Leadership Computing at Oak Ridge National Laboratory

Studham RS, Kuehn JA, White JB, Fahey M, Carter S, Nichols JA

***Abstract*—Oak Ridge National Laboratory is running the world's largest Cray X1, the world's largest unclassified Cray XT3, and a Cray XD1. In this report we provide an overview of the applications requiring leadership computing and the performance characteristics of the various platforms at ORNL. We then discuss ways in which we are working with Cray to establish a roadmap that will provide 100's of teraflops of sustained performance while integrating a balance of vector and scalar processors.**

***Index Terms*— Supercomputers**

I. Introduction

Computing power has become central to scientific leadership in many fields, and over the past five years other nations have begun to outpace the United States in delivering more efficient and powerful computers for open (non-classified) scientific discovery. In May of 2004, the Executive Office of the President of the United States issued the Federal Plan for High-End Computing which called for "Leadership Systems" to provide leading-edge computational capability for high-priority research problems. [i]

> *"The goal of such systems is to provide computational capability that is at least 100 times greater than what is currently available."*
> *"…Leadership Systems are expensive, typically costing in excess of $100 million per year…."*

Computational science capabilities already underpin the research and development that the Department of Energy (DOE) conducts to meet its energy and national security missions. Because these capabilities are central to DOE's missions, and critical to long-term national competitiveness, the DOE's Office of Science brought renewed focus to this challenge. Secretary of Energy Spencer Abraham produced the *Facilities for the Future of Science: A Twenty-Year Outlook*, which listed Leadership Computing (UltraScale Scientific Computing Capability) as the top DOE domestic priority [reference].[ii]

> "*[Leadership Computing] … will increase by a factor of 100 the computing capability available to support open (as opposed to classified) scientific research—reducing from years to days the time required to simulate complex systems, such as the chemistry of a combustion engine, or weather and climate—and providing much finer resolution."*

On February 23, 2004, in response to both the Federal High-End Computing Plan and the DOE twenty-year facilities plan, the DOE Office of Science gave "Notice to SC Laboratories" of a request for proposals for a "Leadership-Class Computing Capability for Science." The main points to address in this solicitation were:

- *The focus of the proposed effort should be on **capability computing** in support of high-end science – rather than on enhanced computing capacity for general science users;*
- *The proposed effort must be a **user facility providing leadership class computing capability** to scientists and engineers nationwide independent of their institutional affiliation or source of funding.*

The National Center for Computational Science (NCCS) at Oak Ridge National Laboratory (ORNL) responded to this call with their proposal, "Establishing a National Leadership Computing Facility: A Partnership in Computational Sciences." ORNL proposed a vision to establish a National Leadership Computing Facility (NLCF) to develop and deploy capability computing for open scientific research at an unprecedented scale and to maintain leadership in capability computing for the nation. The NLCF would become a unique world-class scientific resource that would provide the scientific community with orders of magnitude more computing capability (performance, memory, *etc*.) than is now available. While establishing this new level of scientific capability for the nation, the NLCF would proactively engage the scientific and engineering communities and issue calls for proposals to realize the breakthroughs promised by its extraordinary capabilities.

Years of experience have shown that no one supercomputer architecture is best for all science problems. The NLCF proposed to provide leadership-class capability computing in both vector and scalar architectures in order to cater to the widest possible range of strategic scientific areas and problems. Initially, this is being accomplished through two different systems purpose-built for science – the Cray X1E and the Cray XT3. In 2006, upgrades to these architectures are available which can be fielded at the proposed leadership-

class level. By 2007 these will be merged into a hybrid computer architecture code-named "Rainier" that combines the strengths of the X1E and XT3.[iii]
At the proposed leadership level of funding, the NLCF hardware roadmap has a 100+ TF Cray system deployed in 2006 and a 250 TF Cray system deployed in 2007.

II. Leadership Computing Facility

A. Usage Model

The leadership computing systems will deliver at least 100 times greater computational resources to key problems than what is generally available for advanced scientific and engineering simulations. In 2005 a typical supercomputer center such as the National Energy Research Supercomputer Center (NERSC) at Lawrence Berkeley National Laboratory operates a 10 TF (peak) computational resource and hosts 303 projects.[iv] This represents a delivered 0.027 TF-years per project. With the charter of delivering 100 times greater than what is generally available, a leadership computing facility needs to deliver 2.4 TF-years per project in 2005. These numbers may roughly double every year,[v] which will require the NLCF to focus on a handful of projects and continue to provide some of the largest computing systems in the world.

The NLCF will focus on high priority, challenging, high payoff, and heretofore intractable computationally intensive experiments, where the capability of the NLCF systems can enable new breakthroughs in science. It is expected that the leadership systems will enable the United States to be "first to market" with important scientific and technological capabilities, ideas, and software. A limited set of scientific applications (perhaps 10 per year) will be selected and given substantial access to the leadership systems. Particular consideration will be given to proposals that demonstrate and/or contribute to the creation of computational capabilities that extend the power and reach of computational science in important research domains, and that offer the potential of making those capabilities broadly available to the scientific community. During 2006 this program will deliver a total of 6 million processor hours on a 1,024 vector-processor Cray X1E system and 31 million processor hours on a 5,212 scalar-processor Cray XT3 system, with allocations beginning October 1, 2005.[vi]

B. Platforms

1) Cray X1 and X1E

The Cray X1 is the current generation of high-bandwidth vector systems from Cray. Each multi-streaming processor (MSP) has a peak performance of 12.8 GF, with a peak of 34.1 GB/s of memory bandwidth. The X1 augments this with uniquely high bandwidth to remote memory and to random-stride memory, both measured at over 10 GB/s per MSP. Because of the capability to perform vector load and store operations to remote memory, the X1 also has very low latency communication with multiple processors simultaneously.

