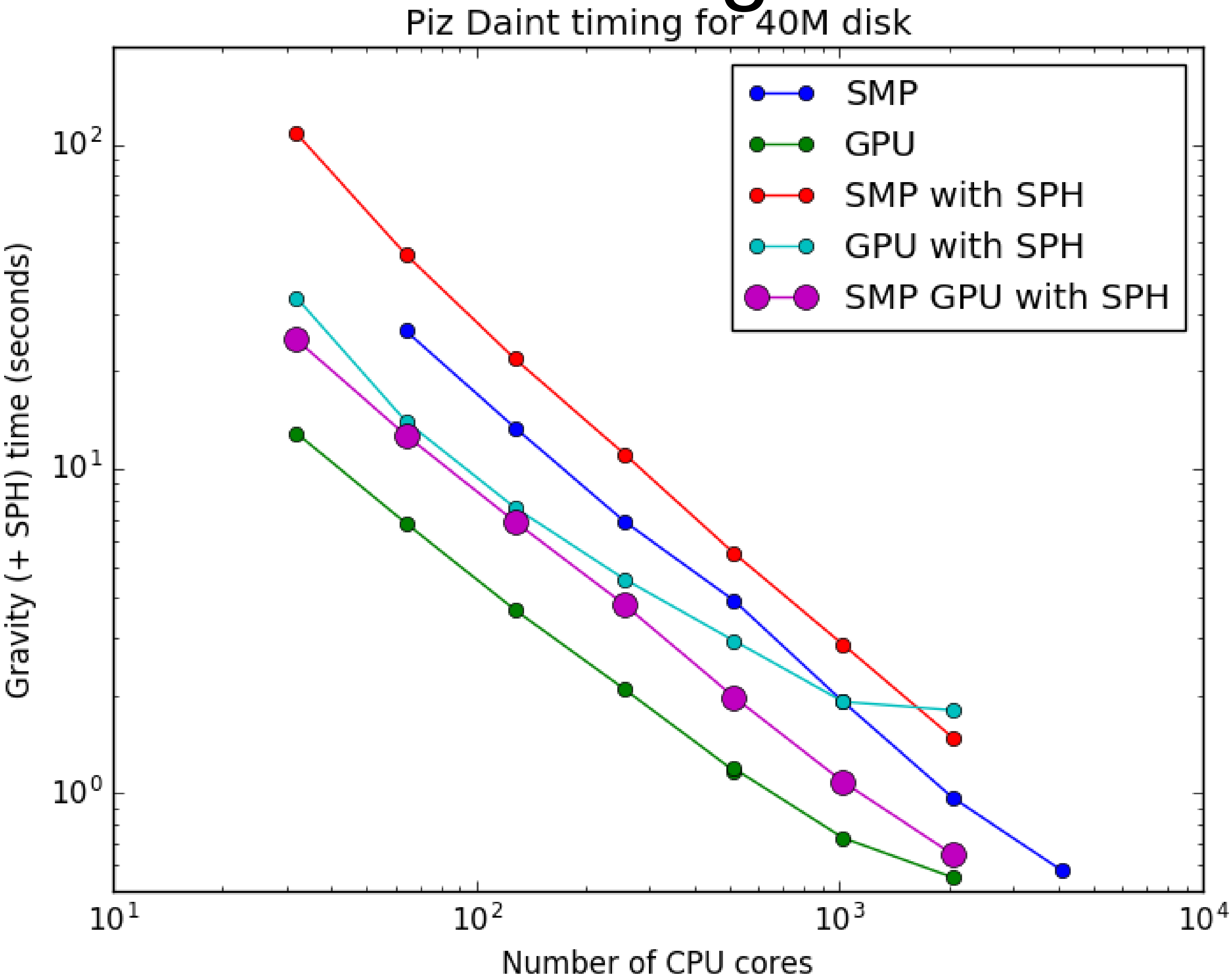
CPU Scaling Summary

- Load balancing the big steps is (mostly) solved
- Load balancing/optimizing the small steps is what is needed:
  - Small steps dominate the total time
  - Small steps increase throughput even when not optimal
  - Plenty of opportunity for improvement

