UNM-NCSA Roadrunner Supercluster Review

May 12, 1999

"Analysis of the Alliance/UNM Roadrunner Linux Supercluster," presented at the NSF/NCSA Alliance Roadmap '99 Meeting, Chicago, IL, May 12, 1999.

Alliance/UNM Roadrunner SuperCluster

Outline

- Architecture and Technology
- System Software
- Applications and Performance
- Users and Projects
- Usage
- How-to-run on RR
- Timeline
- Future Plans

Architecture and Technology: Scalable SuperCluster Design

- **Beowulf design minimizes price per megaflop**
  - Order from "Computer Shopper"
  - Assembly required
  - Last generation of processor
  - Fast Ethernet
- **SuperCluster design maximizes capability**
  - Rely on an integrator
    - packaging, operating system and software, support
  - Lastest processor technology (e.g., Intel/Alpha)
    - SMP nodes, large memory
  - Scalable interconnection network (Myrinet, GigE, ..)
    - Perhaps 40% of the overall price
  - Vendor-Independent

Recent Developments

- Hardware/Software integrators
  - Alta Technology
  - VA Linux Systems
  - ParaLogic
- Vendor support
- Standard environment
- Packaging
- Remote temperature monitoring and reset
- Cloning software
- Scalable networks and systems software

ALTA TECHNOLOGY

VA Linux Systems

PARALOGIC

Xtreme MACHINES

Roadrunner Supercluster

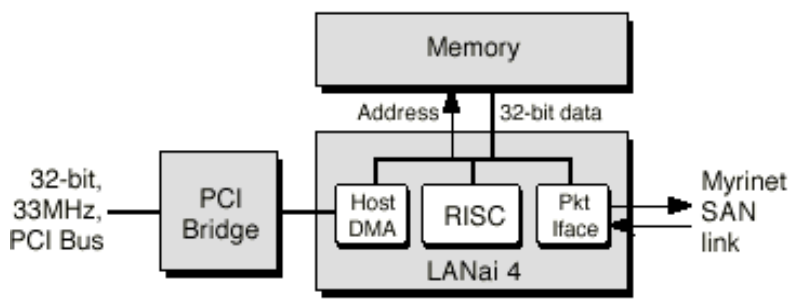
- Strategic Collaborations with
  - Alta Technologies
  - Intel Corp.
- Compute-Node configuration
  - Dual 450MHz Intel Pentium II processors
  - 512 KB cache, 512 MB ECC SDRAM
  - 6.4 GB EIDE hard drive (4 GB scratch space available)
  - Fast Ethernet and Myrinet NIC
  - Low-level diagnostic network
- Peak performance 63 GFLOPs
- Under $6,500 per node for this configuration

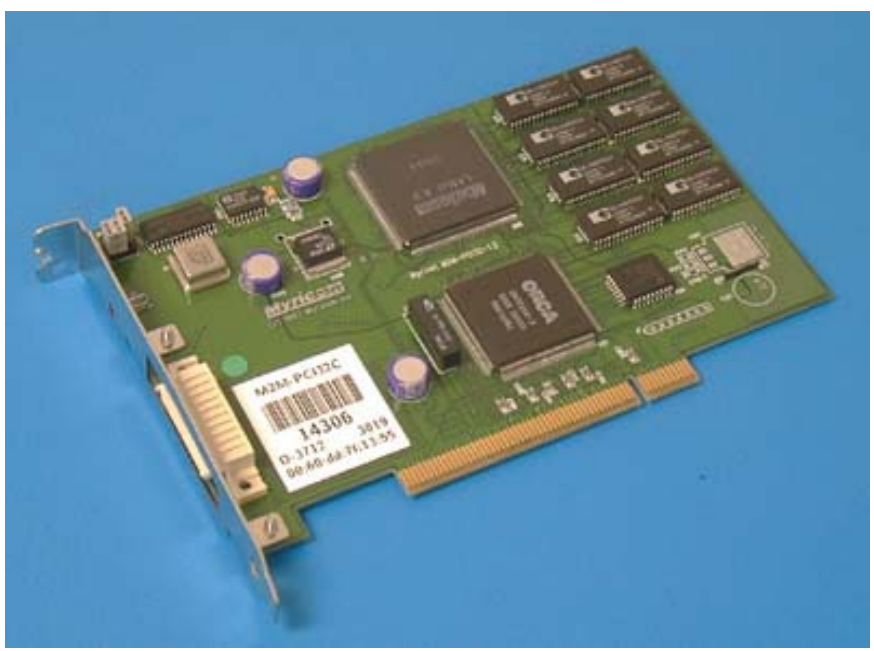
Alliance / UNM Roadrunner

- **Interconnection Networks**
  - **Control: 72-port Fast Ethernet Foundry switch with 2 Gigabit Ethernet uplinks**
  - **Data: Four Myrinet Octal 8-port switches**
  - **Diagnostic: Chained serial ports**

Myricom

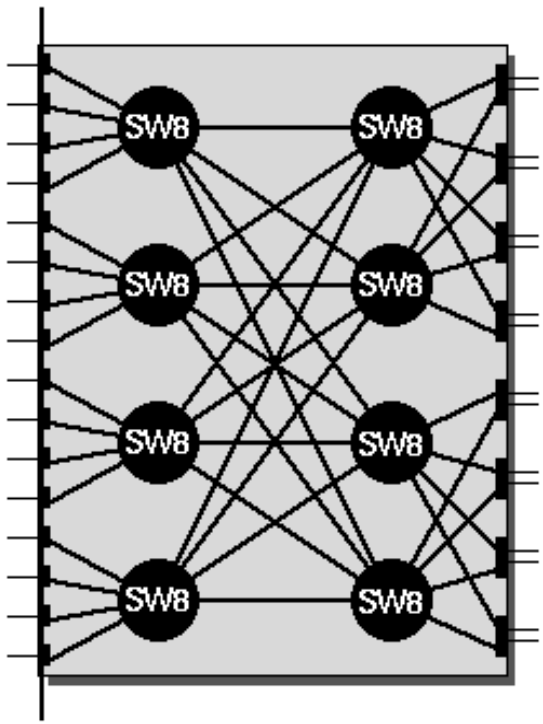
- Full-duplex 1.28 Gbps scalable network
- Low latency (10's of *u*sec) cut-through cross-bar switches





