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

**Paul Muzio**

**Richard Walsh**

May 2003

# Preliminaries

"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."

# Agenda

- User Requirements
- System Requirements
- Hardware Comparison
  - Processor, memory, interconnect
- Software Comparison
  - Operating environment (user and administration)
- Availability and Utilization
  - Ability to support capability computing requirements
- Cost Comparison
  - Purchase + site + operating
- Total Life Cycle Cost per Sustained Mflops
  - Five year TLCC
  - Based on sustained, floating-point, performance from memory
- Conclusions

# User Requirements

- Robust programming environment
  - MPI, Fortran, C optimizing compilers
  - Debugging and performance tuning tools with GUI
  - Optimized scientific libraries
  - Easy-to-use operating system interface
- Capability job requirements
  - 500 gigabytes of memory
  - 650 GFLOPS sustained performance from memory
  - MPI latency of less than 10 microseconds
  - High interconnect bandwidth
  - Fast parallel disk I/O (500+ Mbytes)
  - Robust file system
  - Checkpoint restart (user)
- Support for capability jobs
  - Up to 50% of systems resources on demand
  - Entire system on a scheduled basis

# Operational Requirements

- Effective scheduling
  - Global view of all system resources and queues
  - Flexible and dynamic ability to allocate/re-allocate resources
  - Gang scheduling, swapping, migration, compaction capability
  - Checkpoint/restart (system initiated)
  - Fast I/O to disk (500 Gbytes in less than 30 minutes)
- Single System Image (SSI)
  - Ease of administration/minimal staffing requirements
  - Enhanced security
  - Quick system re-boot
  - Comprehensive diagnostics
  - Parallel I/O system under SSI
  - Global accounting system
- Facility conservation
  - Dense packaging
  - Liquid cooling (more energy/cost efficient than air cooling)

# Study Methodology

- Inventoried applications
  - Most were scalable MPI-based applications
  - Primarily floating-point calculations
  - Often bandwidth limited with medium to large messages
  - At least one critical application with large numbers of small messages at every time step
- Projected future requirements and algorithmic requirements
  - Compared application algorithmic applications to stream triad and SPEC FP2000 benchmarks
  - Concluded that stream triad was a reasonable approximation of application kernels
  - Did not over-emphasize full benchmarking of applications
    - Generally represent past, not future
    - Usually sized for least common denominator
    - Systems not always available for benchmarking
- Vendor specifications/literature review
- Analyzed data and projected results

# Study Methodology

- Sized hypothetical systems based on available data and operational considerations
  - Sustained Gflops from memory (stream triad)
  - Adjusted with estimates of system utilization and availability
- Computed a "Total Life Cycle Cost" to include
  - Acquisition cost
  - Facility modification cost
  - 5-year operating cost
  - Relied on internal data, technical specifications, vendor pricing
- Computed price-performance for 650 Gflops sustained performance target systems

# Study Methodology

## Systems Analyzed

- *Vector Architectures*
  - CRAY X1 (custom vector)
- *Integrated RISC Architectures*
  - HP SuperDome (PA-8700)
  - IBM 690/1600 (Power4)
- *Cluster Architectures*
  - Alpha (EV6 21264 and EV7 21364)
  - Intel IA-64 (Itanium 2)
  - Intel IA-32 (Pentium 4)

Compared today

# Study Methodology

- Why focus on Intel Pentium 4 (IA32) cluster?
  - Typical commodity competitor
  - High flops per processor at very low cost
  - Better stream memory bandwidth than most
  - If you beat it on cost, likely to beat other alternatives

# Floating Point Performance

| *Performance feature* | *Cray X1* | *Pentium 4* |
|---|---|---|
| Sustained Mflops/CPU | 780* | 200** |
| Percent of peak | 24% | 3.4% |
| Peak Mflops/CPU | 3200 | 5600 |
| CPUs needed | 896*** | 3456*** |

* *Stream triad from on-board memory run on one Cray X1 Single Streaming Processor (SSP)*

** *Single 2.8 GHz Pentium 4 CPU*

*** *Processor counts are round up to multiples of 128*

# Floating Point Performance Comments

- Cray X1 schedules 4 hardware-integrated SSPs as a single MSP (Multi-Streaming Processor)
  - MSP has a peak of 12.8 Gflops
  - Each SSP has both a vector and scalar processor
  - Scalar processor has a 400 MHz clock
- Pentium 4 clock is nearly 4x Cray X1 (3 GHz vs 800 MHz)
- But, Cray X1 SSP has 4x the sustained floating point performance of Pentium 4
- Both systems show performance boosts inside cache
- Cray X1's pipelined, vector instruction architecture delivers much higher, sustained performance

# Floating Point Performance

- Was the Stream Floating Point Performance estimate (Cray X1 SSP 4 times Pentium 4) reasonable for our applications?
  - Dramatic differences in clock speeds
- Yes, from current benchmarks of our codes …
  - Achieved 18% of peak on MM5
  - Achieved 22% of peak on CSM application (estimate)
  - Achieved 31% of peak on CFD application
  - Still early in the product cycle, expect to see further improvements in performance
  - Expect to see greater advantages with larger problems

