SuperClusters: A New Approach for High-Performance Computing

Prof. David A. Bader
Department of Electrical and Computer Engineering &
Albuquerque High Performance Computing Center
University of New Mexico
<*dbader@eece.unm.edu*>

SuperCluster Outline

- History of Cluster Computing
- Recent Developments
- SuperCluster Architecture and Technology
- SuperCluster Systems Software
- Computational Grid
- Case Study: Alliance/UNM Roadrunner SuperCluster
  - Performance Analysis
  - Alliance Applications

Brief History of Cluster Computing

- 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)

History of Clusters

- Leveraging of workstation technologies
  - Operating systems
  - Programming languages
  - Compilers
  - Proprietary interconnections networks

DOE ASCI Platforms

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

ALLIANCE

The University of New Mexico

Success Stories

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

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

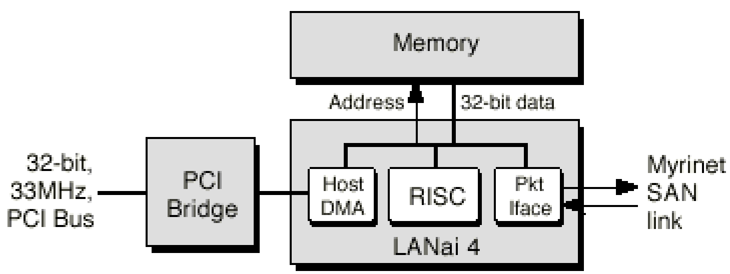
- Hardware/Software integrators
  - Alta Technology
  - VA Linux Systems
  - ParaLogic
- Vendor support
- Standard environment
- Packaging
- Remote temperature monitoring and reset
- Cloning software
- Scalable networks and systems software

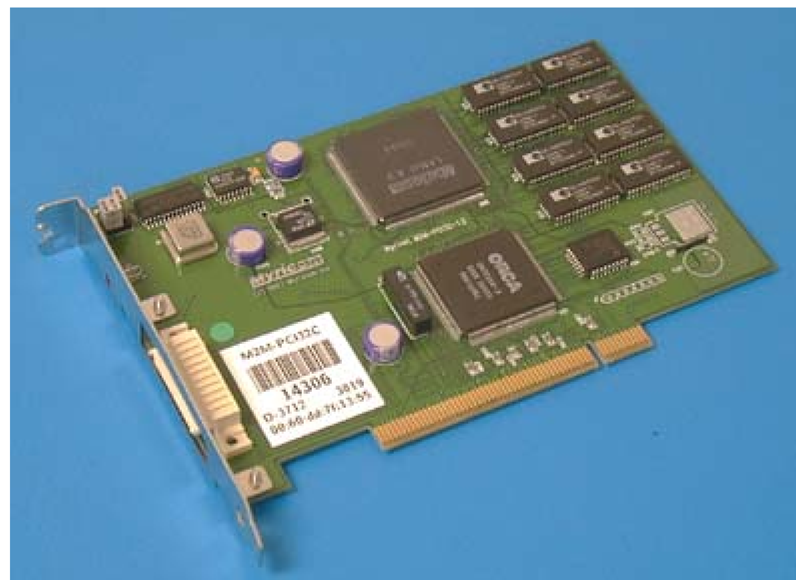
Architecture & Technologies

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

Myricom

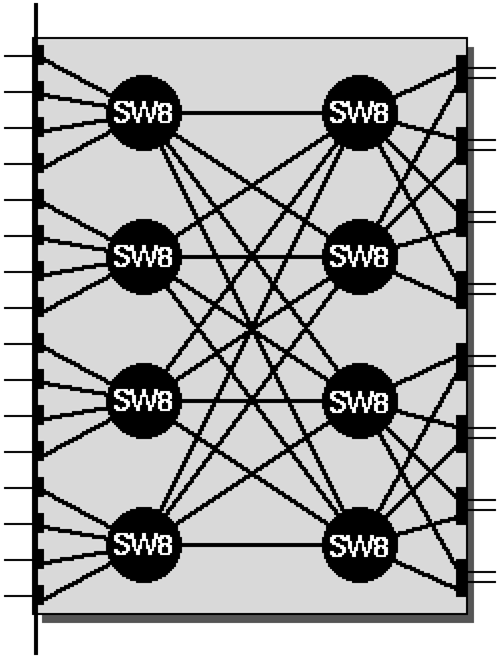
- Full-duplex 1.28 Gbps scalable network
- Low latency (10's of *u*sec) cut-through cross-bar switches



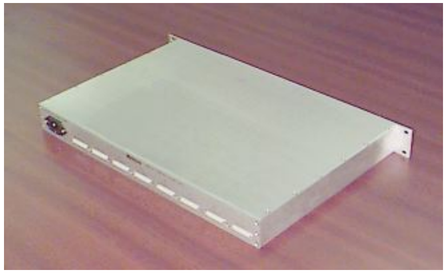


Myrinet

Octal SAN switch



Front



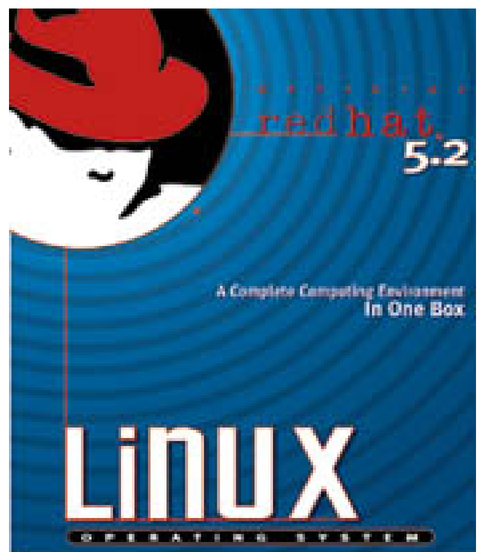
Back

System Software

- Operating Systems
- Compilers
- Parallel Programming Environment
- Job Scheduling