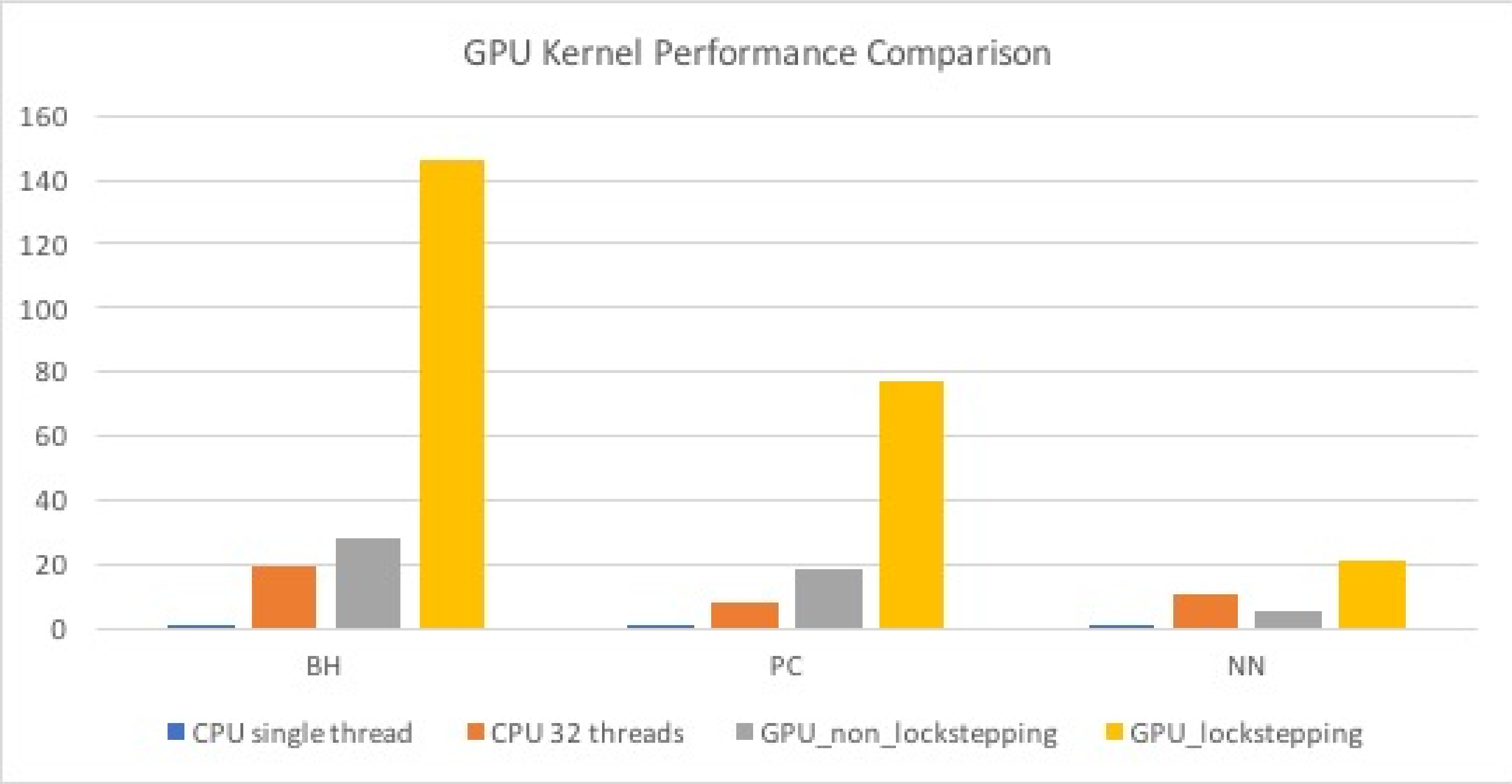
GPU Implementation: Gravity Only

- Load (SMP node) local tree/particle data onto the GPU
- Load prefetched remote tree onto the GPU
- CPUs walk tree and pass interaction lists
  - Lists are batched to minimize number of data transfers
- "Missed" treenodes: walk is resumed when data arrives: interaction list plus new tree data sent to the GPU.