# Floating Point Performance-CFD

## Large Data Set

| | T3E-1200 | EP X1 | | Production X1 | | |
|---|---|---|---|---|---|---|
| *Code Section* | *Time (secs)* | *Time (secs)* | *Gain* | *Time (secs)* | *Gain 1* | *Gain 2* |
| **4CPU** Block | 4,327.3 | 240.1 | ***18.0 x*** | 82.5 | ***52.5 x*** | ***2.9 x*** |
| Block MFlops | *304.9* | *5,495.3* | ***21.5%*** | *15,991.6* | ***31.2%*** | |
| GMRES | 438.3 | 48.6 | ***9.0 x*** | 17.4 | ***25.2 x*** | ***2.8 x*** |
| Total | 5,120.0 | 337.2 | ***15.2 x*** | 117.2 | ***43.7 x*** | ***2.9 x*** |
| % Comm | 0.7 | 3.2 | | 2.8 | | |
| **8CPU** Block | 2,175.5 | 123.7 | ***17.6 x*** | 41.5 | ***52.4 x*** | ***3.0 x*** |
| Block MFlops | *606.5* | *10,663.0* | ***20.8%*** | *31,791.8* | ***31.0%*** | |
| GMRES | 232.9 | 29.5 | ***7.9 x*** | 9.6 | ***24.3 x*** | ***3.1 x*** |
| Total | 2,587.7 | 186.1 | ***13.9 x*** | 61.1 | ***42.4 x*** | ***3.0 x*** |
| % Comm | 0.8 | 5.2 | | 4.2 | | |
| **12CPU** Block | 1,466.4 | 83.7 | ***17.5 x*** | 27.5 | ***53.3 x*** | ***3.0 x*** |
| Block MFlops | *899.8* | *15,757.5* | ***20.5%*** | *47,923.6* | ***31.2%*** | |
| GMRES | 151.9 | 23.2 | ***6.5 x*** | 7.0 | ***21.7 x*** | ***3.3 x*** |
| Total | 1,741.8 | 132.7 | ***13.1 x*** | 42.2 | ***41.3 x*** | ***3.1 x*** |
| % Comm | 0.9 | 7.6 | | 5.8 | | |

*% of Peak*

AHPCRC



NETWORK COMPUTING SERVICES, INC.

