Performance Evaluation of Supercomputers using HPCC and IMB Benchmarks

Subhash Saini
Robert Ciotti, Brian T. N. Gunney, Alice Koniges, Don Dossa
Panagiotis Adamidis, Rolf Rabenseifner, Sunil R. Tiyyagura,
Matthias Mueller, and Rod Fatoohi

NASA Advanced Supercomputing (NAS) Division
NASA Ames Research Center, Moffett Field, California

IPDPS 2006 - PMEO, Rhodes, Greece, April 29
Outline

- **Computing platforms**
  - Columbia System (NASA, USA)
  - NEC SX-8 (HLRS, Germany)
  - Cray X1 (NASA, USA)
  - Cray Opteron Cluster (NASA, USA)
  - Dell POWER EDGE (NCSA, USA)

- **Benchmarks**
  - HPCC Benchmark suite
  - IMB Benchmarks

- **Summary**
NEC SX-8 System
SX-8 System Architecture
SX-8 Technology

- Hardware dedicated to scientific and engineering applications.
- CPU: 2 GHz frequency, 90 nm-Cu technology
- 8000 I/O per CPU chip
- Hardware vector square root
- Serial signalling technology to memory, about 2000 transmitters work in parallel
- 64 GB/s memory bandwidth per CPU
- Multilayer, low-loss PCB board, replaces 20000 cables
- Optical cabling used for internode connections
- Very compact packaging.
SX-8 specifications

- 16 GF / CPU (vector)
- 64 GB/s memory bandwidth per CPU
- 8 CPUs / node
- 512 GB/s memory bandwidth per node
- Maximum 512 nodes
- Maximum 4096 CPUs, max 65 TFLOPS
- Internode crossbar Switch
- 16 GB/s (bi-directional) interconnect bandwidth per node
- Maximum size SX-8 is among the most powerful computers in the world
Columbia 2048 System

- Four SGI Altix BX2 boxes with 512 processors each connected with NUMALINK4 using fat-tree topology
- Intel Itanium 2 processor with 1.6 GHz and 9 MB of L3 cache
- SGI Altix BX2 compute brick has eight Itanium 2 processors with 16 GB of local memory and four ASICs called SHUB
- In addition to NUMALINK4, InfiniBand (IB) and 10 Gbit Ethernet networks also available
- Processor peak performance is 6.4 Gflop/s; system peak of the 2048 system is 13 Tflop/s
- Measured latency and bandwidth of IB are 10.5 microseconds and 855 MB/s.
Columbia System
- Itanium 2@ 1.5GHz (peak 6 GF/s)
- 128 FP reg, 32K L1, 256K L2, 6MB L3

- CC-NUMA in hardware
- 64-bit Linux w/ single system image -- looks like a single Linux machine but with many processors
Columbia Configuration

Front End
- 128p Altix 3700 (RTF)

Networking
- 10GigE Switch 32-port
- 10GigE Cards (1 Per 512p)
- InfiniBand Switch (288port)
- InfiniBand Cards (6 per 512p)
- Altix 3700 2BX 2048 Numalink Kits

Compute Node (Single Sys Image)
- Altix 3700 (A) 12x512p
- Altix 3700 BX2 (T) 8x512p

Storage Area Network
- Brocade Switch 2x128port

Storage (440 TB)
- FC RAID 8x20 TB (8 Racks)
- SATARAID 8x35TB (8 Racks)
Cray X1 CPU: Multistreaming Processor

- New Cray Vector Instruction Set Architecture (ISA)
- 64- and 32-bit operations, IEEE floating-point

**Each Stream:**
- 2 vector pipes
  (32 vector regs. of 64 element ea)
- 64 A & S regs.
- Instruction & data cache

**MSP:**
- 4 x P-chips
- 4 x E-chips (cache)

**Bandwidth per CPU**
- Up to 76.8 GB/sec read/write to cache
- Up to 34.1 GB/sec read/write to memory
Cray X1 CPU: Multistreaming Processor

