Oak Ridge Leadership Computing Facility Snapshot
The Week of June 29, 2009

Spider Up and Spinning Connections to All Computing Platforms at ORNL

'We couldn't phone and order one,' so ORNL, collaborators trail-blaze a file system giant

Spider, the world's biggest Lustre-based, centerwide file system, has been fully tested to support Oak Ridge National Laboratory's (ORNL's) new petascale Cray XT4/XT5 Jaguar supercomputer and is now offering early access to scientists.

An extremely high-performance file system, Spider has 10.7 petabytes of disk space and can move data at more than 200 gigabytes a second. "It is the largest-scale Lustre file system in existence," said Galen Shipman, Technology Integration Group leader at ORNL's National Center for Computational Sciences (NCCS). "What makes Spider different [from large file systems at other centers] is that it is the only file system for all our major simulation platforms, both capable of providing peak performance and globally accessible."

Ultimately, it will connect to all of ORNL's existing and future supercomputing platforms as well as off-site platforms across the country via GridFTP (a protocol that transports large data files), making data files accessible from any site in the system.

Shipman said Spider has demonstrated stability on the XT5 and XT4 partitions of Jaguar, on Smoky (the center's development cluster), and on Lens (the center's visualization and data analysis cluster). "We've had all these systems running on the file system concurrently, with over 26,000 compute nodes (clients) mounting the file system and performing I/O [input and output]. It's the largest demonstration of Lustre scalability in terms of client count ever achieved."

Shipman said the file system is designed to support the latest incarnation of Jaguar, which is capable of 1.64 quadrillion calculations a second (1.64 petaflops). "When they told us they needed a file system to support it, we could not just pick up the phone and order one," he said. "No vendor could deliver such a system, so we essentially trail-blazed."

It was a phased approach. ORNL computer scientists and technicians worked for 3 years with partners Cray Inc., Data Direct Networks (DDN), Sun Microsystems, and Dell to bring Spider online. Cray provided the expertise to make the file system available on both Jaguar XT4 and Jaguar XT5. DDN provided 48 DDN 9900 storage arrays, Sun provided the Lustre parallel file system software, and Dell provided 192 I/O servers. The vendors' collaboration has produced a system that manages 13,000 disks and provides more than 240 gigabytes per second of throughput, a file system cluster that rivals the computational capability of many high-performance compute clusters.

The Spider parallel file system is similar to the disk in a conventional laptop— multiplied 13,000 times. A file system cluster sits in front of the storage arrays to manage the system and project a parallel file system to the computing platforms. A large-scale InfiniBand-based

system area network connects Spider to each NCCS system, making data on Spider instantly available to them all.

"As new systems are deployed at the NCCS, we just plug them into our system area network; it is really about a backplane of services," Shipman said. "Once they are plugged into the backplane, they have access to Spider and to HPSS [the center's high-performance storage system] for data archival. Users can access this file system from anywhere in the center. It really decouples data access and storage from individual systems."

Before Spider each computing platform had its own file system. Once a project ran an application on Jaguar, it then had to move the data to the Lens visualization platform for analysis. Any problem encountered along the way would necessitate that the cumbersome process be repeated. With Spider connected to both Jaguar and Lens, however, this headache is avoided. "You can think of it as eliminating islands of data. Instead of having to multiply file systems all within the NCCS, one for each of our simulation platforms, we have a single file system that is available anywhere. If you are using extremely large data sets on the order of 200 terabytes, it could save you hours and hours."

"A successfully deployed Spider will be one of the most important steps the NCCS has taken toward increasing the scientific productivity of our users," said Bronson Messer, of the Scientific Computing Group and a participant in the "Three-Dimensional Model of SN1987A Frontier" early science project. "Sophisticated users have been asking for this, while new users I have spoken with immediately see the advantages and become very excited."

Spider will change the game for long-distance access too. Data will be able to be moved quickly between the NCCS and other sites via GridFTP, and once at ORNL it will again be able to reside in one place without any data staging.