The X1 in the NCCS currently has 128 nodes; each node has 4 MSPs and 16 GB of memory, for a total of 512 MSPs and 2 TB of memory. The X1 has 32 TB of local disk, connected through 32 1-Gb/s Fibre-Channel (FC) ports.

On February 10, 2004, the Office of Advanced Scientific Computing Research conducted an external review of the Cray X1 evaluation at the NCCS. The review panel found that *"[the] Cray X1 … is indeed a very large system capable of solving very large and important science problems*." Further, "*The committee believes either increasing the size of the current Cray X1 configuration or acquiring a larger version of Cray's planned follow-on systems would be a highly worthwhile investment…*" [vii]

The Cray X1E upgrade planned for the summer of 2005 replaces each processor of the X1 with two faster processors. Each X1E MSP has an improved peak of 18 GF, a faster vector cache, and more efficient use of the memory system. The resulting system will have 1,024 MSPs in 256 nodes. This is a very low risk upgrade that will be accomplished through a processor swap on the existing node boards. The system software and user programming environment will be an update to the existing X1 software.

2) Cray XT3

The Cray XT3 complements the X1E by coupling a high-bandwidth, low-latency interconnect with high-speed scalar processors. The XT3 uses 2.4 GHz AMD Opteron processors connected to a Cray-engineered interconnect through an industry-standard HyperTransport interface. This interface provides a theoretical peak bandwidth of 3.2 GB/s in each direction, though software and control overhead limit measured unidirectional MPI bandwidth to about 1.1 GB/s. Regardless, the XT3's ratio of MPI communication bandwidth to peak performance is surpassed only by the X1 series. The XT3 interconnect is a 3D torus with 7.6 GB/s of peak bandwidth per link. The bandwidth allows the torus to remain scalable while mitigating contention from heavy global communication.

The NCCS XT3 configuration includes the latest available AMD processors, each with 2 GB of memory and scalable I/O connectivity to a globally shared file system. The configurations of the XT3 base and proposed leadership systems are given in Table 1.

Table 1. XT3 base and proposed configurations

Status	Performance	Processors	Memory	I/O Bandwidth
Funded	25 TF	5,304	10.5 TB	15 GB/s
Proposed	50 TF	11,374	23 TB	30 GB/s
Proposed	100 TF	22,748	46 TB	60 GB/s

C. Infrastructure

1) Leadership Networking

To share the world-class resources that it manages, the NCCS is making a significant commitment in both local- and wide-area high-speed networking. This commitment includes a substantial local-area network upgrade to allow multiple 10-Gb/s connections to hosts and adequate backbone capacity to handle the resulting high-speed flows.

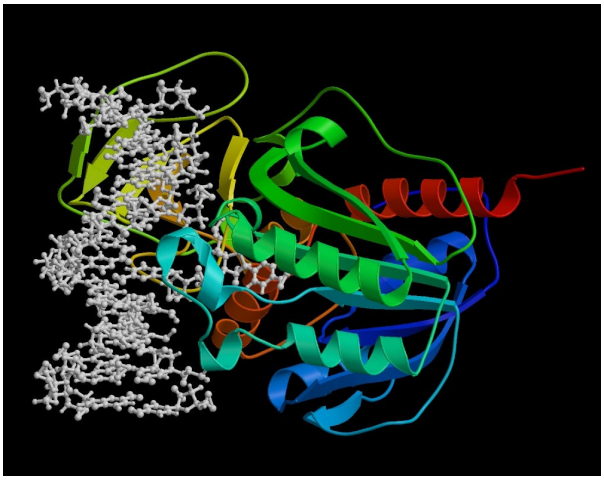
A significant commitment has also been made in the wide area. Currently the wide-area circuits include an OC-192 circuit (10 Gb/s) to the Internet2 aggregation point in Atlanta and an OC-48 circuit (2.4 Gb/s) to the DOE's ESnet, also connecting in Atlanta. Since the science program for NLCF will require moving files and data sets whose sizes easily reach 10's of terabytes today and which will soon be in the range of 100's of terabytes, it is clear that this connectivity is inadequate. Network improvements are being engineered to address this problem.

