Introduction to Clusters

Rocks-A-Palooza II

Ground Rules

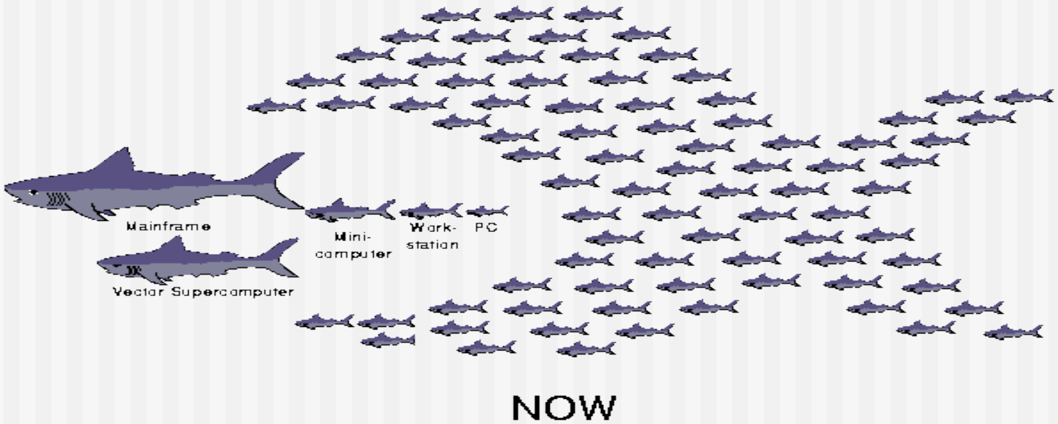
- Interrupt me!
  - If you have a question and need more information
  - Would like me to go into more detail, or skip over some material
  - I already know this stuff
- Tell me to slow down
  - I tend to talk very fast
  - We have about 200 slides to go through (in six hours)
    - But we will skip some, and other are very short
    - We have plenty of time
    - Last session will be unstructured (you've been warned)
- I don't have to use my slides
  - This workshop is for you
  - Other topics are welcome (but also see track2)
- Tomorrow we will go over some of Track2

Introduction

A brief introduction to clustering and Rocks

Brief History of Clustering

(very brief)



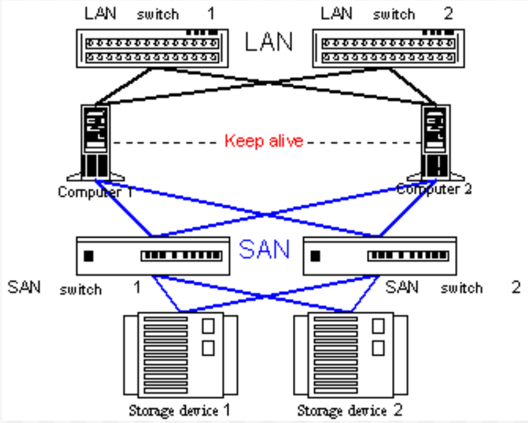
- NOW pioneered the vision for clusters of commodity processors.
  - David Culler (UC Berkeley) started early 90's
  - SunOS / SPARC
  - First generation of Myrinet, active messages
  - Glunix (Global Unix) execution environment
- Beowulf popularized the notion and made it very affordable.
  - Tomas Sterling, Donald Becker (NASA)
  - Linux

Definition: Beowulf

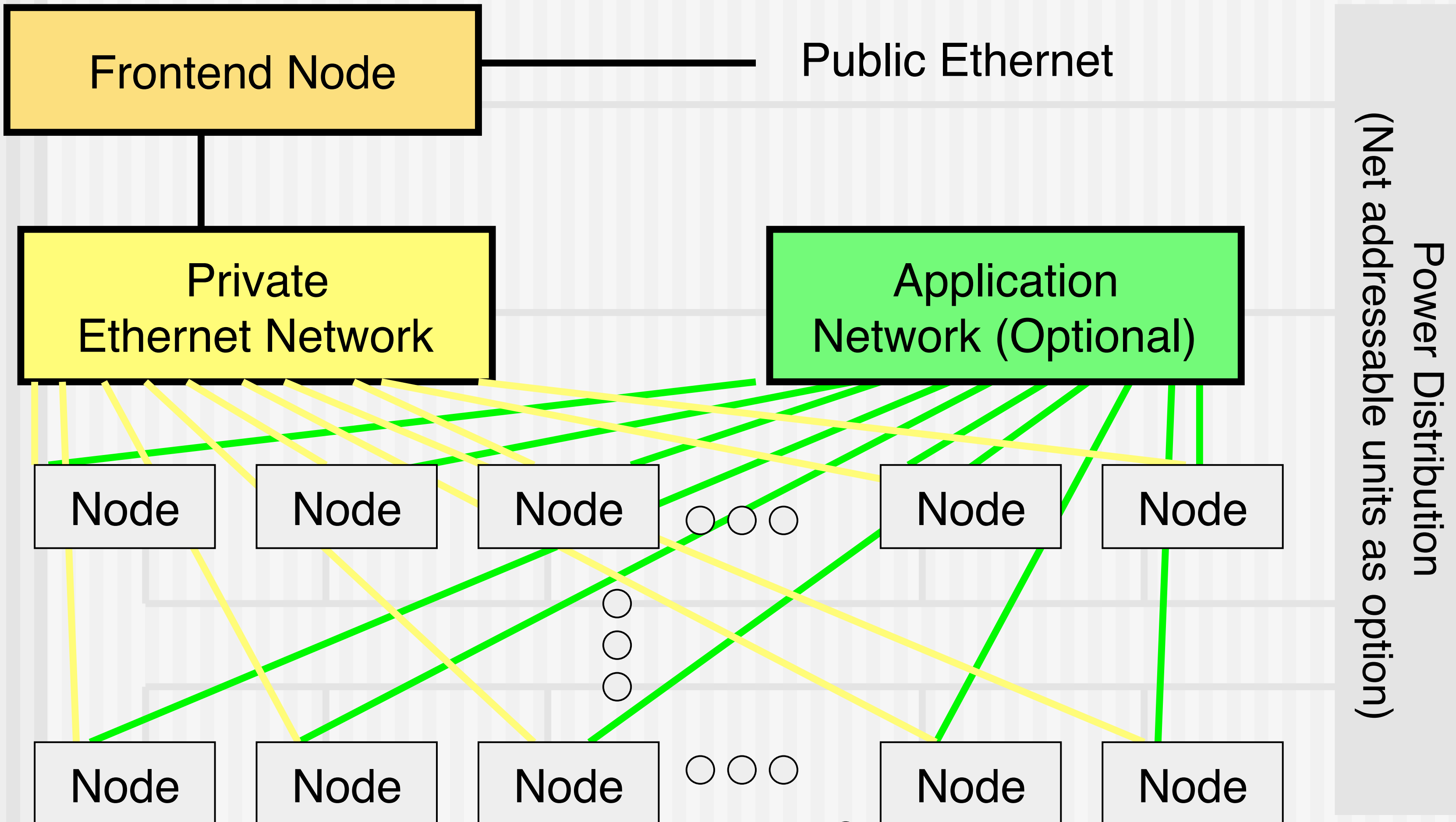
- Collection of *commodity PCs* running an *opensource* operating system with a *commodity network*
- Network is usually Ethernet, although non-commodity networks are sometimes called Beowulfs
- Come to mean any Linux cluster
- www.beowulf.org

Types of Clusters

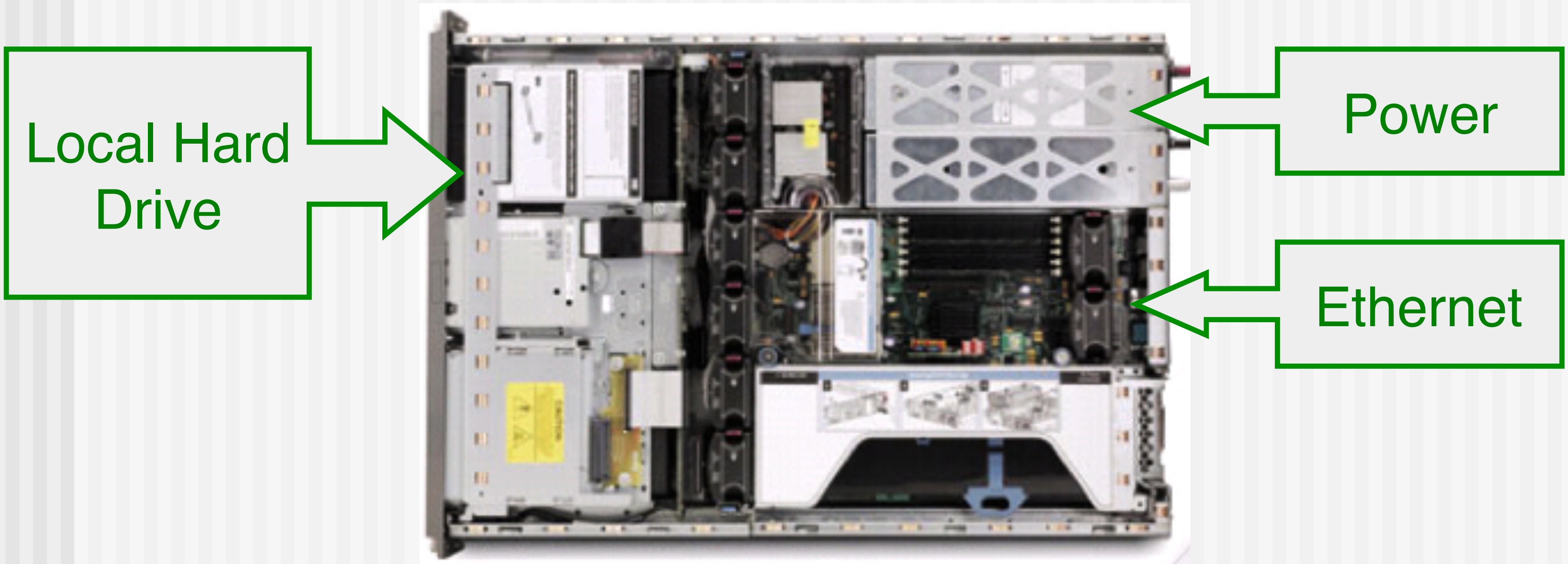
- Highly Available (HA)
  - Generally small, less than 8 nodes
  - Redundant components
  - Multiple communication paths
  - This is not Rocks
- Visualization Clusters
  - Each node drives a display
  - OpenGL machines
  - This is not core Rocks
  - But, there is a Viz Roll
- Computing (HPC Clusters)
  - AKA Beowulf
  - This is the core of Rocks