Myrinet

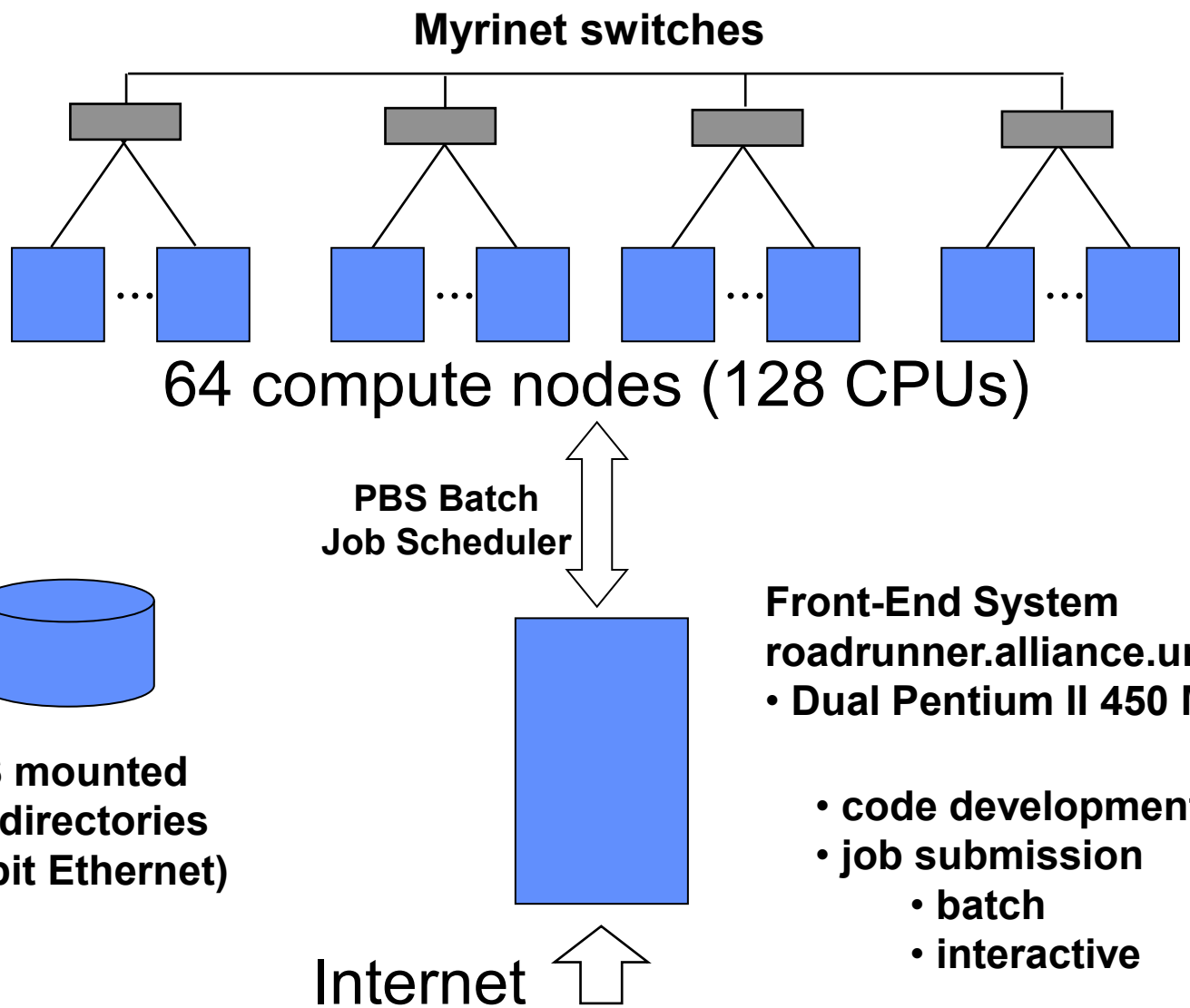
Octal SAN switch



Front

Back

System Layout



A Peek Inside Roadrunner

System Software

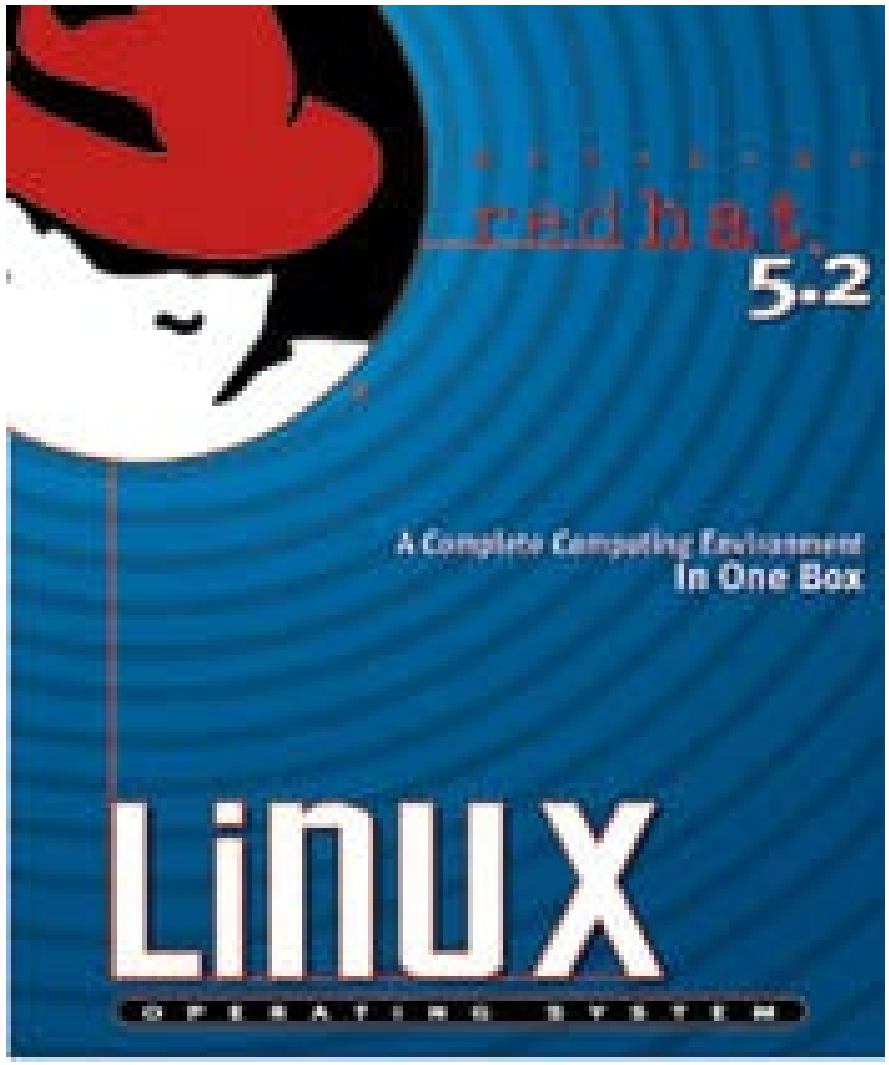
- Operating Systems
- Compilers
- Parallel Programming Environment
- Job Scheduling
- Globus Grid Infrastructure

Why Linux?

- Major vendors are moving to Linux (SGI, IBM, etc.)
- Open source - OS can be modified if necessary
- Active development and support by programmers worldwide resulting in readily available software
- Most software (system and application) is public domain
- Most scientific applications are running on Unix so porting to Linux is not difficult
- Globus is available and running on Linux
- High availability systems, and fast setup/install times

Operating Systems

- Open Source
- Freely Available
- Out of the box
  - RedHat 5.2
  - Linux Kernel 2.2.12



GNU Compilers

- C/C++:
  - gcc
  - egcs
- Fortran 77/9x:
  - g77
  - VAST f90

PGI Compilers

- HPF Parallel Fortran for clusters
- F90 Parallel SMP Fortran 90
- F77 Parallel SMP Fortran 77
  - Fortran supports OpenMP standard
- CC Parallel SMP C/C++
- DBG symbolic debugger
- PROF performanc

Message Passing Interface

- Standard (1.1, June 1995)
- Portable, practical
- Freely-available reference implementations
  - MPICH-GM 1.1.2.3
  - GNU/GM, PGI/GM, GNU/Ethernet, and PGI/Ethernet
- Version 2.0 includes parallel I/O, one-sided communication, etc.

