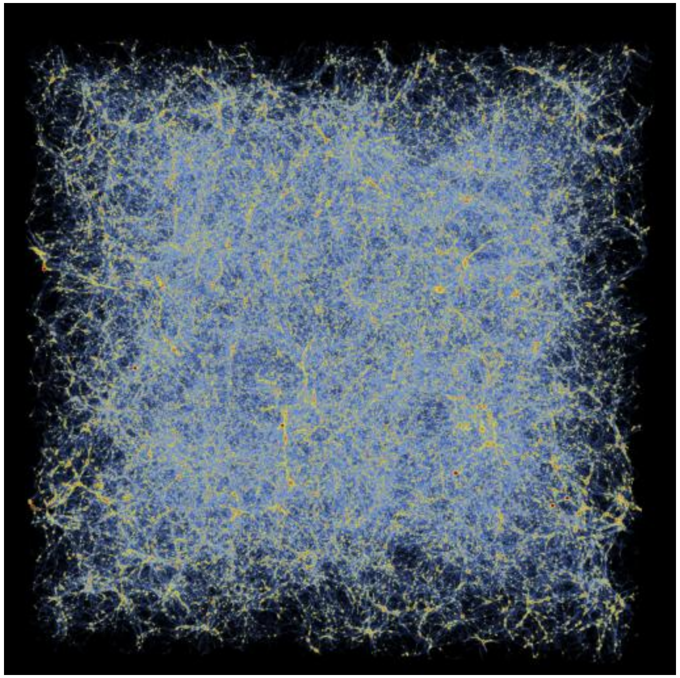
1. Cosmic Frontier/Structure Formation

List of Projects/PI(s)/Institution(s)

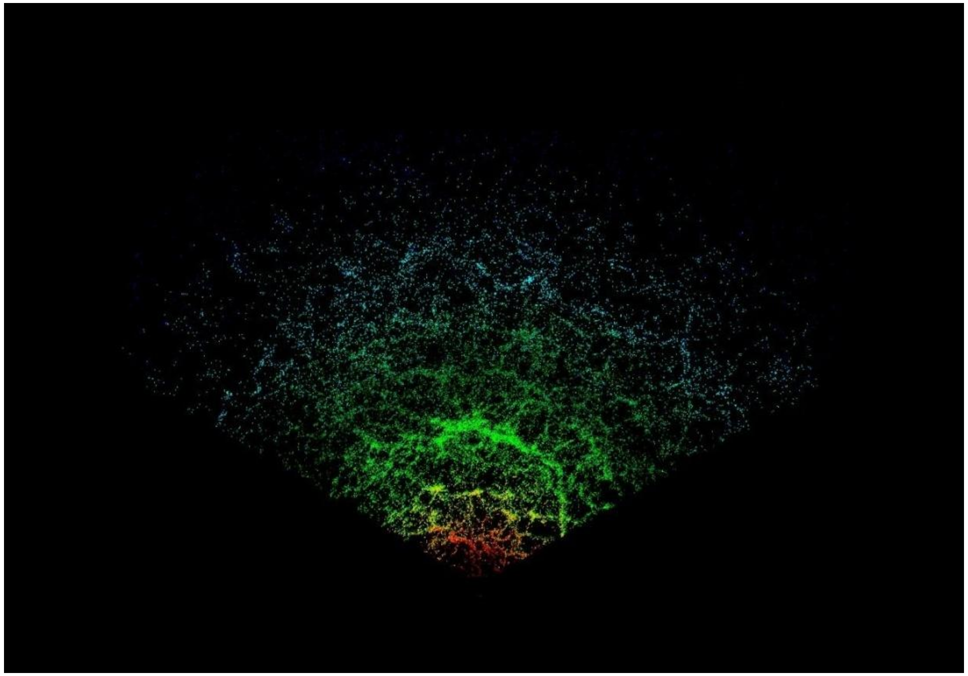
- ENZO simulations of Baryon Acoustic Oscillation (BAO) in the Lyman Alpha Forest and Galaxy Redshift Surveys (M. Norman, USCD)
- No HEP funding (cut from SciDAC2 CAC proposal)