Definition: HPC Cluster Architecture



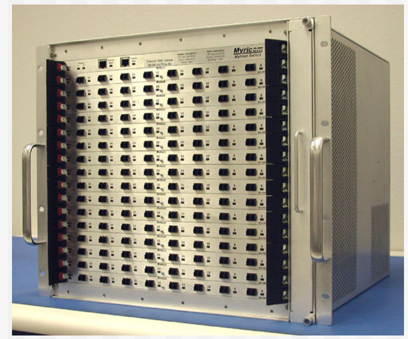
Minimum Components



i386 (Pentium/Athlon)
x86_64 (Opteron/EM64T)
ia64 (Itanium) server

Optional Components

- High-performance network
  - Myrinet
  - Infiniband (Infinicon or Voltaire)



- Network-addressable power distribution unit

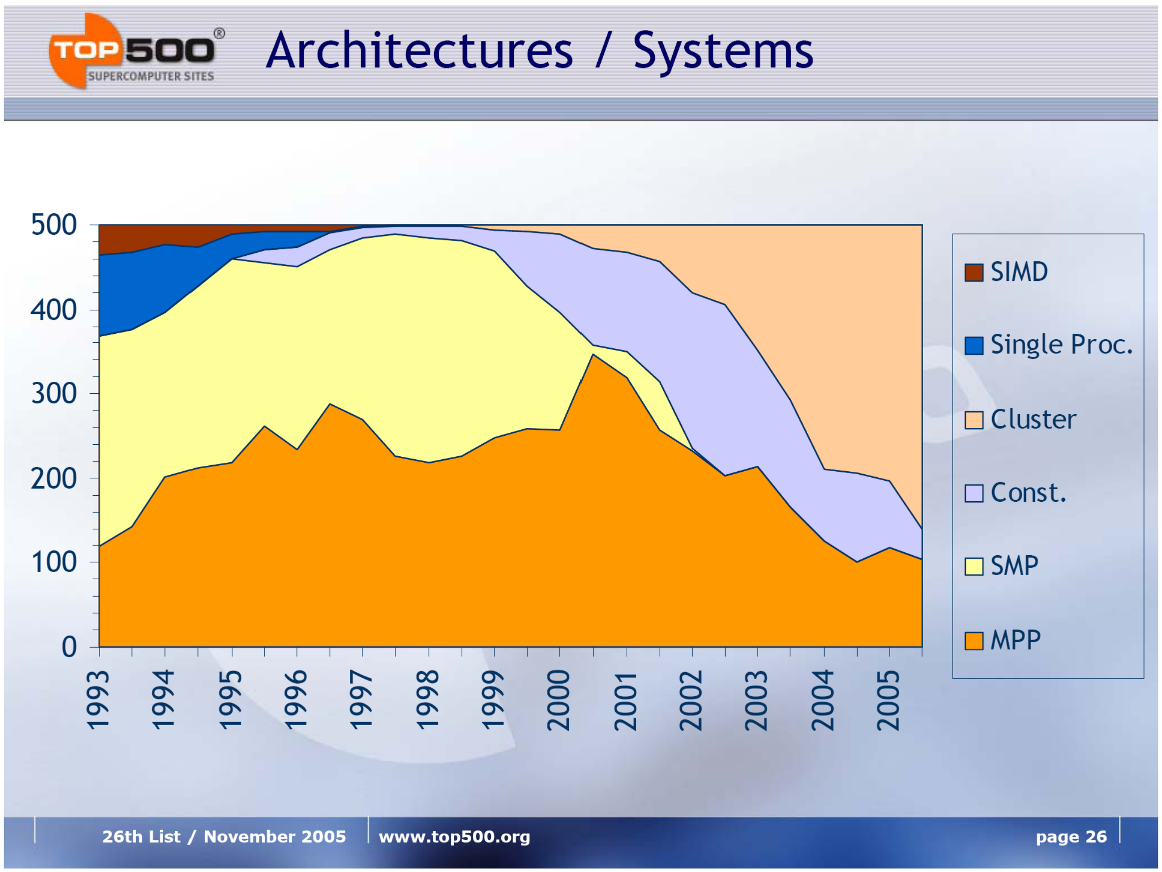
- Keyboard/video/mouse network not required
  - Non-commodity
  - How do you manage your management network?

Cluster Pioneers

- In the mid-1990s, Network of Workstations project (UC Berkeley) and the Beowulf Project (NASA) asked the question:

Can You Build a High Performance Machine From Commodity Components?

The Answer is: Clusters now Dominate High-End Computing



Erich Strohmaier: http://www.top500.org/lists/2005/11/TOP500_Nov2005_Highlights.pdf

Case Scenario

What does 128-node cluster look like?

128 Node cluster [In 6 months]

- Frontend
  - Dual-Processor (e.g. Xeon/Opteron 3.x Ghz) [4P= Dual-Socket Dual-Core @ 2.x GHz]
  - 2GB RAM [4GB/8GB]
  - Dual On board Gigabit Ethernet
  - 500 GB Storage (2 x 250GB SATA Drives) [1TB Storage 2x500GB]
  - CDROM
  - On board video
- Compute Nodes
  - Dual-Processor [4P= Dual-Socket Dual-Core @ 2.x GHz]
  - 2GB RAM [4GB/8GB]
  - Dual On board Gigabit Ethernet
  - 250 GB Storage
  - CDROM [No CDROM]
  - On board video

Additional Components

- Machine Racks
- Power
  - Network addressable power units
  - Power cords
- Network
  - 48 Port gigabit Ethernet switches
  - CAT5e cables
- VGA monitor, PC101 keyboard, mouse

SPEC Benchmark

Processor	GHz	SPECfp	SPECfp Rate	Price
Athlon 64 X2 (1S/2C)	2.4	1634	33.1	649
Pentium 4 EE (1S/2C)	3.7	2236	37.7	1059
Opteron 285 (2S/4C)	2.6	2095	82.4	1049
Opteron 254 (2S/2C)	2.8	2223	53.8	674
Pentium 4 Xeon (2S/2C)	3.6	1868	33.0	640
Itanium 2 (2S/2C)	1.6	2712	51.5	1199
Power5+ (4C)	1.9	3007	133	????

Processors

PowerPC

Power5

Itanium 2

Xeon

Opteron

Opteron

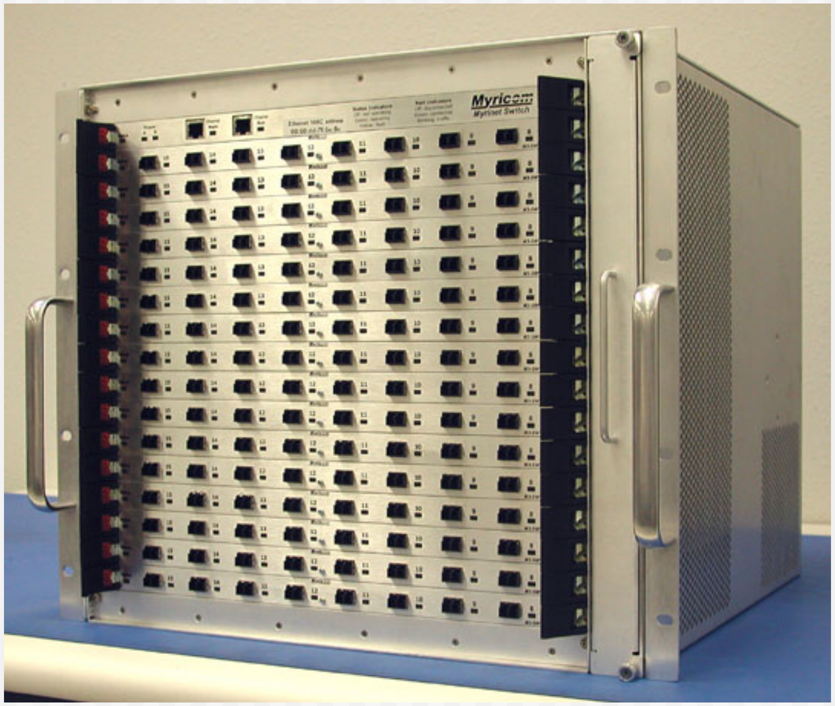
Rank	Site	Computer	Processors	Year	R_{max}	R_{peak}
1	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution IBM	131072	2005	280600	367000
2	IBM Thomas J. Watson Research Center United States	BGW - eServer Blue Gene Solution IBM	40960	2005	91290	114688
3	DOE/NNSA/LLNL United States	ASC Purple - eServer pSeries p5 575 1.9 GHz IBM	10240	2005	63390	77824
4	NASA/Ames Research Center/NAS United States	Columbia - SGI Altix 1.5 GHz, Voltaire Infiniband SGI	10160	2004	51870	60960
5	Sandia National Laboratories United States	Thunderbird - PowerEdge 1850, 3.6 GHz, Infiniband Dell	8000	2005	38270	64512
6	Sandia National Laboratories United States	Red Storm Cray XT3, 2.0 GHz Cray Inc.	10880	2005	36190	43520
7	The Earth Simulator Center Japan	Earth-Simulator NEC	5120	2002	35860	40960
8	Barcelona Supercomputer Center Spain	MareNostrum - JS20 Cluster, PPC 970, 2.2 GHz, Myrinet IBM	4800	2005	27910	42144
9	ASTRON/University Groningen Netherlands	Stella - eServer Blue Gene Solution IBM	12288	2005	27450	34406.4
10	Oak Ridge National Laboratory United States	Jaguar - Cray XT3, 2.4 GHz Cray Inc.	5200	2005	20527	24960

