Jaguar: The World's Most Powerful Computer System – An Update

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

ABSTRACT: *At the SC'09 conference in November 2009, Jaguar was crowned as the world's fastest computer by the web site www.Top500.org. In this paper, we will describe Jaguar, present results from a number of benchmarks and applications, and discuss future computing in the Oak Ridge Leadership Computing Facility. This paper updates the CUG 2009 paper that described Jaguar with new information about the upgrade from 4-core to 6-core processors, and ORNL's plans for exaflops computing by 2018.*

KEYWORDS: Jaguar, XT4, XT5, Lustre, Exascale, OLCF, Top500, Istanbul

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[2] 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[3] in 2005, and an upgrade of that system to a 54 TF dual-core system in 2006, an addition of a 65 TF XT4[4] in 2006 that was combined with the XT3 in 2007 to make Jaguar XT4 a 119 TF system. In 2008 the system was further upgraded to quad-core processors increasing performance to 263 TF with 62 terabytes (TB) of system 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 OLCF took delivery of the first 16 of 200 cabinets of an XT5 system providing an additional 1,375 TF, 300 TB of system memory and over 10,000 TB of total disk space to the OLCF. The final cabinets were delivered on September 17, 2008. Twelve days later on September 29th, this incredibly large and complex system, known as Jaguar XT5, ran a full-system benchmark application that took two and one-half hours to complete. The system completed acceptance testing in December 2008 and was transitioned to operations shortly thereafter. In late July 2009, a phased upgrade of the Jaguar XT5 system from 2.3 GHz 4-core Barcelona AMD processors to 2.6 GHz 6-core Istanbul AMD processors began. This phased approach provided users with a substantial fraction of the Jaguar XT5 platform throughout the upgrade period, minimizing the impact to our users. Completion of this upgrade on November 7th increased performance of Jaguar XT5 to 2,332 TF with improved memory bandwidth due to enhancements in the Istanbul's memory controller. Today, Jaguar XT5 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, astrophysics, and nuclear physics.

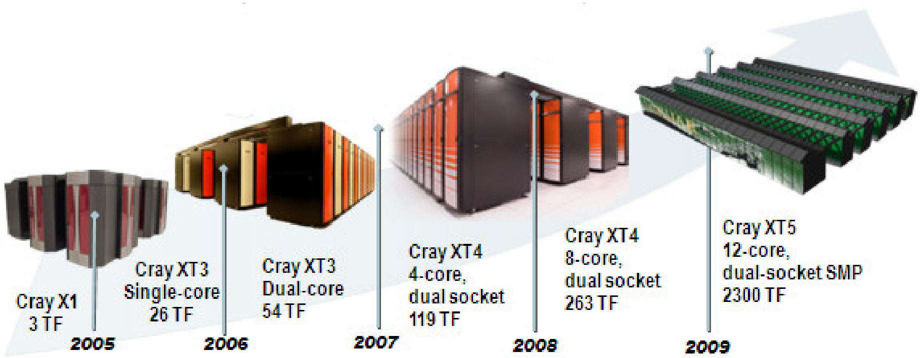

Figure 1: OLCF Petascale Roadmap

The successful deployment of the Jaguar XT5 platform represents a landmark in the OLCF's roadmap to petascale computing. This roadmap, illustrated in Figure 1, was based on phased scaling steps of system upgrades and new system acquisitions, allowing a concurrent phased approach to facility upgrades, system software development, and dramatic application scalability improvements over an acceptable time horizon.

Having successfully transitioned from sustained teraflop to sustained petaflop computing, the OLCF has laid out a similar roadmap to transition to exaflop computing as illustrated in Figure 2. The expectation is that systems on the way to Exascale will be composed of a relatively modest number of CPU cores optimized for single thread performance coupled with a large number of CPU cores optimized for multi-threaded performance and low-power consumption. Images in Figure 2 are placeholders and are meant to illustrate that systems could be quite different at each stage of the roadmap.

2. Scaling Applications and Supporting Infrastructure

Successful execution of our 5-year roadmap to sustained petascale computing required focused efforts in improving application scalability, facility upgrades, and scaling out our supporting infrastructure. In 2005 our initial installation of the Jaguar platform resulted in a 3,748 core XT3, subsequent upgrades and additions have resulted in nearly a 60 fold increase in the number of CPU cores in the current Jaguar system (224,256 cores). Scaling applications and infrastructure over this period, while challenging, has been critical to the success of the OLCF.