# MM5 Standard Benchmark

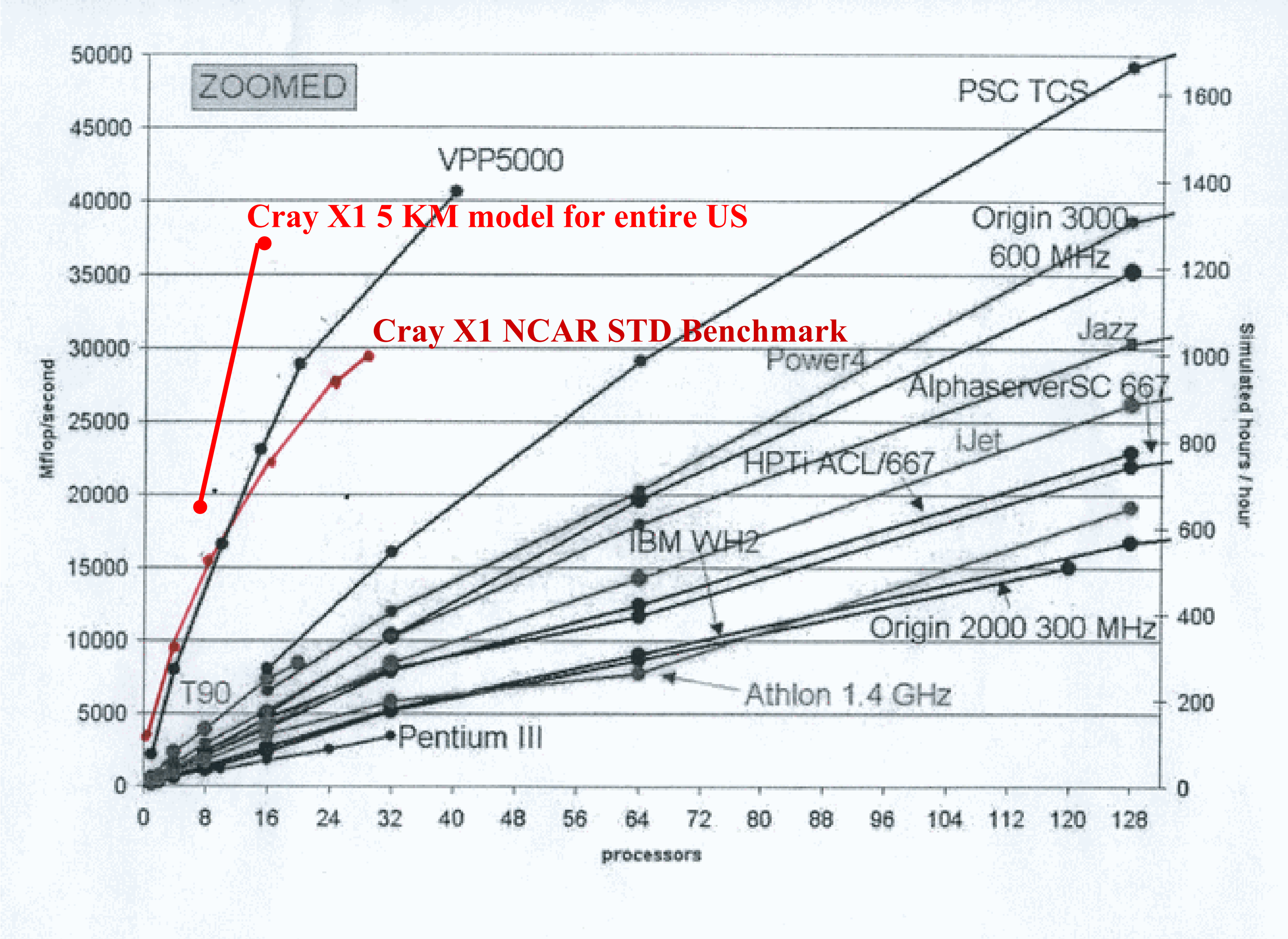



Figure 1b. MM5 floating-point performance on various platforms (zoomed). (Updated March 5, 2003)

AHPCRC

Network Computing Services, Inc.

# Comments on MM5

- Standard benchmark at 10 Gflops
  - 16 Cray X1 SSPs (4 MSPs)
  - 76 Athlon 1.4 GHz processors
  - Reflects (approximately) 4 to 1 advantage
- Standard benchmark at 20 Gflops
  - 48 Cray X1 MSPs (12 MSPs)
  - 128 Athlon 1.4 GHz processors
  - Indicates a drop in Cray X1 performance (benchmark is too small)
- 19 Gflops hybrid comparison
  - 8 Cray MSPs (32 SSPs)
  - 128 Athlon processors
  - Reflects 4 to 1 advantage
  - Cray X1 likes big, capacity jobs

# MM5

- **Operational weather models for United States are typically run at a resolution of about 10 kilometers**
- **AHPCRC demonstrated use of MM5 on Cray X1**
  - 5 kilometers resolution
  - Entire US
  - 33 levels
  - 8X computations
  - 4X memory (20 billion bytes)
- **Cray X1**
  - Sustained 36.7 GFLOPS on 16 MSPs while executing the forecast steps
  - 18 percent of peak
  - Simulated 1 hour of atmospheric physics and dynamics in 8.4 minutes on average, or 24 simulation hours in under 3.5 wall clock hours
- Sites that have clusters to do this work AREN'T

# Application Floating Point Performance

- CFD Block clocked 31.8 Gflops on 32 SSPs (8 MSPs)
  - Stream triad is 780 Mflops per SSP
  - Actual results were 993.4 Mflops per SSP
- MM5 clocked 37.6 Gflops on 64 SSPs (16 MSPs)
  - Stream triad is 780 Mflops per SSP
  - Actual results were 588 Mflops per SSP
- Early in product life cycle
  - Programming environment improvements
  - Additional optimization work

# Memory Designs

| *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 |

**   X1 address is 2**64, MPI board-limited; Pentium 4 is 2**32 – operating system limited*

*** Balance is peak Flops/sustained Mops (lower is better)*

# Memory Design Comments

- Cray X1 delivers its additional available write bandwidth
  - note stream triad benchmark
- Pentium 4 does not
  - reads and writes compete
- Vector loads/stores from memory beat scalar pre-fetching
  - Hide memory latency
  - Delivers designed bandwidth (1/4 peak requirements)
- Memory architecture
  - A perfect system could stream all data needed to sustain peak
  - Pentium 4 streams ~1/20 of what is needed
  - X1 streams ~1/3 of what is needed
- Cray X1 can address any memory location in the system (on-board or off) with a single vector instruction
  - Memory to register and/or
  - Memory to cache

# Interconnect Designs Compared

| ***Performance feature*** | ***Cray X1*** | ***Pentium 4*** |
|---|---|---|
| Interconnect type* | x-bar/switched 2D hypercube | Clos-network (Myrinet) |
| MPI ping-pong bandwidth (Mbytes/sec, off-board, 32Kbyte message) | ~750*2 (two-way) | ~200*2 (two-way) |
| MPI ping-pong latency** (local/remote node, small message) | ~7.5/15 usecs | ~7/10 usecs |
| Co-array Fortran** (local/remote node, small message) | ~6/12 usecs | NA |

* *X1 Interconnect network varies depending on system size; for Pentium 4 a Myrinet-switched, Clos architecture was assumed.*

** *Cray is working to reduce its barrier times to an average of ~2 usecs between MSPs on the same node, success would lower send/receive latencies to the 6 usec range.*

# Interconnect Design Comments

- Cray X1 interconnect (and designed bandwidth) is hierarchical
  - 38.4 Gbytes/sec per SSP on-board (cross-bar to memory)
  - 12.8 Gbytes/sec/per SSP (layer 2, 1 hop, direct board-to-board)
  - 3.2 Gbytes/sec/per process (layer 4, 3 hops, 1024 SSPs)
  - 1.6 Gbytes/sec/per SSP (layer 5, 4 hops, 4096 SSPs)
- Cray X1 bandwidth exceeds Cray T3E and Myrinet by a factor of 3 or 4 as measured by the Pallas MPI Benchmark (worst case)

# Interconnect Design Comments (cont.)

- Myrinet with Clos interconnect required for cluster system of this size
  - Requires 7-8 hops between the most-remote processors
  - Somewhat higher latencies between remote processors
  - Bandwidth is the same at all scales for a Clos network (0.5 Gbytes/sec)
- Cray CAF and UPC models offer direct path to hardware-only latencies via inter-node vector copy instruction
  - X(msize)[1] = X(msize)[2]; call sync; X(msize)[2] = X(msize)[1]
- Other interconnects (Quadrics, SCI) have price-performance profiles similar to Myrinet; Gigabit Ethernet has poor latencies.

