



Cplant: World's Fastest Linux Cluster

Ron Brightwell

Sandia National Laboratories

Scalable Computing Systems

9223

<http://www.cs.sandia.gov/cplant>



Outline

- **Evolution of Cplant**
- **Approach**
- **Hardware**
- **Environment**
- **Status and Plans**



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





Sandia/Univ. of New Mexico OS (SUNMOS)

- **Goals**
 - Develop a high performance compute node OS for distributed memory MPPs
 - Deliver as much performance as possible to apps
 - Small footprint
- **Project started in January 1991 on the nCUBE-2 to explore new message passing schemes and high-performance I/O**
- **Ported to Intel Paragon in Spring of 1993**
- **Used to set world speed record in 1993 (281 GFLOPS on MPLINPACK)**



Puma

- **Successor to SUNMOS**
 - **Added multiprocess support**
 - **Modularized (kernel, PCT)**
- **Developed on nCUBE-2 in 1993**
- **Ported to Intel Paragon in 1995**
- **Ported to Intel TFLOPS in 1996 (Cougar)**
- **Portals 1.0 communication layer (Puma on nCUBE)**
 - **User-/Kernel-managed message buffers**
- **Portals 2.0 communication layer (Puma on Paragon)**
 - **Avoid excessive buffering and memory copies**
- **Used to set the current world speed record (2.38 TFLOPS on MPLINPACK)**



Disadvantages of MPPs

- **Custom hardware is quickly superseded by commodity components**
- **Volume vendors are not the best organizations to create niche products**
- **Scalability requires specialized knowledge and research**
- **Lengthy procurement cycle**



Commodity Clusters

- **Berkeley NOW**
 - 100+ SUN SparcStations
 - Myrinet
 - GLUNIX
- **Beowulf**
 - Commodity PC's
 - Linux
 - Fast Ethernet
- **HPVM**
 - Windows NT workstations
 - Myrinet

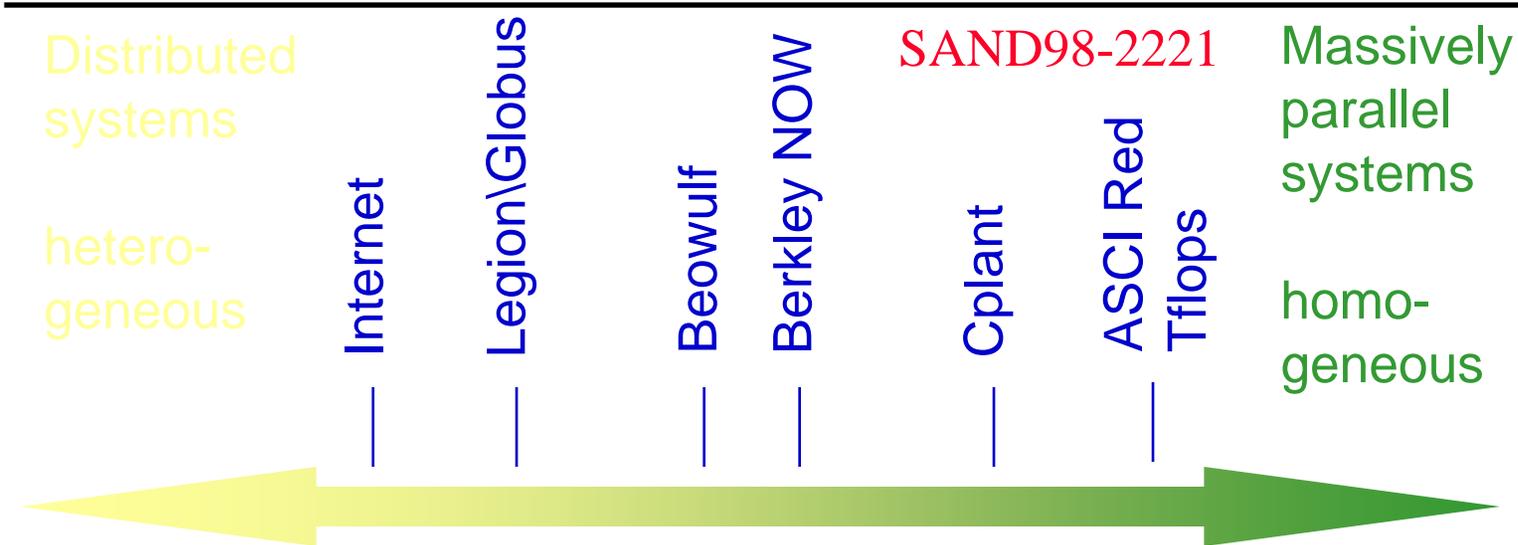


Disadvantages of Clusters

- Performance is not scalable
- Management is not scalable
- Limited user environment
- Not production quality



Distributed & Parallel Systems



- Gather (unused) resources
- Steal cycles
- System SW manages resources
- System SW adds value
- 10% - 20% overhead is OK
- Resources drive applications
- Time to completion is not critical
- Time-shared

- Bounded set of resources
- Apps grow to consume all cycles
- Application manages resources
- System SW gets in the way
- 5% overhead is maximum
- Apps drive purchase of equipment
- Real-time constraints
- Space-shared



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 Metaphors

- **Provide compute cycles like a power plant provides electricity**
- **Grow and prune the plant on a three year cycle**
 - **add the latest hardware every year**
 - **remove the obsolete hardware every year (starting in the fourth year)**



Why Cplant?