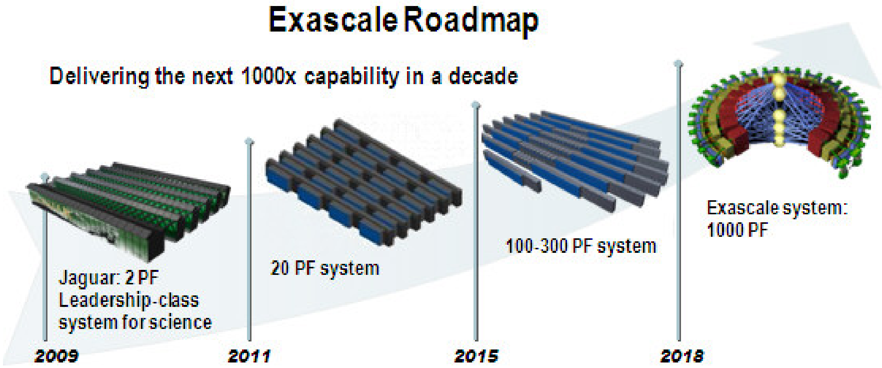
Exascale Roadmap
Delivering the next 1000x capability in a decade
Jaguar: 2 PF Leadership-class system for science
20 PF system
100-300 PF system
Exascale system: 1000 PF
2009
2011
2015
2018

Figure 2: OLCF Exascale Roadmap

2.1. Improving Application Scalability

Recognizing that improving application scalability would require a concerted effort between computational scientists with domain expertise and the science teams and that a gap in computational science expertise existed in many of these science teams, the OLCF created the Scientific Computing group in which group members act as liaisons to science teams helping them improve application scalability, performance, and scientific productivity. To compliment these efforts, Cray established a Supercomputer Center of Excellence (COE) at the OLCF in 2005 providing expertise in performance optimization and scalability on Cray platforms. The liaison model coupled with the Cray COE has proven successful in improving application performance and scalability. As an example, beginning in September 2009, the OLCF worked closely with a number of key applications teams to demonstrate application scalability on the Jaguar XT5 platform. Work on a scalable method for ab-initio computation of free energies in nanoscale systems using the locally self-consistent multiple scattering method (LSMS)[5] resulted in sustained performance of 1.836 Petaflops/sec on 223,232 CPU cores, claiming the 2009 Gordon Bell Award at SC2009. Using the NWChem application, a study[6] of the CCSD(T) calculations of the binding energies of H_2O resulted in sustained performance of 1.4 Petaflops/sec on 224,196 CPU cores, a 2009 Gordon Bell Award finalist at SC2009. This result was particularly notable from a system and software architecture perspective as NWChem's communication model is based on the Global Arrays[7] framework rather than the dominant communication model in large-scale HPC, MPI[8].

2.2. Scaling Supporting Infrastructure

To scale Jaguar from a 26TF system in 2005 to a 2.332 PF system in 2009 required both facility upgrades to our 40,000 square foot facility and scaling out of our supporting infrastructure. Power requirements necessitated facility upgrades from 8 MW in 2005 to 14 MW in 2009. Power efficiency improvements such as transitioning from a traditional 208-volt power distribution system to a 480-volt power distribution system resulted in energy savings estimated at nearly $500,000 while dramatically decreasing the cost of our power distribution installation. Cooling infrastructure required dramatic improvements in efficiency and capacity. Over this four-year period the OLCF's cooling capacity grew from 3,600 tons to 6,600 tons. In order to efficiently cool the Jaguar XT5 system, a new cooling system was designed for the XT5 platform using R-134a refrigerant within the XT5 cabinet and chilled-water heat exchangers to remove the heat. This system requires less than 0.2 Watts of power for cooling for every 1 Watt of power used for computing, substantially better than the industry average ratio of 0.8 : 1.

In addition to facility enhancements, the OLCF required dramatic scalability and capacity improvements in its supporting infrastructure particularly in the areas of parallel I/O and archival storage. Scaling from 8 GB/sec of peak I/O performance and 38.5 TB of capacity on Jaguar XT3 in 2005 to over 240 GB/sec and 10 PB of capacity on Jaguar XT5 in 2009 required improvements in parallel I/O software[9], storage systems, and system area network technologies. Archival storage requirements have grown dramatically with less than 500 TB of data stored in 2005 to over 11 PB as of this writing, requiring improvements in archival storage software[10] and tape archive infrastructure.

3. System Architecture

Jaguar is a descendent of the Red Storm[11] 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 2,332 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[12]. 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[13] 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.

