System Integration Experience Across the Cray Product Line

Arthur Bland, Oak Ridge National Laboratory
Richard Alexander, Oak Ridge National Laboratory
Jeffrey Becklehimer, Cray Incorporated
Steven Carter, Oak Ridge National Laboratory
Nina Hathaway, Oak Ridge National Laboratory
Don Maxwell, Oak Ridge National Laboratory
Robert Silvia, Oak Ridge National Laboratory
Cathy Willis, Cray Incorporated

Abstract: Two years ago, the National Center for Computational Sciences (NCCS) at Oak Ridge National Laboratory installed an early Cray X1. The NCCS has since upgraded this system to 512 MSPs and will upgrade it to an X1E later this year. The NCCS also recently acquired substantial Cray XT3 and XD1 systems, which with the X1 cover the full spectrum of the Cray product line. We describe our experiences with installing, integrating, maintaining, and expanding these systems, along with plans for further expansion and system development.

KEYWORDS: XT3, X1, XD1

1. Overview of the National Center for Computational Sciences

Mission

The mission of the National Center for Computational Sciences is to deliver breakthrough computational science for the nation. We will accomplish this by fielding the most powerful computers for scientific research; building the required infrastructure to facilitate user access to these computers; working with the DOE to select a few time-sensitive problems of national importance that can take advantage of these systems; and working with these teams to achieve their goals.

Description of the National Center for Computational Sciences

The National Center for Computational Sciences (NCCS) is located at the Department of Energy's (DOE) Oak Ridge National Laboratory (ORNL) in Oak Ridge, Tennessee. The NCCS was founded in 1992 as a high-performance computing research center through an open competition. The NCCS was selected by DOE to install and evaluate the new generation of massively parallel supercomputers being developed by several different companies and to bring the systems into production operation for DOE researchers.

The first two systems that were installed were a 64-processor KSR-1 system and a series of Intel Paragon systems ranging from 66 to 3,072 processors. The latter being the third most powerful computer in the world [Dongarra1995.] As the computer systems reached maturity and became stable application platforms, the NCCS began supporting several *Grand Challenge* applications for the DOE including materials science, global climate modeling, and groundwater flow and remediation. In 1999, the NCCS retired the Paragon systems and installed IBM's new Power3 based RS6000 SP system. This system grew to become the first 1-TeraFLOP/s computer installed at a DOE science laboratory. This system was a key part of the human genome project and was used to annotate chromosomes 5, 12, and 19. In 2000, the NCCS installed a 256-processor Compaq AlphaServer SC system for evaluation, the first such system in an unclassified facility. In 2001, the NCCS installed the first IBM Power4-based system. This was upgraded in 2002 to a 4.5 TeraFLOP/s system, making it the eighth fastest system on the Top 500 list [Dongarra2002.] This system has been the primary resource for the DOE Scientific Discovery through Advanced Computing program, providing significant computing resources for many of DOE's most challenging science problems. In 2004, one-half of this system was dedicated for a year to running a series of climate modeling simulations in support of the Intergovernmental Panel on Climate Change's forth assessment. In 2003, the NCCS installed the first Cray X1 in the DOE for evaluation of vector processors on the current generation of DOE science applications. An external review of this evaluation in 2004 recommended expanding this system. [Bailey2004]

On February 23, 2004, the DOE's Office of Science issued a call for proposals for a "*Leadership-Class Computing Capability for Science*". A team led by the NCCS, along with Argonne and Pacific Northwest National Laboratories responded to this competitive request. On May 12, 2004, Secretary of Energy

Spencer Abraham announced the award to Oak Ridge National Laboratory lead team in a speech to the Council on Competitiveness [Abraham2004.] The NCCS proposed to DOE build a national user facility for the nation's open computational sciences communities. The hardware plan proposed was to upgrade the Cray X1 installed at ORNL to 512 processors, and then further upgrade it to an 18.5 TeraFLOP/s system, with 1,024 X1E processors in 2005. Further, the NCCS proposed to install a 25 TeraFLOP/s 5,292 processor Cray XT3 system in 2005. The combination of the two machines provides over 40 TeraFLOP/s of capability computing for both vector and scalar applications. In the longer term, the NCCS proposed to increase the capability to a 100 TeraFLOP/s system in 2006, a 250 TeraFLOP/s system in 2007, and a goal of a 1 PetaFLOP/s system before the end of the decade, as shown in Figure 1.