A fundamental issue is the lack of pre-existing infrastructure to allow significant enhancements to be rolled out. Thus, ORNL is nearing the end of a wide-area infrastructure construction project to provide fiber connectivity from ORNL to Chicago and from ORNL to Atlanta. This project involves procuring optical fiber from the Tennessee Valley Authority (TVA) and from Qwest. The TVA fiber stretches across the state of Tennessee and in particular connects the Oak Ridge DOE reservation with Nashville. The Qwest fiber connects Atlanta to Chicago through Nashville. Dense Wave Division Multiplexing (DWDM) gear is being installed on this fiber, giving the ability to carry up to 96 x 10 Gb/s circuits. Connections to Chicago and Atlanta are significant for their proximity to National Lambda Rail (NLR). NLR is the first user-owned research network being built in the country.[viii] ORNL is providing dedicated circuits to NLR between Atlanta and Chicago in return for dedicated circuits on NLR between Chicago and Sunnyvale, California.

When in place, the new DWDM infrastructure will improve connectivity on two fronts. By Summer, 2005, the Oak Ridge ESnet primary connection will be moved from Atlanta to Chicago and its bandwidth will be increased from OC-48 to OC-192. The current OC-48 circuit will be left in place as both a fallback and an additional source of connectivity for locations in the Southeast. At the same time, we expect to provide two OC-192 circuits from the NCCS to DOE's UltraScience Net. UltraScience Net is a research network project, led by ORNL, to develop circuit-switched (as opposed to packet-switched) techniques in support of next-generation data-transfer requirements. UltraScience Net's backbone consists of two OC-192s providing dedicated channels up to 20 Gb/s between its four main switching hubs: Sunnyvale, California; Seattle, Washington; Chicago, Illinois; and Atlanta, Georgia. The sites connected to these hubs include Stanford Linear Accelerator Center, Pacific Northwest National Laboratory, Fermi National Accelerator Laboratory, and ORNL. It is expected that Argonne National Laboratory and NERSC will also connect to the Chicago and Sunnyvale hubs, respectively, in the near future. The combination of these circuits makes the NLCF accessible by any researcher located on any of the main national research networks, namely ESnet, Internet2, the TeraGrid, and NLR.

2) Visualization

ORNL boasts a unique "tool" for production-scale unclassified scientific discovery – a high-resolution visualization facility known as EVEREST – for unlocking the secrets of DOE science applications. This "PowerWall" exploratory visualization facility includes a 35-megapixel display that fills the room with fine-grained details from scientific simulations and experiments. Driven by leading-edge research in scalable visualization technology, EVEREST is an open venue for interactive large-scale visualization and analysis by DOE Office of Science researchers. Using an integrated software environment and powerful clusters for analysis and rendering, scientists can seamlessly apply EVEREST for rendering on a variety of displays, including remote rendering and grid visualization services, for in-depth analysis of data too large for traditional office workstations.



III. Applications

A. Science Results

The Cray X1 evaluation at the NCCS targeted a range of applications of importance to the DOE Office of Science.[vii] The X1 has enabled breakthrough science in many of these

applications areas, including materials, fusion, atomic physics, climate dynamics, and astrophysics.

At CUG 2004, we described the unique capability that the X1 provides for simulating superconductivity using the dynamical cluster approximation with quantum Monte Carlo.[x] This capability has provided results that may be the first to accurately model cuprate high-temperature superconductors.[xi]

Also at CUG 2004, we described the optimization and resulting performance of GYRO, an application that simulates gyrokinetic processes in fusion plasmas.[xii] Candy provides an update at CUG 2005.[xiii] The unique capabilities of the X1 allowed simulation of fixed-size problems significantly faster, which has led to a string of scientific results.[xiv]

GYRO performance relies on the high bandwidth of the X1 interconnect, and it was unclear if GYRO would continue to perform well on X1E, where the same bandwidth is shared among twice as many processors. Early results are promising, however, showing a consistent 19-21% improvement per processor over a range of processor counts.

Again at CUG 2004, Pindzola, Colgan, *et al.* reported significant progress in computational atomic and molecular physics.[xv] Colgan recently reported a significant result of follow-on work, the first-ever *ab initio* calculation of the double photoionization of the hydrogen molecule.[xvi]

Some of the most visually spectacular science results on the NCCS X1 have come from the area of astrophysics. Blondin used VH-1, which models the hydrodynamics of core-collapse supernovae, for a three-dimensional simulation with 600 million zones.[xvii] The resulting time-dependent data were sent over the network to scientists at North Carolina State for visualization and analysis. The bandwidth requirements to transfer the data drove research and development in networking, which are described in at CUG 2005.[xviii]

Analysis of multiple complex visualizations led to the discovery of a new process through which neutron stars may "spin up" towards supernova ignition, Stationary Accretion Shock Instability (SASI). Examples of SASI visualizations are available at "http://astro.physics.ncsu.edu/TSI/".

B. *Current Applications*

The NCCS Cray X1 has been allocated to four major applications for the remainder of the fiscal year (through September 2005) with the goal of further scientific breakthroughs. These applications include supernova simulation, combustion simulation, computational chemistry, and design of high-energy accelerators.

