Total Life Cycle Cost Comparison: Cray X1 and Pentium 4 Cluster

Paul Muzio *and* **Richard Walsh**,
Army HPC Research Center/Network Computing Services, Inc.
1200 Washington Ave. So.
Minneapolis, MN
April 22, 2003

ABSTRACT: *Network Computing Services, Inc. (NCSI), as the support infrastructure contractor for the Army High Performance Computing Research Center[1] (AHPCRC), acquires, integrates, and operates in its facility, high performance computers for the US Army. Early in 2002, as part of its support infrastructure activities process, NCSI completed an analysis of the high performance computing (HPC) technologies likely to be available in early 2003 to satisfy the growing computational requirements of the AHPCRC and Department of Defense science and technology community. The focus of the analysis was on HPC systems capable of providing a superior capability production-computing environment. Here, we present the part of that analysis comparing the component capabilities (processor, memory, interconnect, etc.), software, and cost of ownership of the Cray X1 system to a representative HPC cluster built using Intel's Pentium 4 processor.*

1. Introduction

NCSI and the AHPCRC Mission

The Army High Performance Computing Research Center (AHPCRC) is a collaborative effort between the United States Army, its university partners, and Network Computing Services, Inc. (NCSI). The university partners include the University of Minnesota, Clark Atlanta University, Florida A&M University, Howard University, Jackson State University and the University of North Dakota**.**

NCSI provides the computational infrastructure and support for AHPCRC's research program. As part of its role within the AHPCRC, NCSI installs, evaluates, and operates HPC resources. Additionally, NCSI provides highly trained staff scientists and research support specialists to assist the Army and its university partners.

Current AHPCRC Resources

In Spring of 2002, NCSI undertook a study of HPC systems that could meet its next-generation performance requirements and would be available in the early 2003 time frame. This paper is based on the material compiled in that effort.

At the time of the study, the AHPCRC's primary computational resource was a CRAY T3E-1200. The AHPCRC's CRAY T3E-1200 system is a parallel, distributed memory scalable system that operates under a single system image (SSI). It has 1088 processors with a peak performance of 1,300 billion floating-point operations per second (Gflops), and 544 gigabytes of memory. The T3E-1200 interconnect network is a three dimensional

[1] "The research reported in this presentation was performed in connection with contract DAAD19-03-D-0001 with the U.S. Army Research Laboratory. The views and conclusions contained in this presentation are those of the authors and should not be interpreted as presenting the official policies or positions, either expressed or implied, of the U.S. Army Research Laboratory or the U.S. Government unless so designated by other authorized documents. Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon."

torus. The T3E-1200 provides *sustained*, floating-point processing speeds of 100 to 200 Mflops per processor (100-200 Gflops total), interconnect bandwidths of 200-300 Mbytes per second, and interconnect latencies as low as 1 to 2 usecs.

Moreover, the operating system is a mature, SSI system with features such as checkpoint restart, and parallel IO, and gang scheduling of parallel work. It is also a resource that is highly utilized, averaging well above 90% over the last several years. Its characteristics were used, by NCSI, as the baseline standard for the study, from both a performance and operational efficiency perspective, in its comparison of next-generation product performance, useability, and utilization.

AHPCRC Next Generation System Requirements

To determine the target capability requirements for the next-generation system, NCSI staff and AHPCRC researchers working on problems in vehicle survivability, interior ballistics, weather forecasting, and dispersion of airborne contaminants were asked to forecast their computational requirements for the next 5-years. The emphasis of the projection was to determine what "capabilities" were required as opposed to what "capacity" was needed. While assessments were made of existing computational kernels and algorithms, emphasis was based on future requirements as opposed to benchmarks of existing applications as benchmarks tend to be backwards looking, not forward looking. Researchers were asked to define the type of problems they would need to solve 5-years from now that would be at the forefront of defense computational science, i.e., the focus was to be on capability computing, not capacity computing.

The researchers established that a target system capable of sustaining 650 Gflops (5 times the power of the CRAY T3E-1200) over a period of 2 to 3 days on a single application was required to enable them to significantly extend the state of their computational research in survivability, ballistics, dispersion of airborne contaminants, and atmospheric science.

Additional hardware requirements included, memory at least equal in size to the CRAY T3E-1200 (preferably doubled) with 2 to 3 times its bandwidth per processor, an interconnect with at least three times the bandwidth and similarly low latencies, and local storage of at least 30 Tbytes with high-speed parallel access and aggregate bandwidth of 450 Mbytes per second for check-pointing large jobs in less than 30 minutes. 100 Tbytes of near-line tape storage would also be required.

Further, important technical requirements included scalability as the system would most likely be purchased and added to in stages over a period of several years perhaps doubling or quadrupling its initial size. The entire hardware package would have to be integrated into a production system with software that would at least approximate the stable and feature-rich environment of the CRAY T3E-1200's Unicos/mk, SSI operating system.

Preference was for candidate systems that would benefit from an SSI operating systems which promoted:

- Ease of use
  - a single IP address to login into,
  - a unified view of process space,
  - automated global scheduling of work
- Ease of administration
  - a single operating system to update,
  - one global file system to manage,
  - a single source tree to update,
  - single system shutdown and reboot

No facilities limitations were placed on the system configuration, but the cost of facilities (UPS, cooling systems, floor space, diesel generators) and utilities above an arbitrary fixed level, identical for both systems, were included in the cost analysis.