# I/O Designs Compared

| *Performance feature* | *Cray X1* | *Pentium 4* |
|---|---|---|
| File server design | Custom Node-to-PCI-X/FC-AL/Raid 5 | GigE Switched Uplink and Local Raid |
| Bandwidth to disk | 600+ Mbytes/sec* | ~200 Mbytes/sec local<br>~100** Mbytes/sec remote |
| Maximum file size | Only file system size limited<br>100+ TB | Block offset limited<br>2-4TB |

** Every disk is visible from all nodes on the X1, performance depends on number of FC-HBA controllers and level of stripping.*

*** This is per node, aggregate depends on number of uplinks and parallel IO*

# IO Design Comments

- Cray X1's design includes 4 SPC 1.2 Gbytes/sec ports per node board (1 per MSP)
- X1's rate limiting component is the number of controllers in FC-AL system (200 Mbytes/sec each with two per FC-HBA card)
- Pentium 4 cluster design includes out-of-band GigE switched file server for non-local storage
- Actual aggregate remote bandwidth for cluster depends on number of Gigabit uplinks to the server (8 minimum)
- SAN solutions, more costly than GigE, are possible for Pentium 4 cluster (PNNL will have one)
- Cray X1's parallel IO software complements its IO hardware design

# Specific Cray X1 Design Advantages

- 16,384 processor (SSP) scalability
- 16-way SMP (cross-bar) character of the X1 node-board
  - Good for mixing of SMP/MPP parallel models
- 64-bit addressable memory
- Highly banked memory architecture (16x4x2/board)
- Logically shared memory space
  - Globally addressable by single vector load/store
- Integrated, high-performance, IO subsystem
- Special instructions (BMM, POP count)

