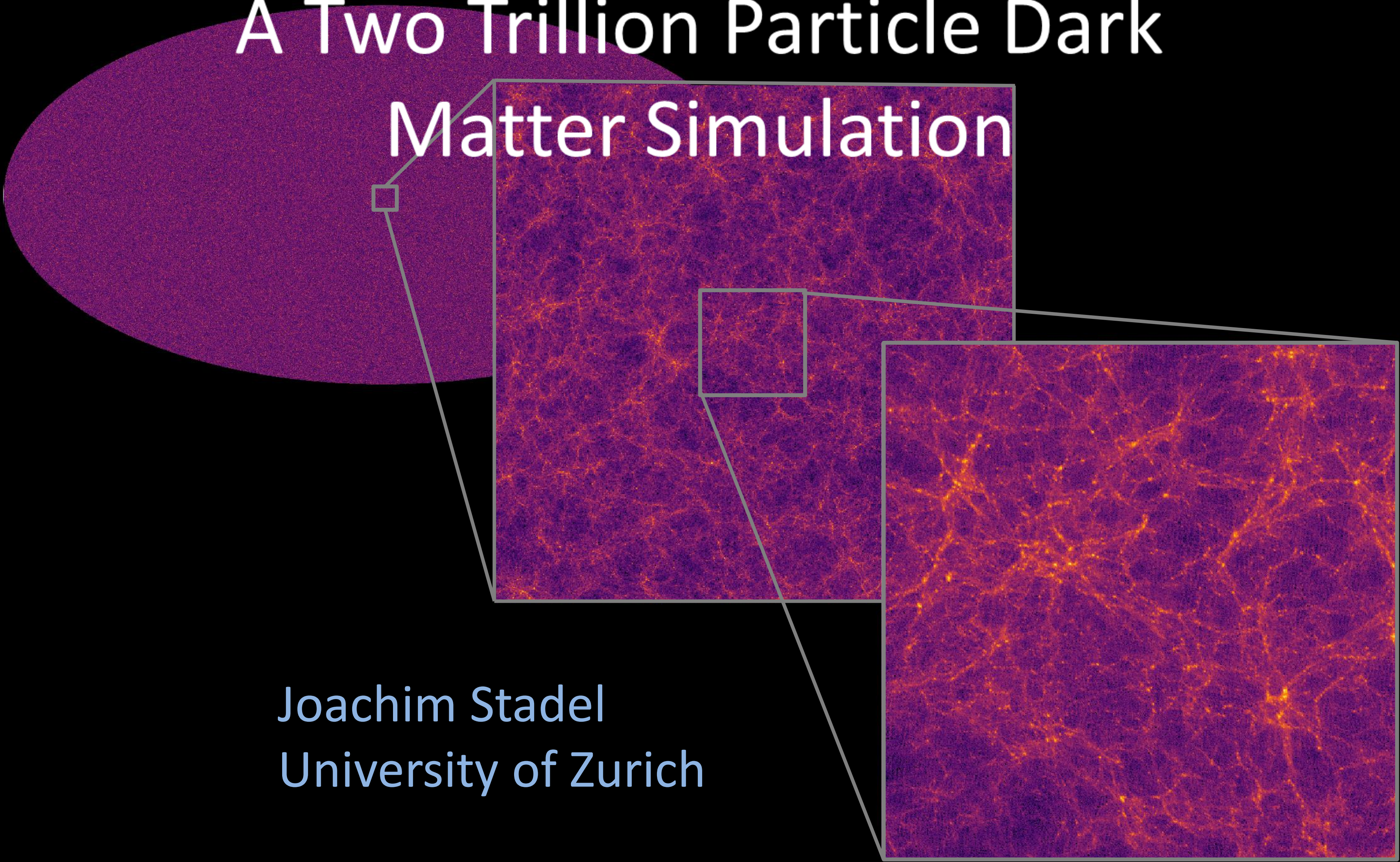
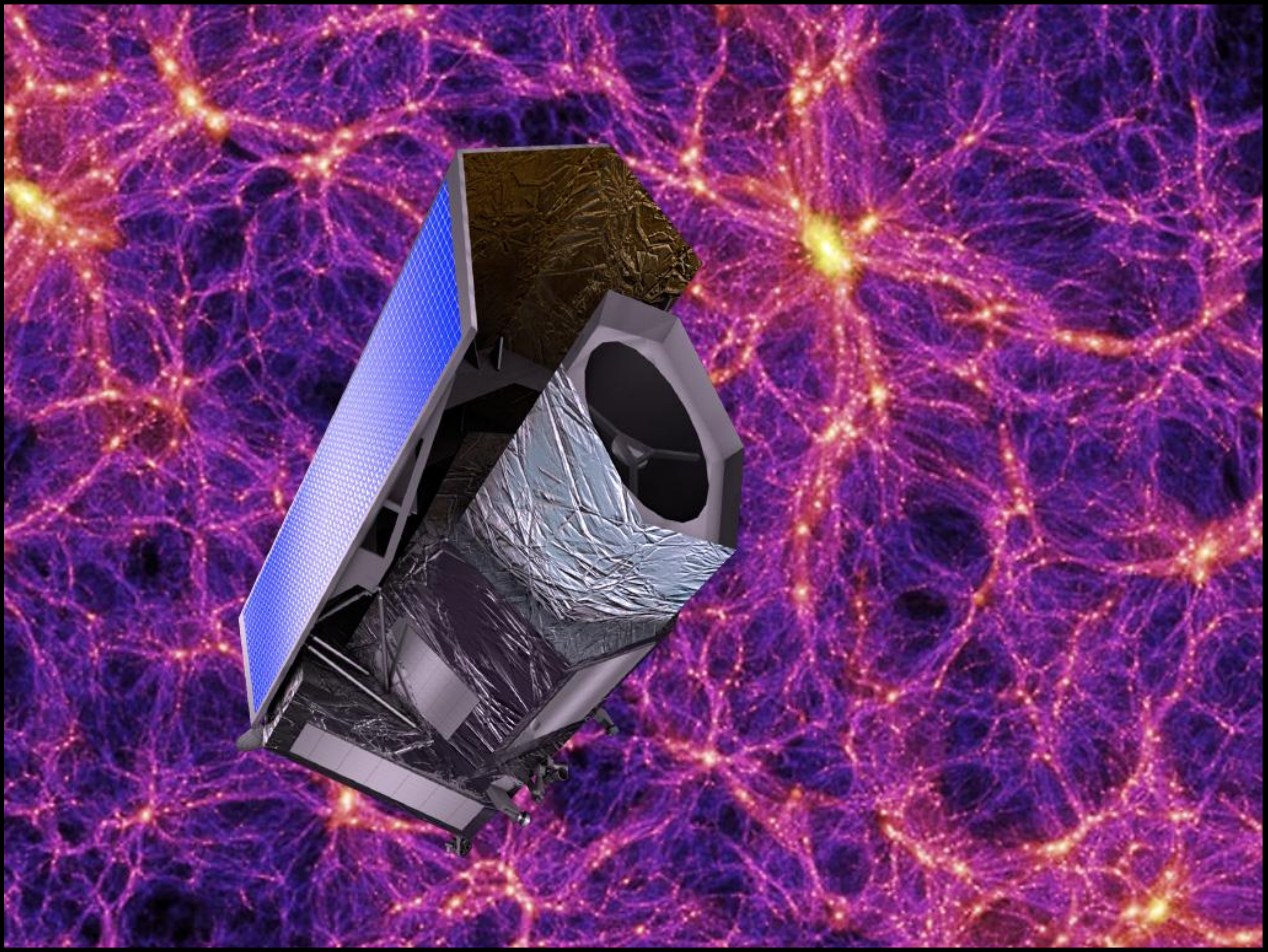
A Two Trillion Particle Dark Matter Simulation

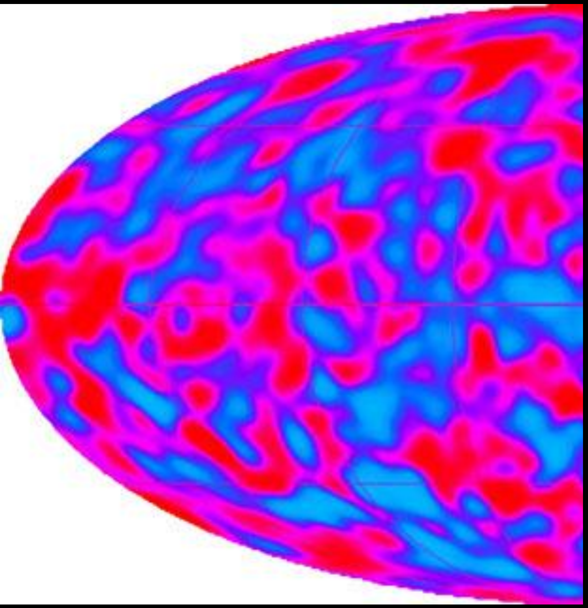


Joachim Stadel
University of Zurich

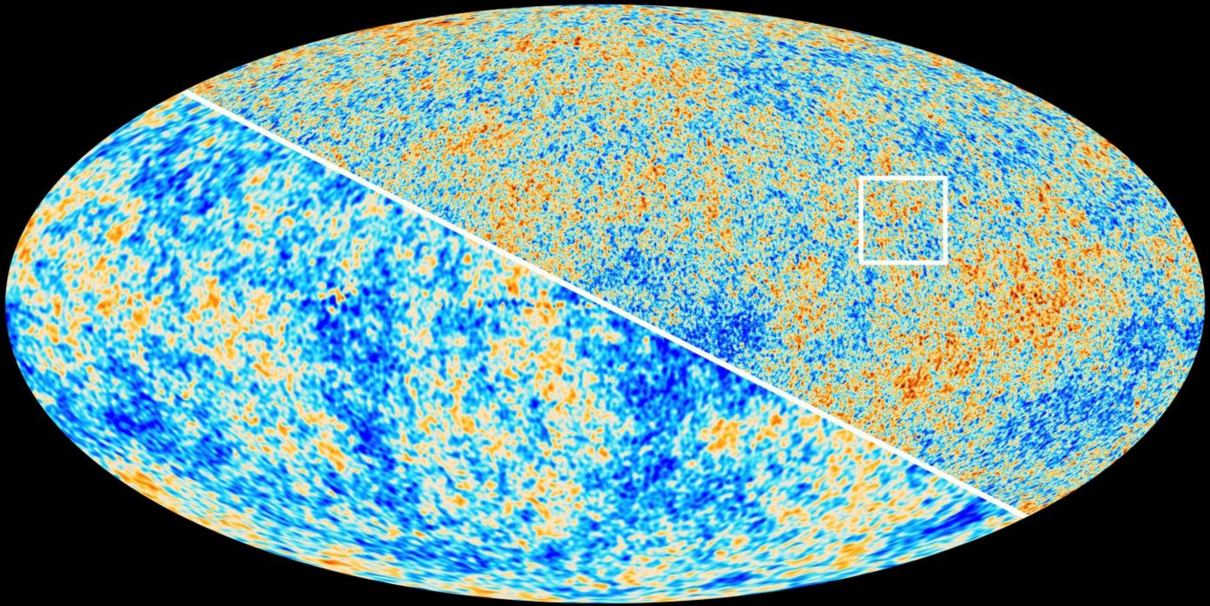
Why does Euclid have that pretty picture in the background?