- **Modeling and simulation, essential to stockpile stewardship, require significant computing power**
- **Commercial supercomputer are a dying breed**
- **Pooling of SMPs is expensive and more complex**
- **Commodity PC market is closing the performance gap**
- **WEB services and e-commerce are driving high-performance interconnect technology**



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

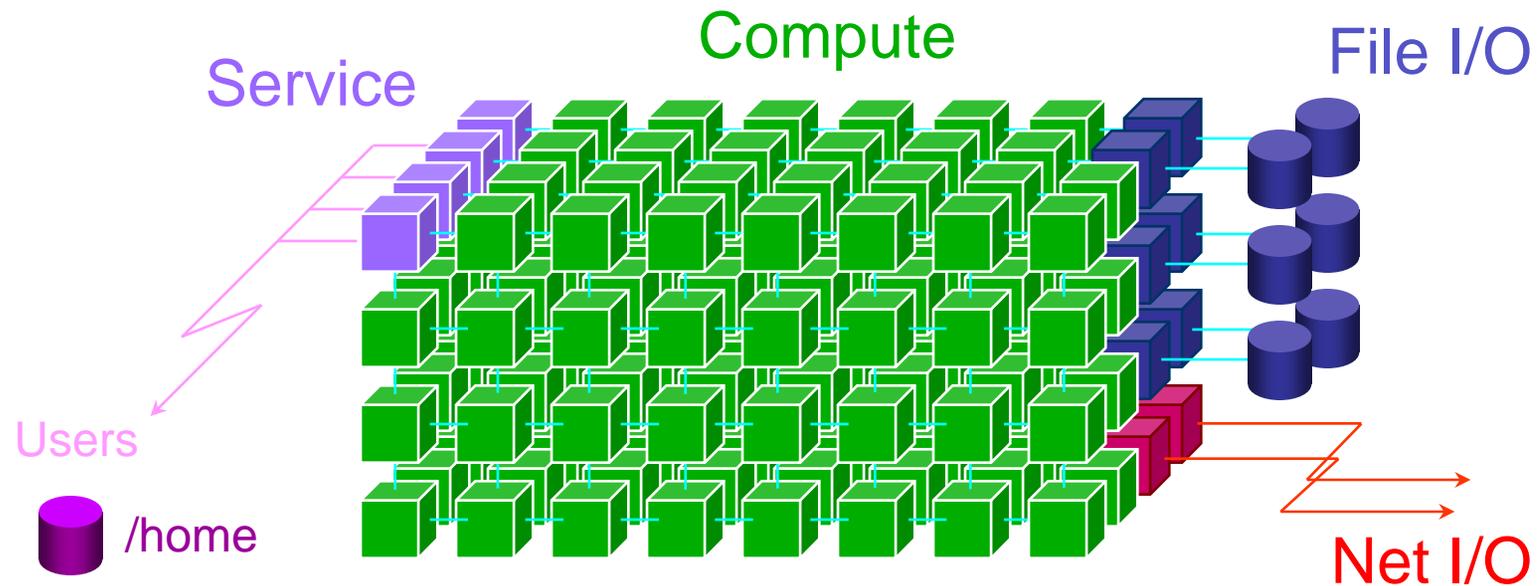


Portals

- **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**

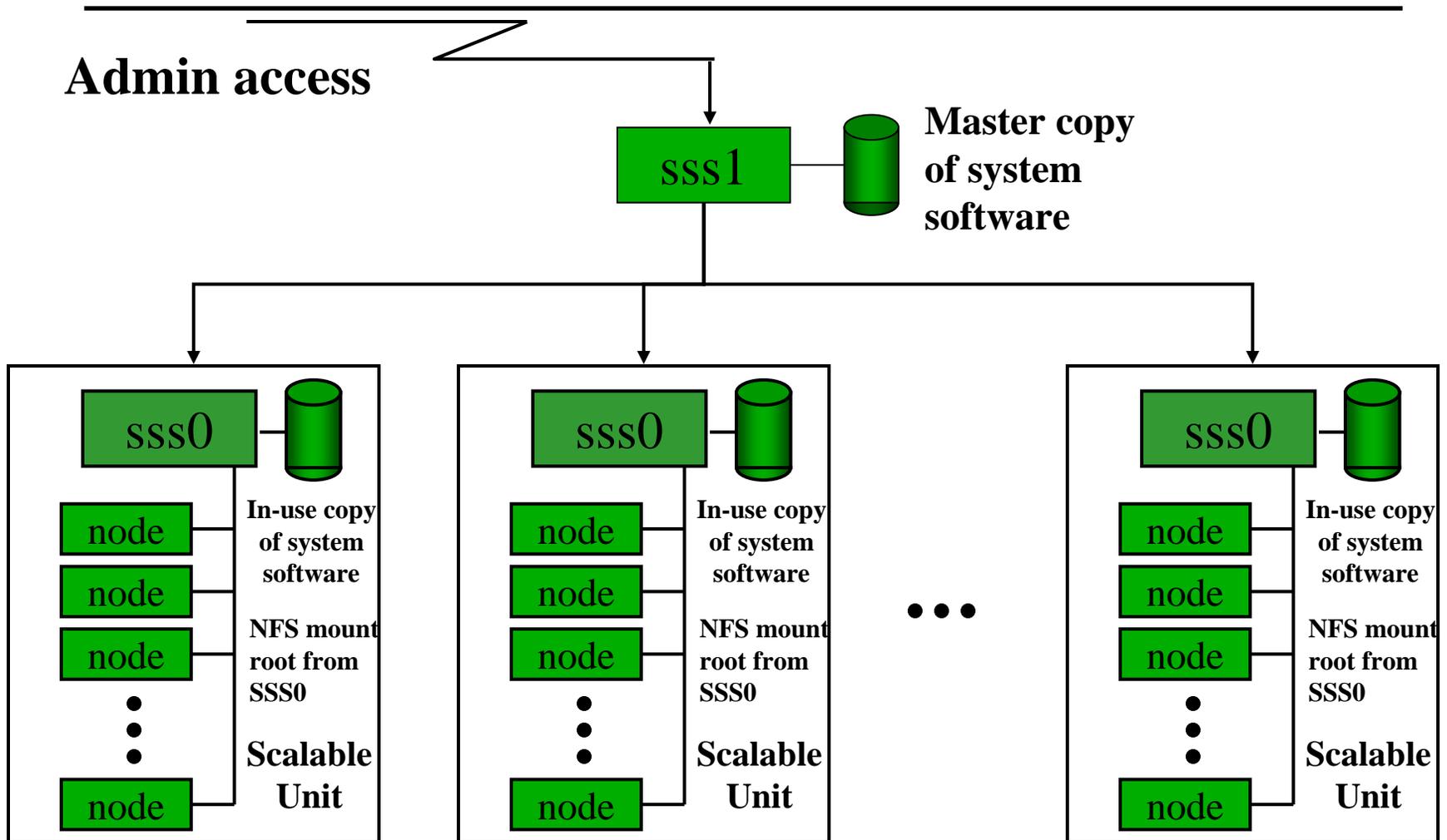


Conceptual Partition Model



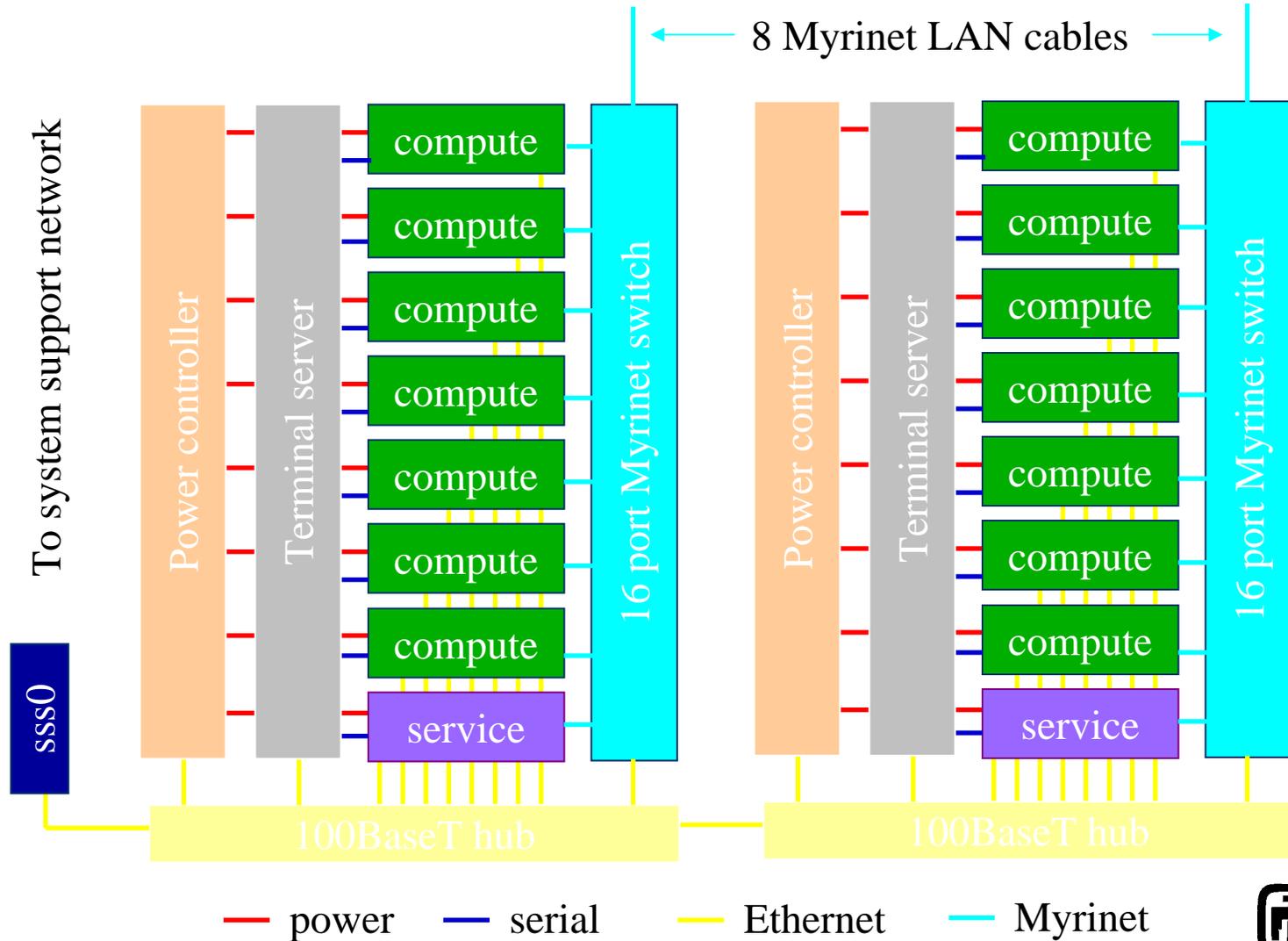


System Support Hierarchy





Scalable Unit



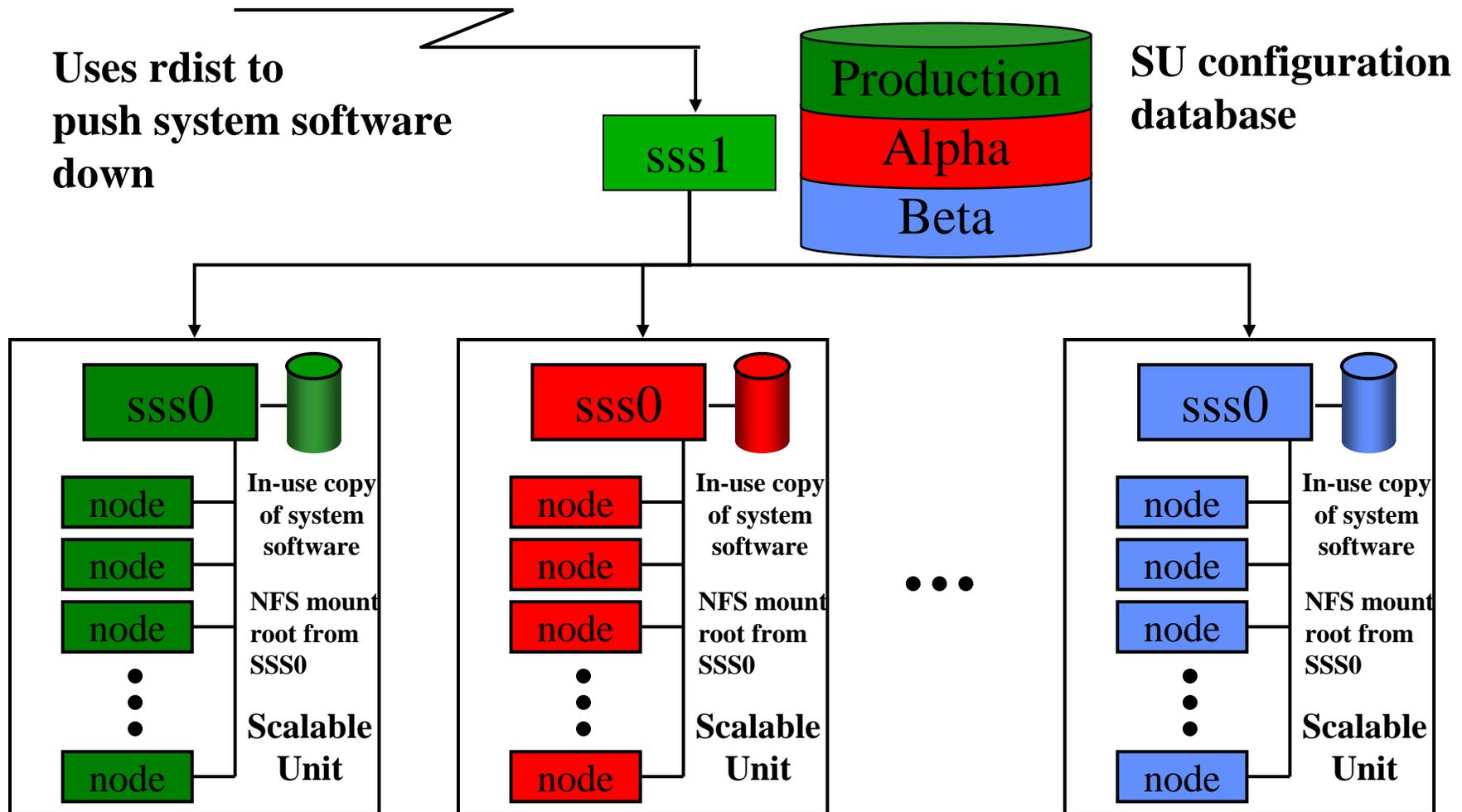


Hardware Management

- **Hardware discovery**