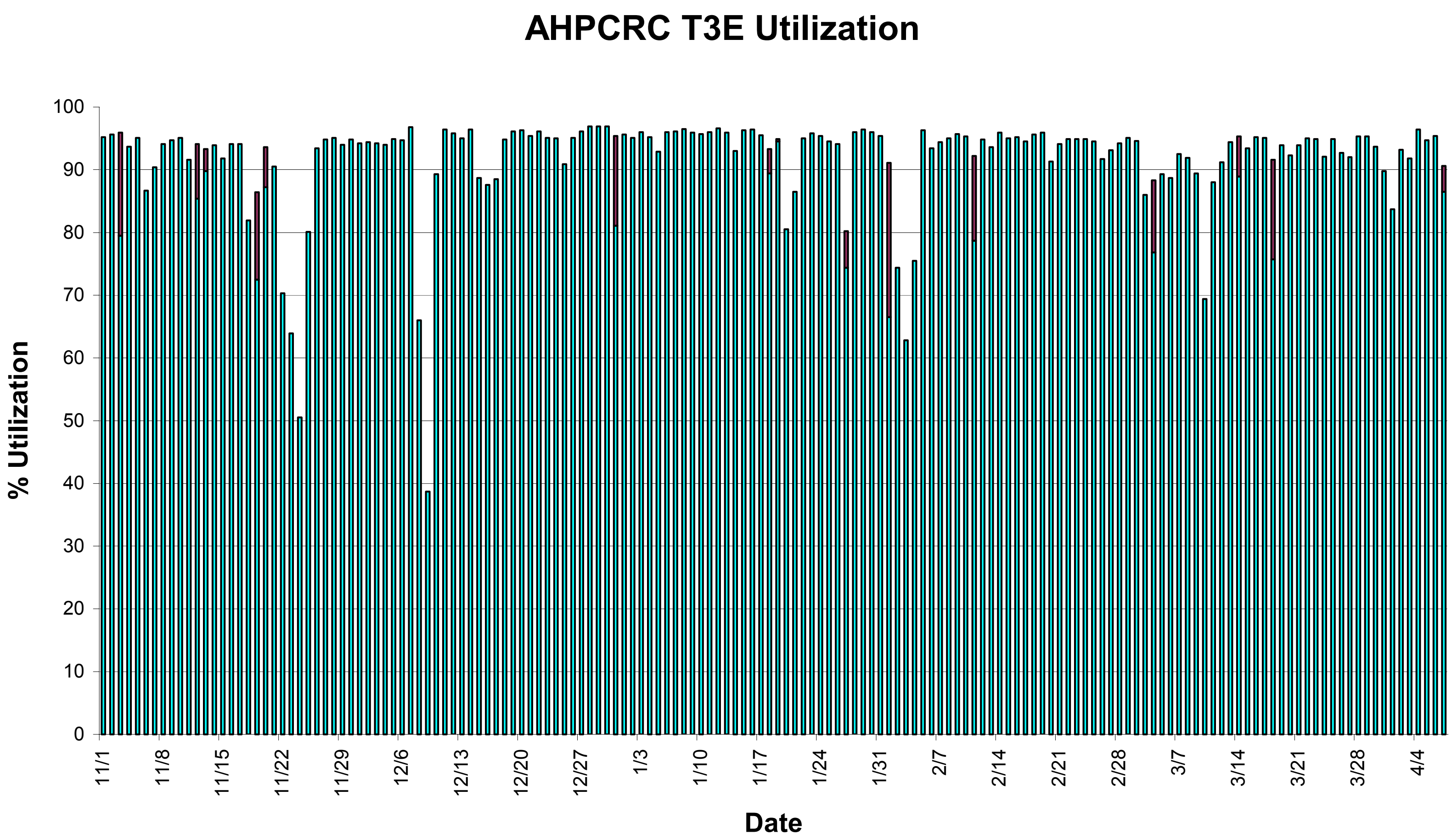
# Utilization

- Must Include Anticipated % Utilization in Analysis
  - A downed/unused system (or node) has zero sustained performance
  - Sustained performance requirement of 650 Gflops must be delivered to a 2-3 day job
  - True costs rise when system is down or job fails, therefore …
  - Systems are scaled in size to compensate for imperfect reliability

# Utilization

- Cray X1 5-Year Utilization Estimate of 90% based on
  – Single system image operating system
  – Custom parallel scheduler and checkpoint restart
  – 95% utilization observed with 1088 cpu T3E-1200
  – Node count scaled up from *896 to 1024*
- Pentium 4 Cluster 5-Year Utilization of 60% based on
  – Multi-system image operating environment
  – Limited scheduling and check-point capability
  – Observed utilization for very large clusters
  – Node count scaled up from *3456 to 5670*

# T3E-1200 Utilization




AHPCRC

NETWORK COMPUTING SERVICES, INC.

# IA32 Cluster Utilization

- Less effective schedulers
  - No system check-pointing
  - Cluster schedulers are less flexible/efficient
    - statically scheduled
    - no gang sceduling/swapping
    - no job migration/compaction
    - no job pre-emption
- Longer time to drain system
- Longer boot times
- System software install time
- Lost jobs
- 60% utilization is a reasonable estimate
- Cluster node count scaled up from *3456 to 5670*

# Target System Sizes/Performance

| *Component Performance Target* | *Cray X1* | *Pentium 4* |
|---|---|---|
| Number of processors | 1024 SSPs | 5760 CPUs |
| Aggregate memory (Gbytes) | 1,024 | 1,440 |
| Memory Bandwidth (Gbytes/sec/cpu stream) | 9.35 | 2.25 |
| Interconnect Bandwidth (two-way, ping-pong, PMB performance 32 Kbyte message) | 2x750 X1 hyper-cube | 2x200 Myrinet Clos network |
| MPI Latency for small messages (average)* | ~10 | ~8 |

** Cray is targeting 6-7 usecs for MPI latencies when development is complete*

# Software Design Comparison

| ***Software Component*** | ***Cray X1 (Unicos/mp)*** | ***Pentium 4 (Linux)*** |
|---|---|---|
| Single System Image (SSI) | Yes | Not yet |
| Global Direct Memory Address Space | Yes | No |
| Global Parallel File System | Yes | Approx |
| Dynamic gang scheduling/swapping, priority pre-emption, job migration/compaction | Yes | Not yet |
| Checkpoint Restart | Yes | No |
| Parallel Programming Models: | | |
| •MPI, Shmem, OpenMP, pthreads | Yes | Yes |
| •CAF, UPC, Streams, Vectors, | Yes | No |
| Modules based control of software | Yes | No |

# Advantages of SSI

- Cray X1's SSI provides
  - Ease of administration
    - One file and operating system tree
    - Easier global monitoring
    - Security and system updates/patches done once globally
    - Fast, total system reboot
    - More efficient scheduling, better utilization of resource
    - Fewer lost jobs
  - Staffing requirements are fixed
    - Four FTEs to support CRAY T3E-1200 and three Cray X1s
- Systems staffing requirements for clusters are higher
  - Staffing requirements tend to scale with system size
  - Estimates are as high as 1 FTE for every 128 processors
    - Over 40 people for the hypothetical cluster
    - For this analysis we assumed a staffing requirement of just 9 people for the cluster

# Programming Environment

- Ease of use and preference by researchers
  - Separately schedule systems and applications nodes
  - One IP address, node blind execution
  - All running processes are easily visible
  - Global file space
- Fewer, more powerful processors are better than more, less powerful processors
  - Easier to program
  - Better scaling
- Added features
  - Direct global memory addressing in hardware provides distinct performance advantage (reduced latency) for UPC and CAF applications
  - UPC and CAF are intuitive and easier to use
  - Special instructions