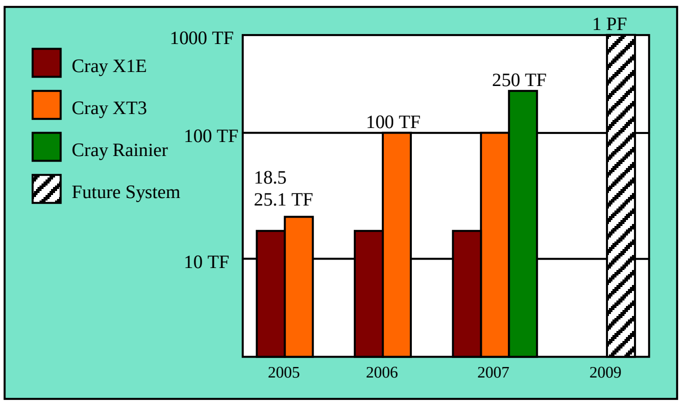

Figure 1 NCCS Computing Capability Plan

Overview of NCCS infrastructure

The NCCS computers cannot exist in a vacuum. The support infrastructure for a national user facility must be world-class. UT-Battelle, the managing partnership for Oak Ridge National Laboratory, recognized the role that high-end computing plays in all types of science. In 2001, the senior management of the laboratory decided to invest in a major new computational sciences facility for the ORNL campus. In June 2003, the NCCS moved from a 1950's vintage computer room into a state-of-the-art 40,000 square-foot computer center designed to house the most powerful computer systems in the world.

The NCCS Building

The NCCS building (see figure 2) is located inside the secure campus of the Oak Ridge National Laboratory. The computer center has a three-foot high raised floor, eight megawatts of power, and 3,600 tons of chiller capacity. Both the power and cooling have redundancy built in to insure minimal disruptions to the computer systems. There is a spare chiller to allow concurrent maintenance and repair, should one of the active chillers fail. We have two power feeds into the building. These are maintained as separate feeds all the way through to the circuit breaker panels in the computer center, thus allowing systems with dual power feeds to continue operation should one side of the power go down for maintenance or repair.

Figure 2 - NCCS Building on the ORNL Campus

The electrical system is backed up by both a 500 KVA uninterruptible power supply, and a 750 KVA diesel generator. While this is insufficient to fully power the entire center, it does keep the networks, servers, and all disk drives in the center powered if there is an electrical outage.

File system strategy

The NCCS has a three-level file system strategy. Each level is designed to provide the required performance and capacity using cost effective disks and tape systems. On each machine, we maintain a high-performance scratch file system using the fastest disk storage available. These disks are configured as a parallel file system to give the maximum I/O bandwidth. The data stored in the scratch file system may be purged to make space available for active jobs. It is the user's responsibility to insure that data are moved to the archival file system if they are to be retained.

The second level of storage is a single home-directory file system that is mounted on each of the computer systems. This system is not designed for high-performance; however, it provides a ubiquitous file system for source code, executable files, batch scripts, and modest amounts of output from the batch jobs. Today, this level of the file system is provided on NFS servers. The NCCS has undertaken a project to convert this level to a higher performing Lustre file system.

Finally, the long-term or very-large file storage layer is provided by the High Performance Storage System (HPSS). HPSS is a hierarchical storage system that migrates data between a disk cache and an archival tape back end. The NCCS has over 550 TB of data stored in HPSS, with this total doubling every 12 months.