Interconnects

- Weak interconnect
  - Gigabit Ethernet
- Strong interconnect
  - Myrinet ($800-$1000/ port)
  - Infiniband ($800-$1000 / port)
- Dual Xeon compute node
  - Node cost $2000
  - All of the above interconnects = $2500
- One of the surprising, but often essential, costs of a cluster

Myrinet

- Long-time interconnect vendor
  - Delivering products since 1995
- Deliver single 128-port full bisection bandwidth switch
- Performance (Myrinet MX):
  - Latency: 2.7 us
  - Bandwidth: 245 MB/s
  - Cost/port (based on 64-port configuration): $1000
    - Switch + NIC + cable
    - http://www.myri.com/myrinet/product_list.html
- Newer Myrinet 10G is Dual Protocol
  - 10GigE or 10G Myrinet



Myrinet

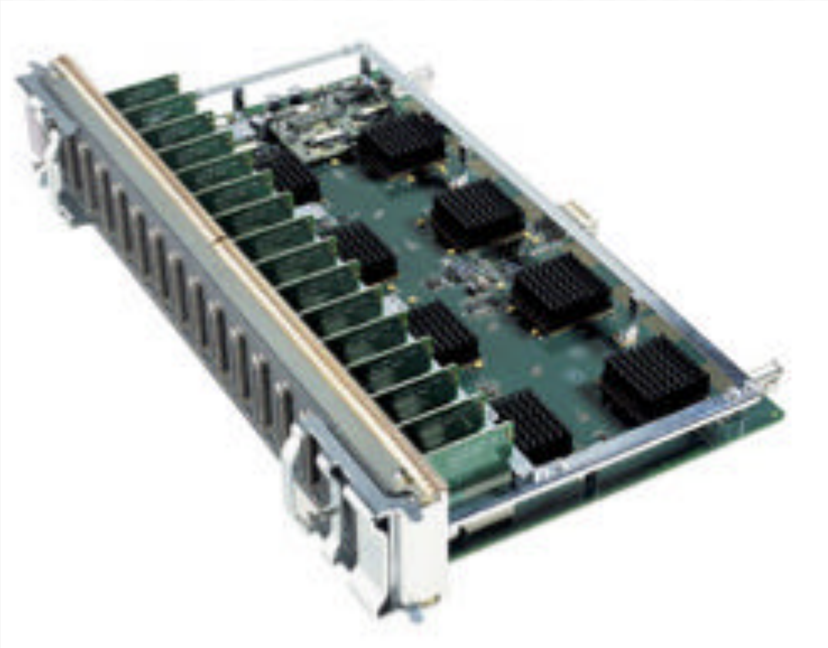
4	**NCSA** United States/2003	*Tungsten* **PowerEdge 1750, P4 Xeon 3.06 GHz, Myrinet** / 2500 Dell	9819 15300

System sustains 64% of peak performance

- But smaller systems hit 70-75% of peak

Quadrics

- QsNetII E-series
  - Released at the end of May 2004
- Deliver 128-port standalone switches
- Performance:
  - Latency: 3 us
  - Bandwidth: 900 MB/s
  - Cost/port (based on 64-port configuration): $1800
    - Switch + NIC + cable
    - http://doc.quadrics.com/Quadrics/QuadricsHome.nsf/DisplayPages/A3EE4AED738B6E2480256DD30057B227



Quadrics

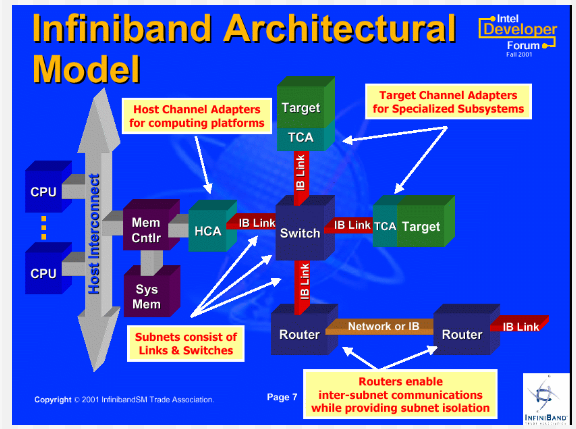
5	**Pacific Northwest National Laboratory** United States/2003	*Mpp2* **Integrity rx2600 Itanium2 1.5 GHz, Quadrics** / 1936 HP	8633 11616

Sustains 74% of peak

- Other systems on Top500 list sustain 70-75% of peak

Infiniband

- Newest interconnect
- Currently shipping 32-port and 96-port switches
  - Requires 32-port switches requires 12 switches (and 256 Cables) to support a full bisection bandwidth network for 128 nodes
- Performance:
  - Latency: 6.8 us (New Adapter from Pathscale takes this to 1.3us without a switch)
  - Bandwidth: 840 MB/s
  - Estimated cost/port (based on 64-port configuration): $1000-$1200
    - Switch + NIC + cable
    - http://www.techonline.com/community/related_content/24364



Infiniband

3	**Virginia Tech** United States/2003	*X* **1100 Dual 2.0 GHz Apple G5/Mellanox Infiniband 4X/Cisco GigE** / 2200 Self-made	10280 17600

- Sustained 58% of peak
  - Other Infiniband machines on Top500 list have achieved 64% and 68%

Ethernet

- Latency: 30-80 us (very dependent on NIC, Switch, and OS Stack)
- Bandwidth: 100 MB/s
- Top500 list has ethernet-based systems sustaining between 35-59% of peak

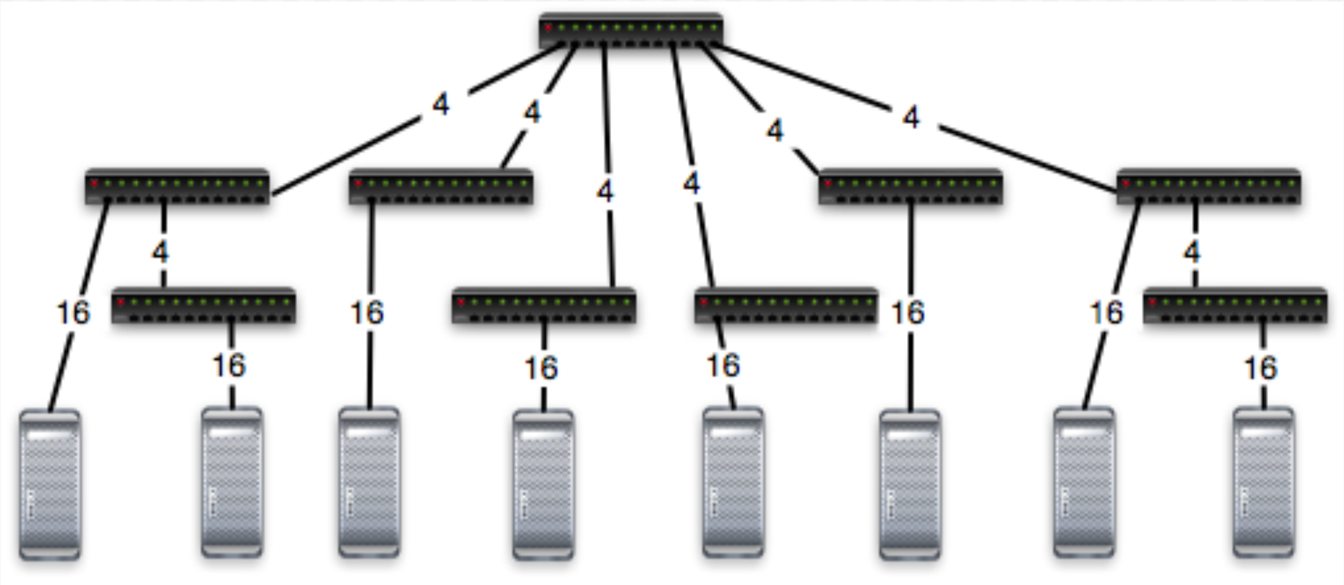
Ethernet

- What we did 3 years ago with 128 nodes and a $13,000 ethernet network
  - $101 / port
  - Sustained 48% of peak

201	UCSD/Cal-IT^2/SDSC United States/2003	**Rocks V60x Cluster 2.8 GHz, Gig Ethernet / 256** Sun	699 1433.6

- With Myrinet, would have sustained 1 Tflop
  - At a cost of ~$130,000
    - Roughly 1/3 the cost of the system

Rockstar Topology (Bisection BW made a difference)