- New Cray Vector Instruction Set Architecture (ISA)
- 64- and 32-bit operations, IEEE floating-point

Each Stream:
- 2 vector pipes (32 vector regs. of 64 element ea)
- 64 A & S regs.
- Instruction & data cache

MSP:
- 4 x P-chips
- 4 x E-chips (cache)

Bandwidth per CPU
- Up to 76.8 GB/sec read/write to cache
- Up to 34.1 GB/sec read/write to memory
Cray X1 Processor Node Module

X1 node board has performance roughly comparable to:
- 128 PE Cray T3E system
- 16-32 CPU Cray T90 system
Cray X1 Node - 51.2 Gflop/s

Interconnect network
2 ports/M-chip
1.6 GB/s/port peak in each direction
= 102.4 GB/s to the network

Local memory
Peak BW = 16 sections x 12.8 GB/s/section = 204.8 GB/s
Capacity = 16, 32 or 64 GB
Cray X1 at NAS

Architecture
- 4 nodes, 16 MSPs (64 SSPs)
- 1 node reserved for system; 3 nodes usable for user codes
- 1 MSP: 4 SSPs at 800 MHz, 2 MB ECache 12.8 Gflops/s peak
- 64 GB main memory; 4 TB FC RAID

Operating Environment
- Unicos MP 2.4.3.4
- Cray Fortran and C 5.2
- PBSPro job scheduler
Intel Xeon Cluster ("Tungsten") at NCSA
## High End Computing Platforms

**Table 2: System characteristics of the computing platforms.**

<table>
<thead>
<tr>
<th>Platform</th>
<th>Type</th>
<th>CPUs/node</th>
<th>Clock (GHz)</th>
<th>Peak/node (Gflop/s)</th>
<th>Network</th>
<th>Network Topology</th>
<th>Operating System</th>
<th>Location</th>
<th>Processor Vendor</th>
<th>System Vendor</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGI Altix BX2</td>
<td>Scalar</td>
<td>2</td>
<td>1.6</td>
<td>12.8</td>
<td>NUMALINK 4</td>
<td>Fat-tree</td>
<td>Linux (Suse)</td>
<td>NASA (USA)</td>
<td>Intel</td>
<td>SGI</td>
</tr>
<tr>
<td>Cray X1</td>
<td>Vector</td>
<td>4</td>
<td>0.800</td>
<td>12.8</td>
<td>Proprietary</td>
<td>4D-Hypercube</td>
<td>UNICOS</td>
<td>NASA (USA)</td>
<td>Cray</td>
<td>Cray</td>
</tr>
<tr>
<td>Cray Opteron Cluster</td>
<td>Scalar</td>
<td>2</td>
<td>2.0</td>
<td>8.0</td>
<td>Myrinet</td>
<td>Fat-tree</td>
<td>Linux (Redhat)</td>
<td>NASA (USA)</td>
<td>AMD</td>
<td>Cray</td>
</tr>
<tr>
<td>Dell Xeon Cluster</td>
<td>Scalar</td>
<td>2</td>
<td>3.6</td>
<td>14.4</td>
<td>InfiniBand</td>
<td>Fat-tree</td>
<td>Linux (Redhat)</td>
<td>NCSA (USA)</td>
<td>Intel</td>
<td>Dell</td>
</tr>
<tr>
<td>NEC SX-8</td>
<td>Vector</td>
<td>8</td>
<td>2.0</td>
<td>16.0</td>
<td>IXS</td>
<td>Multi-stage Crossbar</td>
<td>Super-UX</td>
<td>HLRS (Germany)</td>
<td>NEC</td>
<td>NEC</td>
</tr>
</tbody>
</table>
HPC Challenge Benchmarks

