SIMULATING THE FIRST GALAXIES AND QUASARS: THE BLUETIDES COSMOLOGICAL SIMULATION

Allocation: NSF/2.63 Mnh
PI: Tiziana Di Matteo[1]
Collaborators: Rupert Croft[1]; Yu Feng[1]; Nishikanta Khandai[2]; Nicholas Battaglia[1]

[1]Carnegie Mellon University
[2]Brookhaven National Laboratory

SCIENTIFIC GOALS

Computational cosmology—simulating the entire universe—represents one of most challenging applications for petascale computing. We need simulations that cover a vast dynamic range of space and time scales and include the effect of gravitational fields that are generated by (dark matter in) superclusters of galaxies upon the formation of galaxies. These galaxies, in turn, harbor gas that cools, makes stars, and is being funneled into supermassive black holes that are of the size of the solar system.

We have carried out a full-machine run on Blue Waters, the BlueTides cosmological simulation. It is run with an improved version of the cosmological code P-Gadget. The simulation aims to understand the formation of the first quasars and galaxies from the smallest to the rarest and most luminous, and the role of these processes in the reionization of the universe. The simulation is being used to make predictions for what will be seen by the upcoming WFIRST and James Webb Space Telescope (JWST; successor to Hubble, launch planned for 2018).

ACCOMPLISHMENTS TO DATE

The largest telescopes currently planned aim to study the "end of the Dark Ages" epoch in the early universe, when the first galaxies and quasars form and reionization of the universe takes place.

Our main production run, BlueTides, was started in 2014, and has successfully used essentially the entire set of XE6 nodes on the Blue Waters machine. It is following the evolution of 700 billion particles in a large volume of the universe (600 co-moving Mpc on a side) over the first billion years of the universe's evolution with a dynamic range of 6 (12) orders of magnitude in space (mass). This makes BlueTides by far the largest cosmological hydrodynamics simulation ever run. BlueTides follows not only hydrodynamics, but also includes models for what is colloquially known as the "full physics" of galaxy formation, such as radiative processes, and subgrid models for star and black hole formation and energy release.

BlueTides includes a complicated blend of different physics that is non-linearly coupled on a wide range of scales, which leads to extremely complex dynamics. Without significant investment of effort in code development, including radical improvements in efficiency and load balancing, it would have been impossible to carry out the run. We have also carried out a program of model validation, with several improvements to the physical models in P-Gadget. We detail both of these aspects below.

A pressure-entropy smoothed particle hydrodynamics (SPH) formulation from Hopkins (2013) replaces the old density-entropy formulation which had been in use since 2002 and was known to suppress phase mixing in a non-physical way. We also improved the effective SPH resolution with a higher-order quintic kernel that reduces the shot noise level by a factor of two without additional memory usage.

In the regime of the simulation, the star formation is supply limited, thus it is important to consider the abundance of H_2 molecules that is the direct supply of star-forming interstellar gas. We implemented a molecular H_2 gas model based on work by Gnedin et al. (2008).

The simulation regime also overlaps the interesting Epoch of Reionization, where the entire universe turns from opaque to transparent. Traditionally a uniform ultraviolet (UV) ionizing radiation field is introduced at the same time across space in hydrodynamical simulations. This is no longer a good approximation in a simulation of such large volume. We have incorporated a patchy reionization model from Battaglia et al. (2013); the model introduces a UV field based on a predicted time-of-reionization at different spatial locations in the simulation.

We improved the code infrastructure in several ways:

- **Memory efficiency.** We detached the black hole particle data from the main particle type, reducing the memory usage by one quarter for a problem of the same size. This allows us to model 700 billion particles using all of Blue Waters, while leaving some room for potential node failures.

- **Low maintenance.** The redundant code in all major physical modules has been rewritten based on a new tree walk module. This in turn inspired the following improvement in the threading efficiency.

- **Threading efficiency.** We replaced global critical sections with per-particle (per-node) spin locks. Because the boundaries of thread subdomains are very small, the spin locks hugely improved threading efficiency. Even though the domain decomposition and Fourier transform remain sequential, the wall time improved by about a factor of two at 32 threads. The improved threading efficiency allows us to use fewer domains, which in turn further reduces the complexity of domain decomposition and inter-domain communication, improving the overall efficiency of the code.

