Scalability and Performance of CTH on the Computational Plant

Ron Brightwell
Scalable Computing Systems

H. Eliot Fang
Materials and Process Sciences

Lee Ward
Scientific Computing Systems

Sandia National Laboratories
Outline

• Evolution of Cplant™
• Approach
• Hardware
• Run-time environment
• Platforms
• CTH
• Performance results
Massively Parallel Processors (MPPs)

- Intel Paragon
  - 1,890 compute nodes
  - 3,680 i860 processors
  - 143/184 GFLOPS
  - 175 MB/sec network
  - SUNMOS lightweight kernel

- Intel TeraFLOPS
  - 4,576 compute nodes
  - 9,472 Pentium II processors
  - 2.38/3.21 TFLOPS
  - 400 MB/sec network
  - Puma/Cougar lightweight kernel
Cplant Strategy

• Hybrid approach combining commodity cluster technology with MPP technology
• Build on the design of the TFLOPS:
  – large systems should be built from independent building blocks
  – large systems should be partitioned to provide specialized functionality
  – large systems should have significant resources dedicated to system maintenance
Cplant Approach

- Emulate the ASCI Red environment
  - Partition model (functional decomposition)
  - Space sharing (reduce turnaround time)
  - Scalable services (allocator, loader, launcher)
  - Ephemeral user environment
  - Complete resource dedication

- Use Existing Software when possible
  - Red Hat distribution, Linux/Alpha
  - Software developed for ASCI Red
Conceptual Partition Model
Phase II - Alaska

- 272 Digital PWS 500a (Miata)
- 500 MHz Alpha 21164 CPU
- 2 MB L3 Cache
- 192 MB ECC SDRAM
- 16-port Myrinet SAN/LAN switch
- 32-bit, 33 MHz LANai-4 NIC
- 1 DEC AS4100 compile & user file server
Phase III - Siberia

- 592 Compaq XP1000 (Monet)
- 500 MHz Alpha 21264 CPU
- 4 MB L3 Cache
- 256 MB ECC SDRAM
- 16-port Myrinet SAN/LAN switch
- 64-bit, 33 MHz LANai-7 NIC
Portals

- Data movement layer from SUNMOS and PUMA
- Zero-copy, application bypass mechanism on MPPs
- Flexible building blocks for supporting many protocols
- Elementary constructs that support MPI semantics well
- Linux kernel module that interfaces to a transport layer
  - Ethernet, Myrinet, any Linux network device
Runtime Environment

- yod - Service node parallel job launcher
- bebopd - Compute node allocator
- PCT - Process control thread, compute node daemon
- fyod - Independent parallel I/O
- ENFS – parallel I/O
Runtime Environment (cont’d)

• Yod
  – Contacts compute node allocator
  – Launches the application into the compute partition
  – Redirects all application I/O (stdio, file I/O)
  – Makes any filesystem visible in the service partition visible to the application
  – Redirects any UNIX signals to compute node processes
  – Allows user to choose specific compute nodes
  – Can launch multiple (up to 5) different binaries
• PCT
  – Contacts bebopd to join compute partition
  – Forms a spanning tree with other PCT’s to fan out the executable, shell environment, signals, etc.
  – *fork()*’s, *exec()*’s, and monitors status of child process
  – Cleans up a parallel job
  – Provides a back trace for process faults
Runtime Environment (cont’d)

• Bebopd
  – Accepts requests from PCT’s to join the compute partition
  – Accepts requests from yod for compute nodes
  – Accepts requests from pingd for status of compute nodes
  – Allows for multiple compute partitions
Fyod Parallel Independent I/O

- File yod
- Daemon process
- Runs on nodes in the file I/O partition
- Each compute process manipulates a single file
ENFS Parallel I/O

• Compute nodes bind to an ENFS node at boot time
• Does not maintain coherency or support locking
• Extends NFS protocol to support application control of data caches
• Only NFS calls are used externally
• Performance gain comes from the ability to have a very large number of transactions simultaneously in flight against the external server
• Each compute node can open the same file and coordinate access
• Or use separate files as CTH does
Platforms

- Intel TFLOPS – ASCI Red
  - 9000+ 300 MHz Pentium II Xeon
  - 256 MB per 2-cpu node
  - 400 MB/s network
  - PFS
- IBM SP-2 – ASCI Blue/Pacific
  - 1280 332 MHz PowerPC 604e
  - 1.5 GB per 4-cpu node
  - 150 MB/s network
  - GPFS
- DEC 8400 Cluster
  - 84 622 MHz Alpha 21164
  - 4 GB per 12-cpu node
  - Memory Channel II interface
  - 2 GB local disk
CTH Family of Codes

- Models complex multi-dimensional, multi-material problems characterized by large deformations and/or strong shocks
- Uses two-step, second-order accurate finite-difference Eulerian solution
- Material models for equations of state, strength, fracture, porosity, and high explosives
- Impact, penetration, perforation, shock compression, high explosive initiation and detonation problems
CTH Steps

• CTHGEN
  – Problem setup
    • Create computational mesh, insert materials, calculate volume fraction of each material in cells
  – Assign material properties and run-time controls
    • Broadcasting data is main type of message passing
  – Generate initial restart file, one file per node

• CTH
  – Read initial restart file, one file per node
  – Simulate shock wave physics
    • Many nearest-neighbor communications, a few global reductions per time step
  – Write results to restart, history, and viz files
  – Performance measured in grind time
    • Time to compute all calculations on a single cell for a single time step
Sample Problem

• Two gasses sliding on each other
• Simple enough for a quick performance and scaling check
• All cells are activated throughout the calculation to maintain load balancing
• Restart files average 2 MB
• CTH consumes ~ 42 MB per CPU
Tru64 UNIX/Linux NFS Problem

- Tru64 employs delayed writes in order to bundle many NFS transactions
- Linux 2.2.x doesn’t post I/O request until a timer expires
- For read-ahead and write-behind clients, the I/O request would be fully utilized and the request would get queued before timer expiration
- This doesn’t happen with Linux NFS v2 and 8KB packets
- The additional latency slows I/O requests by as much as two orders of magnitude
CTH Read Time

Number of Nodes

Time (seconds)

Alaska
Siberia
Tflops
DEC
Blue-Pacific
CTH Grind Time

Grind Time (microseconds) vs. Number of Nodes

- Alaska
- Siberia
- Tflops
- DEC
- Blue-Pacific
Phase IV - Antarctica?

- ~1350 Compaq DS-10 Slates
- 466 MHz Alpha 21264
- 256 MB ECC SDRAM
- 64-port Myrinet SAN/LAN switch
- 64-bit, 33 MHz LANai-7 NIC
- Myrinet 33MHz 64bit LANai 9.x

- To be combined with Siberia for a ~1600-node system
- Red, black, green switchable