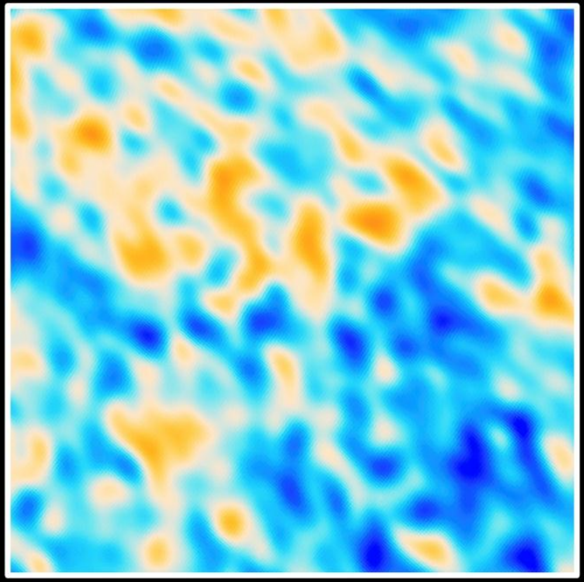


COBE

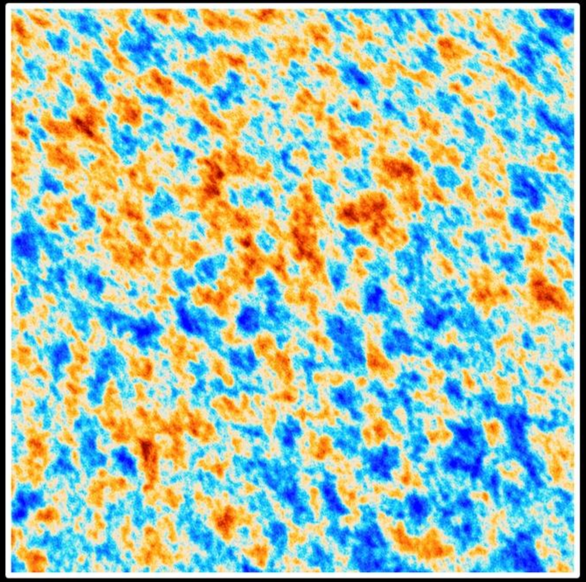


The Cosmic Microwave Background as seen by Planck and WMAP



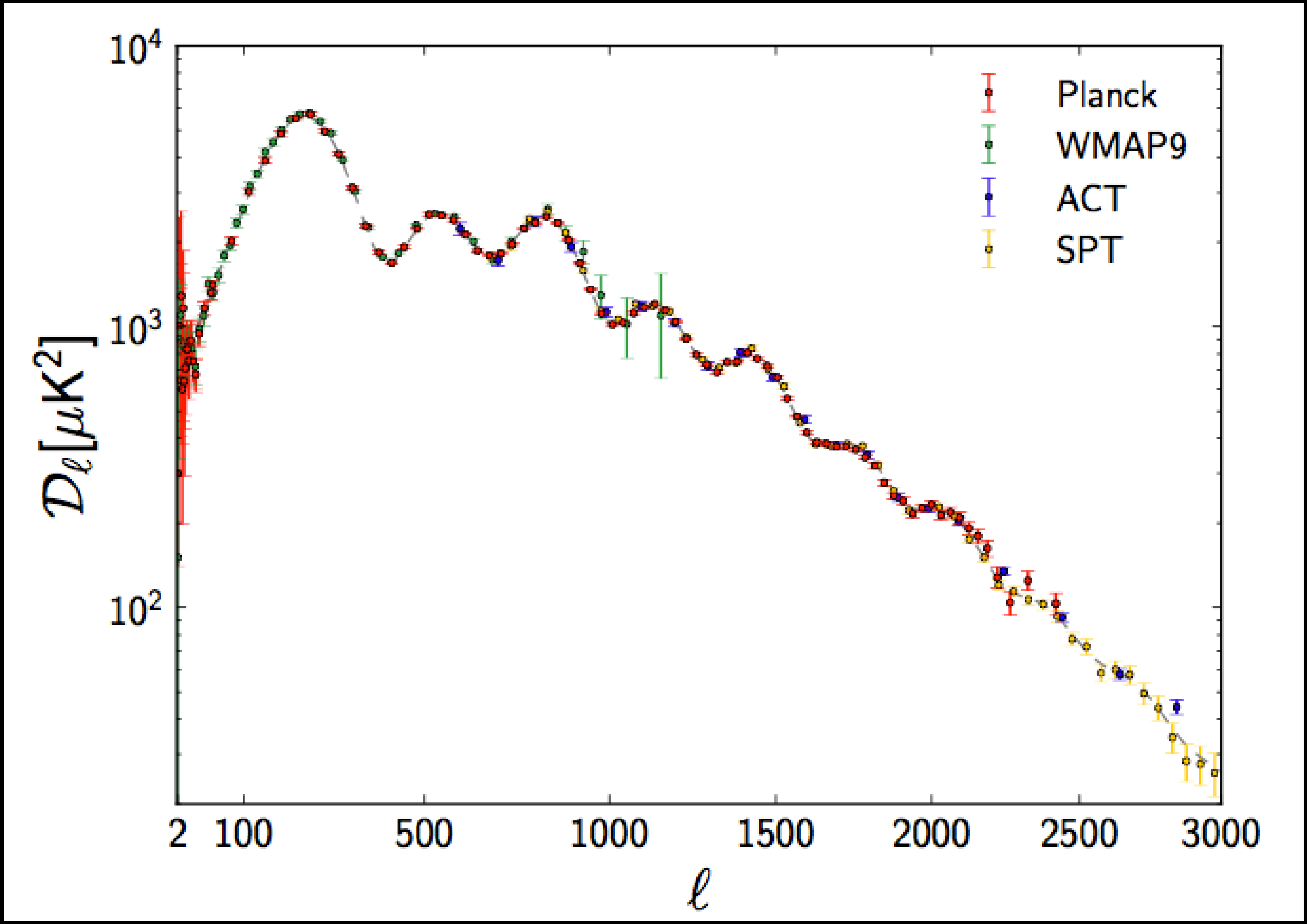


WMAP

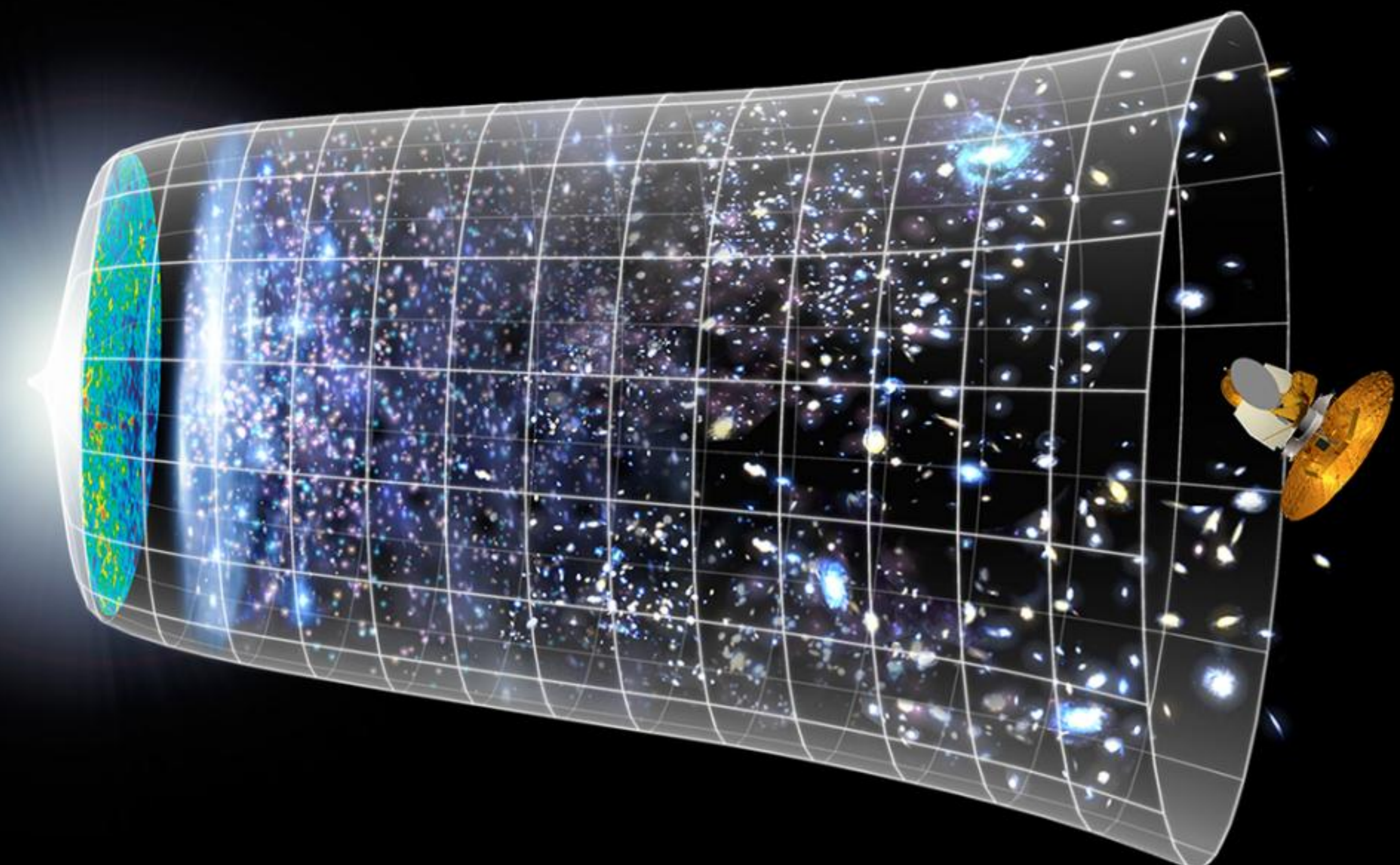


Planck

The CMB fluctuations brought in the era of *precision cosmology*

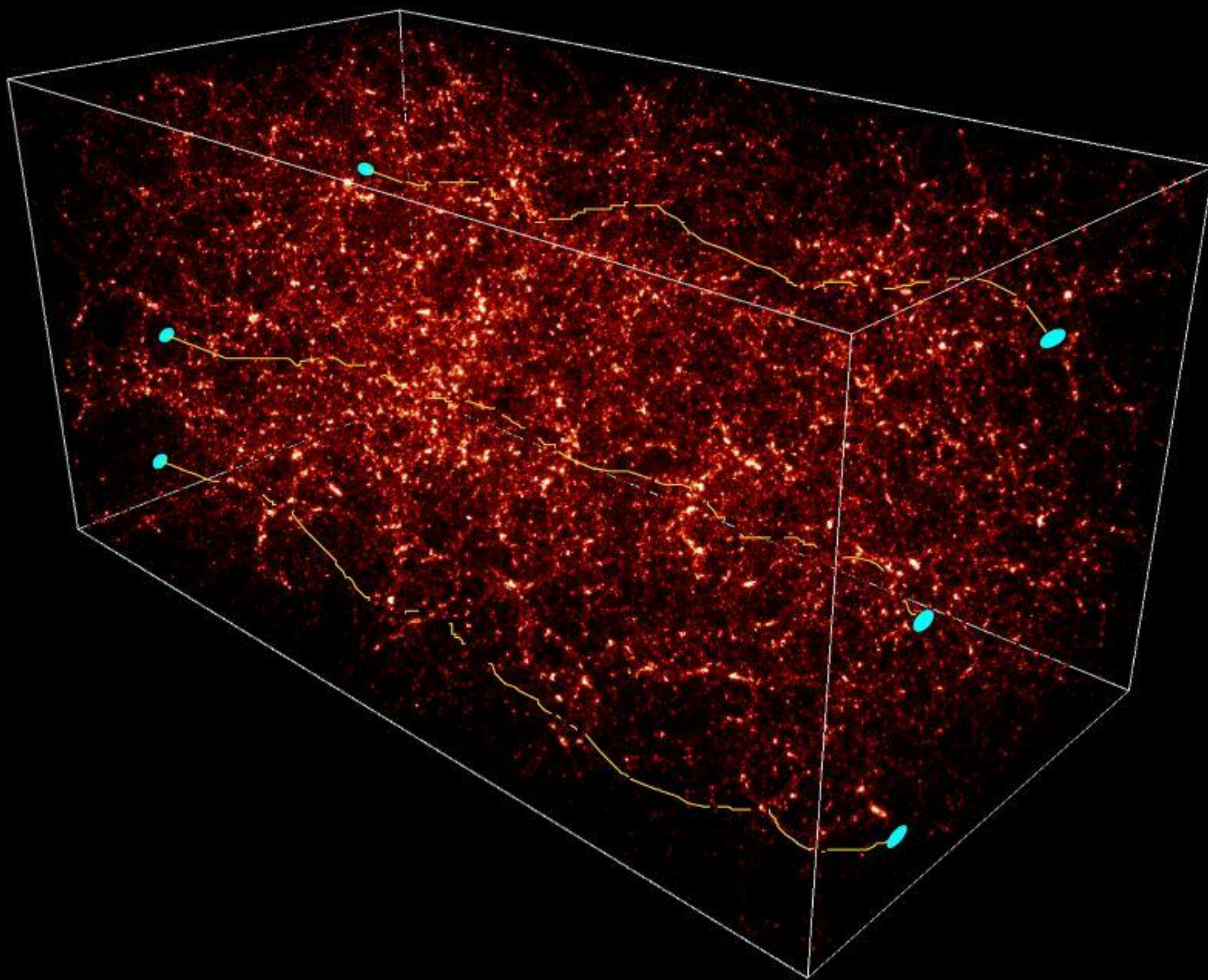


CMB fluctuations tell us about one early epoch.
(indirectly there are ways to get at other epochs too)



