NERSC NEWS

News from the National Energy Research Scientific Computing Center

February 2005

NERSC Staff Help Pave the Way for Running Larger Jobs on Seaborg

As home to one of the largest supercomputers open for unclassified research, the NERSC Center has moved aggressively to devote a greater share of its processing time to jobs running on 512 or more processors.

Since the start of Fiscal Year 2005 on Oct. 1, 2004, more than two-thirds of the processing time available on Seaborg has been utilized by jobs running on 512 or more processors (32 nodes). Seaborg comprises 6,080 computing processors. Through January, 76 percent of Seaborg's processing time had been used for these larger jobs.

Among these jobs was a calculation of an entire year's worth of simulated data from the Planck satellite, which ran on 6,000 processors in just two hours (see accompanying article).

Achieving this rate of utilization has required support from NERSC staff on both the systems side, as well as applications.

“Running larger jobs is a matter of removing bottlenecks,” said David Skinner of the User Services Group. “On the applications side, there is always some barrier to running at a higher scale.”

On Seaborg choosing the right parallel I/O strategy can be important to the scaling of the time spent in I/O for applications. “If you have a code that runs on 16 tasks, there are a lot of ways to do I/O that will perform roughly the same,” Skinner said. “But when you scale up to 4,000 tasks, there is a lot of divergence between different I/O strategies.”

There are two frequently encountered bottlenecks to scaling that come from the computational approach itself. NERSC consultants address removing these bottlenecks by

(continued on page 2)

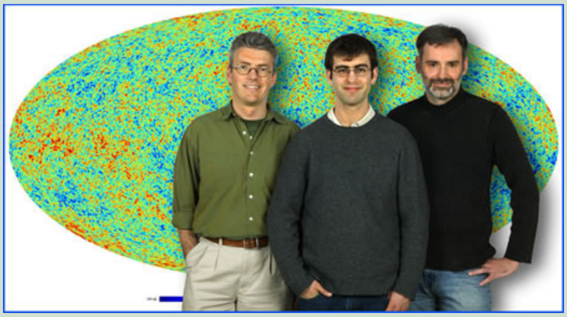
NERSC News

NERSC News, highlighting achievements by staff and users of DOE’s National Energy Research Scientific Computing Center, is published every other month via email and may be freely distributed. NERSC News is edited by Jon Bashor, JBashor@lbl.gov or 510-486-5849.

Simulating a Map of the Cosmic Microwave Background: What the Planck Satellite Will See

In 2007 the European Space Agency, with substantial NASA participation, will launch the Planck satellite on a mission to map the cosmic microwave background (CMB), the remnant radiation believed to be an echo of the Big Bang that started the universe.

Planck is designed to map CMB temperature and polarization fluctuations with unprecedented resolution and sensitivity, but the enormous volume of data this will generate poses a major computational challenge. Can such a mountain of data be efficiently processed?



Julian Borrill, Radek Stompor, and Christopher Cantalupo used the Seaborg supercomputer at NERSC and a year's worth of simulated Planck satellite data to produce the high-resolution map of the CMB behind them. The colors show tiny temperature fluctuations in the CMB, echoing the Big Bang. (Image NERSC, photo Roy Kaltschmidt)

Affirmative.

Berkeley Lab researchers led by astrophysicist Julian Borrill of the Computational Research Division (CRD) have achieved what they say is a "significant milestone" by processing an entire year's worth of simulated Planck data at the single most CMB-sensitive frequency, producing a high-resolution CMB map in just two hours. Using NERSC’s Seaborg supercomputer, the researchers ran their data on 6,000 of Seaborg's 6,080 processors at once, mapping 75 billion observations to 150 million pixels. The run was the first ever to use virtually all of Seaborg’s computing processors on a single code.

“We’ve shown that we can make maps even with the volume of data we will get from Planck,” says Borrill. “There's no computational show-stopper to prevent us from working with the necessary resolution and accuracy. We’ll be able to make CMB maps from the Planck data that astronomers will be able to use for science.”

The map was used to create a poster which was featured at three recent scientific conferences: the Zeldovich-90 International Conference on Cosmology and High Energy Physics in Moscow in December, and the 205th meeting of American Astronomical Society in San Diego and the Planck Consortium meeting in Munich, both in January.

Collaborating with Borrill on the project were Radek Stompor of CRD and Christopher Cantalupo of the Space Sciences Laboratory at UC Berkeley, where Borrill and Stompor also hold appointments. The three colleagues give much credit to the staff at NERSC, who helped shepherd them through all the potential bottlenecks of running 6,000 processors simultaneously.

“The support that the NERSC staff gave us was incredible,” says Borrill. “We’ve shown that we can make high-level, optimal CMB maps at the 217 megahertz frequency” – 217 million cycles per second – “which is the most CMB-sensitive of Planck’s eight detectors and the frequency that will generate the most data. Such maps will be critical for maximizing the mission’s scientific return.”

Out of the cosmic fog

According to the standard model of cosmology, the universe was created 14.7 billion years ago by the Big Bang. In the immediate aftermath of this explosion the universe underwent an enormous expansion at a rate many times faster than the speed of light. Rapid expansion led to a cooling of the universe.

