Jaguar: The World's Most Powerful Computer

Arthur S. Bland, Ricky A. Kendall, Douglas B. Kothe,
James H. Rogers, Galen M. Shipman,
Oak Ridge National Laboratory

ABSTRACT: *The Cray XT system at ORNL is the world's most powerful computer with several applications exceeding one-petaflops performance. This paper describes the architecture of Jaguar with combined XT4 and XT5 nodes along with an external Lustre file system and external login nodes. We also present some early results from Jaguar.*

KEYWORDS: Jaguar, XT4, XT5, Lustre

1. Introduction

In 2004, the National Center for Computational Sciences (NCCS) at the U.S. Department of Energy's Oak Ridge National Laboratory (ORNL) partnered with Cray Inc., Argonne National Laboratory, Pacific Northwest National Laboratory, and others to propose a new national user facility for Leadership Computing.[1] Having won the competition, the new Oak Ridge Leadership Computing Facility (OLCF) has delivered a series of increasingly powerful computer systems to the science community based on Cray's X1 and XT product lines. The series of machines includes a 6.4 trillion floating-point operations per second (TF) Cray X1 in 2004, an upgrade of that system to an 18.5 TF X1e in 2005, a 26 TF single-core Cray XT3 in 2005, and an upgrade of that system to a 54 TF dual-core system in 2006, an addition of a 65 TF XT4 in 2006 that was combined with the XT3 in 2007 to make Jaguar a 119 TF system. In 2008 the system was further upgraded to quad-core processors for a performance of 263 TF with 62 terabytes (TB) of memory. This ten-fold increase in Jaguar's computing power and memory from February 2005 through April 2008 provided a productive, scalable computing system for the development and execution of the most demanding science applications.

On July 30, 2008 the NCCS took delivery of the first 16 of 200 cabinets of an XT5 upgrade to Jaguar that ultimately has taken the system to 1,639 TF with 362 TB of high-speed memory and over 10,000 TB of disk space. The final cabinets were delivered on September 17, 2008. Twelve days later on September 29th, this incredibly large and complex system ran a full-system benchmark application that took two and one-half hours to complete.

Today, Jaguar is running a broad range of time-critical applications of national importance in such fields as energy assurance, climate modelling, superconducting materials, bio-energy, chemistry, combustion, and astrophysics. To date Jaguar has delivered 742 million processor-hours to applications since January 1, 2007.

2. Jaguar's Architecture

Jaguar is the most powerful descendent of the Red Storm[2] computer system developed by Sandia National Laboratories with Cray and first installed at Sandia in 2004. Jaguar is a massively parallel, distributed memory system composed of a 1,375 TF Cray XT5, a 263 TF Cray XT4, a 10,000 TB external file system known as Spider, and a cluster of external login, compile, and job submission nodes. The components are interconnected with an Infiniband network called the Scalable Input-Output Network (SION.) Jaguar runs the Cray Linux Environment system software. The batch queuing and scheduling software is the Moab/Torque system from Cluster Resources, Inc[3]. Spider runs the Lustre file system object storage and metadata servers on 192 Dell Poweredge 1950 servers with dual socket, quad-core Xeon processors, connected to 48 Data Direct Networks S2A9900[4] disk subsystems. The login cluster is composed of eight quad-socket servers with quad-core AMD Opteron processors and 64 GB of memory per node running SUSE Linux.

2.1 Node Architecture

The XT5 nodes are powerful, general purpose symmetric multi-processor, shared memory building blocks designed specifically to support the needs of science applications. Each node has two AMD Opteron model 2356 2.3 GHz quad-core "Barcelona" processors connected to each other through a pair of HyperTransport[5] connections. Each of the Opteron processors has 2 MB of level 3 cache shared among the four cores and a DDR2 memory controller connected to a pair of 4 GB DDR2-800 memory modules. The HyperTransport connections between the processors provide a cache coherent shared memory node with eight cores, 16 GB of memory, and 25.6 GB per second of memory bandwidth. The node has a theoretical peak processing performance of 73.6 billion floating point operations per second (GF). A schematic representation of the node is shown in Figure 1.