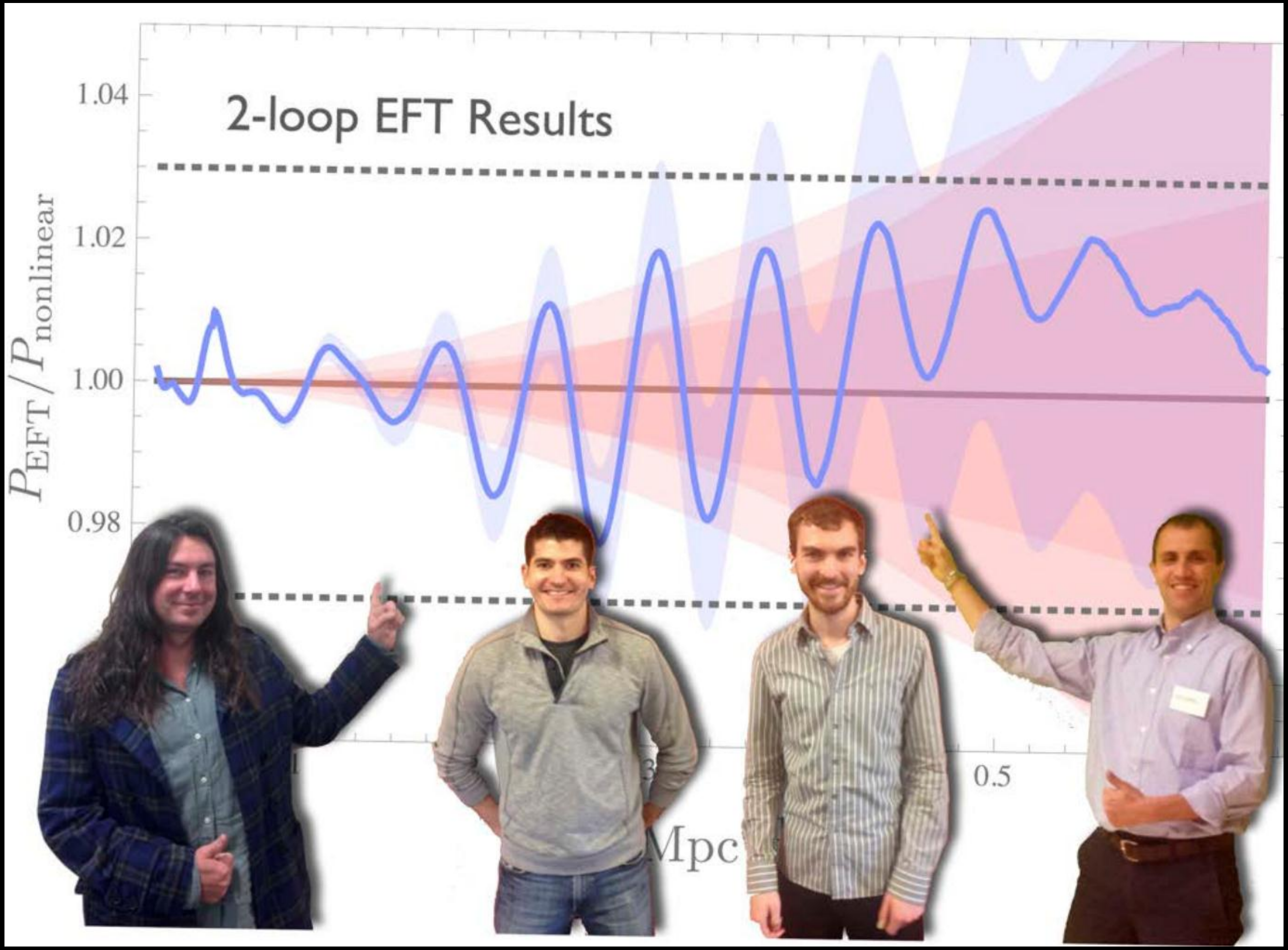
NASA/WMAP Science Team

DEFLECTION OF LIGHT RAYS CROSSING THE UNIVERSE, EMITTED BY DISTANT GALAXIES



SIMULATION: COURTESY NIC GROUP, S. COLOMBI, IAP.

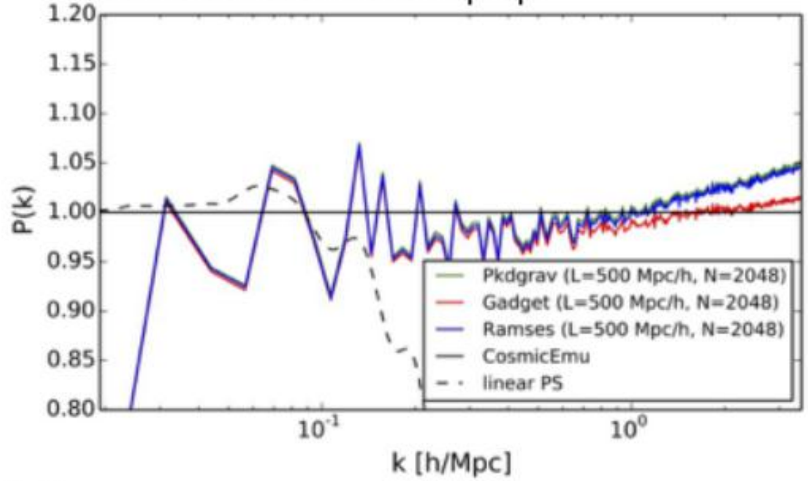
Spoilers? Effective Field Theory of LSS



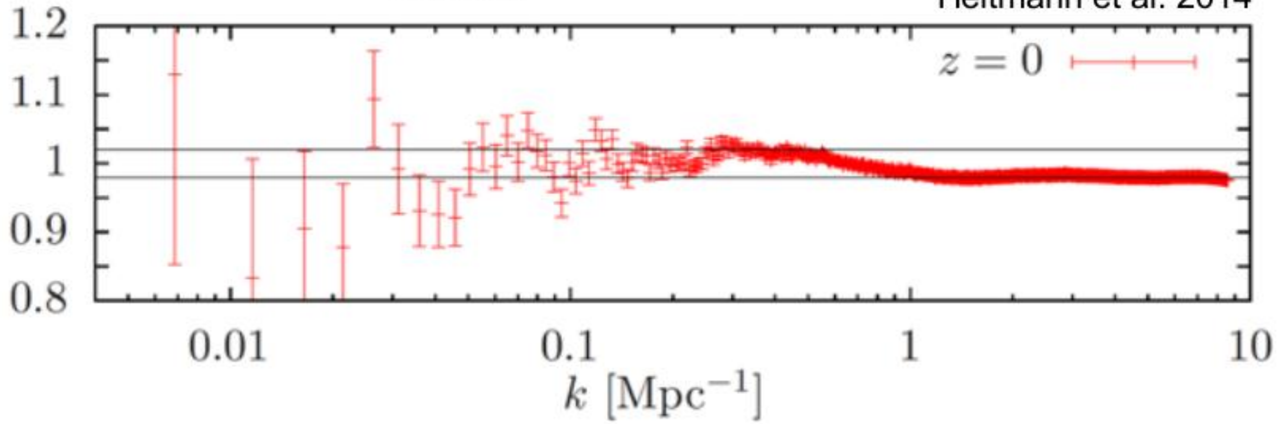
But also the Halo Model can leverage the non-linear regime!

Cosmic Emulators

Schneider et al. 2015 in prep.



Heitmann et al. 2014



Many uses for DM-only Simulations

- Directly extract some observables without worrying about 100s of pages of theory ("data" exploration).
- Creating an independent, updated, cosmic emulator.
- Serve as a check on analytic theories as well as providing input parameters (c_s and c_v in EFT).
- Assess effect of baryons on some direct observables.
- Be an input for galaxy formation theories (SAMs and Halo Occupation Statistics).
- Mocks catalogues for testing survey effectiveness and for survey design.
- Push the forefront of computational methods on the world's largest computers.

But how good are N-body simulations?

University of Zurich[UZH]

