Alliance / UNM Roadrunner Linux Cluster

David A. Bader

Department of Electrical and Computer Engineering &

Albuquerque High Performance Computing Center

University of New Mexico

dbader@eece.unm.edu

http://www.eece.unm.edu/~dbader

May 1999

Outline

- History of Clusters
- Recent Developments
- Cluster Architecture and Technology
- Cluster Systems Software
- Computational Grid
- Applications
- Alliance/UNM Roadrunner Cluster

History of Clusters

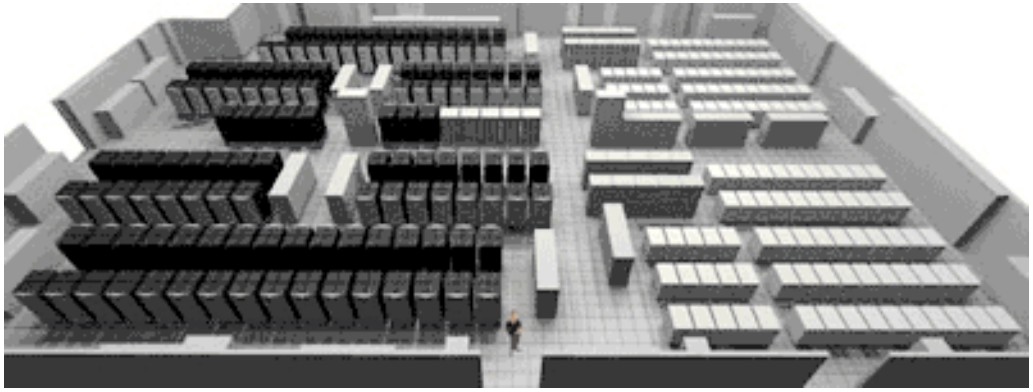
- Commodity microprocessors in supercomputers
  - Thinking Machines CM-5 (SPARC)
  - Intel Paragon (i/860)
  - Cray T3D/E (Alpha)
  - Silicon Graphics Challenge/Origin (R-series)
  - IBM SP (RS6000)
- Leveraging of workstation technologies
  - Operating systems
  - Programming languages & Compilers
  - Proprietary interconnection networks

ASCI Platforms

- Red -- Intel Teraflops
- Blue Mountain -- SGI Origin 2000
- Blue Pacific -- IBM SP-2

SMP Clusters

- Success of DOE Accelerated Strategic Computing Initiative (ASCI) program relies on IBM SP-2 with High Nodes (512 16-way nodes) [Option White]