Lyman alpha forest



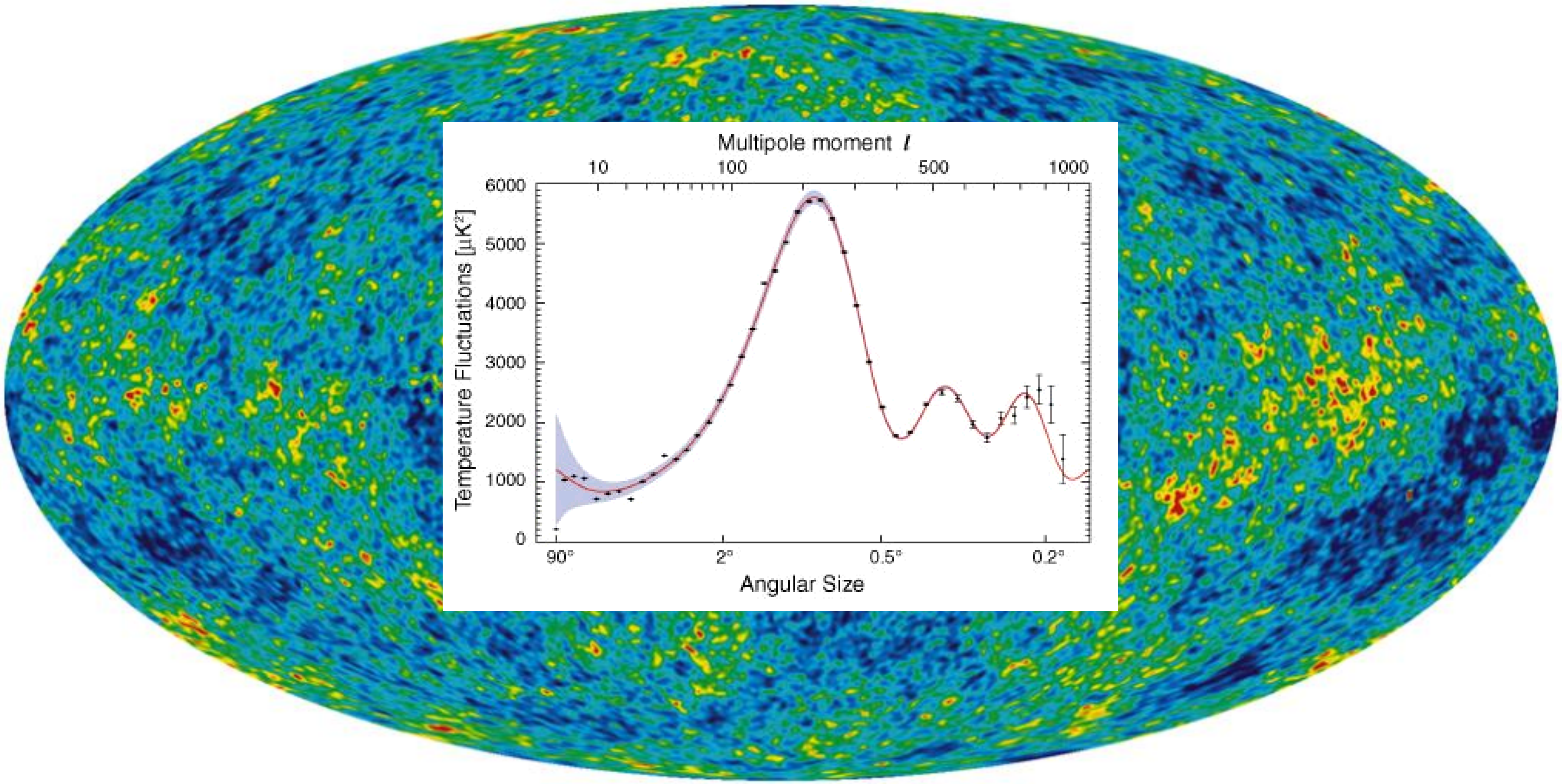
4096^3, 16,384 cores Kraken

Galaxy large scale structure

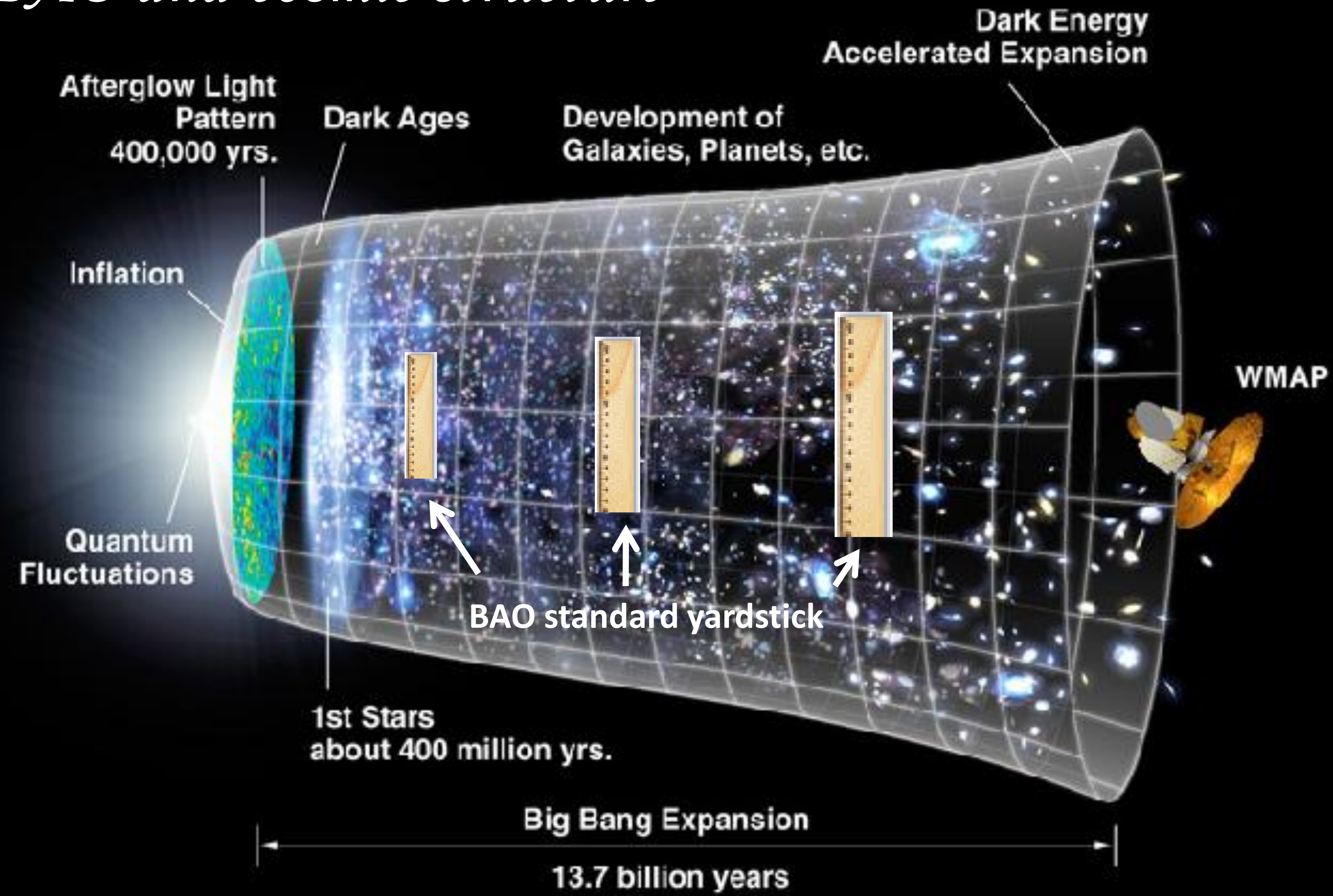


Sloan Digital Sky Survey

Baryon Acoustic Oscillations (BAO) in the Cosmic Microwave Background



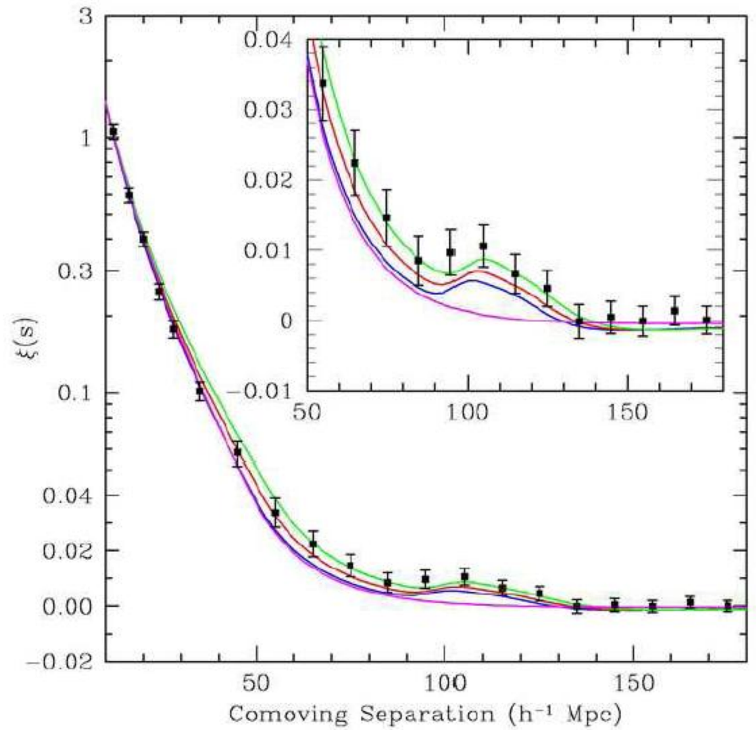
BAO and cosmic structure



Baryon Acoustic Oscillations (BAO)