The supernova simulations are continuations of the work described above to larger spatial dimensions and longer simulated times. Blondin *et al.* plan runs comparing different seeding mechanisms of SASI with and without rotation.[xix]

The combustion work of Sankaran *et al.* targets simulation of a stationary turbulent flame with detailed chemistry. The direct numerical simulation of this system will help establish and validate parameterizations used in indirect methods and may reveal the physical processes describing unexplained experimental results.[xx]

These combustion simulations will use the code S3D with 50 million grid points. Vectorization of S3D was recently completed, leaving inter-processor communication as the performance bottleneck. S3D performs regular nearest-neighbor communication, and the latency of MPI calls was significantly increasing runtime. By promoting some existing communication buffers to co-arrays and replacing a small number of MPI calls with direct copies, we were able to reduce runtime by 38%. Optimization continues, and large-scale production runs are scheduled to begin in June.

The computational chemistry work of Gan *et al.* targets computation of large-scale full-configuration interactions (FCI).[xxi] FCI gives the exact solution to the quantum many-body problem within a finite one-particle basis. Recent runs have performed the equivalent of solving an eigen problem with 65 billion coefficients. A fully vectorized parallel algorithm was developed for the work, yielding a performance of over 5.5 TF on 432 MSPs (60% of peak floating-point performance). Production runs are now underway.

The accelerator-design work of Ko *et al.* targets the low-loss accelerating cavity of the International Linear Collider (ILC).[xxii] They will use the code Omega3D to simulate harmful "wakefields" caused by higher-order modes (HOMs) of the ILC beam. After completing one design iteration at NERSC that used 200 HOMs, they now plan up to four additional design iterations on the NCCS Cray X1 at higher resolution to compare the two design variants under consideration. As of this writing, the initial port of Omega3D has just begun, but progress has been rapid.

C. *Emerging Applications*

In June, the DOE is expected to announce a call for proposals for large-scale computational experiments using the NLCF systems starting in October. In addition to the applications that have produced scientific results and are running now, a number of projects are preparing for the NLCF systems.

Climate modeling is one area in particular that has seen significant application development targeting the Cray X1. This development was described at CUG 2004[xxiii] [xxiv] and is updated at CUG 2005.[xxv] The primary application is the Community Climate System Model (CCSM), which couples separate components simulating the Earth's atmosphere, oceans, land, and sea ice.

CCSM has changed significantly over the past few years with new physical processes and higher-resolution

parameterizations, and the newest model was recently validated on the Cray X1 with moderate vector optimization.

Though many applications have shown exceptional performance on the Cray X1, and several of these applications have already produced breakthrough science, other applications of importance to the DOE Office of Science are better suited for other systems, such as the Cray XT3.

Some applications are limited by memory volume as well as by computation rate, and the extreme memory bandwidth of the X1 is not cost effective. One example is AORSA (All-Orders Spectral Algorithm), which is used to simulate radio-frequency heating and stabilization of fusion plasmas.[xxvi] We expect the XT3 to enable scaling of AORSA to very large processor counts through the capable interconnect, while providing cost-effective aggregate memory volume through significant commodity memory at each processor.

Some of the applications currently experiencing success on the X1 may move to the XT3 because of similar need of memory volume. The FCI application described earlier is now limited by X1 memory. Moving to the XT3 should allow the problem size to grow to where computation rate returns as the bottleneck.

Though VH-1, the astrophysics code also described earlier, is not currently limited by memory on the X1, follow-on applications are adding full simulation of neutrino fluxes in three spatial dimensions, resulting in a dramatic increase in memory. Though the computation and communication needs will continue to be extreme, the system size needed to carry out simulations will not be practical using X1E technology. The XT3 represents a better compromise for this domain, with a highly capable interconnect coupling cost-effective memory and processor technology.

The immediate introduction of the XT3 within the NLCF is also cost effective for some applications because of the current state of their software implementations. These implementations do not vectorize and would be costly and time consuming to refactor into vectorizable implementations.

We have made significant progress with some software that appeared very challenging, such as the PETSc library,[xxvii] but software development can span the lifetimes of multiple systems, and some software is not likely to be vectorized within the lifetime of the X1E. For such
software, the XT3 again provides a better compromise, with competitive *scalar* processors coupled through a highly capable interconnect.

Again, the applications targeted by the NLCF are those that require the highest capability, and the Cray X1E and XT3 provide complementary strategies for providing that capability. Some applications require the most powerful processors and highest bandwidth, while others require memory volume impractical in a vector system, and others would be costly to vectorize. The greatest capability for these latter applications comes from powerful scalar processors with the most capable interconnect.

IV. Synthetic Benchmarks

In the following section we analyze synthetic benchmarks to further understand the performance characteristics of the Cray X1 and XT3.

Application benchmarks tell us a great deal about how a system will run the problems that make up the benchmark test; however, it is very difficult to generalize their performance to predict the performance of other computational problems, even if the same codes are used and only the input is varied. By focusing on more general types of operations, synthetic benchmarks can tell us how classes of operations map to a particular architecture. By understanding this mapping, we can develop an understanding of how other applications dominated by those same types of operations might perform on the system.

The LINPACK benchmark is one such synthetic benchmark, though it suffers from being overly specific, in that it only tests the performance of dense linear algebra. A more recent effort by Dongarra, *et.al.* developed the High Performance Computing Challenge (HPCC) benchmarks[xxviii], which include not only the classic linear-algebra test, but several additional tests which examine computation performance, memory performance, and communication performance from several aspects.