D. Potter

Halo Model

R. Smith

University Of Sussex

R. Teyssier

pkdgrav3

ICs

Ramses

Gadget3

SAMs

baryons

analysis

F. Pearce

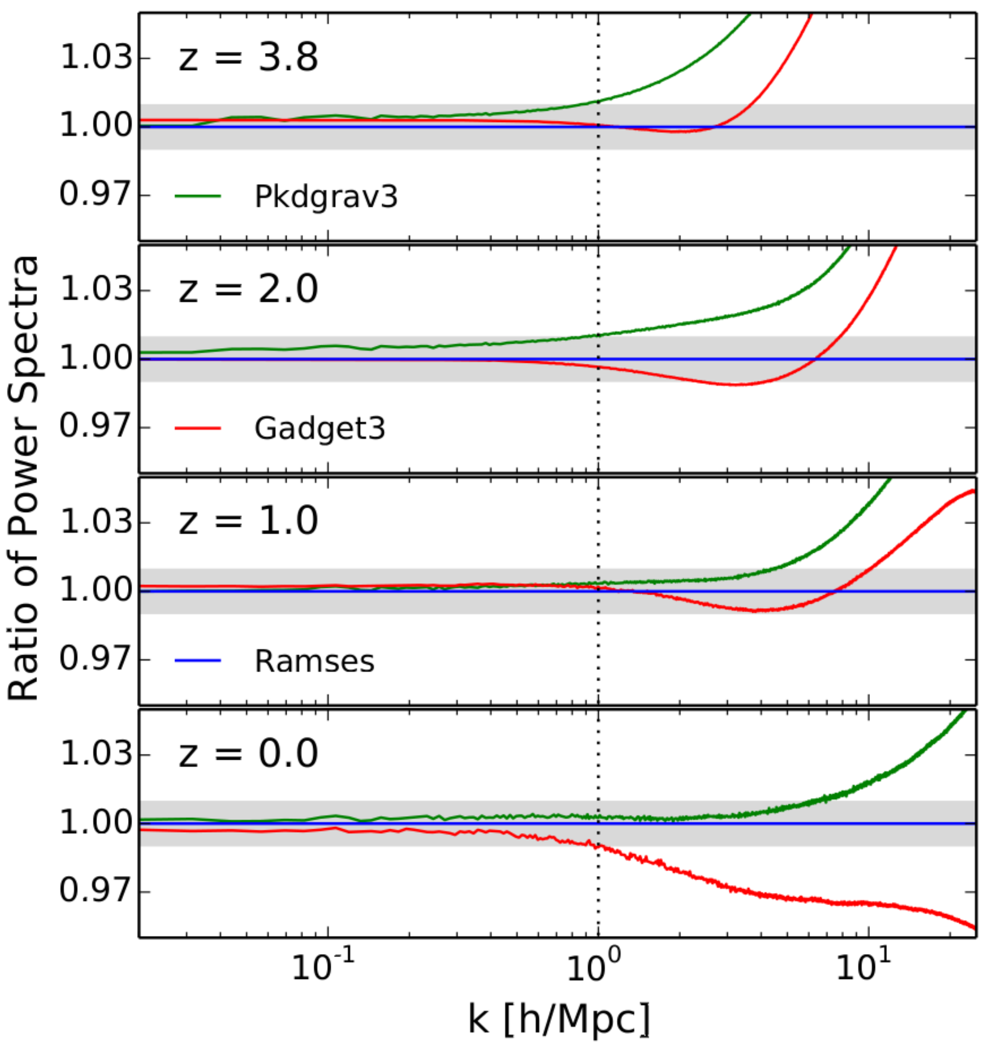
The University of Nottingham

A. Schneider

D. Reed

J. Onions

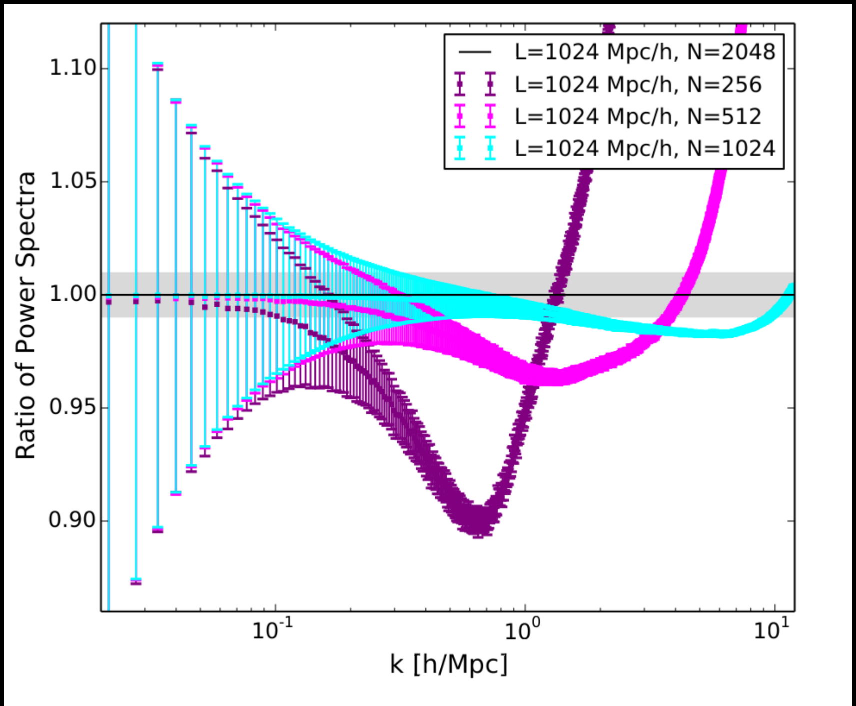
Validation by code comparison…



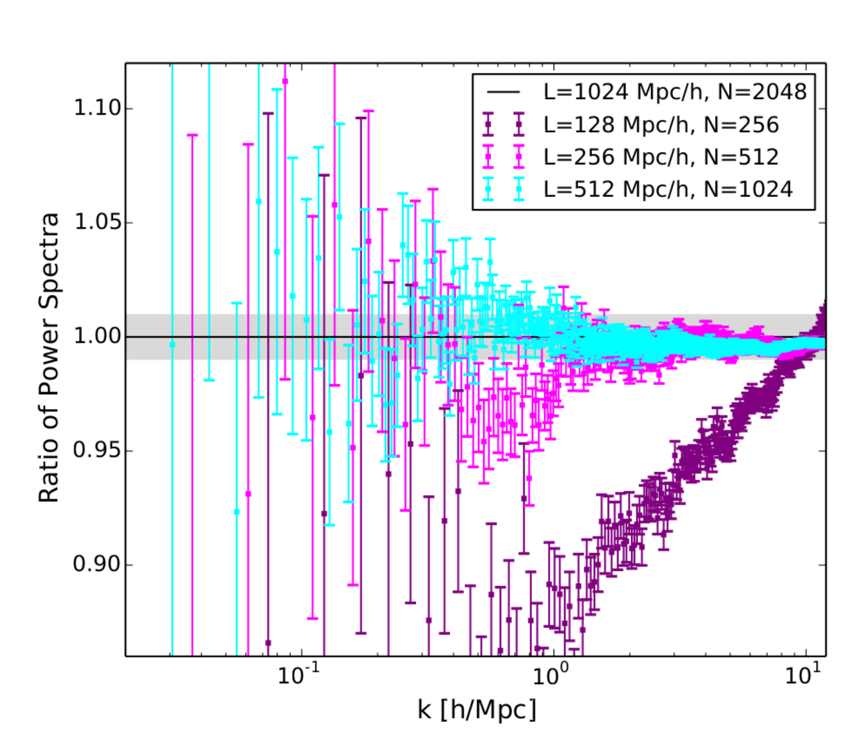
Note: all 3 codes have very different Poisson solvers and integration methods!