Roadrunner System Libraries

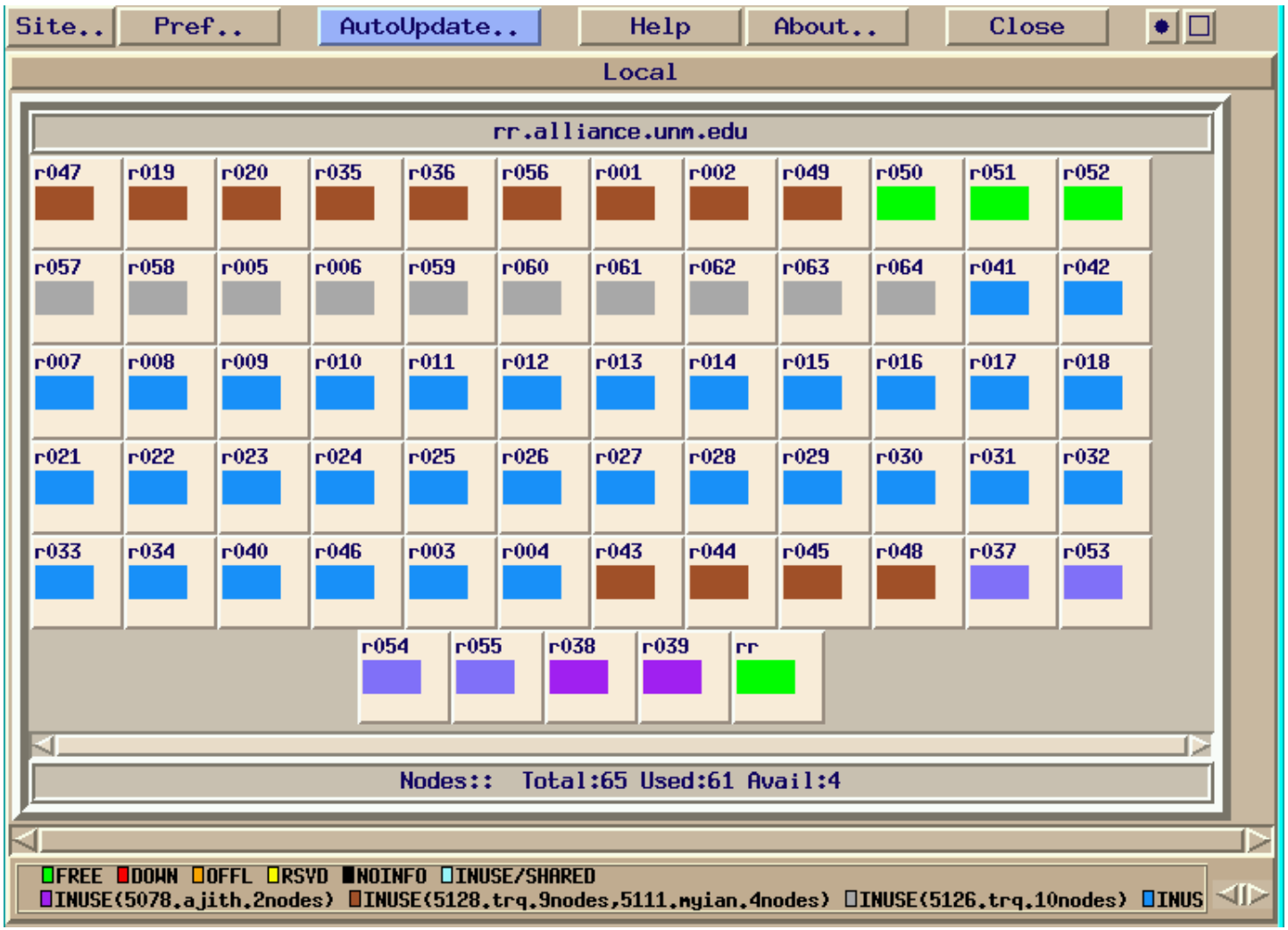
- BLAS
- LAPACK
- ScaLAPACK
- PETSc
- FFTw
- PGPLOT