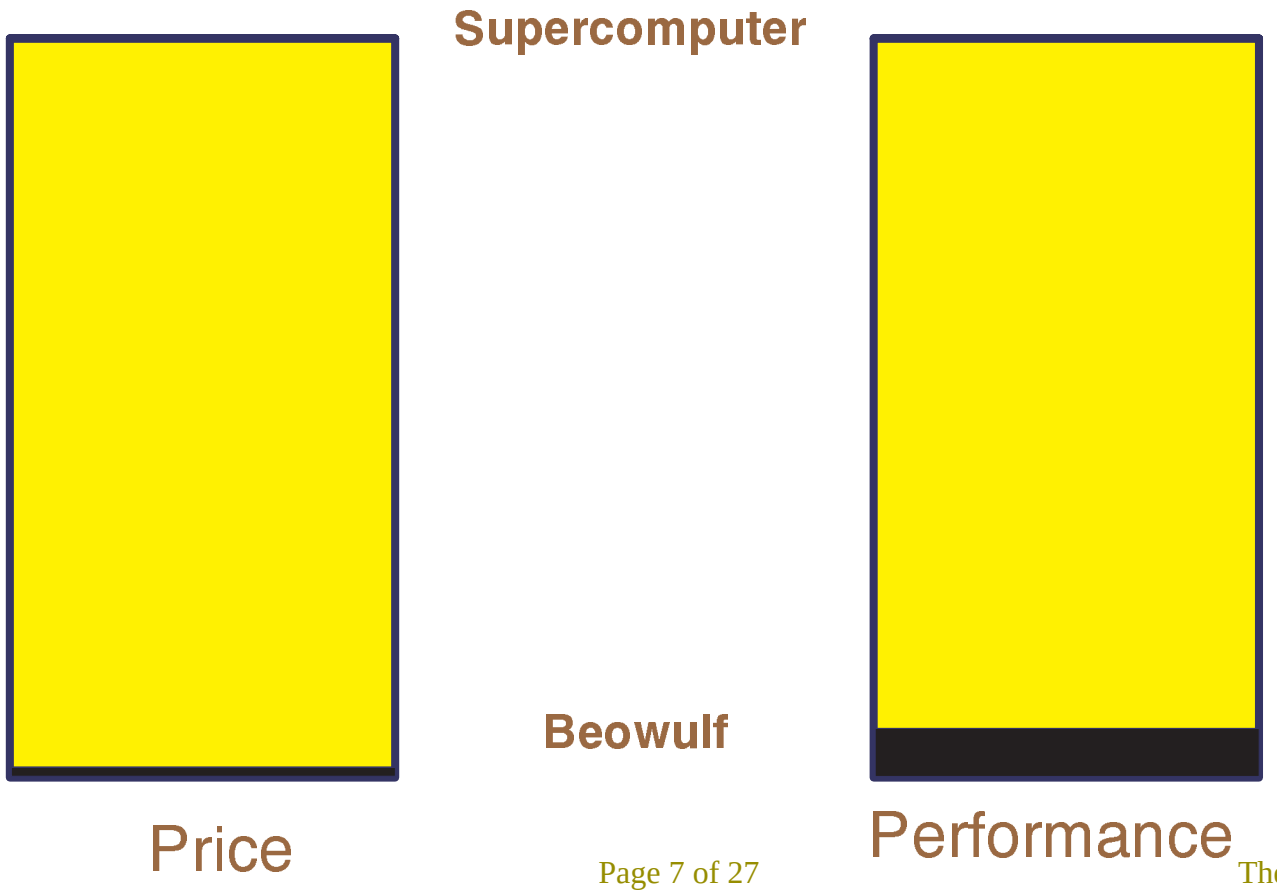


Success Stories

- Networks of workstations (NOW)
  - Cycle stealing
  - Parallel Virtual Machine
  - Condor
- Message Passing Interface
- Beowulf Systems
  - Friendly-user development systems
  - Optimize price (MM-COTS)
  - Home-built

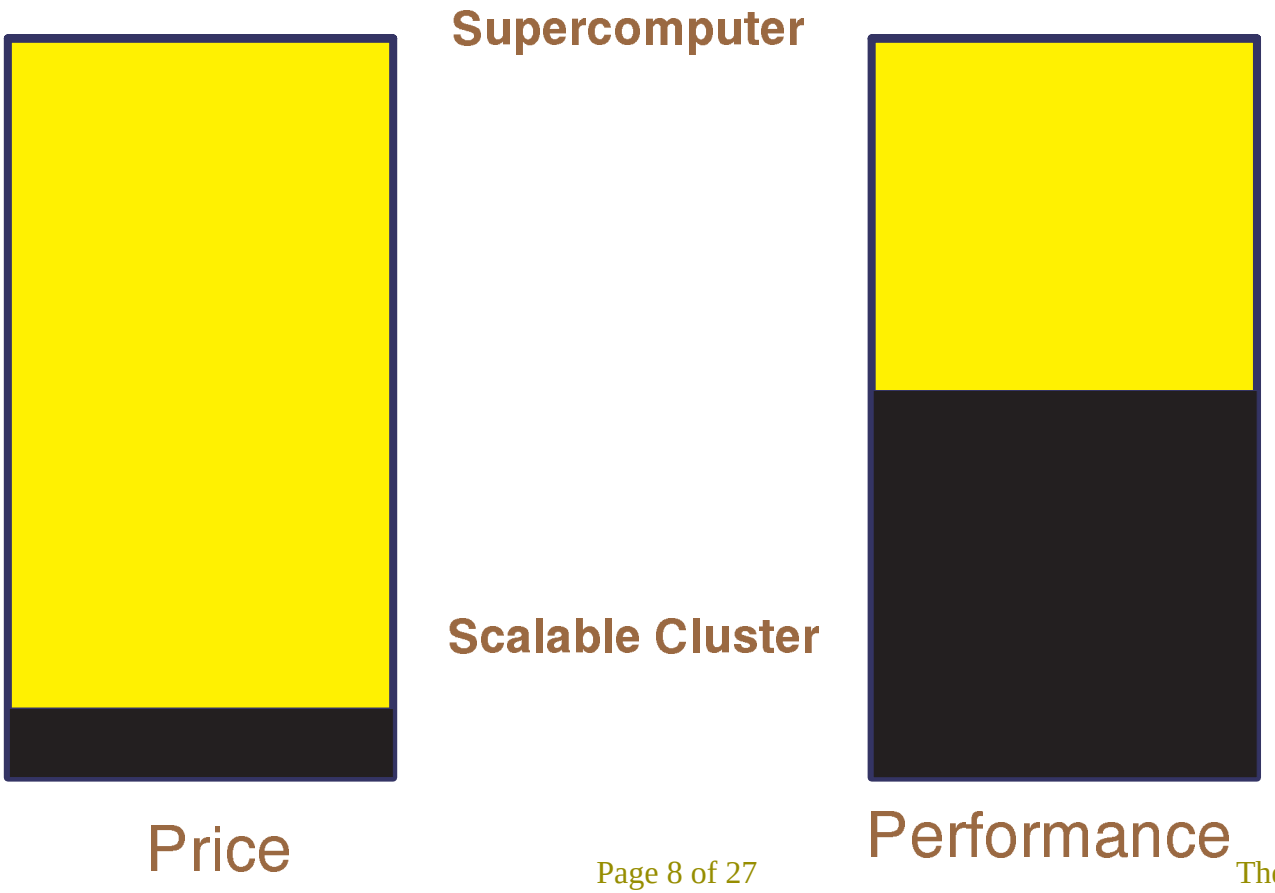
Beowulf Design

Minimize Price Per Mega(fl)ops



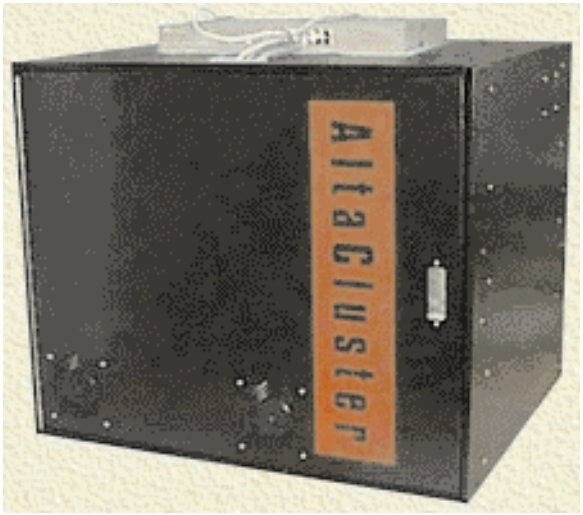
Scalable Cluster Design

Maximize Performance per Price Per Mega(fl)ops



Recent Developments

- Hardware/Software integrators
  - For example, Alta Technologies
- Vendor support
- Standard environment
- Packaging
- Remote temperature monitoring and reset



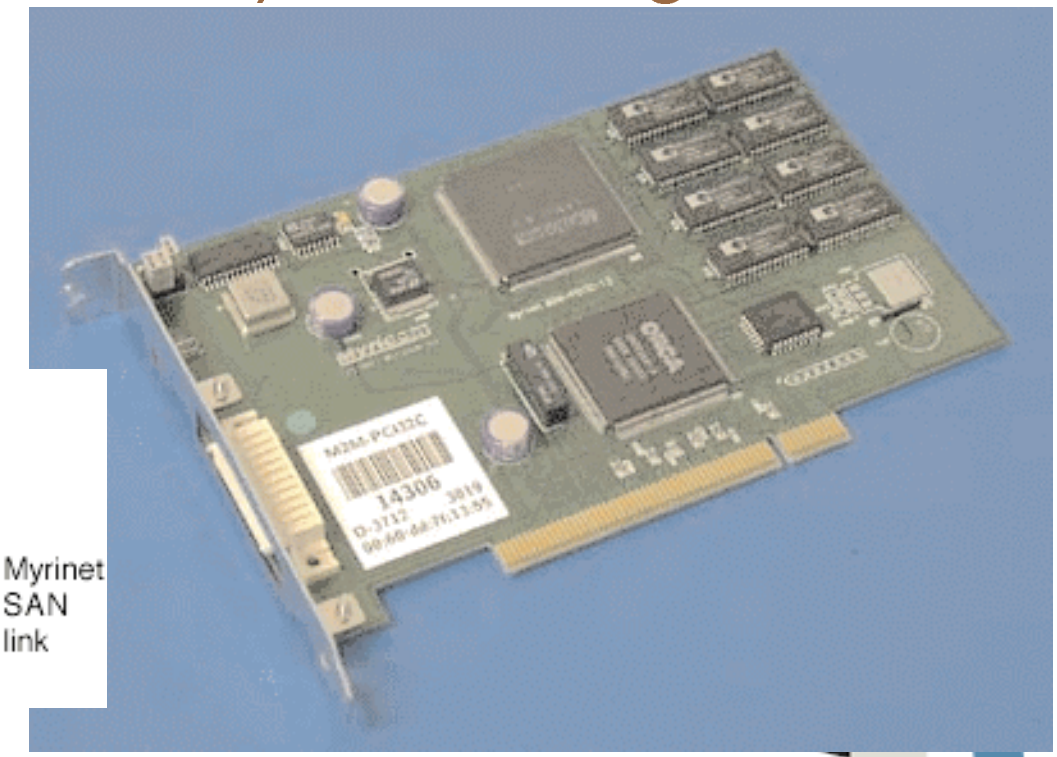
- Scalable networks and systems software