 - Updates database with MAC addresses and other information that can be gathered automatically
- **Database contains rules for node types**
 - Service, support, compute nodes boot different kernels and need different modules
- **Machine partitioning**
 - Support for heterogeneous nodes & sandboxes



“Virtual Machines”





Phase I - Prototype (Hawaii)

- 128 Digital PWS 433a (Miata)
- 433 MHz 21164 Alpha CPU
- 2 MB L3 Cache
- 128 MB ECC SDRAM
- 24 Myrinet dual 8-port SAN switches
- 32-bit, 33 MHz LANai-4 NIC
- Two 8-port serial cards per SSS-0 for console access
- I/O - Six 9 GB disks
- Compile server - 1 DEC PWS 433a
- Integrated by SNL





Phase II - Production (Alaska)

- 400 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
- 6 DEC AS1200, 12 RAID (.75 Tbyte) || file server
- 1 DEC AS4100 compile & user file server
- Integrated by Compaq





Phase III- Production (Siberia)

- 624 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
- 1.73 TB disk I/O
- Integrated by Compaq and Abba Technologies





System Deployment

Current

SNL/NM

- Alpha 21164 nodes - 500 MHz, 192 MB
 - 272 unclassified
 - 96 classified
 - 32 development
- Alpha 21264 nodes - 500 MHz, 256 MB
 - 624 unclassified

SNL/CA

- Alpha 21164 nodes - 433 MHz, 192 MB
 - 128 unclassified
- Alpha 21264 nodes - 500 MHz, 256 MB
 - 128 unclassified

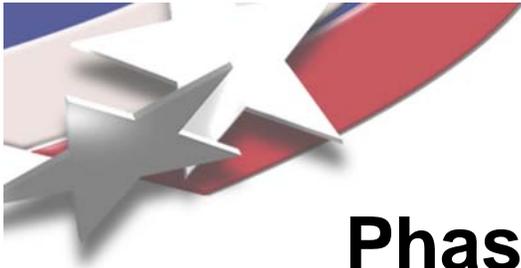
Planned

SNL/NM

- Alpha 21164 nodes - 500 MHz, 192 MB
 - 368 classified
 - 32 development
- Alpha 21264 nodes - 500 MHz, 256 MB
 - 592 unclassified
 - 32 development