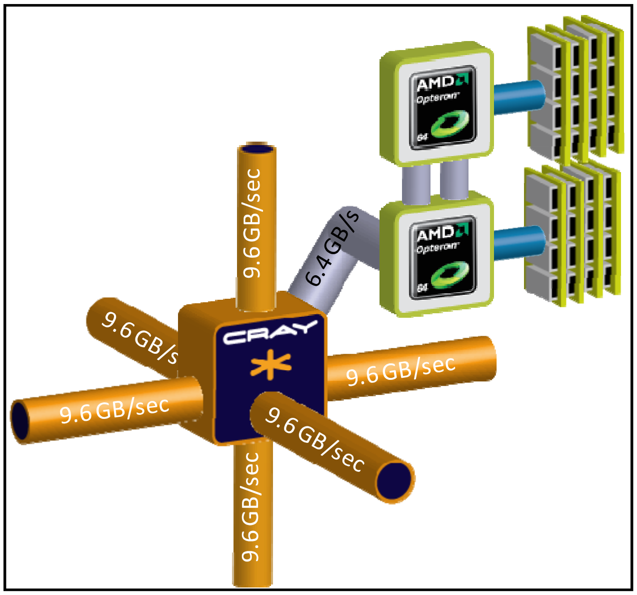

Figure 1 - XT5 Node Configuration

The interconnection fabric that links the nodes is implemented using Cray's SeaStar[6] network interface chip and router. The topology of the network is a three-dimensional torus. Each node is connected to the SeaStar chip through a HyperTransport connection with an effective transfer rate of 2 GB/s in each direction.

The XT4 nodes are similar in design to the XT5 nodes. Each has a single AMD Opteron model 1354 quad-core "Budapest" processor running at 2.1 GHz. The processor has an integrated memory controller linked to either four DDR2-800 or DDR2-667 2 GB memory modules. The XT4 has the same SeaStar interconnect as the XT5, although the two pieces of the system are not connected through the SeaStar links. The XT4 nodes have a peak performance of 33.6 GF and 8 GB of memory.

2.2 Building up the XT systems

Even with very powerful nodes, the XT4 and XT5 require a large number of nodes to get the remarkable performance of Jaguar. With such large numbers of nodes one might expect the high numbers of boards and cables to decrease the overall reliability of the system. To address these concerns, the XT5 and XT4 systems Use very dense packaging to minimize the numbers of cables and connectors. The compute nodes are placed four nodes per "blade" in the system. Eight of these blades are mounted vertically in a chassis. Three chassis are mounted in a single rack with outside dimensions of 22.5 inches wide, by 56.75 inches deep, by 80.5 inches tall. With 192 processors and over 7 TF per rack, the XT5 is among the densest server configurations available.

The XT5 part of Jaguar combines 200 of these cabinets in a configuration of eight rows of 25 cabinets each. The configurations of the XT5 and XT4 parts of Jaguar are shown in Table 1.

	XT5	XT4	Total
Cabinets	200	84	284
Compute Blades	4,672	1,958	6,630
Quad-core Opteron Processors	37,376	7,832	45,208
Cores	149,504	31,328	180,832
Peak TeraFLOPS	1,375	263	1,639
Nodes	18,688	7,832	26,520
Memory (TB)	300	62	362
Number of disks	13,440	2,774	16,214
Disk Capacity (TB)	10,000	750	10,750
I/O Bandwidth (GB/s)	240	44	284