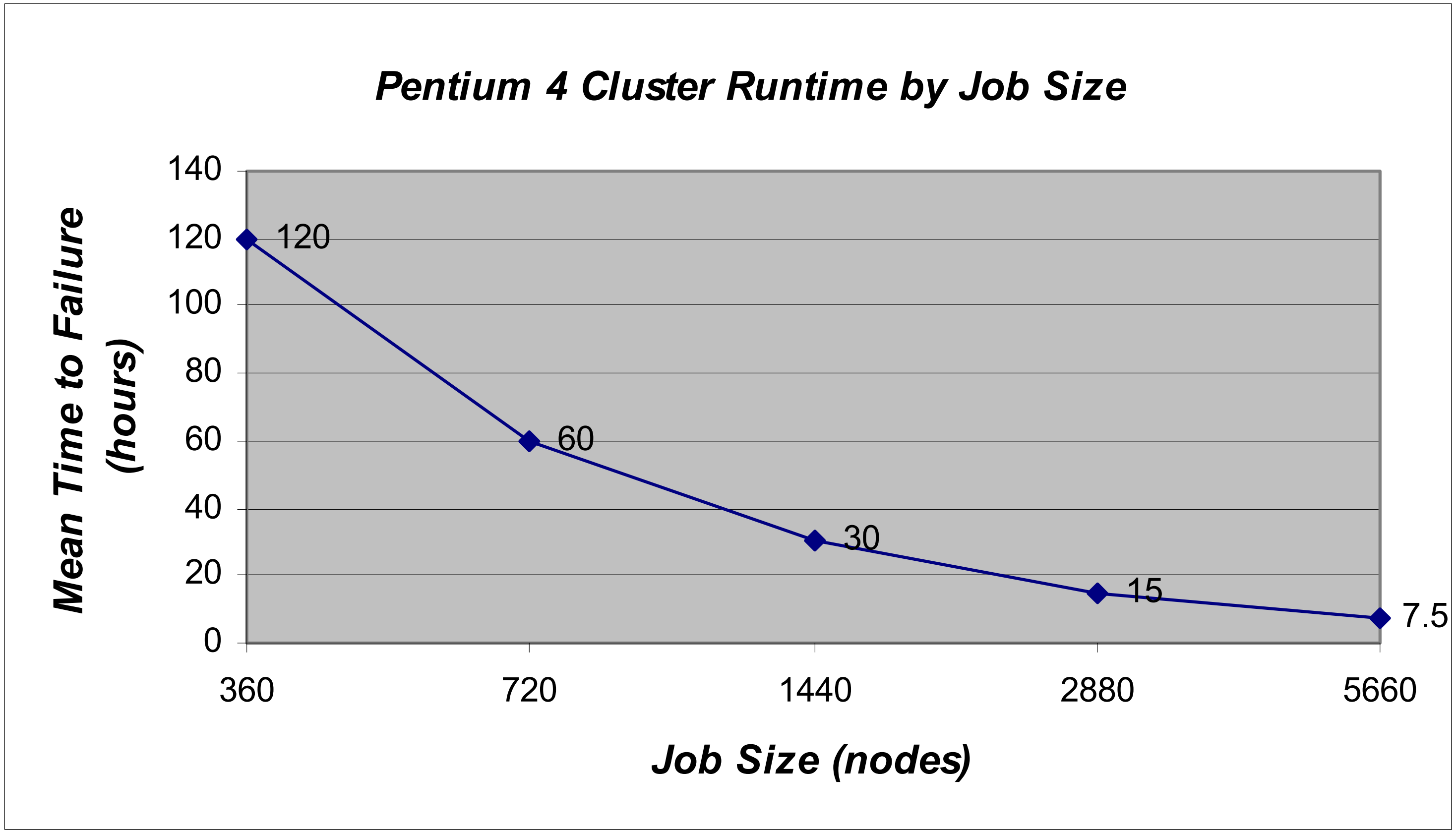
# Cluster Capability Job Sustainability

What is the probability that a job using thousands of commodity processors for days will run to completion?

- Assume every node (5760) has 1 hardware failure once in 5 years
  - Variance of MTTF assumed to be the (mean/4)**2
  - Failure distribution assumed to be random and uniform
  - Number of failures scale
- Assume no checkpoint-restart
- Compute mean time to failure by job size
- Can the system deliver 650+ Gflops on a single application continuously for 2-3 days?

***This is a system performance requirement***

# Capability Job Sustainability



# Capability Job Sustainability

- At this hardware failure rate (~3 per day), the Pentium 4 cluster will not meet the 2-day requirement even 3% of the time.