Networks

As a DOE national user facility, the NCCS must provide high-bandwidth network connections to the users located around the country at laboratories, universities, and in industry. To accomplish this, the NCCS is connected to the two primary research and development networks in the United States: the DOE ESnet and the Internet2. The ESnet [http://www.es.net/] connection is currently a 2.5 gigabit per second connection while the Internet2 [http://www.internet2.org/] link runs at 10 gigabits per second. To tie the NCCS to other major scientific computing centers, the NCCS is a node on the National Science Foundation's TeraGrid. [http://www.teragrid.org] For connections to experimental ultra-high performance networks, the NCCS is a member of the National Lambda Rail [http://www.nlr.net] and has developed the DOE Ultra Science Network test bed. [http://www.csm.ornl.gov/ultranet/]

The National Center for Computational Sciences is making a substantial commitment to improving application-to-application performance to enable its supercomputers to produce the best science possible. This commitment includes significant local and wide area network improvements as well as a concerted effort to optimize applications to best take advantage of each supercomputer¹s I/O characteristics.

The NCCS has recently started a major upgrade of its local area network, allowing it to support 10 Gb/s interfaces throughout the network and accommodate multiple 10 Gb/s transfers over its backbone. In addition, ORNL is deploying an optical network with a total wide area capacity of almost 1Tb/s. This year, six new 10 Gb/s connections will be added to increase the NCCS's ability to share its computing resources with distant researchers.

In concert with the upgrade in network hardware, the NCCS is making a significant commitment in optimizing the operating systems and software to best utilize the available local and wide area bandwidth. Last year, NCCS integrated Net100 technology into the X1 CNS (Cray Network Subsystem) allowing for faster wide area transfers. As part of this work, the TCP assist functionality was re-written and later adopted by Cray for use in the latest versions of the CNS. Currently, researchers are able to get 700-800Mb/s throughput locally and 400-500 Mb/s throughput over wide area networks for a single file transfer through a single CNS on a regular basis. Work is underway within NCCS to increase single CNS throughput to 4 Gb/s.

Visualization

The goal of scientific simulation on large computer systems is to transform data into knowledge and further glean understanding from that knowledge. With the many terabytes of data generated by the computer systems, it is imperative that the NCCS provide high-end visualization tools and facilities to assist the users with these transformations.

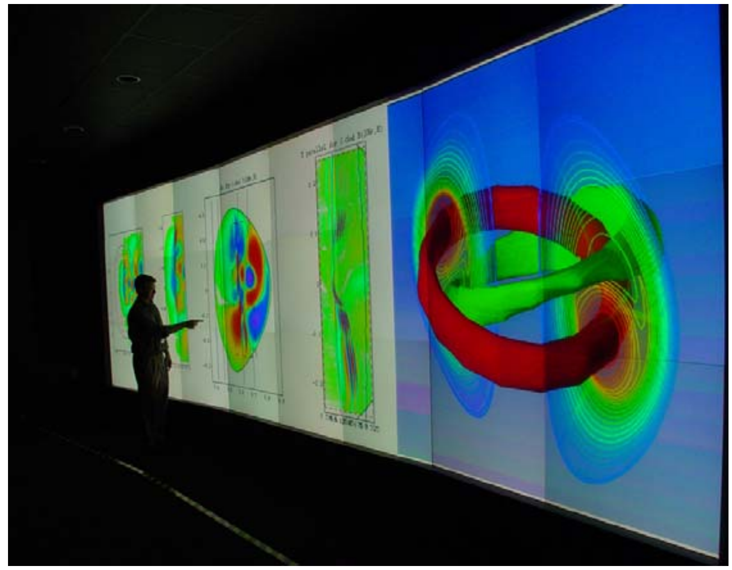

Figure 3 – Jamison Daniel examines a Tokamak simulation in the EVEREST Facility

The Exploratory Visualization Environment for REsearch in Science and Technology (EVEREST) facility, located in the NCCS building, is a large-scale immersive venue for data exploration and analysis. Its screen is 30 feet wide by 8 feet high - comparable in size to 150 standard computer displays - and has a resolution of over 11 thousand by 3 thousand pixels, creating a total pixel space of 35 million pixels (see figure 3). It is integrated with the NCCS computer systems, creating a high-bandwidth data path between large-scale high-performance computing and large-scale data visualization. The immersive qualities of this environment create a powerful discovery tool for research groups and collaborations. [http://www.csm.ornl.gov/viz/]

Users

The users of the NCCS computer systems come from DOE national laboratories, universities, and from industry. Figure 4 shows that the user community is evenly split between DOE national laboratories and universities, with a small number from other agencies and industry.