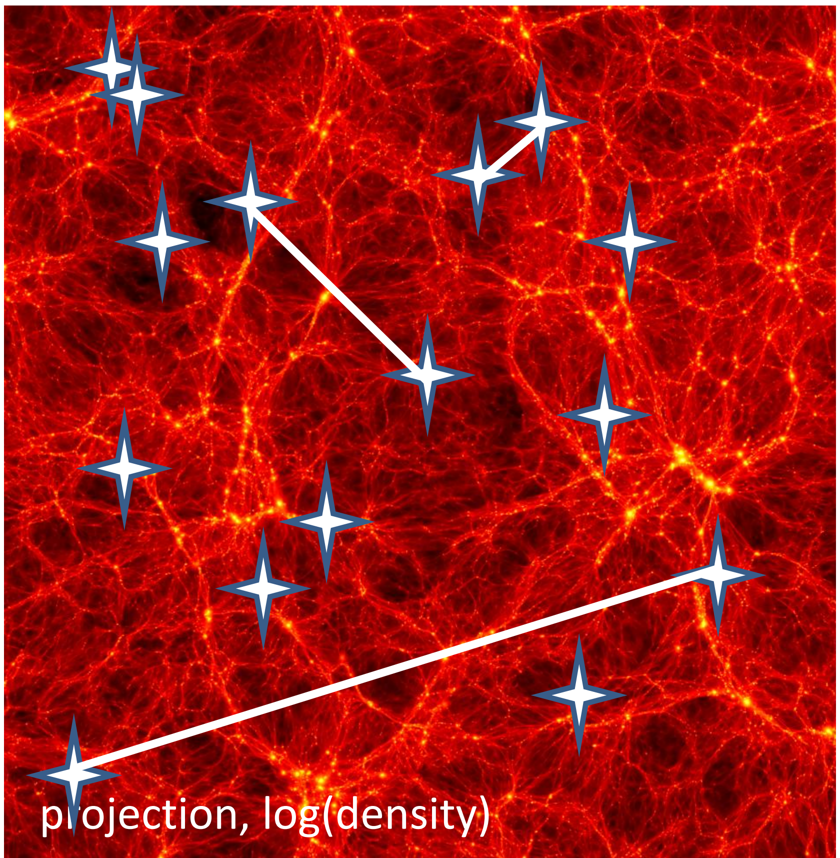
- Imprint on the matter power spectrum P(k) due to acoustic oscillations of the baryon-photon fluid prior to recombination
- Serves as a standard ruler, calibrated by CMB
- Measure $d_A(z)$ and $H(z)$ from large galaxy redshift surveys
- Systematics requiring extreme scale simulation
  - Effect of nonlinearity
  - Redshift space distortions
  - Complex galaxy bias