SNL/CA

- Alpha 21164 nodes - 433 MHz, 192 MB
 - 128 unclassified
- Alpha 21264 nodes - 500 MHz, 256 MB
 - 128 unclassified

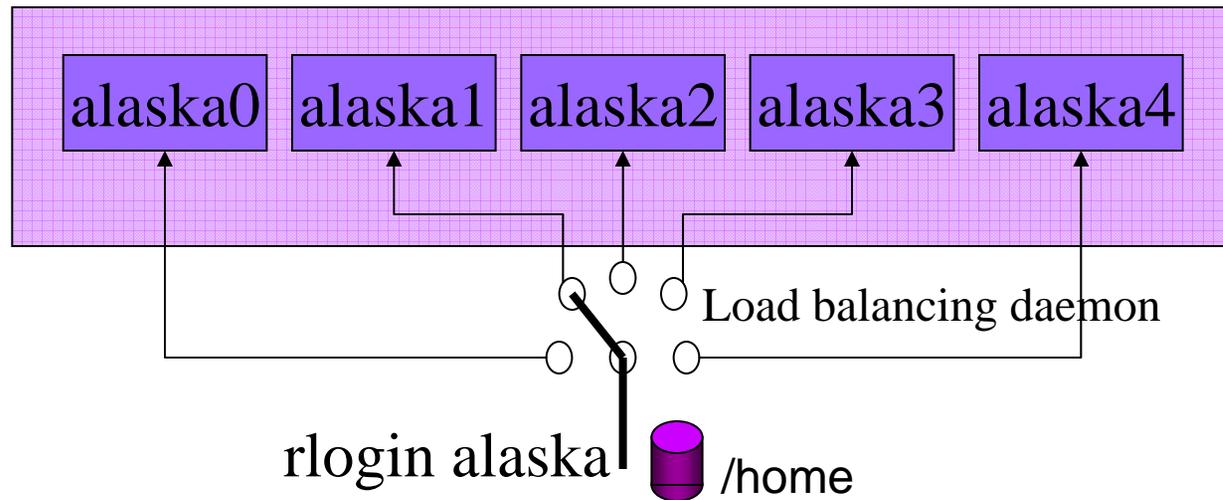
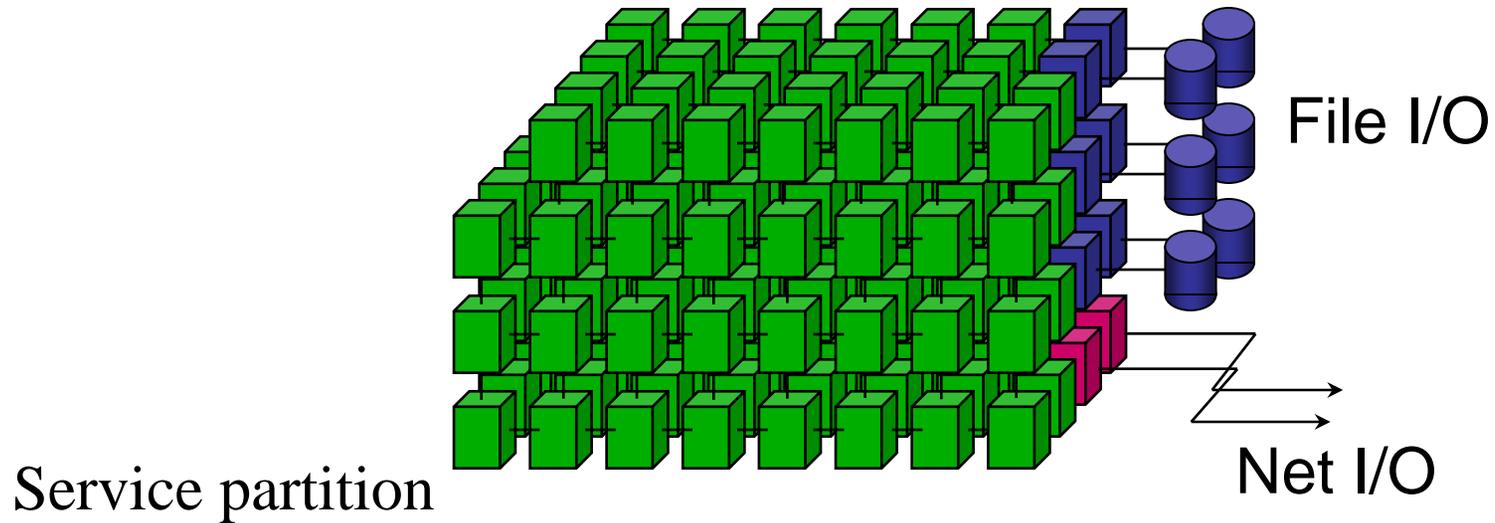


Phase IV – Development (June '00)

- **1024+ Compaq DS-10 (1U Slate)**
- **466 MHz 21264 CPU**
- **256 MB ECC SDRAM**
- **64-port Myrinet SAN/LAN switch**
- **64-bit 33 MHz LANai-7 NIC**
- **Red/Black switching supported**



