Institute for Scientific Computing Research Annual Report, Fiscal Year 2007

ISCR Home | Privacy Policy

LLNL-WEB-405425

Livermore–UC San Diego Scientific Data Collaboration

Beginning in May 2005, the ISCR supervised and administered the Livermore–UC San Diego Scientific Data Management Collaboration (LUSciD), a joint effort in which scientists from LLNL and UCSD create mechanisms for collaborative scientific research involving the massive data sets required for extremely large-scale scientific simulation.

The LUSciD project was specifically selected to create a synergistic collaboration between two major supercomputing centers, LLNL and the San Diego Supercomputing Center at UCSD. The ultimate goal of this pairing is to utilize both institutions' expertise in hardware, software, and large-scale applications to improve the conduct of science through the provision of scientific data management technology that enables the organization, manipulation, and analysis of observational and simulation data.

Three topics highlight this project—two applications areas covering regional climate change and cosmology, as well as scientific data management that enables the technology. Within each of the three topics, the effort is led by a Principal Investigator at each institution, who are tasked with ensuring that the collaboration effectively utilizes the strengths of each participating organization and that the efforts are complementary.

This status report combines the work of the following researchers and organizations:

Tim Barnett and David W. Pierce
Climate Research Division
Scripps Institution of Oceanography

Mike Norman and Robert Harkness
Laboratory for Computational Astrophysics
University of California, San Diego

Randy Banks and Dan Reynolds
Mathematics Department
University of California, San Diego

Reagan W. Moore, Sifang Lu, and Arun Jagatheesan
San Diego Supercomputer Center

Providing Scientific Data Management Technology for Enabling Large-Scale Simulations

1. Introduction

The goal of the joint LLNL/UCSD scientific data management project is to improve the conduct of science through the provision of scientific data management technology that enables the organization, manipulation, and analysis of observational and simulation data in distributed collaborations. This progress report covers two exemplary scientific applications: Global climate modeling to determine the impact of climate change on water supply and Cosmology studies of the structure of the early universe. The Cosmology studies in turn are being used to simulate the images that will be seen by the Large Synoptic Survey Telescope (LSST). The scientific applications drive the requirements for scientific data management by generating large simulation output files at LLNL, moving the data to storage systems at SDSC, and publishing derived data products in digital library technology for use by the broader research community. The data management tasks include the establishment of data grid technology on the Green Data Oasis disk cache at LLNL, and the demonstration of advanced data management technology for use by the LSST project.

2. Climate Simulations

The overall objective of the climate portion of LUSciD is to see if there is a detectable change in the hydrology of the western U.S. and, if so, whether it is due to anthropogenic climate change effects. The climate research is organized in four main areas:

1. changes in temperature
2. changes in streamflow and runoff
3. changes in snowpack
4. new methods for dynamical downscaling of global climate models

The climate simulation collaboration is led by Tim Barnett of UCSD's Scripps Institute and Doug Rotman of LLNL. Work during the second year has concentrated on acquisition and analysis of existing global climate models, downscaling their global fields to 1/8 degree resolution and then using those climate data in hydrological models to estimate changes in western water availability. This has been a huge computational effort, involving researchers at SIO and LLNL.

Results, still tentative, suggest a human-induced signal can be seen in the western snow pack, river flow and temperature. This work is being checked now and papers prepared for immediate submission. Extensive presentation of results took place at the American Geophysical Union annual meeting held in San Francisco, CA on December 10–14, 2007. Coordinated press releases will be set up with LLNL at a special session we our having on our project at the December AGU meeting.

The current research focus is on the Detection and Attribution (D&A) analysis of the river flow, temperature and snow water content. Another round of collaboration is just starting on fine-scaled D&A analyses and air temperature issues. Detailed variations of the D&A analyses are getting full participation from both LLNL and SIO staff.

For summer student participation, we are still waiting for LLNL to get clearances so our postdoc can visit them. Five joint papers are in preparation right now. Adequate funds remain to finish both the D&A work and an analysis of the Colorado River. All in all, the joint project has established a strong collaboration between the researchers at UCSD/SIO and staff at LLNL.