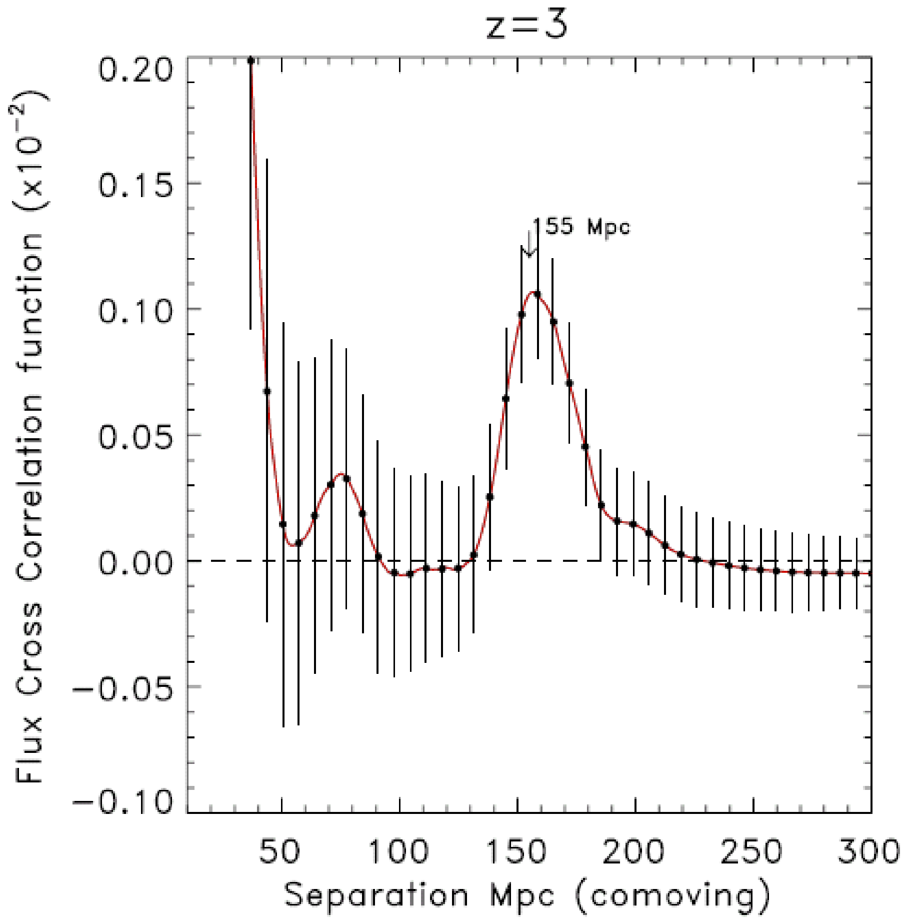
Detection of BAO in SDSS luminous red galaxy LSS Eisenstein et al. (2005)



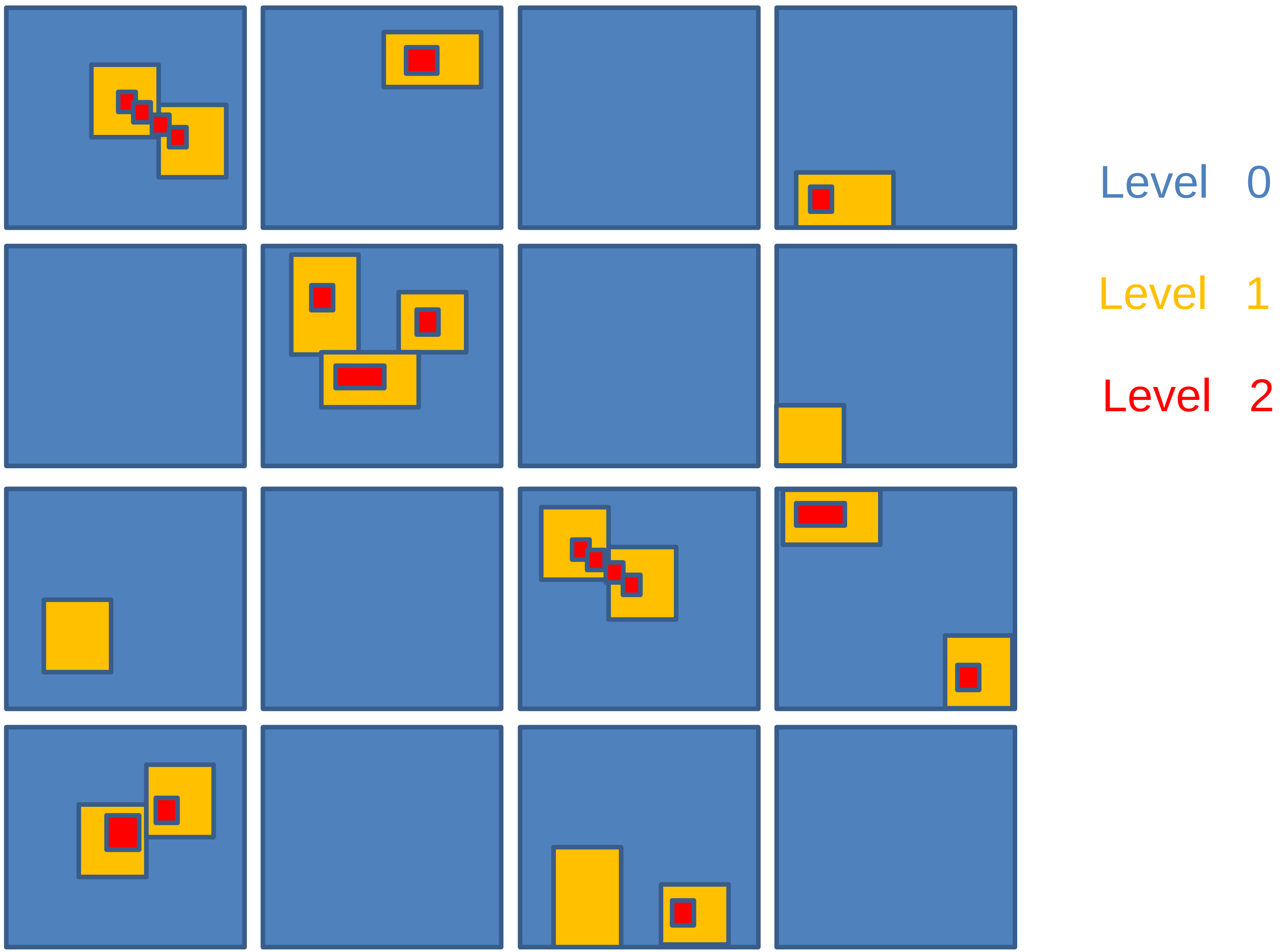
First Self-consistent Simulation of Baryon Acoustic Oscillations in the Intergalactic Medium
Michael Norman, Robert Harkness, Pascal Paschos, UCSD