A. Comparison of Results

The following table and charts summarize HPCC benchmark results obtained for baseline runs on the NCCS Cray XT3 and for comparison purposes the NCCS Cray X1 during early May, 2005. The XT3 software stack included version 1.0 of the operating system, the PGI 6.0.1 compiler, version 2.5 of the ACML math library, and MPT 1.0. The X1 software stack included Programming Environment (and “libsci”) 5.4.0.1, and MPT 2.4.0.3. For the purposes of comparison, HPCC used 64 processors (MSPs) on each system to insure that MPI communication across nodes dominated the MPI measurements and that, for global benchmarks, the problems were at similar scale. Figures 1-5 show the relative performance of the two systems compared on various segments of the benchmark. Table 2 lists the actual results obtained from the baseline runs according to the HPCC rules.

Table 2: A comparison of the Cray XT3 and Cray X1

	Jaguar Cray XT3	Phoenix Cray X1	
number of processors	64	64	
HPL	0.1915	0.6788	TFLOPS
PTRANS	13.0940	34.9082	GB/s
DGEMM	4.2967	12.3813	GFLOPS
*DGEMM	4.2992	12.1718	GFLOPS
STREAM_Add	5.1144	16.6808	GB/s
*STREAM_Add	4.7828	15.1133	GB/s
STREAM_Copy	4.9270	15.2149	GB/s
*STREAM_Copy	4.8315	14.4135	GB/s
STREAM_Scale	4.9702	13.4907	GB/s
*STREAM_Scale	4.8070	12.8722	GB/s
STREAM_Triad	4.8624	16.6929	GB/s
*STREAM_Triad	4.5991	15.1119	GB/s
FFT	0.7757	0.2400	GFLOPS
*FFT	0.7977	0.2401	GFLOPS
MPI-FFT	17.8817	5.0019	GFLOPS
RandomAccess	0.0198	0.0641	GUPS
*RandomAccess	0.0198	0.0641	GUPS
MPI-RandomAccess	0.0023	0.0031	GUPS
MaxPingPongBandwidth	1.1405	9.3177	GB/s
NaturallyOrderedRingBandwidth	0.5770	3.9612	GB/s
RandomlyOrderedRingBandwidth	0.3782	0.9424	GB/s
MinPingPongLatency	31.0689	7.9413	usec
NaturallyOrderedRingLatency	40.8888	15.0680	usec
RandomlyOrderedRingLatency	41.1280	15.0627	usec

Figure 1: Global Benchmark Comparison

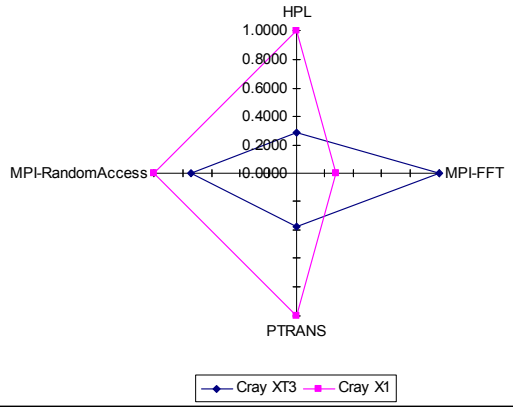


Figure 2: SP Benchmark Comparison

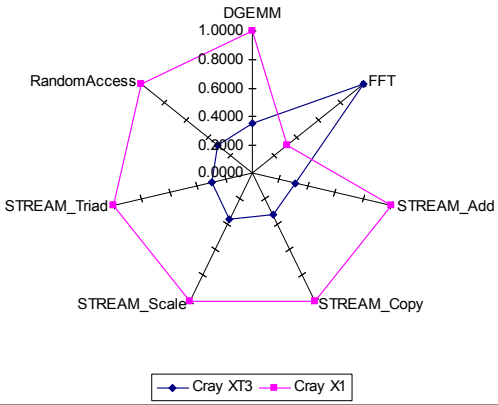


Figure 3: EP (*) Benchmark Comparision

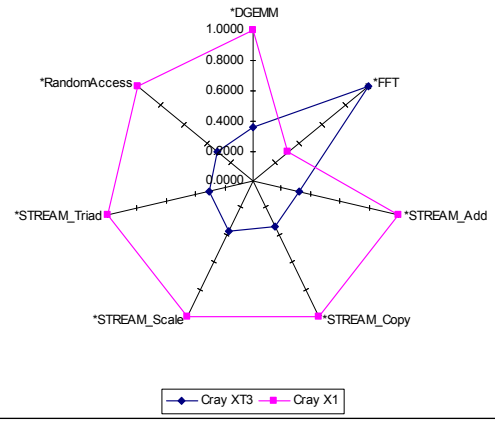
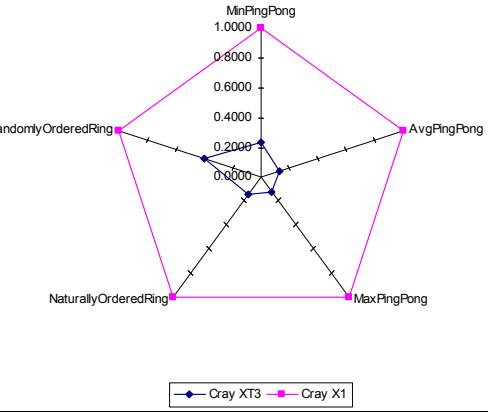
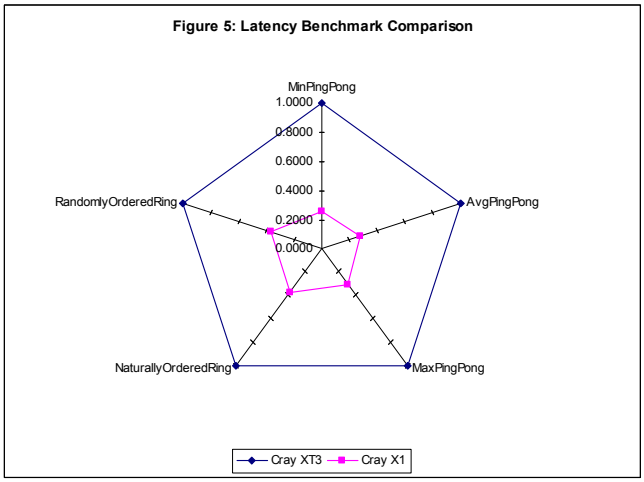