2. Design and Cost Comparison Methodology

Systems Considered

The HPC market is perhaps at a point of maximum variety with systems at the TOP 500 supercomputers website from several genera including distributed memory commodity clusters ("Beowulf" systems); large, custom, distributed-memory parallel processors (CRAY T3E-1200); highly parallel, cc-NUMA, common memory RISC processors (SGI and IBM systems based on MIPS and POWER 4 technology); and older-style, common-memory, moderately parallel, vector processors (Cray SV1, NEC SX5). As such, current and anticipated systems in each of these design categories were considered in light of the AHPCRC technical requirements and with respect total life cycle cost.

Benchmarking all systems was impossible because of time constraints and because the work was in part a forecast of technology likely to be available one year hence, in Q2 of 2003 (several major, "in-flight" adjustments in the analysis were made based on vendor announcements while this work was in progress). All candidate systems were simply not available to test.

As an alternative, AHPCRC applications kernels were inventoried, and their inter-processor communications patterns were reviewed with an eye to finding substitute measures already available or amenable to simulation. The AHPCRC codes exhibited the following general features:

- kernels were typically long vector,
- kernels were largely memory bandwidth limited (lower flops/mops ratio),
- most were MPI message passing codes,
- messages were medium to large in size,

- communication was typically bandwidth limited except that certain key applications required support for low latency communications,
- memory use could be limited to 4 Gbytes per processor although a 64-bit address space was a preference.

Next we evaluated system performance at the micro-architecture level of the processor. The SpecFP2000 and, in particular, the Stream Triad benchmarks scores were obtained or simulated as a substitute for benchmarking every system with AHPCRC codes (some benchmarks were run). In addition, we evaluated processor performance at the design level/ Sufficiently distinct systems based on processors in the top ten places of these standard benchmarks were then considered in detail (only Intel was considered in the IA32 category, for instance). These included the Itanium 2, Pentium 4, Alpha EV6 and EV7 clusters, the IBM Power 4 based 1600 and HP SuperDome parallel RISC systems, and the single stream processor of the Cray X1 parallel, hierarchical vector processor, among others.

Based on our technical analysis of the processors and interconnect performance, target system configurations were established and compared against requirements. Systems unable to meet the standards were dropped from further consideration. Costs were compiled in three parts:

- Acquisition costs
- Basic site upgrade and installation costs
- Five-year operating costs

Cost data was obtained or estimated from a variety of sources including published purchase prices for similar large systems and data available on component-vendor websites in the case of HPC cluster systems (Myricom, for instance). Dollar per sustained Mflops numbers were computed for each system as the component costs were added. System processor counts were also scaled based on estimates of average utilization over the five-year period.

In this document, we summarize only our analysis for the Cray X1, parallel hierarchical vector system and a Pentium 4 cluster system. This allows us to make a comparison between custom engineered HPC systems and commodity assembled HPC systems (so called "Beowulf" clusters). The Pentium 4 processor also offered good bandwidth to memory, and its raw price-performance was the best in the cluster class.

3. Cray X1 and Pentium 4 Architectures

Floating Point Performance

We compared the sustained floating-point performance of the Cray X1 and Pentium systems sized to meet the AHPCRC's requirements. This comparison showed dramatic differences between Intel's IA32, CISC microprocessor and Cray's multi-streaming, vector processor (MSP). To find a common basis for comparison, Cray's MSP was divided into its four, component, single-streaming processors (SSPs) which more closely resemble the standard notion of a processing core as a single, register-functional-unit pairing. Table 1 compares Cray X1 SSP features to Pentium 4 CPU features.

Performance Feature	Cray X1	Pentium 4
Clock Speed	800 MHz	2800 MHz
Sustained Mflops/CPU	780	200
Percent (%) of peak	24%	3.4%
Peak Mflops/CPU	3200	5600
CPUs needed for 650 Gflops Target	896	3456

Table 1: Floating-point performance

Clock speed and peak performance favor the Pentium 4 CPU dramatically, yet in a measure more reflective of application performance, the Stream Triad benchmark, the Cray X1 SSP delivers nearly four times the Mflops for this memory-access intensive kernel. Performance on the stream triad was used in Table 1 to set the number of processors of each type needed to deliver the 650 sustained Gflops of floating-point performance the AHPCRC requires. The Cray X1 system sized for AHPCRC requirements was 896 processors, while the Pentium 4 cluster requires 3456.

Other points of interest in the processor comparison include the presence of cache on both systems, but with the note that cache-based improvements will be much greater on the Pentium 4 because of its fast clock and lower latency cache. Both systems are design to run 32-bit operations at 2x speed. The true vector instructions and memory-to-memory pipeline of the Cray X1 were estimated to deliver 24% of it peak performance, while the Pentium 4's small 128-bit, SSE2-vectors and pre-fetching (viewed here as a means of using hardware bandwidth more effectively or emulating vector loads) return only 3.4% of its almost 2x greater peak performance number.

Each system can be configured to meet the floating-point performance targets as describe above, but the Cray X1 does so in a much more efficient package with about one quarter of the number of processors while providing built-in 16-way SMP capability on its node boards (Pentium 4 offers only 2-way and 4-way SMP with reduced bandwidth to memory).

It should be pointed out that recent benchmarks of AHPCRC applications on a liquid cooled, Cray X1 system validate the assumption that the stream triad was a reasonable model for AHPCRC applications performance. The primary CFD application currently runs at 31% of peak (~1 Gflops), a CSM code at 22%, and the MM5 weather code at 18%. The average is close to the estimated 24% of peak across multiple SSPs. These numbers are expected to improve as the codes are tuned for the system (see Appendix A).