Parallel Job Scheduling

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

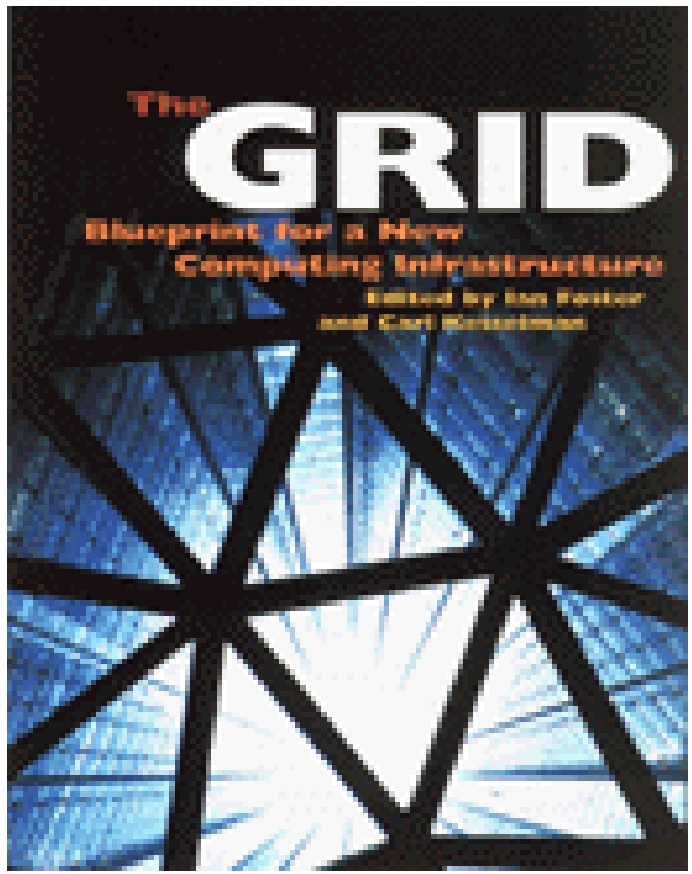
Portable Batch System

Job Monitoring with PBS

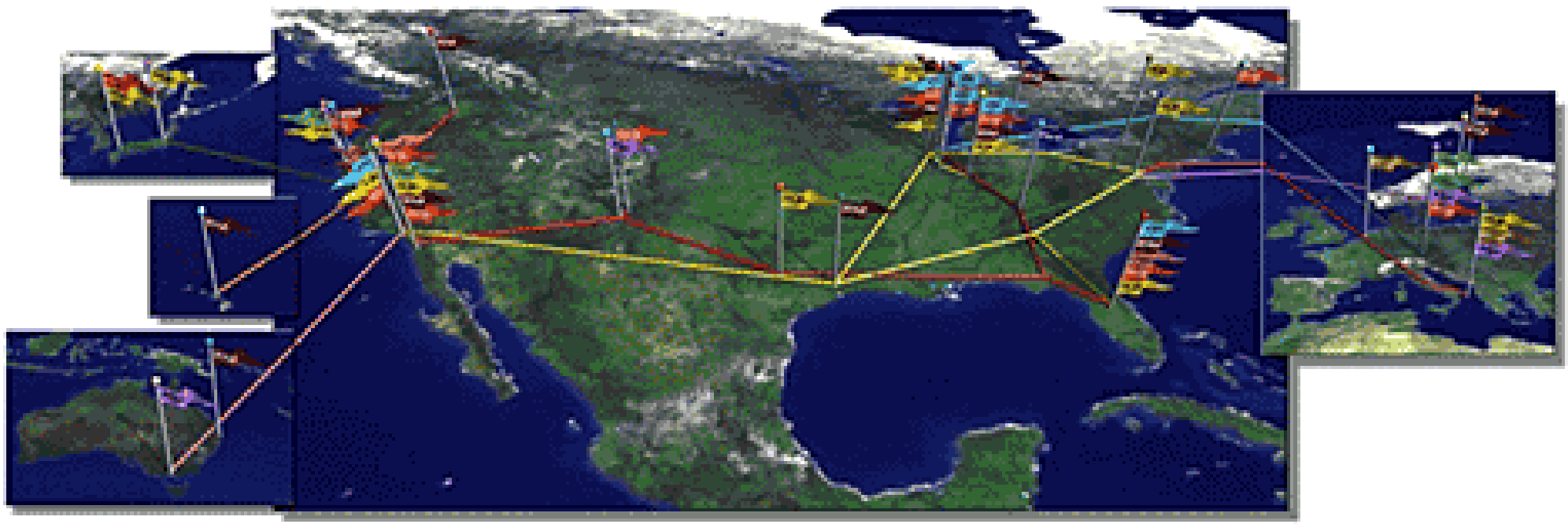


Computational Grid

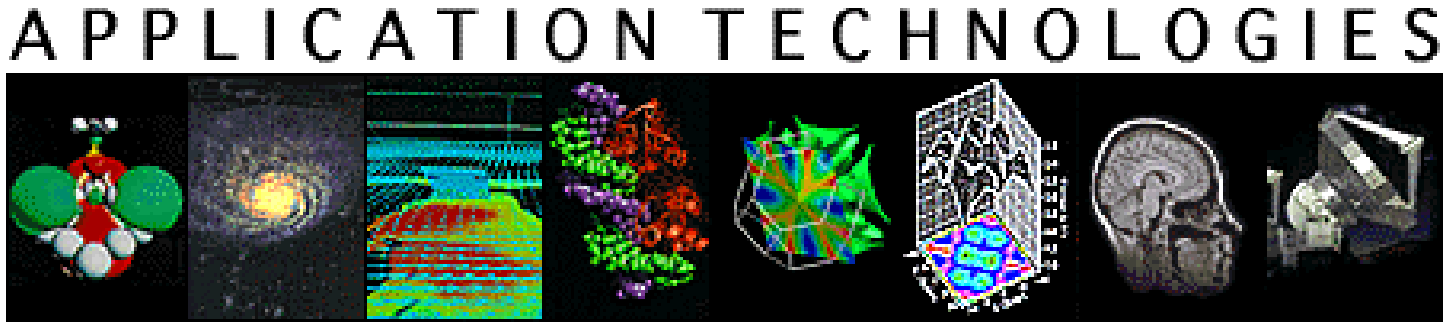
- National Technology Grid
- Globus Infrastructure
  - Authentication
  - Security
  - Heterogenous environments
  - Distributed applications
  - Resource monitoring
- Globus version 1.1.0b16
- Demo at SC ’99



GUSTO Testbed from SC98



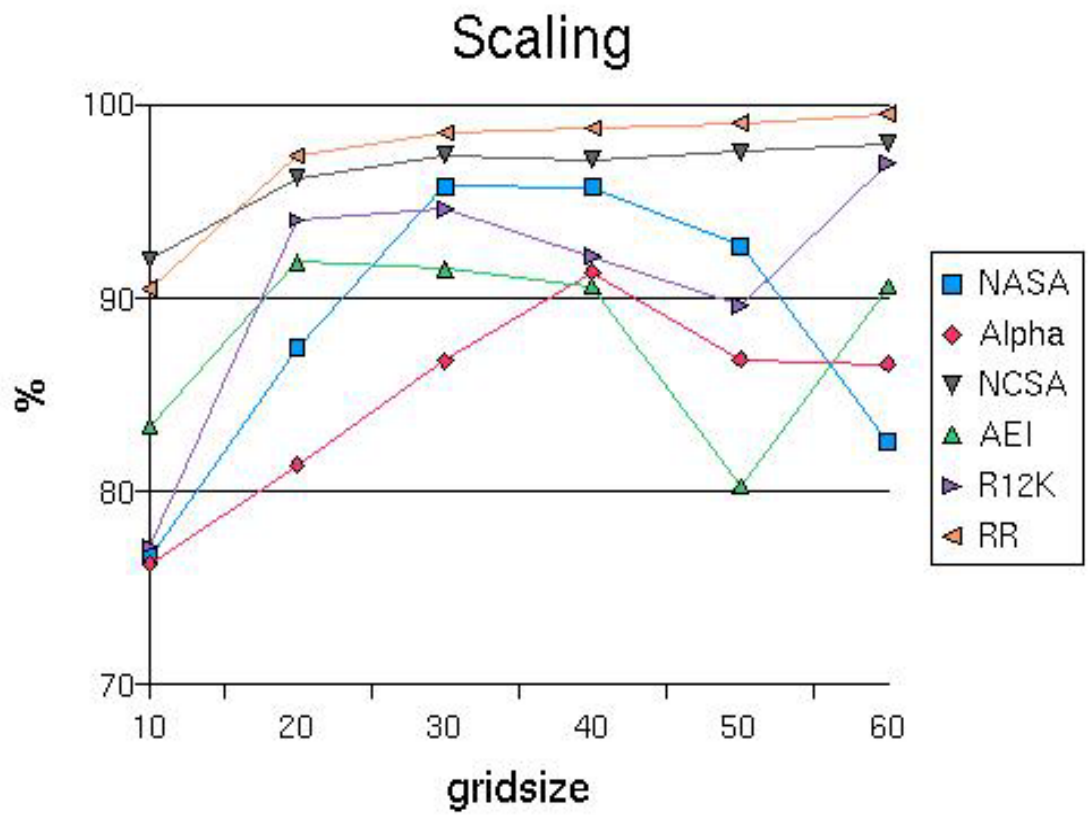
Selected Applications



- **CACTUS - 3D Numerical Relativity Toolkit for Computational Astrophysics**
- **MILC - MIMD Lattice Computation (QCD)**
- **ARPI3D - 3-D numerical weather prediction model**
- **BEAVIS - Dynamic simulation of particle-laden, viscous suspensions**
- **AIPS++ - Astronomical Information Processing System**
- **ASPCG - 2-D Navier Stokes solver**