2.1 Climate Simulation Data Management

Sifang Lu has been collaborating with Dave Pearson and Hugo Hidalgo of Scripps Institute of Oceanography, as well as LLNL researchers, including Bala Govindaswamy and Celine Bonfils, on the management of the simulation output. The Storage Resource Broker (SRB) data grid is used to manage the scientific data for the global climatic change research.

The climate simulation data are organized into five major collections (GFDL, MIROC, PCM, CCSM3, vic) and stored on four storage resources at SDSC (GPFS-WAN file system, HPSS archival storage system, SamQFS archive, srbbrick9 file system). Data is copied onto Green Data Oasis disk storage at LLNL under the direction of Celine Bonfils for collaborative analyses. The total amount of data managed at SDSC in support of the climate simulations is:

- Total number of files: 2,362,771
- Total Size: 17,462,424,050,370 bytes (17 TB)

3. Cosmology Simulations

The overall objective of the cosmology simulations is to generate synthetic “wide and deep” images of the universe such as will be generated by the Large Synoptic Survey Telescope (LSST) when it begins operation in 2013. One of LSST’s key science goals is to map the dark matter distribution of the universe using gravitational lensing measurements of distant galaxies. With an eventual sample of more than 1 billion galaxies, LSST should additionally be able to measure the cosmic dark energy parameters w and w’ provided systematic effects are well understood. The purpose of the cosmology simulations is to calibrate these systematic errors by generating realistic images that can be processed through the LSST science analysis pipeline. The cosmology research is organized into four main areas:

1. Simulating cosmic structure “on the lightcone” for a large area of the sky
2. Generating synthetic LSST images including the effects of gravitational lensing
3. Construction of a digital archive for publishing the lightcone simulation results
4. Improving galaxy formation physical modeling through the incorporation of radiation transport effects into the cosmology simulation code Enzo

The cosmology simulation effort is led by Michael Norman of UCSD’s Center for Astrophysics and Space Sciences, Scott Olivier at LLNL for the LSST-related effort, and Frank Graziani at LLNL for the radiation transport-related effort. Work in the second year has focused on completing the lightcone simulations on LLNL Thunder, moving the data to UCSD for archiving and scientific analysis, and performing verification tests of the radiation transport algorithm recently installed into Enzo. Some highlights follow.

3.1 Lightcone Simulations

The lightcone is constructed from 16 large cosmological simulations, each one simulating cosmic structure over a specific redshift interval. By stacking up the lightcone “tiles” end-to-end, we are able to simulate the distribution of galaxies and galaxy clusters in the universe on a large area of the sky to a redshift depth of three. Each tile represents the largest adaptive mesh refinement (AMR) hydrodynamic cosmological simulations ever carried out. The entire lightcone series consumed over 1 M cpu-hours on Thunder and generated over 100 TB of data. Figure 1 shows an AMR volumetric rendering of one of the

tiles covering an unprecedented range of scales of over 65,000. As a first application of the lightcone, we have simulated distortions in the cosmic microwave background radiation field due to hot gas bound the galaxy clusters via the Sunyaev-Zeldovich effect. These distortions will be measured by the Planck satellite set to begin gathering data in 2008. This work is published in the Astrophysical Journal.