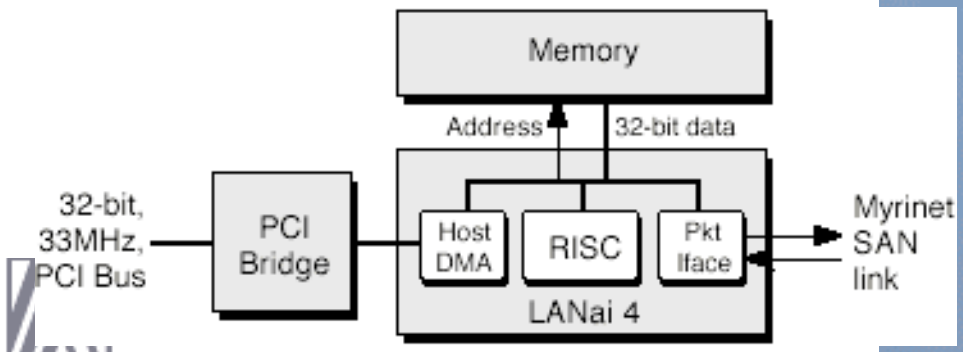
Achitecture & Technologies

- Intel Pentium Processors
- Fast Ethernet
- Gigabit Ethernet
- Myrinet

Myricom

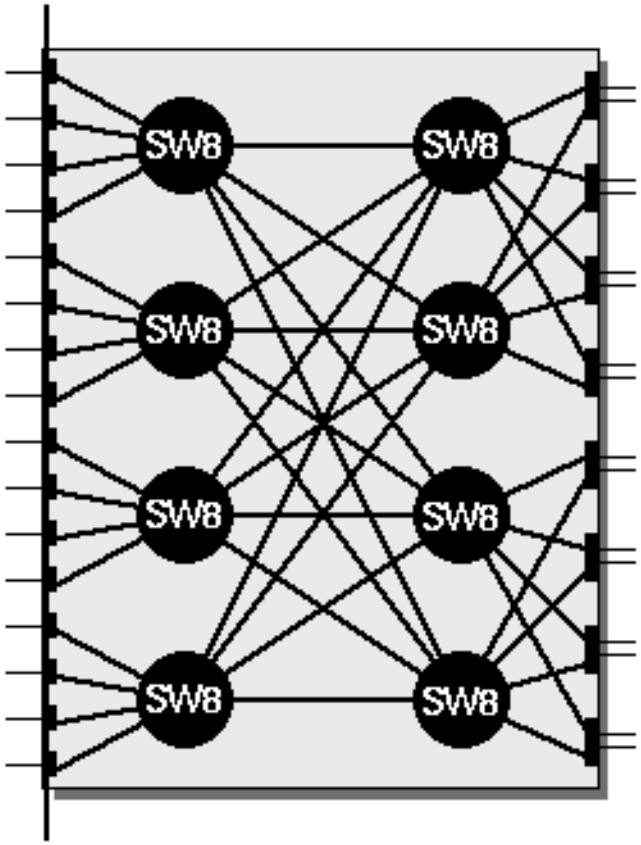
- Full-duplex 1.28 Gbps scalable network
- Low Latency (10's of *u*sec) cut-through cross-bar switches