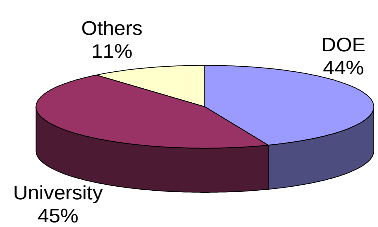

Figure 4 - NCCS Users by affiliation

In fiscal year 2004, the usage of the NCCS systems was divided among several major scientific disciplines. Figure 5 shows the breakout of utilization by the type of science.

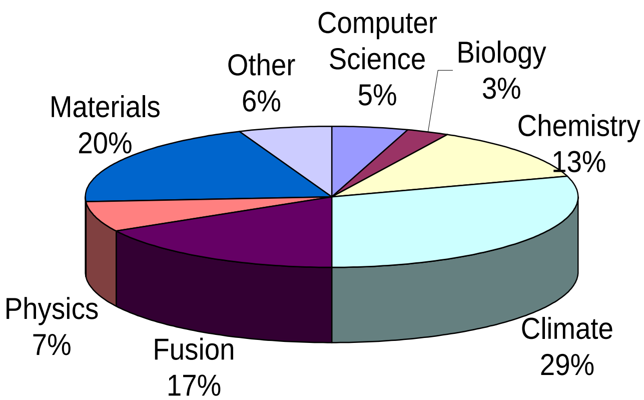

Figure 5 - FY 2004 NCCS Usage by Discipline

The systems in the NCCS are allocated through a peer-reviewed scientific merit process. The model is to make a few, perhaps 10 to 12, very large allocations for substantial fractions of the computers' resources. Perhaps another 30-40 projects will receive small pilot project allocations to prepare for the next round of the proposal process.

Cray systems in the NCCS

Cray X1 - Phoenix

The NCCS has the largest Cray X1 in the world at 512 multi-streaming processors (MSP). The system has grown from a single half-populated cabinet that was installed in March 2003. A second cabinet was added in June 2003. By November 2003, the system had expanded to eight cabinets and 256 MSPs. In September 2004, the eight cabinets were fully populated, taking the system to the current configuration. The system has two terabytes of memory, or four gigabytes for each MSP.

While Phoenix is still an early-life system, our users find it to be an extremely useful system for many applications that have not performed well on typical parallel computers. Figure 6 shows the number of reboots required on Phoenix over the last year categorized by month and cause.

One major operational problem that we see on the system is water quality for the chilled water heat exchangers. We have found that our chilled water supply will occasionally circulate small dirt, rust, or metal particles. We have had several instances when the heat exchanger became plugged with this debris and required back-flushing to continue operation. In Figure 6, the reboots required for environmental reasons are all due to chilled water lines clogging with debris. Our engineers first noticed the problem by seeing the cabinet temperature slowly rise as the heat exchanger clogged. Eventually the problem became severe enough to cause the cabinet to shut down due to overheating. The problem is being addressed by installing filters on the chilled water supply lines to each of the cabinets. When the filter begins to clog, the water is routed through a bypass while the filter is replaced. This operation can be completed without requiring the computer to be shut down. To clean up the source of the problem, large filters are being installed in the chilled water plant to trap the particles. We are working diligently to resolve this issue prior to the X1E modules being installed, as the tolerances get even tighter with these modules.

Another source of recent hardware problems have been the result of memory module failure. Virtually all of the hardware caused reboots over the last three months have been due to replacing failed memory cards.

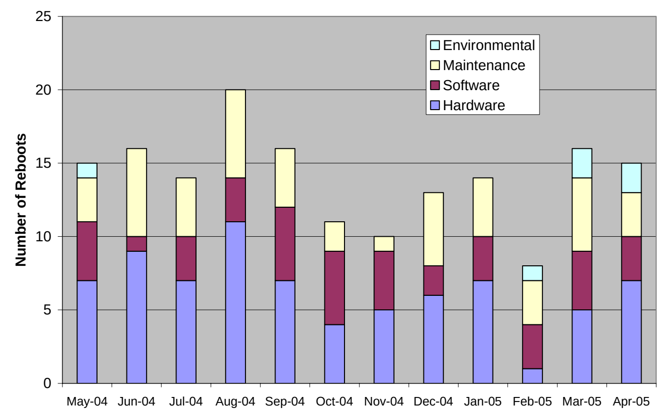

Figure 6 - Cray X1 Reboots by Month and Cause

The NCCS will continue to maintain leadership by upgrading the Cray X1 to an X1E system during the summer of 2005. The upgrade will take place in four stages. One-fourth of the X1 boards will be replaced by X1E boards. Those X1 boards will be returned to the factory to be upgraded to X1E boards. The upgraded X1E boards will then replace another one-fourth of the X1. This will continue until all of the boards have been upgraded. During the upgrade, the X1 and X1E systems will be partitioned into separate systems. The resulting system will be a 1,024 MSP system with a peak speed of 18.5 TeraFLOP/s. This will continue to be the largest Cray vector system and the first time that Unicos/mp has been scaled to over 512 MSPs.

Cray XT3 – Jaguar / Rizzo

The NCCS received its first XT3 cabinet on December 27, 2004. This cabinet has 17 compute cards (68 nodes) and seven service cards installed. Initially, it ran the pre-release development version of the operating system, known as dev-harness, which proved to be very painful for users. With no software to manage processor allocation, users were forced to select processors by node identifier and attempt to avoid stepping on each other. Furthermore, bugs in portals and other software typically meant users could only get one or two runs out of the machine before a reboot was required. However, even with these warts, the system early access did allow the NCCS users to begin testing applications on the system, and to use the PGI compiler.

Cray delivered the first set of 10 cabinets for the 56-cabinet configuration on February 25, 2005. A decision was made to test the follow-on to the dev-harness code called CRMS, which is the released version of the operating system. However, we quickly learned that the code that loads the routing tables at boot time based on available hardware could not properly build the routing table for Jaguar with only 10 of the 56 cabinets being available. Subsequently, another 10 cabinets arrived March 28, 2005.

The April 4th release of the CRMS code was installed and is the version still running today. There were some initial successes with the 20-cabinet configuration, but things soon became unstable as hardware problems were encountered. These problems were very difficult to trace since all Seastar router chips initially had to be available in order to route the machine. A new rtr command was released from the developers that made it possible to disable slots and still route the machine. Troubleshooting the hardware became much easier and got the machine back to the initial stability just in time for 20 more cabinets that arrived on April 28th.

The machine is now configured with 40 cabinets. The network topology is a torus in the X and Z dimension and a mesh in the Y dimension. The Y dimension of the interconnect spans rows of cabinets, but the longer cables needed to complete the torus have not yet been qualified.

Today, Jaguar has 3,748 compute processors and the system boots in less than 15 minutes, down from seven hours during the first week of operation. We can run applications that use the full resources of the system for many hours. Application scaling is exceptional on bandwidth sensitive code, but still lags on codes that are more latency dependent. Cray expects to dramatically improve the latency with a new version of the portals code to be incorporated into the system later this year.

Hardware stability of the machine is currently excellent. Over the past eight days, Jaguar has not had any hardware problems. The software is still very early and has a number of problems. Over the past two weeks, Jaguar has averaged two reboots per day. The most important problem we see today is that jobs will often finish, but for some reason, the nodes will not be released to be reallocated back to another user. This problem typically requires a reboot of the system to recover those nodes for further use.

Another major problem we see is known as the CAM overflow. Each Seastar chip has 256 Content Addressable Memory (CAM) registers. If the number of outstanding messages exceeds this number, then a CAM overflow occurs. When this happens, the application will hang. Cray is working on the problem with a functionality called Basic End-to-End Reliability (BEER) that is scheduled for version 1.1.

The CPU performance and bandwidth of the Seastar network are very good. The message passing latency is currently very poor but expected to be dramatically better with the version 1.1 software. Further progress on latency is also expected with additional releases later in 2005.