Approximately 400,000 years after the Big Bang, temperatures had cooled enough for protons to capture electrons and form hydrogen atoms. At this point photons, no longer

(continued on page 2)

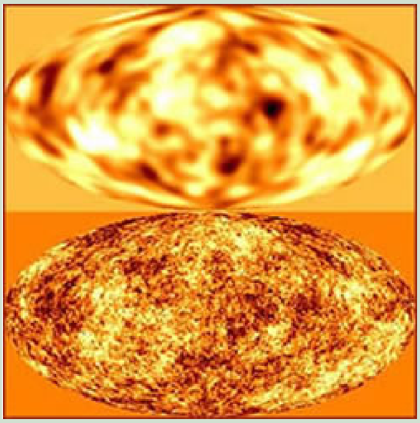
Questions? Comments? Want to subscribe? Contact: JBashor@lbl.gov

Seaborg's 6000 Processors Map Simulated CMB Universe (continued from p.1)

scattered by the dense fog of unattached electrons, were free to move about the universe. Today, photons from the Big Bang have cooled to the very low energy of 2.7 Kelvin, but they permeate the entire universe in the form of microwaves of about one millimeter in wavelength.

When the CMB was first accidentally discovered in 1965 by radio astronomers Arno Penzias and Robert Wilson, it was observed to be isotropic, meaning uniform in all directions — which ran contrary to theory. Subsequent, increasingly sensitive observations revealed that the CMB is actually anisotropic after all: indeed, tiny temperature fluctuations in the earliest radiation were first detected in 1992 in an experiment led by Berkeley Lab physicist George Smoot using NASA's Cosmic Background Explorer satellite, COBE. These temperature fluctuations gave rise to the galaxies and clusters of galaxies that populate our universe today.

The resolution of the COBE images was not sufficient to see any but the largest temperature fluctuations. Subsequent ground and balloon based-experiments, such as MAXIMA and BOOMERANG (whose data were also analyzed at NERSC), have measured smaller-scale temperature fluctuations and some polarization fluctuations as well. But these experiments measure only portions of the sky and aren't sensitive enough to see all the spectral data that the CMB should harbor — data that could help answer a number of important cosmological questions about such things as the Hubble constant and the cosmological constant, important for understanding the expansion of the universe, and the abundance of baryons, the class of massive fundamental particles that includes protons and neutrons.



These CMB maps show COBE's angular resolution of 7 degrees (top) compared to a simulation of the 5 to 10 arc-minute resolution expected from the Planck satellite. (Images NASA, ESA)

MADmap: drinking from the firehose

The Planck satellite, named for Max Planck, the father of quantum theory, is designed to provide enough data to yield full-sky maps in eight frequency bands in the microwave regime between 30 and 857 gigahertz (billions of cycles per second). Such maps would boast an angular resolution of 5 to 10 arc minutes and a sensitivity of a millionth of a degree. (An arc minute is a 60th of a degree; the width of the sun or the full moon viewed from Earth is about half a degree. COBE's best angular resolution was about 7 degrees.)

To produce these maps from the Planck data, Borrill and his colleagues have developed software called MADmap, which stands for Microwave Anisotropy Dataset mapmaker. MADmap capitalizes on massively parallel computation to create millions of pixels from billions of observations.

"We made this map from simulated data that used all 12 of Planck's detectors at the 217-megahertz frequency," Borrill says. "If we can map at this frequency, we should be able to map for all the CMB frequencies that Planck will observe."

By Lynn Yarris, LBNL Public Affairs

Running Large Jobs Requires Turning of Applications, Scheduling (continued from p.1)

rethinking the computational strategy and rewriting portions of the code.

The first area is synchronization, in which all of the calculations in a code are programmed to meet up at the same time. As the code scales to more tasks, this becomes more difficult. Skinner likens it to trying to arrange a lunch with *n* number of people. The larger the desired group, the harder it is to get everyone together at the same time at the same place.

“People think in a synchronous way, about closure,” Skinner said. “But in a code, you often don’t need synchronization. If you remove this constraint, the problems can run unimpeded as long as necessary.”

The other obstacle is in load balancing. By dividing a large scientific problem into smaller segments – and the more uniform the segments, the better – the job can often scale better, Skinner said. “Domain decomposition is important,” he added.

From the perspective of the Computational Systems Group, the issue is one of job scheduling. “Given the nature of a system like Seaborg, this is a difficult task,” said Jim Craw, leader of the Computational Systems Group.

Just as nature abhors a vacuum, Seaborg is programmed to dislike idle nodes. Once processors are freed up, the system “naturally” tries to fill them up with the next appropriate job in the queue. And left unchecked, this would result in lots of small jobs running, filling the nodes and rarely freeing up enough processors to run the larger jobs.

The group created a new system priority formula to run the LoadLeveler queuing system, giving priority to larger jobs, allowing them to work their way to the head of the queue. The system first calculates how many nodes the large job will need, then determines when the required number of nodes will be available. In the meantime, the systems keeps all the nodes utilized by assigns smaller jobs that will be completed before all the needed nodes are available.

While it represents a challenge to the staff, the need for such prioritizing is really a testimony to the success of NERSC as an HPC center of choice. Because there is consistently more demand for computing time than can be allocated, NERSC needs to maintain a very high utilization rate, meaning that as many nodes are in use for as many hours as possible.

“It really boils down to a balancing act between small jobs and large jobs,” Craw said. “We want to fill all the small holes as soon as we can, but we also need to have them available for large jobs. To achieve this, over the years we’ve had to do a lot of tuning of the LoadLeveler scheduling software.”

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California. Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.