Similarly, the AHPCRC CFD application runs at about 4.3% of peak or at about 240 Mflops per processor on a single 2.8 GHz Pentium 4 system.

Memory Design and Performance

The amount of memory addressable by an MPI process is a key feature for AHPCRC's message passing codes. Here, the X1 can use the maximum available on a 16-SSP node-board, while the Pentium 4's 32-bit address space limits the number to 4 Mbytes less space for the Linux kernel. Table 2 compares memory systems. Memory per processor is limited by the 32-bit address space of the Pentium 4 (this is really a per process limit) and by the density of RDRAM memory and the number of board slots on the Cray X1. Because the Pentium 4 cluster will need more processors, it allows for more total memory. The Cray X1's highly banked, RDRAM memory (128 banks per node-board) gives it excellent memory bandwidth and machine balance (peak Mflops divided by sustained Mops, where a smaller number is better).

Performance feature	Cray X1	Pentium 4
Max. Gbytes of physical memory/CPU (SSP)	4	4
Gbytes of addressable memory/MPI process	16 to 64	~3.5
Max. total Gbytes per 650+ Gflops system	3,584	13,824
Peak memory read bandwidth Mbytes/sec (two-thirds of total)	6400	2845
Stream triad Mbytes/sec/CPU (read+write)	9350	2250
Processor balance	2.74	19.9

Table 2: Memory comparison

Another notable point of difference is the fact that the Cray X1's memory is globally addressable such that a single vector instruction issued from any processor on the system can load to its registers data in any memory location. This capability is used in the implicit distributed memory, parallel programming models Unified Parallel C (UPC) and Co-Array Fortran (CAF) on the X1 for "message passing" by direct assignment. Pentium 4 cluster processors can only directly address memory on their own motherboards (one-sided communication in the MPI-2 standard when supported by the interconnect hardware can viewed as a weak approximation to the Cray X1 capability).

Both systems meet the AHPCRC's basic memory size and performance targets although the Cray X1 provides substantially more bandwidth, better balance, and more addressable memory per processor.

Interconnection Network Design and Performance

The processors in the two systems must be interconnected to provide the sustained, floating-point capability required by the AHPCRC. The Cray X1 has a custom, node-board crossbar combined into a router-switched ("bristled") 2D-hypercube. This network's bandwidth is hierarchical with reductions from a peak of 38 Gbytes/sec between processors on the same node-board to 1.6 Gbytes/sec between maximally remote nodes on a very large system. The Pentium 4 considered at the time was configured with a Myrinet, Clos-style interconnect. At this scale, 3456 processors, 5 or 6 hops would be required to move data between remote nodes.

Performance feature	Cray X1	Pentium 4
Interconnect type	X-bar/switched 2D-hypercube	Myrinet Clos/x-bar
MPI ping-pong bandwidth (Mbytes/sec, 32K message)	~750*2 (two-way)	~200*2 (two-way)
MPI ping-pong latency (1 byte, local/remote)	~7.5/15 usecs	~7/10 usecs
Scalability (processors)	16,384	8,192

Table 3: Interconnect comparison

Table 3 shows bandwidth and latencies from runs of the Pallas MPI Benchmark Suite (PMB) and Myrinet performance data from the Myrinet website. The Cray X1 met the interconnect bandwidth requirement easily while the Pentium 4 was close to meeting it. Recent Myrinet product introductions (as well as those from Quadrics and SCI) now meet the AHPCRC minimum bandwidth requirements, but still do not match the Cray X1 numbers.

The MPI latencies for both systems are not far off of the CRAY T3E-1200 numbers and meet the requirements. Latencies are expected to improve in the next several months from both interconnects. The Cray X1 can already provide lower numbers inside its CAF and UPC programming models. Cray indicates that it expects to reduce barrier time in MPI to 2 usec (between MSPs on a single node) to 6 usec (between MSPs on different nodes). Myrinet's faster Lanai card and new MX protocol are expected to provide "zero-length" message latencies in the 4 usec range. These estimates would have to be verified for both these large-scale configurations.

A scalability comparison gives an advantage to the Cray X1 although neither of these maximally sized systems has been built. It should be noted that at 896 processors the Cray X1 meets the sustained, floating-point requirement with much more headroom and fewer interconnect hops than the Pentium 4 cluster at 3456 nodes.

IO Subsystem Design and Performance

The candidate system's IO capabilities must match or exceed those of the CRAY T3E-1200 system currently in use. This implies sufficient bandwidth to checkpoint very large jobs in reasonable times (250-500 Gbytes in 20-30 minutes) to the file server and a complementary parallel file system and parallel IO routines.

The Cray X1 provides a custom designed IO subsystem with 4 x 1.2 Gbytes/sec port channels per node board accessing a configurable number of IO channel adapters which reach RAID disk arrays through PCI-X Fibre Channel adapter cards. This disk space is globally accessible from any processor. Total bandwidth is a function of the number nodes, IOCAs, FC-HBAs, and Raid controllers. The controllers are rated at 200-300 Mbytes per second individually. Their quantity, and how they are striped, define maximum bandwidth to the file server. We use 600+ Mbytes/sec as a reasonable data rate for striped IO on large files.