- **I/O.** We enabled HDF5 compression in the snapshot files. The compression reduces the size of a snapshot by ~30%-40%.

- **Mesh gravity solver.** Domain decomposition was an issue with the huge Fourier transforms (FT; up to 16,384^3) needed to carry out the large-scale mesh computation of the gravitational force. We have been able to improve the speed of this part by a factor of ten (as measured on Blue Waters) by upgrading the FT used to compute the gravitational force from a slab decomposition to a "pencil" decomposition, which allows the computation to be efficiently distributed among the 20,000+ XE6 nodes of Blue Waters.

We generated the initial conditions for the BlueTides simulation for a random realization of the density field as measured by the WMAP satellite. Our new massively parallel version of the Gadget hydrodynamics code, called "MP-Gadget3," was used to evolve these initial conditions forward in time. To date, the BlueTides run has successfully reached redshift z=8, which is 650 million years after the Big Bang. At this epoch, the universe has produced the first generations of stars, galaxies, and black holes, and the intergalactic hydrogen gas which pervades space at this epoch has fully reionized. This represents a major milestone in the history of this simulated universe and means that we have reached our most important goal, the end of the epoch of reionization.

We are carrying out a program of analysis of the simulation outputs so far, finding galaxies and early black holes. We detail below some of this work, which we anticipate will lead to several journal publications. In tandem with this analysis, we hope to evolve the simulation further. Our ultimate goal is to reach redshift z=6, a period that is one gigayear after the Big Bang, when the first bright quasars appeared. These extreme objects powered by black holes of a billion solar masses represent one of the most difficult tests of modern cosmological theories, and our simulation should be the first to have both the size and the mass resolution to include the whole range of galactic systems present at these times, from dwarf galaxies through Milky Way-sized objects all the way up to the hosts of the brightest quasars.

The simulation so far has produced tens of snapshots (40 TB each) of data on particle positions and properties, a total of ~2 PB. After refining our own software to deal with these enormous data volumes, we have carried out the following initial analyses:

- **Imaging.** We have used our stellar modeling to compute the visible light distribution in the simulation, and by projecting this with an appropriate kernel, we have made multi-gigapixel maps of the simulated sky as it would be seen by an observer.