3.1. Node Architecture

The XT5 compute nodes are general-purpose symmetric multi-processor, shared memory building blocks designed specifically to support the needs of a broad range of applications. Each node has two AMD Opteron 2435 “Istanbul” six-core processors running at 2.6 GHz and connected to each other through a pair of HyperTransport[14] connections. Each of the Opteron processors has 6 MB of level 3 cache shared among the six 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 twelve cores, 16 GB of memory, and 25.6 GB per second of memory

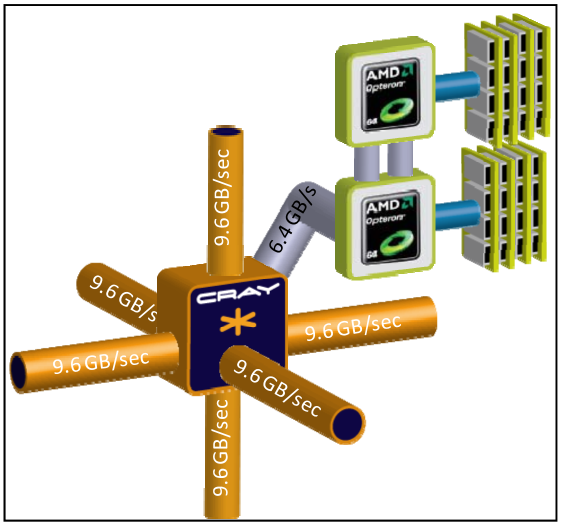

Figure 3: XT5 Node Architecture

bandwidth. The node has a theoretical peak processing performance of 124.8 billion floating-point operations per second (GF). A schematic representation of the node is shown in Figure 3.

The interconnection fabric that links the nodes is implemented using Cray's SeaStar[15] [16] 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 4 GB/s (bidirectional).

3.2. Building up the XT5 System

Even with very powerful nodes, the XT5 requires 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 system uses 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 12 TF per rack, the XT5 is among the densest server configurations available. Jaguar XT5 is made up of 200 of these cabinets in a configuration of eight rows of 25 cabinets each. The configuration of Jaguar XT5 system is shown in Table 1.

System	**XT5**
Cabinets	200
Compute Blades	4,672
Six-core Opteron Processors	37,376
Cores	224,256
Peak TeraFLOPS	2,332
Compute Nodes	18,688
Memory (TB)	300
Number of disks	13,440
Disk Capacity (TB)	10,000
I/O Bandwidth (GB/s)	240

Table 1: Jaguar XT5 System Configuration

3.3. The Spider Parallel File System

Spider[17], 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 OLCF 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 3 terabytes of RAM. Each object storage server can provide in excess of 1.25 GB per second of file system level performance and manages seven Object Storage Targets (RAID 6 Tiers). 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 OLCF. Figure 4 illustrates the Spider architecture.

On the Jaguar XT5 partition 192 service & I/O (SIO) nodes, each with a dual-core 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[18] host channel adaptors and Zarlink CX4[19] optical InfiniBand cables. These SIO nodes are configured as Lustre routers to allow compute nodes within the SeaStar2+ torus to access the Spider file system 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. The Spider parallel file system is mounted by all major OLCF systems, including the XT4 and XT5, 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.

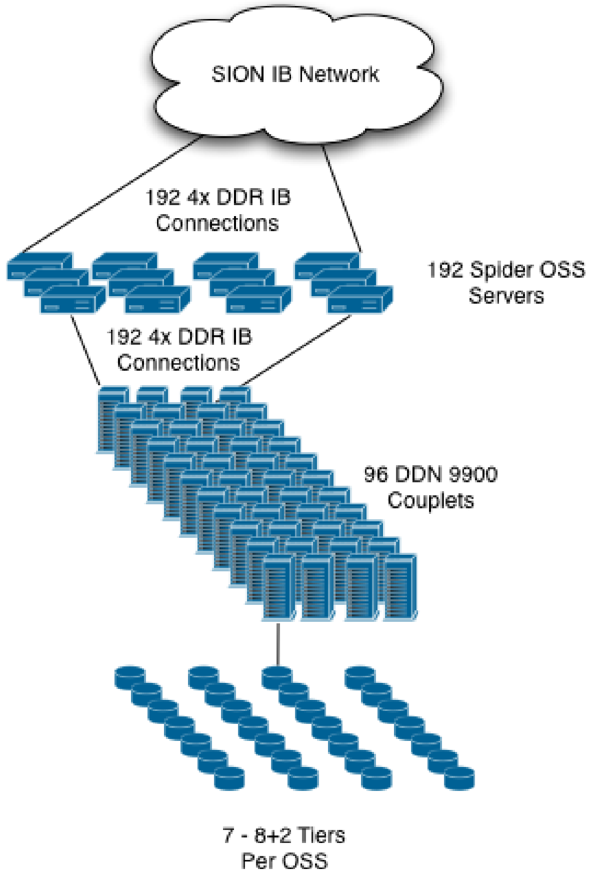

Figure 4: The Spider File System Architecture

3.4. The Scalable I/O 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 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. SION currently connects both segments (XT4 and XT5) of the Jaguar with the Spider file system, external login nodes, Lens (Visualization cluster), Ewok (end-to-end cluster), Smoky (application readiness cluster), and to HPSS and GridFTP[20] servers. SION is a high performance InfiniBand DDR network providing over 889 GB/s of bisection bandwidth. Figure 5 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 OLCF 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.