Table 1 - Jaguar System Configuration

2.3 Spider

Spider[7], a Lustre-based file system, replaces multiple file systems at the OLCF with a single scalable system. Spider provides centralized access to petascale data sets from all NCCS platforms, eliminating islands of data. Unlike previous storage systems, which are simply high-performance raids, connected directly to the computation platform, Spider is a large-scale storage cluster. 48 DDN S2A9900s provide the object storage which in aggregate provides over 240 gigabytes per second of bandwidth, over 10 petabytes of RAID6 capacity from 13,440 1-terabyte SATA drives. This object storage is accessed through 192 Dell dual socket quad core Lustre OSS (object storage servers) providing over 14 teraflops in performance and 3 terabytes of RAM. Each object storage server can provide in excess of 1.25 GB per second of file system level performance. Metadata is stored on two LSI XBB2 disk systems and is served by three Dell quad socket quad core systems. These systems are interconnected via the OLCF's scalable I/O network (SION) providing a high performance backplane for Spider to communicate with the compute and data platforms in the NCCS. Figure 2 illustrates the Spider architecture

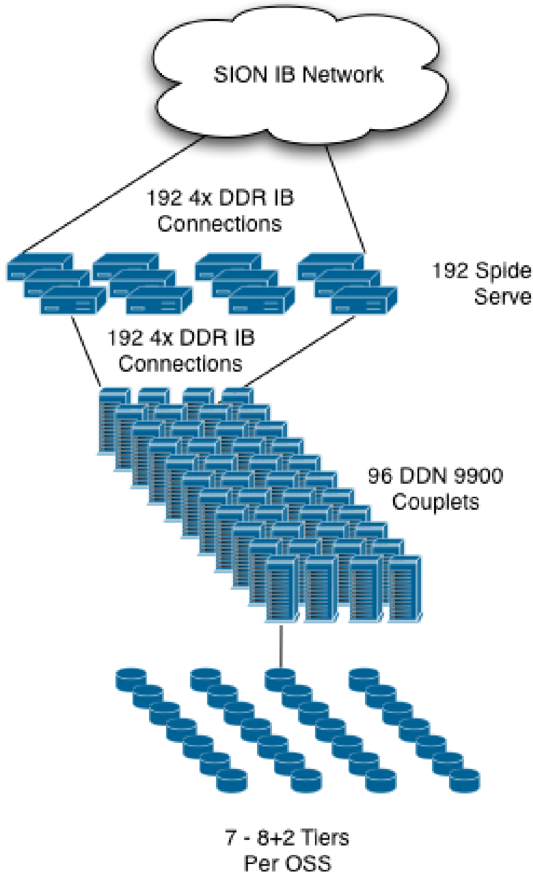

Figure 2 - The Spider File System Architecture

On the Jaguar XT5 partition 192 service I/O (SIO) nodes, each with a dual socket AMD Opteron and 8 GB of RAM are connected to Cray's SeaStar2+ network via HyperTranport. Each SIO is connected to our scalable I/O network using Mellanox ConnectX HCAs and Zarlink CX4 optical cables. These SIO nodes are configured as Lustre routers to allow compute nodes within the SeaStar2+ torus to access the Spider filesystem at speeds in excess of 1.25 GB/s per SIO node. The Jaguar XT4 partition is similarly configured with 48 SIO nodes. In aggregate the XT5 partition has over 240 GB/s of storage throughput while XT4 has over 60 GB/s. Other OLCF platforms are similarly configured with Lustre routers in order to achieve the requisite performance of a balanced platform.

2.4 Login Nodes

Traditional XT system configurations contain heterogeneous combinations of compute nodes, I/O nodes, and login nodes. ORNL has developed an external login node that eliminates the need to dedicate further nodes to login responsibilities. These login nodes are based on the same AMD Quad-Core Opteron processor used in the XT system, but execute a Linux kernel with a more robust set of services than the lightweight Compute Node Linux (CNL) kernel. This ability to extract the login node from a single XT system is coupled with a significant change to the scheduler. From these new external login nodes, users can now submit jobs via Moab to run on more than one XT system. The external login nodes are more fully described in the CUG 2009 paper *Integrating and Operating a Conjoined XT4+XT5 System*[8].