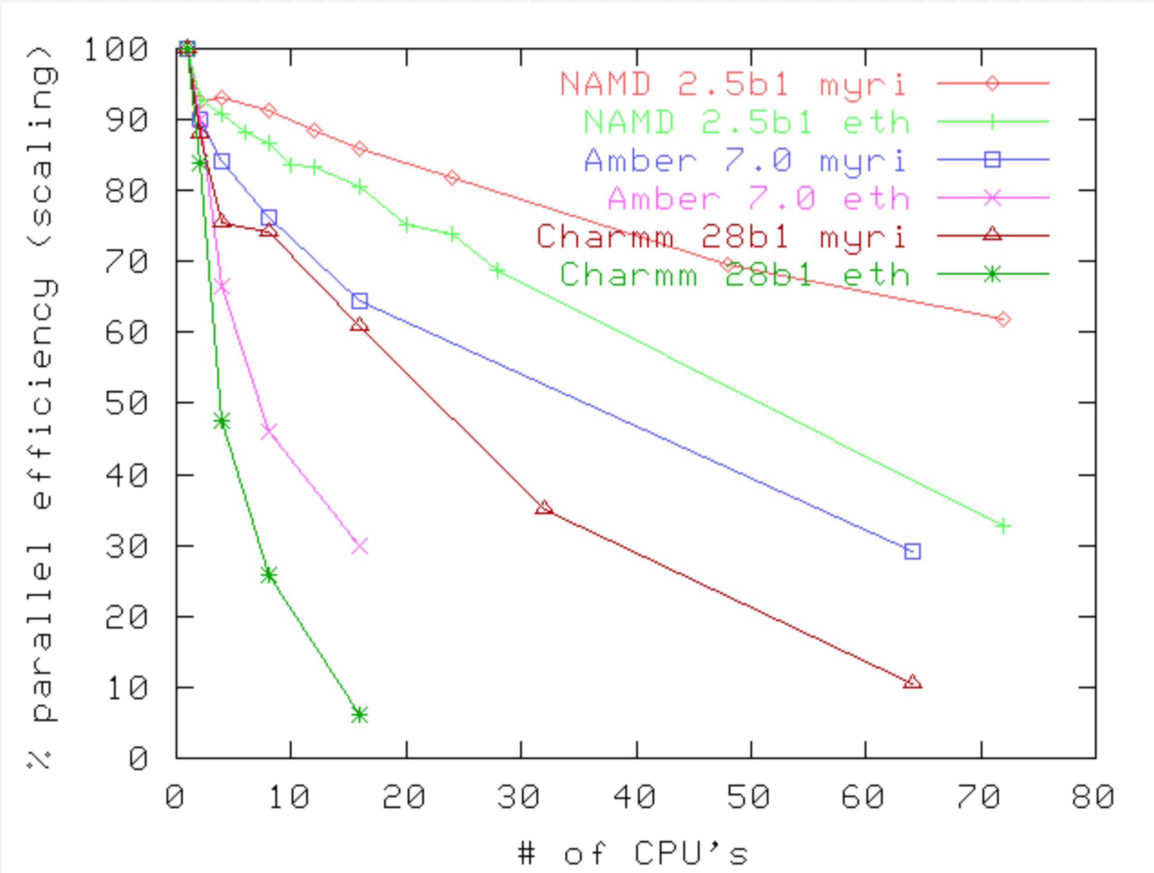


- 24-port switches
- Not a symmetric network
  - Best case - 4:1 bisection bandwidth
  - Worst case - 8:1
  - Average - 5.3:1

Low Latency Ethernet ?

- Bring os-bypass to Ethernet
- Projected performance:
  - Latency: less than 20 us
  - Bandwidth: 100 MB/s
- Potentially could merge management and high-performance networks
- Pioneering Vendor "Ammasso" is out of business
- At 10GigE Force 10 just introduced a 200ns switch (down from ~10us)

Sample Application Benifits



Interconnect Observations

- If your application can tolerate latency, then Ethernet will deliver the best bang for the buck.
- Myrinet, Quadrics and Infiniband all have excellent low latency properties
- Myrinet delivers 2x bandwidth over Ethernet
- Quadrics and Infiniband deliver 2x bandwidth over Myrinet

- Observation: codes are often sensitive first to messaging overhead, then latency, then bandwidth

Details

	Size	Unit Cost	Total Cost		
Compute Nodes	128	2000	$ 256,000		
Frontend Nodes	1	3000	$ 3,000		
Total Node Count	129				
Racks	5	800	$ 4,000		
Ethernet Switches	5	1400	$ 7,000		
Power Cords	135	0	$ -		
Network Cables	130	5	$ 650		
Power Strips	17	100	$ 1,700		
Crash Cart	1	300	$ 300		
Total Hardware Cost			$ 272,650		

- System Cost at **fixed size** is **relatively constant**
  - It is performance that changes
- Memory footprint can change pricing dramatically
- If your application needs low-latency buy a good interconnect

Add KVM

	Size	Unit Cost	Total Cost		
Compute Nodes	128	2000	$ 256,000		
Frontend Nodes	1	3000	$ 3,000		
Total Node Count	129				
Racks	5	800	$ 4,000		
Ethernet Switches	5	1400	$ 7,000		
Power Cords	135	0	$ -		
Network Cables	130	5	$ 650		
Power Strips	17	100	$ 1,700		
Crash Cart	1	300	$ 300		
KVM Cables	129	50	$ 6,450		
KVM Switch	9	1000	$ 9,000		

- $15K USD additional cost (~ 5%)
- KVM's are low volume networks that will require management. Are they worth it?

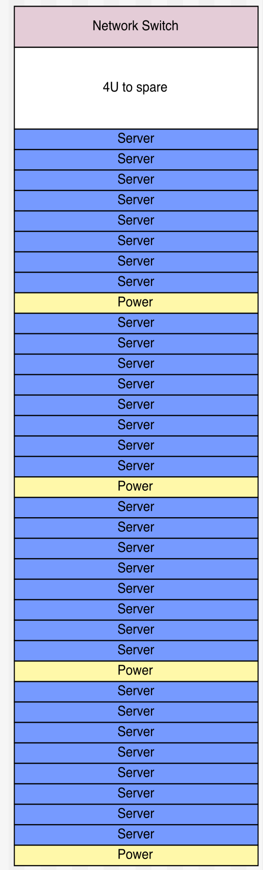
Add Myrinet

	Size	Unit Cost	Total Cost		
Compute Nodes	128	2000	$ 256,000		
Frontend Nodes	1	3000	$ 3,000		
Total Node Count	129				
Racks	5	800	$ 4,000		
Ethernet Switches	5	1400	$ 7,000		
Power Cords	135	0	$ -		
Network Cables	130	5	$ 650		
Power Strips	17	30	$ 510		
Crash Cart	1	300	$ 300		
Myrinet NIC	128	500	$ 64,000		
Myrinet Cables	128	100	$ 12,800		
Myrinet Switch	1	30000	$ 30,000		
Total Hardware Cost			$ 378,260		

- Added $100K USD. ~ 33% of complete system
- Often essential to get codes to scale

1U Servers (Rack of 32 + Frontend)

- 64 Sockets (64-128 Cores)
- 5 electrical circuits (20A, 208V)
- Cable count
  - 65 = power & network
  - 97 with Myrinet
  - 193 with KVM
  - 225 with Serial Port management



Cluster Software Space

Rocks is not alone

Other efforts

Where Rocks fits

The Dark Side of Clusters

- Clusters are phenomenal price/performance computational engines …
  - Can be hard to manage without experience
  - High-performance I/O is still unsolved
  - Finding out where something has failed increases at least linearly as cluster size increases
- Not cost-effective if every cluster "burns" a person just for care and feeding
- Programming environment could be vastly improved
- Technology is changing very rapidly. Scaling up is becoming commonplace (128-256 nodes)

The Top 2 Most Critical Problems

- The largest problem in clusters is *software skew*
  - When software configuration on some nodes is different than on others
  - Small differences (minor version numbers on libraries) can cripple a parallel program
- The second most important problem is adequate job control of the parallel process
  - Signal propagation
  - Cleanup

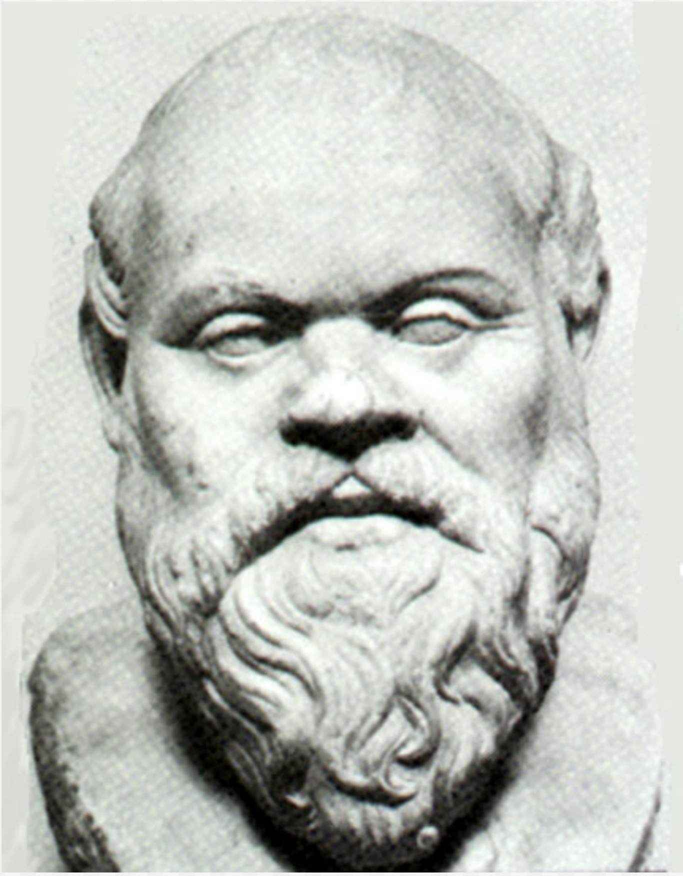
Rocks (open source clustering distribution)

www.rocksclusters.org

- Technology transfer of commodity clustering to application scientists
  - "make clusters easy"
  - Scientists can build their own supercomputers and migrate up to national centers as needed
- Rocks is a cluster on a CD
  - Red Enterprise Hat Linux (opensource and free)
  - Clustering software (PBS, SGE, Ganglia, NMI)
  - Highly programmatic software configuration management
- Core software technology for several campus projects
  - BIRN
  - Center for Theoretical Biological Physics
  - EOL
  - GEON
  - NBCR
  - OptIPuter
- First Software release Nov, 2000
- Supports x86, Opteron/EM64T, and Itanium
- RedHat/CentOS 4.x

Philosophy

- Caring and feeding for a system is not fun
- System Administrators cost more than clusters
  - 1 TFLOP cluster is less than \$200,000 (US)
  - Close to actual cost of a fulltime administrator
- The system administrator is the weakest link in the cluster
  - Bad ones like to tinker
  - Good ones still make mistakes



Philosophy
continued

- All nodes are 100% automatically configured
    - Zero “hand” configuration
    - This includes site-specific configuration
- Run on heterogeneous standard high volume components
    - Use components that offer the best price/performance
    - Software installation and configuration must support different hardware
    - Homogeneous clusters do not exist
    - Disk imaging requires homogeneous cluster

Philosophy
continued

- Optimize for installation
  - Get the system up quickly
  - In a consistent state
  - Build supercomputers in hours not months
- Manage through re-installation
  - Can re-install 128 nodes in under 20 minutes
  - No support for on-the-fly system patching
- Do not spend time trying to issue system consistency
  - Just re-install
  - Can be batch driven
- Uptime in HPC is a myth
  - Supercomputing sites have monthly downtime
  - HPC is not HA

