Argonne NATIONAL LABORATORY

Blasting Through the 10 Petaflops Barrier: HACC on the BG/Q

HACC (Hardware/Hybrid Accelerated Cosmology Code) Framework

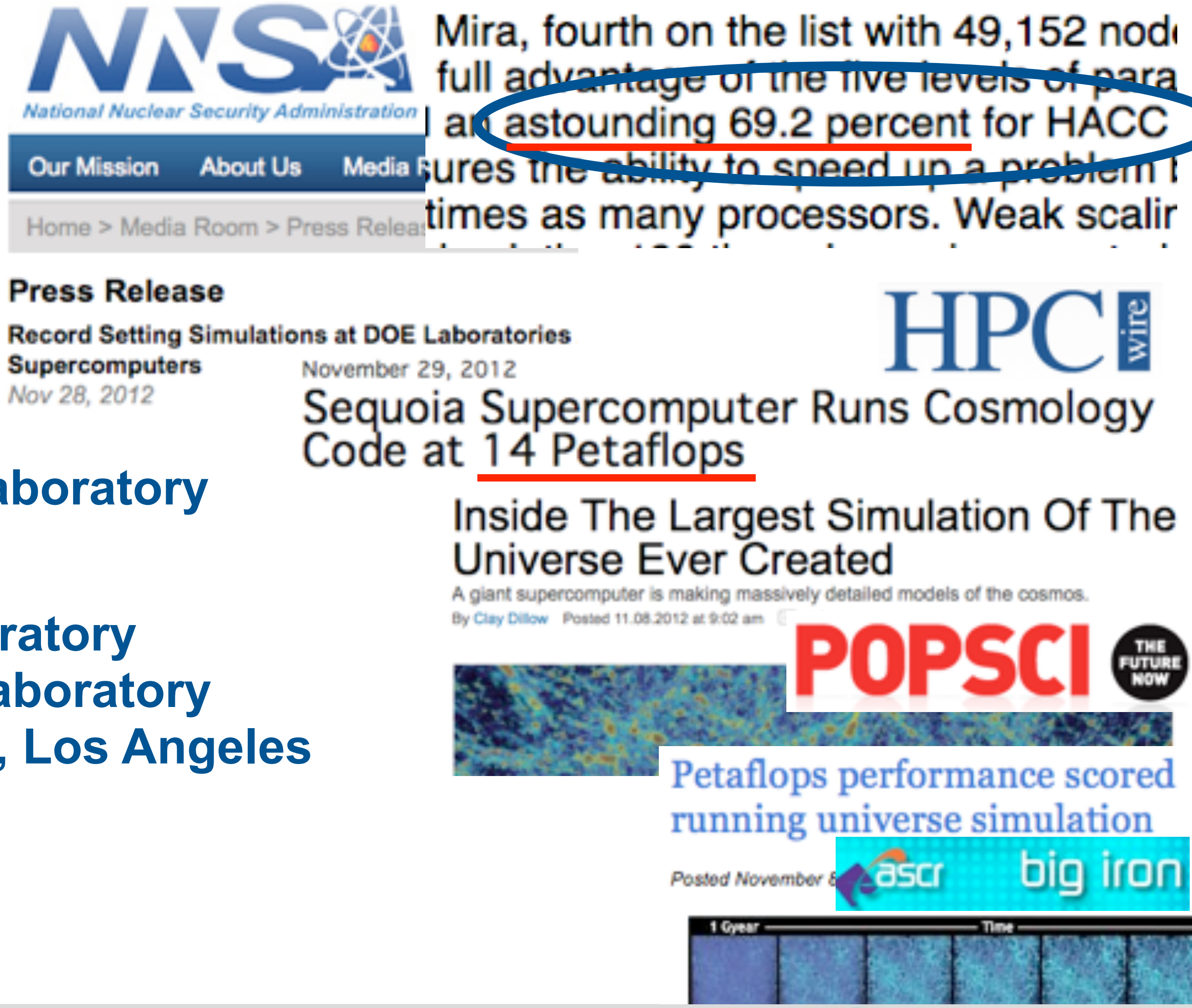
Salman Habib
HEP and MCS Divisions,
Argonne National Laboratory

Vitali Morozov
Hal Finkel
Adrian Pope
Katrin Heitmann
Kalyan Kumaran
Tom Peterka
Joe Insley
Venkat Vishwanath
Argonne National Laboratory

Zarija Lukic
Lawrence Berkeley National Laboratory

David Daniel
Patricia Fasel
Los Alamos National Laboratory

Nicholas Frontiere
Argonne National Laboratory
Los Alamos National Laboratory
University of California, Los Angeles


NNSA National Nuclear Security Administration

Our Mission About Us Media

Home > Media Room > Press Releas

Mira, fourth on the list with 49,152 nod
full advantage of the five levels of para
an astounding 69.2 percent for HACC
sures the ability to speed up a problem
times as many processors. Weak scalir

Press Release

Record Setting Simulations at DOE Laboratories Supercomputers

Nov 28, 2012

HPC wire

November 29, 2012

Sequoia Supercomputer Runs Cosmology Code at 14 Petaflops

Inside The Largest Simulation Of The Universe Ever Created

A giant supercomputer is making massively detailed models of the cosmos.

By Clay Dillow Posted 11.08.2012 at 9:02 am

POPSCI THE FUTURE NOW

Petaflops performance scored running universe simulation

Posted November 8

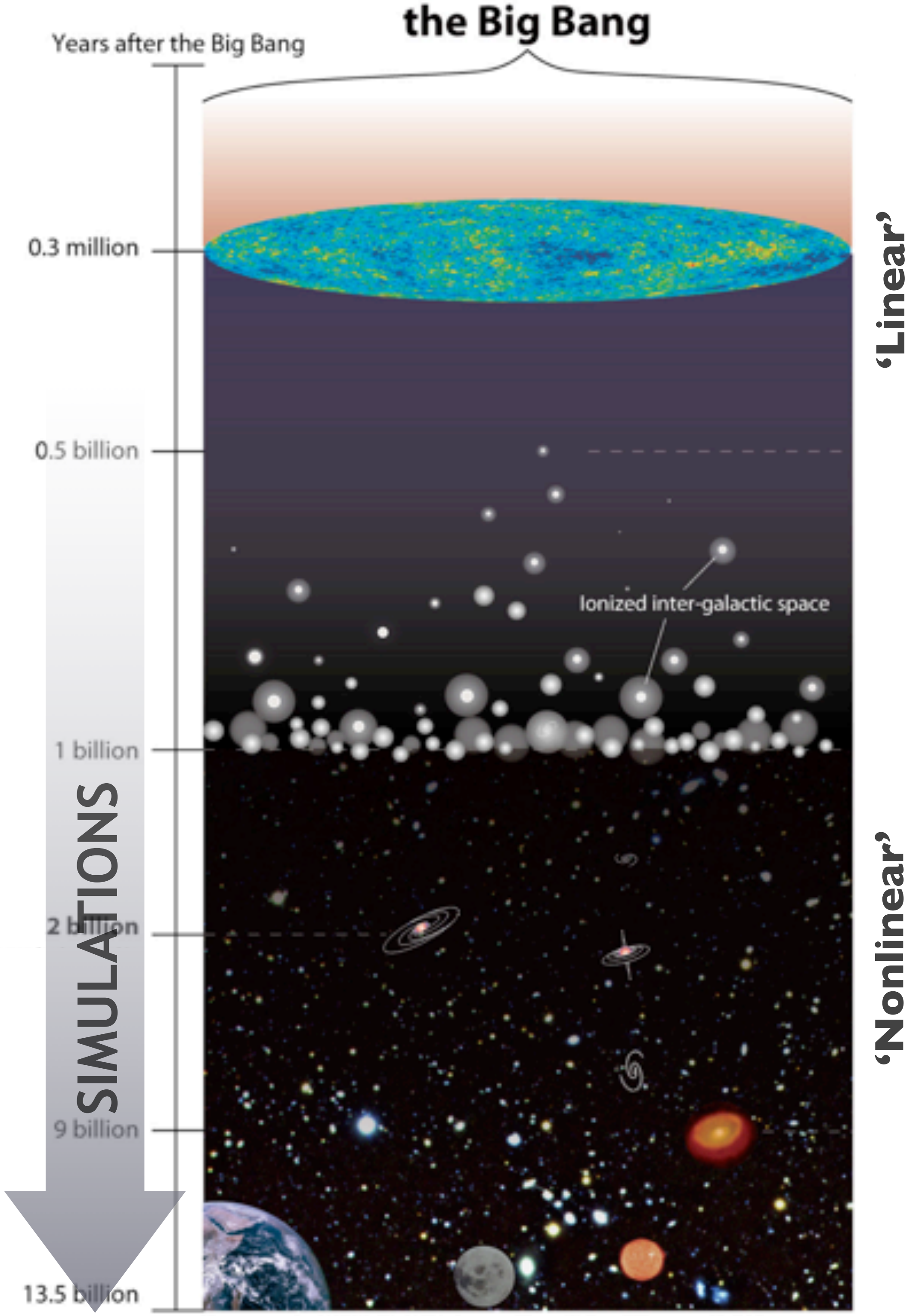
ascr big iron

Argonne NATIONAL LABORATORY BERKELEY LAB THE UNIVERSITY OF CALIFORNIA UCLA IBM Lawrence Livermore National Laboratory Los Alamos NATIONAL LABORATORY EST. 1943

The Dark Universe: Mapping the Sky

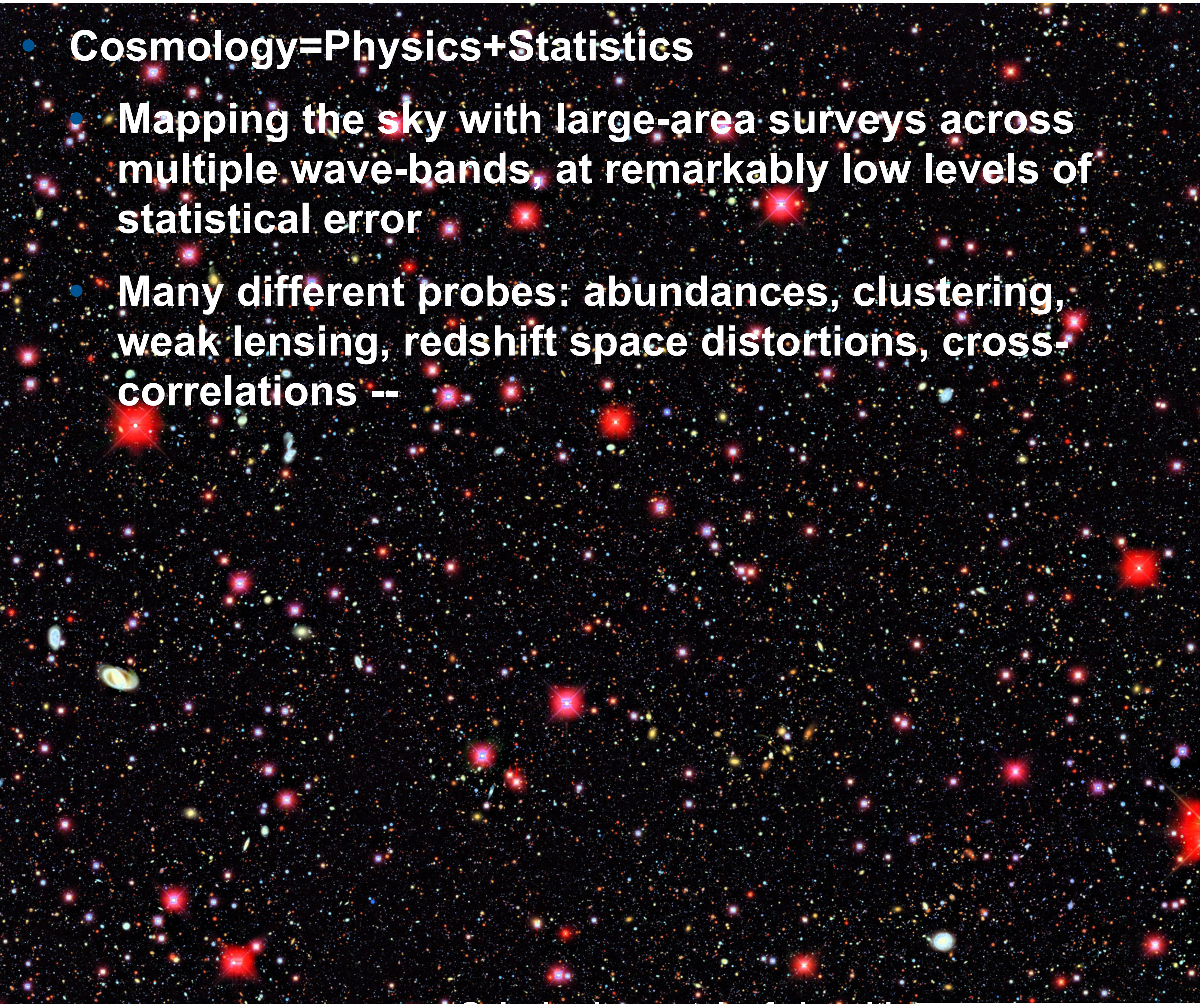
Structure Formation in the Universe: The Basic Paradigm

- **Solid understanding of structure formation is a requirement for cosmic discovery**
  - To high accuracy, initial conditions are given by a Gaussian random field
  - Initial perturbations amplified by gravitational instability in a dark matter-dominated Universe
  - Relevant theory is gravity and atomic physics ('first principles')
- **Early Universe**
  - Linear perturbation theory very successful (Cosmic Microwave Background)
- **The Universe: 'Second Half'**
  - Nonlinear domain of structure formation, impossible to treat without large-scale computing



images credit: NASA / WMAP Science Team, Subaru Telescope/NAOJ

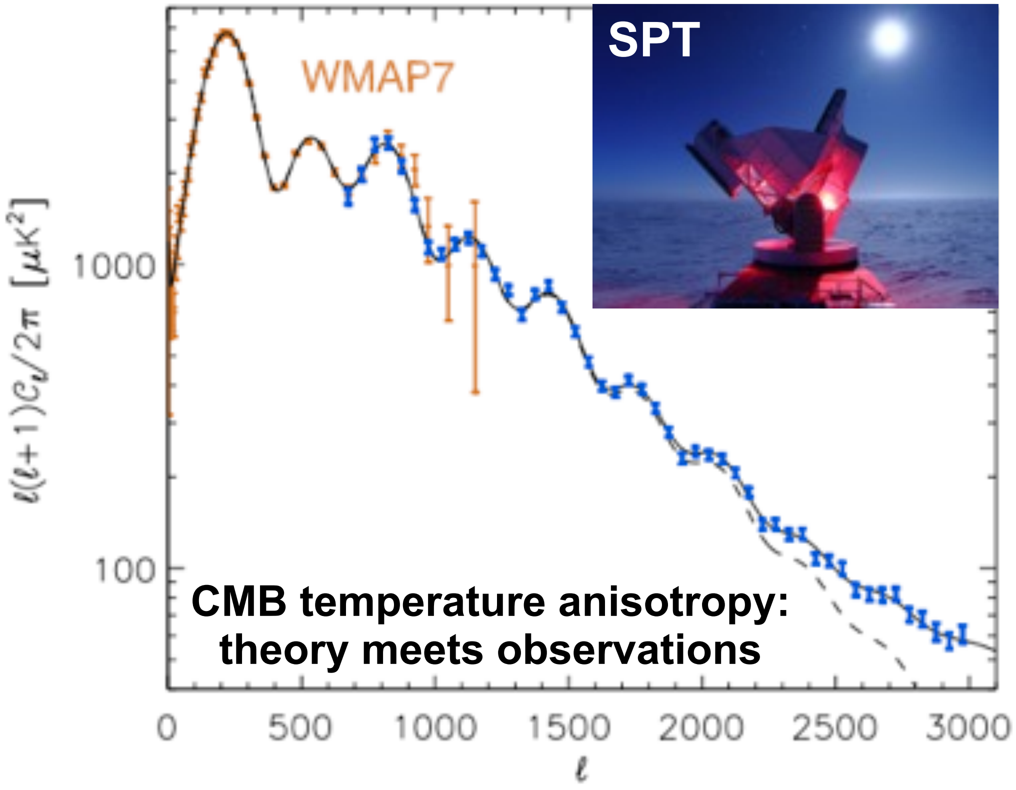
Data 'Overload': Observations of Cosmic Structure