Figure 4: Bandwidth Benchmark Comparison





B. *Synthetic Benchmark Discussion*

1) *HPL and DGEMM*

High Performance LINPACK and Matrix Multiply. HPL is the classic "LINPACK" test, which examines raw computational horsepower of a system by measuring the time required to solve a system of linear equations. DGEMM is one of the highest-computational-intensity fundamental linear-algebra operations. The implementations of these benchmarks are well suited to either vector or scalar computer architectures and take advantage of vendor tuned BLAS libraries. They can be sensitive to the input data (problem size, block size, *etc*.); however, the heavy emphasis on libraries reduces the sensitivity to compiler flags.

While some applications make heavy use of dense linear algebra, many grid-based physical simulations have similar algorithmic constructs, of similar structure: multiply nested loops which sweep through large arrays applying a fixed set of operations. The HPL and DGEMM benchmarks are significant for these kinds of codes. The XT3's HPL score of 191.5 GF on 64 processors represents 62% of the theoretical peak of 4.8 GF per 2.4-GHz Opteron processor. Contrast this with the X1 HPL score of 678.8 GF, which represents 83% of the 12.8 GF per MSP theoretical peak on that machine. Likewise, on the DGEMM benchmark, XT3 reached 90% of peak, whereas X1 reached 98% of peak. Two factors contribute to the X1's performance here: (1) the high-bandwidth memory subsystem, and (2) the use of X1's vector processing feature to **hide memory latency**. The *DGEMM (embarrassingly parallel DGEMM) demonstrates an advantage for single processor per node designs like the XT3: where the X1 efficiency drops to 95% as 63 processors are added to the pool of workers (4 processors per node), the XT3's efficiency remains at 90%.

2) *PTRANS*

Parallel Transpose. This benchmark measures the ability of a parallel and distributed system to move data around in memory, using a combination of memory operations and communication over the interconnect. Like HPL, the results can vary greatly with the problem size and block size selected. This benchmark is also moderately sensitive to compiler flags.

Transpose operations were once of limited necessity because of large flat shared memories which supported strided access patterns at full speed. However, current machines typically suffer large performance impact from such access patterns, and, when the solution method dictates accessing multi-dimensional (and potentially distributed) arrays along different indices (as is the case for atmospheric models computing spectral transforms), it can be more economical to transpose the arrays into a more advantageous memory layout before beginning the operation. For these kinds of applications, the PTRANS results are significant. The X1's reputation as a "bandwidth" machine again makes its presence felt in this benchmark with a whopping 34.9 GB/s on 64 MSPs versus 13.1 GB/s for 64 nodes of the XT3.

While this may seem like a hands-down win for the X1, these numbers should not be considered a comparison of apples to apples. Because the X1 processors can perform strided accesses efficiently and can access "off-node" memory, transpose operations become less significant as a metric for this machine. Moreover, the results of this benchmark should be considered in comparison to the floating-point speed they are intended to optimize. If we consider the ratio of PTRANS performance to HPL performance, we see the X1's balance at 51.4 GB/TF whereas the XT3's is 68.4GB/TF. Thus, relative to the processor speed, XT3 has a higher transposition bandwidth.

3) *STREAM*

The STREAM benchmarks measure the performance of the memory and data cache on operations that access memory with a constant, unit stride. The operations tested are Copy, which copies one vector (one-dimensional array) to another, Add, which adds two vectors, Scale, which multiplies a vector by a scalar constant, and Triad, which multiplies a vector by a scalar constant and adds the result to another vector. The memory system throughput for each of these is for both a single processor and for all of the processors performing in embarrassingly parallel mode. The problem size selected controls the layout of the vectors in memory, which can dramatically affect performance. This benchmark can also be extremely sensitive to compiler flags.