Spider will have both scratch space (short-term storage for files involved in simulations, data analysis, etc.) and long-term storage for each user. Shipman said the technology integration team is now working with Sun to prepare for file system requirements for future NCCS platforms with even more daunting requirements.

Four Supercomputers at Oak Ridge Computing Complex Among World's 25 Fastest

Jaguar XT5 Cray machine remains fastest, most versatile for open science

Jaguar XT5, a Cray high-performance computing (HPC) system component at the Department of Energy's (DOE's) ORNL, remains the world's fastest supercomputer for unclassified research, according to a roster released this week in Hamburg. The TOP500 list (www.top500.org) named four machines at the ORNL computing complex among the world's 25 swiftest. All told, five Oak Ridge machines made the list.

"Researchers from national laboratories, universities, and industry harness Jaguar's supercomputing capability to solve some of the toughest science and engineering problems in the world," said Jeff Nichols, interim associate laboratory director for ORNL's Computing and Computational Sciences Directorate. "Today dozens of large-scale, computationally-

intensive research projects run on Jaguar. The XT5 is often the scientific community's fastest, and sometimes only way to find solutions to grand challenges in climate change, nanoscience, and the energy technology needs in renewable energy, bioenergy, nuclear energy, and fusion energy."

Today's scientific grand challenges are too complex for soloists to solve, and simulations help teams explore complex, dynamic scenarios. "The successes of the Jaguar XT5 and the four other Oak Ridge supercomputers that made the TOP500 list highlight the critical mass of talent we have here," said Arthur Bland, project director of the Oak Ridge Leadership Computing Facility (OLCF), which supports science of critical importance to the nation with the most advanced computational capabilities available. The OLCF is housed in the NCCS at ORNL, which UT-Battelle manages for DOE.

Twice a year the TOP500 list ranks HPC systems on their speed in running High-Performance Linpack, a software code that solves a dense matrix of linear algebra equations. Only two machines have reached calculating speeds exceeding the petaflop range of a quadrillion floating point operations per second.

With a peak speed of 1.382 petaflops, Jaguar XT5 ranked No. 2 on the TOP500 list. The XT5 is part of a larger Cray system, also called Jaguar, that includes a 263-teraflops (trillion floating point operations per second) XT4 component that ranked No. 12.

Another Oak Ridge machine making the TOP500 was Kraken, an XT5 component of a Cray system belonging to the National Institute for Computational Sciences (NICS) and the University of Tennessee (UT). The Kraken XT5 ranked No. 6, becoming the world's fastest academic machine. The other Oak Ridge machines to make the list were the NICS-UT XT4 component, called Athena and ranked No. 21, and the NCCS's Eugene, an IBM Blue Gene/P system that ranked No. 247.

In 2008 Jaguar was a 263-teraflops Cray XT4. It was upgraded with the addition of a 1.4 petaflops Cray XT5 component in the fall of that year. An InfiniBand network connects Jaguar's components for faster data production. With approximately 182,000 AMD Opteron processing cores, the combined system can calculate at a peak rate of 1.64 petaflops. If each person on Earth could perform one mathematical calculation per second, it would take more than 650 years of nonstop work to accomplish what Jaguar XT4/XT5 can in a day.

"The Jaguar system helps the scientific community gain insight into topics critical to DOE and the nation, such as mitigating and adapting to climate change, making efficient photovoltaic materials, producing next-generation biofuels, and controlling plasma in a fusion reactor," said NCCS Director James Hack.

Such simulations have run hundreds of millions of processor hours on the Jaguar system, Bland said.

By a thin margin, Jaguar XT5 trailed the top-ranked machine on the TOP500 list, Roadrunner, a DOE supercomputer at Los Alamos National Laboratory that became the world's first

petascale supercomputer in June 2008. Classified simulations on Roadrunner help ensure the safety, security, and reliability of America's nuclear weapons stockpile.

