Design and evaluation of a TOP100 Linux Cluster

The HPC2N Super Cluster: From Bits and Pieces to a Benchmarked TOP100 Supercomputer System

Robert Granat and Erik Elmroth
High Performance Computing Center North (HPC2N)
Umeå University

Outline

• System overview
• System design and the assembly line
• Benchmark results
  - Compiler test
  - STREAM
  - Pallas MPI Benchmark
  - DGEMM
  - HP Linpack
  - NAS Parallel
• Conclusions
• Acknowledgements

HPC2N Super Cluster - seth.hpc2n.umu.se

• 240 processors in 120 dual nodes
• Nodes interconnected in a 4 x 5 x 6 mesh with "wrap around"
• Self-made
• Rack mounted - 12 racks
• Built in April-May 2002
• Available to users since June 2002
• 83% utilization (24-7) since mid August
• Total peak performance 800 Gflops/s
• On Top500 of June ’02:
  - 94th fastest in the world
  - The fastest in Sweden
• Financed by 5 MSEK Grant from the Kempe foundations

Node Overview

• 2 AMD Athlon MP 2000+ processors per node
  - 1.667 GHz
  - Super-scalar, fully pipelined floating point engine, performing 2 floating point operations per cycle
  - Hardware support for data prefetch and speculative TLBs
  - 384 KB on-chip cache (128 in L1 & 256 in L2)
• 266 MHz system bus: 2.1 Gbytes/sec memory bandwidth
• 66 MHz/64-bit PCI-bus - 528 Mbytes/sec towards SCI network
• Tyan Tiger MPX motherboard (AMD-760 chipset)
• 1(-3) Gbyte of DDR memory
• Serial console
• Fast ethernet
• SCI network card
• 6.7 Gflops/s per node (peak)
Design and evaluation of a TOP100 Linux Cluster

SCI Network
- Wulfkit3 SCI network by Dolphin IS, Norway
- Nodes connected in a $4 \times 5 \times 6$ mesh with "wrap around"
- 667 Mbytes/s bandwidth
- 1.46 µsec latency (peak)
- ScaMPI message passing library (MPI) by Scali AS
- Software for system monitoring and management

The Assembly Line
1. Unpacking 800-1000 boxes of computer hardware parts
2. Mounting power supplies, hard disks. Perform cable and sheet-metal work
3. Static-sensitive assembly: mounting processors, memory, cooling etc on motherboards. Installing boards into cases
4. Node testing
5. Installing SCI cards
6. Racks were prepared and rack-rails attached to the nodes, nodes installed in the racks
7. Electrical installations, including one PDU per frame
8. Cables for fast Ethernet and serial console connected
9. Cables for the SCI network connected in a $4 \times 5 \times 6$ mesh

After 10 000 screws and 700 cables, the system was completed!

General Software
- Linux 2.4.x
- Debian 3.0
- ATLAS
- LAPACK
- ScaLAPACK
- ScaMPI
- PGI Compiler suite (hpf, f77/90, pgcc)
- Intel compilers (ifc, icc)
- Openpbs with Maui scheduler

Compilers and Performance
- Various compilers are available for i86 linux systems:
  - GNU Compiler Collection (C, C++ and Fortran 77)
  - Portland Group Compilers (C, C++ and Fortran 95)
  - Intel Compilers (C, C++ and Fortran 95)
  - Absoft Pro Fortran (Fortran 95)
- Benchmarked on the serial version of the NAS Parallel Benchmarks

- Benchmark results for the LU application (class W):
  - Intel Fortran: 274 Mop/s
  - Portland Group Fortran: 252 Mop/s
  - Absoft Pro Fortran: 249 Mop/s
  - GNU Fortran (g77-3.1): 240 Mop/s
- These results are typical for what we found
Design and evaluation of a TOP100 Linux Cluster

STREAM Benchmark

- Sustainable memory bandwidth in Mbytes/second
  - 4 vector operations:
    - Copy: $a(i) = b(i)$
    - Scale: $a(i) = q * b(i)$
    - Add: $a(i) = b(i) + c(i)$
    - Triad: $a(i) = b(i) * q * c(i)$
  - 1 CPU: 651 – 769 MB/s
  - 2nd CPU: Additional 17 – 45% (up to 964 MB/s)

Network Performance

- PALLAS MPI Benchmarks
  - Maximum MPI Bandwidth (Ping-Pong): 230 Mbytes/s
  - Minimum MPI latency: 3.7 µsec
  - MPI Barrier for 2 processes: 5.3 µsec

- Compare with Myrinet on a dual P4 XEON system
  - Maximum MPI Bandwidth (Ping-Pong): 195 Mbytes/s
  - Minimum MPI Latency (Ping-Pong): 8.8 µsec
  - MPI Barrier for 2 processes: 16.5 µsec

Pallas MPI Benchmark

Atlas DGEMM Performance
**HP Linpack**

- Solving a linear system of equations
- 480.7 Gflops/s for lin.sys. with 116100 unknowns
- Excellent scaling:
  - 60.1% of the theoretical peak performance
  - 75.2% of accumulated processor Dgemm performance
  - 76.9% of accumulated node Dgemm performance
  - 110 times faster than single node HP Linpack

**NAS Parallel Benchmarks (V 2.2)**

- Contains 8 CFD related benchmark codes
  - 3 classes of test problems (different sizes A, B, and C), with sparse matrices (typically with 5 x 5 blocks)
  - Codes in Fortran using MPI
  - We chose not to modify the codes
  - Codes are not well optimized (e.g., for memory hierarchies)

- Today, we present results for
  - BT: Multiple systems of 5 x 5 block tridiagonal equations
  - CG: Conjugate Gradient for solving an unstructured sparse linear system
  - LU: 5 x 5 block lower and upper triangular system solution (from Navier-Stokes)

---

**On the TOP500 List of June 2002**

- 94th fastest computer in the world
- The fastest computer in Sweden
- The world’s 5th fastest “self-made” system (2 Japan, US, Russia)
- 2nd fastest computer in the Nordic countries (Helsinki)
- The world’s fastest computer with Dolphin interconnect
- The world’s 3rd fastest computer with AMD processors (Japan, Germany)

---

... for Cray T3E comparisons ...

- Piles of PC’s give high peak performance at low cost!
- But do we get true supercomputer performance?

- Peak performance ratios (per CPU)
  - AMD MP2000+/Cray T3E-900: 3.71
  - AMD MP2000+/Cray T3E-1200: 2.78

- By scaling benchmark results for the Crays, we can evaluate the cluster’s ability to make use of these cycles in practice!
Design and evaluation of a TOP100 Linux Cluster

**BT**
- Single and dual CPU results
- Performance per processor
- Low effect of 2nd CPU
- Very good scalability

**LU**
- Good effect of 2nd CPU
- Very good scalability
- Extreme cache effects - super linear speedup for large number of processors

**CG**
- Good effect of 2nd CPU
- Good scalability (for large problems)

**Conclusions**
- Extremely cost-effective (cheapest ever on TOP100?)
- True supercomputer performance (not only peak numbers)
- The network does not limit the parallel performance (i.e., bandwidth and latency "good enough")
- With single-CPU nodes: performance similar or better than "frequency-scaled" T3E-numbers
- With dual-CPU nodes: performance from "low" to 98% depending on memory reference patterns (blocking for memory hierarchies as important as ever)
- Quickly accepted by our old and new users!
Many Thanks To...

- **HPC2N personnel**
  - Niklas Edmundsson
  - Bo Kågström
  - Markus Mårtensson
  - Mats Nylén
  - Roger Oscarsson
  - Åke Sandgren
  - Mattias Wadenstein

- **Umdac (for procurement)**
  - Tomas Eklund
  - Jörgen Stenström
  - Lars Wänglund

- **...also on the assembly line...**
  - Lars Bäckström
  - Reine Edmundsson
  - Simon Rönnberg
  - Jonas Sandberg
  - Peter Toneby
  - Tomas Ögren

- **Husservice MIT**
  - Tommy Ericsson
  - Roland Johansson

- **Others (who couldn’t stay away)**
  - Magnus Janlert
  - Helena Lindblom
  - Iradj Rouzbeh

Design and evaluation of a TOP100 Linux Cluster