IO systems on COTS clusters are the typically far less high-performance. For larger cluster systems that demand better-than-COTS performance, custom solutions are created at custom prices. For this comparison, we assumed a largely commodity design of locally controlled raid to SCSI disk and a Gigabit Ethernet switch-based remote file server. Table 4 below summarizes the performance attributes of the two IO subsystems.

Performance Feature	**Cray X1**	**Pentium 4**
File server design	Custom PCI-X/FC-AL/Raid 5	Gigabit switched uplink and local Raid 5
Large file bandwidth (local and remote disks) (aggregate remote)	600+ Mbytes/sec NA ~3000 Mbytes/sec	200 Mbytes/sec 100 Mbytes/sec ~800 Mbytes/sec
Maximum file size	File server size limited 100+ TB	Block offset limited 2-4TB

Table 4. IO subsystem comparison

The Cray X1 IO subsystem (IOS), parallel IO libraries, and full 64-bit size-limited files easily meet the AHPCRC's capability requirements. There is no distinction between local and remote storage on the X1 as all IO is done via the channel ports on the node boards. 600 Mbytes/sec of bandwidth should be achievable for large-file reads and writes from an IOS with 3 IOCAs fully populated with FC-HBAs. These would give an aggregate theoretical bandwidth of 3 x 2 x 2 x ~250 Mbytes/sec or 3 Gbytes/sec.

The Pentium 4 cluster's IO design is two-tiered with a local RAID component on each node running at 200 Mbytes/sec and a remote component supported by several, large, line-speed, Gigabit Ethernet switches running at 100 Mbytes/sec per node. The local pieces could be integrated to some extent using PVFS or similar cluster products via either of the interconnects (Myrinet and Gigabit). Multple uplinks (at least 8) to the remote file server and striping would be used to give the required aggregate bandwidth to the remote file server.

More expensive SAN solutions for the cluster could have been chosen, but this would have had a significant effect on the per node cost of the cluster. The Cray X1 provides a custom SAN as described as part of its purchase price.

4. Cray X1 and Pentium 4 System Utilization

The 650, sustained Gflops AHPCRC performance requirement is implicitly accompanied by availability, utilization, and uptime requirements. To complete the large simulations defining AHPCRC, next-generation system requirements, a 650 Gflops sustained rate will have to be maintained for 2 to 3 days for job completion, or else the job will have to be protected by full check-point restart capability.

As system availability, utilization, and uptime decay, job failures rise and so does the true cost of system ownership. We define utilization as a percentage of the time the system is up and doing useful work, but not idle, out of the total number of theoretical processing hours available. System availability adds any idle time to the utilization, while current uptime or average uptime is the length of time between outright system failures.

Likely utilization differences between the two systems considered here should be corrected for as much as possible to obtain the required next-generation results. Cray systems have a track record of very high utilization at the AHPCRC and elsewhere. The AHPCRC's CRAY T3E-1200 with 1088 processors as mentioned above has had utilization rates that routinely exceeded 95% over the last 2 years (see Appendix A). Similar rates are expected for the Cray X1 when it is fully accepted, installed, and reaches operational steady state. As a new machine, we will be conservative and estimate it will be on average utilized 90% of the time over the five-year time frame considered here.

This figure is further supported by the very low frequency of hardware errors experienced on Cray systems, the integrated SSI operating system, the dynamic queuing and scheduling of work, its job migration and compaction capability, pre-emptive scheduling, the presence of check-point restart on the Cray X1, short shutdown and boot time, and its smaller scale.

Utilization figures for very large clusters are hard to come by as so few have yet reached the scale of the Pentium 4 cluster considered here. Those that come close to this size report mediocre utilization rates when run as typical, multi-operating system image (MSI) environments. Utilization is improved when such large clusters are run with SSI-emulating operating systems such as those from Scyld. Utilization can be expected to be substantially lower for clusters than for the Cray X1. We conservatively estimate that a cluster of the scale required to meet requirements will have an average utilization of two-thirds of the Cray X1's 90% or 60% over the five-year period considered here.

This figure is supported by the higher frequency of hardware errors expected on a cluster of commodity components of this scale, the higher rate of software failures expected with 1000s of individual operating systems to maintain and run, the absence of check-point restart capability, the relatively primitive scheduling capabilities available for clusters (no job migration/compaction, dynamic gang scheduling, job swapping, and pre-emption, etc), and longer draining, shutdown, and boot times. This is a difficult number to estimate, and one likely to improve on a per year basis over this five-year time frame, but it is set conservatively at 60% in this analysis.

Accordingly, the processor counts of our study systems are scaled and round up to multiples of 128 before being fully priced. This brings the Cray X1 system up to 1024 SSP processors and the Pentium 4 cluster up to 5760 processors. The effect of rounding up to multiples of 128 actually gives the Cray X1 and additional 15% of compute capability, 5% more than the utilization estimate requires.

We present a summary of the architecture component comparisons relative to the AHPCRC target requirements at this utilization-adjusted scale in Table 5 below.

Component Performance	Cray X1	P4 Cluster
Number of processors	1024 SSPs	5760 CPUs
Total Memory (Gbytes)	1,024 Mbytes	1,440 Mbytes
Memory Bandwidth (Gbytes/sec/CPU)	9.35	2.25
Bandwidth (two-way, ping-pong, PMB performance 32K bytes Mbytes/sec)	~2x1500	~2x220
MPI Latency for small messages (usecs) (local/remote)	7.5/15	7/10

Table 5. Overall characteristics of systems designed to deliver 650 Gflops sustained.