Operating Systems

- Open Source
- Freely Available
- Linux



Parallel Programming Environment

- Message Passing Standard: MPI
  - Enforces a shared-nothing paradigm between tasks
  - Communication via explicit messaging, perhaps through shared member 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

MPI

- 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

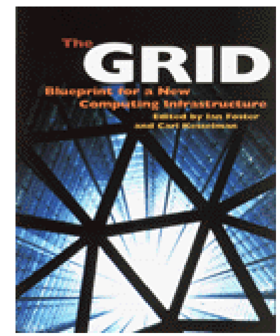
Parallel Job Scheduling

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

Portable Batch System

Computational Grid

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

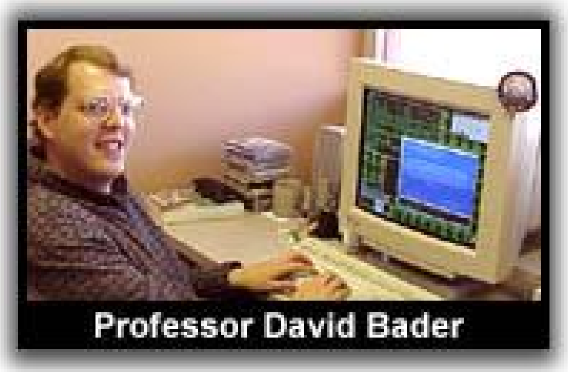


National Technology Grid

GUSTO Testbed from SC98

Clusters on the Grid

- Smooth transition from the desktop to cluster, supercluster, and supercomputer
- Think global, act local!



Alliance/UNM Roadrunner SuperCluster

Alliance/UNM Roadrunner SuperCluster

- 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 IDE hard drive
  - Fast Ethernet and Myrinet NICs