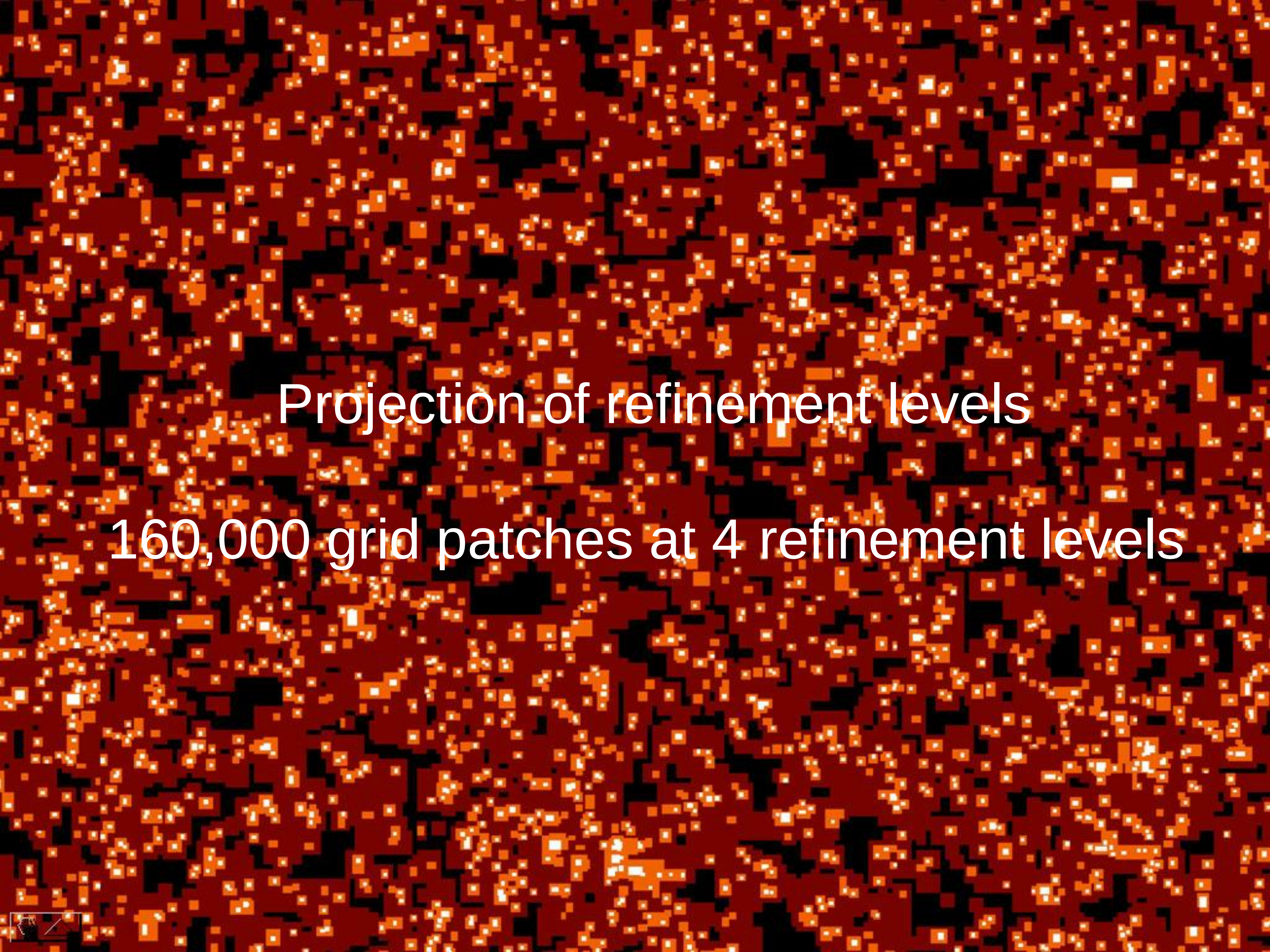
2048^3 cell/particle hybrid simulation, 330 Mpc volume, ENZO code
2048 procs, 1.2 million CPU-hrs, 6 TB RAM, 200 TB output, 6 month job , NERSC Seaborg
2006 INCITE award; 2006 Joule Metric application

