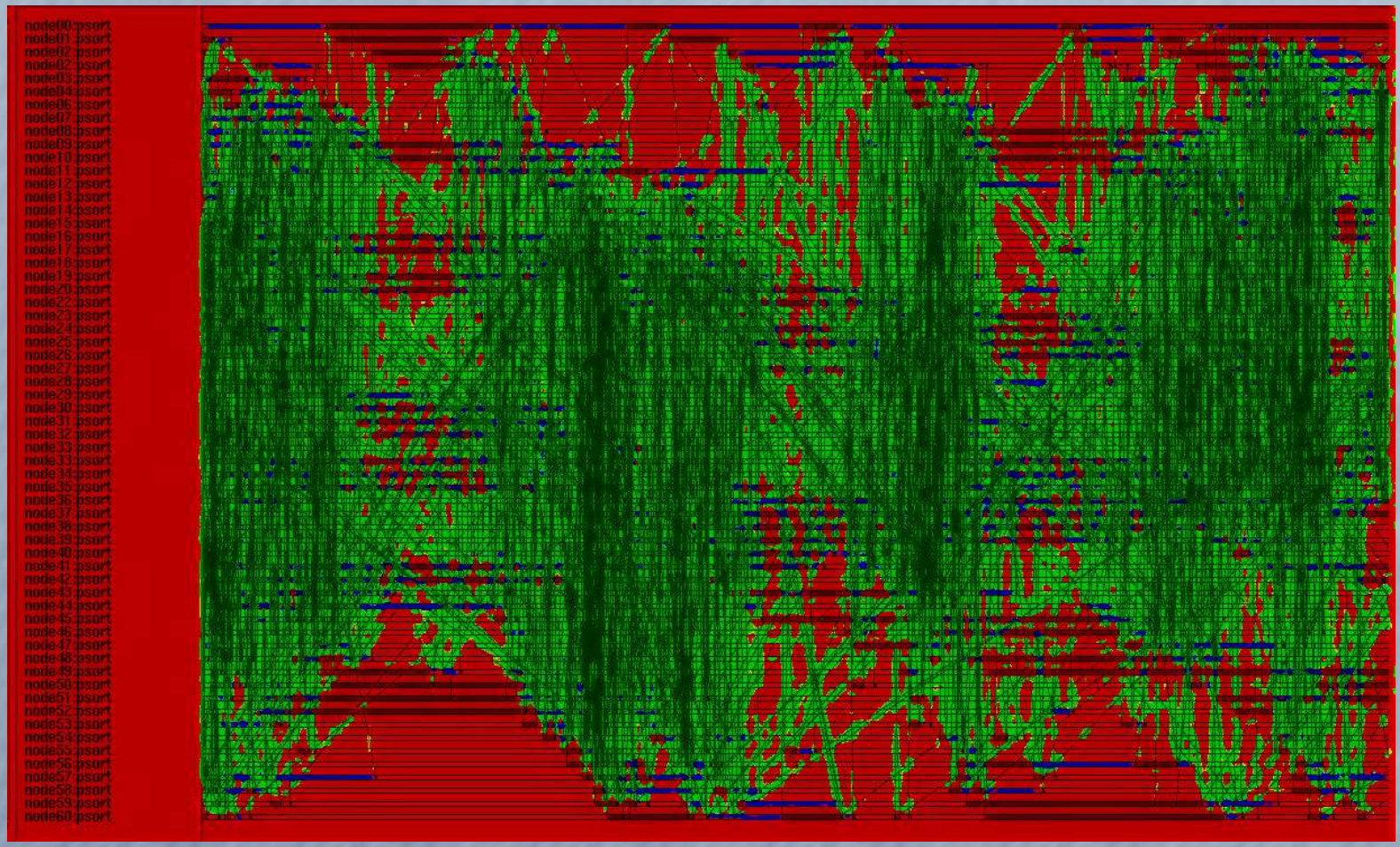
Boise State Beowulf Cluster

Amit Jain
amit@cs.boisestate.edu
Computer Science
College of Engineering
Boise State University



Project Goals

- *Establish a resource for researchers on campus with large computing needs.*
- *Help researchers convert their programs to run on the cluster.*
- *Research performance bottlenecks.*
- *Develop tools to improve the usability of clusters.*

The Beowulf Cluster Lab is funded by the National Science Foundation Major Research Infrastructure Award No. 0321233.

Cluster Specifications

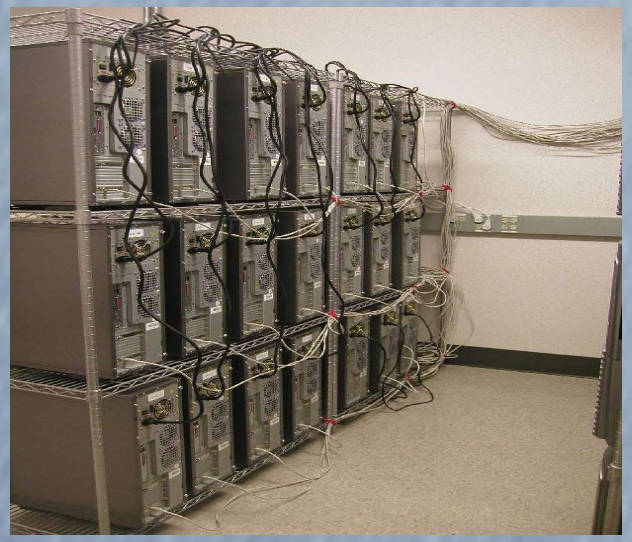
The main Beowulf Cluster (beowulf.boisestate.edu)

- 61 nodes
- 122 2.4 GHz Intel Xeon processors
- 64GB RAM
- 2.4 TB disk space
- private Gigabit network
- Gigabit connection to the campus backbone

Other clusters:

- 6 processor developmental cluster (tux.boisestate.edu)
- 32 processor teaching cluster (onyx.boisestate.edu)

Beowulf Cluster Lab



Cluster Hardware

Compute nodes (about $1400/node for 64 nodes = $90,000)

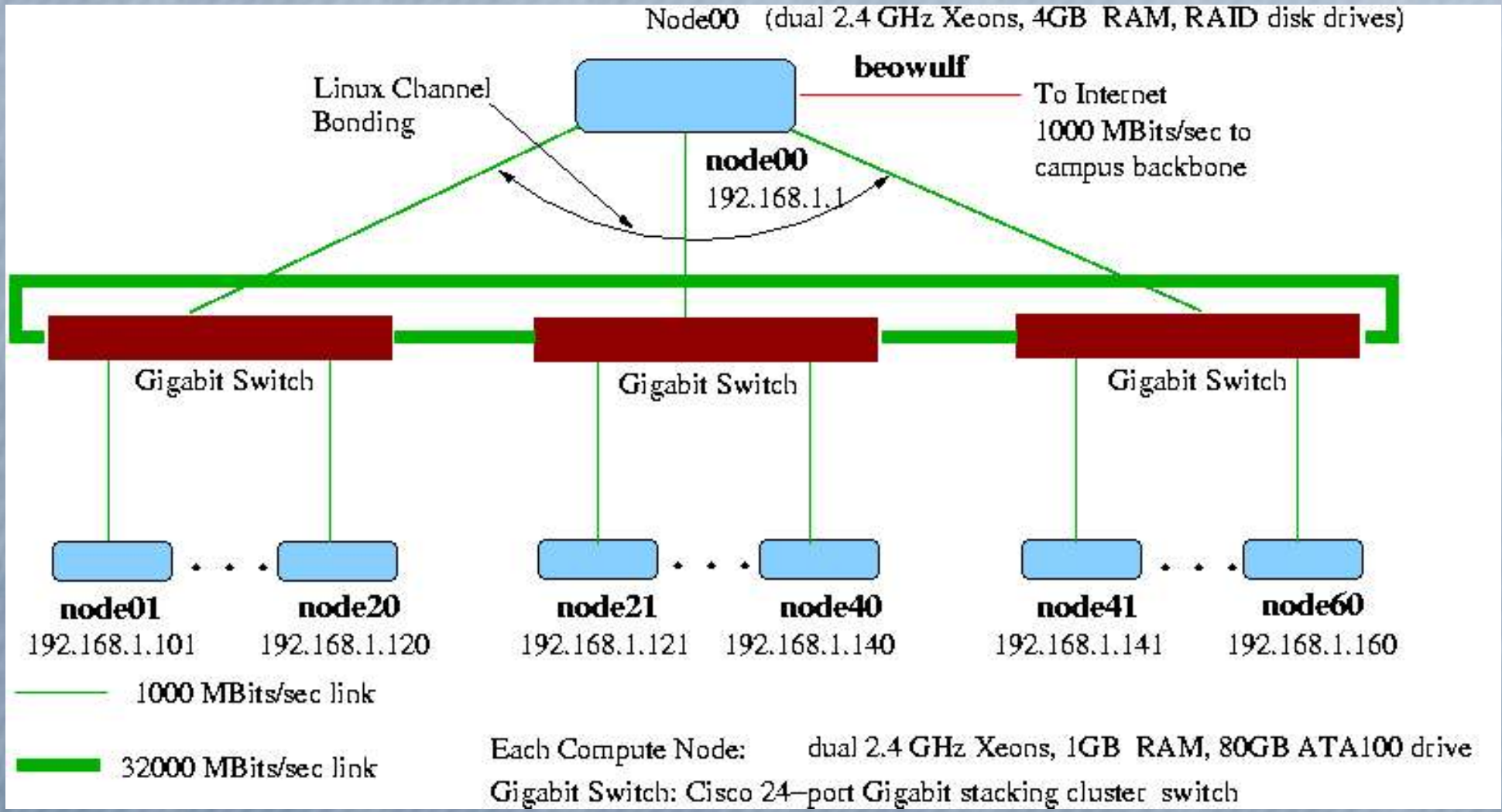
- Tyan i7505 S2665ANF dual-533 MHz FSB
- dual 2.4 Ghz Intel Xeon CPUs with 512K Cache 533MHz FSB
- 2 x 512MB Micron Technology Memory Module 184-pin DIMM PC2100 DDR 266 MHz, unbuffered, non-parity
- Samsung SP4002H disk drives, 80GB 7200RPM ATA100
- HP Broadcom NetXtreme 5782 Gigabit card
- Antec 1080 Plus AMG case with Antec True Power 550W Supply
- Master node: same, except with 4GB RAM and SATA drives with RAID

Networking: (about $12,000)

- 3 x Cisco 3750G 24-port Stacking Cluster Gigabit Switches with redundant power supply

Facilities: Liebert A/C, power setup to handle up to 300 Amps ($29,000)

Cluster Architecture



Cluster Software

- Red Hat Linux 9.0 with custom 2.4.24 -bigmem SMP kernel (Fedora Core 1 Linux with stock kernel on the cluster used for teaching)
- Portable Batch Scheduling for job scheduling
- Parallel Programming Libraries and Tools
  - PVM, MPICH/MPI, LAM/MPI
  - XPVM and XMPI
- Portland Group Cluster Development Toolkit
  - HPF, Fortran 90, Fortran 77, C, C++
  - Parallel graphical debugger
  - Parallel graphical profiler
- GNU C, C++ and Fortran 77 compilers and related tool set like ddd (Data Display Debugger)
- Full suite of other tools available under Linux.

Cluster Setup Experiences

- YACI (Yet Another Cluster Installer) was used for automated installation. The 61- node cluster went from bare disks to fully operational in 12 minutes! YACI is available from Larwence Livermore National Lab.
- Design choice to go with boxes instead of blades since cooling boxes is easier and real estate was a relatively smaller issue.
- Evaluated AMD Athlon, AMD Opteron, Intel Xeon for Performance/Power/Price (PPP) factor to choose Intel Xeons.
- Chose to go with a regular PC assembler rather than a “cluster” company to keep costs down and have more control of what goes in each node.

People

- *Faculty*: Amit Jain (Computer Science) and Paul Michaels (Geophysics)
- *Graduate Students*: Kevin Nuss, Hongyi Hu and Mason Vail
- *Undergraduate Students*: Joey Mazzarelli, Brady Catherman, Luke Hindman, Charles Paulson, Jason Main and Oralee Nudson.

The project uses a model of teaming up computer scientists with researchers from other fields to create a synergistic environment.

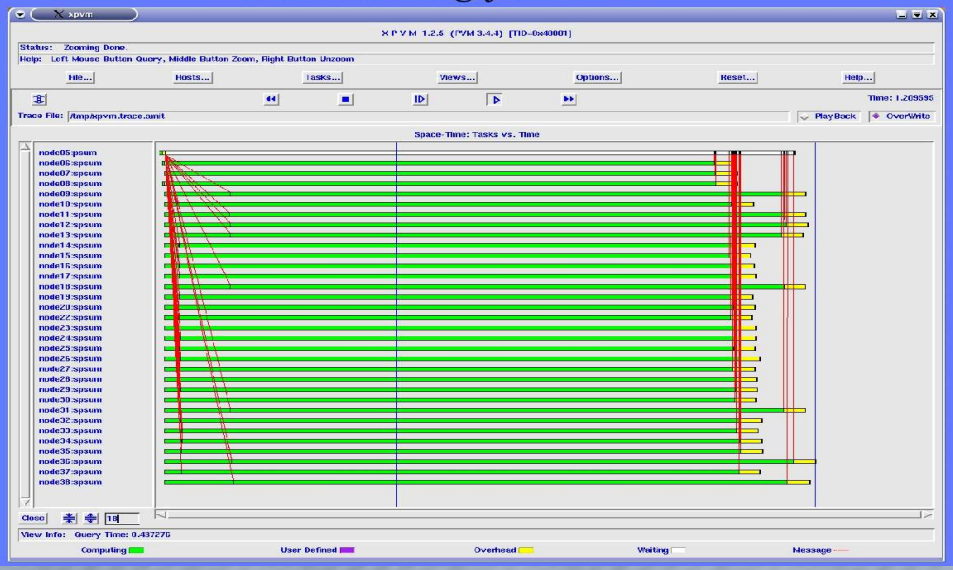
Projects

Some applications running on the cluster.

- *Air Quality Modeling*. Paul Dawson (Mechanical Engineering), Kevin Nuss and Charles Paulson.
- *Modeling of Ocean Currents*. Jodi Mead (Mathematics) and Hongyi Hu.
- *Waveform Relaxation*. Barbara Zubik-Kowal (Mathematics) and Hongyi Hu.
- *Hydraulic Tomography*. Tom Clemo (Geophysics) and Kevin Nuss.
- *Bioinformatics: Bayesian Analysis of Phylogeny*. James Smith (Biology) and Amit Jain.
- *Basic Seismic Utilities* package. Paul Michaels (Geophysics) and Amit Jain.
- *Biologically Inspired Computing*. Crowley Davis Research (private company)

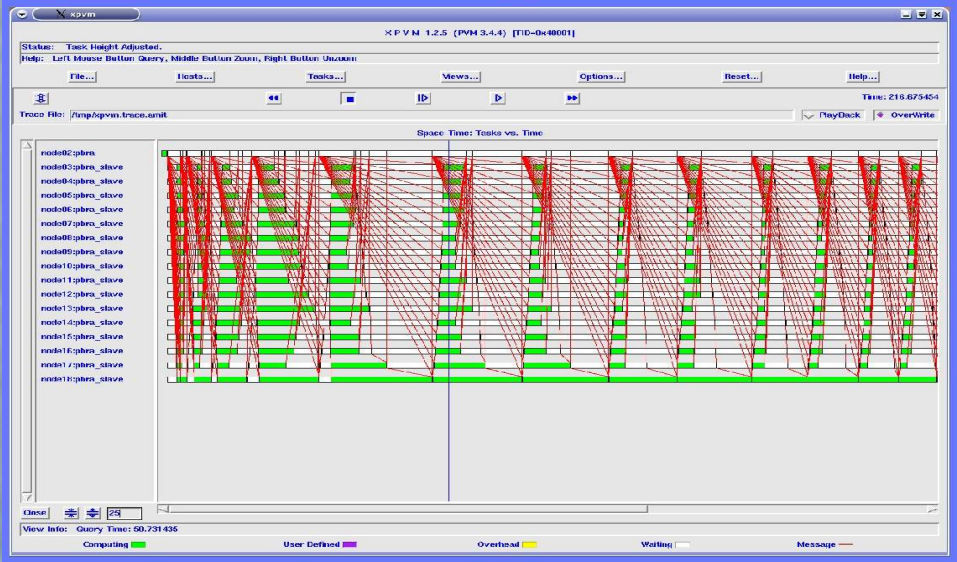
Design Patterns

Embarrassingly Parallel

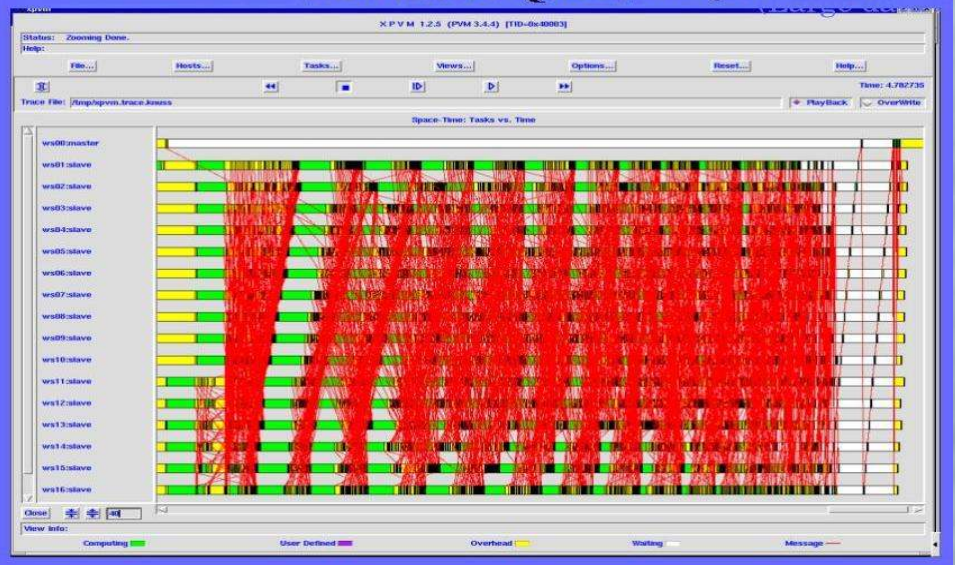


Gossip (All-to-all communication)

Load Balancing Problem



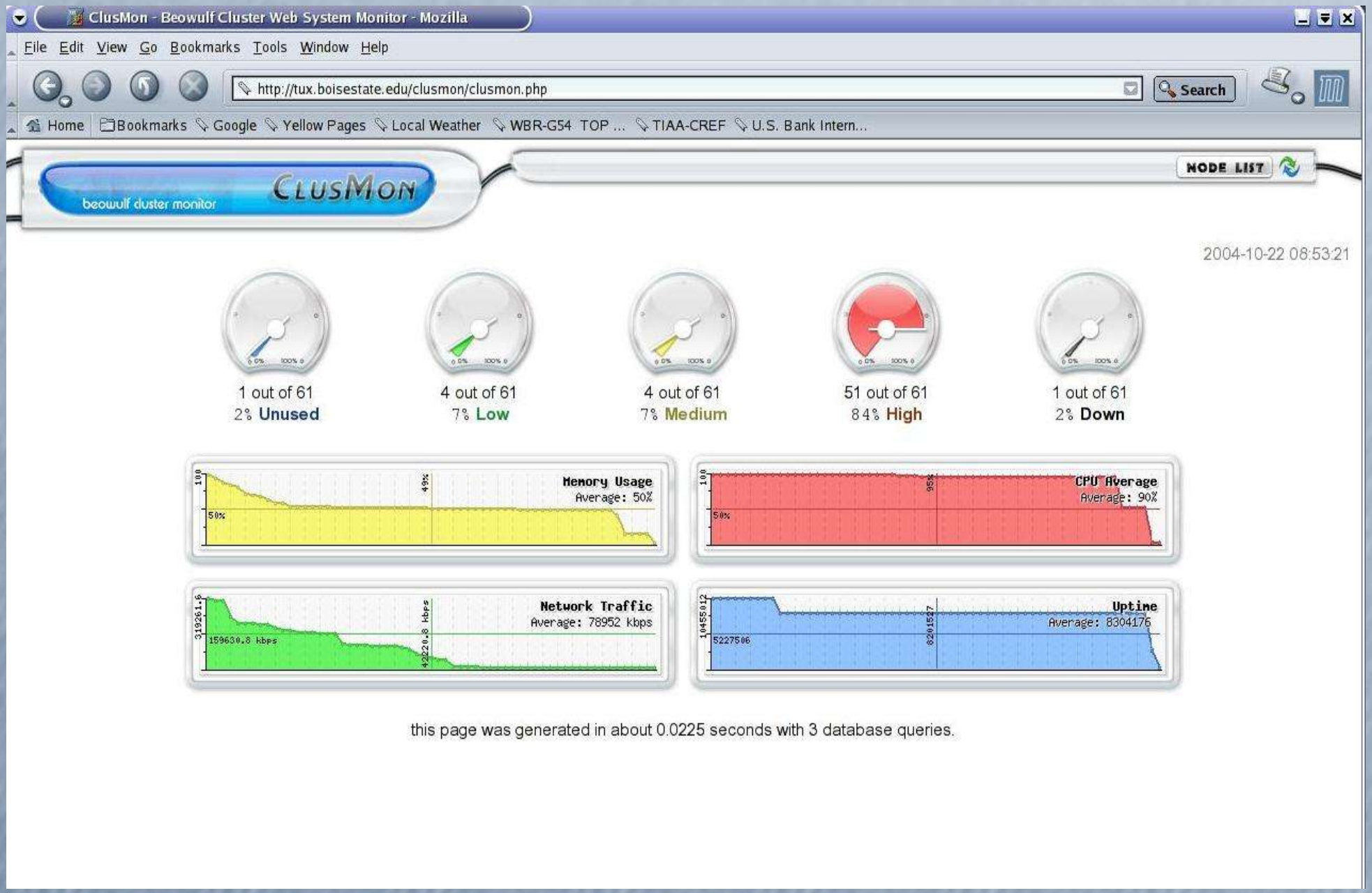
Distributed Queue (Large data set)



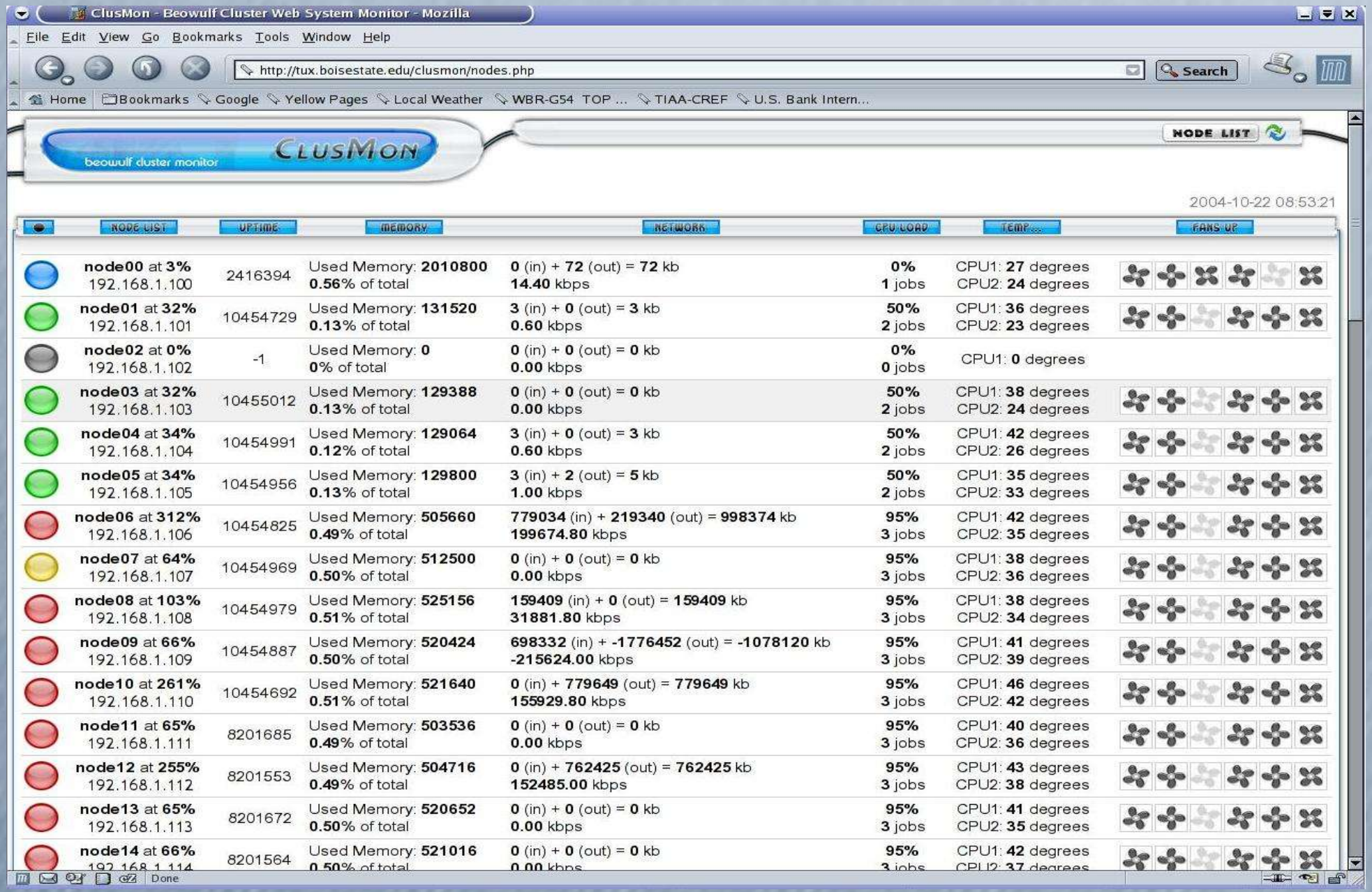
Projects (contd.)

- *Clusmon.* A comprehensive web-based cluster monitoring software. (Joey Mazzarelli, Computer Science senior)
- *Remote Power Control.* A cluster of smart power strips to enable remote hard power on/off, cascaded power on/off etc. (Brady Catherman, Computer Science junior)
- *Parallel Shell*. A more capable parallel shell for system administration. (Mason Vail, Computer Science graduate student)