User View





Compile Environment

- **Shell scripts that call Tru64 UNIX compilers**
- **Support libraries - startup module, PUMA library, redirected I/O and C libraries, TFLOPS MPI library**
- **Centrally located for any Tru64 UNIX machine**
- **Dedicated Tru64 UNIX machine**
- **Works on any Tru64 UNIX machine**
- **Avoids Linux-specific configurations and immature GNU FORTRAN compiler**
- **Compaq Alpha/Linux compilers in future**



Runtime Environment

- **yod - Service node parallel job launcher**
- **bebopd - Compute node allocator**
- **PCT - Process control thread, compute node daemon**
- **pingd - Compute node status tool**
- **fyod - Independent 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**



Runtime Environment (cont'd)

- **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



Runtime Environment (conc'd)

- **Pingd**
 - Displays list of available compute nodes
 - Displays list of compute nodes in use
 - Displays owner, elapsed time of jobs
 - Allows users to kill their jobs
 - Allows administrators to kill jobs and free up specific nodes
 - Allows administrators to remove nodes from the compute partition
- **Showmesh**
 - Massages pingd output into TFLOPS-like showmesh



Parallel I/O

- **Fyod**
 - Runs on nodes in the file I/O partition
 - Parallel independent file I/O
 - Each compute process opens a single file
- **Third party solution**
 - Use I/O nodes as proxies
 - Use a third party filesystem (currently SGI's CXFS)
 - CTH Performance
 - 64 compute nodes -> 4 I/O proxies -> 1 SGI O2K
 - 38 MB/s writing, 30 MB/s reading



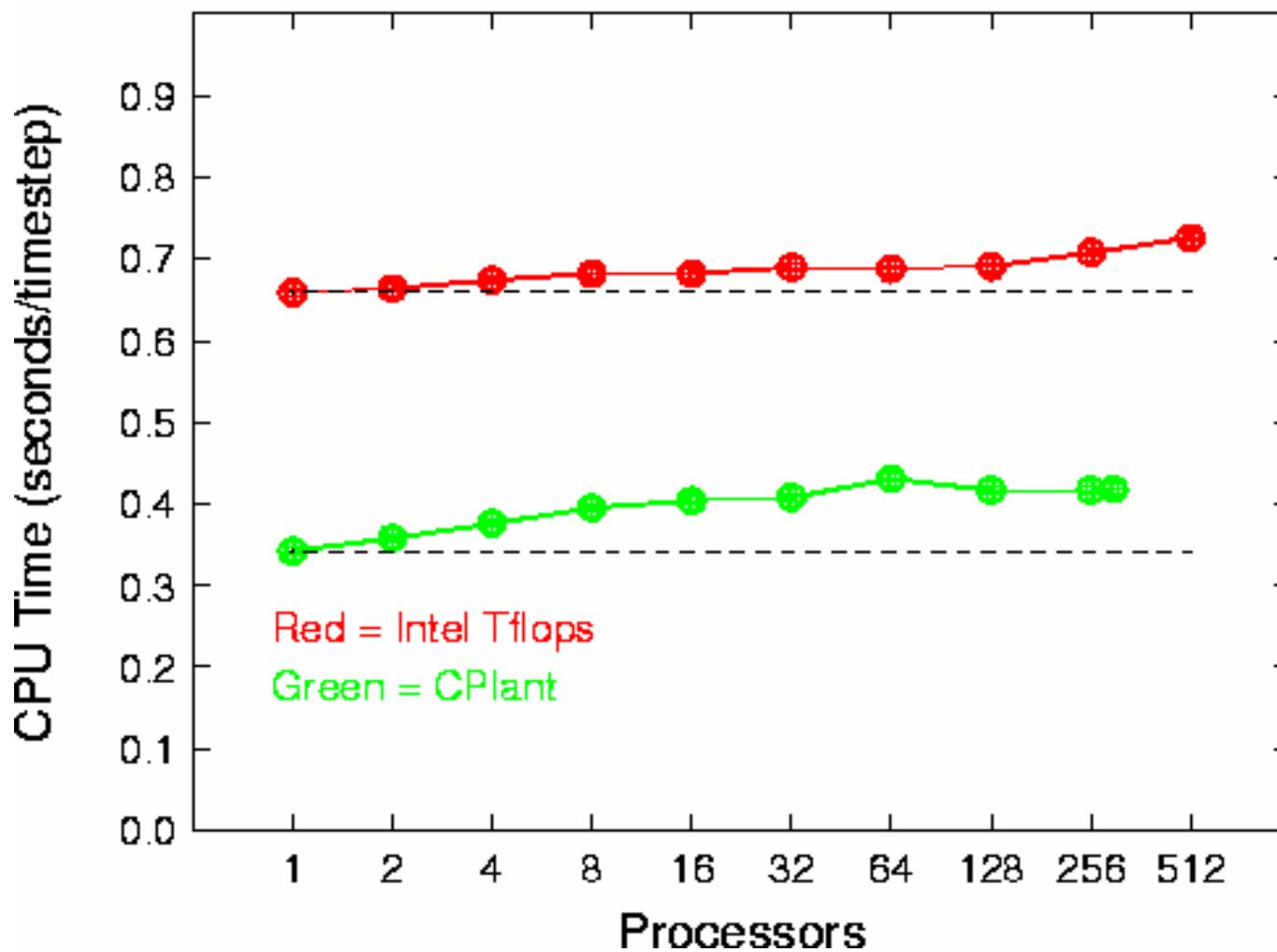
Cplant Performance

- **350 Alaska nodes**
 - 125.2 GFLOPS on MPLINPACK
 - would place 53rd on June 1999 Top 500
- **572 Siberia Nodes**
 - 247.6 GFLOPS on MPLINPACK
 - would place 29th on June 1999 Top 500
- **Comparison of Cplant to TFLOPS:**
 - LAMMPS molecular dynamics benchmark
 - CTH shock physics code