5. Cray X1 and Pentium 4 Software

A complete comparison of the differences in both user and system software on these two systems could easily run for many pages. For our purposes it is sufficient to consider highlights that most directly affect the use and support of the system for the execution of the AHPCRC's very large, next-generation capability computing workload.

Many of the important differences spring from the fact that, like the CRAY T3E-1200, the Cray X1 has an SSI operating system, Unicos/mp, while current clusters are almost exclusively multi-system image (MSI) operating systems, sometimes unified with an administrative interface (ClusterWorks) or at the process table level (Scyld). This is an area of research, development, and investment that has produced some initial SSI products for clusters this year (Unlimited Scale).

Numerous advantages, which are missing or only partially implemented on MSI systems, flow from the SSI feature. These include user advantages:

- One system IP address
- System process table visible from one location
- Ease of queuing and dynamic scheduling
- Fast parallel file system and parallel IO
- Check-point/restart capability
- Easy support of multiple parallel programming model (MPI, UPC, CAF, SHMEM, OpenMP)
- Modules based control of programming environment
- Globally addressable memory (hardware support also required)

They also include system administration advantages:

- Single file system and source tree to manage and backup
- System boots/halts quickly as a single unit
- Easy patching/updating
- Disk space in globally accessible for fast and efficient use
- Uniformity promotes reliability

Clusters have features that recommend them, but their ease of use and administration are not among them. Practical experience with even small MSI cluster systems at AHPCRC under scores this. They require more people to run and, if given a choice, the user community prefers to use the stable and simpler environment of SSI systems.

Users of smaller to modest sized cluster systems with less demanding and mixed workloads do not require the extras provided by SSI, but experience shows they are critical in the AHPCRC setting. Notwithstanding that, the growing interest in and application of larger clusters (at PNNL, for instance) to very large problems from mixed applications environments will continue to stimulate SSI development for clusters and promote a convergence of the cluster operating environment with that of large, customized, parallel, SSI systems like the Cray X1's Unicos/mp.

For the near term, however, the Cray X1 in expected to offer an important operational advantage from its SSI Unicos/mp operating system.

6. Cray X1 and Pentium 4 Total Costs

With two HPC systems configured to meet AHPCRC performance requirements as much as possible, the five-year, total life cycle costs were estimated. These cost were accumulated in three parts—purchase price, site preparation and installation costs, and five-year operating expenses. The site preparation costs assumed the availability of a computer room with sufficient raised floor space and assumed a certain level of in place mechanical and electrical equipment, but are otherwise not specific to any particular facility. Consequently, certain facility costs are identically excluded from the total cost estimates for both system configurations. Estimates did not include a need for any major building modifications. Actual costs would be expected to vary somewhat from facility to facility.

Purchase Price

The system purchase prices include all processors, memory, disk, and interconnection equipment required. It also includes our estimates of discounts from list, and in the case of the cluster, assumes that the purchase would be made from one of the larger suppliers of cluster systems (HP, Dell, LinuxNetworx, etc.) because of the scale and long-term commitment required. Estimated purchase prices for both systems scaled to meet the sustained performance and utilization requirements of the AHPCRC are presented in Table 6 below. The sustained Gflops numbers are reduced by each system's estimated utilization factor.

System	Sustained Gflops	CPUs	$/CPU	$/Sustained Mflops	System Cost
Cray X1	718	1024	$41,000	$58	$42M
Pentium 4	691	5760	$6,000	$50	$35M

Table 6: Estimated purchase prices

The per processor prices are per SSP for the Cray X1 and per single processor node on the Pentium 4 cluster because single processor nodes will deliver the best bandwidth to memory per processor—a key AHPCRC requirement. The processor counts used are those scaled-up based on the utilization estimates of 90% and 60% for the Cray X1 and Pentium 4 cluster, respectively. The system-wide, sustained Gflops values are derived from the single processor stream triad performance for each system multiplied by the processor count and the utilization factors.

The impact of the cluster's poor sustained performance from memory as a percentage of peak (3.5%) and low utilization estimates are clear from the table. While the Pentium 4 cluster's per processor purchase price is close to one-seventh that of the Cray X1 its price per sustained-utilized Mflops is only 15% less at purchase.

Site Preparation Costs

Estimates provided below include the cost to acquire and install power distribution and chilled-water cooling capacity as well as UPS and diesel engine backup equipment above some assumed level that is identical for both systems. Table 7 below compiles the estimated costs of each for the installation of each system.

System	Electric Work	PDU	UPS	Diesel Backup	Cooling	Total
Cray X1	$77K	$19K	$140K	$205K	$150K	$591K
Pentium 4	$223K	$56K	$345K	$300K	$550K	$1,474K

Table 7: Site preparation and installation costs

The differences in site preparation costs are driven directly by the scale of the Pentium 4 cluster. We estimated average power requirements at 150 watts per node plus the power required to cool the system. At this rate, the cluster needs ~3 times the power of the Cray X1 which drives up not only electrical equipment purchase and wiring costs, but those of UPS, diesel and cooling systems as well. With 3 times the power required, Table 7 shows site preparation expenses to be roughly 3 times those of the Cray X1.

On the other hand, site preparation costs are a small part (1-3%) of the total five-year cost of ownership of either system. Some of the other HPC systems evaluated had power and cooling site-preparation costs substantially greater than the Pentium 4 cluster.