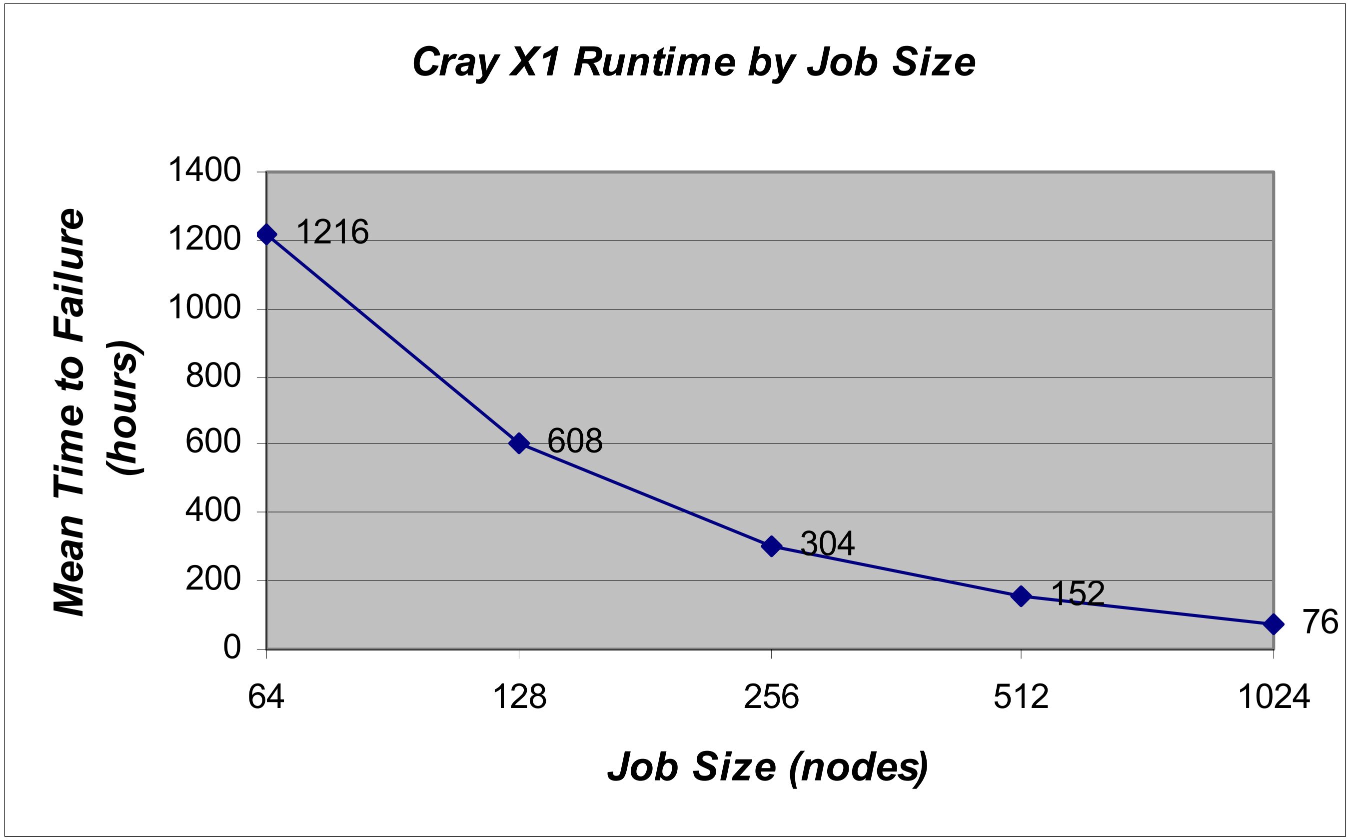
# Capability Job Sustainability

- What is the probability that a job using thousands of Cray X1 SSPs for days will run to completion?

- Assume failure rate of 1/10th of Pentium 4 cluster
  - Failure rate on the X1 is expected to be **much lower**
    - **Two early production systems experienced only 1 hardware failure in a total of 12 months of operation**
  - Variance of MTTF assumed to be the (mean/4)**2
  - Failure distribution assumed random and uniform
  - Checkpoint-restart is available (but not factored into the analysis)
  - Compute mean time to failure by job size
  - Can the system deliver 650+ Gflops for 2 days?

***This is a system performance requirement***

# Capability Job Sustainability



# Capability Job Sustainability

- Cray X1 Gflops Delivered versus Gflops Required
  - 97% of the time, under these conservative assumptions, the Cray X1 will deliver the 2-day requirement
  - This is an extremely conservative estimate

# Cost Comparison Components

## Three Core Components

- Purchase price
  - Includes discounts
  - Assumes top tier vendor capable of full support
- Site preparation costs
  - Suitable building assumed
  - Some implicit site specificity
- Operating costs over 5-years
  - Based on local experience and staff expertise

# Cost Comparisons

- Core Purchase Price for Target Systems of 650+ Sustained Gflops
  - includes processor, memory, interconnect and disk
  - Pentium 4 cluster has one CPU per node to maximize bandwidth
  - Utilization is factored into total Gflops estimates

| 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 |

**   Running stream triad from on-board memory*

*** 1024 SSPs, 256 MSPs*

# Cost Comparisons

- Site Preparation Cost for Each System
  - Fully powered, cooled, and generator backed-up
  - No major structural site modifications assumed