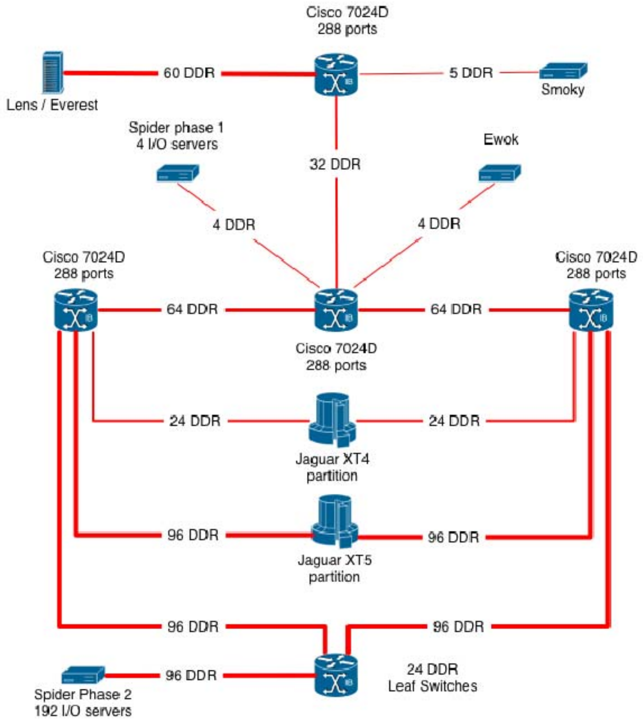

Figure 5: The SION Infrastructure

3.5. External Login Nodes

Traditional Cray XT system configurations contain heterogeneous combinations of compute nodes, I/O nodes, and login nodes. ORNL has developed an external login node[21] that eliminates the need to dedicate further nodes to login responsibilities. These login nodes are based on AMD Quad-Core Opteron processors. They 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 Cray 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 Cray XT system.

4. 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 US$1,000,000 on the site preparation and installation, and was essential to be able to cool the system.

4.1. 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 OLCF identified some innovative opportunities during site preparation that reduced up-front material costs by over US$1,000,000, and will reduce operational costs due to voltage losses for the life of the system. Electrical distribution to ORNL 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 OLCF chilled water plant includes five separate chillers configured to automatically adjust to load conditions up to 6,600 tons, the equivalent of 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.

4.2. 480 Volt Electrical Infrastructure

When the OLCF project began in 2004, 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, ORNL 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 7.6 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.

4.3. 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

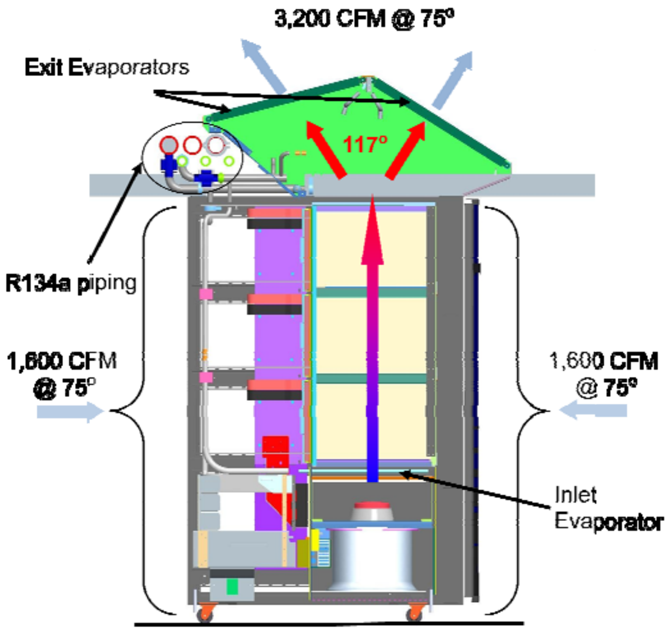

Figure 6: Cray XT5 Cabinet with ECOphlex Cooling

liquid cooling using a technology they have named ECOphlex™.[22] 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 6. 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. The heat exchanger is an enhanced version of Liebert's XDP[23] 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 US$1,000,000 per year in the operational cost of Jaguar over using traditional air-cooling.