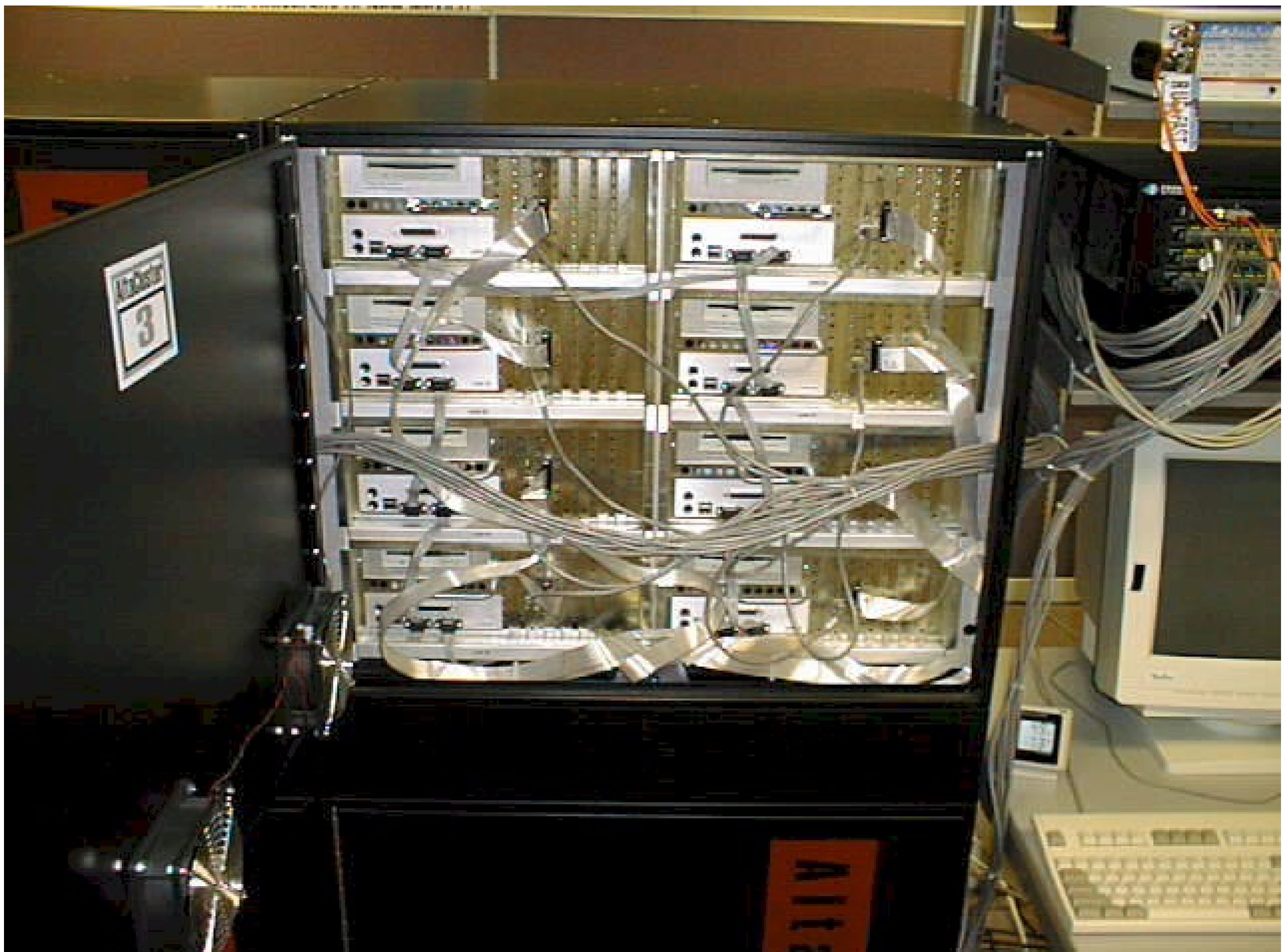
Alliance / UNM Roadrunner

- 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

CHAUTAUQUA ALLIANCE

ALBUQUERQUE High Performance Computing Center

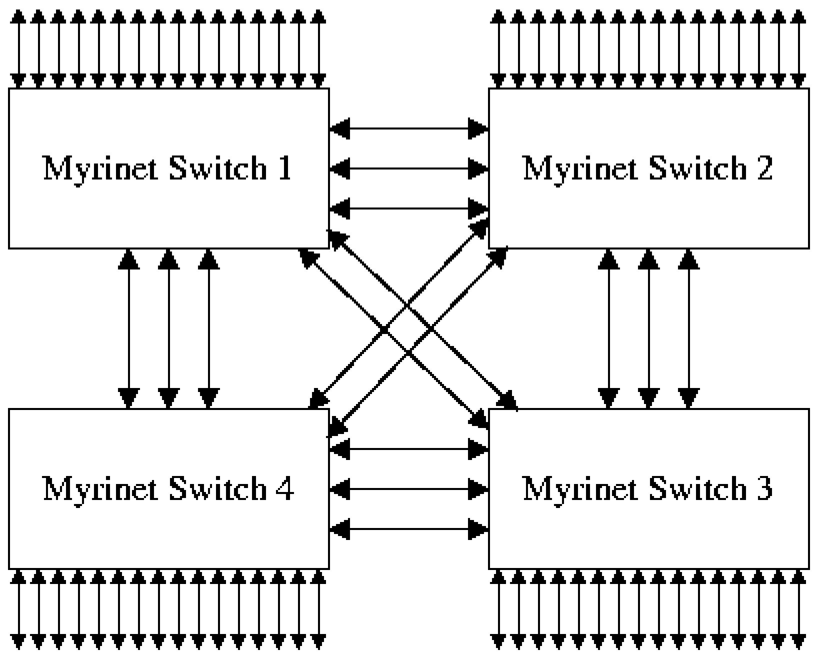
A Peek Inside Roadrunner



ALLIANCE

The University of New Mexico

Myrinet Topology



Roadrunner SuperCluster Timeline

1999

- **3/17** — Received the first 32 nodes
- **3/31** — Received the next 32 nodes
- **4/8** — Running benchmarks on 64 nodes with Fast Ethernet
- **5/3** — Running Alliance test applications and benchmarks on 64 nodes with Myrinet
- **6/1** — Begin test production
- **6/15** — Begin test production with Alliance Grid (Globus)
- **7/1** — Allocations start

Roadrunner System Software

- Redhat Linux 5.2 (6.0)
- SMP Linux kernel 2.2.10
- MPI (Argonne's MPICH 1.1.2.3)
- Portland Group Compiler Suite
- Myricom GM Drivers (1.04) and
- MPICH-GM (1.1.2.3)
- Portable Batch Scheduler (PBS)