Unlike PTRANS, this benchmark focuses strictly on the memory subsystem's performance and emphasizes cache performance for stride-one access. While few programs perform stride-one sweeps through large one-dimensional arrays, this benchmark is representative of programs with a high degree of locality, finite-difference algorithms being a

classic example. The XT3 delivered 4.8-5.1 GB/s in single processor mode for a memory subsystem rated at 6.4 GB/s; this represents 75-80% of the peak memory performance. The X1 delivered 13.5-16.7 GB/s in single processor mode on a memory subsystem with a theoretical peak of 34.1 GB/s per CPU, an efficiency of 40-49%.

As with PTRANS, these results should also be considered relative to the processors' performance on a single processor "FLOP" benchmark such as DGEMM. The XT3 delivered 1.12-1.19 STREAM GB/ DGEMM GF and the X1 delivered 1.09-1.35 STREAM GB/DGEMM GF. So why does the X1 sustain a low percentage of its peak memory performance? Several possible answers arise, including problems with the compiler or the benchmark. But given that DGEMM is running very near theoretical peak with a STREAM/DGEMM ratio similar to XT3, and that the *STREAM results only degrade by 5-10%, it is clear that the X1 still has memory bandwidth above and beyond the HPCC STREAM results. (It should also be noted that results from the standalone version of STREAM returns somewhat better results; however, these are not directly comparable to the STREAM code in HPCC because it includes compiler directive additions disallowed by the rules for an HPCC baseline run.) It remains to be seen whether or not the X1 memory system will be fast enough to support the X1E processor upgrade, though early results are promising.[xxix]

4) FFT

Fast Fourier Transform. The FFT benchmark measures floating-point performance of a one dimensional (1D) FFT. This benchmark is run in three separate modes: a single processor performing the computation alone (FFT), all of the processors performing the same computation in embarrassingly parallel mode (FFT*), and all of the processors collaboratively computing a single FFT using MPI (MPIFFT). This benchmark displays moderate sensitivity to compiler flags.

The 1D-FFT is an example of an algorithm which structurally does not lend itself to efficient use of the vector processing features of the X1. The 1D-FFT is heavily used in signal processing. This benchmark should be considered representative of other compute-intensive algorithms which will not be able to take advantage of vector features for raw processor performance and memory latency hiding. For these types of applications, the XT3 demonstrates a significant advantage: 0.78-0.80 GF versus 0.24 GF for the X1. The XT3 is ~3x faster on this benchmark, roughly the ratio of Opteron's peak performance to the X1's scalar peak performance.

5) RandomAccess

The RandomAccess benchmark is similar to STREAM in that it measures memory bandwidth, but unlike STREAM, which essentially measures the unit stride performance of the data cache, RandomAccess measures the lower limit of performance on non-unit stride access to memory. It does this by creating a large table of data and randomly updating individual data elements, thus the performance is reported in billions of updates per second (GUPS). Examples of application domains which depend on such memory access performance include cryptography and emerging algorithms for a new generation of computational chemistry and biology. Performance on this benchmark is of interest to many random walk algorithms.

Like the FFT benchmark, the RandomAccess benchmark is performed in three modes: single processor, embarrassingly parallel, and global in which all of the processors cooperate to update a single table using MPI. For this benchmark to actually measure the memory performance, rather than the cache performance, the problem size must be defined to be much larger than the cache; if the selected problem size is too small, the results will be artificially high. This benchmark displays low sensitivity to compiler flags.

This is a case where the X1 has a strong advantage, delivering 0.0641 GUPS versus XT3's 0.0198 GUPS – more than 3x in X1's favor. However, one must discount a comparison of the MPI version of RandomAccess on the X1 since other approaches to this problem (such as those using UPC) would be far more appropriate to the X1 architecture. The number stands as a baseline number only according to the rules of the HPCC benchmarks; optimized results would be significantly improved.

6) Communication

The communication benchmarks measure time required for a message to be sent to and returned from another processor (a.k.a. "ping-pong"). These benchmarks are performed in several different fashions. When the test is performed on very small messages, the result is reported as communication latency (in microseconds). When the test is performed on very large messages, the message size is divided by the message transit time and is reported as bandwidth (in GB/s). The tests are performed in an all-pairs mode in which each processor exchanges a message with multiple partners, in natural ring mode in which each processor exchanges messages with the processors whose MPI rank immediate precedes or succeeds it, and a randomly ordered ring in which the processors are randomly order into a ring but only exchange messages with their logically nearest neighbor within the ring. This benchmark is a direct reflection of the interconnect performance and the communications libraries. It is difficult to draw any conclusion from the current results of the HPCC communication benchmarks on either of these machines, since, as may be noted from the latency measurements, the MPI implementations on both machines are not yet mature. Thankfully, we note that both are also under active development. The HPCC tests will need to be re-run as the software stack matures on both architectures. Improvements in the MPI implementations will also impact the results of HPL, PTRANS, MPI-FFT, and MPI-RandomAccess. Applications which are highly sensitive to MPI performance (bandwidth,

latency, or both) should expect to see improvements with future versions of the software stacks on both machines.