The Spider parallel file system will be mounted by a large number of systems, including the XT4 and XT5 partitions, as well as visualization and analysis clusters. This configuration provides a number of significant advantages, including the ability to access information from multiple systems without requiring an intermediate copy, and eliminating the need for a locally mounted file system on, for example, the XT4 portion to be available if work is being executed on another system or partition. This configuration allows maintenance activities to proceed on a compute partition without impacting the availability of data on the Spider file system.

2.5 Scalable Input-Output Network (SION)

In order to provide integration between all systems hosted by the OLCF, a high-performance large-scale Scalable I/O Network (SION) has been deployed. SION is a multi-stage fat-tree InfiniBand network and enhances the current capabilities of the OLCF. Such capabilities include resource sharing and communication between the two segments of Jaguar and real time visualization as data from the simulation platform can stream to the visualization platform at extremely high data rates.

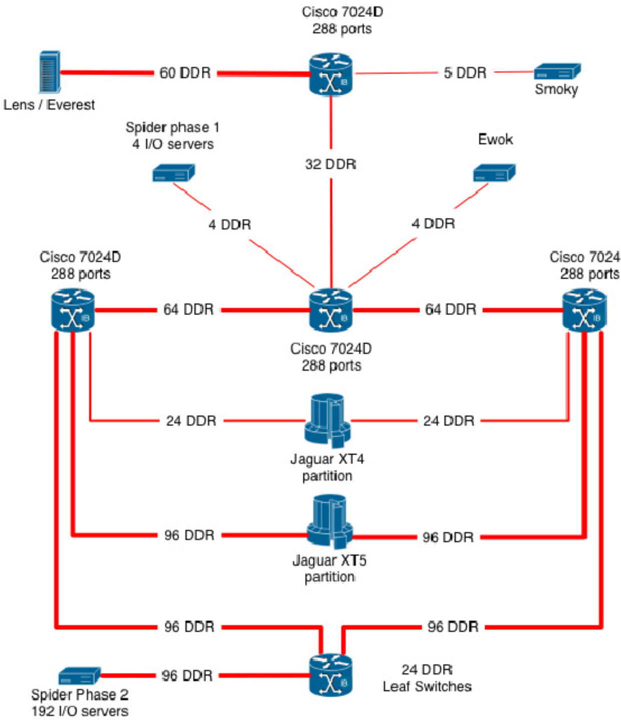

Figure 3- The SION Infrastructure

SION currently connects both segments (XT4 and XT5) of the Jaguar with the Spider file system, Lens (Visualization cluster), Ewok (end-to-end cluster), Smoky (application readiness cluster), and to HPSS and GridFTP servers. SION is a high performance InfiniBand DDR network providing over 889 GB/s of bisection bandwidth. Figure 3 illustrates the core SION infrastructure. The core network infrastructure is based on four 288-port Cisco 7024D IB switches. One switch provides an aggregation link while the other two switches provide connectivity between the two Jaguar segments and the Spider file system. The forth 7024D switch provides connectivity to all other LCF platforms and is connected to the single aggregation switch. Spider is connected to the core switches via 48 24-port Flextronics IB switches allowing storage to be accessed directly from SION. Additional switches provide connectivity for the remaining OLCF platforms.

3. XT5 Mechanical and Electrical Systems

As computer systems continue to get larger, the requirement to engineer the electrical and cooling infrastructure of the machines becomes more important. With a machine the size of Jaguar, the proper power distribution saved over one-million U.S. dollars on the site preparation and installation, and was essential to be able to cool the system.