4.4. Jaguar XT5 Power Usage

Our experience shows that Jaguar XT5 has an average power consumption of 5 MW when running a typical mix of production jobs, but can exceed 7 MW for applications that keep all the floating point units active such as during an HPL Linpack run. The absolute power consumption is large because Jaguar is just so much larger in both FLOPS as well as memory than the other top systems, as shown in Table 2. When compared to other large-scale systems using normalized metrics of MFLOPS/Watt and Mbytes/Watt, the Cray XT5 is the third most efficient system. Only the IBM BG/P and the accelerator based Roadrunner are more efficient.

		Linpack	System	Power	MFlops/	MBytes/
System	Platform	Rmax (TFs)	Memory (GB)	(kW)	Watt	Watt
Jaguar XT5 (ORNL)	Cray/XT5	1759	300,000	6950.6	253.07	43.16
Road Runner (LANL)	IBM/Cell	1042	103,600	2345.5	444.25	44.17
Jugene (JSC)	IBM/BGP	826	144,000	2268	363.98	63.49
Pleiades (NASA)	SGI/Altix ICE	544	74,700	2348	231.81	31.81
Blue Gene/L (LLNL)	IBM/BGL	478	53,248	2329.6	205.19	22.86
Ranger (TACC)	Sun/Constellation	433	15,640	2000	216.6	7.82

Table 2: Power Usage on Leadership Systems

5. Performance Results

The XT5 portion of Jaguar is currently allocated to scientific users through the DOE Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program[24]. In 2010, the Jaguar XT5 partition will deliver nearly 950 million processor-hours to the INCITE program alone. This supports at-scale breakthrough science applications in a "leadership usage" model, whereby large jobs are encouraged and favored. During the 2009 Jaguar XT5 transition to operations period, over 55% of the processor hours used on the Jaguar XT5 partition were consumed by jobs using 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.

After its formal acceptance at the end of 2008, the Jaguar XT5 partition went through a "transition to operations" (T2O) period, until mid-July 2009, when the system was transitioned to support large-scale science applications as part of the DOE INCITE Program. Several key activities were 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 were undertaken for two reasons:

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

In early 2009, 26 projects were selected for the T2O period. These applications 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 and nuclear energy;
- Materials and nanoscience: Structure of nanowires, nanorods and strongly correlated materials;
- Fundamental science: Astrophysics, chemistry and nuclear physics.

Science Domain	Code	Cores	Total Performance
Materials	DCA++	213,120	1.9 PF
Materials	WL-LSMS	223,232	1.8 PF
Seismology	SPECFEM3D	149,784	165 TF
Weather	WRF	150,000	50 TF
Climate	POP	18,000	20 sim yrs/ CPU day
Combustion	S3D	210,000	120 TF
Fusion	GTC	153,600	27 billion Particles / sec
Nano Materials	OMEN	222,720	860 TF
Materials	LS3DF	147,456	442 TF
Chemistry	NWChem	224,196	1.4 PF
Chemistry	MADNESS	140,000	550+ TF

Table 3: Application Performance and Scaling

These 26 projects represent total Jaguar XT5 usage of over 350 million processor-hours and over 50 different scientific applications. During the months of T2O operation, 57% of the usage on Jaguar XT5 was with applications running on greater than 30,000 cores, and 14% of the usage was on greater than 90,000 cores. In addition, five 2009 Gordon Bell submissions were 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.

Following the Jaguar XT5's T2O period and subsequent production phase, Jaguar went through a processor upgrade from AMD quad-core to six-core processors, increasing the available compute cores to 224,256 and increasing peak performance by 69%. This upgrade put Jaguar at the top of the TOP500 list, making it, at 2.332 petaflops peak speed, the world's fastest computer. The resulting additional processor-hours and compute capabilities will accelerate science discovery to an even greater degree.

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 in recent months. In Table 3, we highlight a few of them.

The applications above include three Gordon Bell winners, WL-LSMS, DCA++[25] and LS3DF,[26] and two Gordon Bell finalists, NWCHEM and SPECFEM3D.[27.] WL-LSMS won the Gordon Bell 2009 peak performance award for highest performance for a scientific application. The NWChem code, a Gordon Bell 2009 finalist, is significant because it uses the Global Arrays toolkit which is a one-sided asynchronous programming model. The SeaStar network is optimized for message passing and presented some unique challenges to obtaining efficient and scalable performance when using Global Arrays. These challenges were solved by a collaborative effort among Pacific Northwest National Laboratory, ORNL and Cray.

The OLCF Jaguar system was designed to further the science goals of the DOE Advanced Scientific Computing Research (ASCR) Program, which are:

- Energy security,
- Nuclear security,
- Scientific discovery and innovation, and
- Environmental responsibility.