The file system for the XT3 is Lustre. With the 1.0 software, we have been able to get Lustre working and demonstrated scalability of the bandwidth with the number of object storage targets up to four OSTs. While there is still much work to be done, the basic functionality is working. The NCCS hosted a workshop this month to write a Lustre primer for those who are new to this file system. This will include an appendix on how to get Lustre running on the XT3 system.

In June 2006, the NCCS will add 16 cabinets to Jaguar, taking it to 56 cabinets in a 4x14 configuration. The system will have a total of 5,212 compute processors, 82 service and I/O processors, 10.7 terabytes of memory, and 120 terabytes of disk space. This will be a 25.1 TeraFLOP/s system. Our plan is to expand the system to 120 cabinets with dual-core processors in 2006. The machine will then be over 100 TeraFLOP/s with over 40 terabytes of memory and one-half petabyte of disk space.

Cray XD1 – Tiger

Tiger is a 12 chassis, 144 processor Cray XD1 system that the NCCS has had since September 27, 2004, when the first six chassis arrived. The next six chassis arrived November 3, 2004. The processors are 2.2 GHz AMD Opterons and each processor has 4 GB of memory. One of the chassis has six of the optional FPGA based application accelerators. The system has four RapidArray links per node, giving a bandwidth of 8 GB/s per two-processor node.

Physical Installation and Startup

The physical installation went fairly well with only one problem. Diagnostics did not work as planned. The XD1 is designed to be installed at sites without on-site engineers; therefore diagnostics do not normally ship with the system. There was a conflict between the diagnostic software and the Active Manager software. Both the diagnostics and Active Manager would modify configuration files on the system without accounting for the presence of the other. This has since been corrected.

When our first six chassis arrived, the largest supported single configuration was three chassis, so initially Tiger was configured as two separate systems. When the second set of six chassis arrived six weeks later, it came with a beta release of version 1.1.7 of the system software that allowed systems of up to six chassis. The new six chassis system was installed as a stand-alone system. Once it worked, the previous six chassis were integrated into a single system. On May 4, 2005, we installed a limited availability release of version 1.2.37 that supports systems of up to 12 chassis. With this working, we were able to integrate all of the nodes into a single system.

Commissioning

On the first system to be commissioned, we ran into the issue of having to obtain 19 contiguous IP addresses for the system. We made the decision to commission it with three addresses to see how that would work. We quickly found that neither LDAP nor external NFS access were possible with that configuration. We reconfigured the system to use NIS and internal NFS home directories while investigating a solution for the need for 38 IP addresses. We were given our own subnet, so we were able to configure the second system with a range of 19 contiguous IP addresses and LDAP authentication.

The second commissioning only had minor problems:

1. When asked for the range of IP addresses the example shows "upperlimit - lowerlimit" but the manual indicates it should be "lowerlimit - upperlimit". It was entered as lowerlimit - upperlimit.
2. In the DNS section, an IP address was entered where a name was expected. This was caught before finishing the section, so we CTRL-C'd out of the commission. When commission resumed it noted that DNS information had already been entered and could not be changed. We had to decommission and start over.
3. A typo was made in the subnet address range and ActiveManager told us the entry was invalid, but gave no indication that a new entry could be made. We made the new entry anyway and it was accepted.

Post-Commissioning

The main problems that needed to be solved for us to allow general users on the system were related to Active Manager and authentication.

1. Many of the activities as described in the System Administrators guide, such as 'amshutdown', 'mvsmp', 'amreboot', etc. caused corruption of the Active Manager database. This manifested itself as GigE configuration problems and with IP addresses being assigned but then not released along with various SMPs hanging.

2. The GigE cards that were installed, but not connected to a network, caused L2 forwarding to be configured improperly. The unused GigE cards were removed to resolve this issue. When Active Manger saw a GigE card in a chassis, it would set the default IP route through that card, even if it was inactive.

3. A major issue was the integration of the OpenLDAP used on the XD1 with the LDAP server running on AIX. We had quite a bit of difficulty getting groups to authenticate. Since active manager authenticates access to partitions based on groups it has to know each group that a user is a member of and the mechanism between XD1 and AIX did not work. We worked around this by maintaining a flat /etc/groups file and turning off LDAP searches for groups in /etc/nsswitch.conf.