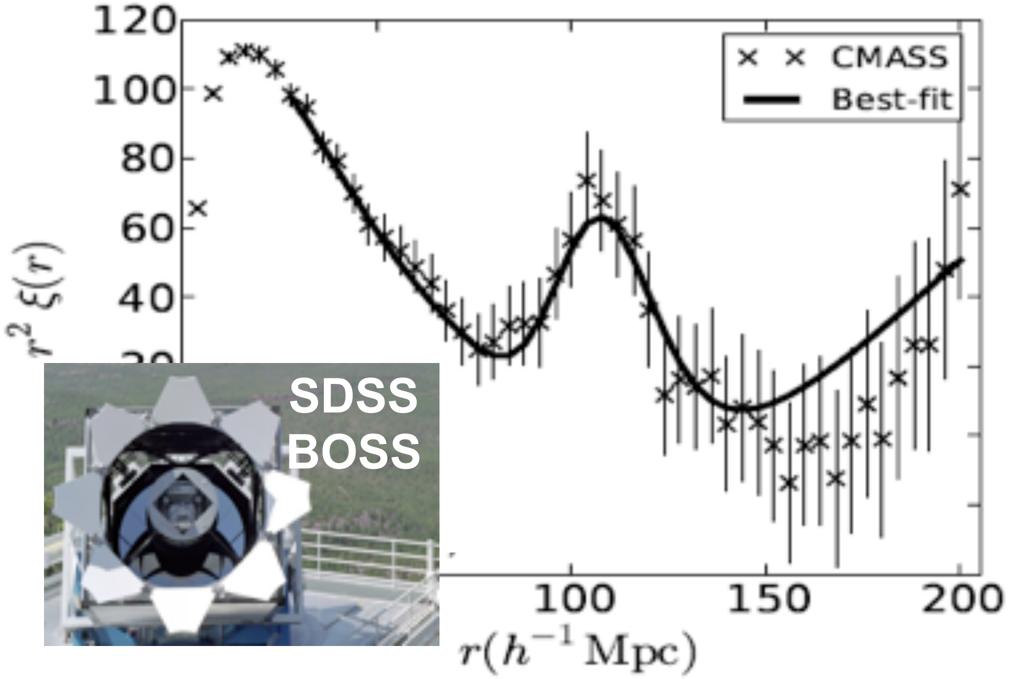
- Cosmology=Physics+Statistics
  - Mapping the sky with large-area surveys across multiple wave-bands, at remarkably low levels of statistical error
  - Many different probes: abundances, clustering, weak lensing, redshift space distortions, cross-correlations --

Galaxies in a patch of sky with area roughly the size of the full moon as seen from the ground (Deep Lens Survey). LSST will cover an area 50,000 times this size (and go deeper)

~300 PB database

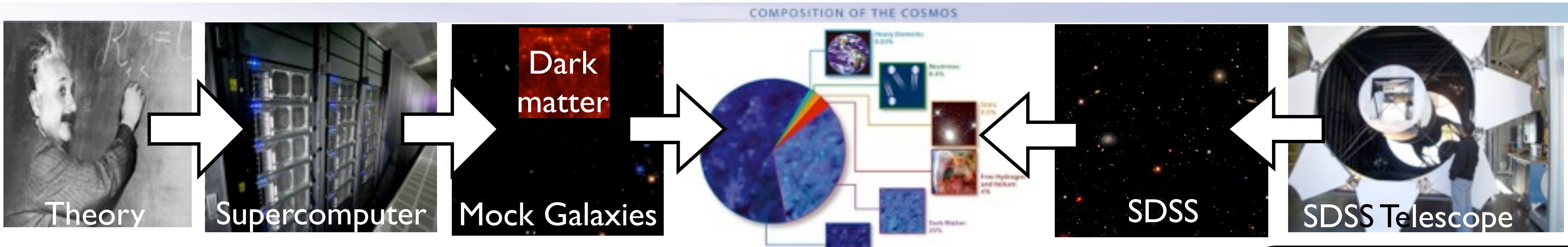


CMB temperature anisotropy: theory meets observations

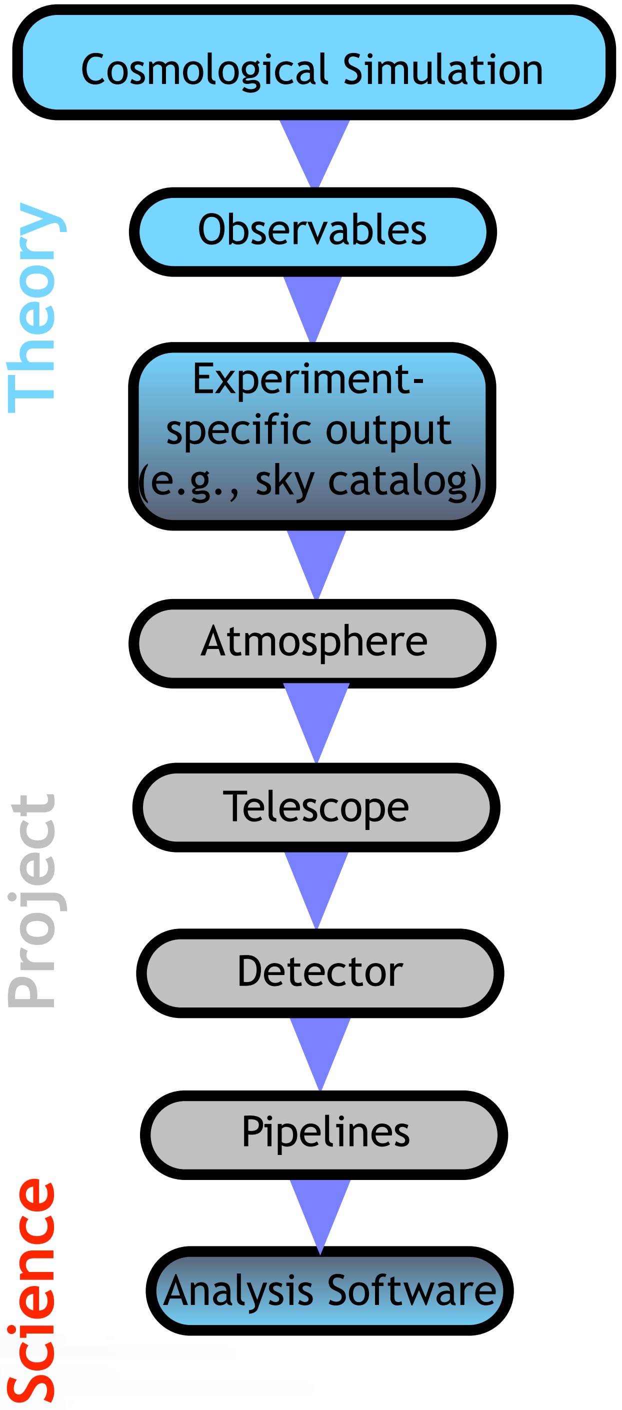


The same signal in the galaxy distribution

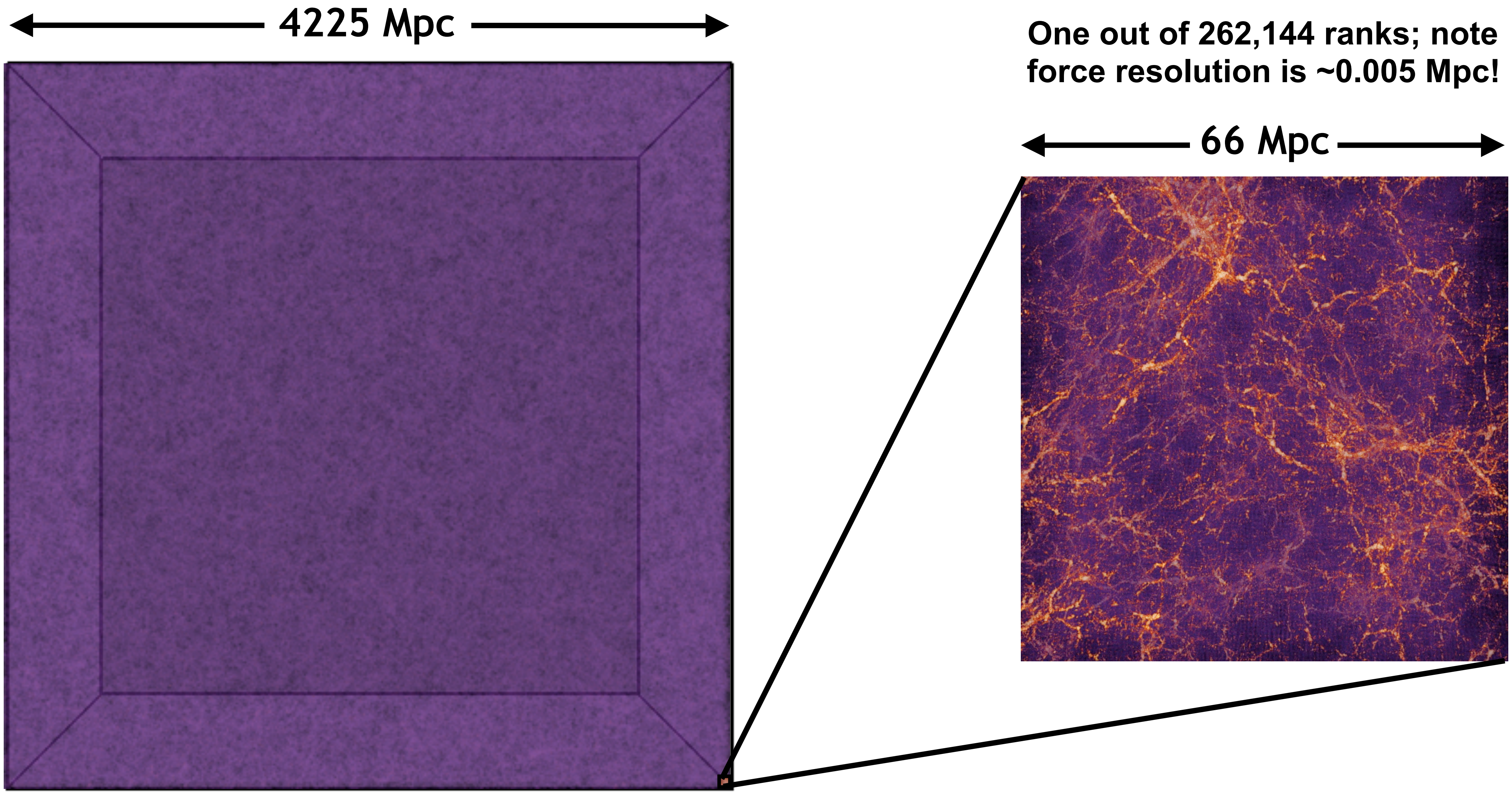
Key Role of Computational Theory/Modeling



- **Three Roles of Cosmological Simulations**
  - Basic theory of cosmological probes
  - Production of high-fidelity ‘mock skys’ for end-to-end tests of the observation/analysis chain
  - Essential component of analysis toolkits
- **Extreme Simulation and Analysis Challenges**
  - Large dynamic range simulations; control of subgrid modeling and feedback mechanisms
  - Design and implementation of complex analyses on large datasets; new fast (approximate) algorithms
  - Solution of large statistical inverse problems of scientific inference (many parameters, ~10-100) at the ~1% level



Capturing Sky Surveys: Trillion Particles in a ‘Box’



One out of 262,144 ranks; note force resolution is ~0.005 Mpc!

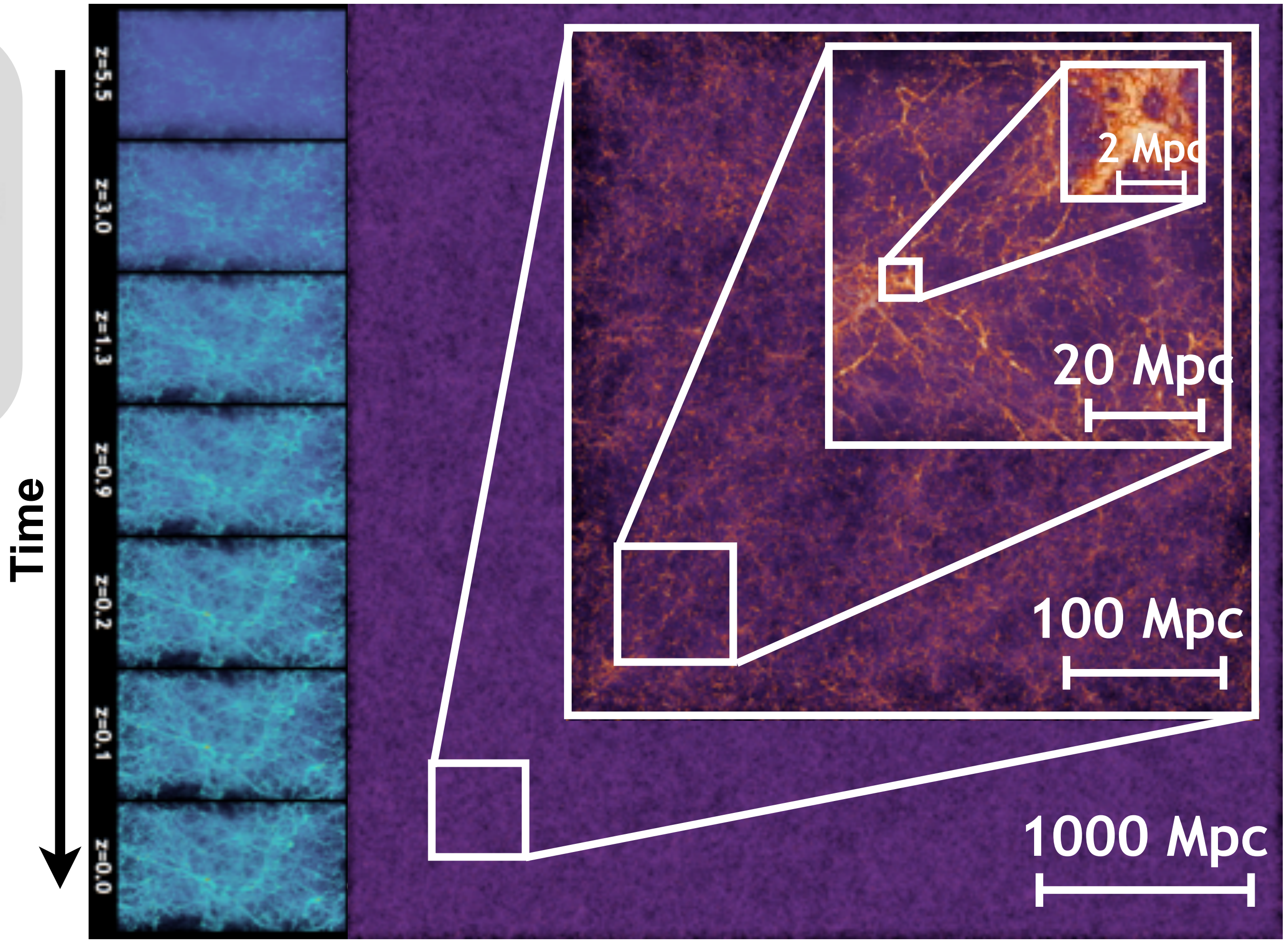
- **Size:** Volumes = ~100’s of cubic Gpc (1 pc = 3.26 light-years)
- To capture individual galaxy mass concentrations over this volume, need trillions of particles (billions of objects with thousands of sampling particles per object) -- simple numerical algorithms useless