To support these goals, multiple science areas and applications are supported on OLCF

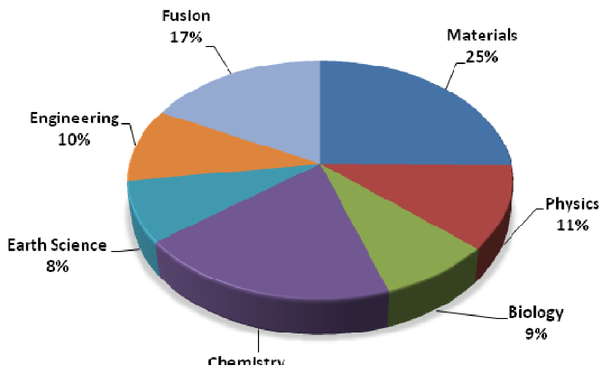

Figure 7: Jaguar Usage by Science Category

platforms. Figure 7 illustrates Jaguar XT5 usage by science area for the T2O period. The diversity of usage for different science areas highlights the need for leadership computing systems to perform effective science across a diverse portfolio of science domains.

Delivering breakthrough science through the Jaguar XT5 system requires a well-balanced selection of hardware subsystem capabilities. Figure 8 illustrates, for the Jaguar XT5 quad-core system, the proportion of application time spent in CPU, memory access, interprocessor communication and I/O for selected application runs and processor counts in key science areas. The results show that applications tend to stress multiple hardware subsystems; thus, leadership-class platforms must deliver good performance for a range of hardware characteristics, such as interconnect bandwidth, node peak flops, memory bandwidth, mean time to interrupt, node memory capacity, memory latency and interconnect latency. Though I/O performance requirements are not stressed in these examples as check pointing and other I/O tasks were minimized or disabled, the need for fast, reliable I/O is also extremely important to support the ever-growing data requirements for leadership-class systems.