V. Roadmap to 100's sustained TF

By 2007, there will be a hybrid computer architecture code-named "Rainier", containing both vector and scalar processors. The hardware interconnect and system software are being designed using new technology based on the Cray XD1 system. The Cray Rainier design combines the strengths of the X1 and XT3 in one flexible integrated architecture. It incorporates the next generation of Cray vector processors (code-named BlackWidow), scalar processors, and scalable I/O. This system is expected to be a significant advance in bandwidth, scalability, and price/performance. NLCF officials believe it is possible to deliver 250 TF in 2007 and reach a balanced PetaFLOP/s by 2008.

VI. Conclusion

The Leadership Computing Facility at ORNL will deliver at least 100 times greater computational resources to key problems than what is generally available for advanced scientific and engineering simulations. The NLCF will do this by fielding complementary computing resources that are tailored to the science. The resources in the NLCF will focus on only a handful of grand challenges at a time allowing scientific discoveries that would be otherwise intractable.

Acknowledgment

This research is sponsored by the Mathematical, Information, and Computational Sciences Division; Office of Advanced Scientific Computing Research; U.S. Department of Energy. The work was performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. De-AC05-00OR22725

References

[i] High-End Computing Revitalization Task Force. *Federal Plan for High-End Computing*. May 10, 2004. http://www.hpcc.gov/pubs/2004_hecrtf/20040702_hecrtf.pdf

[ii] DOE Office of Science. *Facilities for the Future of Science: A Twenty-Year Outlook.* November 10, 2003. http://www.er.doe.gov/Sub/Facilities_for_future/20-Year-Outlook-screen.pdf

[iii] Dave Kiefer. "Cray Product Roadmap: Hardware and Software," *CUG 2004*.

[iv] http://www.nersc.gov/nusers/accounts/allocations/awards/alloc2005.php

[v] http://www.top500.org/

[vi] http://nccs.gov

[vii] A copy of the full review report is available from: http://www.csm.ornl.gov/DOE/Feb2004/X1ReviewReport.final.doc

A copy of the CCS Evaluation Report of the Cray X1 is available from: http://www.ornl.gov/~webworks/cppr/y2001/rpt/119526.pdf

[viii] http://www.nlr.net/

[x] T Maier, JB White III, and T Schuthess. "Towards Full Simulation of High-Temperature Superconductors," *CUG 2004.*

[xi] T Maier. "Does the 2D Hubbard model describe high-temperature superconductors?" *APS March Meeting*, 2005.

[xii] M Fahey and J Candy. "GYRO– Analyzing New Physics in Record Time on the Cray X1," *CUG 2004*.

[xiii] J Candy and M Fahey. "GYRO Performance on a Variety of MPP Systems," *CUG 2005*.

[xiv] C Estrada-Mila, J Candy, and RE Waltz. *Phys. Plasmas* 12, 022305 (2005).

[xv] M Pindzola, JP Colgan, *et al.* "Computational Atomic and Molecular Physics," *CUG 2004*.

[xvi] J Colgan, MS Pindzola, and F Robicheaux. "Time-dependent close-coupling calculations for the double photoionization of He and H2," *J. Phys. B: At. Mol. Opt. Phys.* 37 (2004) L377–L384.

[xvii] "Neutron Star Spin-Up Discovered with 3D Simulations on Cray X1." http://astro.physics.ncsu.edu/TSI/

[xviii] S Carter. "High-Speed Networking with Cray Supercomputers at ORNL," *CUG 2005.*

[xix] J Blondin and A Mezzacappa. "Three-Dimensional Studies of the Newly Discovered Stationary Accretion Shock Instability in Core Collapse Supernovae and Its Ramification for the Supernova Mechanism and Observables," DOE proposal.

[xx] R Sankaran, ER Hawkes, and JH Chen. "Stationary Direct Numerical Simulation (DNS) of Turbulent Premixed Combustion in the Thin Reaction Zones Regime," DOE proposal.

[xxi] RJ Harrison, Z Gan, M Gordon, and K Ruedenberg. "Benchmark many-body molecular electronic structure calculations for open-shell and excited state systems," DOE proposal.

[xxii] K Ko. "Computational Design of the Low-Loss Accelerating Cavity for the International Linear Collider," DOE proposal.

[xxiii] J Drake, P Worley, M Cordery, and I Carpenter. "Experience with the Full CCSM," *CUG 2004.*

[xxiv] F Hoffman, M Vertenstein, and JB White III. "Adventures in Vectorizing the Community Land Model," *CUG 2004.*

[xxv] G Carr, Jr., M Cordery, J Drake, M Ham, F Hoffman, and P Worley. "Porting and Performance of the Community Climate System Model (CCSM3) on the Cray X1," *CUG 2005.*

[xxvi] EF Jaeger, LA Berry, DB Batchelor, and MD Carter. "All-orders spectral calculation of radio-frequency heating in two-dimensional toroidal plasmas," *Phys. Plasmas* 8 (2001) 1573-1593.

[xxvii] R Mills, M Fahey, and E D'Azevedo. "Optimization of the PETSc Toolkit and Application Codes on the Cray X1," *CUG 2005.*

[xxviii] J Dongarra and P Luszczek. *Introduction to the HPCChallenge Benchmark Suite*. Computer Science Department Tech Report 2005, UT-CS-05-544.
[xxix] J Levesque. "Comparisons of the X1 to X1E on Major Scientific Applications," *CUG 2005.*