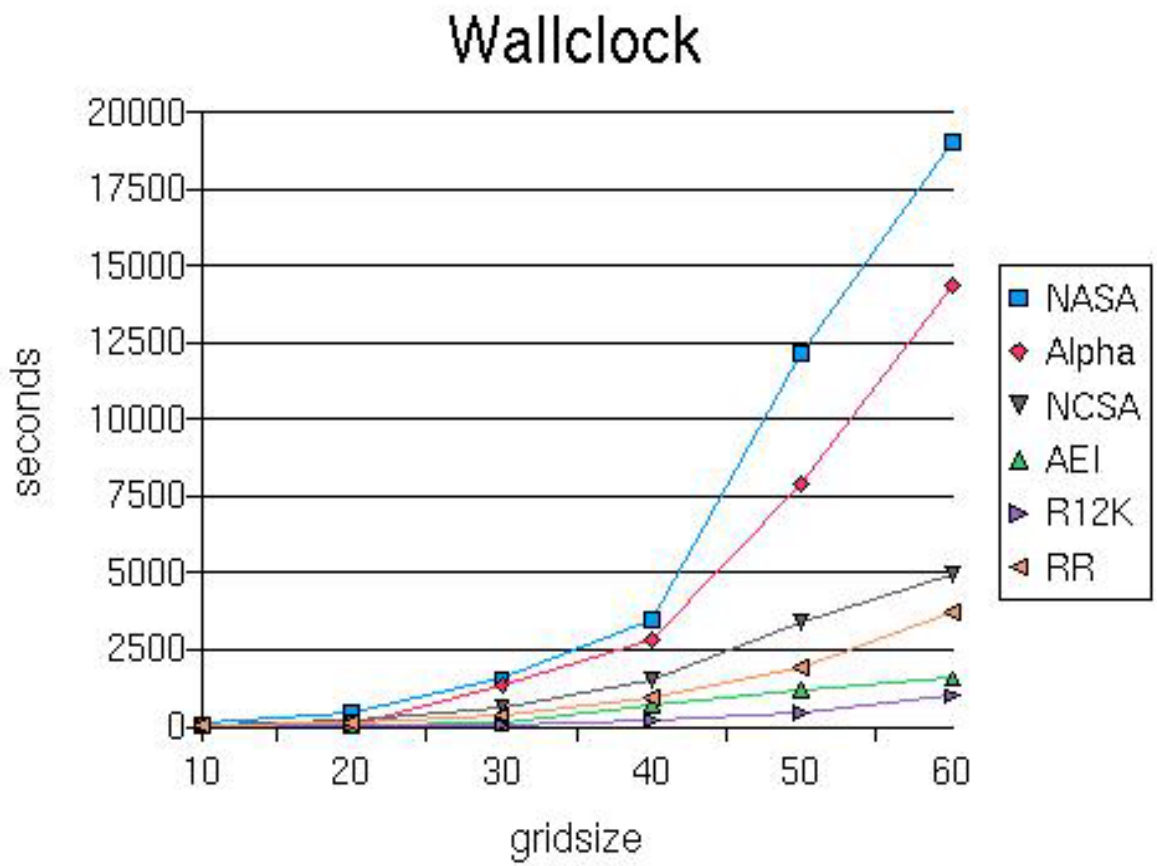
Cactus

- A modular manageable high-performance 3D tool for solving Einstein equations numerically
- Fast CPUs and high scalability are demonstrated
- Graph represents scaling of various sized problems on 32 processors



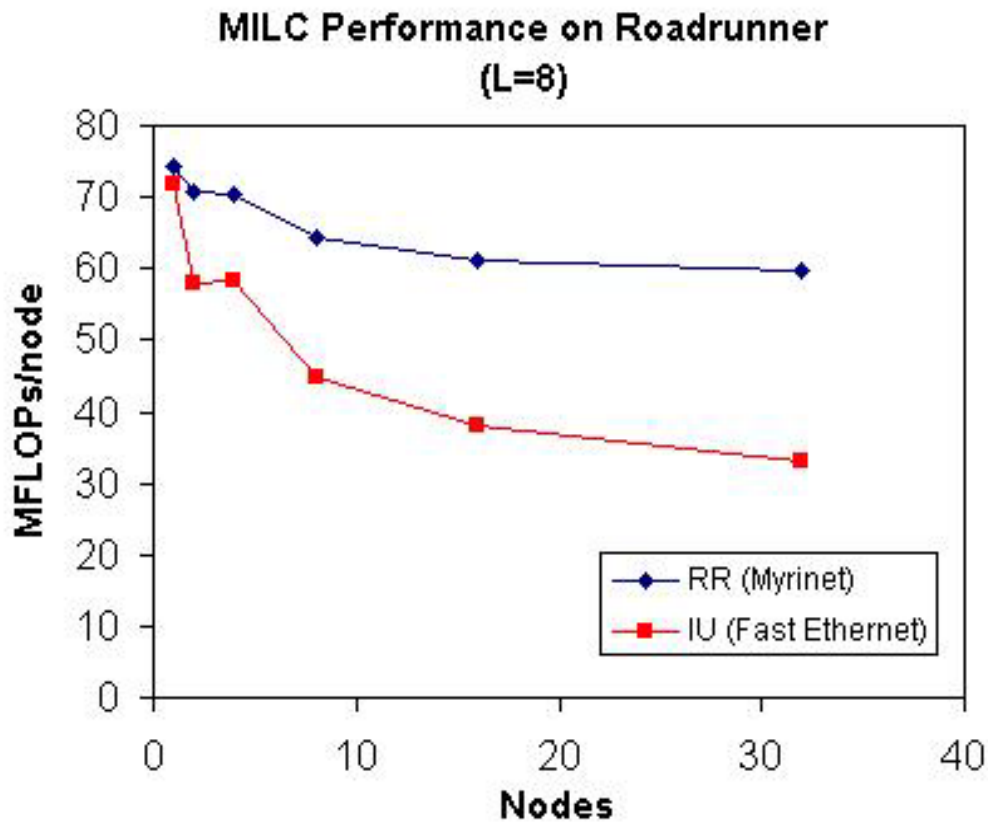
Cactus (Cont.)

- NASA
  - 64 dual PPro 200, 64MB RAM
- Alpha/Linux
  - 48 DEC Alpha 300 XL
- NCSA
  - 32 dual PII 333 512 MB RAM, 64 dual PII 300, 512 MB
- Origin2000@AEI
  - 32 R10K, 195 MHz, 4 MB Cache, 8 GB RAM
- Origin2000@SGI (R12K)
  - 32 R12K, 300 MHz
- Roadrunner
  - 64 dual PII 450 MHz, 512 MB RAM