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

# Cost Comparisons

- Operating Cost Drivers

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

- Annual Operating Cost

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

# Cost Comparisons

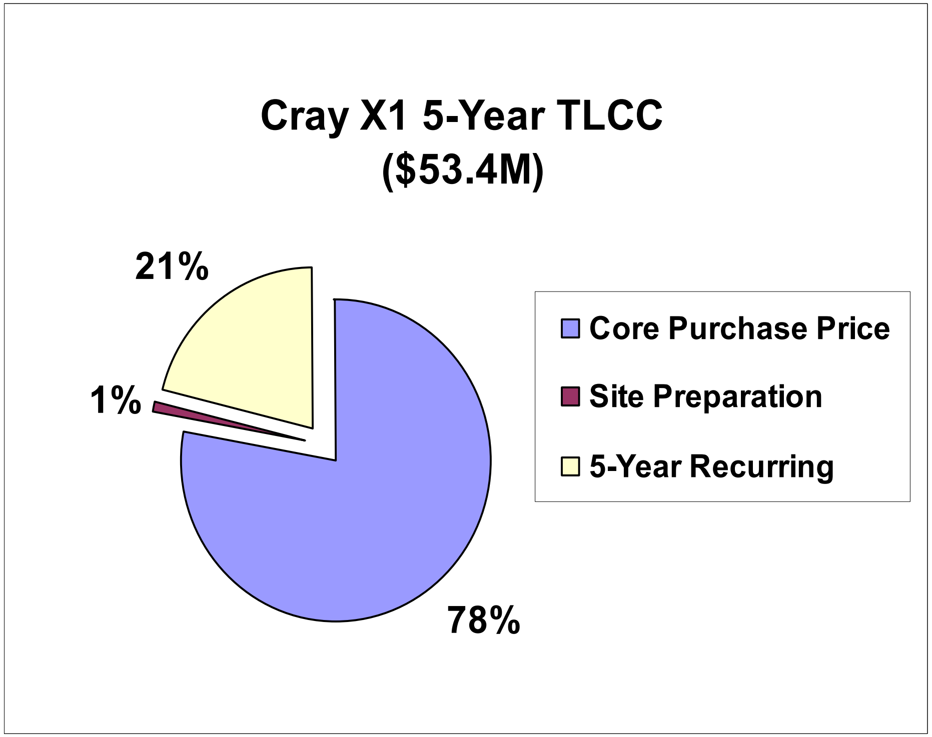
- Cray X1 Cost

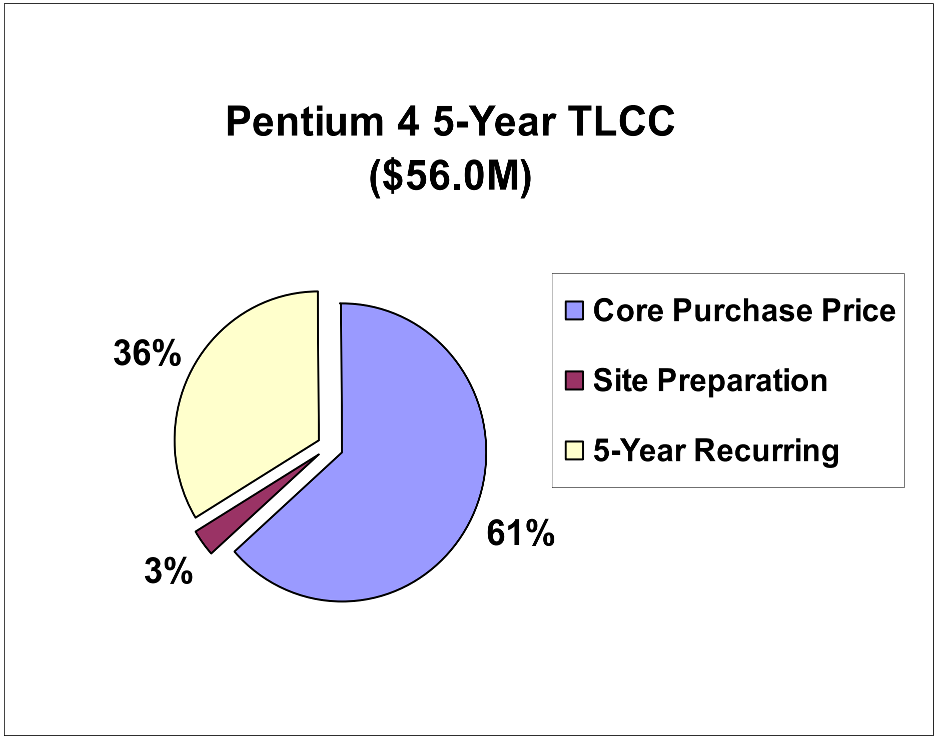
| | |
|---|---|
| $41,537,000 | System |
| $ 591,000 | Site Prep |
| $11,278,000 | 5 Year Operating |
| $53,406,000 | Total Cost |

- Pentium 4 Cluster Cost

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

# Cost Comparisons



# Cost Comparisons

- Tracking Dollars per Mflops through TLCC

| System | $/Mflops Peak (@purchase) | $/Mflops Sustained (@purchase)* | $/Mflops Sustained (installed) | $/Mflops Sustained (5-year) |
|---|---|---|---|---|
| Cray X1 | $12.70 | $57.80 | $58.65 | $74.35 |
| Pentium 4 | $1.10 | $50.00 | $52.20 | $81.20 |

** Adjusted with utilization estimates*

# Summary

- Cost estimates are conservative
  - Acquisition cost estimate for the Cray X1 can be bettered
  - Labor cost for the Cluster support was understated
- Performance model (stream triad) was validated in recent benchmarking on X1
- Capability computing model is validated for Cray X1
  - Demonstrated ability
  - Effective global schedulers
  - High reliability and availability
- Capability computing model is weak for large cluster configurations
  - Cluster COTS model does not, at present, scale well for capacity computing
  - Scheduling robustness, flexibility, and efficiency is lacking
  - Difficulty in maintaining high availability across a defined “capability” block of components to complete very large job

# Summary

- Production Support
  - Cray X1 has features required in a production environment
    - Single system image
    - Effective queuing systems including gang-scheduling and fast job migration
    - Global accounting
    - Enhanced security
  - Cluster lacks production strength features
    - Sites are investing human resources to develop system or manual fixes
    - User productivity suffers
    - Encourages submission of multiple small jobs, rather than new technology jobs
- Cray X1 compares favorably on a cost basis with the least expensive cluster alternative at capability scales