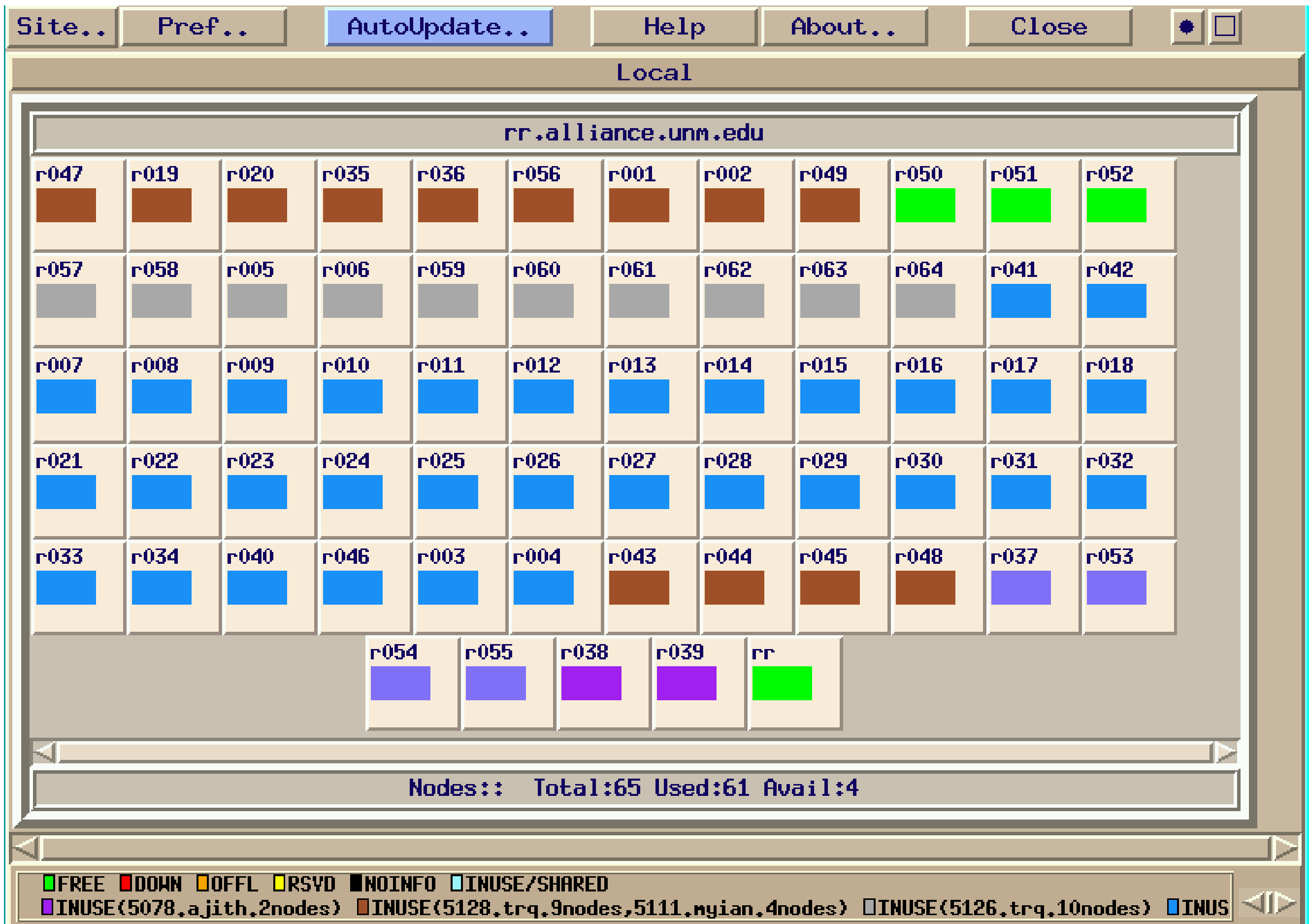
Roadrunner System Libraries

- BLAS
- LAPACK
- ScaLAPACK
- Petsc
- FFTw
- SPRNG
- Globus Grid Infrastructure

Ease of Use

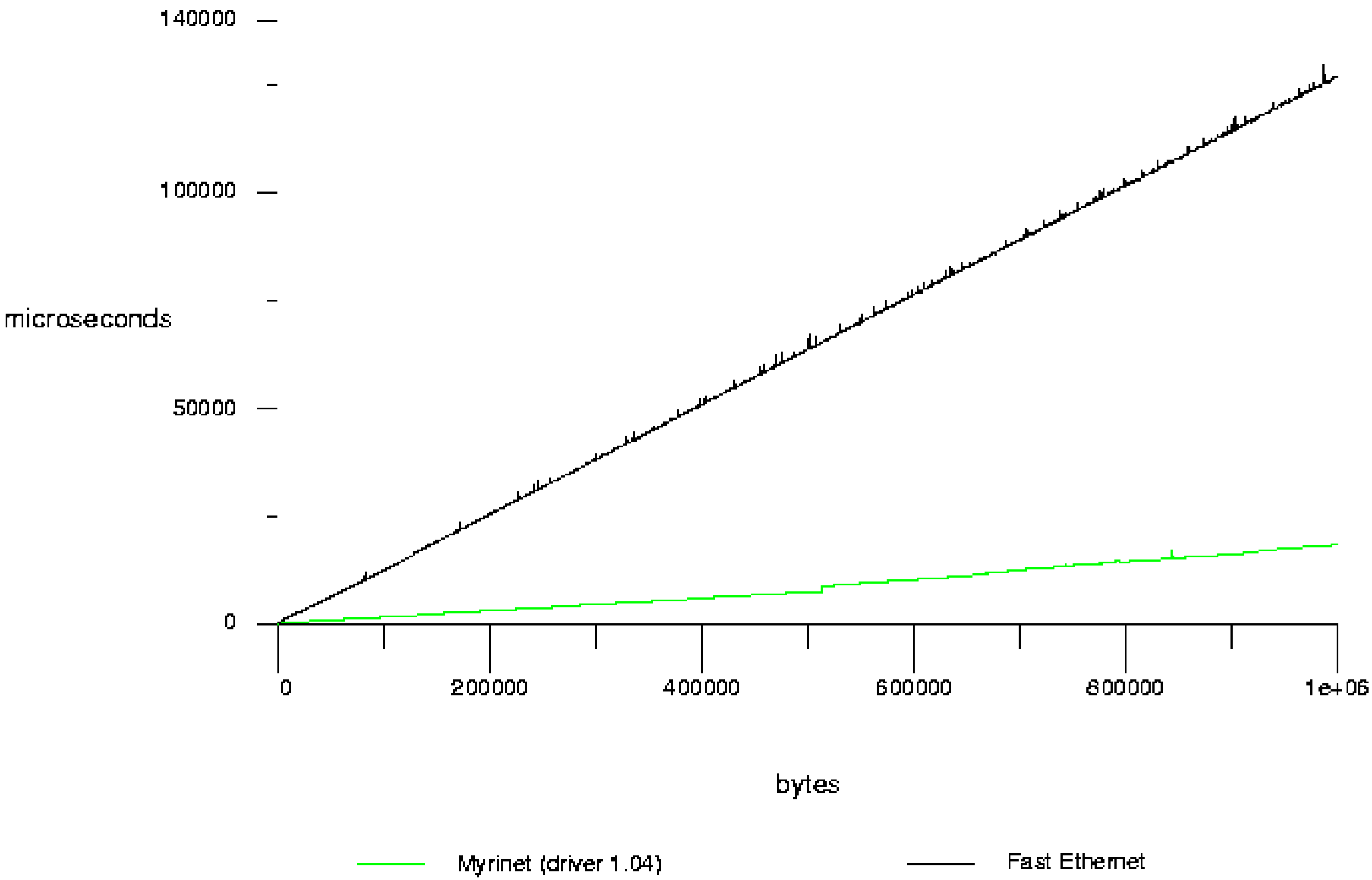
```
% ssh -l username rr.alliance.unm.edu
% mpicc -o prog helloWorld.c
% qsub -I -l nodes=64
% mpirun prog
```

Job Monitoring with PBS

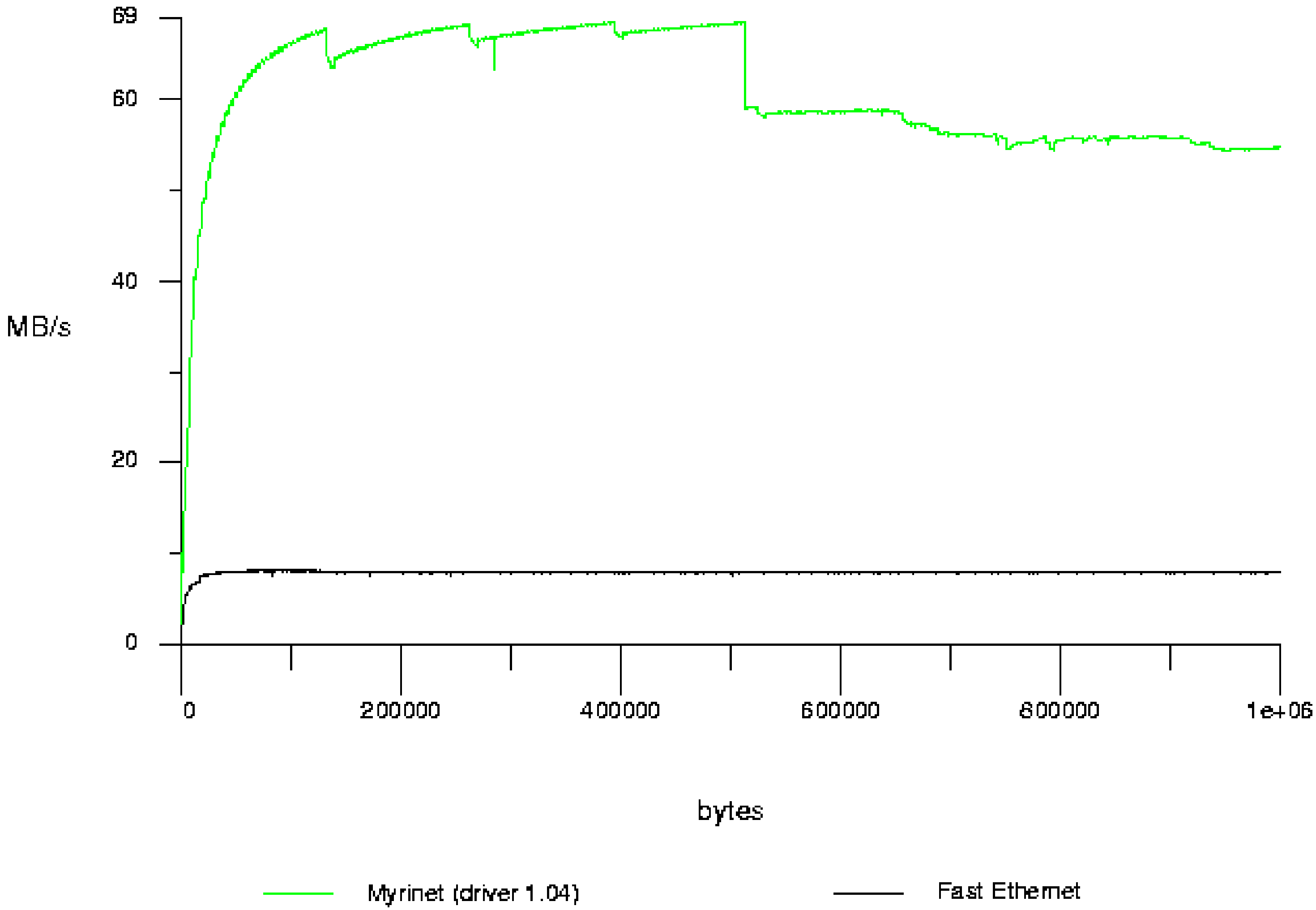


Roadrunner Performance

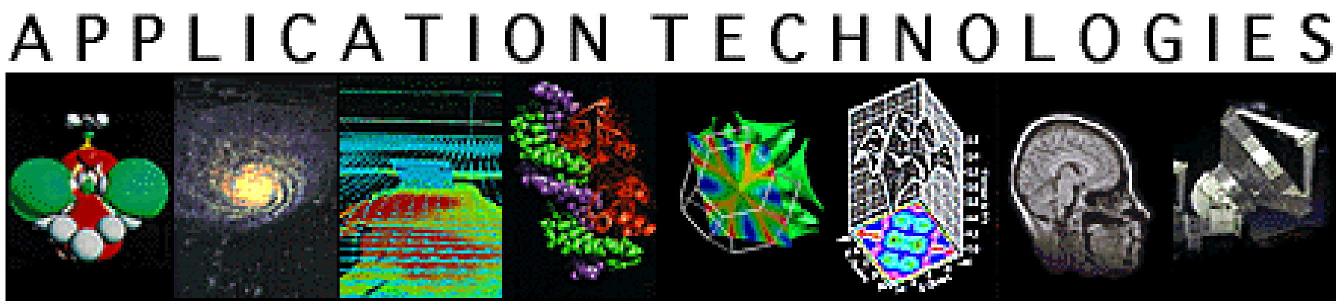
Time for Large Messages



Bandwidth for Large Messages



Alliance Applications



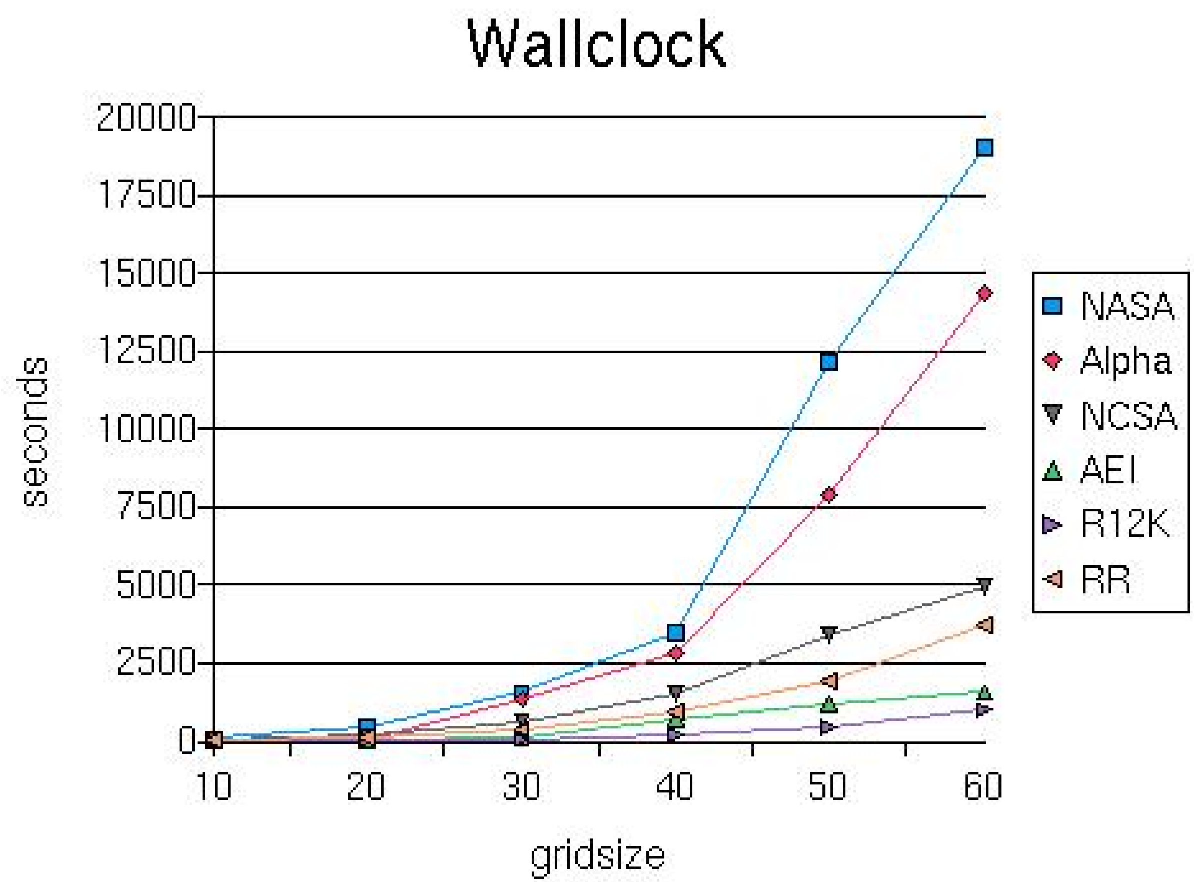
Applications: CACTUS

- 3D Numerical Relativity Toolkit for Computational Astrophysics
- (Courtesy of Gabrielle Allen and Ed Seidel)

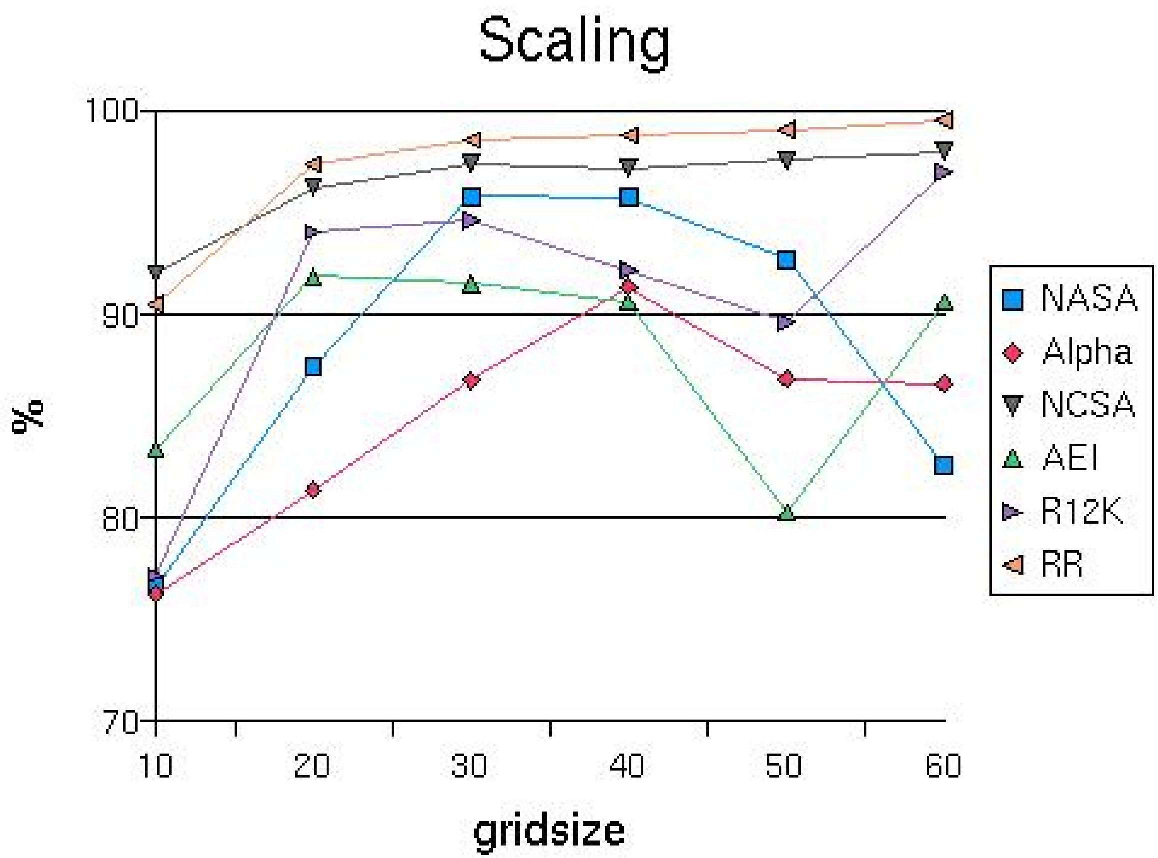
