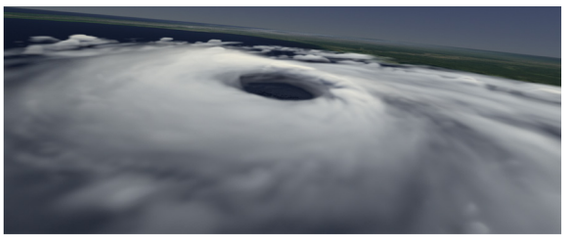
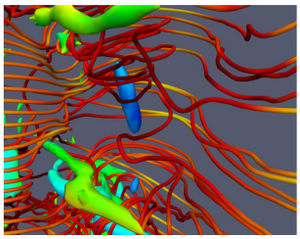
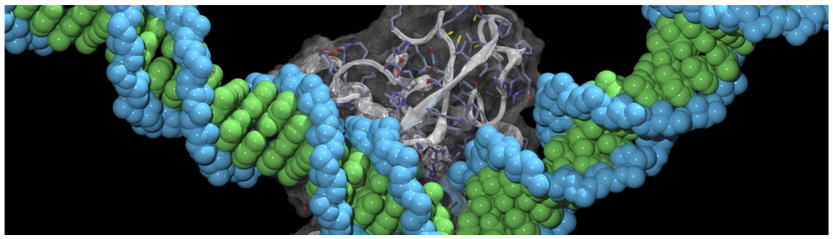
BLUE WATERS PROJECT
NEWSLETTER







Volume 3
Issue 3
December 2012

One Year Later: Delivering Sustained Petascale Science

ONE YEAR AGO, the National Science Foundation approved NCSA's request that the sustained petascale computing system being designed and deployed in the Blue Waters project be based on leading-edge technology from Cray, Inc. We want to bring you up to date on the project and summarize all that has occurred in the last 12 months—a remarkable year that has resulted in the deployment of a computer system that is both the largest ever fielded by Cray and an outstanding resource for the most challenging science and engineering research.

Installation

The most basic accomplishment is that manufacturing and installation were completed in approximately nine months. A small development system arrived in late November 2011 and was in use just one week later for a training class. By February 2012, NCSA and Cray had installed a 48-rack Early Science System with both XE (2-socket CPU-CPU) and XK (2-socket CPU-GPU) nodes. Although it was only one-sixth of the full Blue Waters computational system, the Early Science System was more powerful than any other U.S. academic computing system. Fifteen NSF-approved research teams used the Early Science System during its four months of operation, preparing their applications for use on the full Blue Waters system as well as producing new scientific discoveries.

By the beginning of the summer, all 315 racks were installed at the University of Illinois' National Petascale Computing Facility—a LEED Gold facility that includes many energy efficient innovations—and full-scale testing began. By September's end, all of the final components were in place and operating. To reach this point, NCSA and Cray had to overcome not only all the usual challenges of installing an extremely large and complex computing system but also unanticipated challenges, such as the flooding in Thailand that created a worldwide shortage of disk drives.

Scaling

With all of the equipment in place, everyone's focus turned to scaling the system hardware and software to unprecedented levels. The full Blue Waters system consists of 237 XE6 racks (45,504 AMD Interlagos chips), 32 XK7 racks (3,072 AMD Interlagos chips; 3,072 NVIDIA Kepler chips), and 7 I/O racks. Blue Waters also includes 1.5 PB high-speed memory—4 GB per core for the CPU nodes and 6 GB for the GPUs. This unique resource makes it easier to use the system to simulate the most challenging complex systems. All the memory can be referenced from any core in the system with advanced languages if this is required for a particular application.