In contrast, Jaguar is serving the diverse demands of unclassified science applications. Such codes and algorithms explore topics including batteries, combustion, carbon capture and storage, medicine, nanotechnology, astrophysics, aeronautical engineering, groundwater, and fundamental physics. Balancing superlative speed with impressive memory, in November the Jaguar XT5 ran the fastest scientific application ever, sustaining more than a petaflop in a simulation of superconductors and earning the prestigious Gordon Bell Prize.

DOE's Office of Science, the largest supporter of basic research in the physical sciences in the United States, provides HPC resources to the scientific community. Its leadership computing facilities at Oak Ridge and Argonne national laboratories, supported by its Office of Advanced Scientific Computing Research, make large awards of supercomputing time to researchers through the Innovative and Novel Computational Impact on Theory and Experiment (or INCITE) program. In 2009 approximately 470 million processor hours were allocated on Jaguar. In 2010, to further accelerate transformational research, the allocation will increase to 700 million processor hours.

Climate Visualization Team Wins SciDAC Award for ORNL

Researchers use Jaguar to run simulations

"GEOS-5 Seasonal CO^2 Flux," a climate visualization by two scientific computing researchers at ORNL, has received an award in the 2009 Outstanding Achievement in Scientific Visualization category from SCiDAC, DOE's Scientific Discovery through Advanced Computing office. The award was announced at the annual SciDAC 09 conference, held June 14-18 in San Diego.

Jamison Daniel of the Scientific Computing Group at the NCCS and David Erickson of the Computational Earth Sciences Group in the Computer Science and Mathematics Division received the award for their visualization, which describes the seasonal flux of carbon dioxide in the atmosphere over North America.

"The visualization illustrates the seasonal atmospheric carbon dioxide boundary fluxes in the NASA [National Aeronautics and Space Administration] GEOS-5 climate simulation," Daniel explained. The simulations, which are currently running on ORNL's Jaguar XT4 computational platform, will make it possible to resolve down to regional scale the predictions and projections of climate change using global models that will contribute to the 2011 United Nations Intergovernmental Panel on Climate Change 5th Assessment Report.

Work on the visualization was completed in December 2008 and submitted to SciDAC in May.

Supercomputing Tests the Waters

Simulations explore mysterious properties of Earth's most abundant molecule

The mysterious properties of water molecules keep our planet warm and full of life. Although water is one of our most common and vital resources, many of its characteristics are still incompletely understood.

David Ceperley and John Gergely, physicists in the Department of Physics and at the National Center for Supercomputing Applications at the University of Illinois, Urbana-Champaign, will use the Cray XT Jaguar supercomputer at ORNL to create one of the most accurate descriptions of water's microscopic properties.

"Our project is to calculate the energies and forces of water molecules. These types of calculations are part of a long-term goal of being able to design new materials," Ceperley said. "If you're looking for materials with certain properties, then you search through lots of different compounds and get just a few likely candidates. Computers should help the search for viable new compounds."

Water's presence in all biological matter and its molecular properties are of interest in many different fields. One such area is trying to understand its existence and behavior in extreme conditions such as those found on other planets. Another is to shed light on how proteins surrounded by water in our bodies are affected by the forces and movement of the water molecules. And understanding the nature of water in highly confined spaces, such as nanotubes, could lead to the development of technology to supply clean water. The possibilities are almost endless.

Because it is a versatile substance, water is complicated to study. Many factors can alter it at the atomic scale, and these tiny alterations are reflected in properties on a much larger scale. For example, in liquid water, molecules cluster in a tetrahedral, or pyramid-shaped, structure, but water molecules in ice group as a hexagonal structure like a snowflake. Thus, the rearrangement of hydrogen bonds between molecules is the difference between solid and liquid water.

Water's complex nature has slowed water research. Many models are accurate under only specific conditions and cannot be transferred to other phases of water or environments. Ceperley will address the melting and freezing points of water, which in some theoretical models are different from those found experimentally. This means that other predictions coming from computer calculations will not necessarily be correct.