- Roadrunner performance under the Cactus application benchmark shows near-perfect scalability compared to:

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

CACTUS Performance

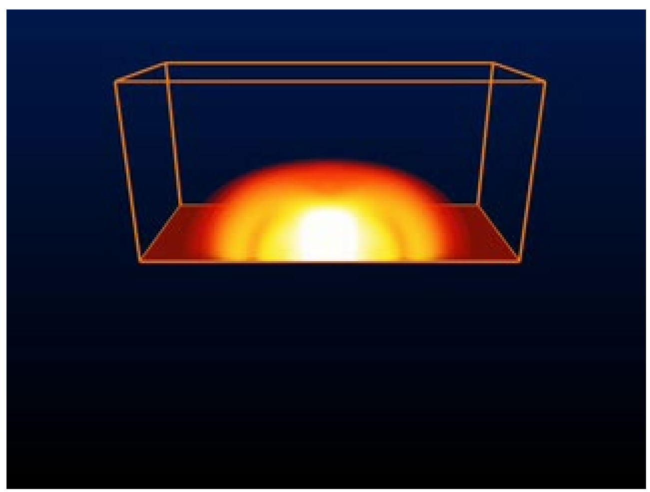


CACTUS Scaling



CACTUS: The evolution of a pure gravitational wave

A subcritical Brill wave (Amplitude=4.5), showing the Newman-Penrose Quantity as volume rendered 'glowing clouds'. The lapse function is shown as a heighth field in the bottom part of the picture.