OpenMosix

- Overview
  - Single system image - all nodes look like one large multiprocessor
  - Jobs migrate from machine to machine (based on machine load)
  - No changes required for apps to use system
- Interconnects supported
  - All IP-based networks
- Custom Linux Kernel
  - Download a new kernel
  - Or patch and compile
  - Install kernel on all nodes
- Supports
  - Diskfull
  - Diskless

Warewulf

- Overview
  - Install frontend first
    - Recommend using RPM-based distribution
  - Imaged based installation
    - “Virtual node filesystem”
  - Attacks problem of generic slave node management
- Standard cluster software not included
  - Added separately
  - Use ‘chroot’ commands to add in extra software
- Supports
  - Diskfull
  - Diskless

Scyld Beowulf

- Single System Image
  - Global process ID
  - Not a global file system
- Heavy OS modifications to support BProc
  - Patches kernel
  - Patches libraries (libc)
- Job start on the frontend and are pushed to compute nodes
  - Hooks remain on the frontend
  - Does this scale to 1000 nodes?
- Easy to install
  - Full distribution
  - Often compared to Rocks

SCore

- Research group started in 1992, and based in Tokyo.
- Score software
  - Semi-automated node integration using RedHat
  - Job launcher similar to UCB's REXEC
  - MPC++, multi-threaded C++ using templates
  - PM, wire protocol for Myrinet
- Development has started on SCore Roll

Scalable Cluster Environment
(SCE)

- Developed at Kasetsart University in Thailand
- SCE is a software suite that includes
  - Tools to install, manage, and monitor compute nodes
    - Diskless (SSI)
    - Diskfull (RedHat)
  - A batch scheduler to address the difficulties in deploying and maintaining clusters
  - Monitoring tools (SCMSWeb)
- User installs frontend with RedHat and adds SCE packages.
- Rocks and SCE are working together
  - Rocks is good at low level cluster software
  - SCE is good at high level cluster software
  - SCE Roll is now available for Rocks

Open Cluster Group
(OSCAR)

- OSCAR is a collection of clustering best practices (software packages)
  - PBS/Maui
  - OpenSSH
  - LAM/MPI
- Image based installation
  - Install frontend machine manually
  - Add OSCAR packages to frontend
  - Construct a “golden image” for compute nodes
  - Install with system imager
  - “Multi-OS” – Currently only supports RPM-based Distros
    - Dropping “Mandriva” ..
- Started as a consortium of industry and government labs
  - NCSA, ORNL, Intel, IBM, Dell, others
  - Dell now does Rocks. NCSA no longer a contributor. IBM?

System Imager

- Originally VA/Linux (used to sell clusters) (now “bald guy software)
- System imaging installation tools
  - Manages the files on a compute node
  - Better than managing the disk blocks
- Use
  - Install a system manually
  - Appoint the node as the golden master
  - Clone the “golden master” onto other nodes
- Problems
  - Doesn’t support heterogeneous
  - Not method for managing the software on the “golden master”
  - Need “Magic Hands” of cluster-expert admin for every new hardware build

Cfengine

- Policy-based configuration management tool for UNIX or NT hosts
  - Flat ASCII (looks like a Makefile)
  - Supports macros and conditionals
- Popular to manage desktops
  - Patching services
  - Verifying the files on the OS
  - Auditing user changes to the OS
- Nodes pull their Cfengine file and run every night
  - System changes on the fly
  - One bad change kills everyone (in the middle of the night)
- Can help you make changes to a running cluster

Kickstart

- RedHat
  - Automates installation
  - Used to install desktops
  - Foundation of Rocks
- Description based installation
  - Flat ASCII file
  - No conditionals or macros
  - Set of packages and shell scripts that run to install a node

LCFG

- Edinburgh University
  - Anderson and Scobie
- Description based installation
  - Flat ASCII file
  - Conditionals, macros, and statements
    - Full blown (proprietary) language to describe a node
- Compose description file out of components
  - Using file inclusion
  - Not a graph as in Rocks
- Do not use kickstart
  - Must replicate the work of RedHat
- Very interesting group
  - Design goals very close to Rocks
  - Implementation is also similar

Rocks Basic Approach

- Install a frontend
  1. Insert Rocks Base CD
  2. Insert Roll CDs (optional components)
  3. Answer 7 screens of configuration data
  4. Drink coffee (takes about 30 minutes to install)
- Install compute nodes:
  1. Login to frontend
  2. Execute insert-ethers
  3. Boot compute node with Rocks Base CD (or PXE)
  4. Insert-ethers discovers nodes
  5. Goto step 3
- Add user accounts
- Start computing

Optional Rolls
- Condor
- Grid (based on NMI R4)
- Intel (compilers)
- Java
- SCE (developed in Thailand)
- Sun Grid Engine
- PBS (developed in Norway)
- Area51 (security monitoring tools)
- Many Others …

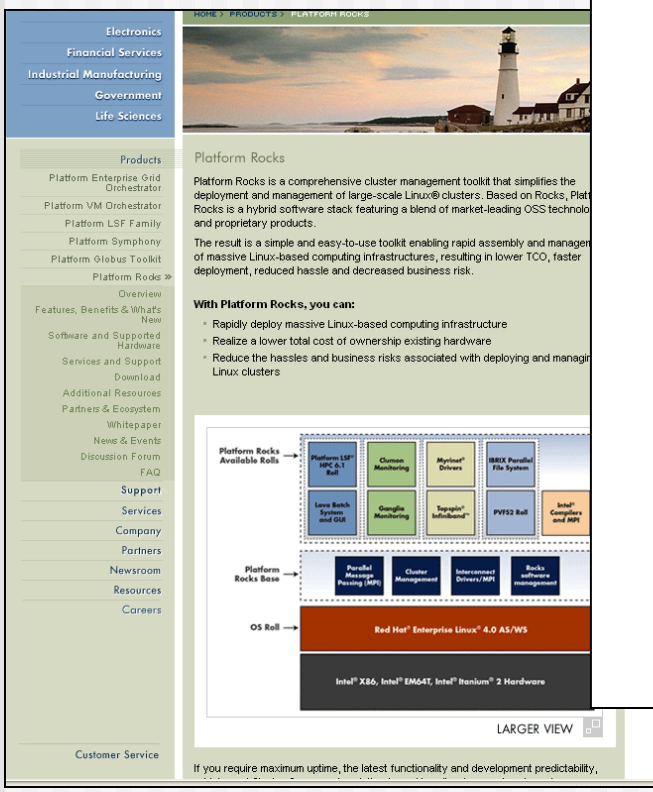
Minimum Requirements

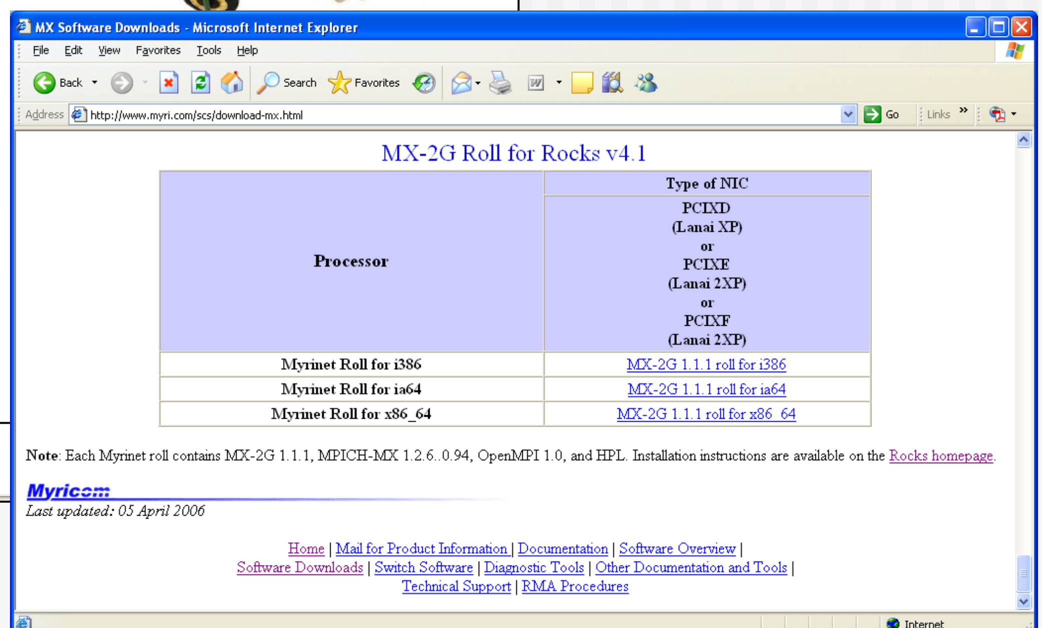
- Frontend
  - 2 Ethernet Ports
  - CDROM
  - 18 GB Disk Drive
  - 512 MB RAM
- Compute Nodes
  - 1 Ethernet Port
  - 18 GB Disk Drive
  - 512 MB RAM
- Complete OS Installation on all Nodes
- No support for Diskless (yet)
- Not a Single System Image
- All Hardware must be supported by RHEL

HPCwire Reader's Choice Awards for 2004/2005

- Rocks won in Several categories:
  - Most Important Software Innovation (Reader's Choice)
  - Most Important Software Innovation (Editor's Choice)
  - Most Innovative - Software (Reader's Choice)

Commercial Interest





MX-2G Roll for Rocks v4.1

Processor	Type of NIC: PCIXD (Lanai XP) or PCIXE (Lanai 2XP) or PCIXF (Lanai 2XP)
Myrinet Roll for i386	MX-2G 1.1.1 roll for i386
Myrinet Roll for ia64	MX-2G 1.1.1 roll for ia64
Myrinet Roll for x86_64	MX-2G 1.1.1 roll for x86_64