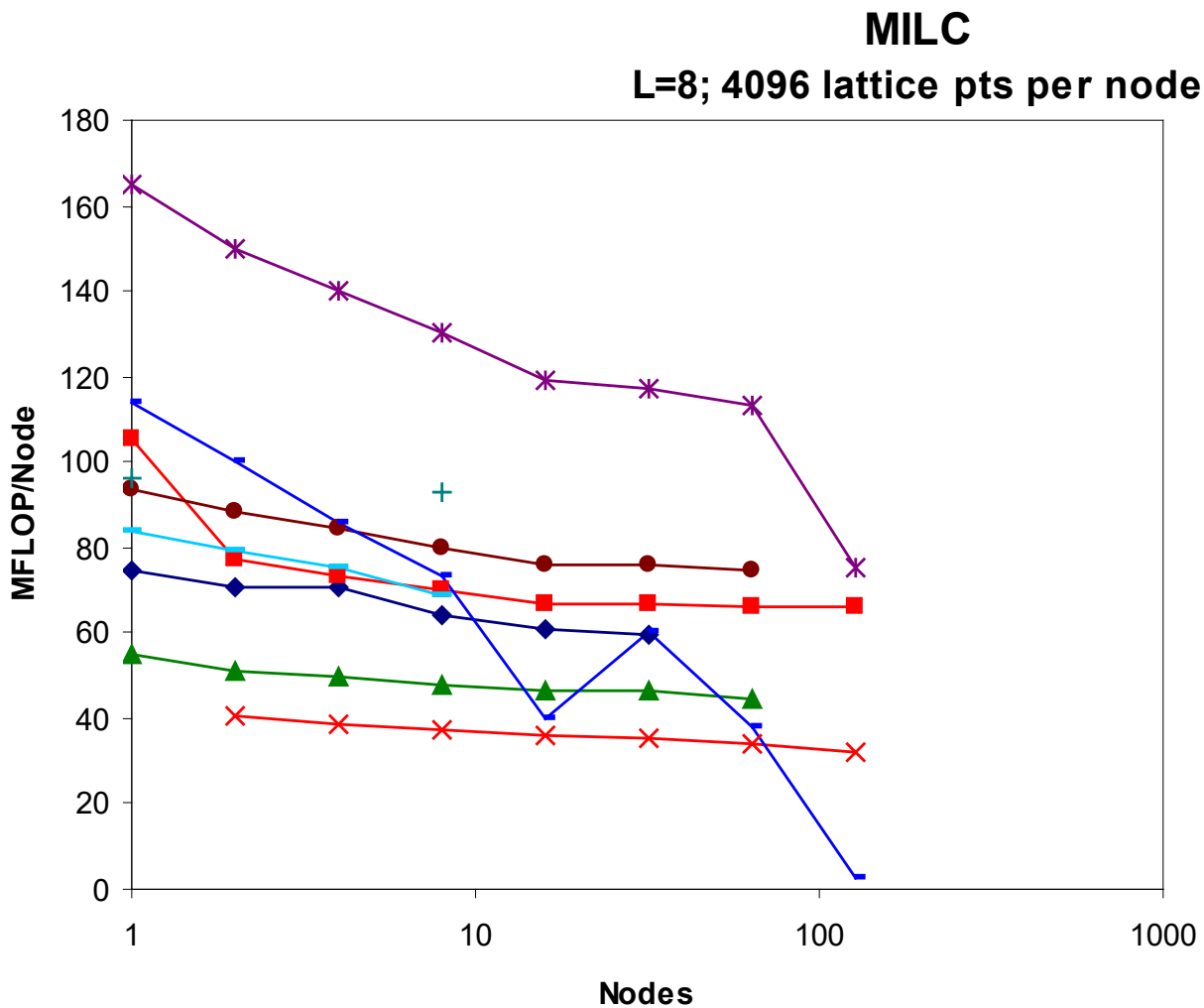


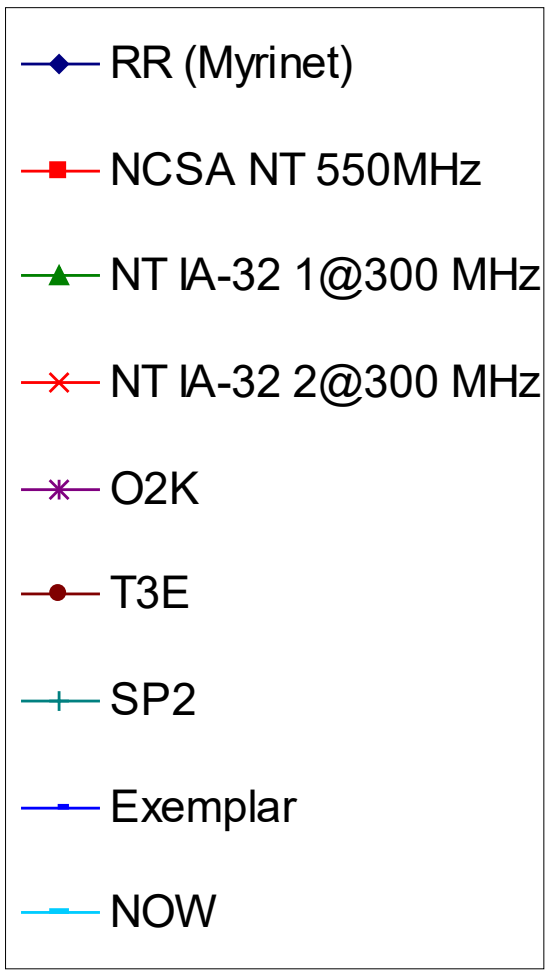
MILC

- The MILC benchmark problem is a conjugate gradient algorithm for Kogut-Susskind quarks
- L=4 means that there is a 4^4 piece of the domain on each node.
- For larger problems, MILC achieves > 60 MFLOPs/node
- Linux cluster at IU is used to compare Myrinet and ethernet