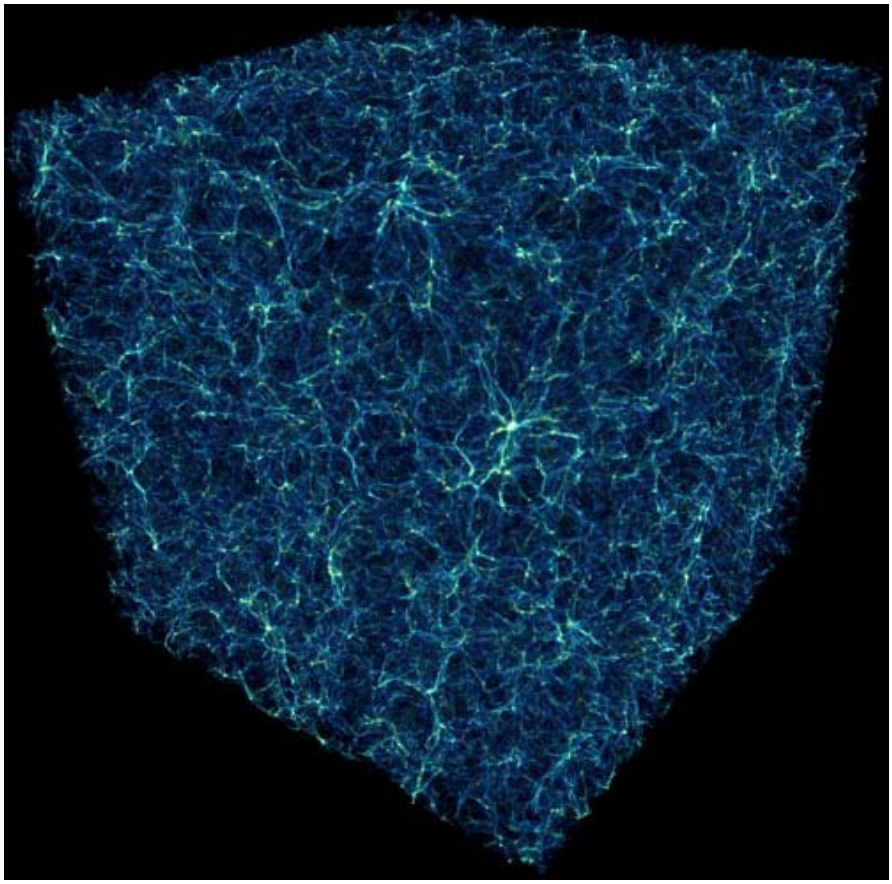

Figure 1. Volumetric rendering of cosmological structure in a portion of the lightcone simulation carried out at LLNL as a part of the LUSciD project. The AMR code Enzo was used to simulate the distribution of galaxy clusters in a volume 2 billion light years on a side with an effective resolution in the cluster cores of a $(65,536)^3$ grid.

3.2 Radiation Transport in Enzo

Radiation transport has been installed into the Enzo code and is undergoing verification testing at the present time in collaboration with LLNL B-division's John Hayes. The algorithm is novel in that flux-limited radiation diffusion is implicitly coupled to both the gas energy equation and gas ionization kinetics. This promises greater accuracy in AMR simulations covering a broad range of physical conditions. The algorithm was designed and implemented by Dan Reynolds, formerly a postdoc in CASC and now at UCSD. The algorithm is performing well on an extensive battery of verification tests including radiation-matter equilibration, nonequilibrium Marshak wave, radiating shock wave, and ionization front tests. Fig. 2 shows a preliminary comparison between numerical and analytic solutions for the ionization front test on a low-resolution, uniform grid. ISCR summer student Geoffrey So assisted in the verification testing.

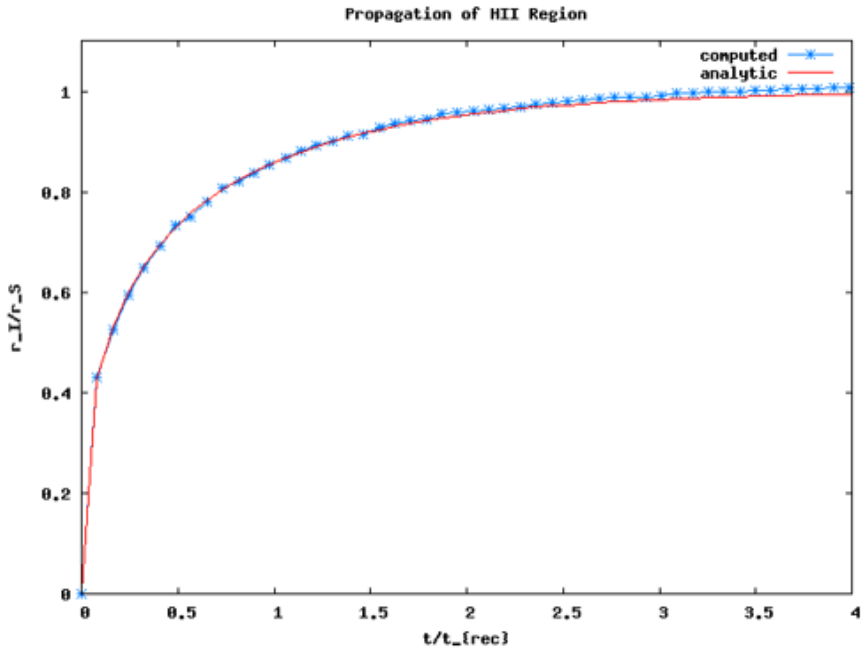

Figure 2. Stromgren sphere verification test of Enzo's coupled radiation transport and gas ionization model. Here we compare analytic and computed expansion of an ionization front in an initially neutral hydrogen region.