Quantify systematics

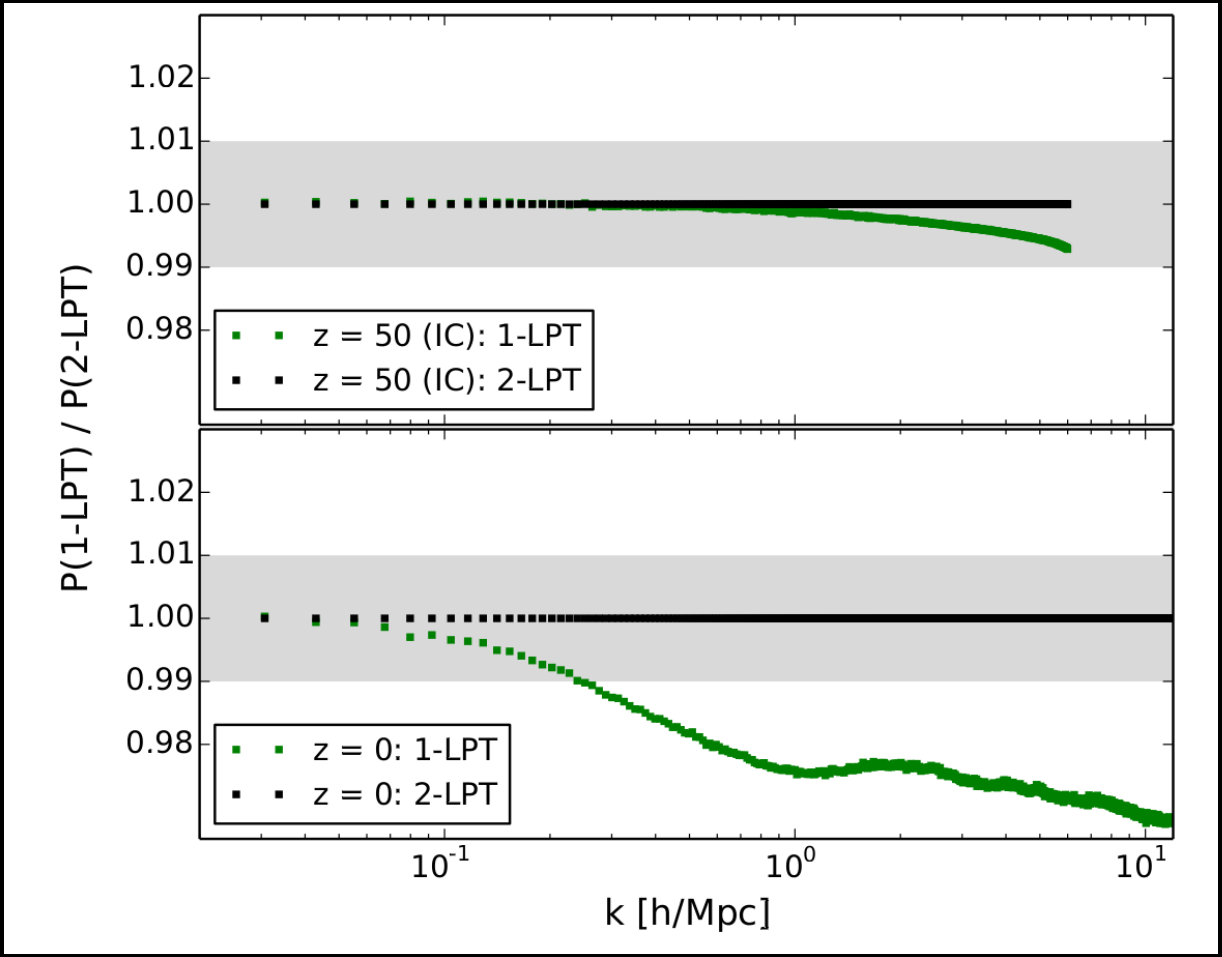
Convergence with resolution



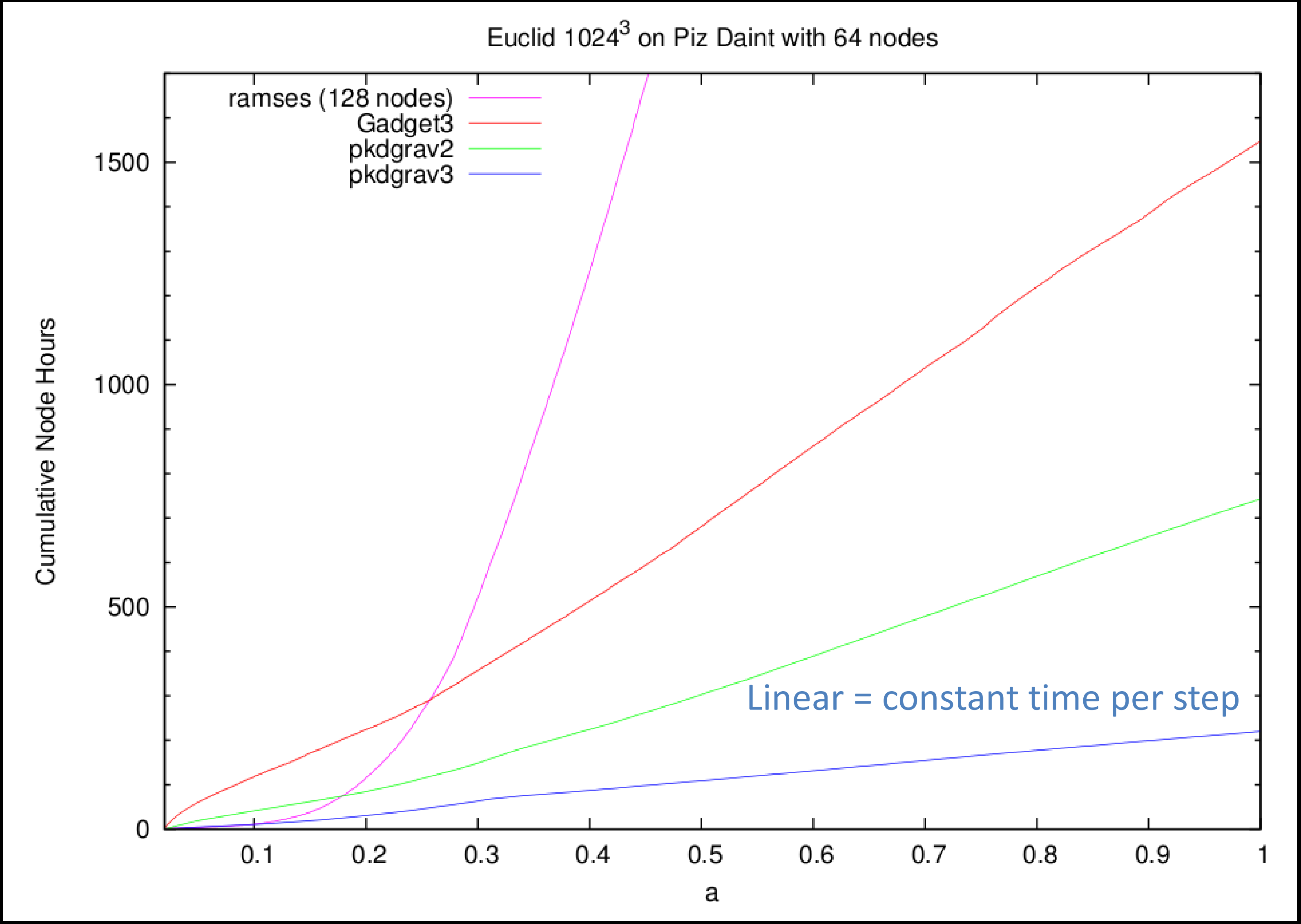
Convergence with box size



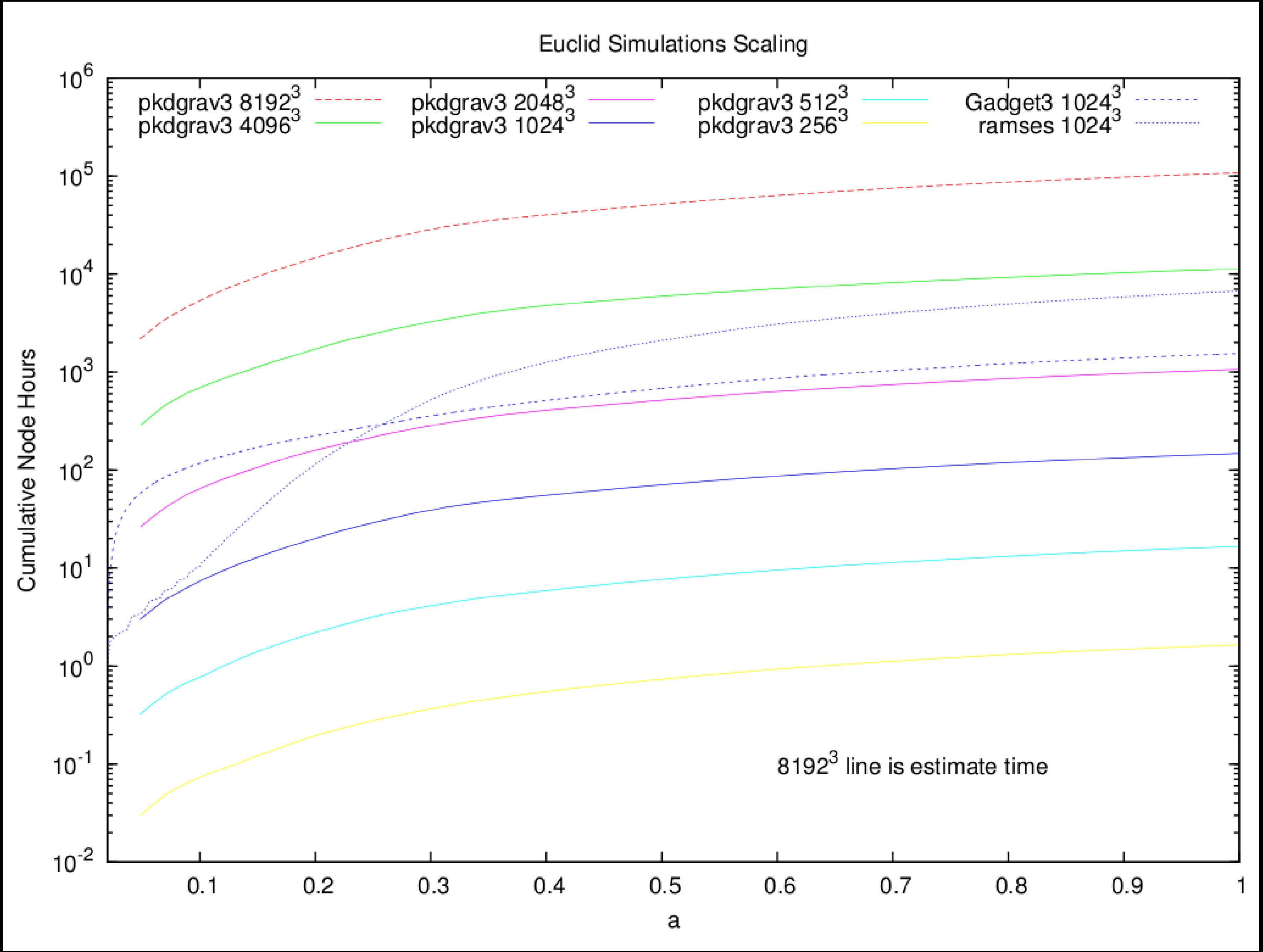
Test IC systematics

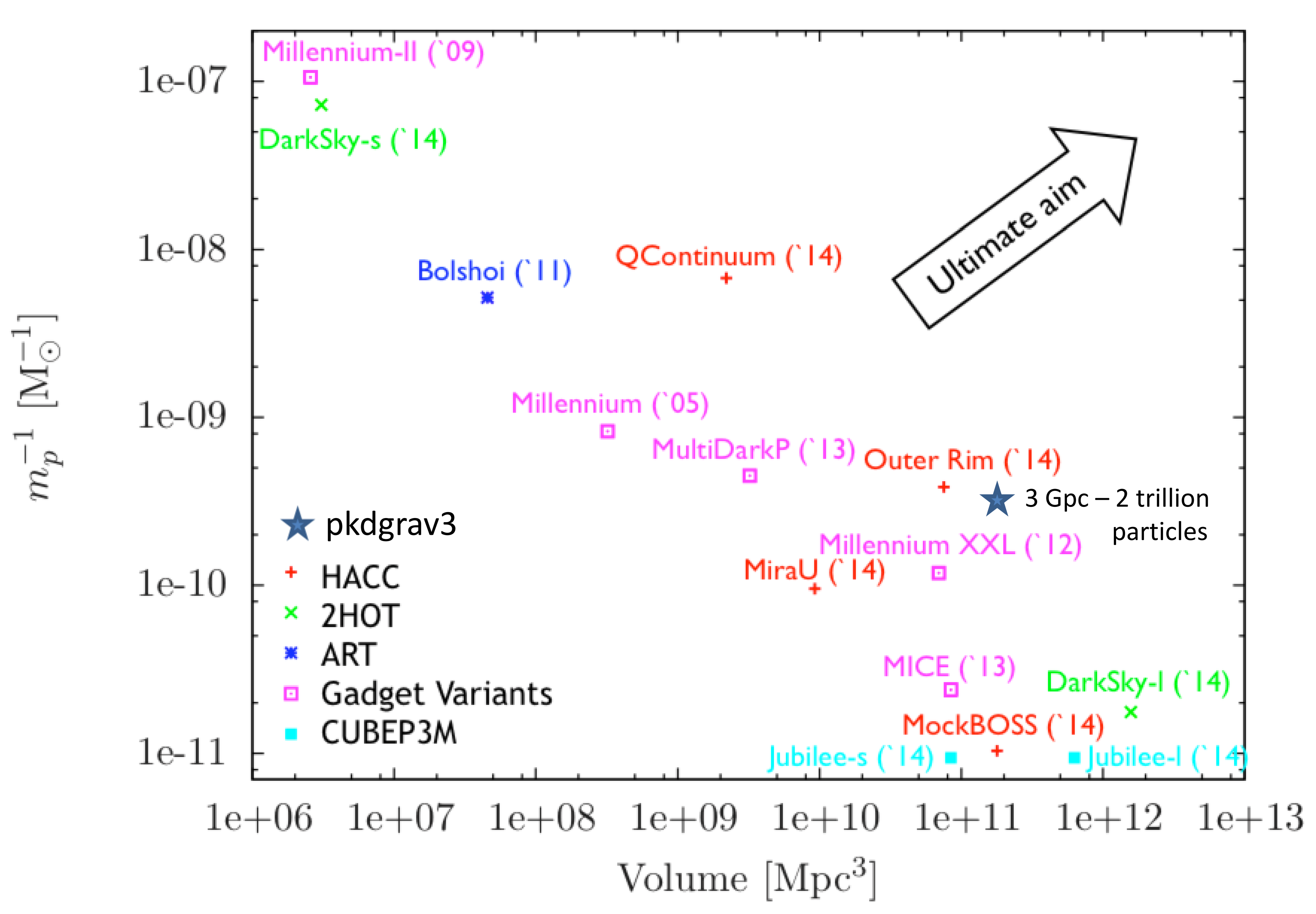


Speed of the Codes



Speed of the Codes – log scale!



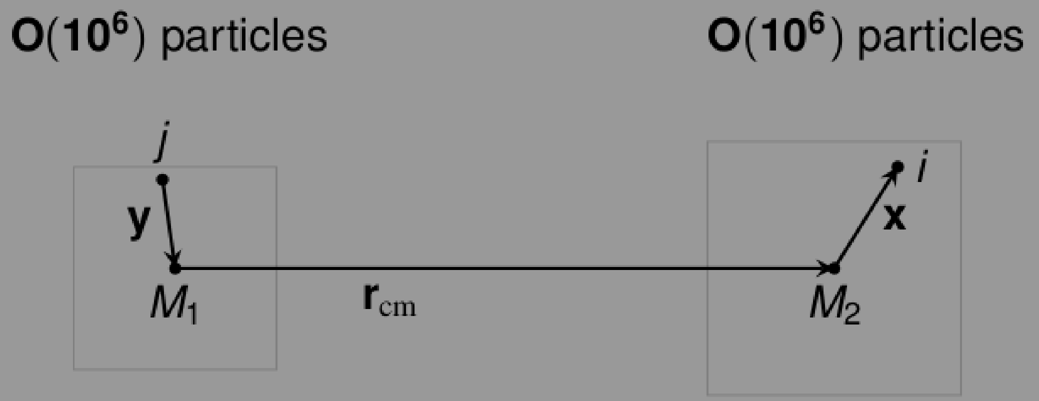

Millennium-II (`09)
DarkSky-s (`14)
Bolshoi (`11)
QContinuum (`14)
Ultimate aim
Millennium (`05)
MultiDarkP (`13)
Outer Rim (`14)
3 Gpc – 2 trillion particles
pkdgrav3
HACC
2HOT
ART
Gadget Variants
CUBEP3M
Millennium XXL (`12)
MiraU (`14)
MICE (`13)
DarkSky-l (`14)
MockBOSS (`14)
Jubilee-s (`14)
Jubilee-l (`14)
1e-07
1e-08
1e-09
1e-10
1e-11
1e+06
1e+07
1e+08
1e+09
1e+10
1e+11
1e+12
1e+13
m_p^{-1} [$M_\odot^{-1}$]
Volume [Mpc^3]

The pkdgrav3 N-Body Code

1. Fast Multipole Method, O(N), 5th order in Φ
2. GPU Acceleration (Hybrid Computing)
3. Hierarchical Block Time-Stepping
4. Dual tree gravity calculation for very active particles
5. Very efficient memory usage per particle
6. On-the-fly analysis
7. Asynchronous direct I/O for checkpoints, the light cone data and halo catalogs.
8. Available on www.pkdgrav.org (bitbucket.org)

pkdgrav3 and Fast Multipole

Quick explanation of FMM



Direct **O**(**10**^{12}) interactions to calculate! $O(N^2)$ code.

Tree Use a multipole approximation for the mass at M_2 to calculate the force at each j: **O**(**10**^{6}) interactions to calculate. $O(N \log N)$ code.

FMM Use a multipole approx for the mass at M_2 to approximate the "potential landscape" at M_1 (n^{th} order gradients of the potential): **O**(**1**) interaction to calculate. $O(N)$ code!

Data Locality in pkdgrav3

- Note that as we proceed deeper in the tree, the data we need to fetch becomes ever more local! As long as data is stored in a kind of “tree order”.
- This is what makes FMM very efficient on systems with many cache levels in the memory hierarchy (slowest: off-node mem).
- FMM algorithm achieves a kind of minimal amount of data movement within the entire computing architecture.
- The periodic BCs are handled by multipole Ewald summation technique. Instead of 4 transposes, 3 FFTs, 3 IFFTs, a single independent (GPU) calculation for each particle is done.

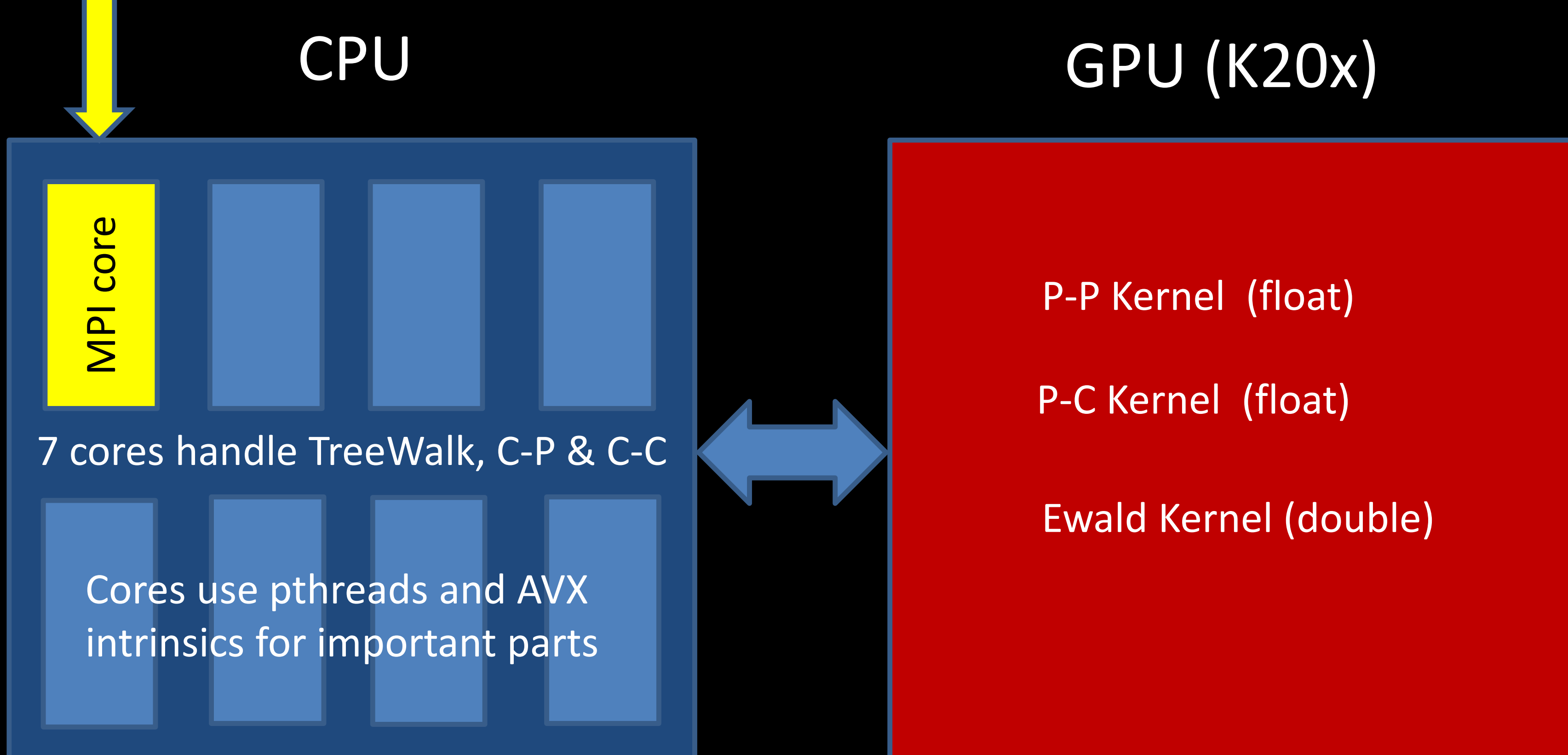
Piz Daint – over 5000 GPU Nodes

6^{th} Fastest Computer in the World. Upgrade to Haswell & P100 (now)…

GPU Hybrid Computing

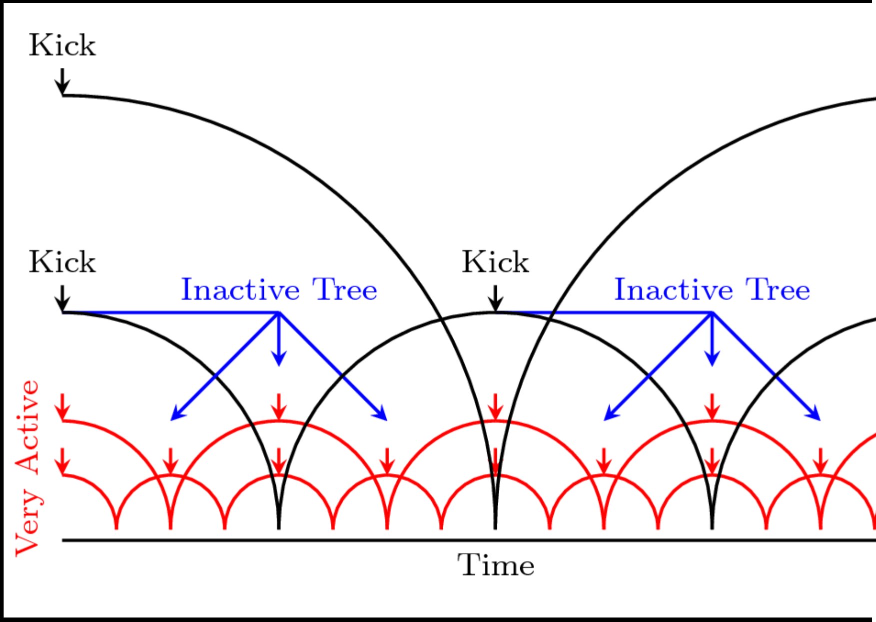
Piz Daint example

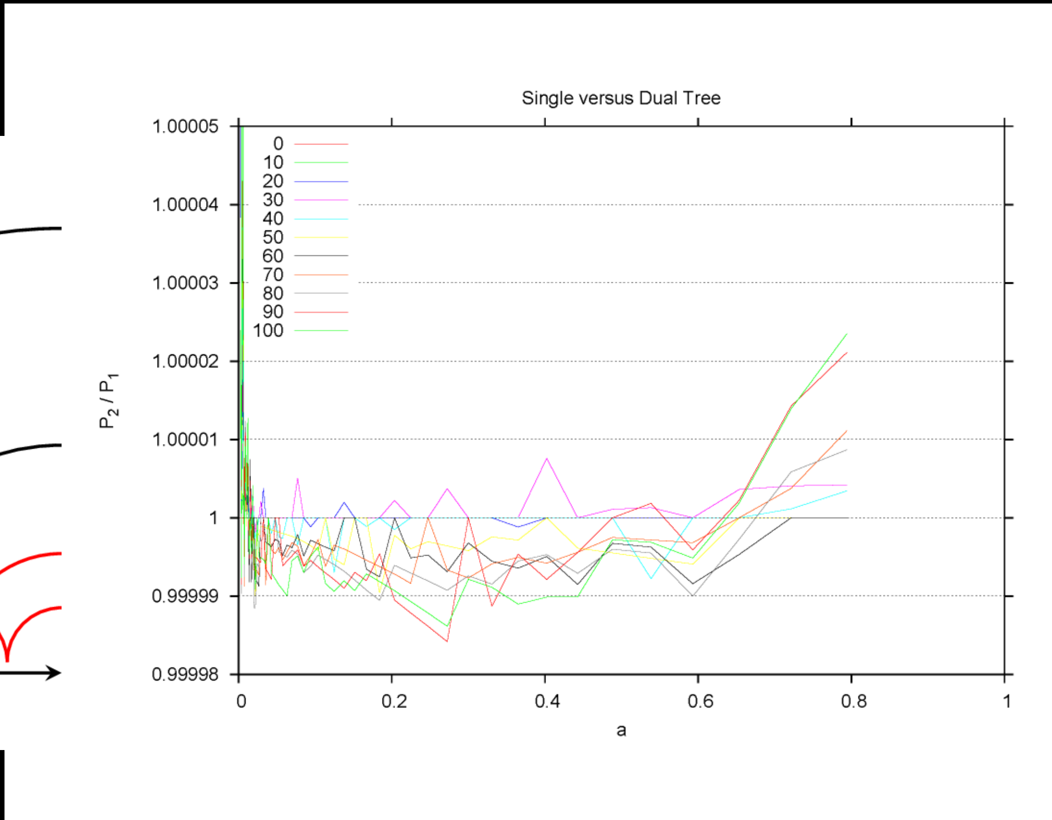
We don’t directly use MPI or pthreads in the main code, but our own MDL library which implements a software cache to access data on remote nodes or threads.



Dual Tree Implementation

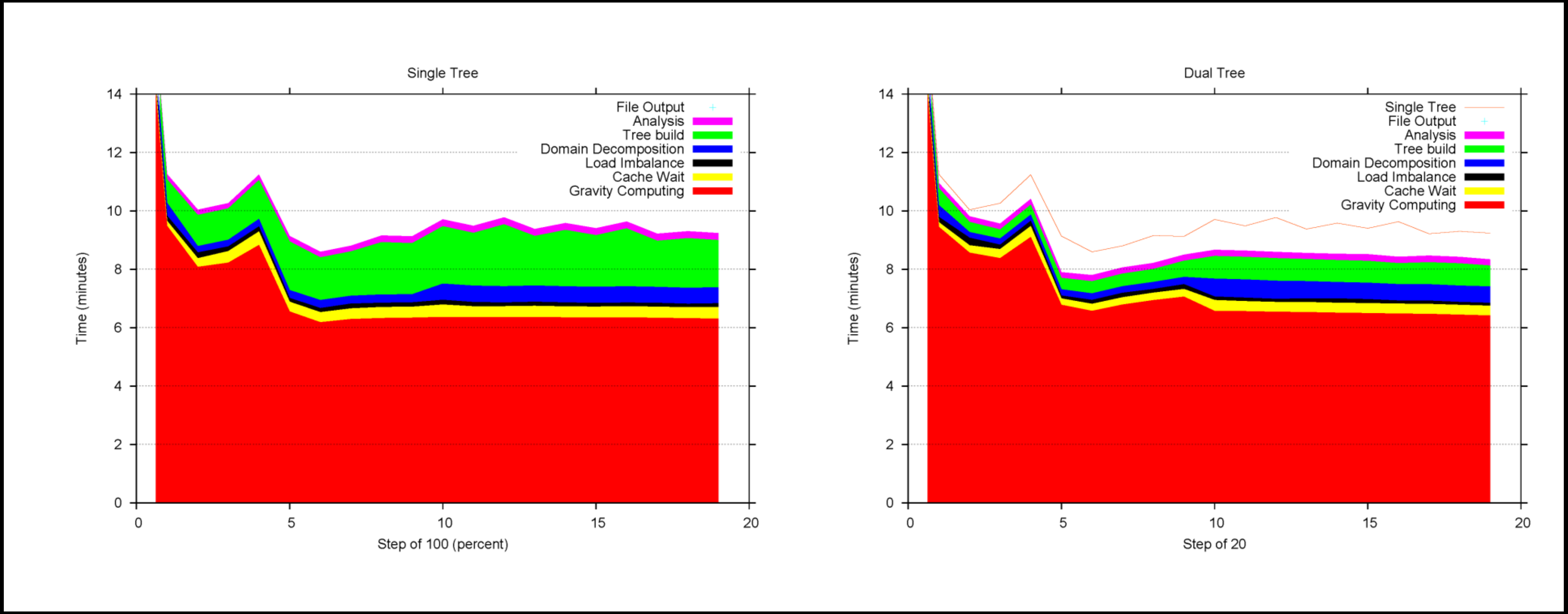
- We create a fixed tree of all particles which are on longer timescales.
- This *fixed* inactive tree is built **time centered** for the very active time-steps.
- Both trees are walked to obtain the force.
- Typically we define the very active rungs as <5% of the most active particles.





Dual Tree Performance

Note: without GPU here!



Memory Usage in pkdgrav3

0.5 billion particles can fit on a 32 Gbyte Node like Piz Daint

28 bytes persistent

6 bits: old rung 24: group id	
pos[0]	int32_t
pos[1]	int32_t
pos[2]	int32_t
vel[0]	float
vel[1]	float
vel[2]	float

<28 bytes / particle

Tree Cells
Binary Tree

4th order
Multipoles
(float prec)

~5 bytes / particle

Cache/Buffers

0-8 bytes ephemeral

Group finding
Other analysis

ClAoS is used for the particle and cell memory which makes moving particles around simple

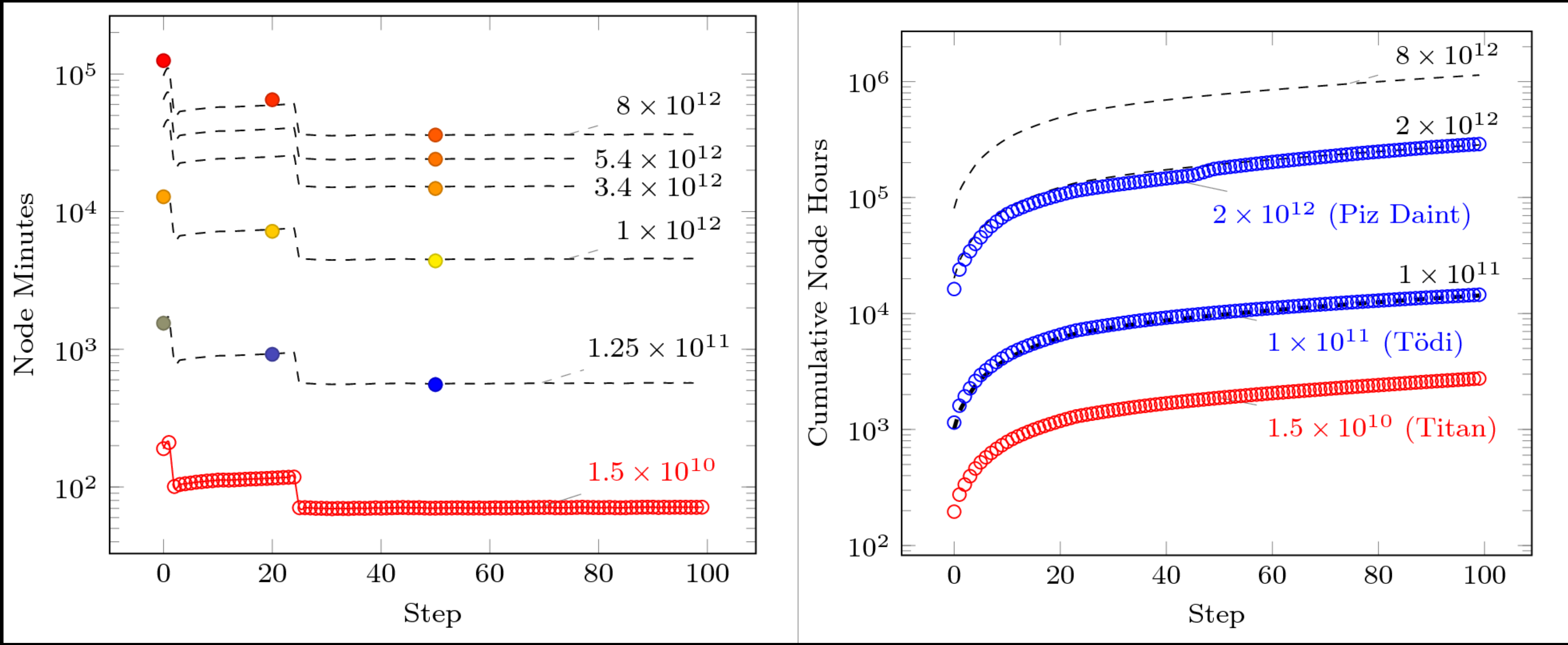
AoSoA is used for all interaction lists which are built by the TreeWalk algorithm.