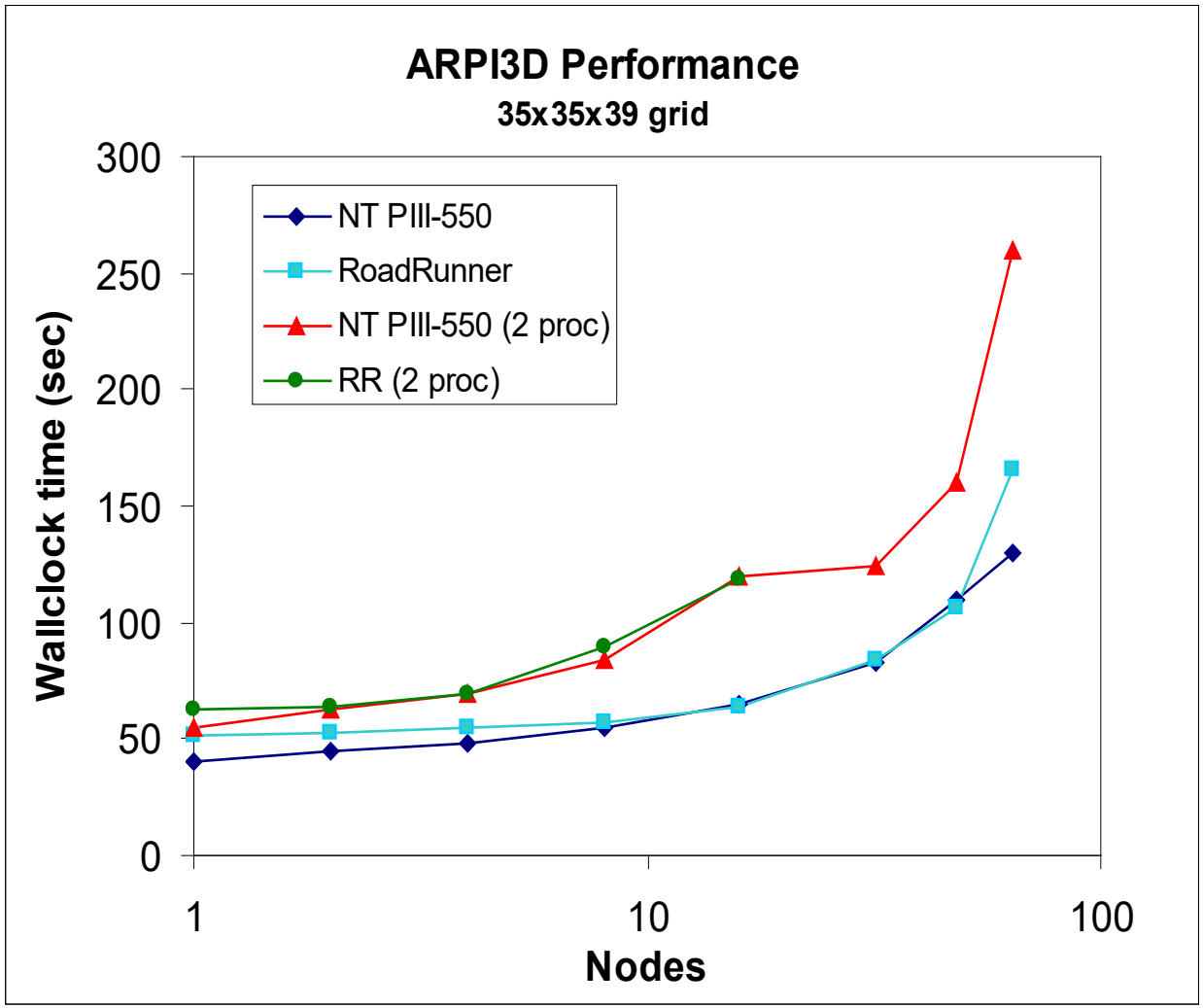
MILC Performance





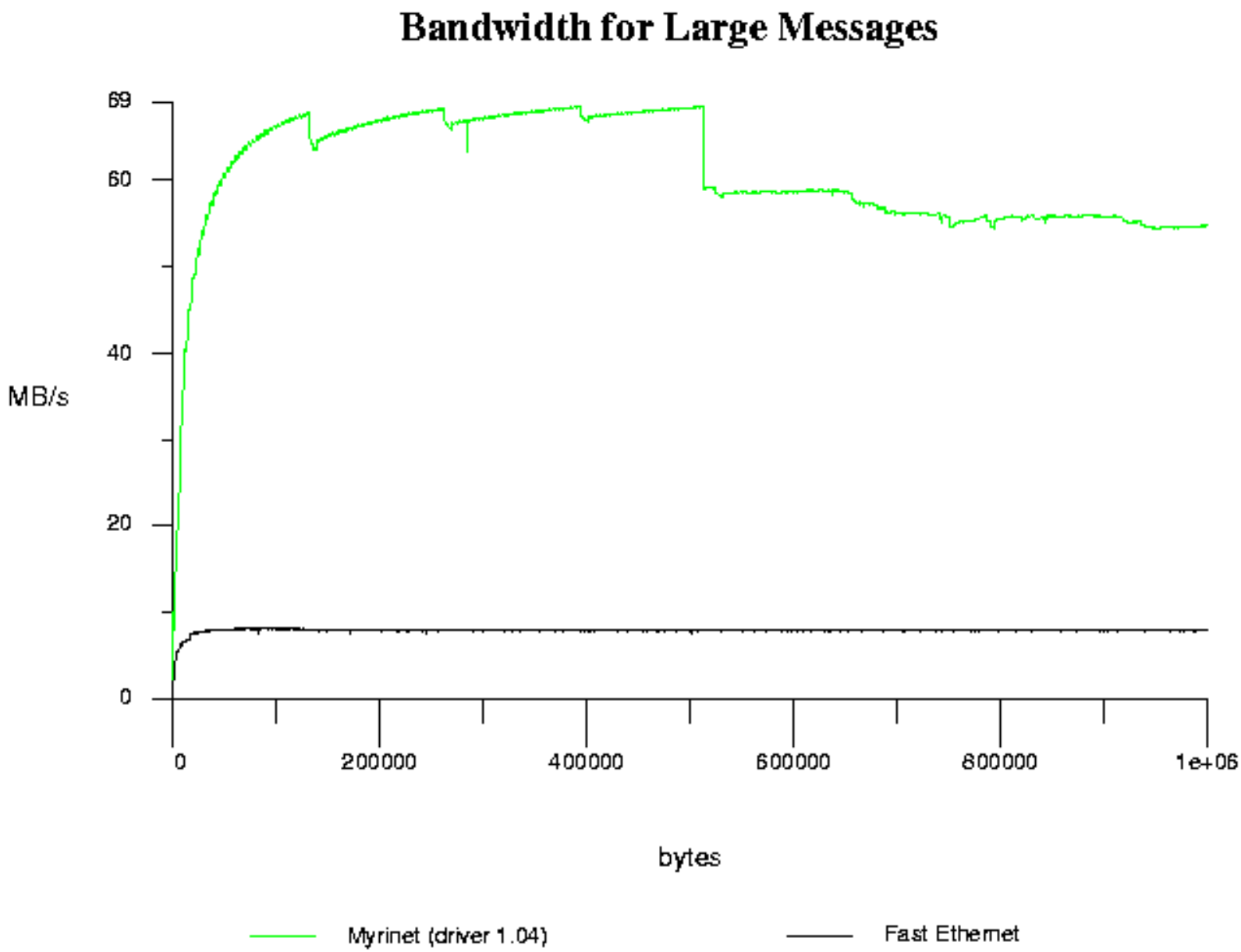
ARPI3D

- 1-process/node and 2-process/node tests
- 2-process/node times are approximately 20% larger (in the computational part of the code due to memory/bus competition) than the 1-processor/node results
  (for both Linux and NT)
- *Performance problems with 2proc/nodes timings on 48,64,128 tests*

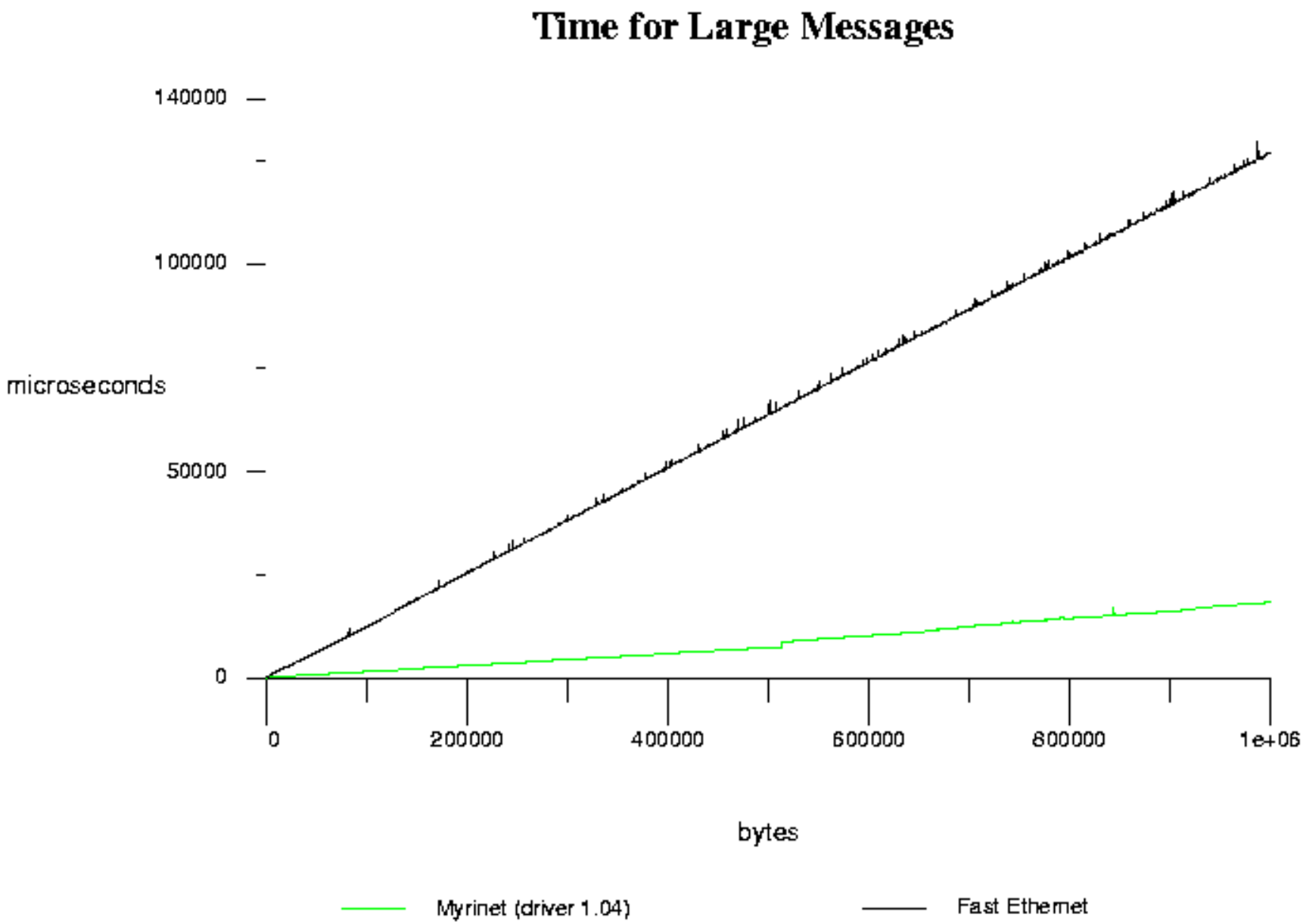


courtesy of Dan Weber

Roadrunner Bandwidth



Roadrunner Ping-Pong Time



Usage

- Usage statistics from PBS go here
  - 100% happy users ;-)