The Portland Group compilers have been installed and are working with interface files from Cray benchmarking. It has been recommended that we rebuild the MPICH libraries with the Portland Group compilers to allow seamless compatibility and are awaiting MPICH source from Cray Canada.

Overall Impressions

The Cray XD1 is an excellent performer. The RapidArray fabric has both very high bandwidth and very low latency. We have just switched from using the Sun Grid Engine for a workload manager to using PBSpro. This will make the system interface similar to the other systems that we run in the NCCS. We have experimented with both the Pathscale and PGI compilers for the Opteron processor. We have found that each is better for certain code, sometimes dramatically better. It can pay to try both if application performance is not why is expected.

NCCS Future Plans

As shown in Figure 1, the NCCS does not intend to stand still with its two large Cray systems. In 2006, we plan to expand the Cray XT3 system from 25 TeraFLOP/s to over 100 TeraFLOP/s by doubling the number of cabinets, installing dual-core processors, doubling the memory, and installing four times as much disk space. The resulting system will have over 20,000 processors, over 40 TB of memory, and one-half *petabyt*e of high-speed parallel disk space attached to the system. In the usage model of the NCCS, this expanded system will be used to provide substantial additional capability to the same small number of projects rather than trying to support additional projects.

In 2007 or 2008, the NCCS expects to provide a new system to replace the Cray X1E vector platform based on Cray's Rainier system. The goal for this system is a 250 TeraFLOP/s machine. The NCCS is planning for a PetaFLOP/s system by the end of the decade.

About the Authors

Arthur Bland is the Director of Operations for the NCCS and has been at ORNL for 20 years. He can be reached at BlandAS@ornl.gov.

Richard Alexander has been a systems administrator and programmer with the NCCS for nine years. Currently his is working on the Lustre file system on the Cray XT3. He can be reached at alexar@ccs.ornl.gov.

Jeffrey Becklehimer is a on-site systems analyst with Cray Inc. He is the primary Cray systems administrator for the XT3. He can be reached at jlbeck@cray.com.

Steven Carter is the networking lead for the NCCS. He can be reached at scarter@ornl.gov.

Nina Hathaway is a system administrator with the Joint Institute for Computational Sciences, a joint venture of ORNL and the University of Tennessee. She has been working with ORNL for four years and can be reached at nina@ccs.ornl.gov.

Don Maxwell has been a systems administrator with the NCCS for five years. He is currently the primary ORNL systems administrator for the Cray XT3. He can be reached at maxwellde@ornl.gov.

Robert Silvia is the group leader for HPC operations in the NCCS. After spending over a dozen years working with Cray machines at the North Carolina Supercomputing Center, he joined ORNL about a year and a half ago He can be reached at silviarj@ornl.gov.

Cathy Willis is an on-site systems analyst with Cray, Inc. at ORNL. Her primary task is administration of the Cray X1. She can be reached at willis@cray.com.

References

[Abraham2004] Spencer Abraham, *Leadership Class Computer Announcement*, May 12, 2004. http://www.energy.gov/engine/content.do?PUBLIC_ID=15869&BT_CODE=PR_SPEECHES&TT_CODE=PRESSSPEECH.

[Bailey2004] F. Ronald Bailey, et. al., *Cray X1 Review Report*, February 26, 2004, Report to DOE Office of Advanced Scientific Computing Research

[Dongarra1995] Jack J. Dongarra, et. al., *Top 500 Supercomputer Sites*, June 22, 1995. http://www.top500.org/list/1995/06/

[Dongarra2002] Jack J. Dongarra, et. al., *Top 500 Supercomputer Sites*, June 20, 2002. http://www.top500.org/lists/2002/06/

Acknowledgment

This work was sponsored by the U.S. Department of Energy's Office of Advanced Scientific Computing Research and performed at the Oak Ridge National Laboratory, managed by UT-Battelle, LLC under contract number DE-AC05-00OR22725.