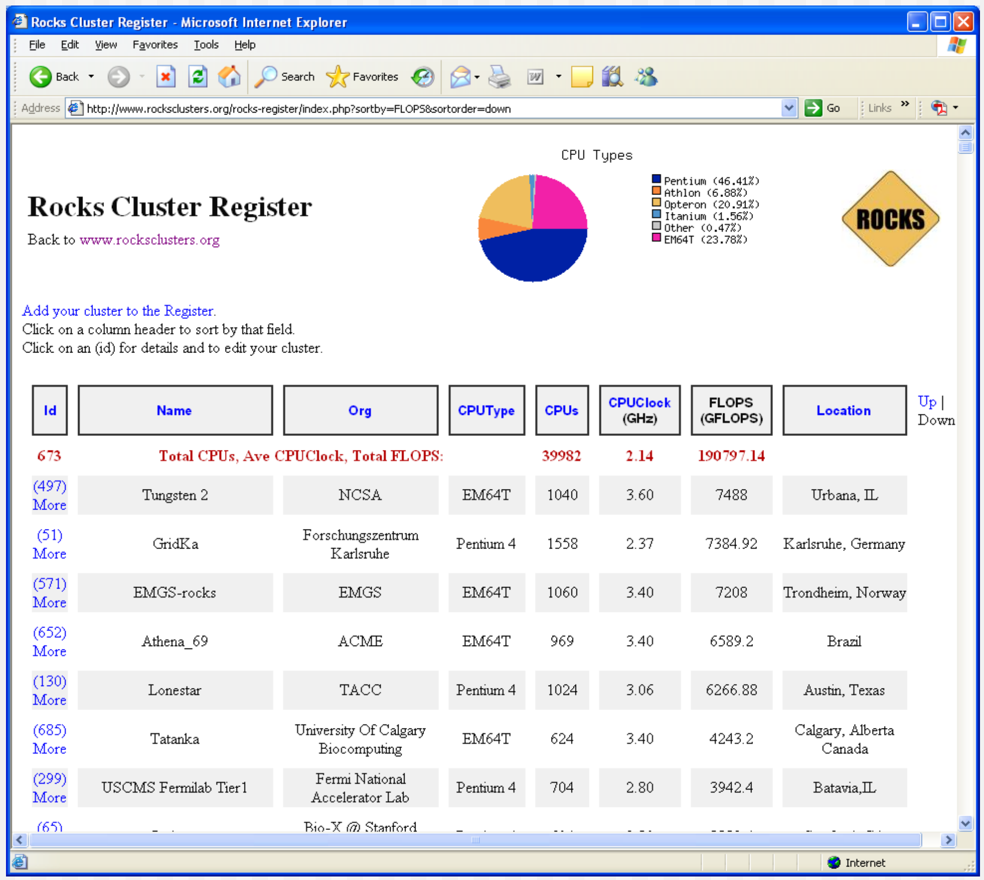
Note: Each Myrinet roll contains MX-2G 1.1.1, MPICH-MX 1.2.6..0.94, OpenMPI 1.0, and HPL. Installation instructions are available on the Rocks homepage.

Last updated: 05 April 2006

Home | Mail for Product Information | Documentation | Software Overview | Software Downloads | Switch Software | Diagnostic Tools | Other Documentation and Tools | Technical Support | RMA Procedures

Registration Page

(optional)



Rocks Cluster Register - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.rocksclusters.org/rocks-register/index.php?sortby=FLOPS&sortorder=down

Rocks Cluster Register

Back to www.rocksclusters.org

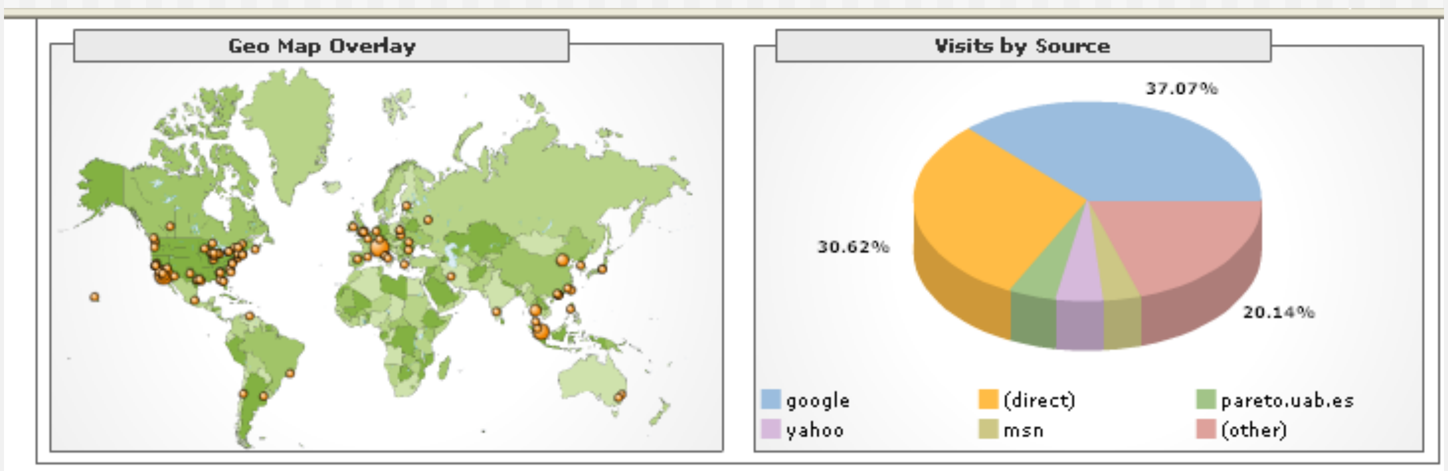
Add your cluster to the Register.
Click on a column header to sort by that field.
Click on an (id) for details and to edit your cluster.

Id	Name	Org	CPUType	CPUs	CPUClock (GHz)	FLOPS (GFLOPS)	Location	Up | Down
673	Total CPUs, Ave CPUClock, Total FLOPS:			39982	2.14	190797.14		
(497) More	Tungsten 2	NCSA	EM64T	1040	3.60	7488	Urbana, IL	
(51) More	GridKa	Forschungszentrum Karlsruhe	Pentium 4	1558	2.37	7384.92	Karlsruhe, Germany	
(571) More	EMGS-rocks	EMGS	EM64T	1060	3.40	7208	Trondheim, Norway	
(652) More	Athena_69	ACME	EM64T	969	3.40	6589.2	Brazil	
(130) More	Lonestar	TACC	Pentium 4	1024	3.06	6266.88	Austin, Texas	
(685) More	Tatanka	University Of Calgary Biocomputing	EM64T	624	3.40	4243.2	Calgary, Alberta Canada	
(299) More	USCMS Fermilab Tier1	Fermi National Accelerator Lab	Pentium 4	704	2.80	3942.4	Batavia,IL	
(65)		Bio-X @ Stanford						

Internet

User Base

- > 1300 Users on the Discussion List
- 5 Continents
- **University**, Commercial, Hobbyist

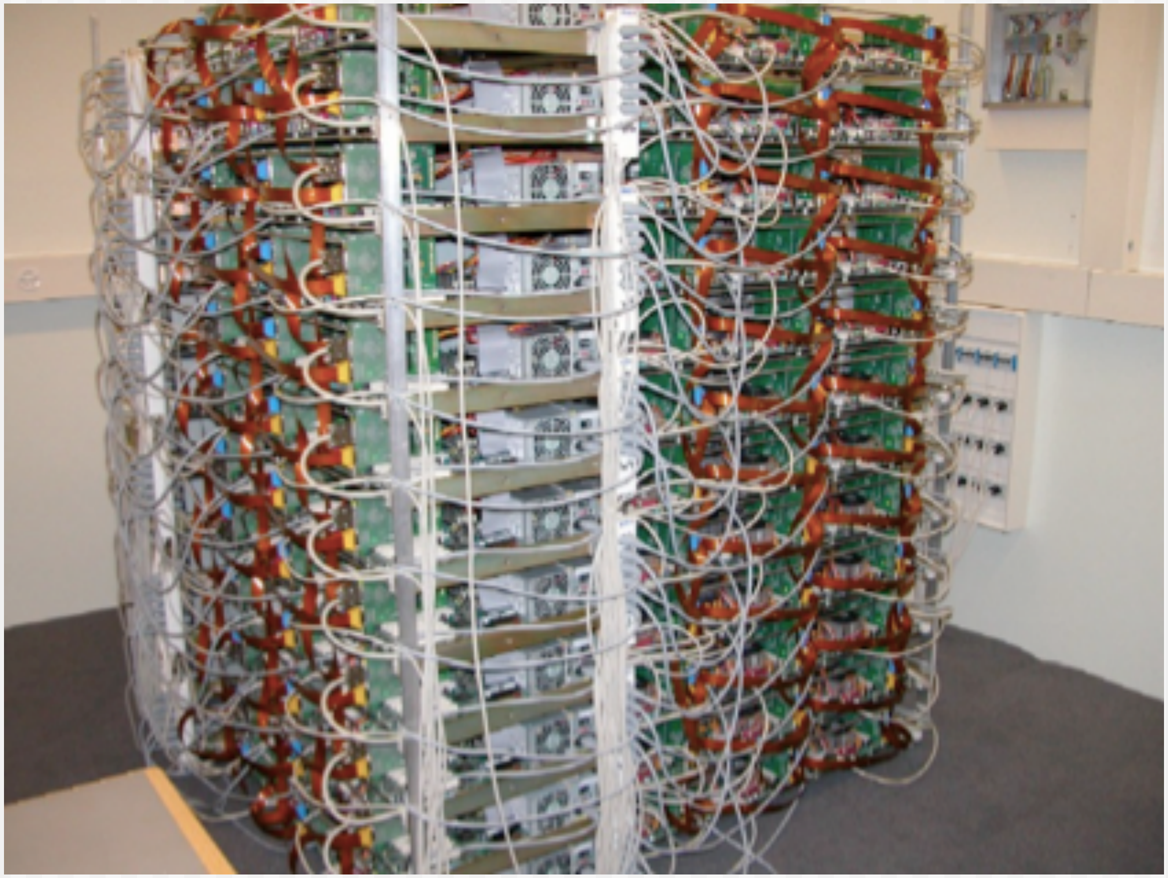


Beowulf Mentality

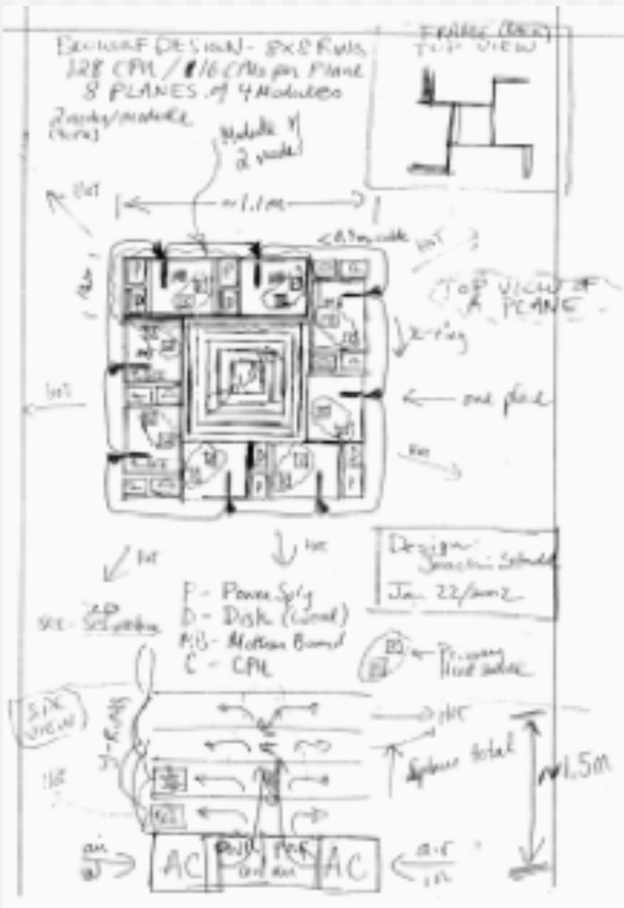
Why DIY is wrong

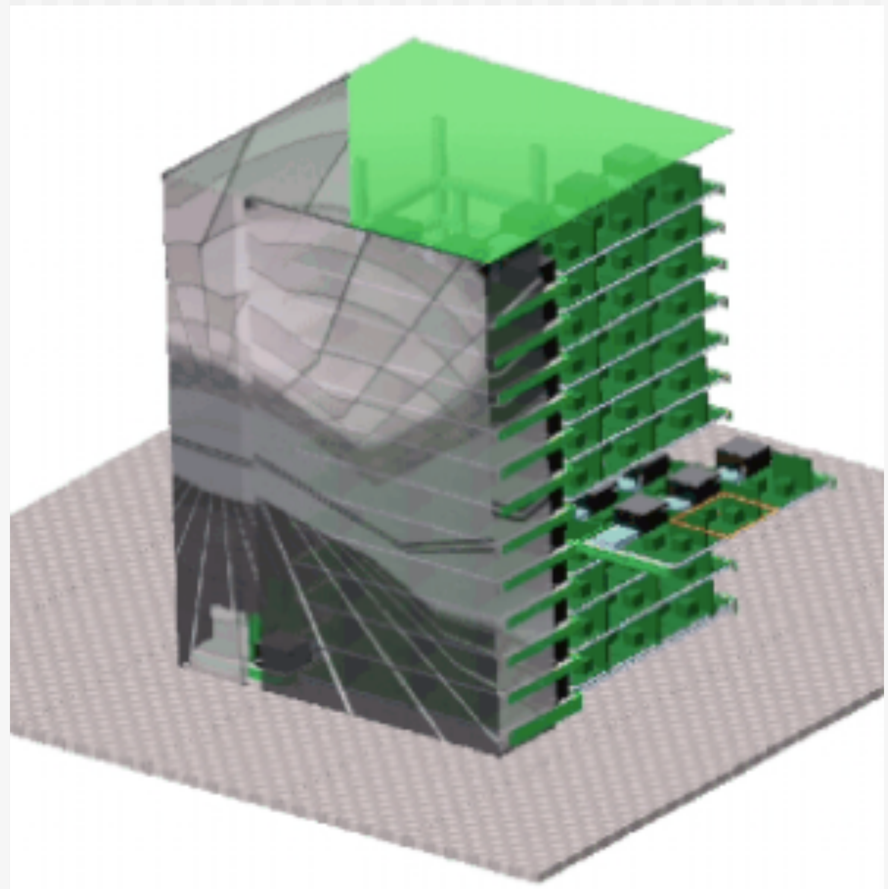
A Tale of a Cluster Tuner

(288 AthlonMP Hand Built Machine)

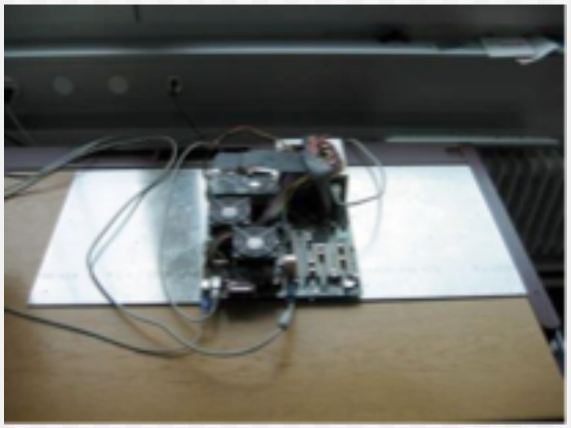


07.2002: The Idea





08.2002 - 11.2002: Construction

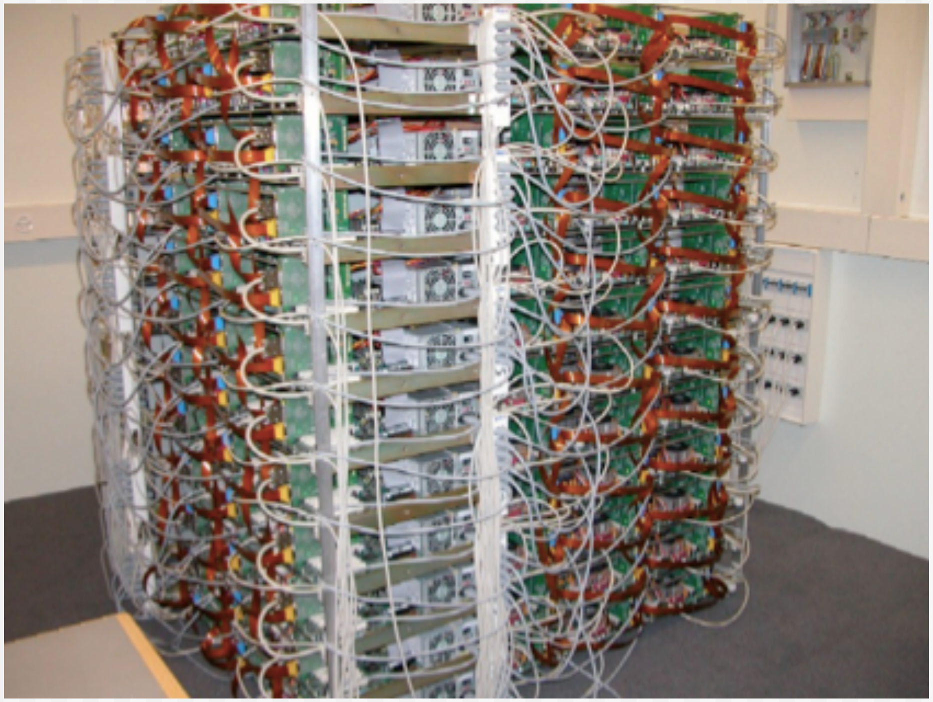


12.2002: Build Complete & Celebration

- Machine only 50% operational
- But, they are getting results
- Machine is fully operational 3 months later

Summary

- 07.2002
  - Design system
- 08.2002 - 11.2002
  - Build system
- 03.2003
  - System in Production
- **7 months** (maybe 8)
  - **Concept to Cluster**
  - Still just a Beowulf
  - Moore-cycle is 18 months
    - Half life for performance
    - Half life for cost
  - Useful life is 36-48 months
- What did they optimize for?