- Basically consists of 7 benchmarks
  - **HPL**: floating-point execution rate for solving a linear system of equations
  - **DGEMM**: floating-point execution rate of double precision real matrix-matrix multiplication
  - **STREAM**: sustainable memory bandwidth
  - **PTRANS**: transfer rate for large data arrays from memory (total network communications capacity)
  - **RandomAccess**: rate of random memory integer updates (GUPS)
  - **FFTE**: floating-point execution rate of double-precision complex 1D discrete FFT
  - **Latency/Bandwidth**: ping-pong, random & natural ring
HPC Challenge Benchmarks

- Top500: solves a system
  \[ Ax = b \]
- STREAM: vector operations
  \[ A = B + s \times C \]
- FFT: 1D Fast Fourier Transform
  \[ Z = FFT(X) \]
- RandomAccess: random updates
  \[ T(i) = XOR(T(i), r) \]

HPCS program has developed a new suite of benchmarks (HPC Challenge)
- Each benchmark focuses on a different part of the memory hierarchy
- HPCS program performance targets will flatten the memory hierarchy, improve real application performance, and make programming easier
Spatial and Temporal Locality

- Programs can be decomposed into memory reference patterns.
- Stride is the distance between memory references.
  - Programs with small strides have high “Spatial Locality”.
- Reuse is the number of operations performed on each reference.
  - Programs with large reuse have high “Temporal Locality”.
- Can measure in real programs and correlate with HPC Challenge.
Spatial/Temporal Locality Results

- HPC Challenge bounds real applications
- Allows us to map between applications and benchmarks
1. **Barrier:** A barrier function `MPI_Barrier` is used to synchronize all processes.

2. **Reduction:** Each processor provides $A$ numbers. The global result, stored at the root processor is also $A$ numbers. The number $A[i]$ is the results of all the $A[i]$ from the $N$ processors.

3. **All_reduce:** `MPI_Allreduce` is similar to `MPI_Reduce` except that all members of the communicator group receive the reduced result.

4. **Reduce scatter:** The outcome of this operation is the same as an MPI Reduce operation followed by an MPI Scatter

5. **Allgather:** All the processes in the communicator receive the result, not only the root
Intel MPI Benchmarks Used

1. **Allgatherv**: it is vector variant of MPI_ALLgather.
2. **All_to_All**: Every process inputs $A*N$ bytes and receives $A*N$ bytes ($A$ bytes for each process), where $N$ is number of processes.
3. **Send_recv**: Here each process sends a message to the right and receives from the left in the chain.
4. **Exchange**: Here process exchanges data with both left and right in the chain
5. **Broadcast**: Broadcast from one processor to all members of the communicator.
Accumulated Random Ring BW vs HPL Performance
Accumulated Random Ring BW vs HPL Performance

- NEC SX8
  - 32-576 cpus
- Cray Opteron
  - 4-64 cpus
- SGI Altix Numalink3
  - 64-440 cpus
- SGI Altix Numalink4
  - 64-2024 cpus
Accumulated EP Stream Copy vs HPL Performance

- NEC SX8 32-576 cpus
- Cray Opteron 4-64 cpus
- SGI Altix Numalink3 64-440 cpus
- SGI Altix Numalink4 64-2024 cpus
Accumulated EP Stream Copy vs HPL Performance

- NEC SX8 32-576 cpus
- Cray Opteron 4-64 cpus
- SGI Altix Numalink3 64-440 cpus
- SGI Altix Numalink4 64-2024 cpus
## Normalized Values of HPCC Benchmark

<table>
<thead>
<tr>
<th>Ratio</th>
<th>Maximum value</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-HPL</td>
<td>8.729 TF/s</td>
</tr>
<tr>
<td>G-EP DGEMM/G-HPL</td>
<td>1.925</td>
</tr>
<tr>
<td>G-FFTE/G-HPL</td>
<td>0.020</td>
</tr>
<tr>
<td>G-Ptrans/G-HPL</td>
<td>0.039 B/F</td>
</tr>
<tr>
<td>G-StreamCopy/G-HPL</td>
<td>2.893 B/F</td>
</tr>
<tr>
<td>RandRingBW/PP-HPL</td>
<td>0.094 B/F</td>
</tr>
<tr>
<td>1/RandRingLatency</td>
<td>0.197 1/μs</td>
</tr>
<tr>
<td>G-RandomAccess/G-HPL</td>
<td>4.9e-5 Update/F</td>
</tr>
</tbody>
</table>
HPCC Benchmarks Normalized with HPL Value