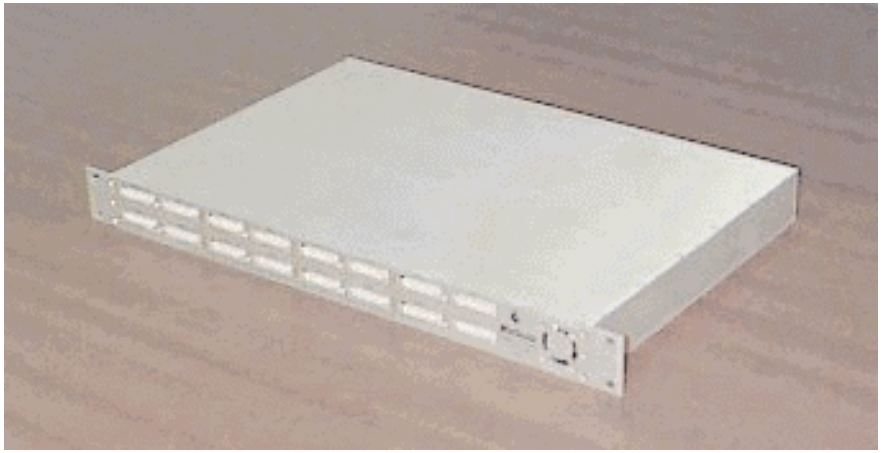


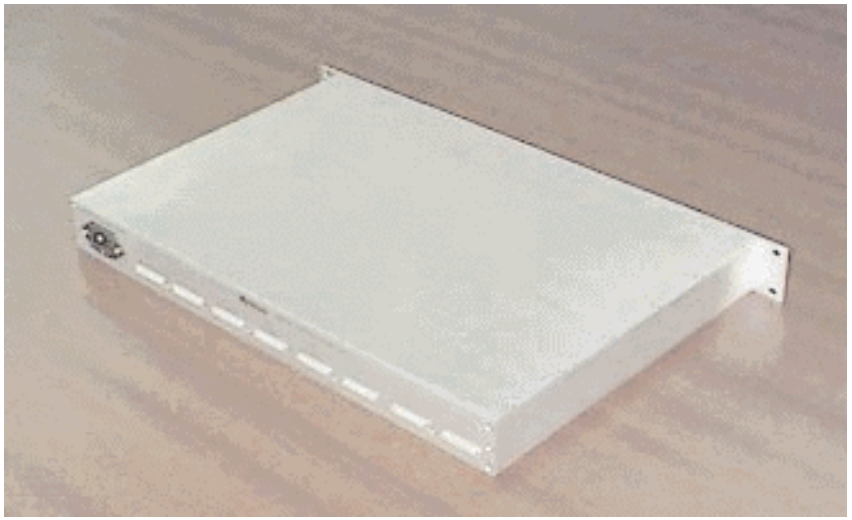
Myrinet

Octal SAN switch



Front





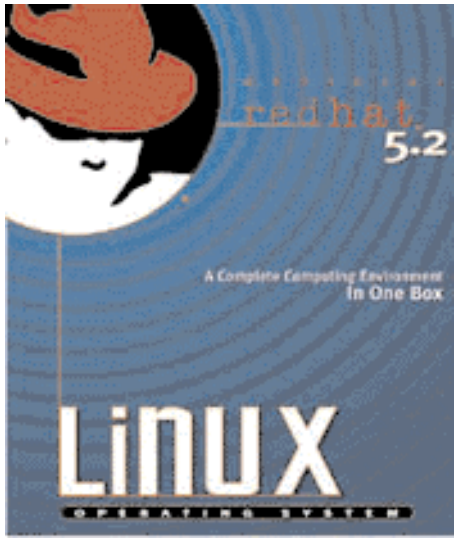
Back

System Software

- Operating Systems
- Compilers
- Parallel Programming
- Job Scheduling

Operating Systems

- Open Source
- Freely Available
- Linux



Parallel Programming

- Message Passing Standard: MPI
  - Enforces a shared-nothing paradigm between tasks
  - Communication via explicit messaging, perhaps through shared memory buffers when processors are on the same SMP node
- Shared Memory Paradigm
  - Coordinate accesses to shared memory
  - Simulate global shared address space via software-based distributed shared memory

Message Passing Interface

- Standard (1.1, June 1995)
- Portable, practical
- Freely-available reference implementations
- Version 2.0 includes parallel I/O, one-sided communication, etc.

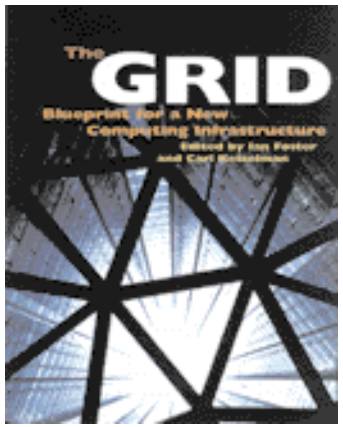
- HPF Parallel Fortran for clusters
- F90 Parallel SMP Fortran 90
- F77 Parallel SMP Fortran 77
- CC parallel SMP C/C++
- DBG symbolic debugger
- PROF performance profiler

Parallel Job Scheduling

- Node-based resource allocation
- Job monitoring and auditing
- Resource reservations