Five-Year Operating Expenses

Table 8 shows the drivers used for operating expenses. The large scale and air-cooled nature of the Pentium 4 cluster give it a large footprint (1980 sq.ft.) made by 162 standard 42U racks. The Cray X1's smaller scale gives it a smaller foot print by comparison and smaller floor space expenses (Table 9).

System	Chasses/ Racks	Floor Space (sq. ft.)	Power (KWs)	Coolin g (KWs)	Staffing (FTEs)
Cray X1	4	840	269	89	4
Pentium 4	162	1980	914	301	9

Table 8: Operating cost drivers

System	Floor Space	Power/ Coolin g	Staffing	System (maint.)	Totals (annual)
Cray X1	$37K	$126K	$780K	$1,312K	$2,255K
Pentium 4	$88K	$426K	$1,700K	$1,772K	$3,985K

Table 9: Annual Operating costs

The power and cooling drivers and operating expenses simply recapitulate the figures from the site preparation discussion above.

Looking at system support, the number of full time equivalents estimated to support the cluster is estimated to be more than twice that of the Cray X1. We feel very confident in our Cray X1 numbers (4 FTEs) based on AHPCRC experience, expertise in maintaining Cray systems. The figure for the cluster (9 FTEs) idepends on several factors including the software running on the cluster over the five-year term, the expertise of the support team, the strategy used to maintain availability (hot swap and trash replacement versus fix and replace), and the quantity of built-in vendor support. Nine FTEs may be a conservative figure for this very large cluster system as some argue that staffing needs to be on the order of one person per 128 nodes and some sites with large clusters report as many as 15 FTEs for system support.

System hardware maintenance figures are from Cray Inc. for the X1 and from a ~5% per node per year estimate for the cluster. Total annual system maintenance for the

Pentium 4 cluster is estimated to be higher than that of the X1.

Single-year and five-totals for each system show the Cray X1 to be ~45% cheaper to operate. The total difference over the five-year term is ~$8.5 million or about 15% of the total life cycle costs for either system.

Total Five-Year Life Cycle Costs

Each system's component costs from the prior tables are summed below to give total five-year costs. The results give a small total life cycle cost advantage for the 1024 processor Cray X1 (about 4%) over the 5760 processor Pentium 4 cluster during a five-year life time.

While most would predict that the custom-engineered Cray X1 would deliver substantial advantages in both hardware and software technology over a commodity-assembled cluster for a premium price, the idea that the Cray X1 might have an absolute cost advantage in any real-world operational setting is not as obvious. Error bars on such multi-factor estimates are significant, but even so, for very large-scale systems, the data here suggest that commodity HPC solutions lose much of their price-performance advantage at very large scale when utilization and sustained performance are fully factored into the cost of ownership.

Cray X1 Costs Summed

$41,537,000	System
$ 591,000	Site Prep
$11,278,000	5 Year Operating
$53,406,000	**Total Cost**

P4 Cluster Costs Summed

$34,560,000	System
$ 1,473,000	Site Prep
$19,929,000	5 Year Operating
$55,962,000	**Total Cost**

Looking back at the technical analysis presented above in the context of the AHPCRC requirements as of Q2 of 2002, the Cray X1 clearly meets the expectation that it would offer significant hardware and software design advantages over a system based on commodity technology. These include better sustained performance, more memory per MPI process, higher bandwidth interconnect, better IO capabilities, and the many faceted benefits of its SSI operating system which lead to higher utilization and probably more satisfied users. Since, the conclusions on cost are more surprising, it make sense to recapitulate some of the important points.

First, the key requirement for high, sustained floating-point performance from memory dictated by most of the AHPCRC's applications maps into the Cray X1's vector instruction set, bandwidth from memory advantage, and the prediction that the X1 would deliver high percentages of peak performance on AHPCRC codes based on the Stream Triad performance. Using the Stream Triad as a predictor has since proved valid with key AHPCRC codes getting 20-30% or more of peak on the X1, while they get only 4-5% of peak on Pentium 4 processors. The Pentium 4 cluster processor count was scaled up to compensate for the Cray X1's 4 to 1 sustained, floating point advantage driving up its purchase and operating costs.

Second, the scale and far less integrated nature of the Pentium 4 cluster led us to predict a 50% difference in utilization between the two systems. This prompted another scaling of node count and the costs node count drives.

These two effects combined brought the estimated purchase prices to within 20% of each other. Finally, estimated total operating costs over the five years and, to a lesser extent, site preparation costs brought estimated prices for the two systems to within a few percent for total five-year life cycle costs leaving the Cray X1 at a slight advantage.

This progression is reflected more cleanly in Table 9 below, which traces dollars per Mflops through each scaling and TLCC component.

System	$/Mflops Peak (acquisition)	$$/Mflops Sustained-Utilized	$$/Mflops Installed	$$/Mflops 5-years
Cray X1	$12.70	$57.80	$58.65	$74.35
P4 Cluster	$1.10	$50.00	$52.20	$81.20

Table 10: Dollars per Mflops

The initial 12x dollar per *peak* Mflops advantage that the Pentium 4 is a theoretical peak advantage, not a real sustained advantage, particularly when operational support costs are factored in. As user requirements and operational costs are layered onto the price-performance equation, it becomes readily apparent that the cost is comparable. This leaves the systems performance differences as the deciding factor in the selection process. And, in the area of performance, the authors believe that there is a difference in kind between the Cray X1 and cluster technology that cannot be overlooked.