1.1 trillion particle HACC science run at z=3 illustrating the dynamic range of a large, high-resolution, cosmological N-body simulation

Large Scale Structure Simulation Requirements

$$
\begin{aligned}
\frac{\partial f_i}{\partial t} + \dot{\mathbf{x}}\frac{\partial f_i}{\partial \mathbf{x}} - \nabla\phi\frac{\partial f_i}{\partial \mathbf{p}} &= 0, \qquad \mathbf{p} = a^2\dot{\mathbf{x}}, \\
\nabla^2\phi &= 4\pi G a^2(\rho(\mathbf{x},t) - \langle\rho_{\rm dm}(t)\rangle) = 4\pi G a^2 \Omega_{\rm dm}\delta_{\rm dm}\rho_{\rm cr}, \\
\delta_{\rm dm}(\mathbf{x},t) &= (\rho_{\rm dm} - \langle\rho_{\rm dm}\rangle)/\langle\rho_{\rm dm}\rangle, \\
\rho_{\rm dm}(\mathbf{x},t) &= a^{-3}\sum_i m_i \int d^3\mathbf{p} f_i(\mathbf{x},\dot{\mathbf{x}},t).
\end{aligned}
$$

Cosmological Vlasov-Poisson Equation



Gravitational Jeans Instablity

- **Resolution:**
  - Force resolution has to be ~kpc, a dynamic range of a million to one, also controls time-stepping
  - Local overdensity variation is ~million to one
- **Physics:**
  - Gravity dominates at scales greater than ~Mpc
  - At small scales: galaxy distribution modeling
- **Computing 'Boundary Conditions':**
  - Total memory in the PB+ class
  - Performance in the 10 PFlops+ class
  - Wall-clock of ~days/week, in situ analysis

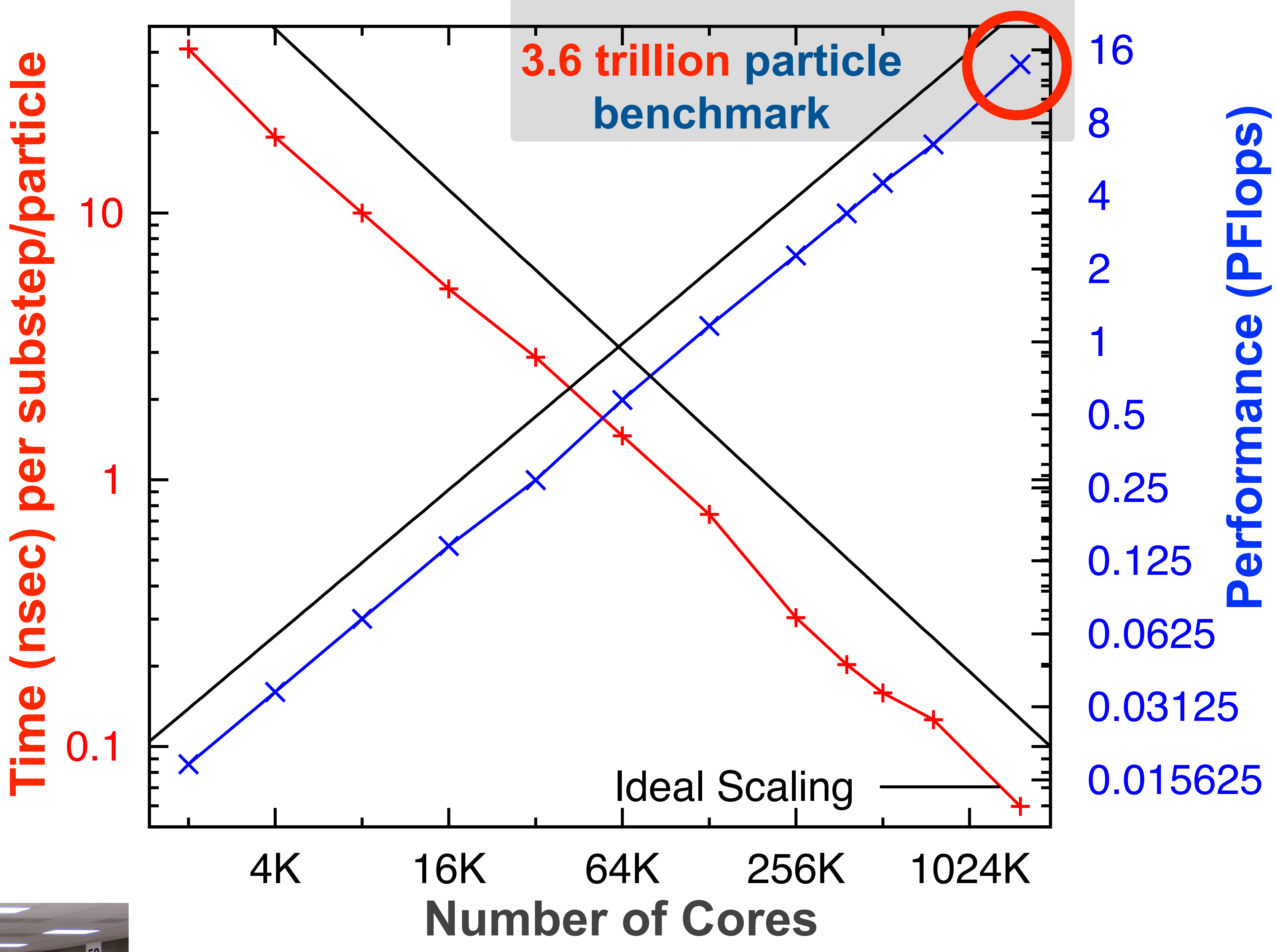
→

Can the Universe be run as a short computational 'experiment'?

Meeting the Challenge: HACC on the BG/Q

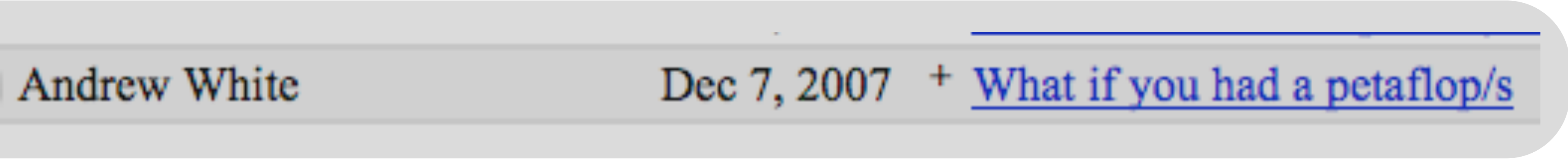
- New Cosmological N-Body Framework
  - Designed for extreme performance AND portability, including heterogeneous systems
  - Supports multiple programming models
  - Memory efficient
  - In situ analysis framework
  - Production science code

13.94 PFlops, 69.2% peak, 90% parallel efficiency on 1,572,864 cores/MPI ranks, 6.3M-way concurrency



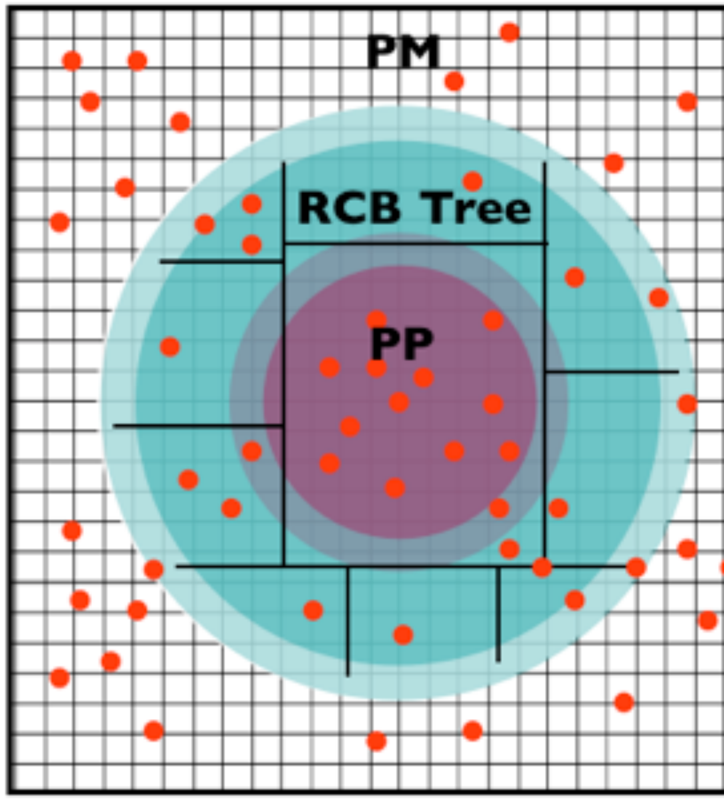
HACC weak scaling on the IBM BG/Q (MPI/OpenMP)

Opening the HACC 'Black Box': Design Principles



Andrew White Dec 7, 2007 + What if you had a petaflop/s

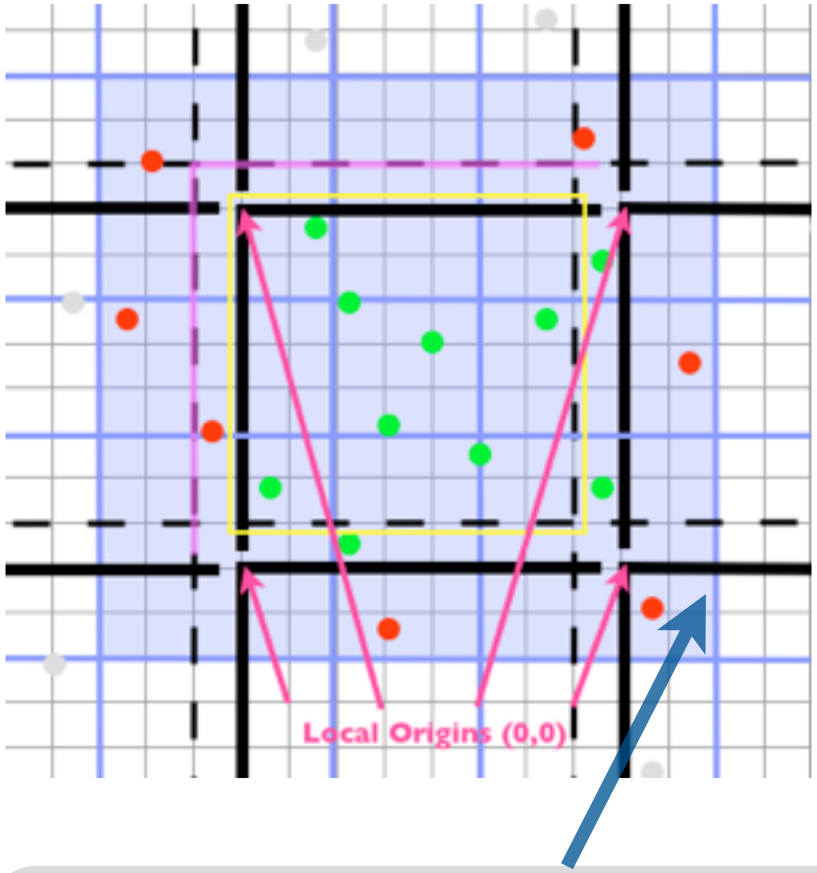
- **Optimize Next-Generation Code 'Ecology':** Numerical methods, algorithms, mixed precision, data locality, scalability, I/O, in situ analysis -- life-cycle significantly longer than architecture timescales
- **Framework design:** Support a 'universal' top layer + 'plug-in' optimized node-level components; minimize data structure complexity and data motion -- support multiple programming models
- **Performance:** Optimization stresses scalability, low memory overhead, and platform flexibility; assume 'on your own' for software support, but hook into tools as available (e.g., ESSL FFT)
- **Optimal Splitting of Gravitational Forces:** Spectral Particle-Mesh melded with direct and RCB tree force solvers, short hand-over scale (dynamic range splitting ~ 10,000 X 100)
- **Compute to Communication balance:** Particle Overloading
- **Time-Stepping:** Symplectic, sub-cycled (uses Hamiltonian Maps)
- **Force Kernel:** Highly optimized force kernel takes up large fraction of compute time, no look-ups due to short hand-over scale
- **Production Readiness:** runs on all supercomputer architectures



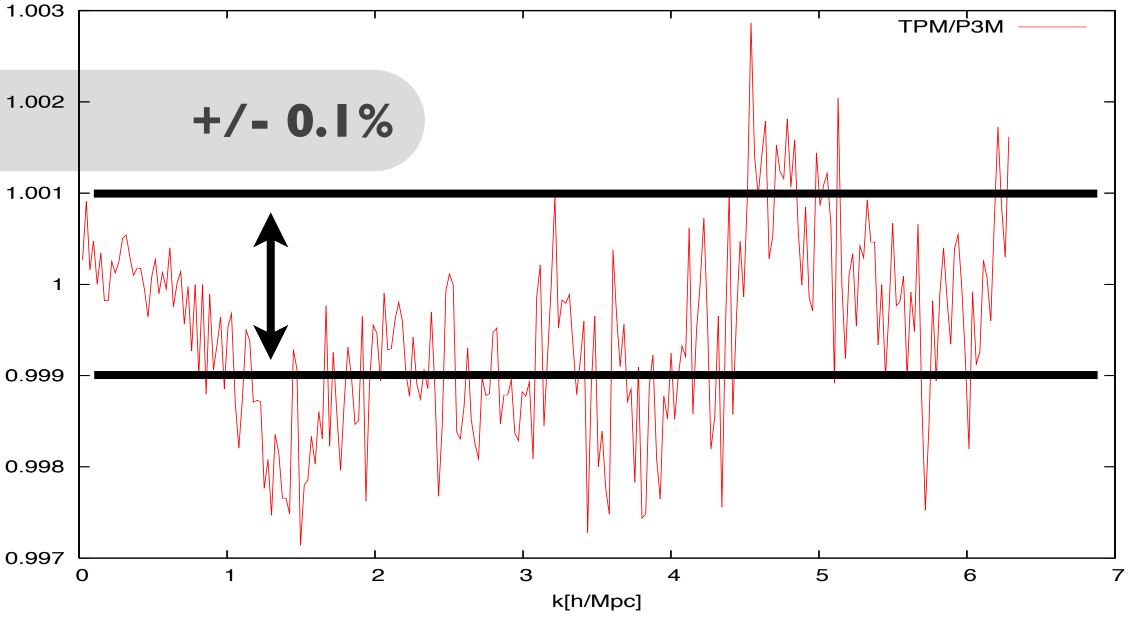
HACC force hierarchy (PPTreePM)