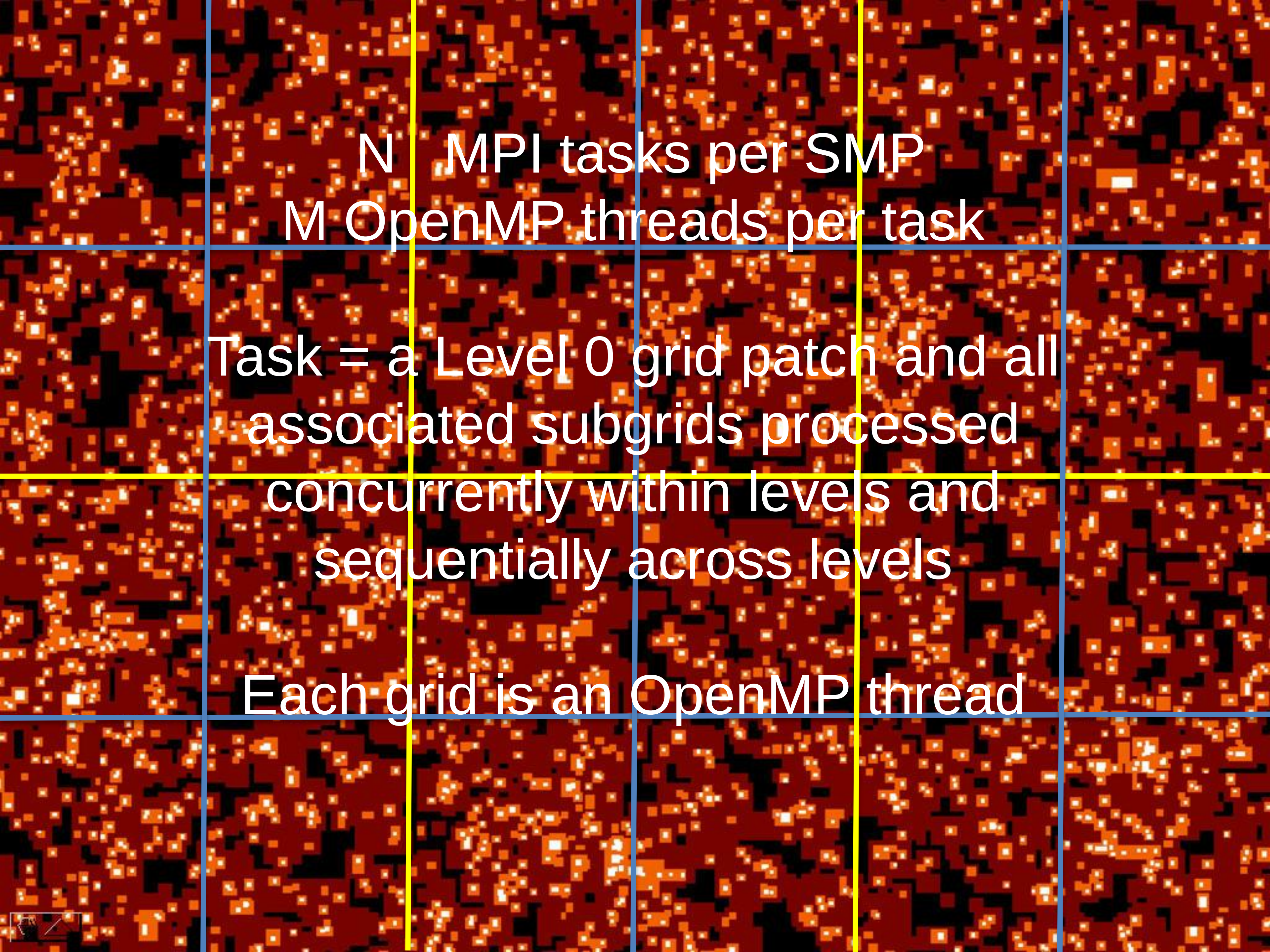




AMR = collection of grids (patches); each grid is a C++ object




Projection of refinement levels
160,000 grid patches at 4 refinement levels


N MPI tasks per SMP
M OpenMP threads per task
Task = a Level 0 grid patch and all associated subgrids processed concurrently within levels and sequentially across levels
Each grid is an OpenMP thread

2. Current HPC Requirements

- Architectures

  Cray XT, IBM PowerX
- Compute/memory load

  1024^3 AMR→2 MSU, 4 TB RAM

  4096^3 non-AMR→4 MSU, 15 TB RAM
- Data read/written

  1024^3 AMR→4 TB restarts, 100 TB saved

  4096^3 non-AMR→8 TB restarts, 200 TB saved
- Necessary software, services or infrastructure