7. Conclusion

The Cray X1 has a decided performance advantage in almost every hardware category relevant to AHPCRC applications requirements. Its processors offer higher sustained floating-point performance, more bandwidth to memory, 16-way SMP capability, and 64-bit addressing. The Pentium 4's best features, high peak performance and faster cache, do not deliver significant benefits to AHPCRC applications.

The Cray X1 memory architecture offers much larger common workspaces that deliver larger memory maximums to individual MPI processes (at least 4 to 1) and do so at higher bandwidths. The Cray X1's vector instruction set and memory architecture allow its memory to be globally addressed while hiding memory reference latency in its vector pipeline. Scaled to its size here, the Pentium 4 offers only the potential advantage of more total memory.

While interconnect latencies between the two systems are approximately equal, the Cray X1 provides a factor of 3 or 4 more bandwidth—the interconnect feature most relevant to AHPCRC application performance.

The Cray X1's fully integrated IO subsystem is customized to deliver high aggregate bandwidth to disk from each of its processors, supported by parallel IO libraries, and without operating system file size constraints. The IO subsystem on a cluster of this scale, if commodity based, would lack the X1's aggregate bandwidth and software and hardware integrity. Higher performance, custom IO subsystems are being built for large clusters with IO features similar to those of the X1, but they drive up costs and beg the question, "Why design a one-off approach out of untried technology when it could be purchased as part of a package at similar prices?"

Finally, the Cray X1 offers the significant advantage of a mature SSI operating system that yields benefits in improved utilization, ease of administration, and ease of use.

Specific items include check-point restart, better job scheduling, global parallel file system, rapid shutdown and reboot, more parallel programming models (including the new implicit DM models CAF and UPC).

While these software conveniences are being designed and in some measure delivered for large-scale cluster operations, they are not fully featured or provided as an integrated package today.

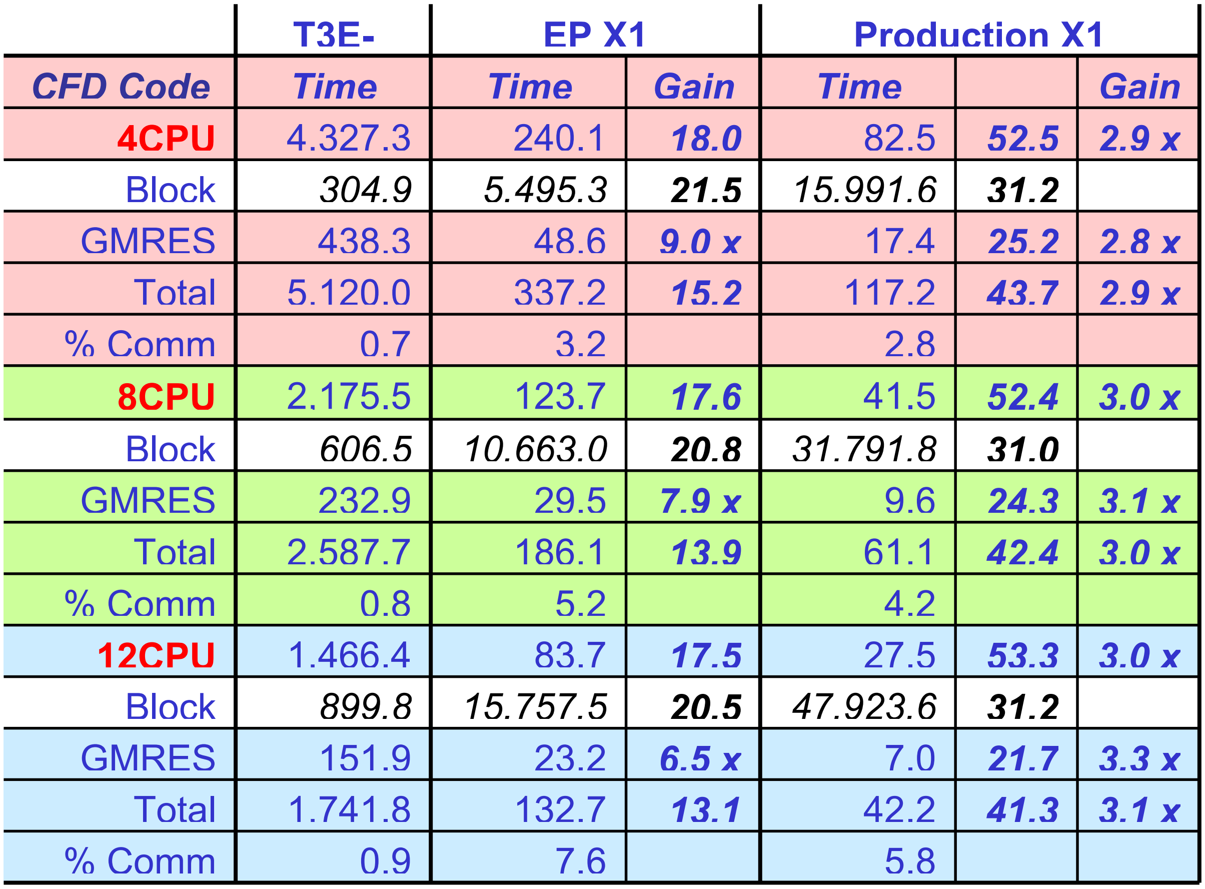
We conclude that, for the very large scale, multi-user environment requirements of the AHPCRC, and for its set of applications and high utilization requirements, the Cray X1 is the best next-generation system and the best replacement for the 1088 processor CRAY T3E-1200. Furthermore, in similar settings elsewhere, regimes where very large-scale problems must be solved in a multi-user environment, we expect the Cray X1 to compete with clusters on total life cyclic costs and to excel in terms of system software and hardware features required to support production capability computing requirements.