Particle Overloading and Short-Range Solvers

- **Particle Overloading:** Particle replication instead of conventional guard zones with 3-D domain decomposition -- minimizes inter-processor communication and allows for swappable short-range solvers
- **Short-range Force:** Depending on node architecture switch between P3M and PPTreePM algorithms (pseudo-particle method goes beyond monopole order), by tuning number of particles in leaf nodes and error control criteria, optimize for computational efficiency
- **Error tests:** Can directly compare different short-range solver algorithms



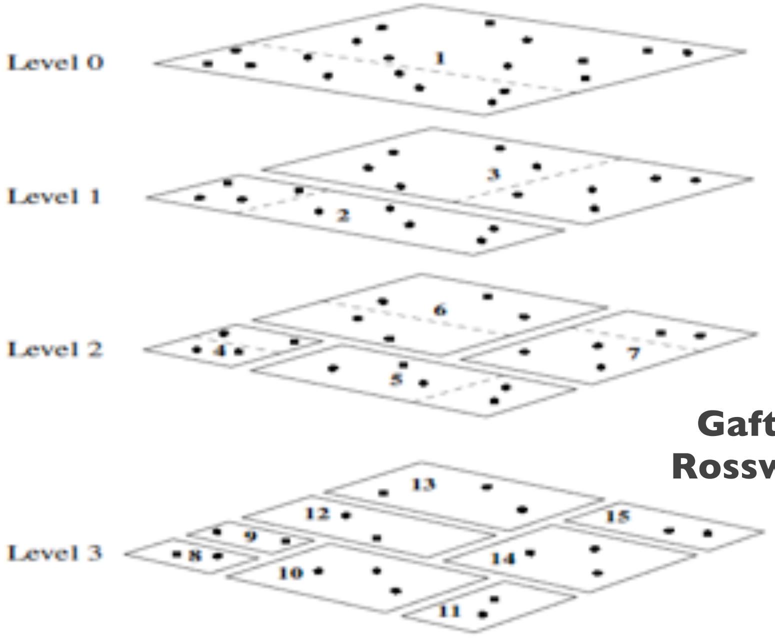
Overload Zone (particle 'cache')



+/- 0.1%

P(k) Ratio

HACC Force Algorithm Test: PPTreePM vs. P3M



Gafton and Rosswog 2011

RCB Tree Hierarchy

HACC: BG/Q Implementation

- **HACC BG/Q Algorithms:**

  1) Long-range force with base HACC FFT-based SPM (excellent performance)

  2) Short-range force: Particle-Particle + RCB Tree + highly tuned force kernel

- **Data Locality:** At rank-level, enforced by particle overloading, at tree-level use the RCB grouping to organize particle memory buffers (all P-P interactions are in nearby leaf nodes, this also increases accuracy)
- **Tree Build/Walk Minimization:** Every particle has an interaction list -- constructing this is an overhead (‘treebuild’); reduce tree depth in two ways: (i) rank-local trees, (ii) shortest possible hand-over scale, (iii) bigger P-P component than is usual, using the optimized force kernel
- **Force Kernel:** Because of the compactness of the short-range interaction, the kernel can be represented as

$$f_{SR} = (s+\epsilon)^{-3/2} - f_{grid}(s)$$

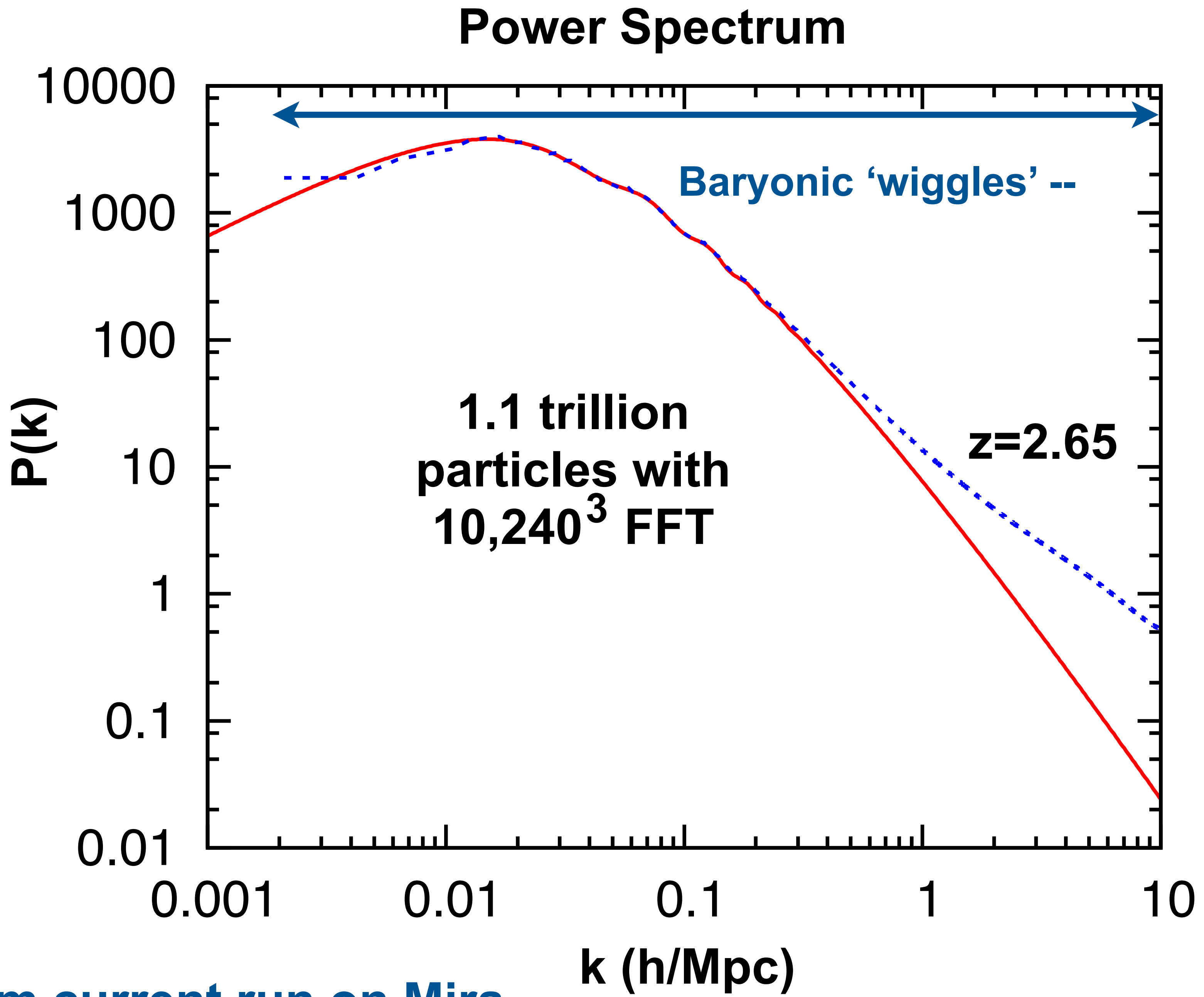
where

$$s = \mathbf{r}\cdot\mathbf{r}, \quad f_{grid}(s) = poly[5](s)$$

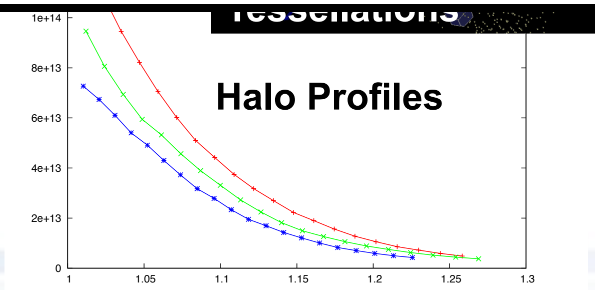
- **Kernel Evaluation:** This consists of three parts: (i) Filtering, (ii) Inverse square root evaluation, and (iii) Polynomial evaluation

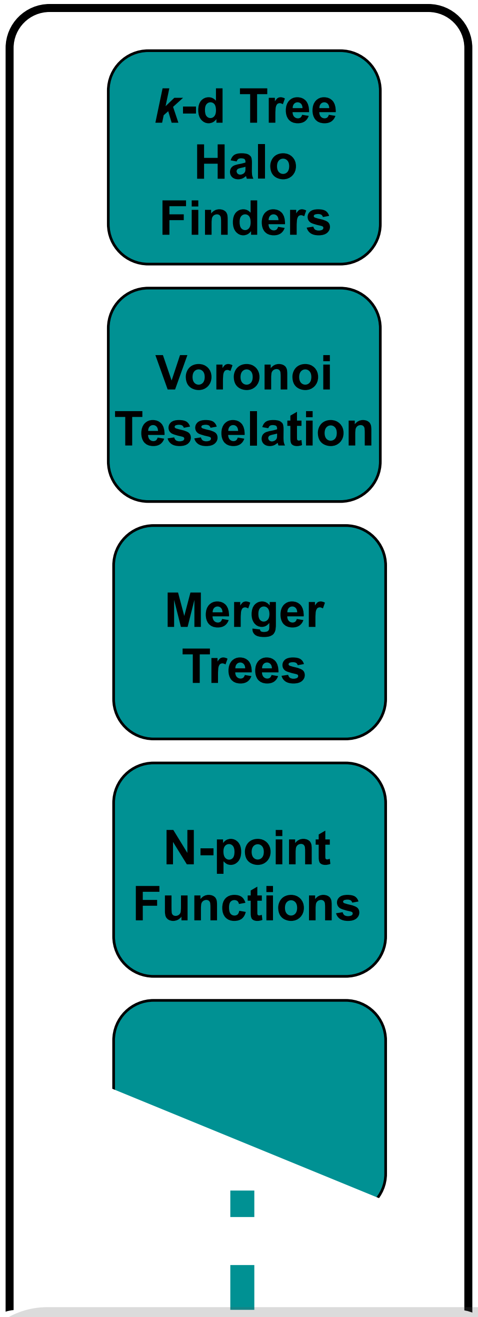
HACC: *Fast* In Situ Analysis

- **Data Reduction:** A trillion particle s analysis requirem analysis
- **I/O Chok** analyses time > ar schedulir
- **Fast Alg** time is o simulatio
- **Ease of** analyses post-proc



From current run on Mira --





k-d Tree Halo Finders

Voronoi Tesselation

Merger Trees

N-point Functions

Predictions go into Cosmic Calibration Framework that solves the Cosmic Inverse Problem

