Computational Cosmology Collaboration Proposal

FNAL LBNL SLAC BNL

- Overview by Stuart Marshall

Modern Cosmological Simulations

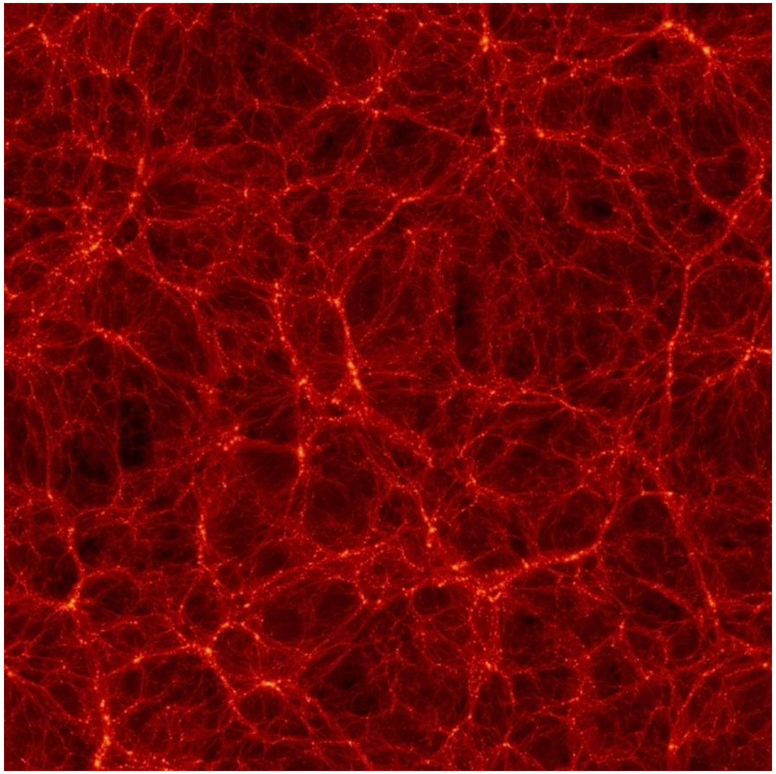
- **N-body**
- Dark matter particles
- Large scales
- Useful for statistical studies of matter distribution, groups, clusters
- For generating synthetic catalogs for DES, LSST, etc.
- “Gadget” widely used

- **Hydro**
- N-body + gas-dynamics
- Gas cooling
- Star formation, feedback
- Radiative transfer
- Other physics
- “Enzo”

Modern Cosmological Simulations

N-body ⟺ **Hydro**

- Hydrodynamic simulations significantly more expensive (10x)
- Hydro sims needed to model baryonic physics
- Hydro results allow modifying N-body results to include baryonic effects (galaxies added)



Computing Scales

- Millions of cpu-hours per simulation
- 10's TB per run
- Many sim runs to cover parameter space
- Coordination/Data-management becoming important
- Data storage/access critical to using and sharing results

Collaboration Principals

- FNAL Frieman, Dodelson, Gnedin
- LBL White, Seljak, Linder
- SLAC Wechsler, Abel
- BNL Slosar, Ma, Sheldon
- KICP Kravtsov

Computational Cosmology Collaboration Goals

- Carry out and organize computational cosmology research efforts hosted by the participating labs in support of DOE-HEP experiments and science goals
- Engage observational and experimental physicists with the theoretical and computational cosmology community
- Develop, maintain, and support simulation analysis tools and public state-of-the-art adaptive mesh refinement cosmological hydrodynamics codes for the community.

These tasks have been taking place independently to date.

Computational Cosmology Collaboration Implementation

- Install and maintain medium-scale computational facilities enabling
  - development and optimization of cosmological codes for petascale machines
  - applied computations that require fast turnaround
  - interactive data analysis and visualization
  - analysis of the largest data sets produced at the leadership class facilities
- Deliver the Cosmology Data Grid
  - a data repository of curated numerical simulations using standardized data formats, derived data products, and observational data.
  - a user facility to share and disseminate simulation and observational data within and outside the collaboration.
- Provide three FTEs, whose responsibilities will include
  - training and user support for the software developed by the collaboration
  - performance testing and supporting ongoing research to scale codes for leadership class computers
  - supporting the development of the cosmological data grid including curating the data products generated and organized by the collaboration.

Organize

DES, BOSS, LSST etc. require a variety of simulations to extract science.
A project-like approach is needed.

Table 1: Required Simulation Resources for Primary Cosmological Probes over 5 years

Probe	Science	CPU Resources (Million CPU Hours)	Storage Resources (TBytes)
Galaxy Distribution	DE, Inf., m_ν	120	800
Weak Lensing	DE	60	400
Lyman-alpha	m_ν, Inf.	150	1,500
21-cm	DE, m_ν, Inf.	100	400
DM Detection	DM	100	300
Total		530	3,400

Access, curation, planning, standardization needed to exploit these results.

Engage

- Software support and development
  - Support state-of-the-art hydrodynamical codes
    - ART (Fermilab)
    - Enzo (SLAC)
    - Nyx (LBNL)
  - Bring codes to the petascale
  - Develop and support a suite of simulation analysis tools (e.g. halo finding,

  ray tracing, merger trees, mock catalog creation)
    - building on yt analysis framework (http://yt.enzotools.org)
- Data curation through the cosmology data grid
  - Currently a lot of duplication and no standardization, hard to make existing simulations public even when desired.
  - Coordinated effort with standard products

The Ask

- Support for local medium-scale facilities
- enables fast turnaround for intermediate runs
- enables analysis of largest leadership class runs
- enables visualization and interactive analysis

Substantial increase in productivity when these machines are designed to the problem and locally maintained.

Model is 4 sites with 2 major, 2 minor. Assuming (late 2010) $350/core and $500/TB and using the cpu-hours shown above, cost is $1.9M/yr for 5 years.

Support for 3 FTE's also requested.

Status of Proposal

- Presented several times to DOE, most recent at Argonne in Feb 2011.
- Need for data products seems to be accepted
- Somewhat less consensus on need for a “project” to organize the groups and generate the data and tools.
- Difficulty getting backing for local resources approach. Making the case for effectiveness of local facilities over national scale facilities has not been fully successful.

Future prospects

- Phase the discussion: first develop consensus that this needs to be carried out as an organized effort (as opposed to outcome of these groups working independently)
- Document the workflows for these simulations. Examples of existing DES mock catalog generation sims, also numerous cases of published results from our local systems.
- Enlist assistance from OCIO to support user needs in transferring data to SLAC, and in making it available to collaborators outside. (This is for “average” users rather than large projects) Current rates in/out of SLAC make this cumbersome.
- Study some examples of computing on local ~1000's core systems vs. large facilities like TACC.
- Document how these codes are developed and tested at small and medium scale and what the constraints are on the larger facilities.
- Ask community for help in making the case. KIPAC's MPI resources are now pretty old and we need 5 times the capacity to continue the successes of these programs. (These resources support cosmology simulations and many other simulations and analyses)