Roadrunner "Friendly" User Phase

Users	Institution
27	University of New Mexico
5	Rice University
4	NCSA
4	National Radio Astronomy Observatory
3	University of Texas at Austin
3	University of Oklahoma
3	University of Washington
2	Brown University
2	University of Illinois
2	Sandia National Laboratories

plus users from Arizona State University, Indiana University, Iowa State University, Los Alamos National Lab, MIT, National Radio Astronomy Observatory, Penn State University, Princeton, University of Kentucky, and New Mexico State University (78 total users)

“Friendly” User Snapshot

Dan Weber - University of Oklahoma,

- Benchmarking of weather prediction code
- *"Porting code to the Supercluster is transparent. I had models running in less than 30 minutes."*

George Karamanos and George Karniadakis - Brown University

- Benchmarking of Fourier-algorithm bluff body simulation code

Steve Gottlieb - Indiana University

- MILC QCD benchmarking

Selected User and Projects

- James Adams - ASU - Modelling Adhesion/Adhesive Wear
- Matt Challacombe - LANL - Irregular parallel computation in linear scaling quantum chemistry
- Romeel Dave - Princeton Univ. - PTreeSPH Benchmarking
- Steve Gottlieb - Indiana Univ. - Benchmarking MILC QCD code
- Marc Ingber - UNM - Multiphase Flow
- George Karniadakis - Brown Univ. - F15 Turbulence Simulation

Selected User and Projects (cont.)

- Athol Kemball - NRAO - AIPS++
- Richard Matzner - UT-Austin - Gravitational Waves from Black Hole Collisions
- George Phillips - Rice Univ. - Evaluation of Treadmarks (Molecular Dynamics)
- Tom Quinn - UWashington - Galaxy Formation
- Prasada Rao - UKentucky - Parallel Least Squares Finite Element Methods
- Ken Summers - Visualization of Parallel Programs in a Virtual Environment
- Timothy Thomas - UNM - Particle Physics Simulations
- Dan Weber - UOK - Weather Modeling

Getting an Account

- **To Apply for an Account**
  - http://www.alliance.unm.edu/accounts
  - accounts@alliance.unm.edu

- **Contact Information**

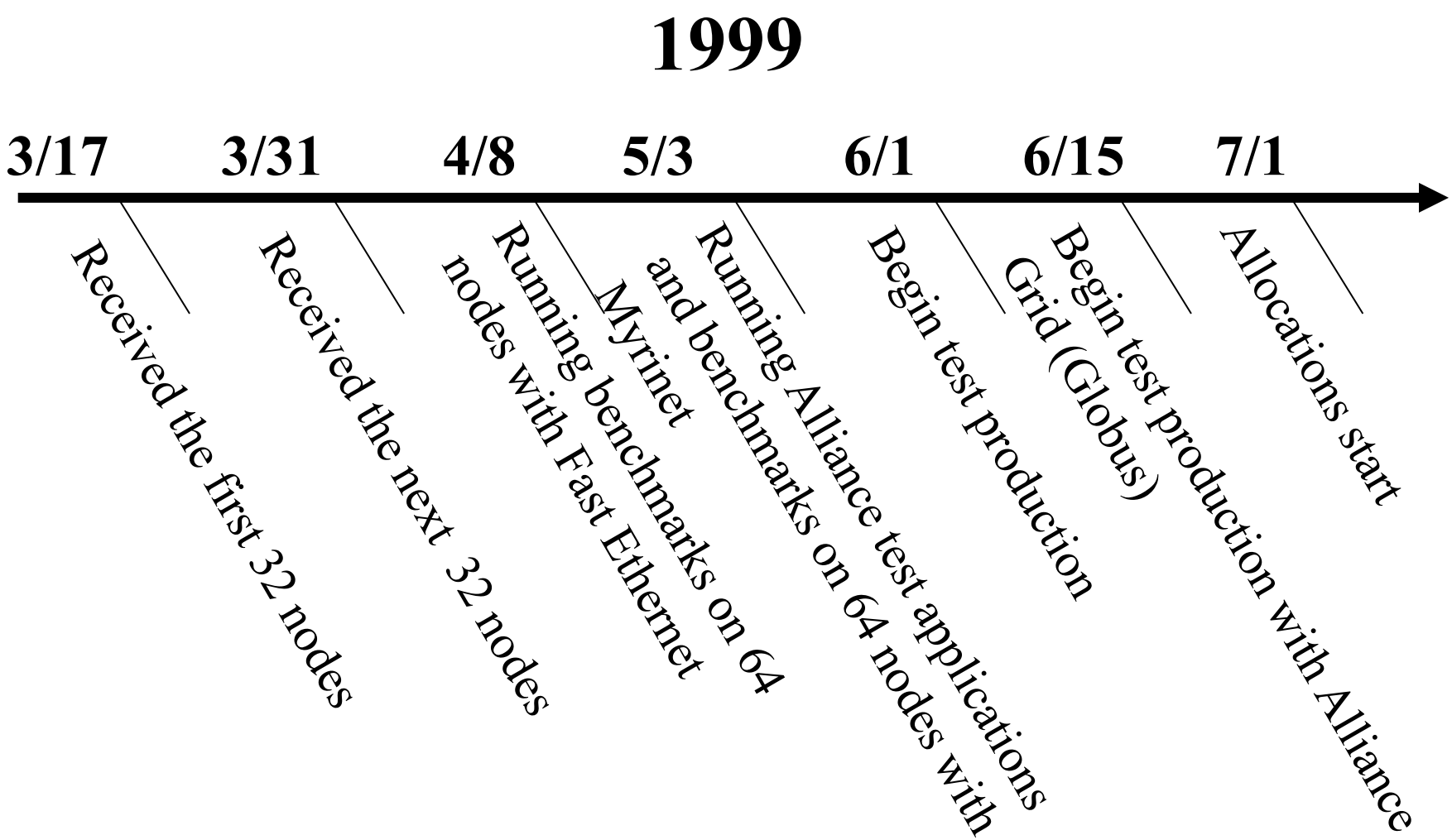
  http://www.alliance.unm.edu/

  help@alliance.unm.edu

Easy to Use

- `ssh -l username rr.alliance.unm.edu`
- `mpicc -o prog helloWorld.c`
- `qsub -I -l nodes=64`
- `mpirun prog`

Roadrunner SuperCluster Timeline



Future Plans

- **Software:**
  - Install PGI 3.1 F90 when it becomes available. (next month)
  - Install Maui Scheduler when it becomes available in the next two months.
  - Install VAMPIR, a performance monitoring tool.
  - Install the Etnus MPI TotalView Debugger as soon as it becomes available. We offered to be a beta site...(next couple of months)
  - Evaluating NAG F95.
  - Improving the PBS accounting system.
  - Create additional batch queues in PBS.
  - Upgrade to Red Hat 6.0.
  - Working with Myricom to resolve remaining Myrinet issues.

Future Plans (cont.)

- **Hardware:**
  - Mass Storage (> 0.5 TB in the next month)
  - Adding an eight node (16 processor) 333MHz Pentium II development machine in the next month.
- **SC '99:**
  - Will become a part of the Globus 1.1 Virtual Machine Room demo.
  - Working on Globus-Condor-Roadrunner demo.
  - Working with GK at Brown for F15 demo-movie.
  - Working with other users for demos.
- **User Support:**
  - Working on a weekly newsletter for users.
  - Creating additional on-line documentation.