Ly-A Forest Simulations

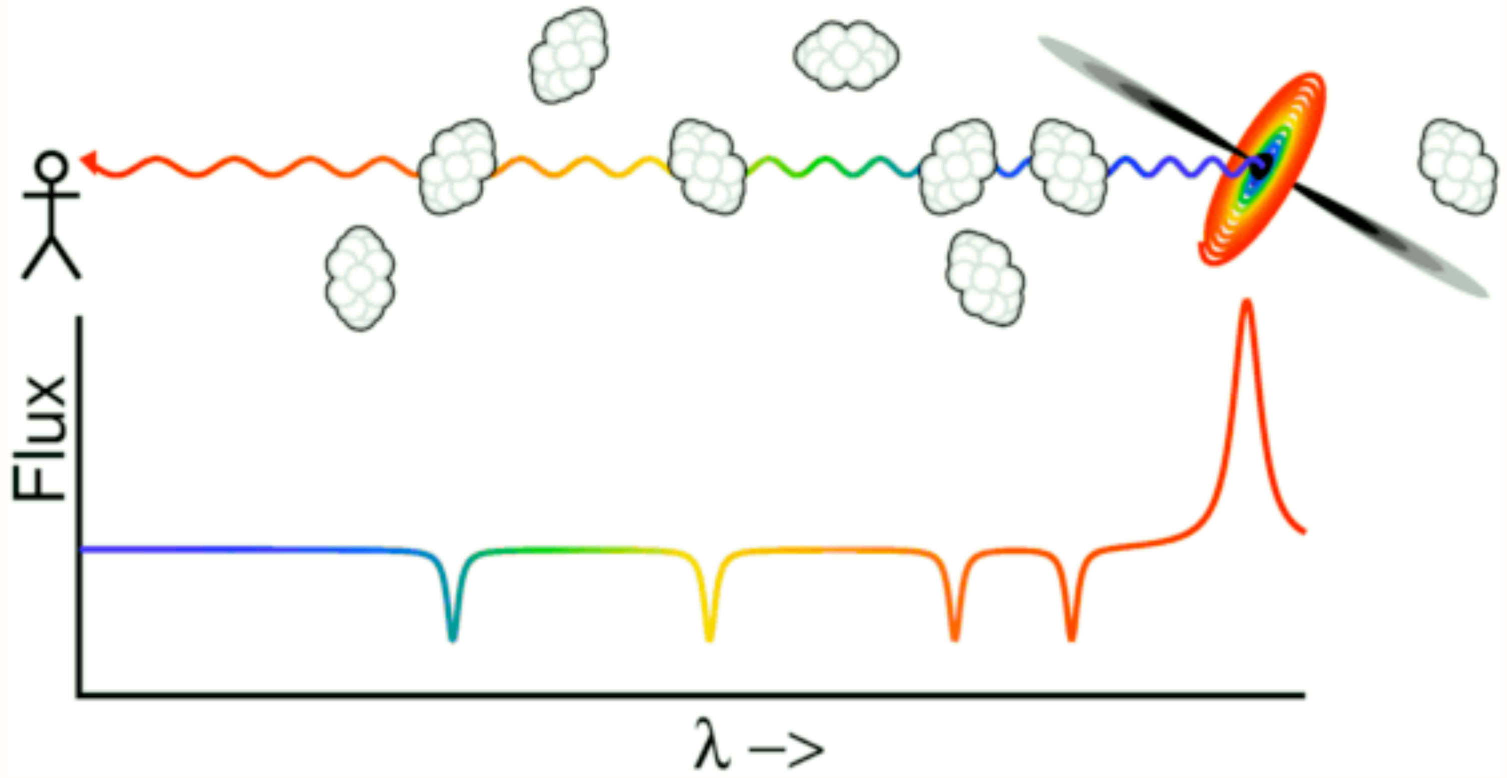
Zarija Lukić

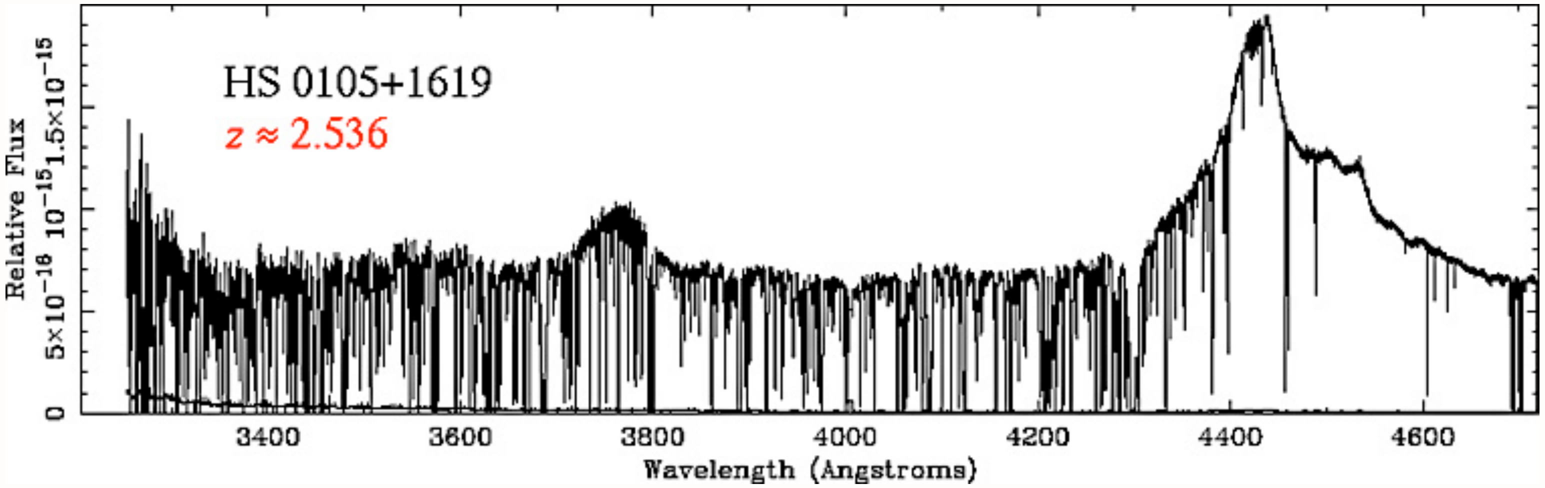
Ann Almgren, Peter Nugent, Casey Stark, Martin White

BERKELEY LAB

C3
COMPUTATIONAL
COSMOLOGY
CENTER

Lyman-alpha forest



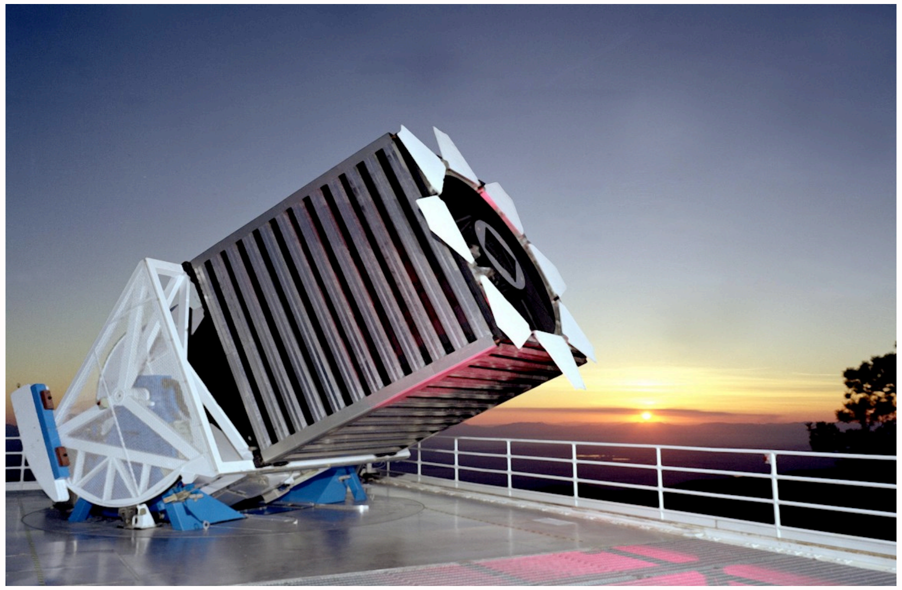


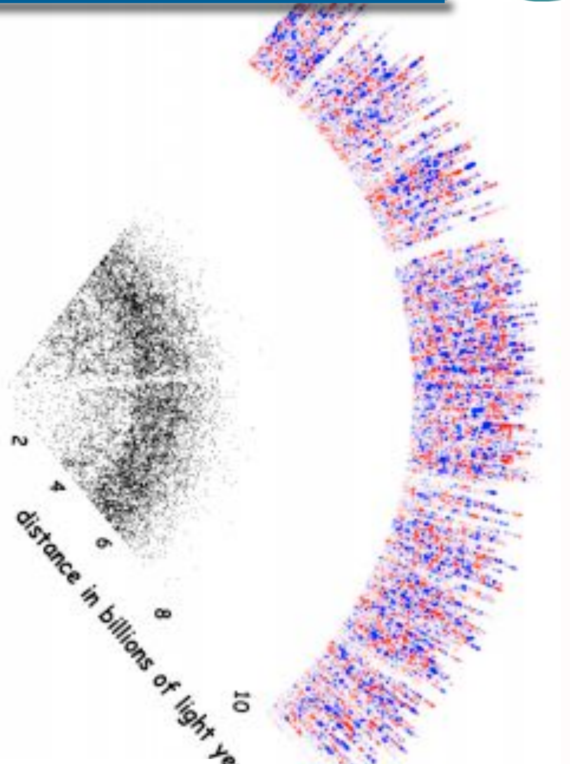
- Quasars emit featureless spectrum with a few broad emissions
- Neutral hydrogen absorbs light at its rest-frame Ly-A
- HI traces gas, which traces dark matter...
- Each "skewer" is a 1-D map of density field

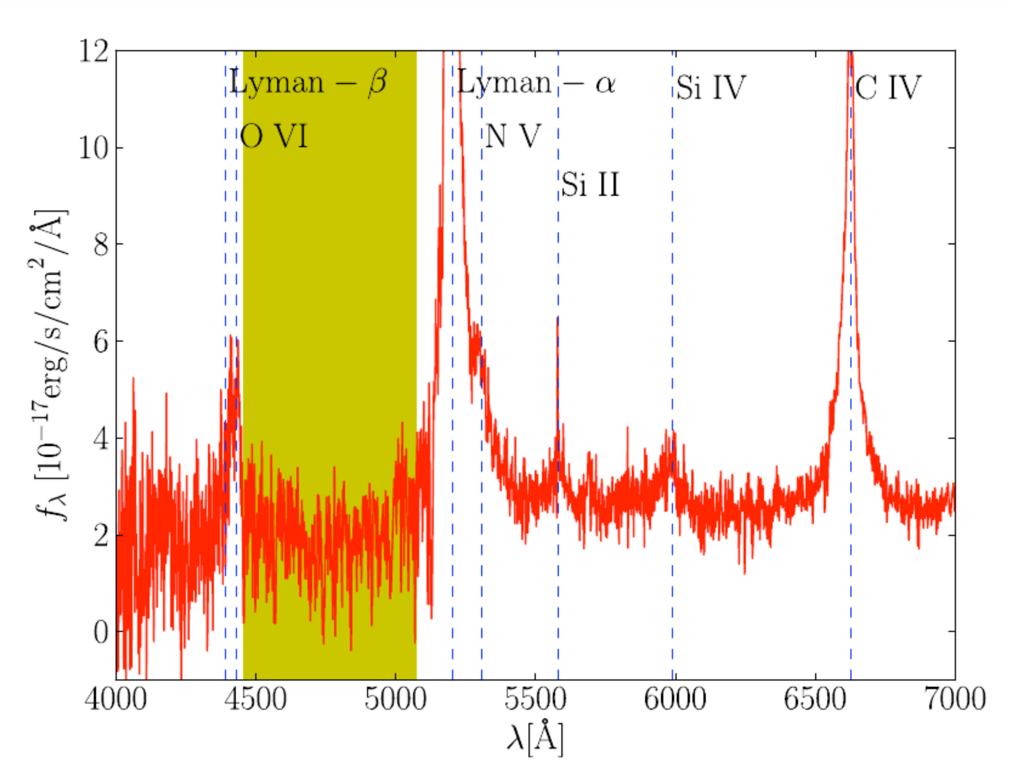
Surveys

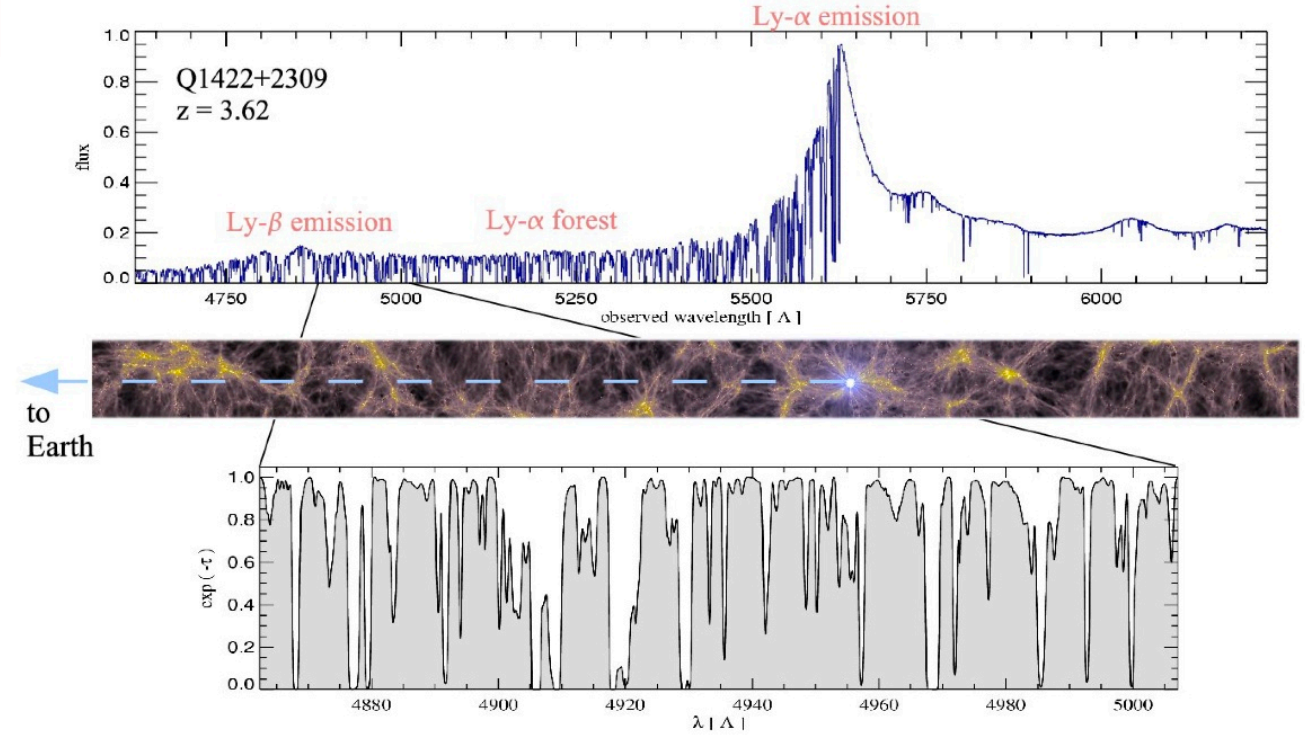
1. BOSS (2009-2014):
   ~160,000 quasars
2. MS-DESI (2018+):
   ~1,000,000 quasars

Low resolution!









O(100) high-res VLT/Keck spectra

NYX code

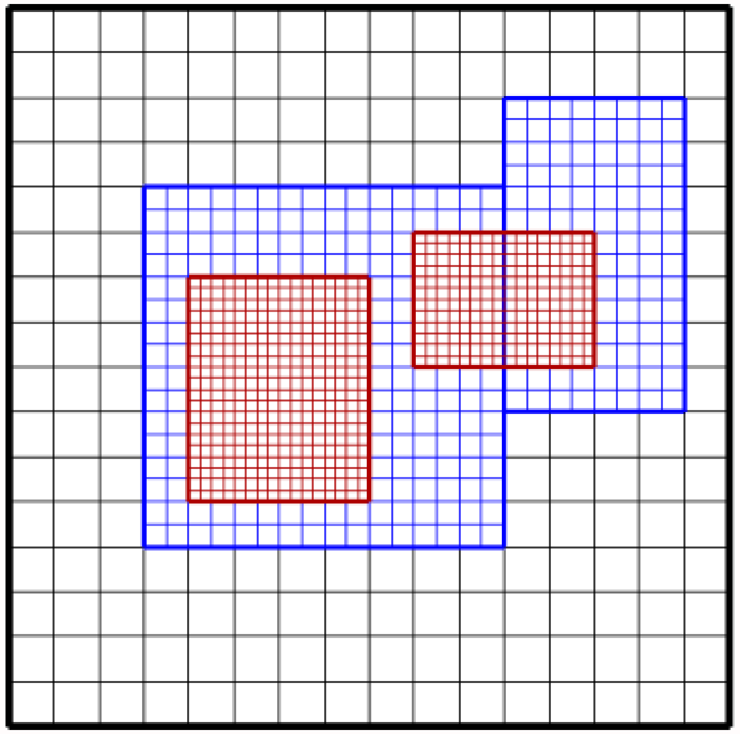
- 3-D Cartesian grid, finite volume representation
- Evolve dark matter as collisionless Lagrangian fluid
- Evolve baryons as ideal gas using unsplit, Godunov-type methodology
- Adaptive mesh refinement (AMR) to extend dynamic range
- Uses BoxLib software framework developed by CCSE group @ LBL
- Code paper: ApJ, 765, 39 (2013)