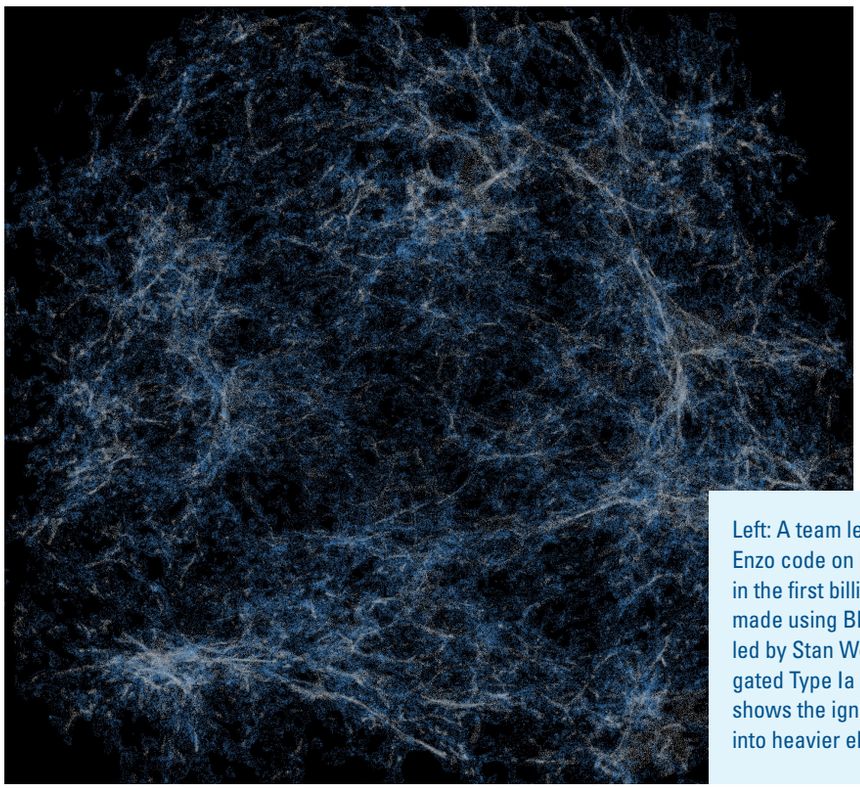
Blue Waters has an interconnect network that is one-third larger than any other deployed Cray system. Blue Waters also has the largest, most robust online storage system in the open research world, with more than 25 PB of usable online storage. The sheer size of the online storage system is impressive, but the Cray Sonexion file system also provides sustained average, aggregate performance over 1.1 TB/s. This unprecedented level of performance is a substantial achievement. The near-line data storage system is also in place and provides more than 300 PB of RAIT-protected data on the floor.

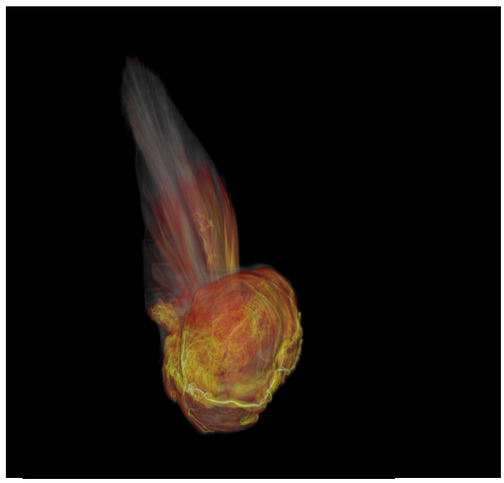
To achieve these results, a multi-organizational team of expert computer technologists made hundreds of improvements to existing technologies and developed new capabilities for storage and I/O. Connectivity to the national networks in Chicago began at 55 Gbits/s and will be scaled to hundreds of Gbits/s as needed by the research community using Blue Waters.

Testing

Throughout the challenge of fielding a system of this size and complexity we have been guided by our long-term interactions with the research teams. Consistent with Blue Waters' goal of sustained petascale performance on real science and engineering applications, we are validating the system by using it as a researcher would—with full, real world applications. We are judging performance based on the elapsed time required to perform real work rather than simplified benchmarks. Testing

Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign, its National Center for Supercomputing Applications, Cray, and the Great Lakes Consortium for Petascale Computation. It is supported by the National Science Foundation, the University of Illinois at Urbana-Champaign, and the state of Illinois.





Left: A team led by Brian O'Shea, Michigan State University, used the Enzo code on Blue Waters to study the development of the universe in the first billion years after the Big Bang. This simulation image was made using Blue Waters' VisIt visualization software. Above: A team led by Stan Woosley, University of California Observatories, investigated Type Ia supernovae with Blue Waters. This simulation image shows the ignition point, which converts large amounts of Carbon 12 into heavier elements by nuclear fusion.

is not yet complete, but we would like to highlight some of the amazing accomplishments achieved so far.

Blue Waters has run all the original NSF application benchmarks. Three of these tests are full-scale petascale benchmarks—complete codes whose runtimes range from 16 hours to over 60 hours, on virtually the entire system—more than 25,000 nodes (or 400,000 floating point cores). These benchmarks solve real science problems, and we measure the entire time from the start to the completion of the problem, including all required I/O (for both scientific results and defensive I/O) and the time required for checkpoints and restarts if failures occur! We are very pleased to report that Blue Waters is meeting or exceeding expectations for these stringent tests.

SPP test

The Blue Waters team set an even higher challenge by creating the Sustained Petascale Performance test. The SPP test is a collection of 12 application benchmarks that are complete applications drawn from the research teams that will use the system! Again these are truly representative tests as they include all start-up processing, I/O, computation, and post-processing, just as these tasks would be performed for a real science run.