Enzo's new radiation transport algorithm rests atop the *hypre* linear solver library developed by CASC. Up to now, verification tests have been done on uniform Cartesian grids. AMR tests are scheduled to begin by year's end. The AMR development is being carried out by James Bordner at UCSD in collaboration with the *hypre* group led by Rob Falgout.

3.3 Cosmology Simulation Data Management

The Storage Resource Broker (SRB) data grid is used to manage the scientific data for the cosmology research. Lightcone simulation data is moved to SDSC and placed into an SRB-managed data collection designed and implemented by UCSD graduate student Rick Wagner. Work this year has focused on developing a database data model for describing the simulations that is general enough to encompass a broad class of simulations while at the same time serving the needs of LSST. A preliminary design (Figure 3) organized around the notion of object catalogs is being implemented at the present time.

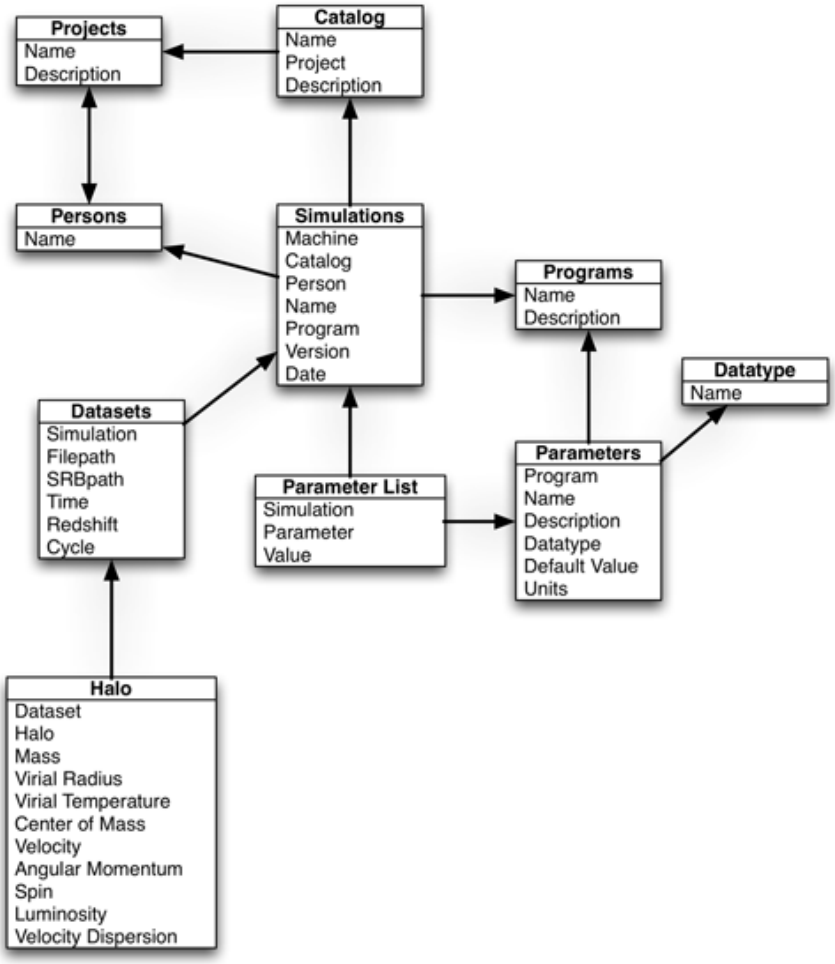

Figure 3. Model for the database schema implemented for the lightcone simulation data. It allows the cataloging of both galaxy clusters and the simulations that generated them. While it lacks detailed information about algorithms, it is very extensible.

4. Scientific Data Management

The LUSCiD project has a unifying goal of demonstration of advanced scientific data management technology in support of large-scale simulations. The scientific data management has five components:

- creation of a digital library of Global Climate Change simulation results in support of the extended data analysis conducted at SIO of simulations run at LLNL on the Thunder computer
- creation of a digital library of Cosmology simulation results for simulations run at LLNL on the Thunder computer
- support for evaluation of the Storage Resource Broker (SRB) data grid technology and the integrated Rule-based Data System (iRODS) for use at LLNL on the Green Oasis disk cache
- support for the Large Scale Synoptic Telescope data management requirements and subsequent data challenge demonstrations
- demonstration of generic infrastructure that supports data sharing, data publication, and data preservation.