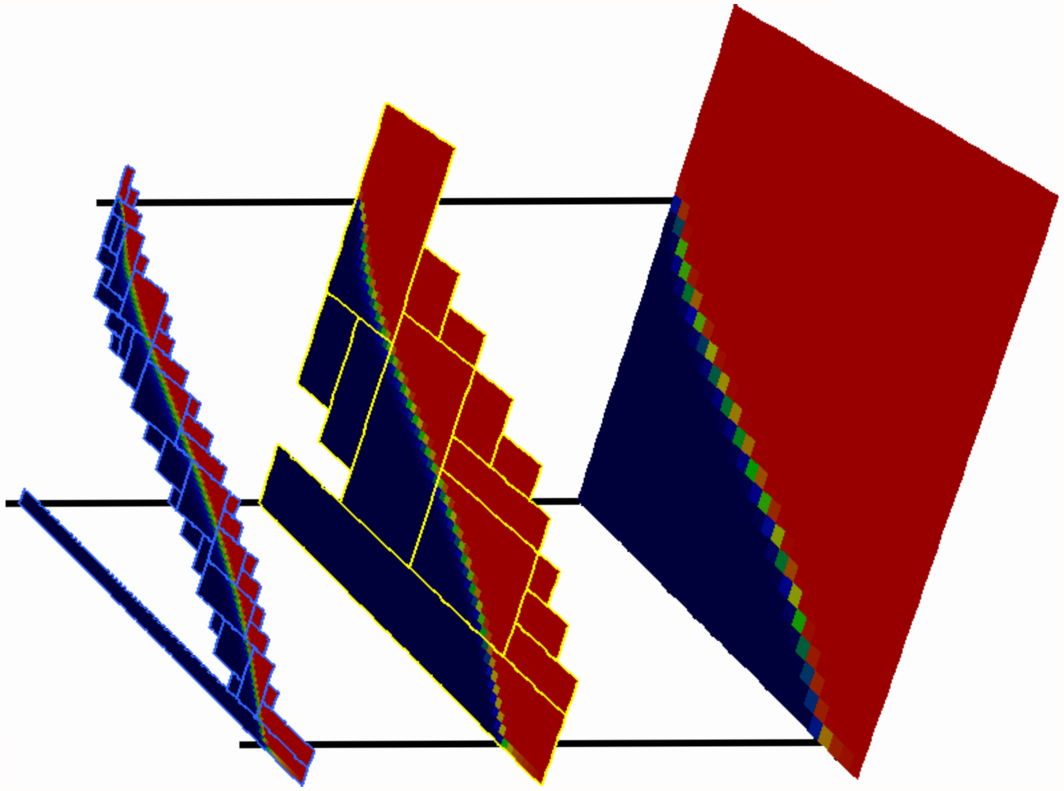
Mesh (AMR) code

- BoxLib framework
- Adaptive refinement
- Works as:

1. Tag cells for refinement on a desired criteria
2. Group cells into optimal rectangular grids
3. Chunk grids & distribute them to processes

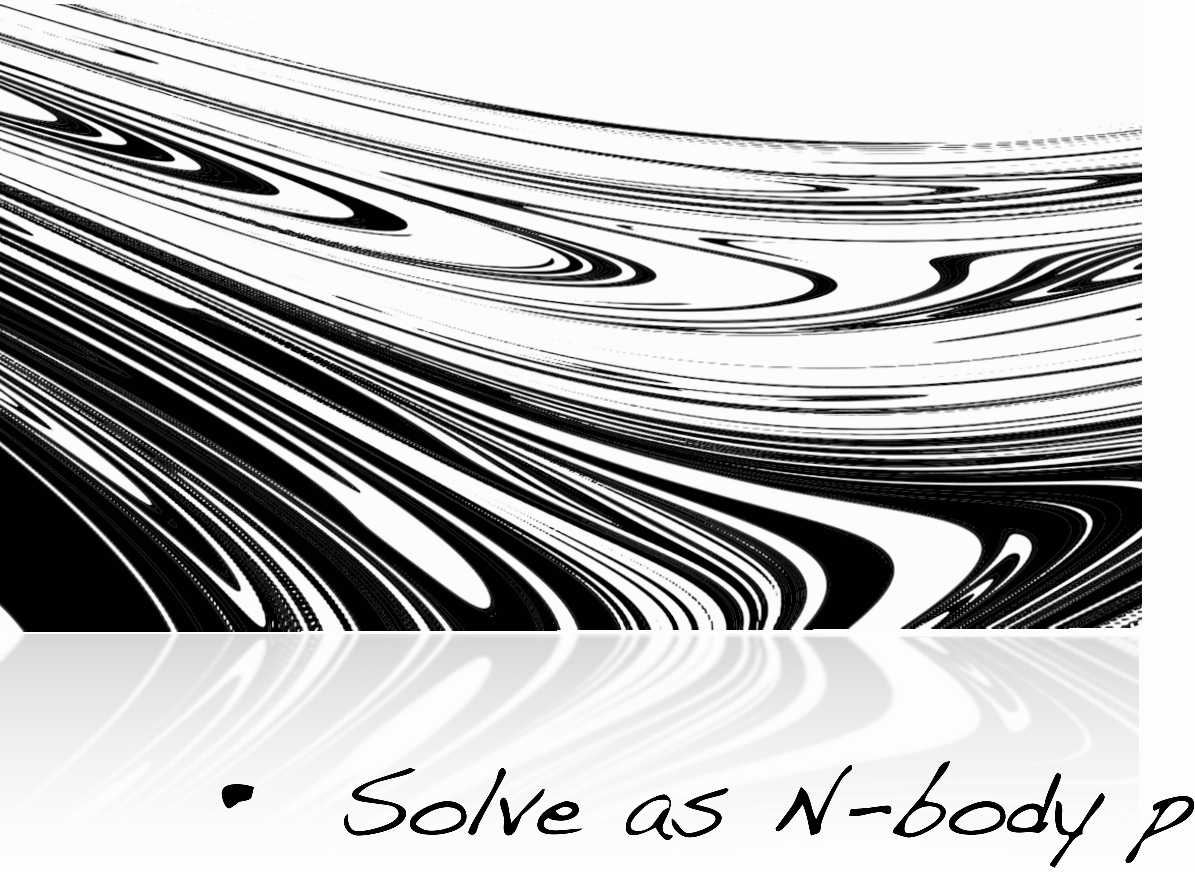
- Refinement factor 2 or 4
- No strict parent-child relation between patches





Dark matter

- Collisionless fluid evolving under self-gravity



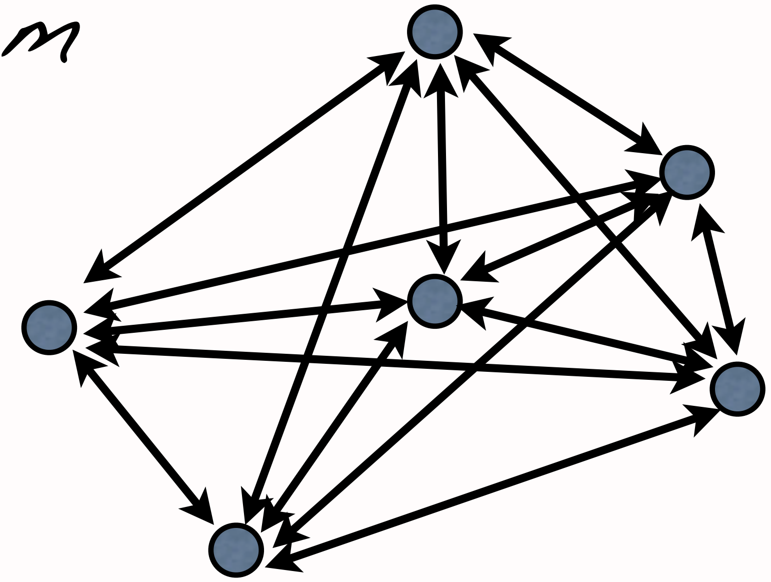
$$\frac{\partial f}{\partial t} + \frac{1}{ma^2}\mathbf{p}\cdot\nabla f - m\nabla\phi\cdot\frac{\partial f}{\partial \mathbf{p}} = 0$$

$$\nabla^2\phi = \frac{4\pi G}{a}(\rho_{tot} - \rho_0)$$

- Solve as N-body problem

$$\frac{d\mathbf{x}_i}{dt} = \frac{1}{a}\mathbf{u}_i$$

$$\frac{d(a\mathbf{u}_i)}{dt} = \mathbf{g}_i$$

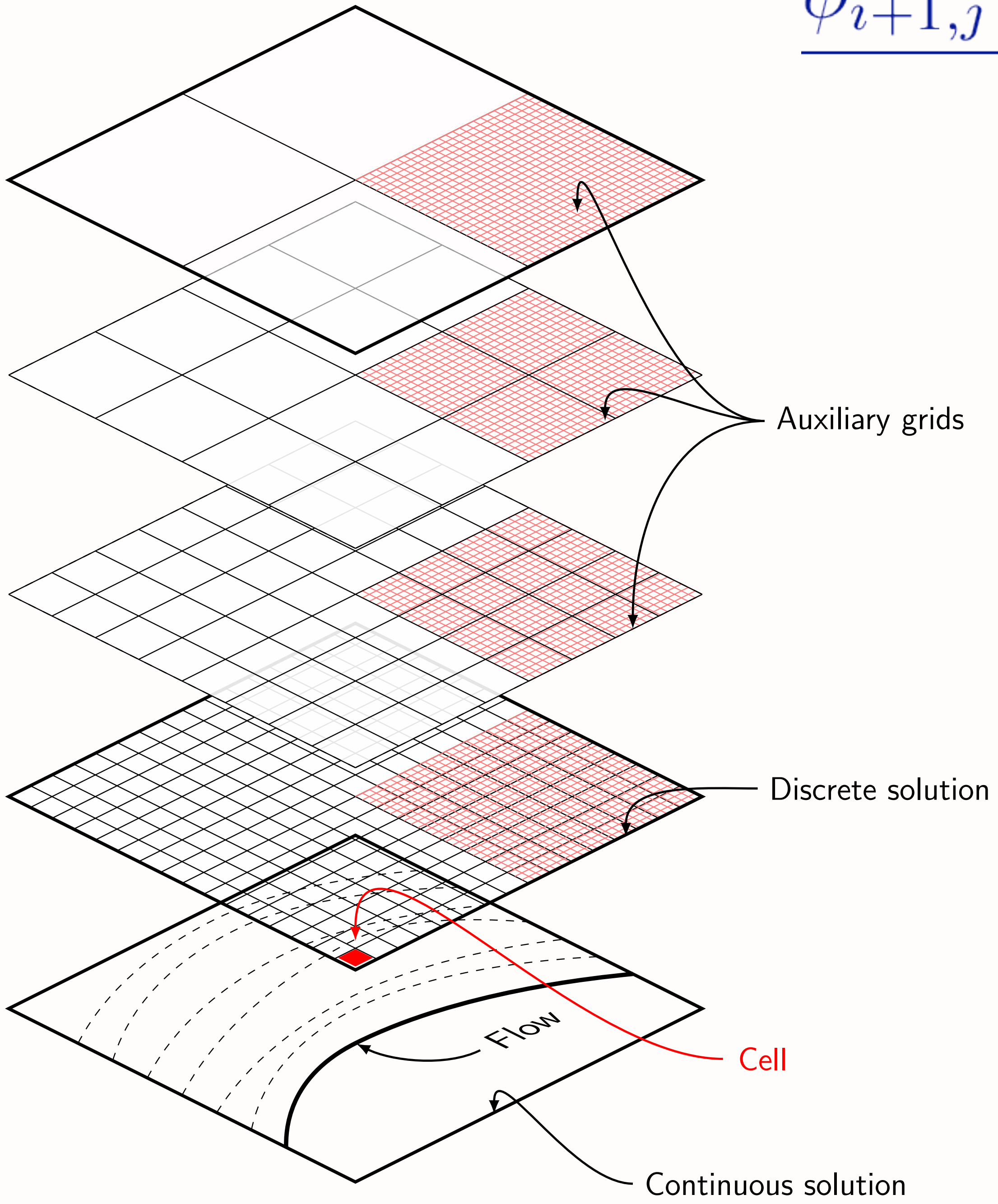


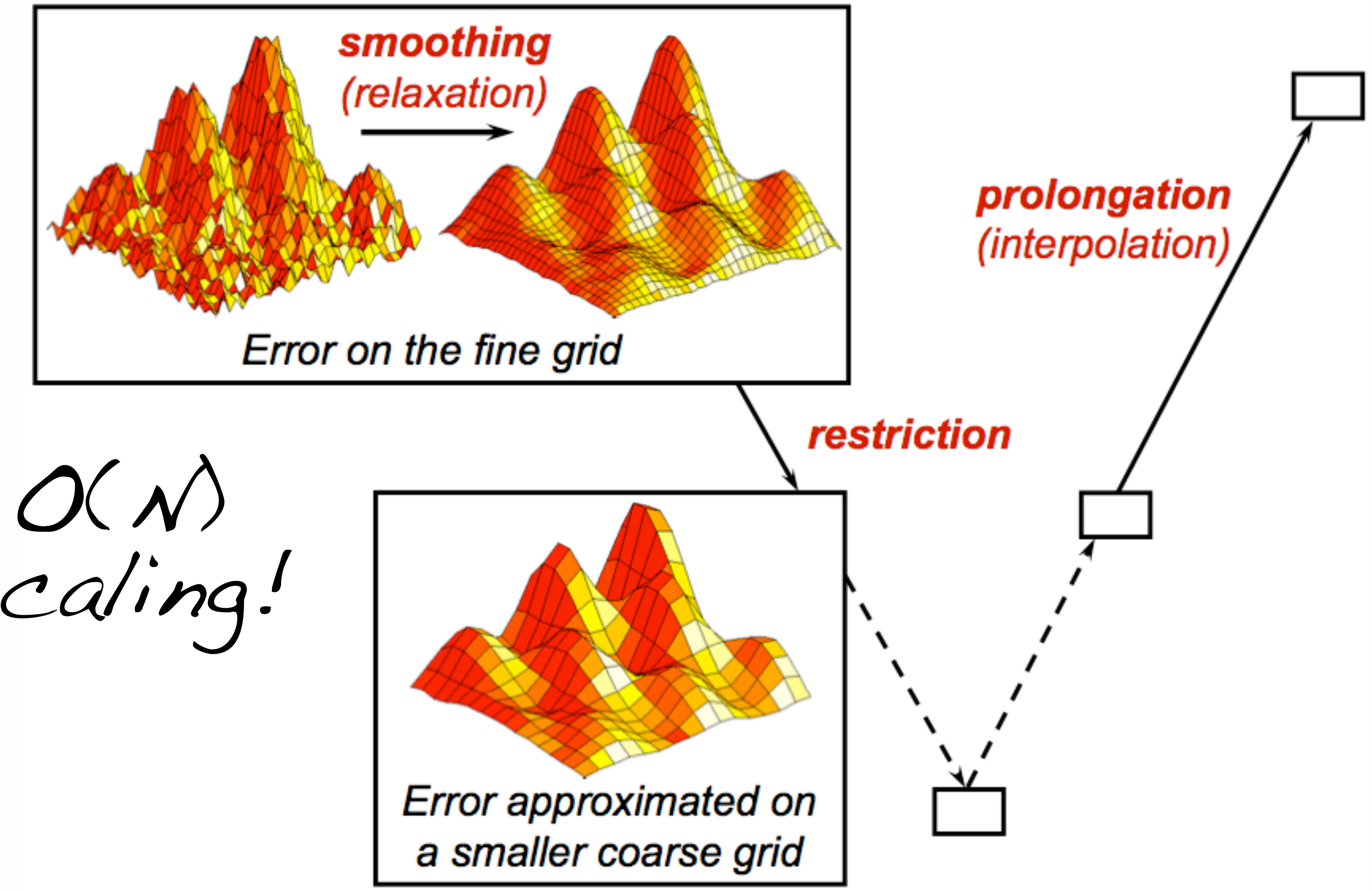
$O(N^2)$ scaling

Gravity solve

- Deposit mass on a grid, and solve linear system:

$$\frac{\phi_{i+1,j} + \phi_{i,j+1} - 4\phi_{i,j} + \phi_{i-1,j} + \phi_{i,j-1}}{\Delta x^2} = \rho_{i,j}$$