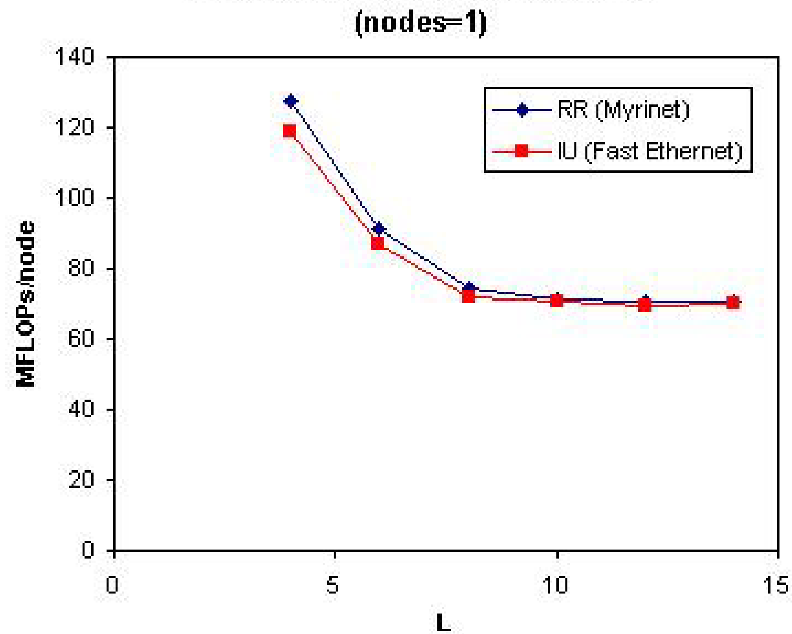


(Courtesy of Werner Benger)

Application: MILC Grand Challenge

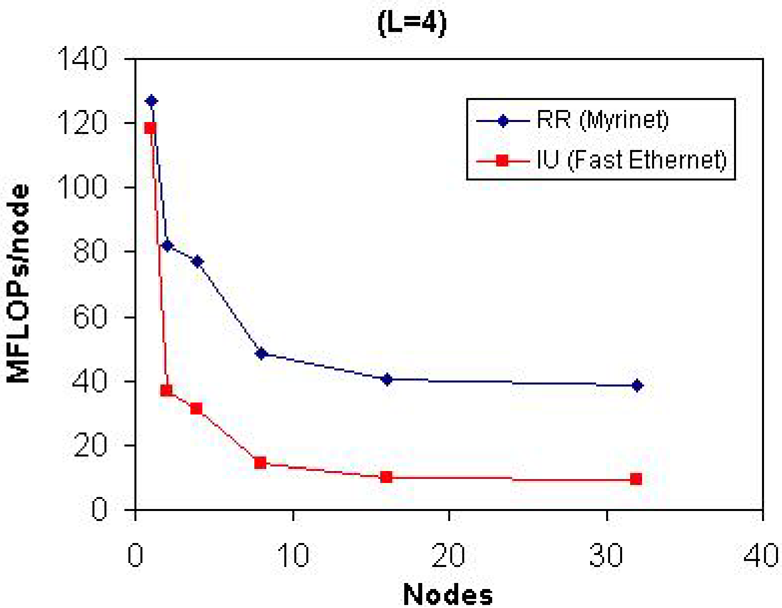
- (Courtesy of Steven Gottlieb, Indiana University, and Robert Sugar, University of California, Santa Barbara)
- The MIMD Lattice Computation (MILC) benchmark problem is a conjugate gradient algorithm for Kogut-Susskind quarks. L=4 means that there is a 4^4 piece of the domain on each node.
- The MILC benchmark was run on two Linux clusters. The Roadrunner SuperCluster and the Indiana University (IU) Physics Linux cluster, a 32-node Fast Ethernet cluster with a 350 MHz Pentium II processor, 64 MB and 4.3 GB disk per node.
- (Non-Roadrunner data - courtesy of NCSA NT Cluster Group.)

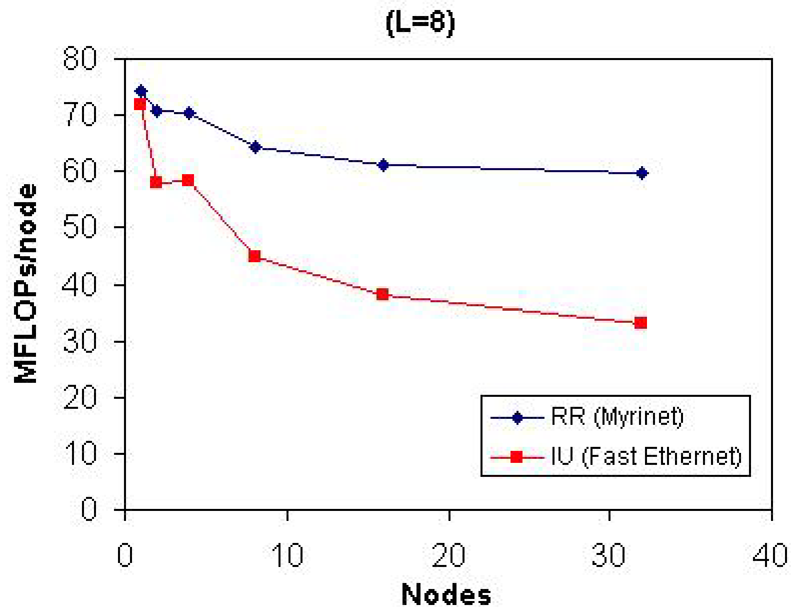
MILC Performance



- Looking at the single-node benchmarks, we see that the for small problem sizes that fit completely or mostly in the cache, the Roadrunner cluster with its 450 MHz processor is faster than the 350 MHz system. For L > 6, however, memory access becomes a limiting factor, and there is little perfomance difference.