These areas can be categorized as support for production simulation runs including the transport of data from LLNL to SDSC and the archiving of the results; support for data

management challenges conducted as part of the LSST project; and demonstration of generic infrastructure that supports data grids, digital libraries, and persistent archives.

4.1 Installation of Data Grid Technology on Green Data Oasis

Under funding support from the National Science Foundation and the National Archives and Records Administration, SDSC has completed a security assessment of the Storage Resource Broker data grid. SDSC integrated the use of bind variables for accessing the metadata catalog, and collaborated with an evaluation process led by Dr. Barton Miller at the University of Wisconsin. The effort required the modification of over 50,000 lines of code in the system. The resulting system will be released as version 3.5 of the SRB during the fall quarter, 2007. This will be the preferred release for production use on the Green Data Oasis.

LUSciD researcher, Arun Jagatheesan is working with Jeff Long from the LLNL's Computation Directorate to use SDSC SRB to manage unclassified collaborative data at LLNL. This collaborative effort is part of the Green Data Oasis (GDO) that is maintained by Computation. The use of data grid technologies at Computation could potentially improve LLNL's data management operations beyond the scope of the LUSciD project. In particular, micro-services are being implemented in the iRODS technology that will enforce the LLNL data exchange policies, provide audit trails on data exchange operations, and support parsing of the audit trails to verify compliance.

The first production release of the iRODS technology was scheduled for fall quarter 2007. This system automates the execution of management policies, minimizing the amount of effort required to maintain a distributed shared collection. Rules can be defined that control all operations performed within the data grid, and that enforce management policies for access, distribution, retention, and disposition.

4.2 SDSC and LLNL Partner in LSST Data Management

Researchers from LLNL and SDSC have been working together to design and develop the cyberinfrastruture that will be required by the Large-scale Synoptic Survey Telescope (LSST). More than 150 PB of data will be moved between South and North America and organized for distribution from a Data Access Center. The LSST software includes multiple sub-systems—scientific pipeline software to process LSST data, real-time data monitoring software, alert-generation and distribution software, database systems to manage the very large number of records, data grid software to manage shared distributed collections, and middleware that provides a simple interface to applications by hiding the highly distributed cyberinfrastructure.

The LSST team evaluates the software system through yearly data challenges. In 2006, SDSC researchers worked with other LSST partners from NCSA to use the SDSC Storage Resource Broker (SRB) to simulate intercontinental data management. This year, Jagatheesan and Don Dossa from LLNL have been working together on LSST technology assessment. While Dossa focused on the hardware infrastructure for the mountain base in Chile to manage LSST data and delivery alerts, Jagatheesan worked on the infrastructure required by the LSST Data Access Center (DAC).

In this year's Data Challenge, intercontinental data transfer experiments are planned using different file sizes/formats, protocols and tools. The group has already identified the metrics and features that need to be studied. Jagatheesan is collaborating with researchers from NCSA and CC-IN2P3 in France to conduct these experiments.

In the NSF Conceptual Design Review (CoDR) for LSST conducted in September 2007, the review panel gave strong approval for LSST to move ahead. Computer science researchers from both LLNL and SDSC, including Dossa, Jagatheesan and LLNL's Celeste Matarazzo, attended the CoDR to share their expertise on LSST Data Management. One of the potential middleware systems that is of great interest to LSST is the open source iRODS data grid software. An iRODS testbed has also been created for testing automation of management policies for LSST.

4.3 Integrated Rule-based Data Systems

The iRODS release planned for Fall quarter 2007, will be a production capable system. The technology has been demonstrated in tutorials for the Teragrid and Grid2007 conferences. The extensions that have been developed to the system include support for:

- loop constructs for server-side workflows
- notification server for high-transaction rate message passing
- audit trails
- generic interface for accessing structured information

The rule-based system promises to be an excellent candidate for joint research activities on data management technology between SDSC and LLNL, based on interest at NSF.

Back to Top of Page