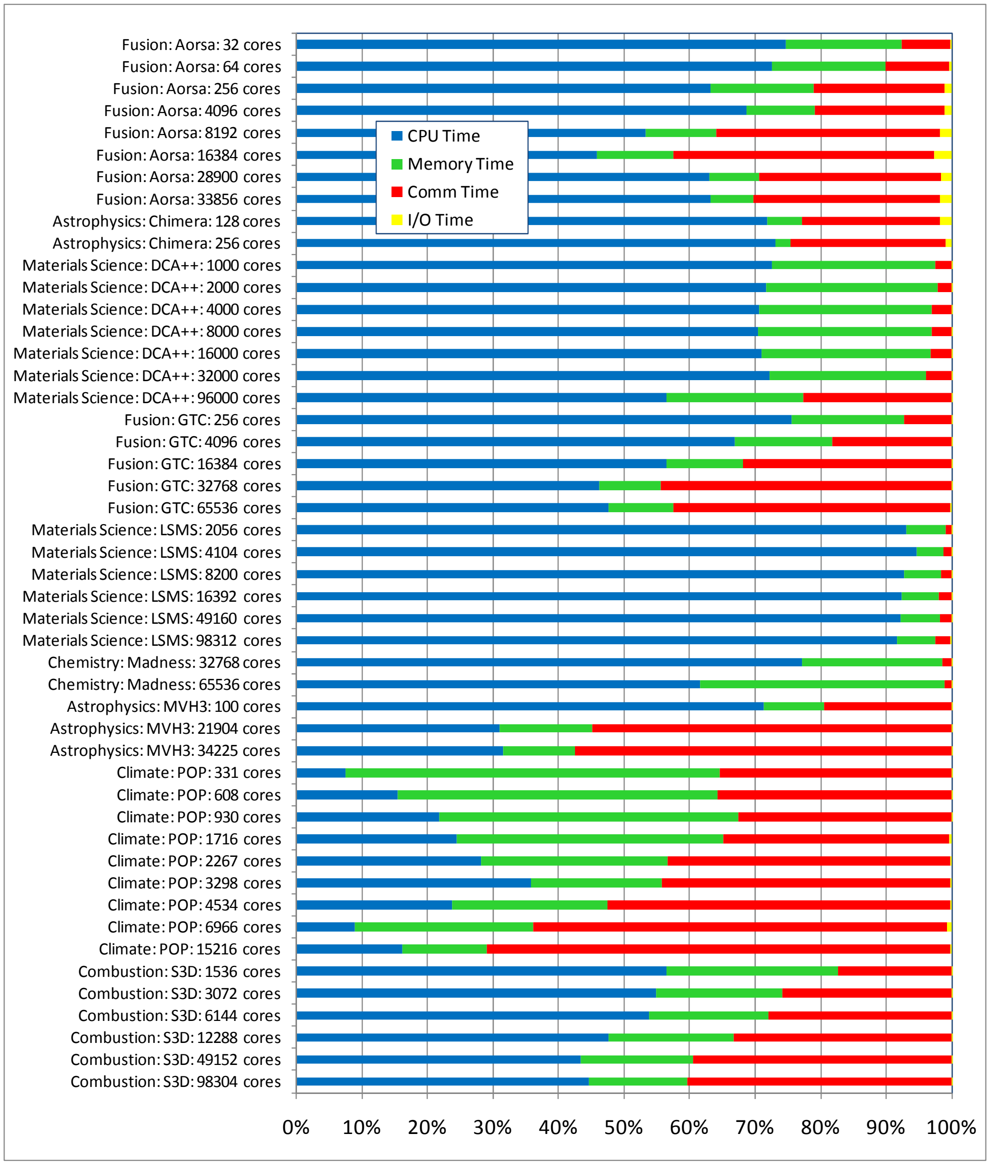

Figure 8: Application Runtime by Hardware Subsystem

Power consumption is a growing concern for HPC systems. Figure 9 illustrates power consumption per processor core for selected dedicated application runs on the Cray XT5 quad-core platform. Power consumption profiles such as this can help inform hardware and software decisions to optimize "science output per watt", a critical factor in the path to exaflop computing.

6. Roadmap to Exascale

Having successfully deployed a series of high performance systems, transitioning from the teraflop to the petaflop era, the OLCF is poised to lead the transition from the petaflop to exaflop era. The transition from the teraflop to the petaflop era required focused effort in a number of key areas such as power and cooling efficiency, system software scalability, and most importantly, dramatically increasing the scalability of applications. This transition was marked primarily by an increase in system scale as single thread performance increased by less than 10%, core count increased by nearly 60x.

The transition from the petaflop to exaflop era will be dramatically different. A primary constraint in the transition to exaflop computing will be total power consumption of no more than 40 MW. This constraint coupled with heat dissipation challenges in multi-core architectures optimized for single threaded performance will require a transition to hybrid multi-core architectures. These systems will be composed of a relatively modest number of CPU cores optimized for single thread performance coupled with a large number of CPU cores optimized for multi-threaded performance and low-power consumption. Given this landscape, the OLCF is embarking on our roadmap to exaflop computing with the delivery of a hybrid multi-core system in 2011/2012. This system will present unique challenges in achieving optimal performance of scientific applications. Whereas the OLCF has traditionally focused efforts on improving application performance through improved scalability across a large number of CPU cores, efforts are now underway to optimize applications for hybrid-multicore architectures. These efforts include work at all levels of the hierarchy including compilers, scientific libraries, application kernels, and algorithms. While a number of challenges remain, the deployment of this system will usher in a new era, providing a baseline for next generation systems and allow scientific applications to embark on a successful transition to exaflop computing.

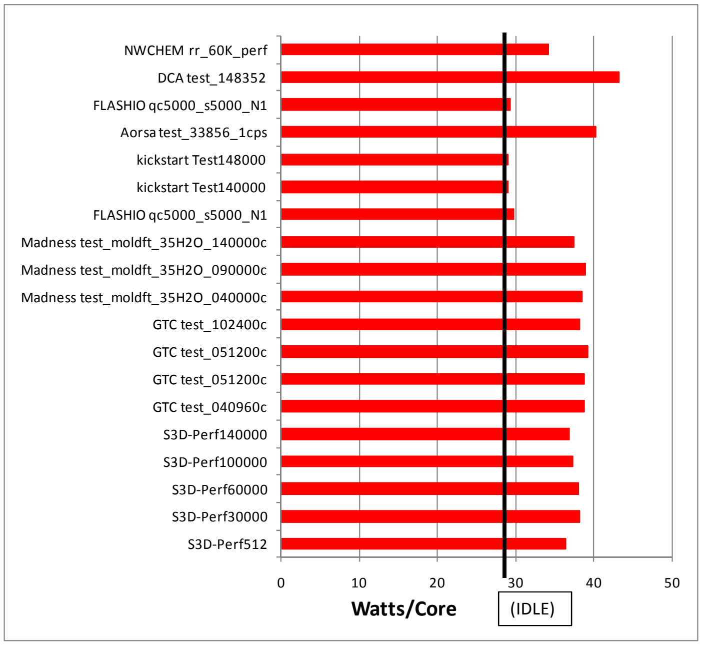

Figure 9: Application Power on Jaguar XT5

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**.