Grav/SPH scaling with GPUs

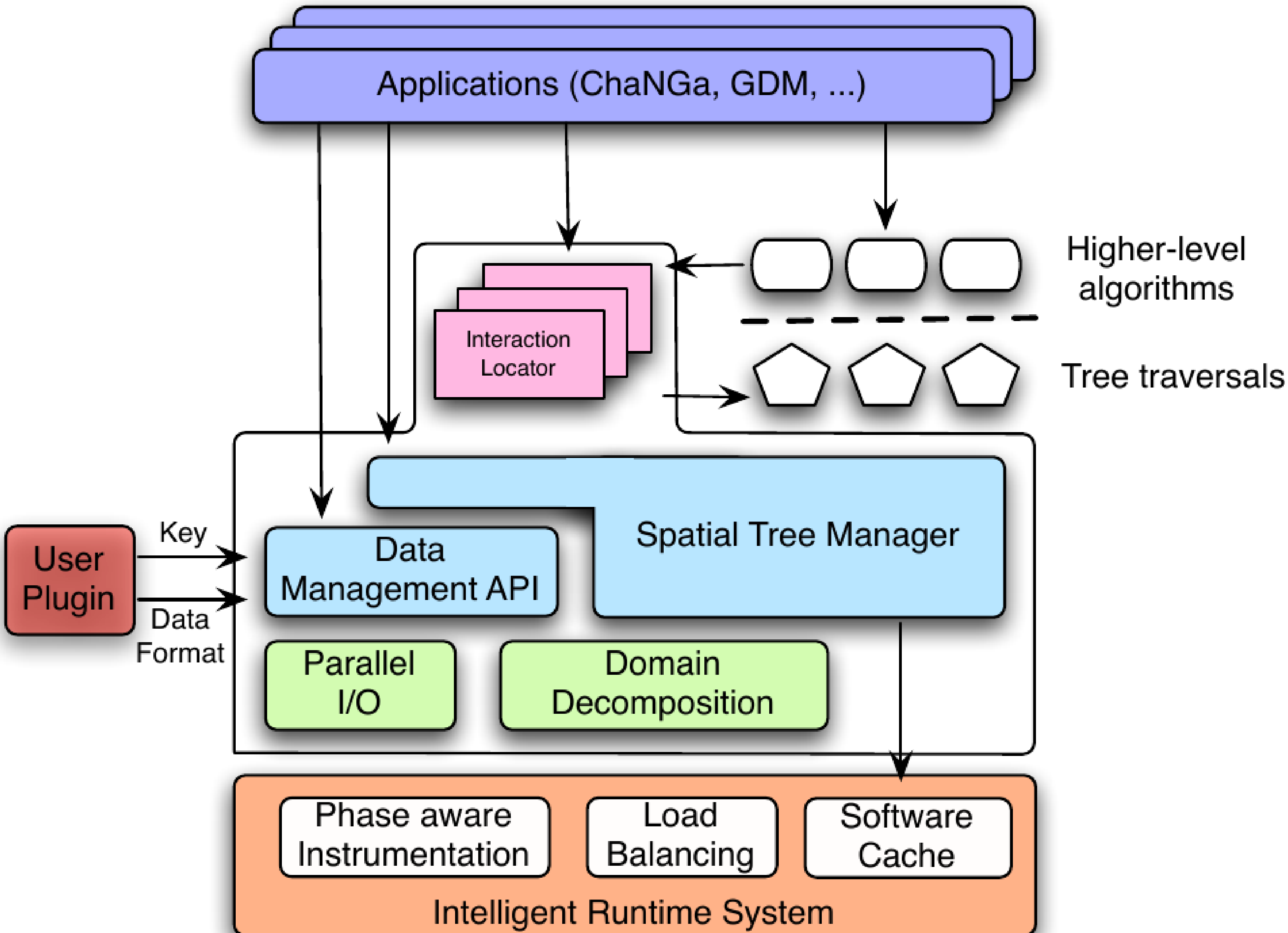


Tree walking on the GPU



Jianqiau Liu, Purdue University

Paratreet: parallel framework for tree algorithms



Availability

- ChaNGa: http://github.com/N-bodyShop/changa
  - See the Wiki for a developer's guide
  - Extensible: e.g. ChaNGa-MM by Phil Chang
- Paratreet: http://github.com/paratreet
  - Some design discussion and sample code

Acknowledgments

- NSF ITR
- NSF Astronomy
- NSF XSEDE program for computing
- BlueWaters Petascale Computing
- NASA HST
- NASA Advanced Supercomuting