MILC: Performance vs. Nodes

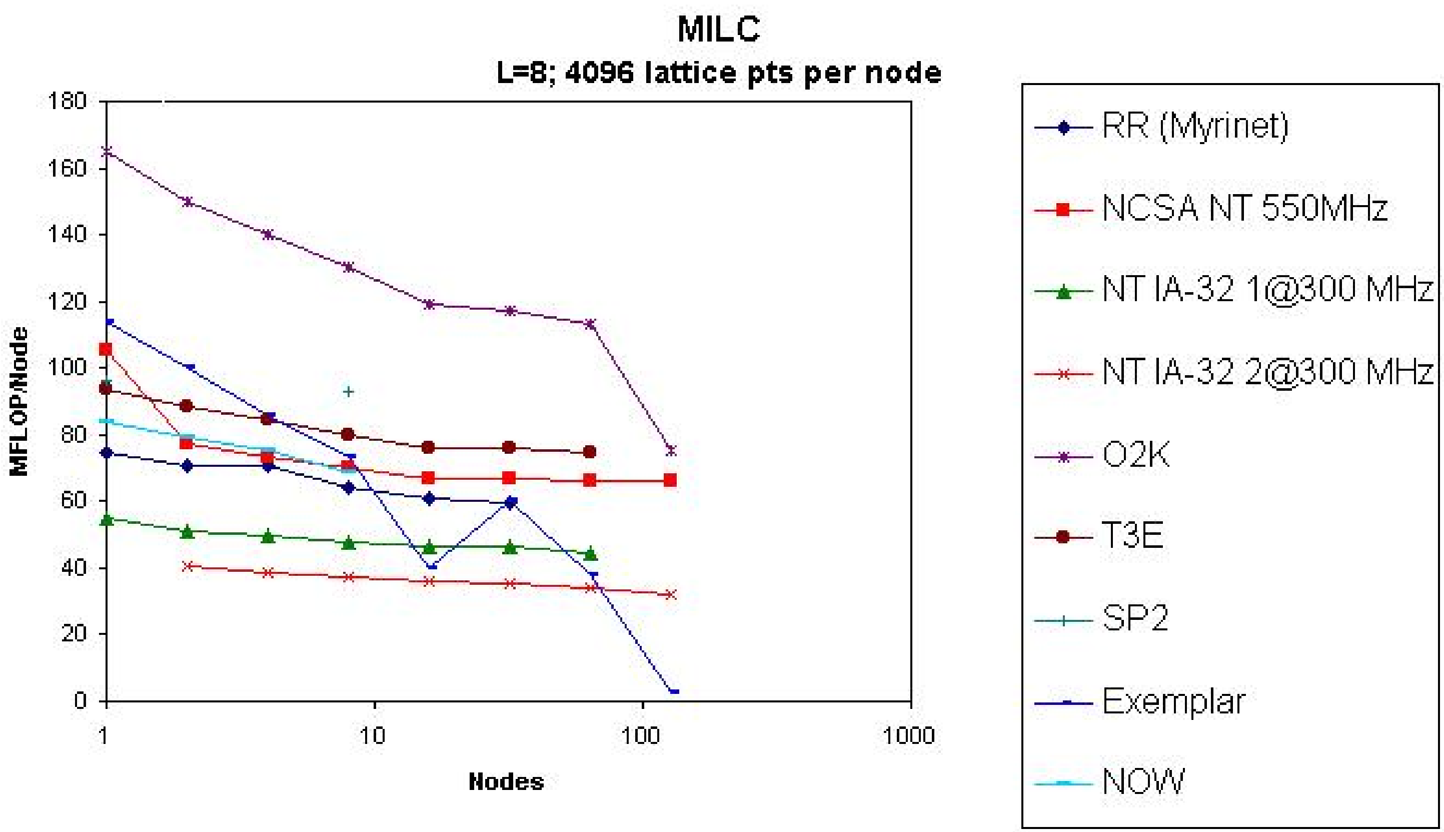




For L >= 6, the Myrinet cluster is achieving > 60 MF/node for almost all cases. For Fast Ethernet, for L >= 8, on up to 32 nodes, the performance of is near 50% of the Myrinet performance.

As a point of reference for L=8 on 16 nodes, MILC achieves 76 MF/node on an Cray T3E-900 and 119 on an SGI Origin 2000 (250 MHz).

MILC: Performance vs. Architecture



SuperClusters: What's Next?

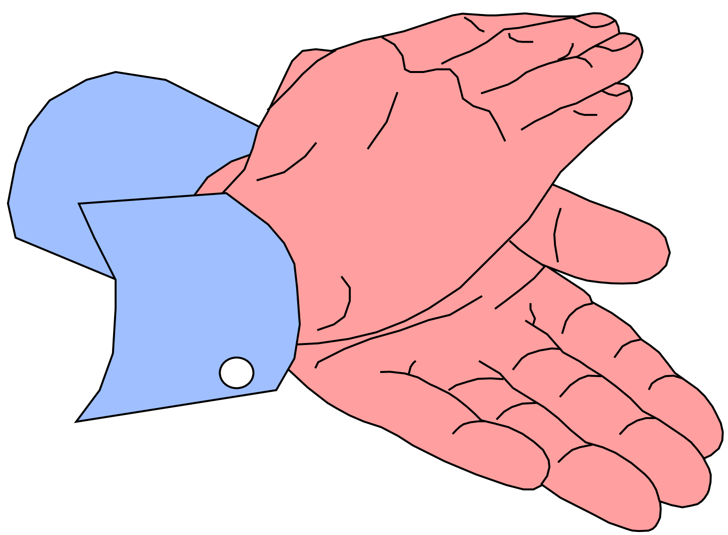
- Alliance computational scientists are needed!
- Improving advanced programming models and toolkits
- System and application tuning for SuperClusters
- Combining access and computational grids and post-web environments.

Future Directions

- TeraScale computing
- “A SuperCluster in every lab”
- Efficient use of SMP nodes
- Scalable interconnection networks
- High-performance I/O
- Advanced programming models for hybrid (SMP and Grid-based) clusters

For more information:

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



Thank you!