3.1 ORNL's Computing Facility

Each Cray XT5 cabinet uses a single direct-attached, 100 amp, 480 volt electrical connection. With a total system configuration of 200 cabinets, the NCCS identified some innovative opportunities during site preparation that reduced up-front material costs by over $1,000,000, and will reduce operational losses due to voltage losses for the life of the system. Electrical distribution to the NCCS is based on 13,800V service. These distribution lines are then stepped down using 2,500kVA transformers located within the facility. These transformers provide 480V power to three separate switchboards, located inside the computer room, immediately adjacent to the Cray XT5. By locating the transformers in the building, the distance from the transformers to the main switchboards is significantly reduced. Reducing this distance reduces material cost, and reduces voltage losses across the copper cable. Positioning the switchboards inside the computer room additionally reduces the distance from the switchboards to the individual cabinets.

The NCCS chilled water plant includes five separate chillers configured to automatically adjust to load conditions up to 6,600 tons, the equivalent of more than 23MW of heat. Chiller plant efficiency is maintained by running specific chillers up to their individual optimum efficiency prior to starting another chiller and subsequently rebalancing the load. The control systems are automated so that chillers can enter and exit service without intervention. The chiller plant delivers the chilled water at a steady inlet temperature of 42 degrees Fahrenheit. ORNL will separate the computer chillers from the building cooling later this year so that the chilled water routed to the computer room can have a higher temperature, thus reducing the amount of electricity needed to cool the systems and improving overall efficiency.

3.2 480 Volt Electrical Infrastructure

When the LCF project began in 2003, one of the first items that we discussed with Cray was our growing concern over the increasing power that all systems were using. The goal was to reduce the total power used and the cost of distributing that power on a petascale system. At the time, the NCCS was operating an IBM Power4 P690 system called "Cheetah" that used 480 volt power. The relatively high voltage to the cabinet had two important impacts on that system. The first was that with higher voltage, the same amount of power could be delivered to the cabinet at a much lower current therefore using smaller wires and circuit breakers. This saved several thousand dollars per cabinet in site preparation costs. The second impact was that with lower current, the electrical losses on the line were smaller, thus saving money every month on the electrical bill. Both of these were important factors as we considered building a petascale system.

The Cray XT5 cabinets at ORNL draw approximately 37 KW per cabinet when running a demanding application such as the high-performance Linpack benchmark, and are rated to as high as 42 KW per cabinet. The maximum power we have seen on the full XT5 system as configured at ORNL is 7 MW. With this power requirement, standard 208 volt power supplies were not an option. The 480 volt supplies that the XT5 uses today allow a 100 amp branch circuit to each cabinet rather than a 200+ amp circuit needed at 208 volts.

3.3 ECOphlex™ Cooling

Equally important to finding efficient ways to get the power in to the computer system is the need to get the heat back out. Cray computer systems have always used innovative cooling techniques to allow the electrical components to be tightly packed, giving the maximum performance possible on each successive generation of technology. With the Cray XT5, Cray has introduced its seventh generation of liquid cooling using a technology they have named ECOphlex™.[9] This cooling system circulates low pressure, liquid R-134a refrigerant (the same use in car air conditioners) through three evaporators where the heat from the cabinet boils the R-134a absorbing the heat through a change of phase from a liquid to a gas as shown in Figure 4. After leaving the cabinet, the gaseous R-134a and any remaining liquid are returned to a heat exchanger at the end of each row of cabinets. In the heat exchanger, up to 100 gallons per minute of cold water circulate to condense the R-134a gas back into a liquid, thus undergoing a second phase change. The R-134a is recirculated in a closed loop. The water is returned to the building chiller plant through a second closed loop system.

The design of ECOphlex was a collaboration of Cray with Liebert, a maker of high-density electrical distribution and cooling equipment for computer centers.