About the Authors

Paul Muzio is Vice-President HPC Programs at Network Computing Services, Inc. and Director of Support Infrastructure for the Army High Performance Computing Research Center.

Richard Walsh is a Project Manager at Network Computing Services, Inc. that has worked in the HPC industry and with Cray systems for 20 years. His expertise includes computer architecture and parallel programming, Linux cluster design and assembly, and computational chemistry and finance.

Appendix A



	T3E-	EP X1		Production X1		
CFD Code	*Time*	*Time*	*Gain*	*Time*		*Gain*
4CPU	4.327.3	240.1	***18.0***	82.5	***52.5***	***2.9 x***
Block	*304.9*	*5.495.3*	***21.5***	*15.991.6*	***31.2***	
GMRES	438.3	48.6	***9.0 x***	17.4	***25.2***	***2.8 x***
Total	5.120.0	337.2	***15.2***	117.2	***43.7***	***2.9 x***
% Comm	0.7	3.2		2.8		
8CPU	2.175.5	123.7	***17.6***	41.5	***52.4***	***3.0 x***
Block	*606.5*	*10.663.0*	***20.8***	*31.791.8*	***31.0***	
GMRES	232.9	29.5	***7.9 x***	9.6	***24.3***	***3.1 x***
Total	2.587.7	186.1	***13.9***	61.1	***42.4***	***3.0 x***
% Comm	0.8	5.2		4.2		
12CPU	1.466.4	83.7	***17.5***	27.5	***53.3***	***3.0 x***
Block	*899.8*	*15.757.5*	***20.5***	*47.923.6*	***31.2***	
GMRES	151.9	23.2	***6.5 x***	7.0	***21.7***	***3.3 x***
Total	1.741.8	132.7	***13.1***	42.2	***41.3***	***3.1 x***
% Comm	0.9	7.6		5.8		

Figure 1. AHPCRC computational fluid dynamics applications are achieving 31% of peak for the key computational kernels on the Cray X1.

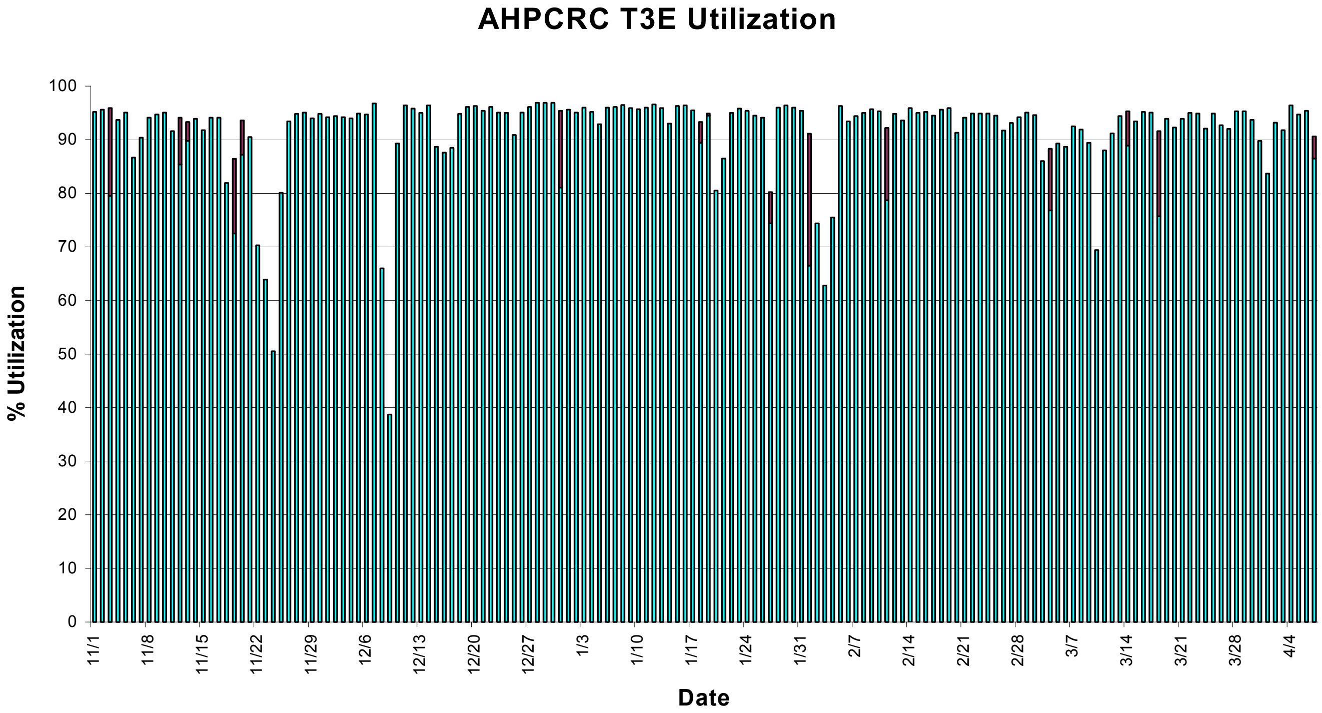


Figure 2. The AHPCRC's CRAY T3E-1200 operates with a single system image and routinely sustains over 90% utiliziation.