Molecular Dynamics Benchmark

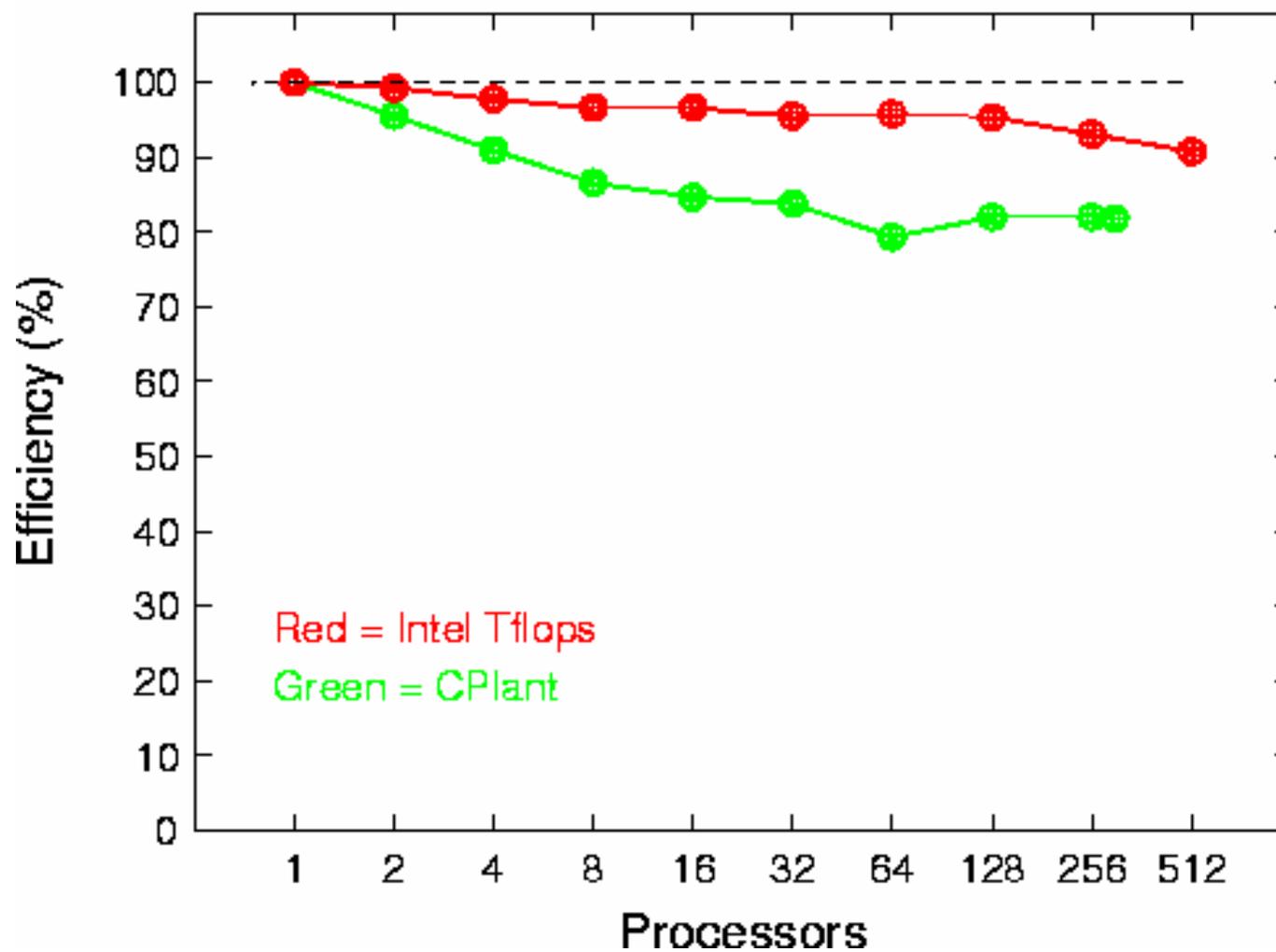
Scaled-Size Performance, N = 32000 atoms/proc