We are timing not just parts of the programs or the computationally intensive kernels in order to judge performance—the SPP is a measure of the real sustained research potential of Blue Waters. The SPP codes run on one-fifth to one half of the total system, and several also have run at full scale. Four of the codes are used to show that the XK GPU nodes improve sustained performance (and therefore time to solution) and do so using 600 to 1,500 XK7 nodes. Another test—a "small" test by Blue Waters standards—is the largest Weather Research & Forecasting (WRF) simulation ever documented.

We are pleased to report that four of these codes already run above 1 PF of sustained performance. All 12 of the SPP tests are running at their largest scale, and three are taking good advantage of the XK7 GPU nodes. We expect to report additional achievements and improvements as the SPP benchmarking continues. The most impressive achievement, however, is the work of the science and engineering teams with whom Blue Waters collaborates. As noted previously, 15 of these science and engineering teams used the Blue Waters Early Science System, with many of them making substantial science accomplishments. The teams used the ESS to explore how HIV infects cells, how stars explode, how the most basic constituents of matter behave, and how severe storms occur.

Friendly users

Starting in early November, the full Blue Waters system became available to the 33 NSF-approved science and engineering teams. These "friendly users" have access to the entire system; their work on the system will help test and evaluate the system and will expedite the teams' ability to use Blue Waters productively as soon as it is in full production.

Many people and organizations—including our funders and stakeholders (the National Science Foundation, State of Illinois, and University of Illinois) and our suppliers and sub-awardees (in particular Cray, AMD, Xyratex, NVIDIA, the Great Lakes Consortium for Petascale Computation, INRIA, and the University of Tennessee)—have contributed tremendous efforts to help the Blue Waters project achieve the outstanding progress described above. We have also had enormous help from many of the science teams and look forward to a long and mutually beneficial partnership with them.

Finally, we must make a special acknowledgement of the outstanding contributions of all the Blue Waters project staff at NCSA and the University of Illinois who are working 24 by 7 to create the unique Blue Waters system. Blue Waters project staff have also produced over 50 invited and peer reviewed papers, almost 100 presentations, and have conducted many training and educational activities.

Blue Waters recently began its acceptance testing period to ensure it is ready for all of the challenging use cases of the NSF science and engineering community. We hope to end our deployment stage and enter our full service production phase early next year. A celebration honoring this accomplishment is scheduled for March 28, 2013.

Thom H. Dunning, Jr., project director
William T. Kramer, deputy project director

Assessing sustained performance

The TOP500 list was introduced over 20 years ago to assess supercomputer performance at a time when a major technical challenge was solving dense linear algebra problems on high-end computers. Now, most consider the list a marketing tool or an easy, if simplistic, way to impress politicians, funding agencies and the naïve rather than having much technical value.

The Top500 list is based on the floating point computational performance assessed by a single benchmark, Linpack, which performs a dense matrix calculation. Unfortunately, the TOP500 does not provide comprehensive insight for the achievable sustained performance of real applications on any system—most compute-intense applications today stress many features of the computing system and data-intensive applications require an entirely different set of metrics. Yet, many sites feel compelled to submit results to the Top500 list "because everyone has to."

It is time to ask, "Does everyone have to?" and more specifically, "Why does the HPC community let itself be led by a misleading metric?" Some computer centers, driven by the relentless pressure for a high list ranking, skew the configurations of the computers they acquire to maximize Linpack, to the detriment of the real work to be performed on the system.

The TOP500 list and its associated Linpack benchmark have multiple, serious problems. A few of the issues and possible solutions are briefly listed below.

The TOP500 list disenfranchises many important application areas.
All science disciplines use multiple methods to pursue science goals. Linpack only deals with dense linear systems and gives no insight into how well a system works for most of the algorithmic methods (and hence applications) in use today. The lack of relevance to many current methods will get worse as we move from petascale to exascale computers since the limiting factor in performance in these systems will be bandwidth of memory and interconnects.

Possible improvement: Create a new, meaningful suite of benchmarks that are more capable of representing achievable application performance. Several SERPOP metrics are in use today, such as the NERSC SSP test series, the DOD Technology Insertion benchmark series, and the NSF/Blue Waters SPP test that use a composite of a diverse set of full applications to provide a more accurate estimate of sustained performance for general workloads. These composite measures indicate critical architectural balances a system must have.

There is no relationship between the TOP500 ranking and system usability.
In a number of cases, systems have been listed while being assembled at factories or long before they are ready for full service, leaving a gap of months between when a system is listed and when it is actually usable by scientists and engineers. This perturbs the list's claim of historical value and gives misleading reports.