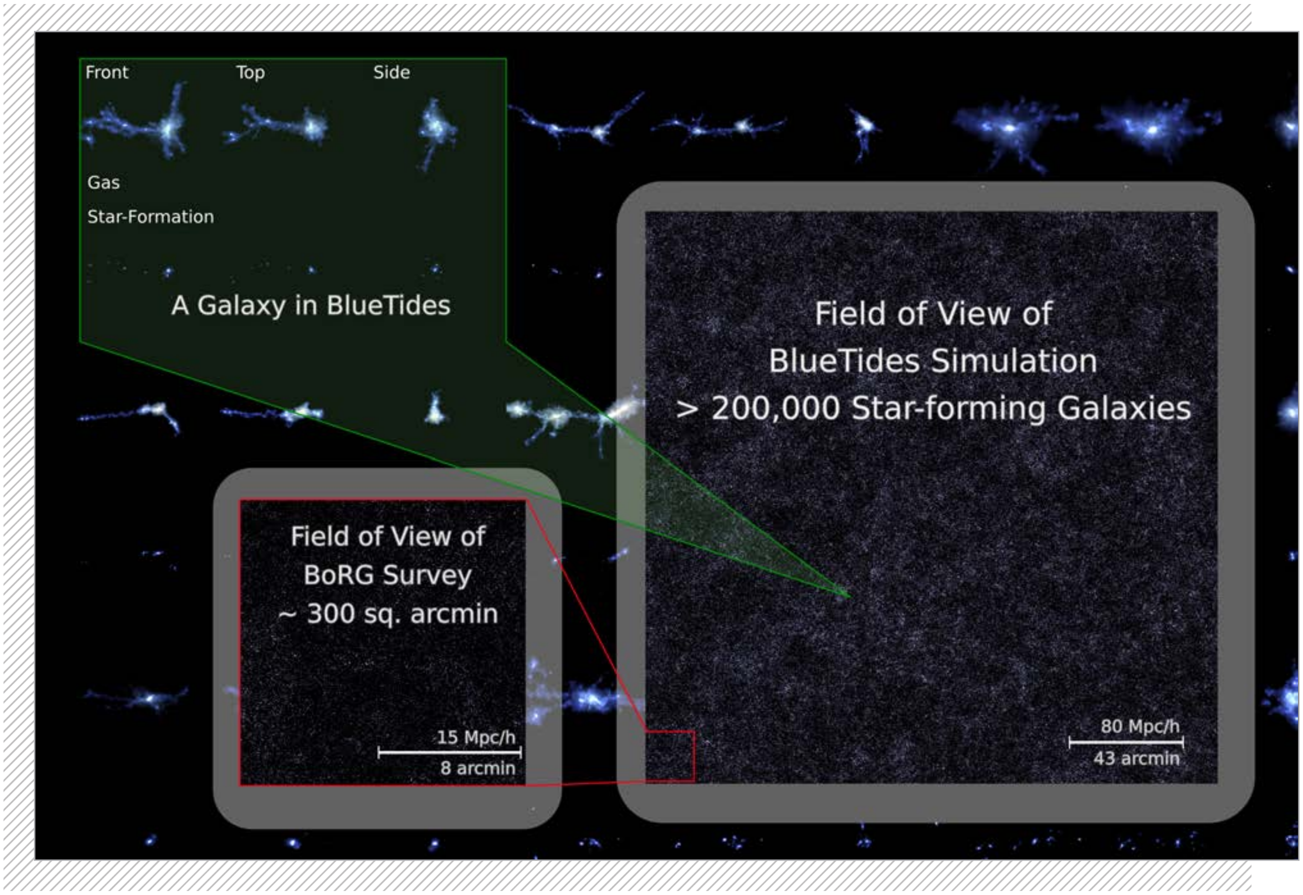
- **Galaxy selection.** We have applied observational selection algorithms (the widely used SourceExtractor software) to the simulated sky maps and created catalogs of millions of galaxies (to our knowledge the first time this has been done on this type of data). We have also used friends-of-friends algorithms on the 3D stellar particle data to create catalogs of galaxies, showing that this technique (the one commonly used by simulators) gives very similar results in most cases.

- **Galaxy and quasar luminosity functions.** From our simulated catalogs, we have computed statistical measures of the population of galaxies and quasars at different redshifts, showing consistency with some early data from the Hubble Space Telescope, and making predictions for upcoming surveys.

- **Galaxy morphology.** Using the high resolution of BlueTides, we have made detailed images of individual galaxies, uncovering a striking and unexpected population of large Milky Way-sized disk galaxies when the universe was 5% of its present age.

HOW BLUE WATERS PROJECT STAFF HELPED

Running such large jobs on a regular basis in a very timely fashion obviously requires advanced resource management, and the way the Blue Waters system and technical staff have been set up made this possible. The project staff also helped our team with MPI+OpenMP development of the Gadget simulation code and assisted with file handling issues including HDF and Lustre tuning. Processing the petabytes of simulation



output necessitated some radically new ways of thinking and Blue Waters staff helped greatly; for example, they participated in the development of a new parallel sorting algorithm for BlueTides. This was published with Blue Waters personnel in ACM's *Transactions on Parallel Computing.*

WHY THE RESEARCH MATTERS

Our simulations of the early universe blaze a trail for future calculations. We are evolving models forward in time for 1 billion years, rather than the 14 billion years necessary to cover the history of the universe to the present day. The science case for the early universe (which requires an enormous number of particles) allows us to carry out memory-limited computations on Blue Waters, thus pioneering the running of petascale simulations and handling and analysis of petabyte-scale data stores, but without the enormously longer runtime that would be needed to reach redshift z=0, the present day universe (which for the BlueTides run would take an unfeasible billion core-hours or more on Blue Waters).