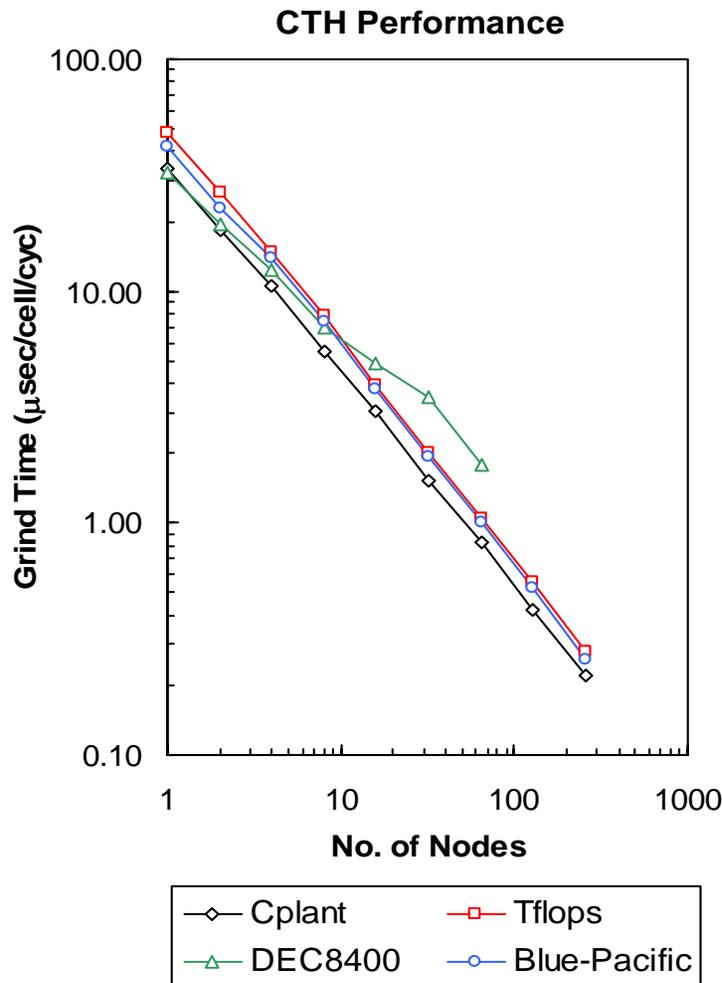
Molecular Dynamics Benchmark

Scaled-Size Efficiency, N = 32000 atoms/proc

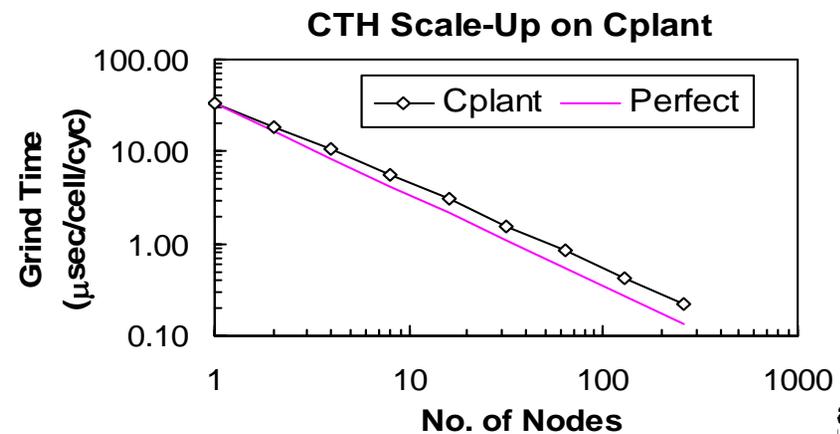




CTH Performance



# of nodes	Grind Time ($\mu\text{sec}/\text{cell}/\text{cyc}$)			
	Cplant	Tflops	DEC8400	Blue-Pacific
1	34.06	48.58	32.41	41.65
2	18.37	26.73	19.33	22.69
4	10.53	14.78	12.34	14.03
8	5.54	7.78	7.02	7.35
16	3.05	3.96	4.91	3.77
32	1.52	2.02	3.47	1.93
64	0.83	1.05	1.78	1.01
128	0.42	0.56		0.52
256	0.22	0.28		0.26





Applications Work In Progress

- **CTH**
 - 3D Eulerian shock physics
- **ALEGRA**
 - 3D arbitrary Lagrangian-Eulerian solid dynamics
- **GILA**
 - Unstructured low-speed flow solver
- **MPQuest**
 - Quantum electronic structures
- **SALVO**
 - 3D seismic imaging
- **LADERA**
 - Dual control volume grand canonical MD simulation
- **Parallel MESA**
 - Parallel OpenGL
- **Xpatch**
 - Electromagnetism
- **RSM/TEMPRA**
 - Weapon safety assessment
- **ITS**
 - Coupled Electron/Photon Monte Carlo Transport
- **TRAMONTO**
 - 3D density functional theory for inhomogeneous fluids
- **CEDAR**
 - Genetic algorithms



Applications Work In Progress

- **AZTEC**
 - Iterative sparse linear solver
- **DAVINCI**
 - 3D charge transport simulation
- **SALINAS**
 - Finite element modal analysis for linear structural dynamics
- **TORTILLA**
 - Mathematical and computational methods for protein folding
- **EIGER**
- **DAKOTA**
 - Analysis kit for optimization
- **PRONTO**
 - Numerical methods for transient solid dynamics
- **SnRAD**
 - Radiation transport solver
- **ZOLTAN**
 - Dynamic load balancing
- **MPSALSA**
 - Numerical methods for simulation of chemically reacting flows

<http://www.cs.sandia.gov/cplant/apps>



Near-Term Plans

- **Performance**
 - Double number of Myrinet switches for Siberia (Node to switch ratio = 4:1)
 - Third party tool for parallel I/O
- **User Tools**
 - TotalView
- **Distributed Resource Management**
 - PBS
 - Globus



Short-Term Plans

- **Linux**
 - physically contiguous memory
 - in-memory `exec()`
- **Intelligent compute node allocator**
- **Parallel filesystem**
- **Portals 3.0**
- **Compaq Linux compilers**



Long-Term Plans

- **Alternate interconnect technologies**
 - Quadrics
 - InfiniBand Trade Association (FIO/NGIO/SIO)
 - Servernet II
 - Giganet
- **IA-64-based clusters**
- **Heterogeneity**
- **Lightweight Linux**



Cplant Contributors

System Software Development

- **Ron Brightwell**
- **Pang Chen**
- **Lee Ann Fisk**
- **Bob Davis**
- **Tramm Hudson**
- **Jeanette Johnston**
- **Jim Otto**
- **Rolf Riesen**
- **Lee Ward**
- **Robert Clay**
- **David Evensky**
- **Pete Wyckoff**

Operations

- **Mike McConkey**
- **Jim Laros**
- **Geoff McGirt**
- **Mike Kurtzer**
- **John Laroco**
- **Rob Armstrong**

Applications

- **20+ applications**
- **see `./cplant/apps`**