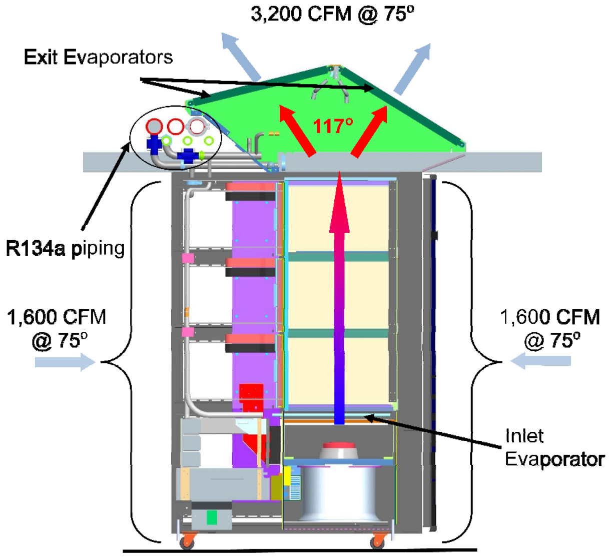

Figure 4 - Jaguar's XT5 Cabinet with ECOphlex™ cooling

The heat exchanger is an enhanced version of Liebert's XDP[10] system in which Cray and Liebert teamed to increase the thermal capacity of each XDP unit. Cray's engineers designed the evaporators and distinctive stainless steel piping that connects each XDP unit to four or five XT5 cabinets. The result is a highly energy efficient method of removing the heat from the computer room and a savings of at least one-million dollars per year in the operational cost of Jaguar over using traditional air-cooling.

4.0 Applications Scaling and Results

The XT4 portion of Jaguar is currently allocated to scientific users through the DOE Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program[11]. Since January 1, 2008, the Jaguar XT4 partition to date has delivered 182 million processor-hours to the INCITE program alone. This partition supports at-scale breakthrough science applications in a "leadership usage" model, whereby large jobs are encouraged and favored. In 2008, for example, 52% of the jobs executed successfully on the Jaguar XT4 partition consumed at least 20% of its total processors. This system is truly a "capability" system in that the majority of its usage is by at-scale applications pursuing breakthrough science goals.

Since its formal acceptance at the end of 2008, the Jaguar XT5 partition has been in a "transition to operations" (T2O) period, where it will remain until early August when the system will be transitioned over to support large-scale science applications as part of the DOE INCITE Program and the Intergovernmental Panel on Climate Change Program. Several key activities are carried out during the T2O period:

- "Science-at scale" simulations with a few selected pioneering applications;
- System monitoring and stabilization; and
- Production workload support and assessment through invited, external friendly usage of the resource.

These activities are undertaken for two reasons:

- Deliver early breakthrough science results for DOE Office of Science before the system enters general availability; and
- Further harden and stabilize the system by subjecting it to more extended usage outside of the bounds and scope of acceptance testing.

Since early 2009, 28 projects have been selected for this T2O period that span virtually all areas of science, but can be categorized in three basic areas:

- Energy for environmental sustainability: Climate change, bioenergy, solar energy, energy storage, energy transmission, combustion, fusion, nuclear energy;
- Materials and nanoscience: Structure of nanowires, nanorods, and strongly correlated materials;
- Fundamental science: Astrophysics, chemistry, nuclear physics.

These 28 projects represent a total Jaguar XT5 allocation of over 540 million processor-hours and over 50 different scientific applications. In its first three months of T2O operation, where almost 150 million processor hours have been consumed by these projects, 32% of the usage on Jaguar XT5 has been with applications running on greater than 45,000 cores and 18% of the usage has been on greater than 90,000 cores. In addition, five 2009 Gordon Bell submissions have been based in part on simulations executed on the Jaguar XT5 during this T2O period. Science results are being generated daily, with the expectation that numerous advances will be documented in the most prestigious scientific journals. Jaguar is truly the world's most powerful computer system and a great system for science.

Porting applications from most parallel systems to Jaguar is very straightforward. The inherent scalability of these applications is a function of the nature of each application and the porting exercises have uncovered new bottlenecks in some of the applications due to the fact that Jaguar is often the largest system on which these applications have been run. Broad spectrums of applications have been running on the system for the last few months. In Table 2 we highlight a few of them. More details of some of these applications and others will be presented in other talks at this meeting.