- NEC SX-8
- Cray Opteron
- SGI Altix Numalink3
- SGI Altix Numalink4

Graph showing normalized performance of various benchmarks across different systems, with a focus on HPL.
Barrier Benchmark

![Graph showing the performance of different processors in a barrier benchmark.](image)

- SGI Altix BX2
- Cray Opteron Cluster
- Cray X1 (SSP)
- Cray X1 (MSP)
- NEC SX-8
- Xeon Cluster

![Barrier Benchmark Chart](image)

- Time (in msec)
- Number of Processors

![Barrier Benchmark Chart](image)
1 MB Reduction

![Graph showing time in milliseconds versus number of processors for different computer systems. The graph includes SGI Altix BX2, Cray Opteron Cluster, Cray X1 (SSP), Cray X1 (MSP), NEC SX-8, and Xeon Cluster.]
1 MB Allreduce

Number of Processors vs. Time (in msec)

- SGI Altix BX2
- Cray Opteron Cluster
- Cray X1 (SSP)
- Cray X1 (MSP)
- NEC SX-8
- Xeon Cluster
1 MB Reduction_scatter

![1 MB Reduction scatter graph](image)

- SGI Altix BX2
- Cray Opteron Cluster
- Cray X1 (SSP)
- Cray X1 (MSP)
- NEC SX-8
- Xeon Cluster
1 MB Allgatherv

![Graph showing the performance of different processors for 1 MB Allgatherv.](image)

- **Time (in msec)**: The y-axis measures time in milliseconds, ranging from $10^0$ to $10^7$.
- **Number of Processors**: The x-axis represents the number of processors, ranging from 1 to 1000.
- **Legend**:
  - SGI Altix BX2
  - Cray Opteron Cluster
  - Cray X1 (SSP)
  - Cray X1 (MSP)
  - NEC SX-8
  - Xeon Cluster

The graph illustrates how the time taken for 1 MB Allgatherv scales with the number of processors for different systems.
1 MB All_to_All

![Graph showing the time in msec for different number of processors for various systems.](image)

- **SGI Altix BX2**
- **Cray Opteron Cluster**
- **Cray X1 (SSP)**
- **Cray X1 (MSP)**
- **NEC SX-8**
- **Xeon Cluster**
1 MB SendRecv

Bandwidth (in MB/seconds)

1.0 × 10^5

1.0 × 10^4

1.0 × 10^3

1.0 × 10^2

1.0 × 10^1

1.0 × 10^0

Number of Processors

1

10

100

1000

SGI Altix BX2

Cray Opteron Cluster

Cray X1 (SSP)

Cray X1 (MSP)

NEC SX-8

Xeon Cluster
1 MB Exchange

Bandwidth (in MB/seconds)

Number of Processors

- SGI Altix BX2
- Cray Opteron Cluster
- Cray X1 (SSP)
- Cray X1 (MSP)
- NEC SX-8
- Xeon Cluster
1 MB Broadcast

Time (in msec)

Number of Processors

1.E+00 1.E+01 1.E+02 1.E+03 1.E+04 1.E+05 1.E+06

1 10 100 1000

SGI Altix BX2
Cray Opteron Cluster
Cray X1 (SSP)
Cray X1 (MSP)
NEC SX-8
Xeon Cluster
Summary

- Performance of vector systems is consistently better than all the scalar systems
- Performance of SX-8 is better than Cray X1
- Performance of SGI Altix BX2 is better than Dell Xeon cluster and Cray Opteron cluster
- IXS (SX-8) > Cray X1 network > SGI Altix BX2 (NL4) > Dell Xeon cluster (IB) > Cray Opteron cluster (Myrinet).