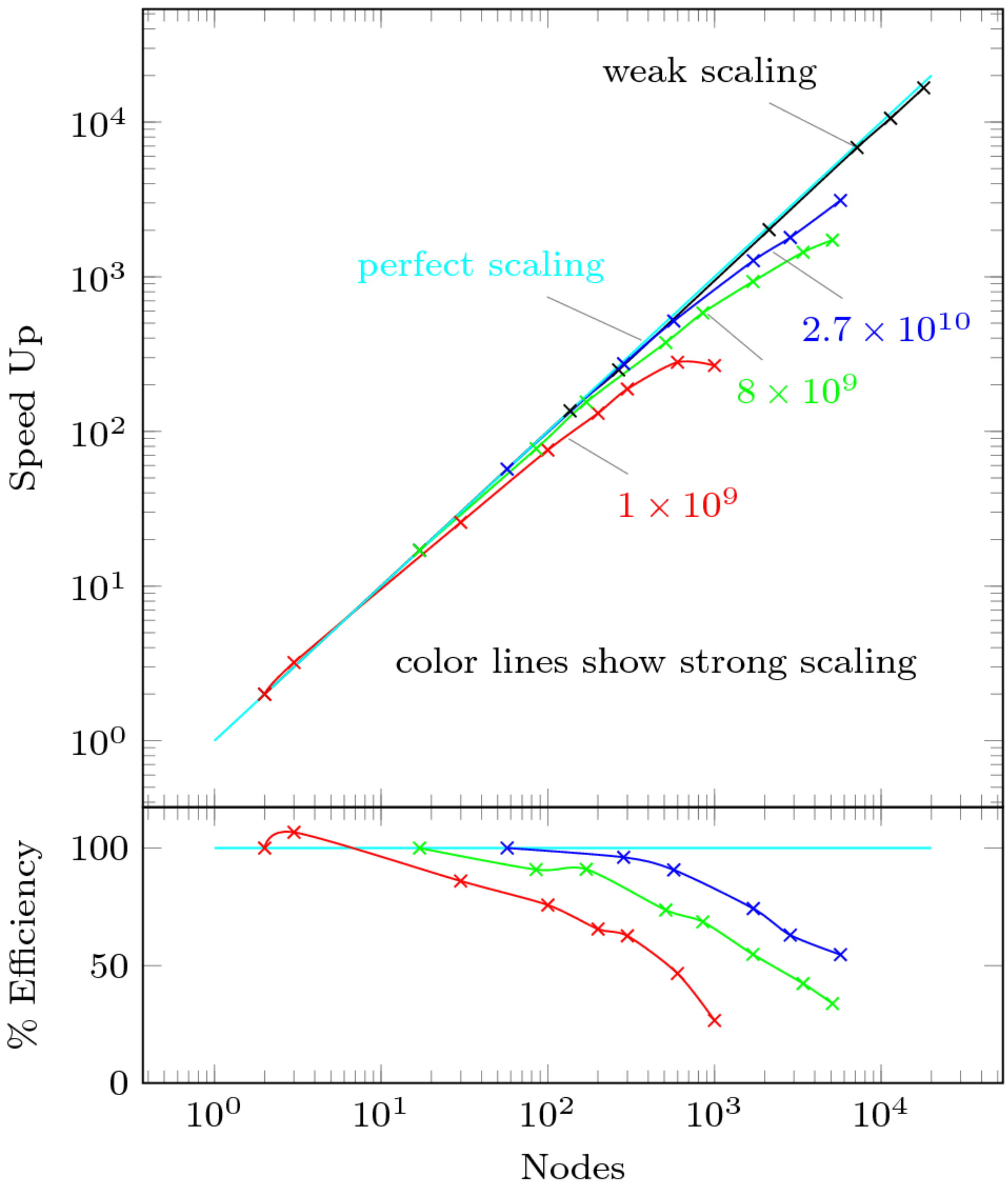
Reducing memory usage increases the capability of existing machines, but also increases performance somewhat. Simulations are limited more by memory footprint.

Benchmarking on Titan and Piz Daint

Nearly Perfect Weak Scaling makes performance prediction very accurate for these simulations.
120 seconds for an all N gravity solve!

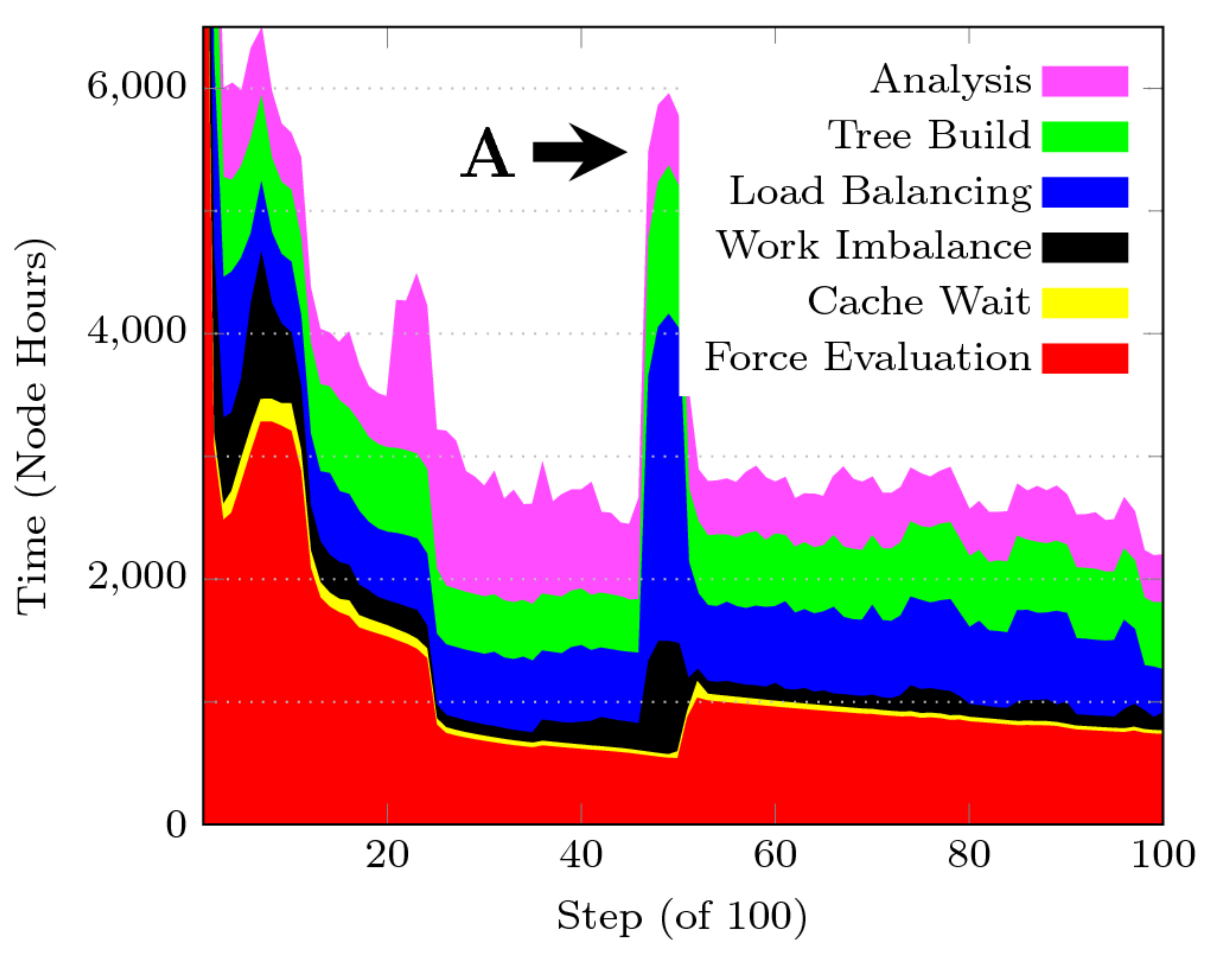
We show that it is quite feasible to run 8 trillion particles on Titan with a little over 1 million node hours. **10 PFlops**



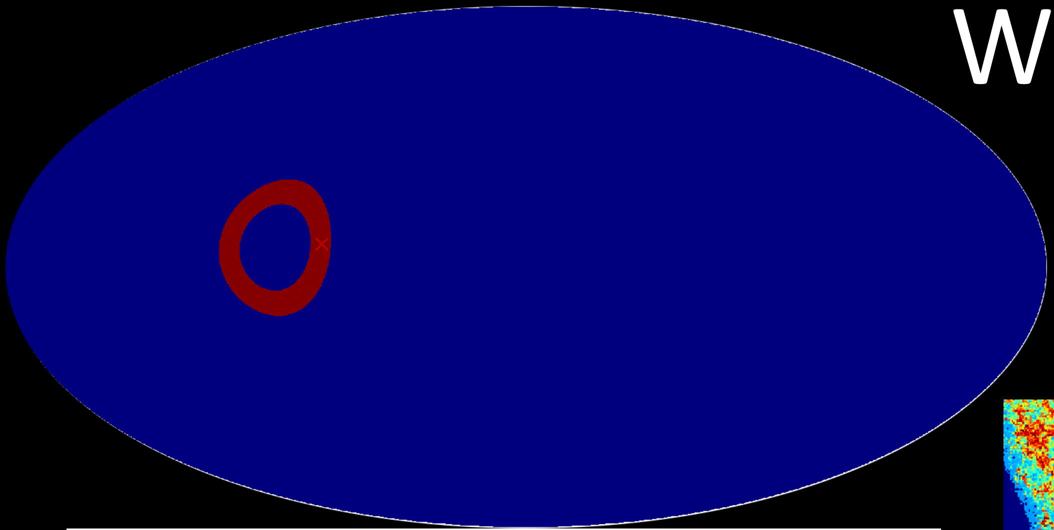


weak scaling
perfect scaling
2.7×10^{10}
8×10^{9}
1×10^{9}
color lines show strong scaling
Speed Up
% Efficiency
Nodes

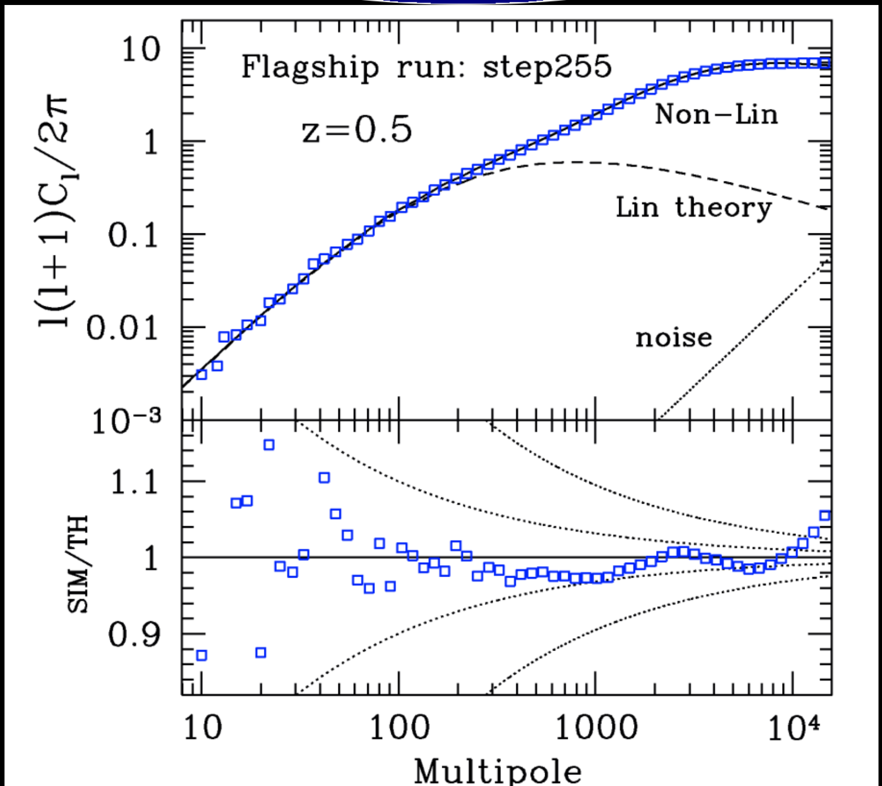
Profile of the 2 trillion particle production simulation (Piz Daint)