Ceperley's group will use quantum Monte Carlo methods to compute how the electrons arrange themselves in a small sample of water.

"The idea of quantum Monte Carlo is that you move the electrons around randomly," Ceperley said. "Monte Carlo is named after the game of chance or roulette, but with rules.

Quantum just means you're using the algorithms to address the quantum mechanical properties of the electrons. This is the first time quantum Monte Carlo methods have been used to do a problem like this, meaning a liquid."

Ceperley and Gergely will be able to see where the discrepancies lie between the theoretical models and experimental results of freezing and melting points. This more accurate model of electron distribution and movement could, unlike current models, be used for ice, water, steam, and other environments. Other researchers will use these models to make their own models and experiments more accurate.

Another little-understood property of water is the quantum effects of protons in it. Funded by DOE's SciDAC program, Ceperley's team will use 32 million processor hours on Jaguar in 2009 to try to figure out how the protons affect water. This will be a challenge because the effects of protons are very similar to those caused by electrons.

This project will provide more accurate information than has previously been available and bring theoretical models closer to what is seen in experiments.

"The goal is to go beyond pure water and do more complex things," Ceperley said. "In the long term we hope to learn how to do the electronic calculations for any type of system made out of ions and electrons—because that's what everything is made out of."

ORNL Hosts Lustre Part II

Workshop looks ahead to 2015

Users, engineers, and developers converged on ORNL May 19–20 for part two of the Lustre Scalability Workshop.

Lustre is an open source cluster file system developed and supported by Sun Microsystems and popular in HPC environments due to its scalability and open-source nature.

Sponsored by ORNL, Sun, and Cray, Inc., the workshop brought together representatives from the world's largest Lustre deployments to identify key scalability issues and develop a roadmap for the future, namely bandwidth in the terabytes-per-second range and the manageability of exabytes of storage by 2015.

"The focus of the workshop was to identify long-term—2015 and beyond— I/O and storage requirements for high-performance computing and to discuss how the Lustre file system can meet these requirements," said Galen Shipman, group leader for technology integration at the NCCS.

Speakers included Instrumental Incorporated's Henry Newman, who discussed the Defense Advanced Research Projects Agency's (DARPA's) 14 I/O scenarios and their implications for parallel file systems. DARPA, the Department of Defense's main research and development office, is creating High Productivity Computing Systems (HPCS) for national security purposes and to ensure U.S. leadership in critical technologies. For example, HPCS, or trans-

petaflop systems, will allow DARPA to more effectively develop models for weather prediction, ocean and wave prediction, ship design, climate modeling, nuclear-stockpile management, and weapons integration. Improved I/O is crucial in the future HPCS that will be used in this research.

Andreas Dilger from Sun showcased the latest Lustre designs to meet DARPA's future file system I/O goals. Dilger outlined these goals for the HPCS project and the architectural improvements and performance enhancements necessary to achieve them. Among the latter are end-to-end data integrity, file-system-integrity checking, and recovery improvement, to name a few. Necessary performance enhancements include improvements in scalability and the combination of multiple network faces. Overall, said Dilger, the Lustre file system is capable of meeting DARPA's HPCS roadmap, and the HPCS program symbiotically provides a motivation to continue growing Lustre.

ORNL's Shipman discussed the NCCS's current Lustre-based Spider file system and presented a roadmap for delivering an exascale system within the decade and the I/O requirements necessary to achieve the 2015 goals of a number of mission partners, including ORNL, Lawrence Livermore National Laboratory, Pacific Northwest National Laboratory, and others.

In all the conference hosted 30 attendees and was held at ORNL's Joint Institute for Computational Sciences.

"The workshop provided a great opportunity for us to involve our most demanding users in setting the direction we will take with Lustre over the next several years. This is an important part of our ongoing commitment to meet the most demanding I/O and storage needs of the HPC community" said Peter Bojanic, director of Sun's Lustre Group.