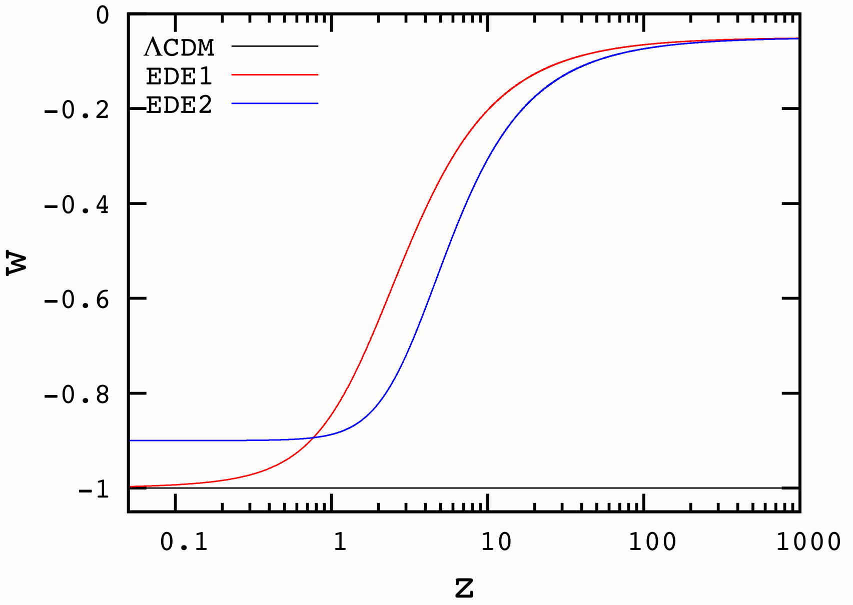


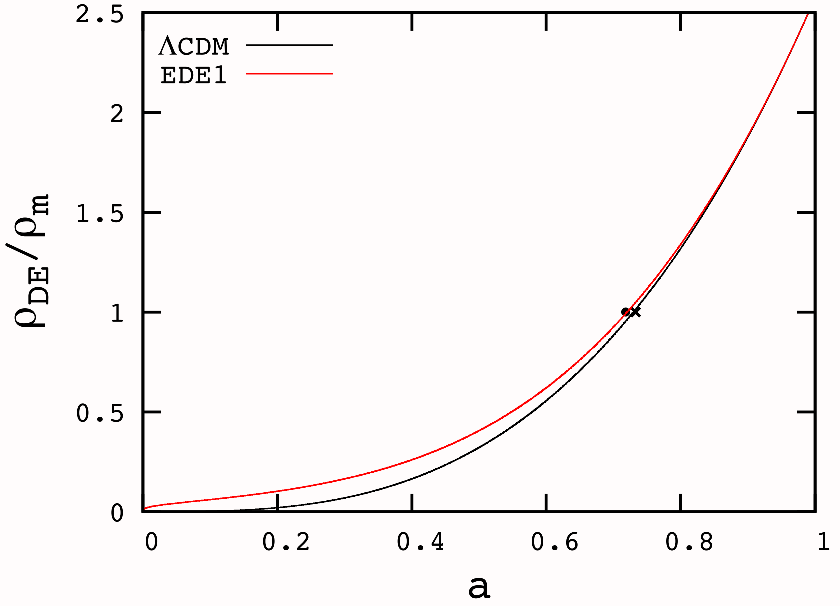
O(N) scaling!

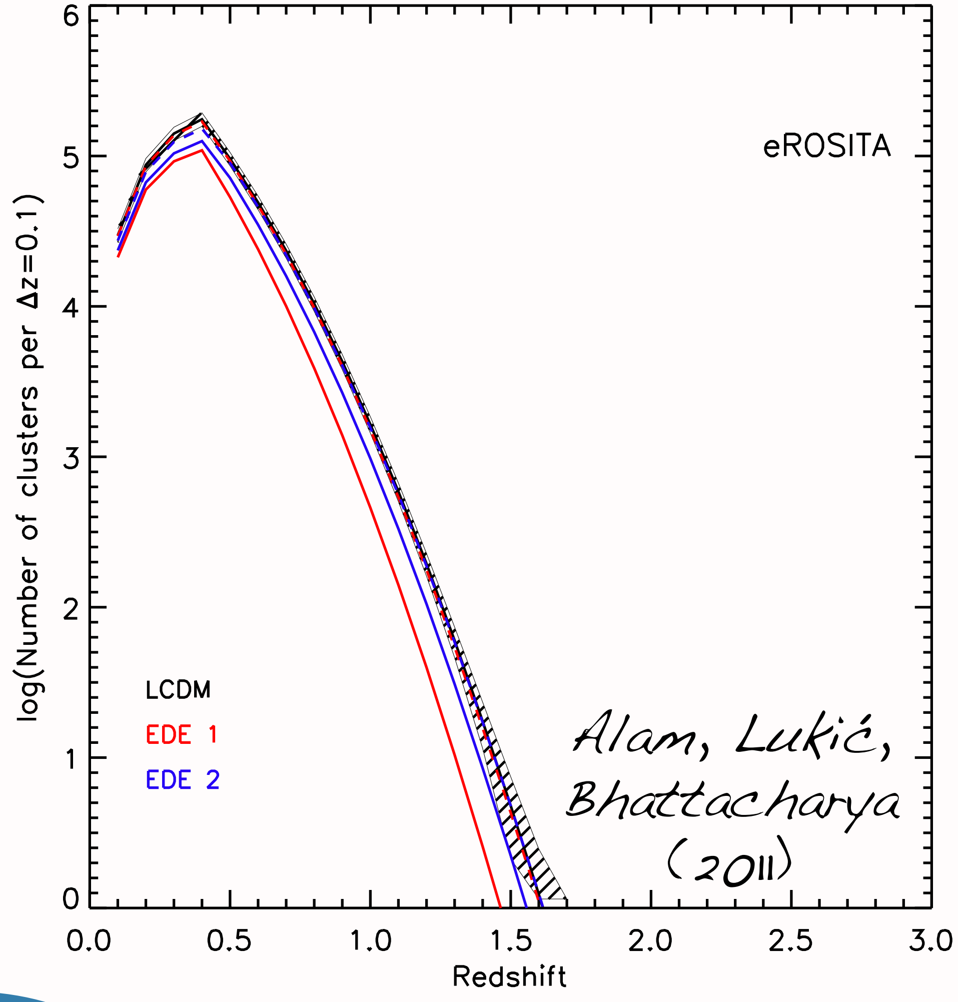
Excellent engine for non-linear Poisson equation

Dynamical models*

- Dark energy equation of state $w \neq -1$
- Modifications of gravity







Alam, Lukić, Bhattacharya (2011)

$$ds^2 = a^2(\eta)\left[(1+2\Psi(\mathbf{x},\eta))d\eta^2 - (1+2\Phi(\mathbf{x},\eta))\delta_{\alpha\beta}dx^\alpha dx^\beta\right]$$

$$\delta_i' = -3\mathcal{H}(c_{s,i}^2 - w_i)\delta_i - \left[1+3\mathcal{H}(c_{s,i}^2 - c_{a,i}^2)\right](1+w_i)\frac{v_i}{k} \boxed{- 3(1+w_i)\Psi'}$$

$w \neq -1$ perturbations!

$$v_i' = -\mathcal{H}(1-3c_{s,i}^2)v_i + \frac{kc_{s,i}^2}{(1+w_i)}\delta_i - kA$$

* Footnote slide

Baryons

- Modeled as inviscid ideal fluid
- Solve Euler equations of gasdynamics:

$$\frac{\partial \rho_b}{\partial t} = -\frac{1}{a}\nabla \cdot (\rho_b \mathbf{u})$$

$$\frac{\partial (a\rho_b \mathbf{u})}{\partial t} = -\nabla \cdot (\rho_b \mathbf{u}\mathbf{u}) - \nabla p + \rho_b \mathbf{g}$$

$$\frac{\partial (a^2\rho_b E)}{\partial t} = -a\nabla \cdot (\rho_b \mathbf{u} E + p\mathbf{u}) + a\rho_b \mathbf{u} \cdot \mathbf{g} + a\dot{a}\left((2 - 3(\gamma - 1))\rho_b e\right) + a\Lambda_{HC}$$

+ gamma-law equation of state

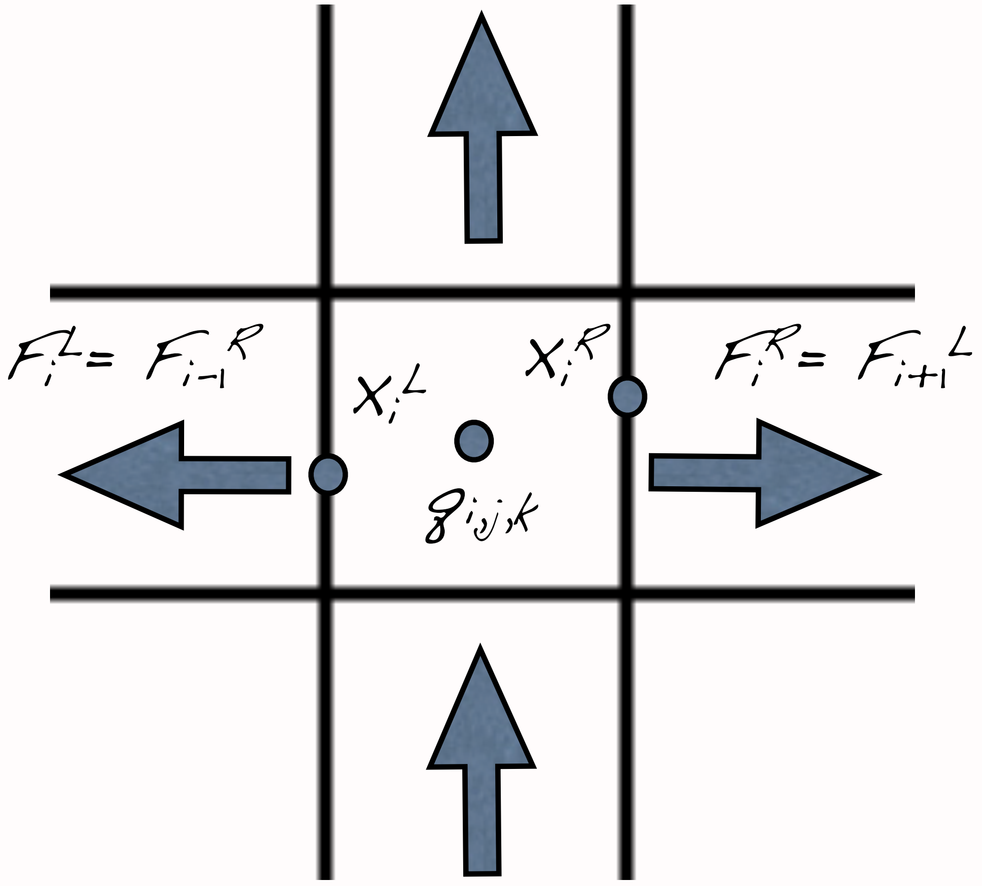
$$p = (\gamma - 1)\rho e$$

Finite volume method

- Calculate "face" values of primitive variables from cell averages using high-order interpolation
- Reconstruct profile of each variable within the cell
- Predict average values on edges over the time step using characteristic extrapolation.
- Compute fluxes by solving exact Riemann problem
- Use these fluxes to update solution to the next timestep

Conservation law:

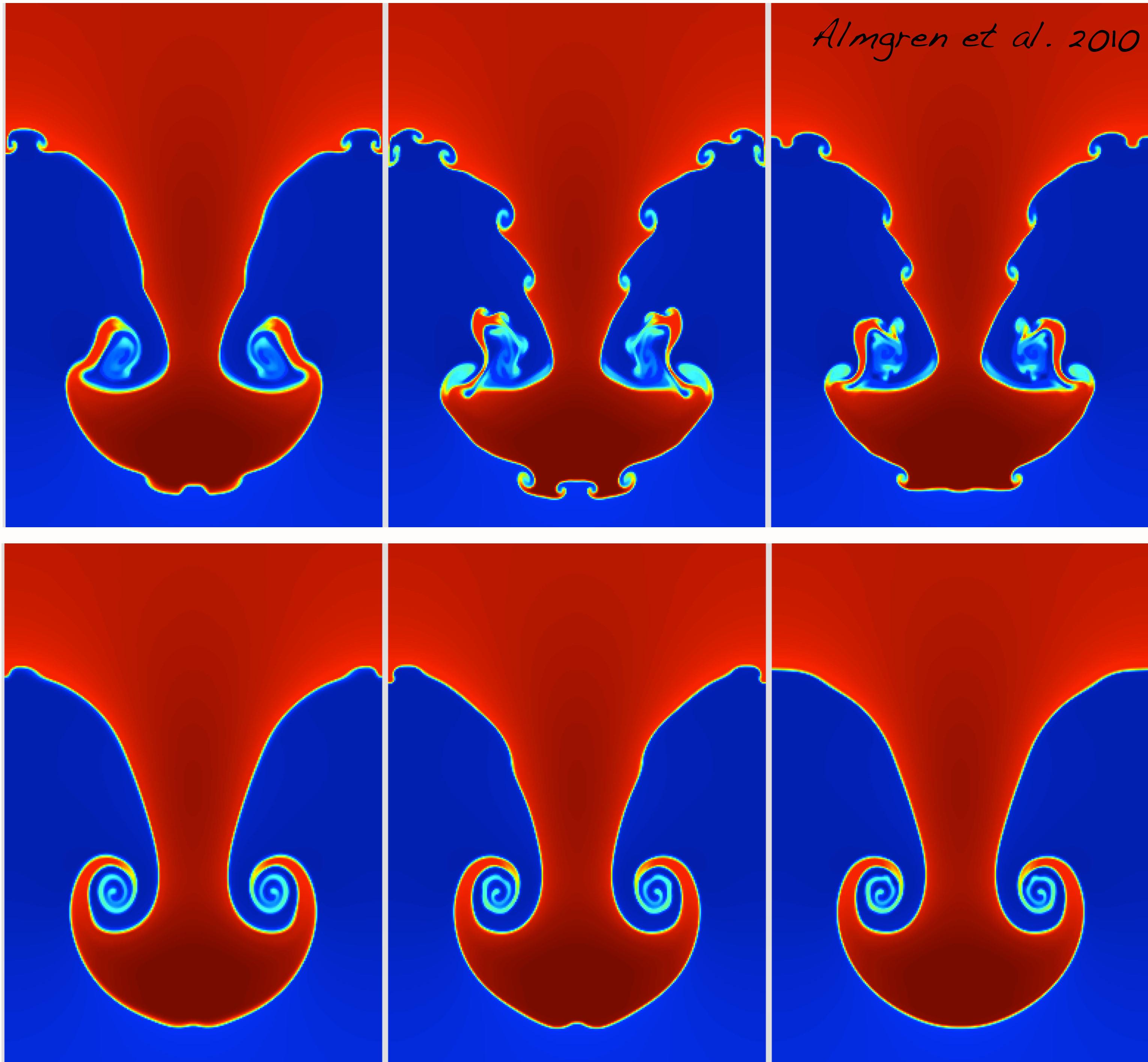
$$\frac{\partial \mathbf{q}}{\partial t} = -\nabla \cdot \mathbf{F}(\mathbf{q}, t)$$



Multiple dimensions

Rayleigh-Taylor instability:

Dimensionally split methods induce secondary instabilities


Almgren et al. 2010

Unsplit methods don't; price: ~2x slower in 3D

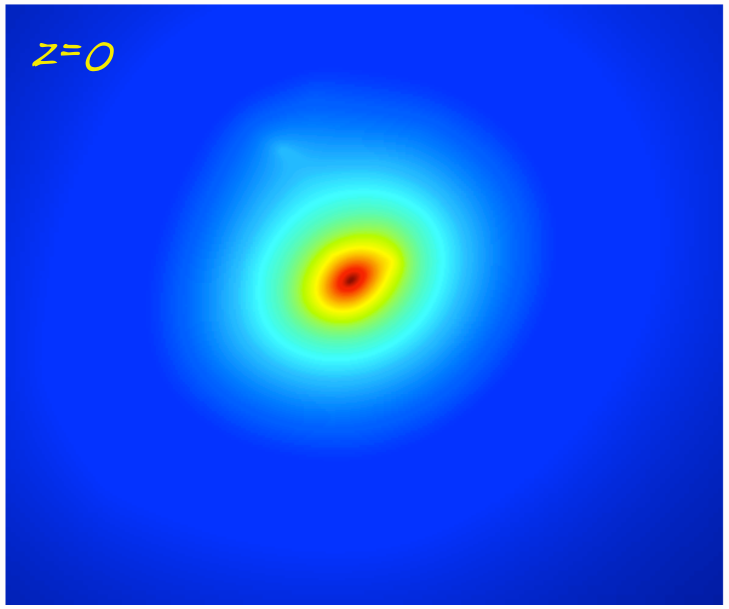
Validation

Santa Barbara cluster:

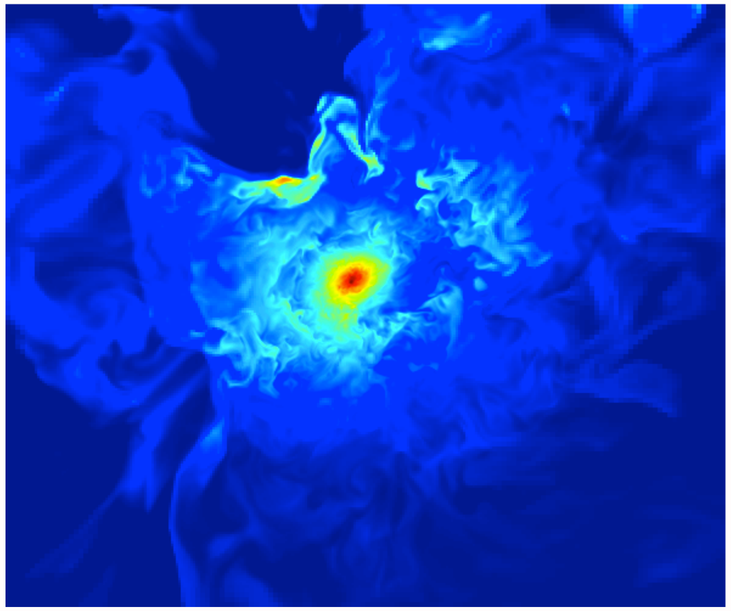
$L=64 Mpc$ $\Omega_{tot}=1$

$z_{ini}=63$ $\Omega_L=0$

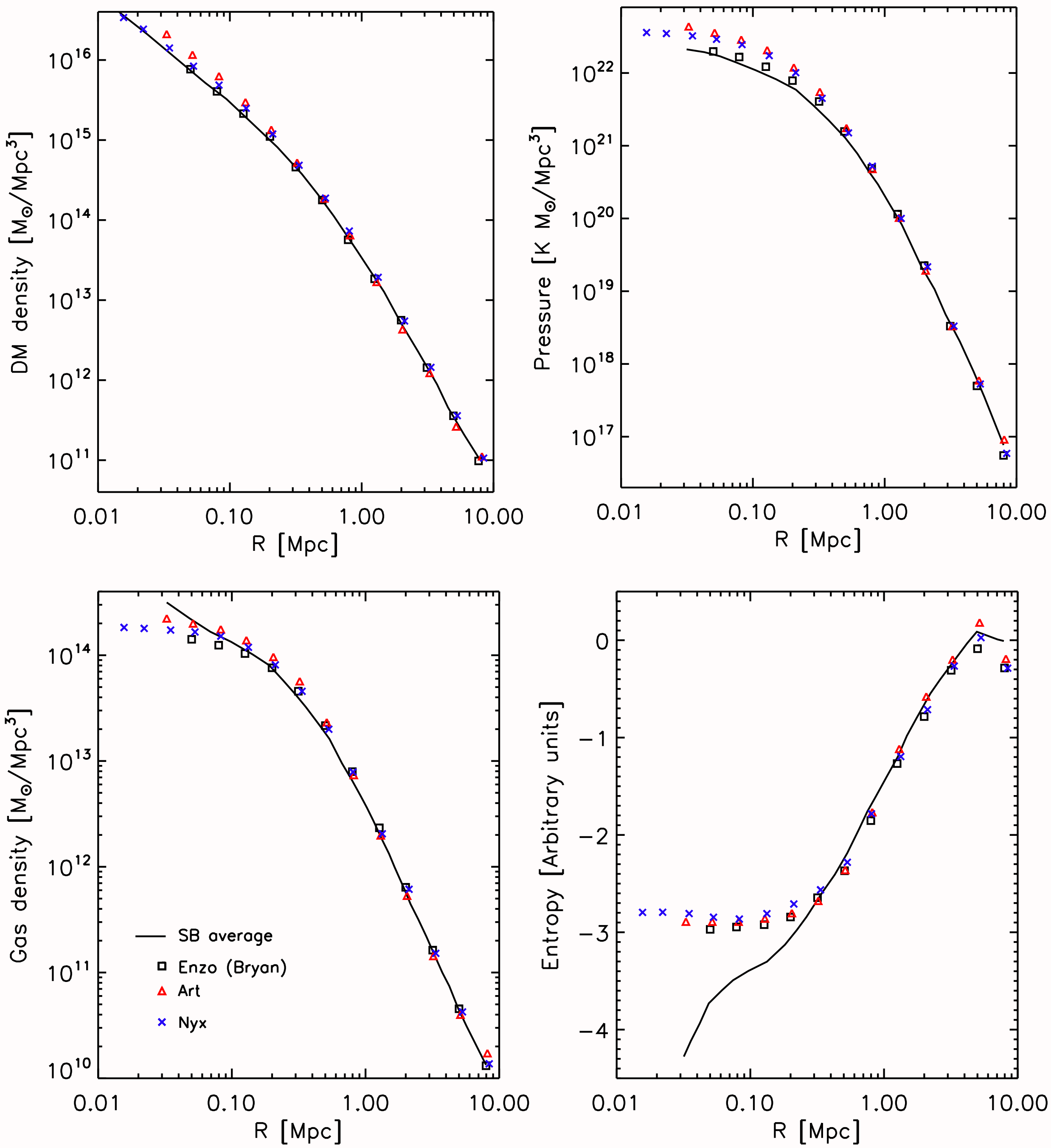
$h=0.5$ $\Omega_b=0.1$



Grav. potential

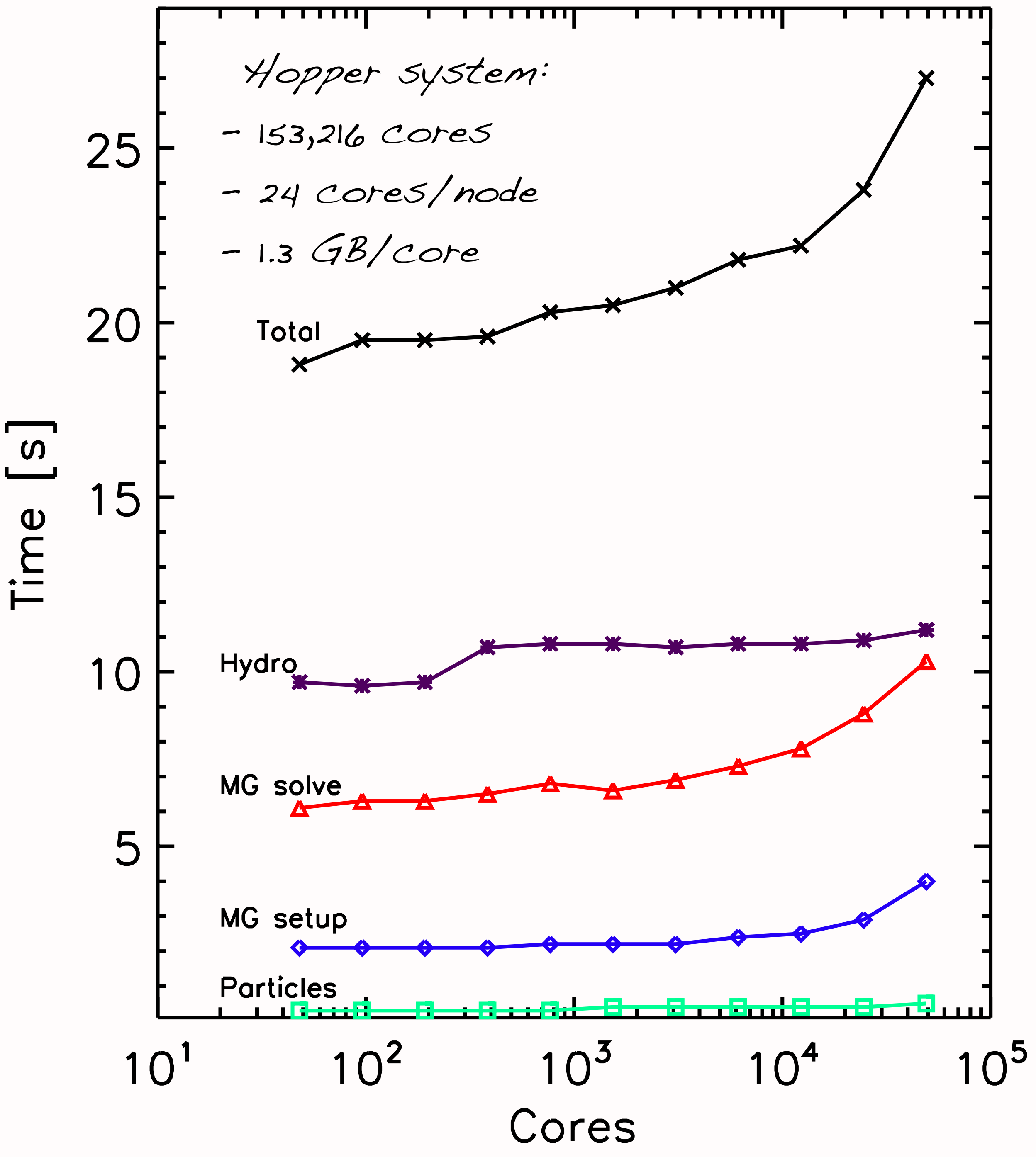


Gas temperature

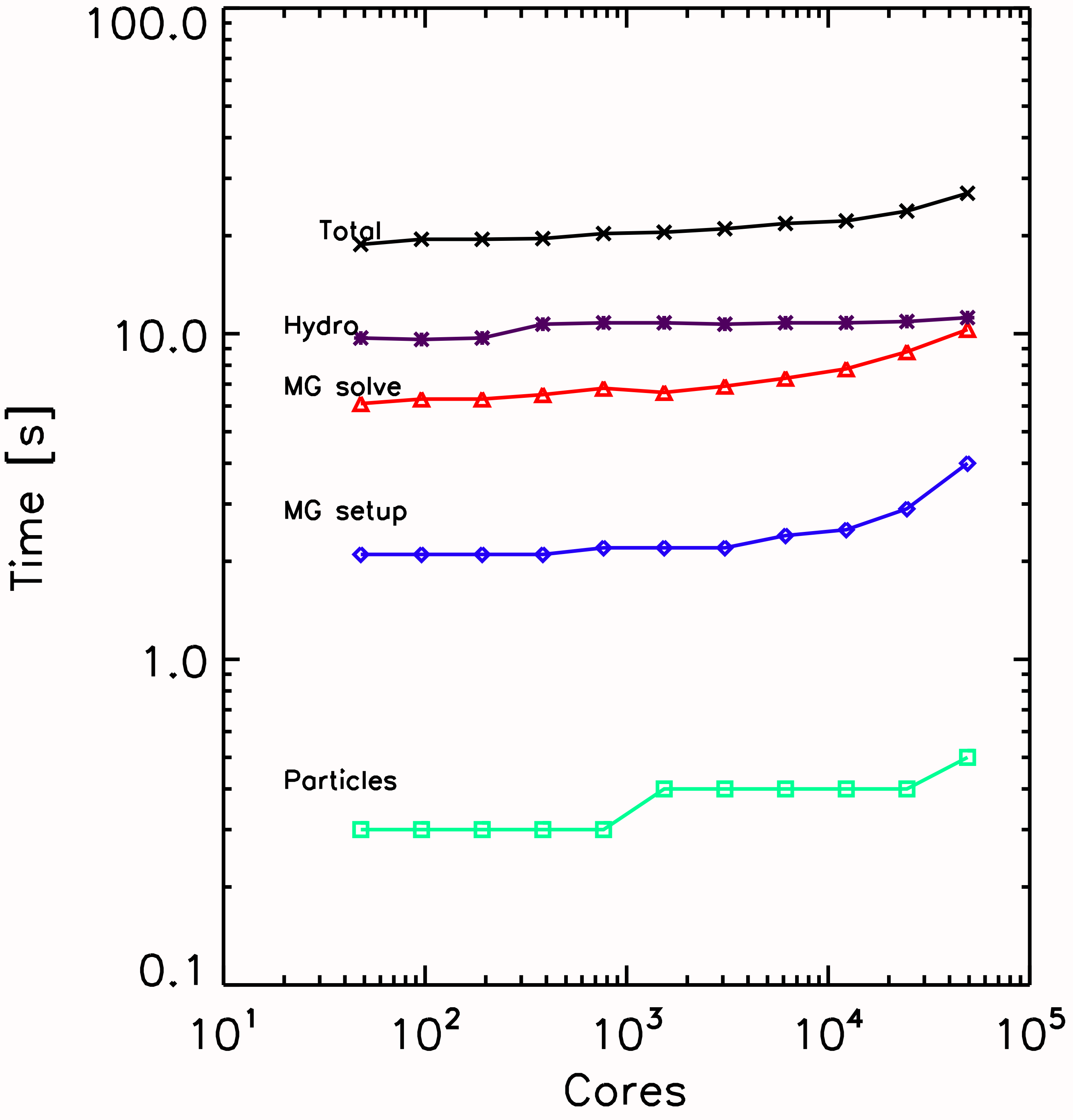


Almgren, Bell, Lijewski, Lukić, Van Anden: ApJ, 765, 39 (2013)

Scaling



Hopper system:
- 153,216 cores
- 24 cores/node
- 1.3 GB/core
Total
Hydro
MG solve
MG setup
Particles
Time [s]
25
20
15
10
5
10^1
10^2
10^3
10^4
10^5
Cores



Total
Hydro
MG solve
MG setup
Particles
Time [s]
100.0
10.0
1.0
0.1
10^1
10^2
10^3
10^4
10^5
Cores