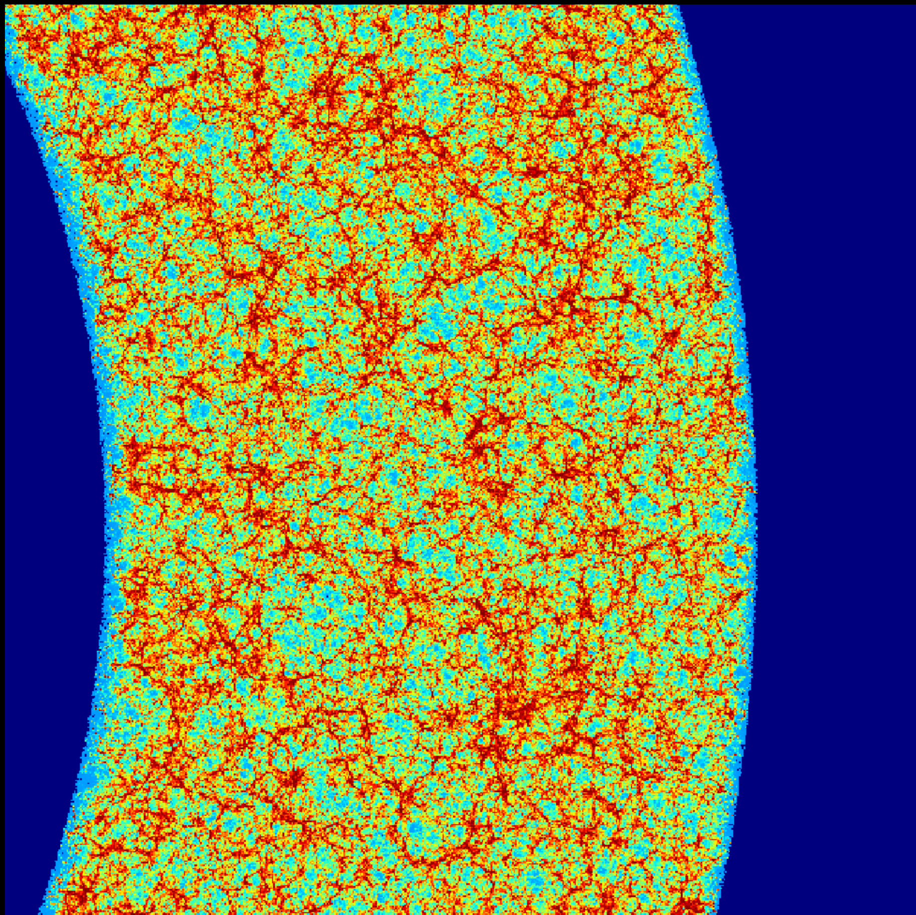


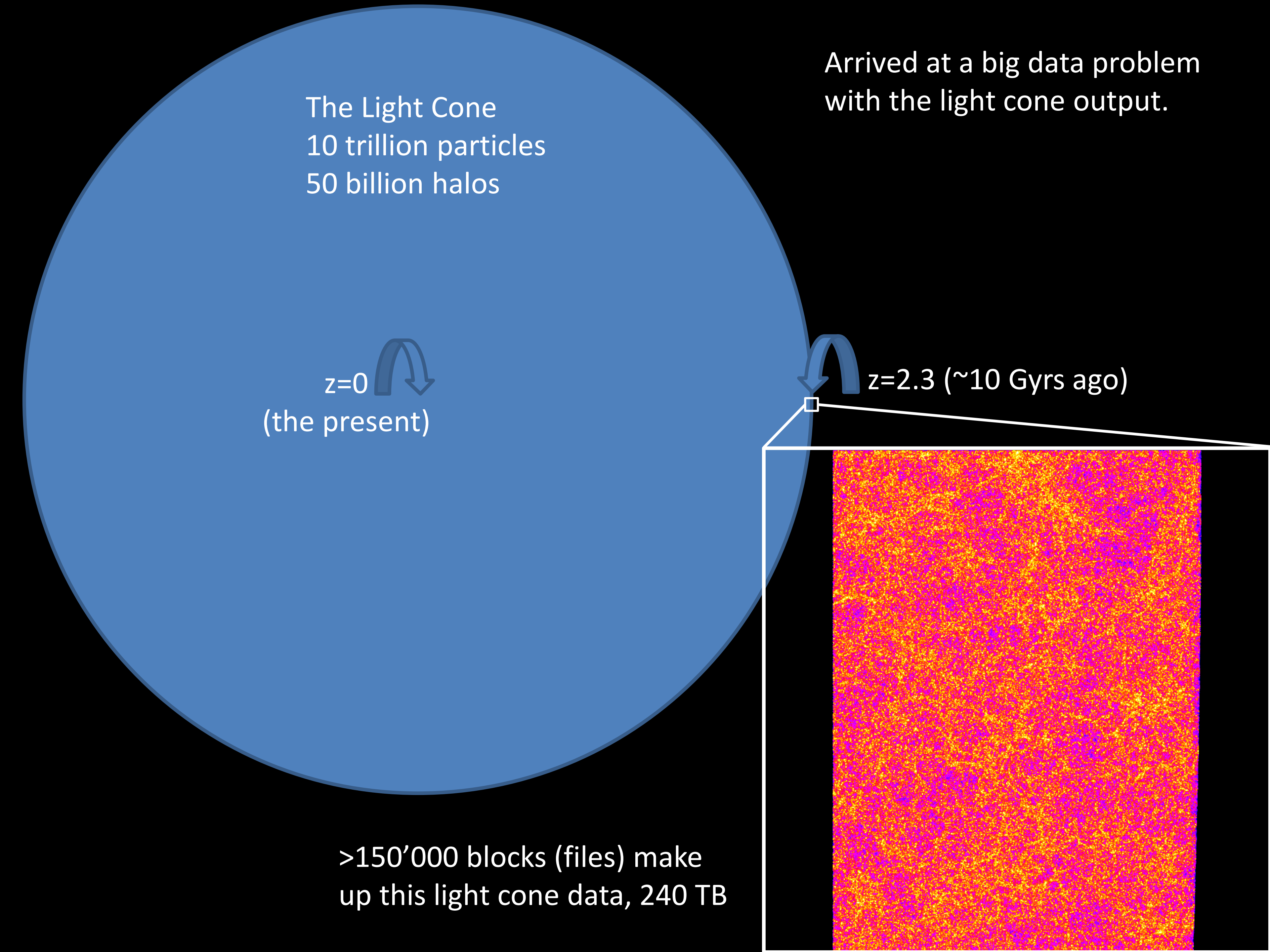
Weak lensing maps

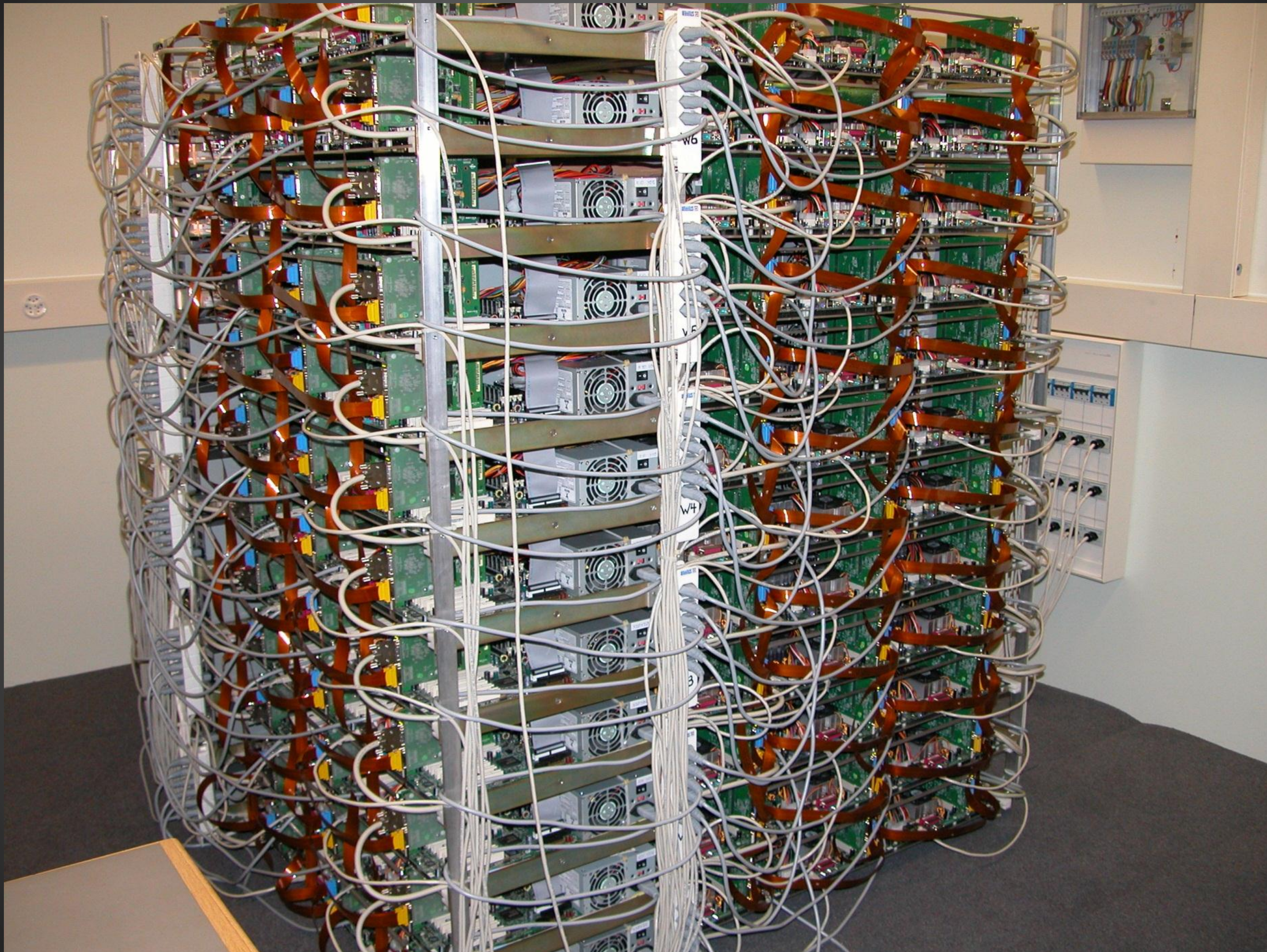


There are O(400) such maps which form a set of spherical, concentric, lensing planes to distort the shapes of the background galaxies.






Arrived at a big data problem with the light cone output.
The Light Cone
10 trillion particles
50 billion halos
z=0
(the present)
z=2.3 (~10 Gyrs ago)
>150'000 blocks (files) make up this light cone data, 240 TB



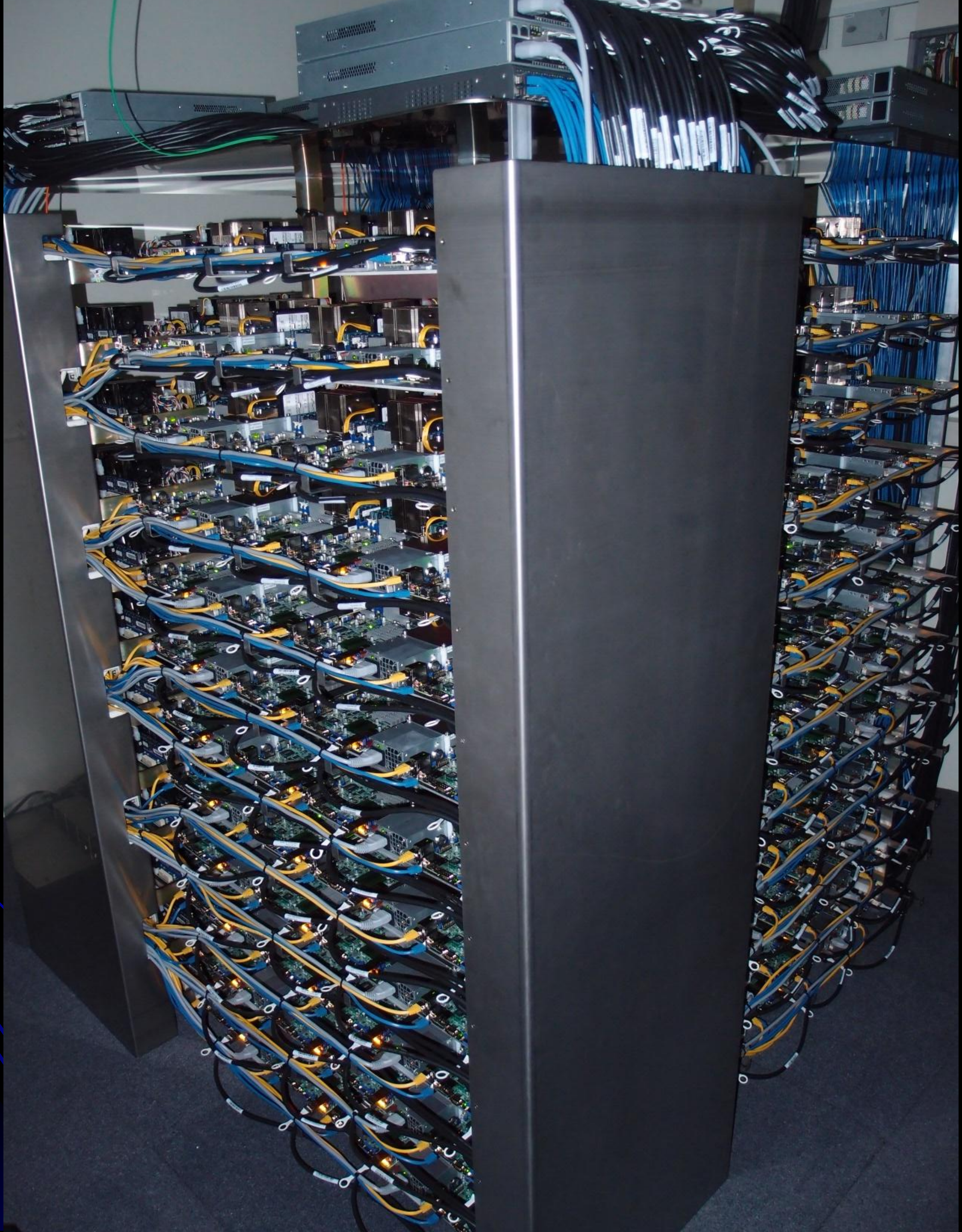
zBox4

3200 cores, 3 GHz
14 TB of RAM

> 2 PB Disk
25 TB Local SSD

2 Tb/s X-section Bandwidth

45 kW



2012

zBox4

2016

3200 cores, 3 GHz
14 TB of RAM

> 2 PB Disk
25 TB Local SSD

2 Tb/s X-section Bandwidth

45 kW

upgrade

+ 192x GTX 950

+ 800 TB Local HD

+ 15 kW

3x pkdgrav3 speedup

The future?

- Will we pursue even bigger cosmological simulations? Likely, but the priority is improving the physics at this resolution.
- Creating a new emulators will require a large number of (somewhat) smaller simulations – cheap high throughput computing!
- New *analysis intruments* to develop so that we don't have to do large scale I/O anymore!
- EUCLID launch 2020! Dark Energy and Inflation