  SPRNG, 3D FFT, HYPRE solver
- Current primary codes and their methods or algorithms

  ENZO: block AMR, PM N-body, PPM gas dynamics, FFT+MG Poisson solver; hybrid MPI/OpenMP
- Known limitations/obstacles/bottlenecks

  Scalability of ENZO AMR infrastructure and N-body solver

  Massive I/O can be mitigated by inlining, but not eliminated

3. HPC Usage and Methods for the Next 3-5 Years

- Upcoming changes to codes/methods/approaches

  OO redesign/reimplementation ENZO AMR infrastructure for petascale

  UPC reimplementation of ENZO N-body solver
- Changes to Compute/memory load

  ENZO runs increasing in core counts to 100,000; 100 TB MEM; 100 M SU
- Changes to Data read/written

  4096^3 AMR→20 TB restarts, 200 TB saved

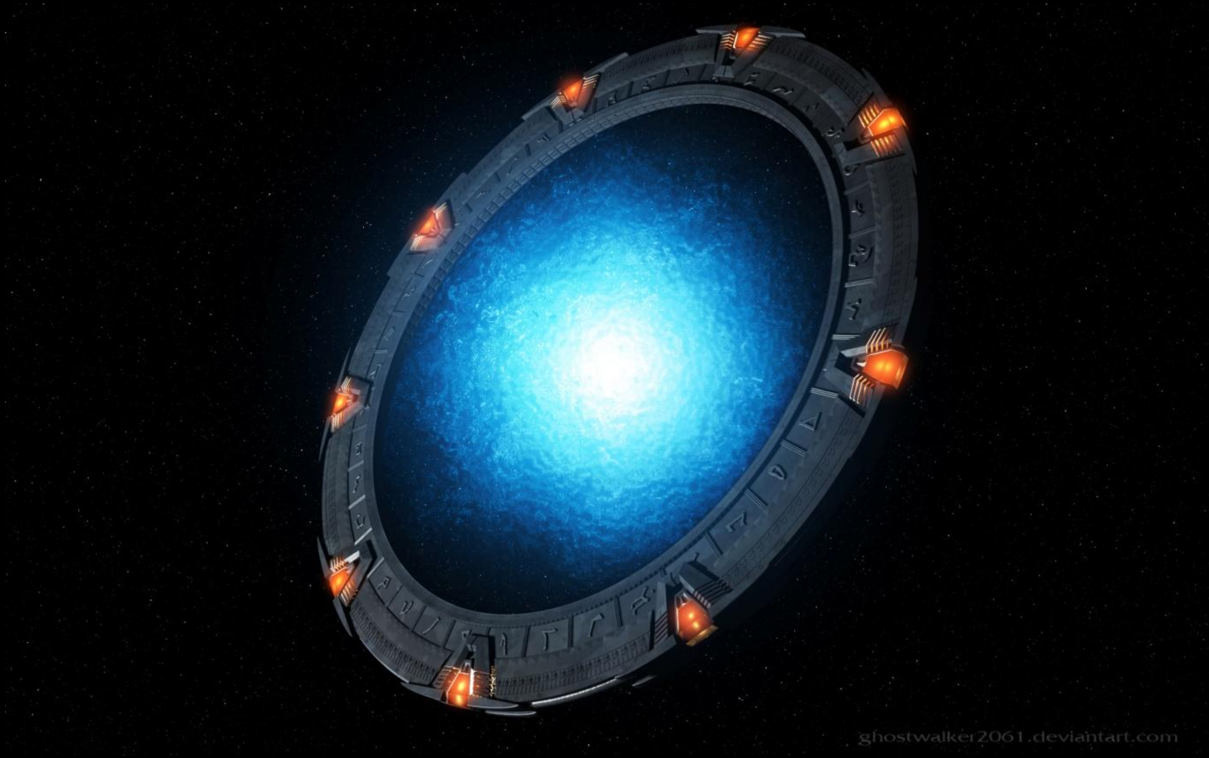
  8192^3 N-body→35 TB, 350 TB saved
- Changes to necessary software, services or infrastructure

  exploit UPC/UPC++ to re-engineer ENZO

  more inlining of analysis functions

  automatic data migration to SDSC over ESnet

Project StarGate

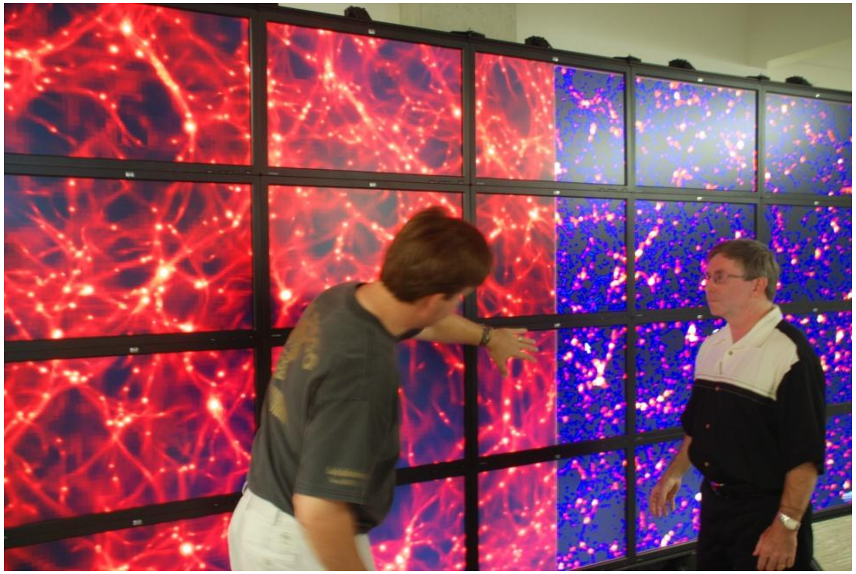


ANL * Calit2 * LBNL * NICS * ORNL * SDSC

Project StarGate Goals

- Explore Use of OptIPortals as Petascale Supercomputer "Scalable Workstations"
- Exploit Dynamic 10 Gbs Circuits on ESnet
- Connect Hardware Resources at ORNL, ANL, SDSC
- Show that Data Need Not be Trapped by the Network "Event Horizon"