Science Domain	Code	Cores	Total Performance
Materials	DCA++	150,144	1.3 PF*
Materials	LSMS	149,580	1.05 PF
Seismology	SPECFEM3D	149,784	165 TF
Weather	WRF	150,000	50 TF
Climate	POP	18,000	20 sim yrs/ CPU day
Combustion	S3D	144,000	83 TF
Fusion	GTC	102,000	20 billion Particles / sec
Materials	LS3DF	147,456	442 TF
Chemistry	NWChem	96,000	480 TF
Chemistry	MADNESS	140,000	550+ TF

Table 2 - Application Scaling on Jaguar

The applications above include two Gordon Bell winners, DCA++[12] and LS3DF[13] and a Gordon Bell finalist SPECFEM3D[14.] DCA++ won the Gordon Bell 2008 peak performance award where it achieved the first sustained petaflop performance ever for a scientific application and LS3DF won the Gordon Bell 2008 special category award for innovative algorithms. The LSMS code is also of note since it has achieved more than a Petaflop running on Jaguar. Also the NWChem code is significant because it uses the Global Arrays toolkit which is a one-sided asynchronous programming model. The SeaStar network is a message passing network and getting Global Arrays to work efficiently has been a significant challenge solved by a collaborative effort among PNNL, ORNL and Cray.

Acknowledgments

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.

About the Authors

Arthur S. "Buddy" Bland is the Project Director for the Leadership Computing Facility project at Oak Ridge National Laboratory. He can be reached at **BlandAS at ORNL.GOV**.

Ricky A. Kendall is the group leader for Scientific Computing at the National Center for Computational Sciences at ORNL. He can be reached at **KendallRA at ORNL.GOV**.

Douglas B. Kothe is the Director of Science for the National Center for Computational Sciences at ORNL. He can be reached at **Kothe at ORNL.GOV**.

James H. Rogers is the Director of Operations for the National Center for Computational Sciences at ORNL. He can be reached at **jrogers at ORNL.GOV**.

Galen M. Shipman is the Group Leader for Technology Integration of the National Center for Computational Sciences at ORNL. He can be reached at **gshipman at ORNL.GOV**.

[1] Zacharia, Thomas, et.al.; *National Leadership Computing Facility: A Partnership in Computational Science*, April 2004; ORNL/TM-2004/130

[2] http://www.cs.sandia.gov/platforms/RedStorm.html

[3] http://www.clusterresources.com/

[4] http://www.datadirectnet.com/pdfs/S2A9900Brochure041408A.pdf

[5] Anderson, D., Trodden, J., *HyperTransport System Architecture,* Addison-Wesley, 2003, ISBN 0321168453, 9780321168450

[6] Brightwell, R.; Pedretti, K.; Underwood, K.D., "Initial performance evaluation of the Cray SeaStar interconnect," *High Performance Interconnects, 2005. Proceedings. 13th Symposium on* , vol., no., pp. 51-57, 17-19 Aug. 2005, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1544577&isnumber=32970

[7] Shipman, G., et. al: *The Spider Center Wide File System; From Concept to Reality;* CUG 2009 Proceedings

[8] Maxwell, D. et. al.; *Integrating and Operating a Conjoined XT4+XT5 System; CUG 2009 Proceedings*

[9] Gahm, D., Laatsch M.; *Meeting the Demands of Computer Cooling with Superior Efficiency: The Cray ECOphlex™ Liquid Cooled Supercomputer offers Energy Advantages to HPC users*; Cray Inc.; WP-XT5HE1; 2008; http://www.cray.com/Assets/PDF/products/xt/whitepaper_ecophlex.pdf

[10] http://www.liebert.com/product_pages/Product.aspx?id=206&hz=60; Emerson Network Power

[11] http://hpc.science.doe.gov/; US DOE INCITE Program

[12] http://doi.acm.org/10.1145/1413370.1413433

[13] http://doi.acm.org/10.1145/1413370.1413437

[14] http://doi.acm.org/10.1145/1413370.1413432