Wayne Joubert is a staff member in the Scientific Computing Group at ORNL. He can be reached at **joubert 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] Dunigan, Jr., T. H, et. Al.; *Early Evaluation of the Cray X1*, SC '03: Proceedings of the 2003 ACM/IEEE conference on Supercomputing, 18 (2003).

[3] Alam, Sadaf, et.al.; *An Evaluation of the Oak Ridge National Laboratory Cray XT3*, International Journal of High Performance Computing Applications, Volume 22, Issue 1 (2008).

[4] Alam, Sadaf, et.al.; *Cray XT4: an Early Evaluation for Petascale Scientific Simulation*, SC '07: Proceedings of the 2007 ACM/IEEE conference on Supercomputing, 1 (2007).

[5] Eisenbach, Markus, et.al.; *A Scalable Method for ab initio Computation of Free Energies in Nanoscale Systems*, SC '09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, Article 64 (2009).

[6] Apra, Edoardo, et.al., *Liquid Water: Obtaining the Right Answer for the Right Reasons*, SC '09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, Article 66 (2009).

[7] Nieplocha, Jaroslaw, et.al., *Global Arrays: a Non-uniform Memory Access Programming Model for High-Performance Computers*, The Journal of Supercomputing, Volume 10, Issue 2, 169 (1996).

[8] The MPI Forum, *MPI: A Message Passing Interface*, Proceedings of the 1993 ACM/IEEE conference on Supercomputing, 878 (1993)

[9] Sun Microsystems Inc, *Lustre File System. High-Performance Storage Architecture and Scalable Cluster File System*, http://www.sun.com/servers/hpc/docs/lustrefilesystem_wp.pdf

[10] Danny Teaff, Dick Watson,Bob Coyne, *The Architecture of the High Performance Storage System (HPSS)*, In Proceedings of the Goddard Conference on Mass Storage and Technologies, 28, 1995.

[11] W. J. Camp and J. L. Tomkins, *Thor's Hammer: The First Version of the Red Storm MPP Architecture*, Proceedings of the SC 2002 Conference on High Performance Networking and Computing, (2002).

[12] Cluster Resources Inc.; http://www.clusterresources.com/

[13] Data Direct Networks Inc.; http://www.datadirectnet.com/pdfs/S2A9900Brochure041408A.pdf

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

[15] Brightwell, R.; Pedretti, K.; Underwood, K.D., *Initial Performance Evaluation of the Cray SeaStar Interconnect*, HOTI '05: Proceedings of the 13th Symposium on High Performance Interconnects, 51 (2005).

[16] Brightwell, R., Pedretti, K. T., Underwood, K. D., and Hudson, T,; *SeaStar Interconnect: Balanced Bandwidth for Scalable Performance*, IEEE Micro, Volume 26, 3 (2006)

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

[18] Sur, S., Koop, M. J., Lei, and Panda, D. K.; *Performance Analysis and Evaluation of Mellanox ConnectX InfiniBand Architecture with Multi-Core Platforms*. In Proceedings of the 15th Annual IEEE Symposium on High-Performance interconnects, 125 (2007).

[19] Products, Computer, Volume 36, Issue 9, 94 (2003).

[20] Allcock, W., Bresnahan, J., Kettimuthu, R., Link, M., Dumitrescu, C., Raicu, I., and Foster,; *The Globus Striped GridFTP Framework and Server*., In Proceedings of the 2005 ACM/IEEE Conference on Supercomputing, 54 (2005).

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

[22] 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

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

[24] US Department of Energy; http://hpc.science.doe.gov/

[25] Alvarez, G., Summers, M. S., Maxwell, D. E., Eisenbach, M., Meredith, J. S., Larkin, J. M., Levesque, J., Maier, T. A., Kent, P. R., D'Azevedo, E. F., and Schulthess,; *New Algorithm to Enable 400+ TFlop/s Sustained Performance in Simulations of Disorder Effects in high-T_c Superconductors*. In Proceedings of the 2008 ACM/IEEE Conference on Supercomputing, Article 61 (2008).

[26] Wang, L., Lee, B., Shan, H., Zhao, Z., Meza, J., Strohmaier, E., and Bailey, D. H.; *Linearly scaling 3D fragment method for large-scale electronic structure calculations*, In Proceedings of the 2008 ACM/IEEE Conference on Supercomputing, Article 65 (2008).

[27] Carrington, L., Komatitsch, D., Laurenzano, M., Tikir, M. M., Michéa, D., Le Goff, N., Snavely, A., and Tromp, J.; *High-Frequency Simulations of Global Seismic Wave Propagation using SPECFEM3D_GLOBE on 62K Processors*. In Proceedings of the 2008 ACM/IEEE Conference on Supercomputing, Article 60 (2008).