OptIPortal@SDSC



Rick Wagner Mike Norman

- ANL * Calit2 * LBNL * NICS * ORNL * SDSC

StarGate Network & Hardware

ALCF

DOE Eureka
100 Dual Quad Core Xeon Servers
200 NVIDIA Quadro FX GPUs in 50
Quadro Plex S4 1U enclosures
3.2 TB RAM

ESnet

Science Data Network (SDN)
> 10 Gb/s fiber optic network
Dynamic VLANs configured
using OSCARS

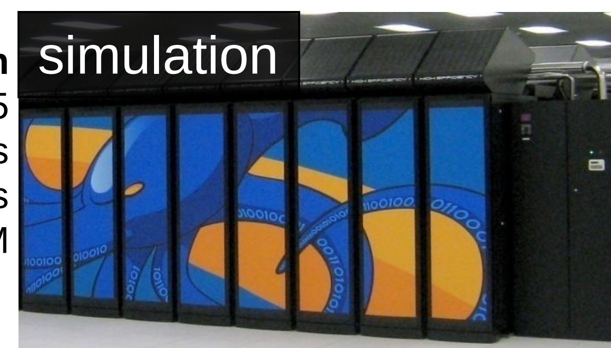
SDSC

Calit2/SDSC OptIPortal1
20 30” (2560 x 1600 pixel) LCD panels
10 NVIDIA Quadro FX 4600 graphics
cards > 80 gigapixels
10 Gb/s network throughout

Challenge: Kraken is not on ESnet

NICS

NSF TeraGrid Kraken
Cray XT5
8,256 Compute Nodes
99,072 Compute Cores
129 TB RAM



- ANL * Calit2 * LBNL * NICS * ORNL * SDSC

StarGate Streaming Rendering

ESnet

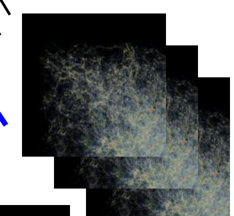
ALCF Internal

3 A media bridge at the border provides secure access to the parallel rendering streams.

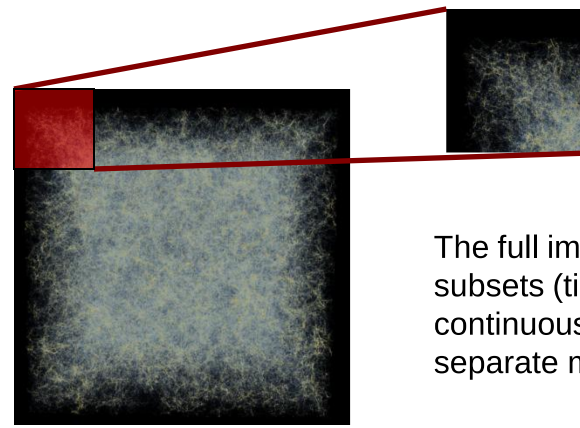
gs1.intrepid.alcf.anl.gov

SDSC

ALCF



5 Updated instructions are sent back to the renderer to change views, or load a different dataset.



2

The full image is broken into subsets (tiles). The tiles are continuously encoded as a separate movies.

4 flPy, a parallel (MPI) tiled image/movie viewer composites the individual movies, and synchronizes the movie playback across the OptIPortal rendering nodes.

1 Simulation volume is rendered using vl3 , a parallel (MPI) volume renderer utilizing Eureka's GPUs. The rendering changes views steadily to highlight 3D structure.

3. HPC Usage and Methods for the Next 3-5 Years

- Anticipated limitations/obstacles/bottlenecks on 10K-1000K PE system.

  petascale AMR must distribute grid hierarchy metadata and improve mem/cpu workload➔dynamic process migration

- Strategy for dealing with multi-core/many-core architectures

  In principle, AMR has more than enough work to keep 1000 core nodes busy, but will need very sophisticated runtime support for placing/moving data ➔UPC?

4. Summary

- Recommendations on NERSC architecture, system configuration and associated service requirements needed for your science:

  - fundamental problem of all gravity calculations is workload and memory imbalance as structure formation proceeds

  - Architecture must have sufficiently high "memory ceiling" per node, and sufficiently high intra/internode BWs to accommodate dynamic but imperfect load balancing

  - Provide highly optimized PGAS to express global AMR grid metadata and N-body particle lists

- What significant scientific progress could you achieve over the next 5 years with access to ~50X NERSC resources?

  Calibrated, validated dark energy surveys

- What "expanded HPC resources" are important for your project?

  Parallel programming models expressing multilevel parallelism, abstract data structures, and process migration → UPC++?

- Any other special needs or NERSC wish lists?

  Hardware support for small messages