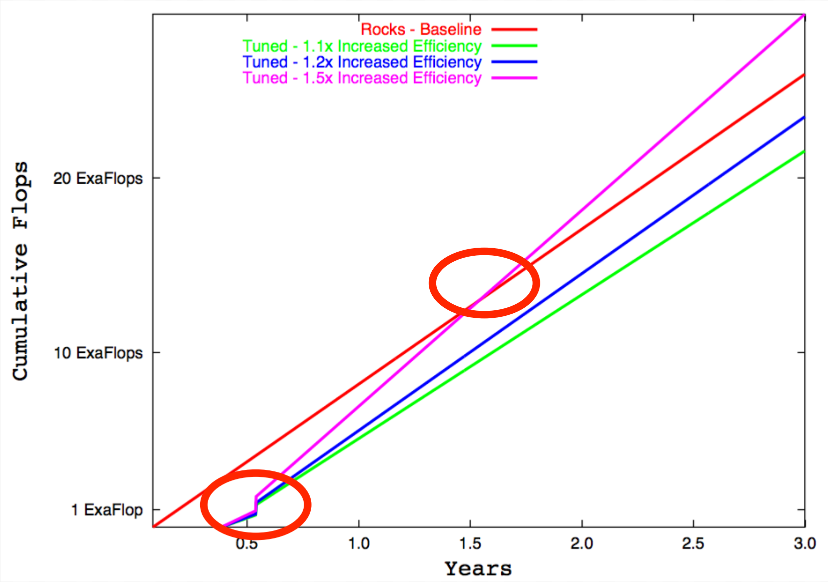
Rocks Cluster Timeline

- Day 1 - Idea
- Day 30 - Production

- Not just us, world wide user base has done the same

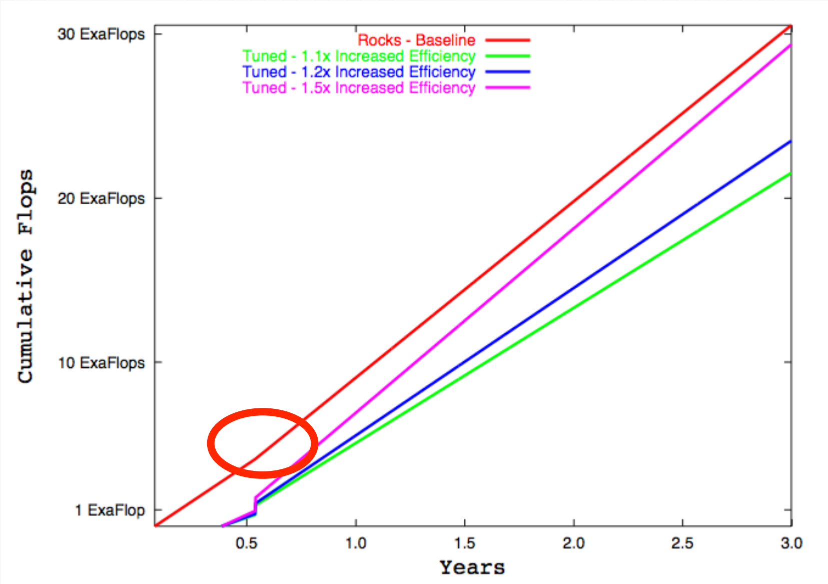
Lost Time = Lost Computation

- Assumption
  - Rocks
    - 256 2.2 GHz Pentium IV
    - 1,126 GFlops
    - Available at same time as tuner build
    - 1 month to build
  - Tuner
    - 144 - 264 Athlon-MP 2200+
    - 512 - 950 Gflops
    - 5 - 7 months to build
- Baseline of 50% CPU efficiency for Rocks
- Tuner improvement beyond baseline
  - 10% (55% efficiency)
  - 20% (60% efficiency)
  - 50% (75% efficiency)
- Tuner must have 50% gain to catch baseline after 1.5 years



Invest in Hardware not People

- Assumptions
  - Two salaried tuners
  - “Full burden” (salary, grant overhead, office space, etc) is $180k / year.
- Invest
  - 5 months salary into baseline
  - $150k (5 months)
  - Just buy more nodes
    - $2500k / node
- Month 7
  - Baseline cluster grows
  - 54 2.2 GHz servers
  - Ignoring Moore’s Law!
- Baseline wins



Other Tuners

- Kernel Tuning
  - “My handcrafted kernel is X times faster.”
- Distribution Tuning
  - “Distribution Y is X times faster.”
  - RFP: “Vendor will be penalized for a Red Hat only solution”
- White-box Tuning
  - “White-box vendor Y has a node that is X times cheaper.”

Rocks

Making Clusters Easy

When You Need Power Today

Young Frankenstein - Gene Wilder, Peter Boyle

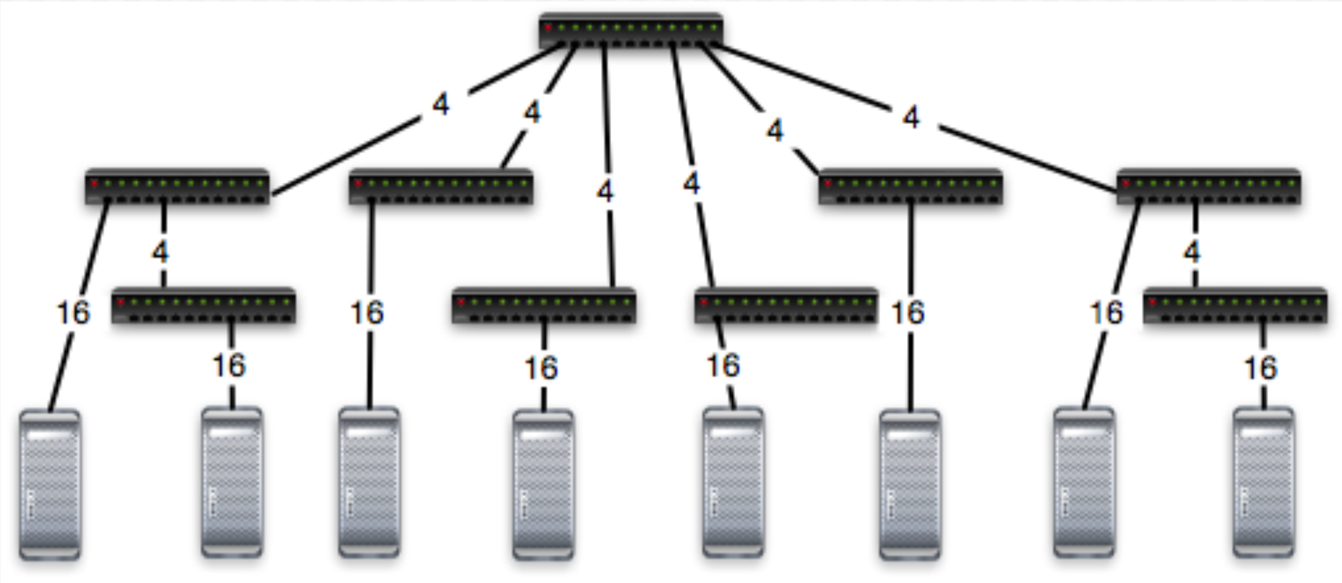
Two Examples

- Rockstar - SDSC
- Tungsten2 - NCSA

Rockstar Cluster

- 129 Sun Fire V60x servers
  - 1 Frontend Node
  - 128 Compute Nodes
- Gigabit Ethernet
  - $13,000 (US)
  - 9 24-port switches
  - 8 4-gigabit trunk uplinks
- Built live at SC'03
  - In under two hours
  - Running applications
- Top500 Ranking
  - 11.2003: 201
  - 06.2004: 433
  - 49% of peak

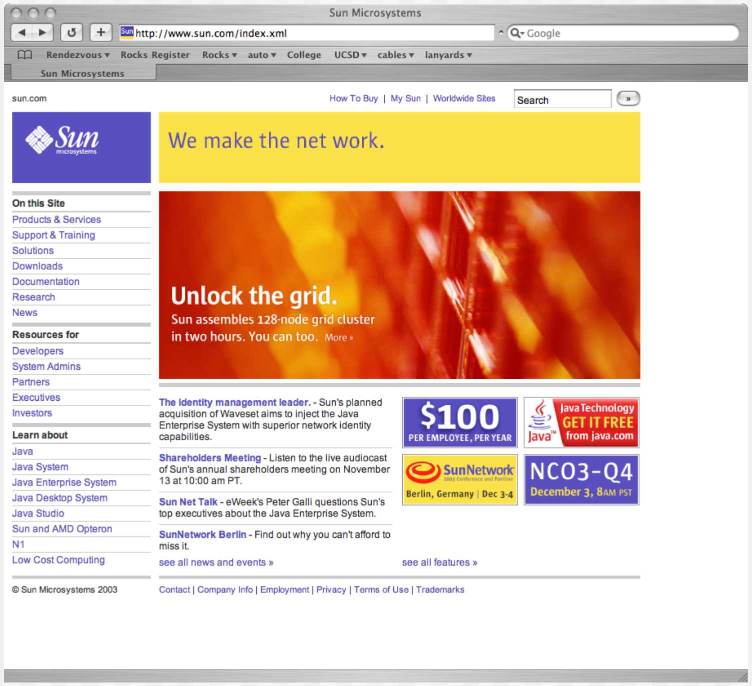
Rockstar Topology



- 24-port switches
- Not a symmetric network
  - Best case - 4:1 bisection bandwidth
  - Worst case - 8:1
  - Average - 5.3:1

Super Computing 2003 Demo

- We wanted to build a Top500 machine live at SC'03
  - From the ground up (hardware and software)
  - In under two hours
- Show that anyone can build a super computer with:
  - Rocks (and other toolkits)
  - Money
  - No army of system administrators required
- HPC Wire Interview
  - **HPCwire**: What was the most impressive thing you've seen at SC2003?
  - **Larry Smarr**: I think, without question, the most impressive thing I've seen was Phil Papadopoulos' demo with Sun Microsystems.



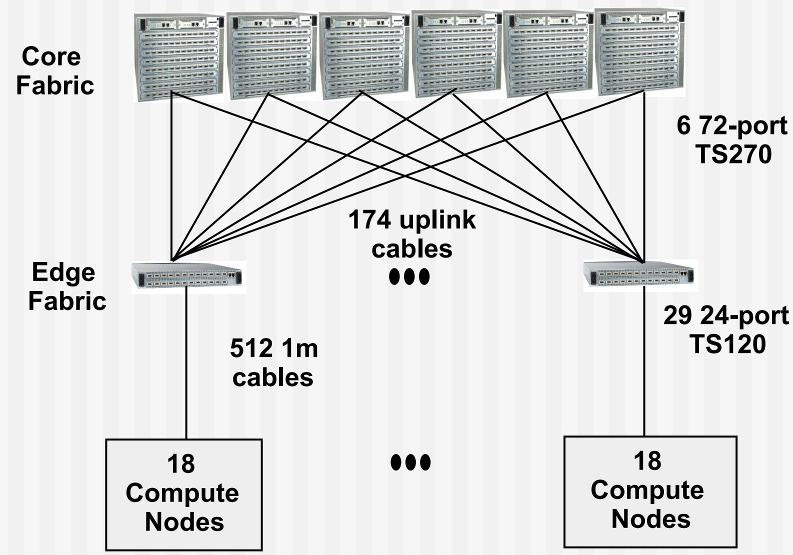
SC2003 IGNITING INNOVATION
PHOENIX | AZ | NOVEMBER 15 - 21

Building Rockstar

NCSA

National Center for Supercomputing Applications

- Tungsten2
  - 520 Node Cluster
  - Dell Hardware
  - Topspin Infiniband
- Deployed 11.2004
- Easily in top 100 of the 06.2005 top500 list
- “**We went from PO to crunching code in 2 weeks**. It only took another 1 week to shake out some math library conflicts, and we have been in production ever since.” -- *Greg Keller, NCSA (Dell On-site Support Engineer)*



Id	Name	Org	CPUType	CPUs	CPUClock (GHz)	FLOPS (GFLOPS)	Location
435	Total CPUs, Ave CPUClock, Total FLOPS:			26571	2.02	117134.22	
(497) More	Tungsten 2	NCSA	EM64T	1040	3.60	7488	Urbana, IL

Largest registered Rocks cluster

source: topspin (via google)