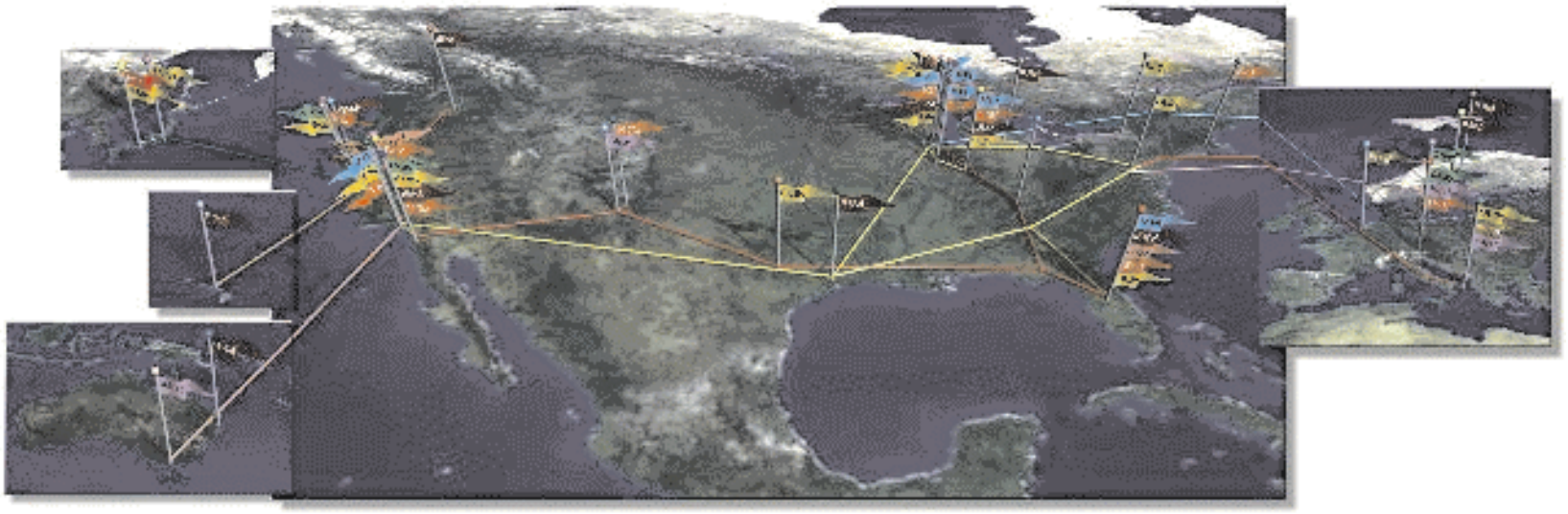
Computational Grid

- National Technology Grid
- Globus Infrastructure
  - Authentication
  - Security
  - Heterogenous environments
  - Distributed applications
  - Resource monitoring



National Technology Grid

GUSTO Testbed from SC98



Clusters on the Grid

- Develop applications locally, run large problems remotely

Applications

- AZTEC: iterative solver for solving sparse linear systems (ASCI)
- CACTUS: numerical relativity (astrophysics)
- HEAT: diffusion PDE using a conjugate gradient solver (ASCI)
- HYDRO: Lagrangian hydrodynamics code (ASCI)
- MILC: MIMD lattice computation, hybrid Monte Carlo (DOE Grand Challenge)

Applications

- NPB: NASA Ames Numerical Aerodynamics Simulation benchmarks
- SLbench: numerical linear algebra (MP-LINPACK)
- SPRNG: Scalable pseudo-random number generator
- SWEEP3D: 3-d wavefront in rectangular grid (ASCI)
- et cetera...

Alliance / UNM Roadrunner

- Strategic Collaborations with
  - Alta Technologies
  - Intel Corp.
- Node configuration
  - Dual 450MHz Intel Pentium II processors
  - 512 KB cache, 512 MB ECC SDRAM
  - 6.4 GB Hard drive
  - Fast Ethernet and Myrinet NICs
- Interconnection Network
  - Control: 72-port Fast Ethernet switch
  - Data: Myrinet


AltaCluster

Roadrunner System Software

- Redhat Linux 5.2
- MPI (Argonne's mpich 1.1.2)
- Portland Group Compiler Suite
- Myricom GM Drivers and MPICH/GM
- Portable Batch Scheduler (PBS)

Research Topics

- Shared-memory (SMP) algorithms
- Message-passing algorithms
- SMP Cluster algorithms
- Grid-enabled applications
- Parallel and distributed job scheduling across a cluster or grid nodes
- Performance-engineered applications