Possible improvement: List only systems that are fully accepted and fully performing their mission in their final configurations.

The TOP500 encourages organizations to make poor choices.
There have been notable examples of systems being poorly configured in order to increase list ranking, leaving organizations with systems that are imbalanced and less efficient. Storage capacity, bandwidth, and memory capacity were sacrificed in order to increase the number of peak (and therefore Linpack) flops in a system, often limiting the types of applications that it can run and making systems harder for science teams to use. For example, for the same cost, Blue Waters could have been configured to have 3 to 4 times the peak petaflops by using all GPU nodes and having very little memory and extremely small storage. This would have made Blue Waters very hard to program for many science teams and severely limited what applications could use Blue Waters, but almost certainly would have guaranteed being at the top of the Top500 list for quite a while.

Possible improvement: Require sites to fully specify their system capacities and feeds. For example, the amount and speed of memory and the amount and speeds of the I/O subsystems should be reported. This would allow observers to assess how well a system is balanced and would also document how different types of components influence the performance results.

The TOP500 measures the amount of funding for a system—it gives no indication of system value.

The dominant factor for list performance is how much funding a site received for a computer. Who spends the most on a system influences list position as much as (or more than) programming skill, system design, or Moore's Law. Without an expression of cost listed alongside the performance metric it is impossible to understand the relative value of the system, inhibiting meaningful comparisons of systems.

Possible improvement: Require all list submissions to provide a system cost. The cost estimate could be the actual cost paid, a cost estimated from pricing tables, or, at the worst, a component-wise estimate. The cost of a system contract is often publically announced by sites, or IDC (or others) can help calculate a typical "street" selling price for most systems. A cost estimate along with a ranking would provide much more insight and value. Remember, in the past every system listing in the NAS Parallel Benchmark reports was required to have a cost estimate with it.

Adopting these and other improvements would be steps in the right direction if the list continues. However, it is time the community comes to agreement to entirely replace the Top500 with new metrics, or multiple lists, that are much more realistically aligned with real application performance. Or the HPC community could just say "No more" and not participate in the list. Many government and industry sites already do this, we just never hear about them (which further limits the use of the list for historical information).

As our HPC community strives for more and more powerful systems, and as we cope with having to implement more exotic architectural features that will make future systems harder to use for sustained application performance, it is critical we have measures to guide us and inform our decision making rather than divert our focus and adversely influence our decisions.

Because of the issues discussed here, and with the National Science Foundation's blessing, Blue Waters will intentionally not submit a listing to the Top500 list this fall or any other time. NCSA will continue to pursue new ways to assess sustained performance for computing systems.

Upcoming Events of interest

International Conference on Computational Science
Barcelona, Spain, June 5-7, 2013
http://www.iccs-meeting.org/iccs2013/

International Supercomputing Conference
Leipzig, Germany, June 16-20, 2013
http://www.isc-events.com/isc13/

SC13
Denver, Colorado, November 17-22, 2013
http://sc13.supercomputing.org

About the Blue Waters Project

Blue Waters is one of the most powerful supercomputers in the world for open scientific research. Scientists will create breakthroughs in nearly all fields of science using Blue Waters.

For more information visit:
www.ncsa.illinois.edu/BlueWaters

Blue Waters Project Team

Project Director
Thom Dunning | tdunning@ncsa.illinois.edu

Deputy Project Director
Bill Kramer | wkramer@ncsa.illinois.edu

Senior Project Manager
Cristina Beldica | cbeldica@ncsa.illinois.edu

Chief Software Architect
Marc Snir | snir@illinois.edu

Chief Applications Architect
Bill Gropp | wgropp@illinois.edu

Chief Hardware Architect
Wen-mei Hwu | w-hwu@illinois.edu

Applications Technical Program Manager
Greg Bauer | gbauer@ncsa.illinois.edu

Blue Waters Supporting Hardware Systems Technical Program Manager
Mike Showerman | mshow@ncsa.illinois.edu

Education, Outreach & Training Technical Program Manager
Scott Lathrop | scott@shodor.org

Industrial Engagement
Merle Giles | mgiles@ncsa.illinois.edu

Building & Facilities Technical Program Manager
Tom Durbin | tdurbin@ncsa.illinois.edu

Director, Virtual School of Computational Science and Engineering
Sharon Glotzer | sglotzer@umich.edu

Communications
Trish Barker | tlbarker@illinois.edu

President, Great Lakes Consortium for Petascale Computation
Steve Gordon | sgordon@osc.edu