Clusmon: Cluster Monitor



Clusmon: Cluster Monitor



NODE LIST	UPTIME	MEMORY	NETWORK	CPU LOAD	TEMP...	FANS UP
node00 at **3%** 192.168.1.100	2416394	Used Memory: **2010800** **0.56%** of total	**0** (in) + **72** (out) = **72** kb **14.40** kbps	**0%** **1** jobs	CPU1: **27** degrees CPU2: **24** degrees	
node01 at **32%** 192.168.1.101	10454729	Used Memory: **131520** **0.13%** of total	**3** (in) + **0** (out) = **3** kb **0.60** kbps	**50%** **2** jobs	CPU1: **36** degrees CPU2: **23** degrees	
node02 at **0%** 192.168.1.102	-1	Used Memory: **0** **0%** of total	**0** (in) + **0** (out) = **0** kb **0.00** kbps	**0%** **0** jobs	CPU1: **0** degrees	
node03 at **32%** 192.168.1.103	10455012	Used Memory: **129388** **0.13%** of total	**0** (in) + **0** (out) = **0** kb **0.00** kbps	**50%** **2** jobs	CPU1: **38** degrees CPU2: **24** degrees	
node04 at **34%** 192.168.1.104	10454991	Used Memory: **129064** **0.12%** of total	**3** (in) + **0** (out) = **3** kb **0.60** kbps	**50%** **2** jobs	CPU1: **42** degrees CPU2: **26** degrees	
node05 at **34%** 192.168.1.105	10454956	Used Memory: **129800** **0.13%** of total	**3** (in) + **2** (out) = **5** kb **1.00** kbps	**50%** **2** jobs	CPU1: **35** degrees CPU2: **33** degrees	
node06 at **312%** 192.168.1.106	10454825	Used Memory: **505660** **0.49%** of total	**779034** (in) + **219340** (out) = **998374** kb **199674.80** kbps	**95%** **3** jobs	CPU1: **42** degrees CPU2: **35** degrees	
node07 at **64%** 192.168.1.107	10454969	Used Memory: **512500** **0.50%** of total	**0** (in) + **0** (out) = **0** kb **0.00** kbps	**95%** **3** jobs	CPU1: **38** degrees CPU2: **36** degrees	
node08 at **103%** 192.168.1.108	10454979	Used Memory: **525156** **0.51%** of total	**159409** (in) + **0** (out) = **159409** kb **31881.80** kbps	**95%** **3** jobs	CPU1: **38** degrees CPU2: **34** degrees	
node09 at **66%** 192.168.1.109	10454887	Used Memory: **520424** **0.50%** of total	**698332** (in) + **-1776452** (out) = **-1078120** kb **-215624.00** kbps	**95%** **3** jobs	CPU1: **41** degrees CPU2: **39** degrees	
node10 at **261%** 192.168.1.110	10454692	Used Memory: **521640** **0.51%** of total	**0** (in) + **779649** (out) = **779649** kb **155929.80** kbps	**95%** **3** jobs	CPU1: **46** degrees CPU2: **42** degrees	
node11 at **65%** 192.168.1.111	8201685	Used Memory: **503536** **0.49%** of total	**0** (in) + **0** (out) = **0** kb **0.00** kbps	**95%** **3** jobs	CPU1: **40** degrees CPU2: **36** degrees	
node12 at **255%** 192.168.1.112	8201553	Used Memory: **504716** **0.49%** of total	**0** (in) + **762425** (out) = **762425** kb **152485.00** kbps	**95%** **3** jobs	CPU1: **43** degrees CPU2: **38** degrees	
node13 at **65%** 192.168.1.113	8201672	Used Memory: **520652** **0.50%** of total	**0** (in) + **0** (out) = **0** kb **0.00** kbps	**95%** **3** jobs	CPU1: **41** degrees CPU2: **35** degrees	
node14 at **66%** 192.168.1.114	8201564	Used Memory: **521016** **0.50%** of total	**0** (in) + **0** (out) = **0** kb **0.00** kbps	**95%** **3** jobs	CPU1: **42** degrees CPU2: **37** degrees	

Remote Power Control

Cluster Statistics

- 1179 jobs since July, adding up to about 88000 CPU-hours.
- Average CPU temperatures: 77F at low load and 100F at full load. The A/C is set to 65F with tolerance of 4F.
- Hardware failures: Extremely low...
  - One disk drive failed right after installation.
  - The memory for one node failed.
- Only one unscheduled "downtime" in the last three months. The A/C compressor was cycling more than the factory set limit. As a result, it shut itself off. The CPU temperatures still remained below 115F after several hours! (as the air flow was maintained) The cluster was shut down as a precaution. The solution was to simply set a higher tolerance (4 degrees instead of 2 degrees)
- The experiences gained in this project were used to help Geophysics set-up a 10 processor cluster and Mathematics a 20 processor cluster.

Further Work

- Integrate Beowulf clusters with Condor grids.
- Develop a complete catalogue of programs illustrating each design pattern in PVM and MPI.
- Continue to team with researchers to help get their code up and running on clusters.