WHY BLUE WATERS

A complete simulation of the universe at the epochs we are studying requires a small enough particle mass (i.e. high particle density) to model the dwarf galaxies that contribute significantly to the summed ionizing photon output of all sources. It also requires an enormous volume, of the order of one cubic gigaparsec (1 Gpc^3 is 3x10^{19} cubic light years) in order to capture the rarest and brightest objects, the first quasars. Previous calculations on smaller HPC systems have either fulfilled the first, but in a small volume, or the second, but with large particle masses and so only resolved large galaxies. The science that could be carried out with these earlier runs (which includes our previous MassiveBlack simulation, a 64 billion particle hydrodynamics simulation carried out on Kraken at NICS) was therefore incomplete. With Blue Waters, however, we have reached the point where the required number of particles (about one trillion) could be contained in memory, and the petascale computing power was available to evolve them forward in time. The Blue Waters project therefore made possible this qualitative advance, making possible what is arguably the first complete simulation (at least in terms of the hydrodynamics and gravitational physics) of the creation of the first galaxies and large-scale structures in the universe.

The application runs required essentially the full system. We used 20,250 nodes (648,000 core equivalents; the new version of the code can scale higher, but we left a safety margin) using 57 GB/node (89%) of memory. This application thus used 1.15 PB of memory—something only Blue Waters can provide, and which is nearly 90% of the available memory.

PRE-PETASCALE PREPARATION

Our team was given the only cosmology award in the first round of NSF Petascale Applications funding (OCI-0749212, PI T. Di Matteo) with a proposal to develop petascale hydrodynamics simulations by building on the Gadget code. Since 2007, we have been working on simulation algorithms, on-the-fly analysis methods, and all the other activities needed to make this happen. We have steadily built up the capabilities of Gadget with science production runs on Teragrid and XSEDE machines, from Big Ben through Ranger and Kraken until we have reached our goal on Blue Waters of petascale computations with the BlueTides run.

LOOKING FORWARD TO THE NEXT TRACK-1 SYSTEM

Evolution of the universe further in time could be carried out on future, more powerful Track-1 systems. As simulations reach later epochs, they overlap with larger and larger observational datasets, which allows more detailed comparisons and tests of new physics. Even without going to later redshifts, higher mass and force resolution could be achieved with more memory, which would allow the properties of galaxies to be resolved in greater detail.

COMMUNITY IMPACT

In the coming decade, a new generation of astronomical instruments, all in the billion dollar class will start making observations of the universe during the period of the first stars and quasars, and opening up the “last frontier” in astronomy and cosmology. Those that are specifically targeting this epoch as their highest priority include the Square Kilometer Array radio telescope, the NASA James Webb Space Telescope, the successor to Hubble, and several huge ground-based telescopes, such as the Thirty Meter Telescope, the European Extremely Large Telescope, and the Giant Segmented Meter Telescope, each of which has a collecting area an order of magnitude larger than the current largest telescopes. The scientific community has obviously decided that research targeting this epoch matters enormously. However, observing and experimenting without theories to test and models to build understanding will give a very limited return on this investment. It is therefore crucial to build equally powerful theoretical tools. The combination of petascale HPC resources and the newest simulation codes are the equivalent to billion dollar-class theory programs in cosmology. Our research matters because without it there are no reliable ways to know what cold dark matter cosmology predicts for the first stars and galaxies and their properties. The best, most complete, and largest simulations must be carried out in concert with these ambitious upcoming observational programs.

PUBLICATIONS

Battaglia, N., H. Trac, R. Cen, and A. Loeb, Reionization on Large Scales. I. A Parametric Model Constructed from Radiation-Hydrodynamic Simulations. *Astrophys. J.*, 776:2 (2013), 81, doi:10.1088/0004-637X/776/2/81.

Hopkins, P. F., A General Class of Lagrangian Smoothed Particle Hydrodynamics Methods and Implications for Fluid Mixing Problems. *Mon. Not. R. Astron. Soc.*, 428 (2013), pp. 2840-2856, doi:10.1093/mnras/sts210.

This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications.