



NPACI Rocks Cluster Distribution

Mason Katz

Group Lead, Cluster Development

San Diego Supercomputer Center

Gelato Meeting UI 05-24-04

Overview of San Diego Supercomputer Center

- Founded in 1985
 - **Non-military access to supercomputers**
- Over 400 employees
- Mission: Innovate, develop, and deploy technology to advance science
- Recognized as an international leader in:
 - **Data Management**
 - **High Performance Computing**
 - **Grid and Cluster Computing**
 - **Networking**
 - **Visualization**
- Primarily funded by NSF



SDSC Top500 Machines

■ Blue Horizon

- **IBM SP Power3**
- **AIX**
- **Federation Switch (54% peak)**

137	<u>UCSD/San Diego Supercomputer Center</u> United States/2000	SP Power3 375 MHz 8 way / 1152 IBM	929 1728
-----	--	--	-------------

■ TeraGrid Cluster

- **IBM Intel Itanium2**
- **Suse Linux / XCat**
- **Myrinet (79% peak)**

171	<u>UCSD/San Diego Supercomputer Center</u> United States/2003	TeraGrid Cluster Itanium 2 1 GHz - Myrinet / 252 IBM	798.3 1008
-----	--	--	---------------

■ Rockstar Cluster

- **Sun Intel x86**
- **Redhat Linux / Rocl**
- **Ethernet (49% peak)**

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



NPACI 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 Hat Linux**
- **Clustering software (PBS, SGE, Ganglia, NMI)**
- **Highly programmatic software configuration management**



- Core software technology for several campus projects

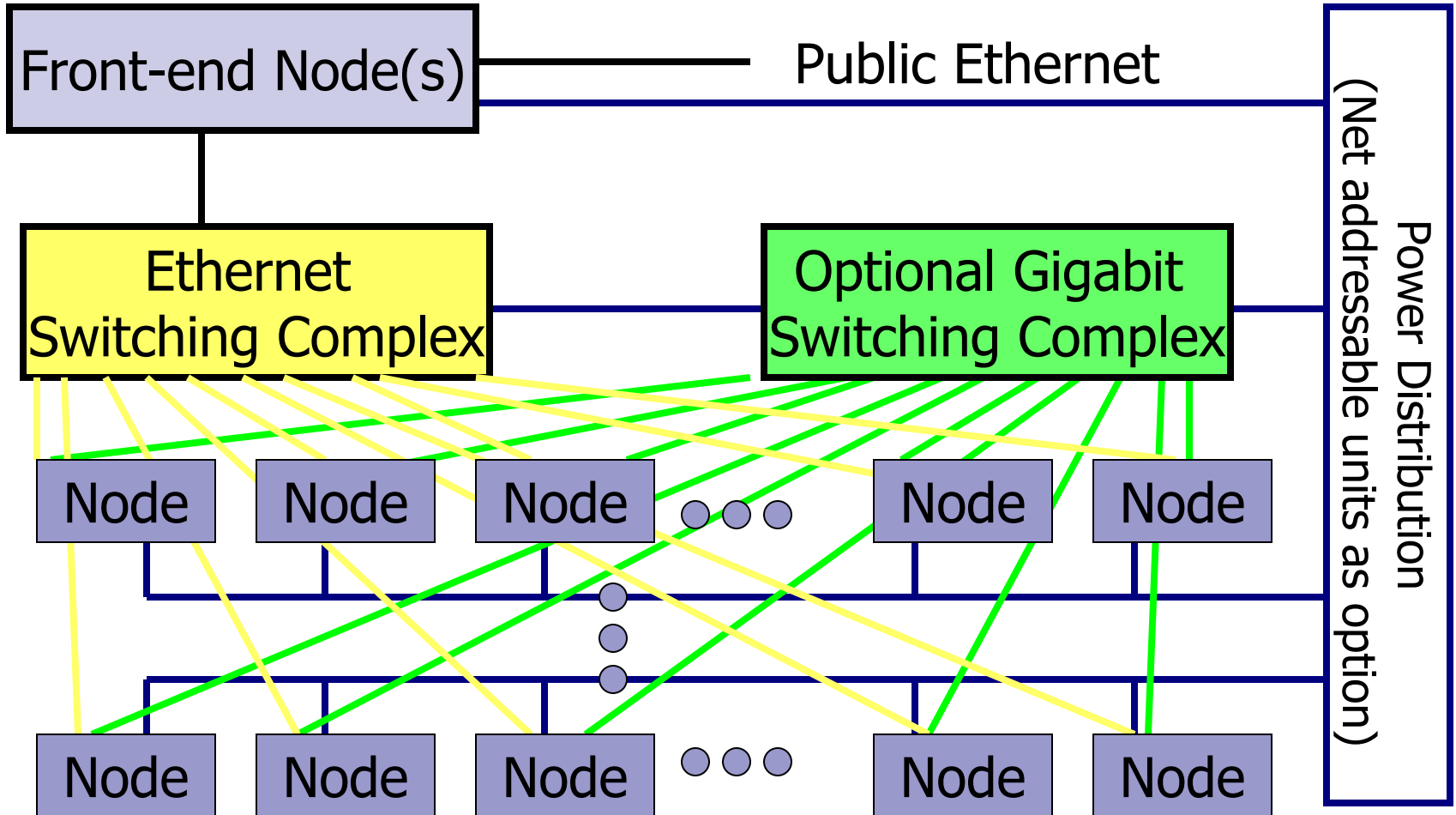
- **BIRN**
- **Center for Theoretical Biological Physics**
- **EOL**
- **GEON**
- **NBCR**
- **OptIPuter**



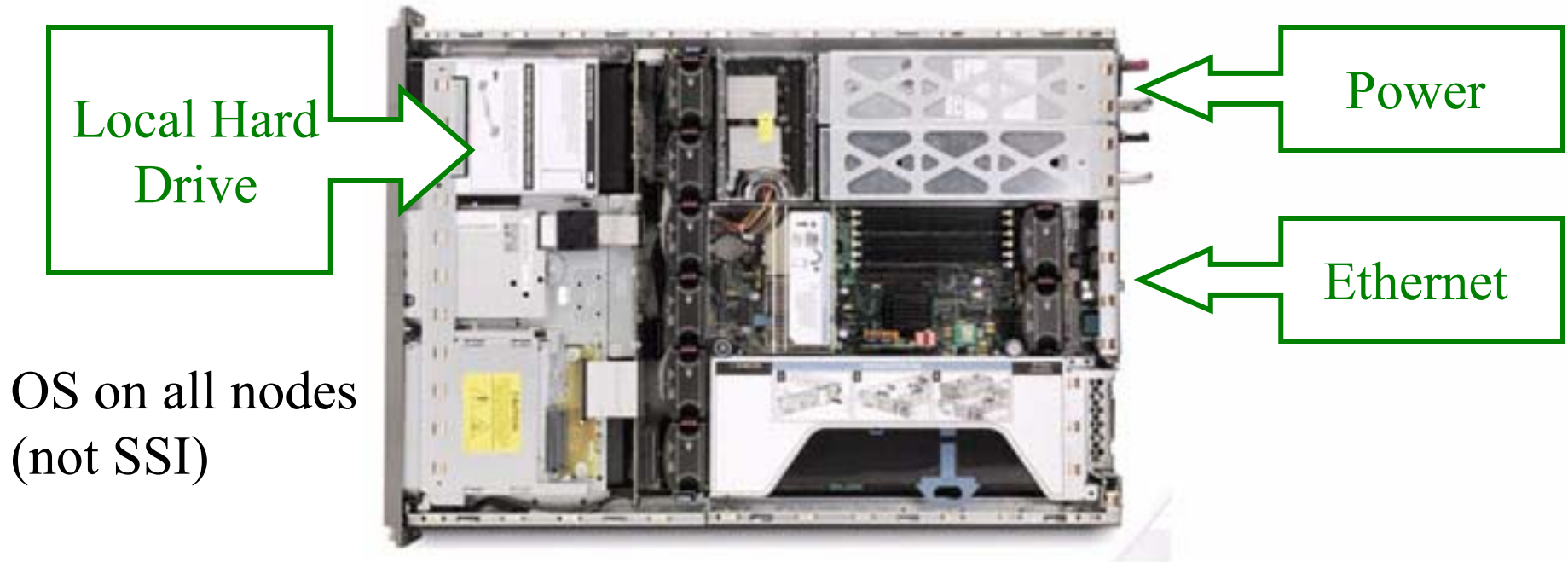
- First Software release Nov, 2000
- Supports x86, Opteron, and Itanium



Basic Architecture



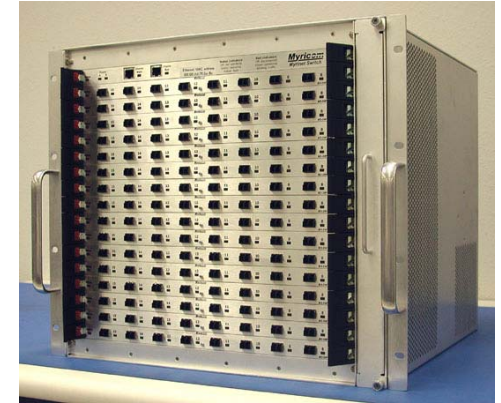
Minimum Components



X86, Opteron, IA64 server

Optional Components

- Myrinet high-performance network
- Network-addressable power distribution unit
- keyboard/video/mouse network not required
 - **Non-commodity**
 - **How do you manage your management network?**
 - **Crash carts have a lower TCO**



Rocks Registry (optional for users)

Top500
June '03

Top500
Nov'03

#319 →

#300 →
#242 →

Rocks Cluster Register								Up Down
Id	Name	Org	CPUType	CPUs	CPUClock (GHz)	FLOPS (GFLOPS)	Location	
Total CPUs, Ave CPUClock, Total FLOPS:				13099	1.80	47791.92		
(51) More	GridKa	Forschungszentrum Karlsruhe	Pentium 4	782	2.41	3769.24	Karlsruhe, Germany	
(130) More	lonestar	TACC	Pentium 4	600	3.06	3672	Austin, Texas	
(65) More	Iceberg	Bio-X @ Stanford University	Pentium 4	604	2.80	3382.4	Stanford, CA	
(231) More	Femilab reconstruction farms	Femilab	Pentium 3	1504	1.50	2256	Batavia, IL	
(151) More	UN TOTS	Intel	Pentium 4	388	2.66	2064.16	Hillsboro, OR	
(210) More	Matterhorn	University of Zurich	Opteron	522	1.80	1879.2	Zurich	
(181) More	lodestone	Scripps Institution of Oceanography	Pentium 4	278	2.80	1556.8	San Diego	
(148) More	RockStar	UCSD, CalIT ² , SDSC	Pentium 4	258	2.80	1444.8	Phoenix / La Jolla	
(52) More	Chihuahua	North Carolina State University College of PAMS	Pentium 4	311	2.25	1399.5	Raleigh, North Carolina	
(204) More	Voyager	US Government	Opteron	320	1.80	1152	Washington, DC	
(225) More	*Private*	*Private*	Opteron	256	2.20	1126.4	Paris, France	
(203) More	USCMS and CDF Cluster	UCSD HEP	Pentium 4	200	2.60	1040	La Jolla	
(124) More	CTBP	Center for Theoretical Biological Physics	Pentium 4	160	2.80	896	San Diego	
(174) More	Nadelhorn	University of Zurich	Pentium 4	152	2.80	851.2	Zurich, Switzerland	
(115) More	HPCCC TDG ROCKS	High Performance Cluster Computing Centre, HK Baptist University	Pentium 4	128	2.80	716.8	Hong Kong	
(82) More	Dell PE2650	Dell Computer Corp	Pentium 4	140	2.40	672	Round Rock, TX	
(14) More	The Extinction Machine	Biological Sciences, University of South Carolina	Athlon MP	194	1.73	671.24	Columbia, South Carolina	

← #26

← #201

← #408

← #176

Over 47 TFlops Aggregate Peak

- Earth Simulator - 41 TFlops
- ASCI Q - 20 Tflops
- Not a real comparison



Rank	Site Country/Year	Computer / Processors Manufacturer	R_{max} R_{peak}
1	Earth Simulator Center Japan/2002	Earth-Simulator / 5120 NEC	35860 40960
2	Los Alamos National Laboratory United States/2002	ASCI Q - AlphaServer SC45, 1.25 GHz / 8192 HP	13880 20480
3	Virginia Tech United States/2003	X 1100 Dual 2.0 GHz Apple G5/Mellanox Infiniband 4X/Cisco GigE / 2200 Self-made	10280 17600

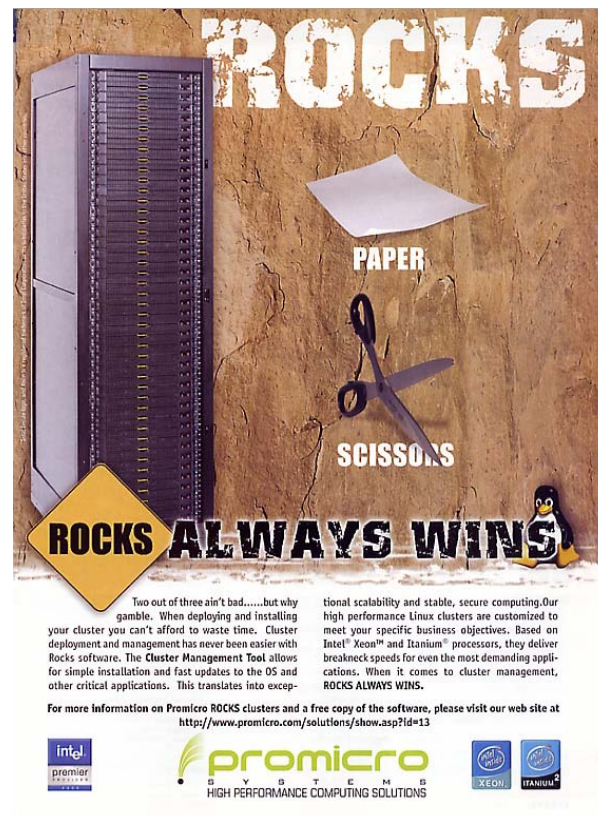
Rocks in the commercial world

■ Rocks Cluster Vendors

- Cray**
- Dell**
- Promicro Systems**
 - See page 79 of April's '03 Linux Journal
 - New add from SCS expected June
- SCS (in Singapore)**
 - Contributed PVFS, SGE to Rocks
 - Active on the Rocks mailing list

■ Training and Support

- Intel is training customers on Rocks**
- Callident is offering support services**



The advertisement features a server rack on the left. The word "ROCKS" is written in large, white, distressed letters at the top. Below it, a piece of paper is shown being cut by a pair of scissors. The words "PAPER" and "SCISSORS" are written below the paper and scissors respectively. At the bottom, a yellow diamond-shaped sign contains the text "ROCKS ALWAYS WINS". A small penguin icon is visible in the bottom right corner of the advertisement.

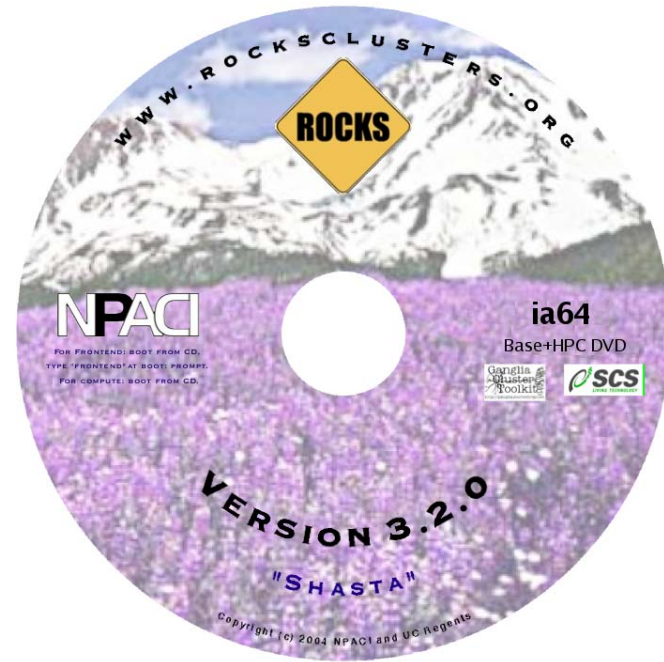
Two out of three ain't bad.....but why gamble. When deploying and installing your cluster you can't afford to waste time. Cluster deployment and management has never been easier with Rocks software. The **Cluster Management Tool** allows for simple installation and fast updates to the OS and other critical applications. This translates into exceptional scalability and stable, secure computing. Our high performance Linux clusters are customized to meet your specific business objectives. Based on Intel® Xeon™ and Itanium™ processors, they deliver breakneck speeds for even the most demanding applications. When it comes to cluster management, **ROCKS ALWAYS WINS.**

For more information on Promicro ROCKS clusters and a free copy of the software, please visit our web site at <https://www.promicro.com/solutions/show.asp?id=13>

Logos for Intel Premier, promicro (BYBTEMS), and Intel Xeon/Itanium are displayed at the bottom.

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 installs)**
- 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



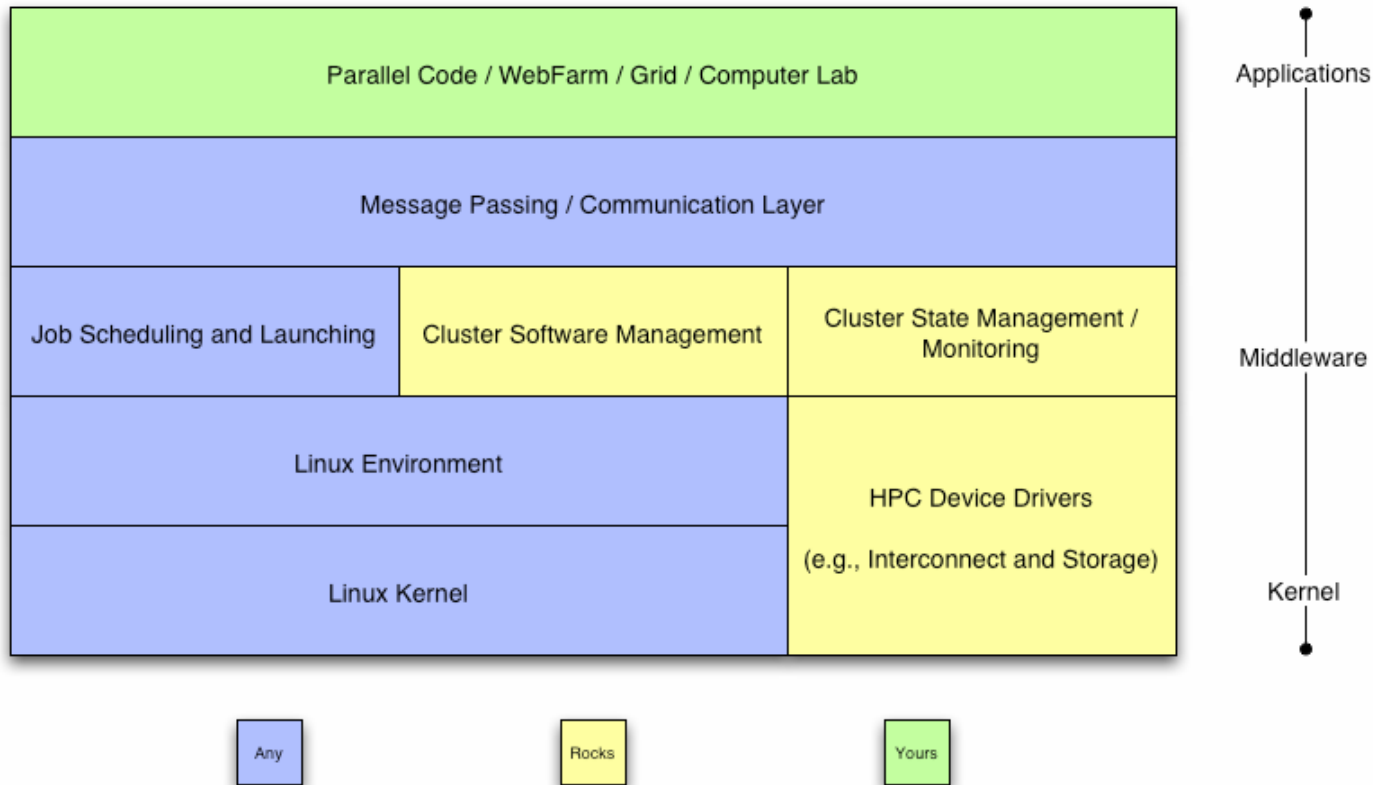
Optional Rolls

- Condor**
- Grid (based on NMI R4)**
- Intel (compilers)**
- Java**
- SCE (Thailand)**
- Sun Grid Engine**

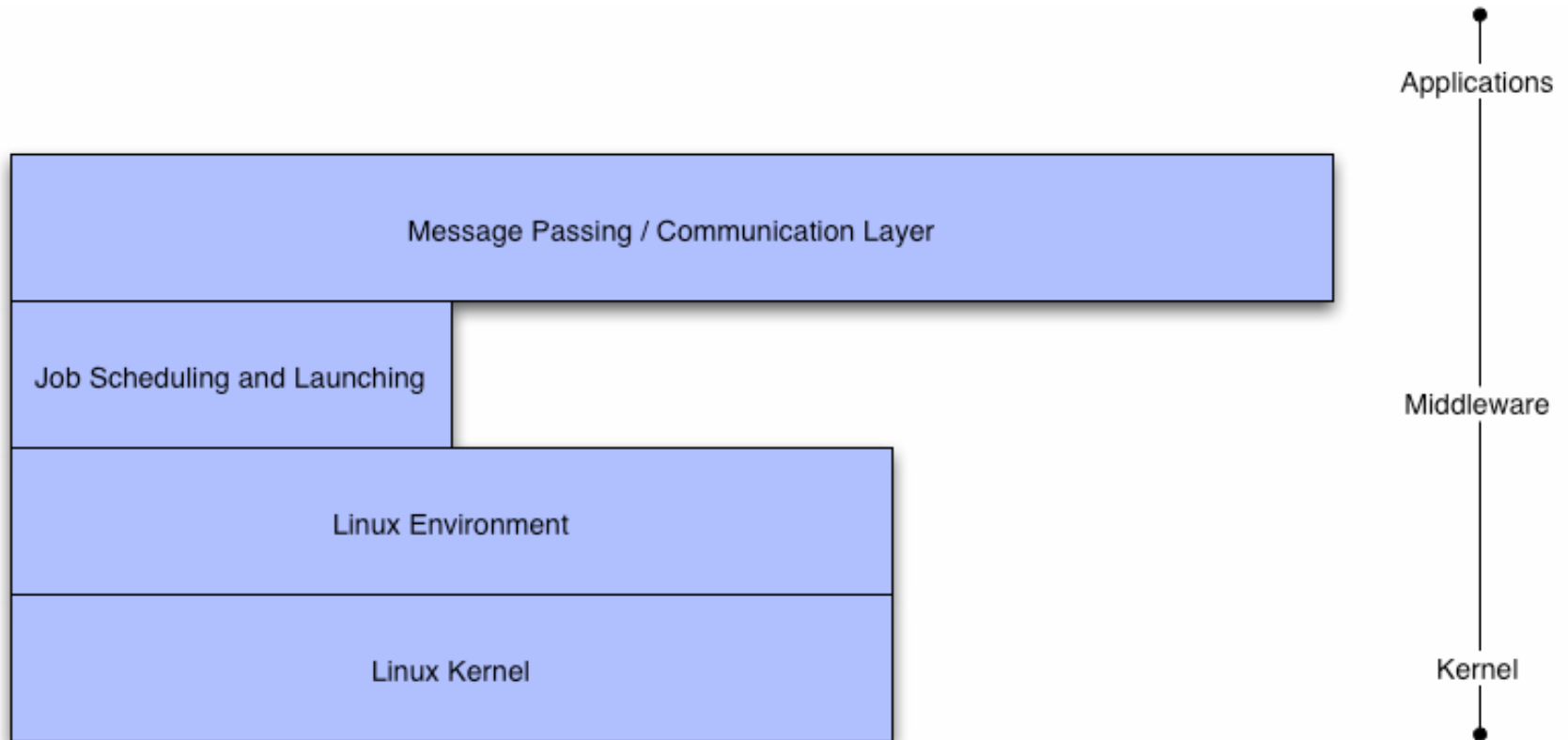
Rocks 3.2 (released May 2004)

CPUs	x86	IA64	Opteron
Queuing System	Sun Grid Engine		Condor
Grid	NMI R4	Simple CA	SGE GRAM

Cluster Software Stack



Common to Any Cluster



Red Hat

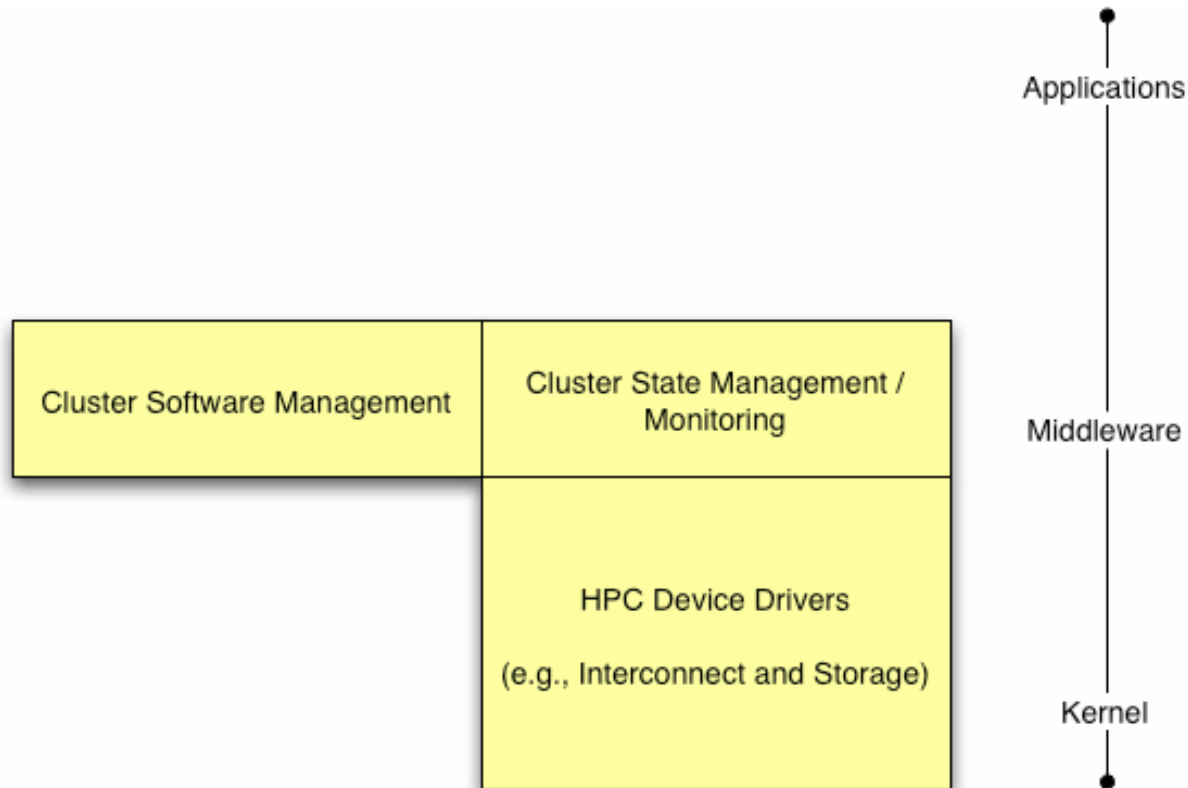
- Enterprise Linux 3.0
 - **Recompiled from public SRPMS, including errata updates (source code)**
 - **No license fee required, redistribution is also fine**
 - **Recompiled for all CPU types (x86, Opteron, Itanium)**
- Standard Red Hat Linux kernel
 - **No Rocks added kernel patches**
- No support for other distributions
 - **Red Hat is the market leader for Linux**
 - In the US
 - And becoming so in Europe
 - **Trivial to support any Anaconda-based system**
 - **Others would be harder (SuSe ~ 3 months work)**
- Excellent support for automated installation
 - **Scriptable installation (Kickstart)**
 - **Very good hardware detection**

Batch Systems

- Portable Batch System and Maui
 - **Long time standard for HPC queuing systems**
 - **Maui provides backfilling for high throughput**
 - **PBS/Maui system can be fragile and unstable**
 - **Multiple code bases:**
 - PBS
 - OpenPBS
 - PBSPro
 - Scalable PBS
- Sun Grid Engine
 - **Rapidly becoming the new standard**
 - **Integrated into Rocks by SCS (Singapore)**
 - **Now the default scheduler for Rocks**
 - **Robust and dynamic**
 - **Currently 5.3, moving to 6.0 when out of Beta**



Rocks Cluster Software

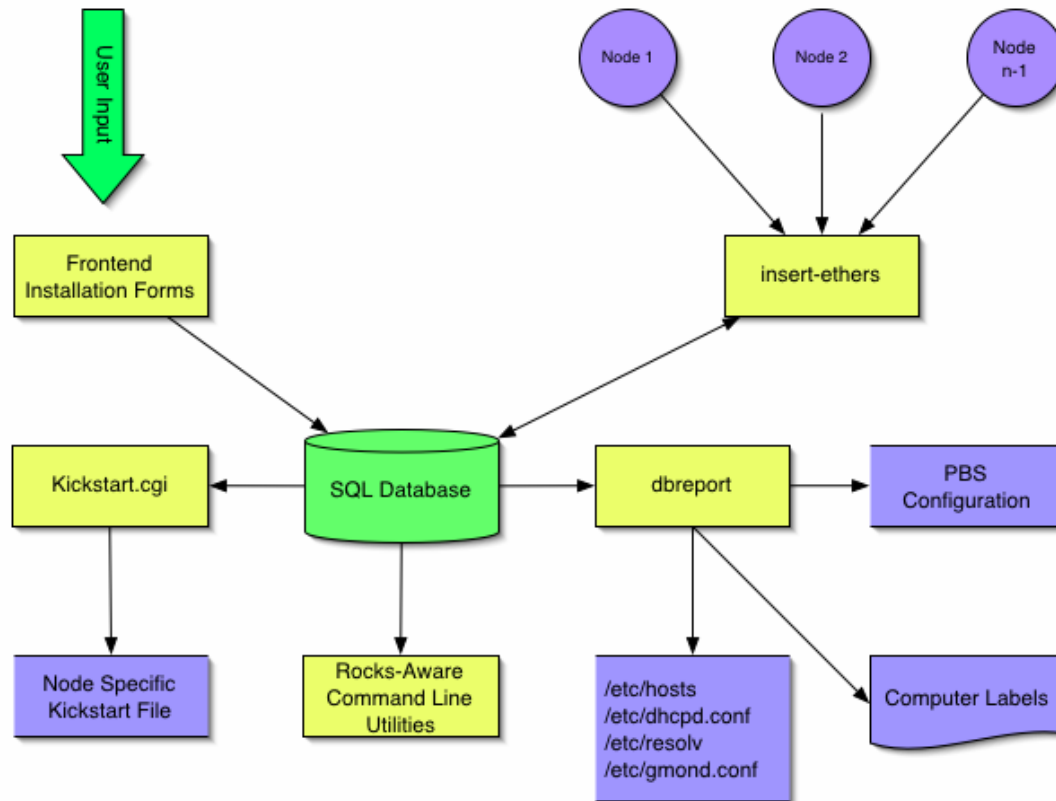


Cluster State Management

- Static Information
 - **Node addresses**
 - **Node types**
 - **Site-specific configuration**
- Dynamic Information
 - **CPU utilization**
 - **Disk utilization**
 - **Which nodes are online**

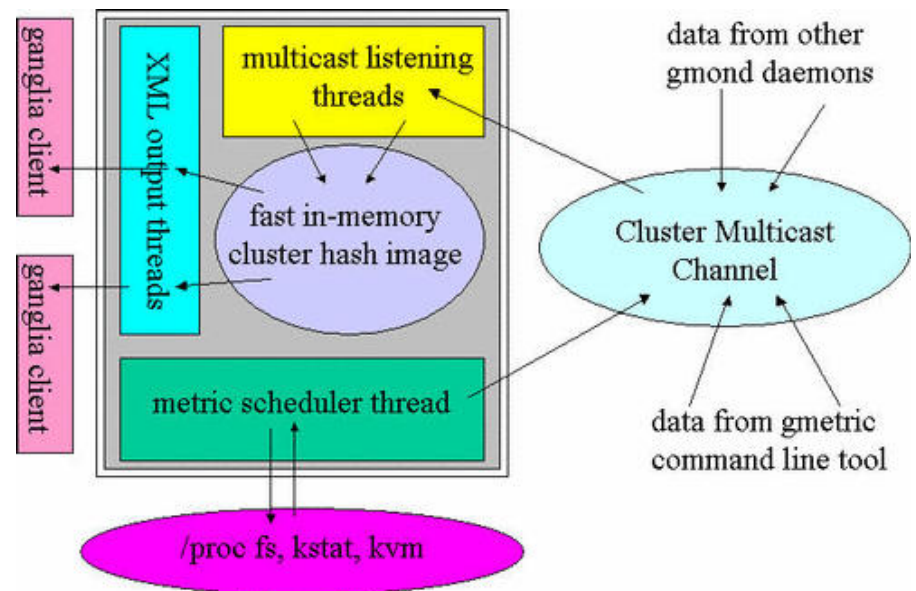


Cluster Database



Ganglia

- Scalable cluster monitoring system
 - **Based on ip multi-cast**
 - **Matt Massie, et al from UCB**
 - <http://ganglia.sourceforge.net>
- Gmond daemon on every node
 - **Multicasts system state**
 - **Listens to other daemons**
 - **All data is represented in XML**
- Ganglia command line
 - **Python code to parse XML to English**
- Gmetric
 - **Extends Ganglia**
 - **Command line to multicast single metrics**



Ganglia Screenshot



Host Report for Tue, 18 Mar 2003 01:28:58 +0000

Get Fresh Data



Last

Node View

Our Cluster > britannic

britannic Overview



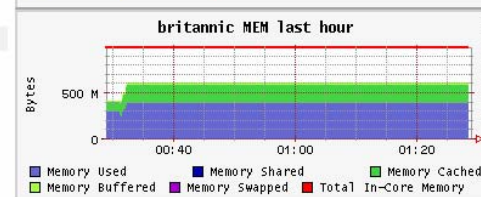
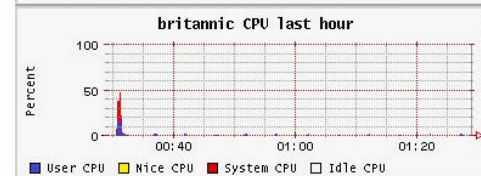
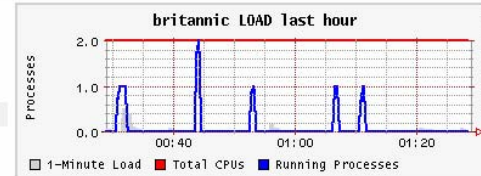
This node is up and running

Time and String Metrics

Name	Value
boottime	Tue, 18 Mar 2003 00:23:20 +0000
gexec	OFF
machine_type	ia64
os_name	Linux
os_release	2.4.18-e.12smp
sys_clock	Tue, 18 Mar 2003 00:25:34 +0000
uptime	0 day, 1:5

Constant Metrics

Name	Value
cpu_idle	97.1 %
cpu_num	2
cpu_speed	900 MHz
mem_total	1011568 KB
mtu	1500 B
swap_total	1048544 KB



Cluster Software Management

Software Packages

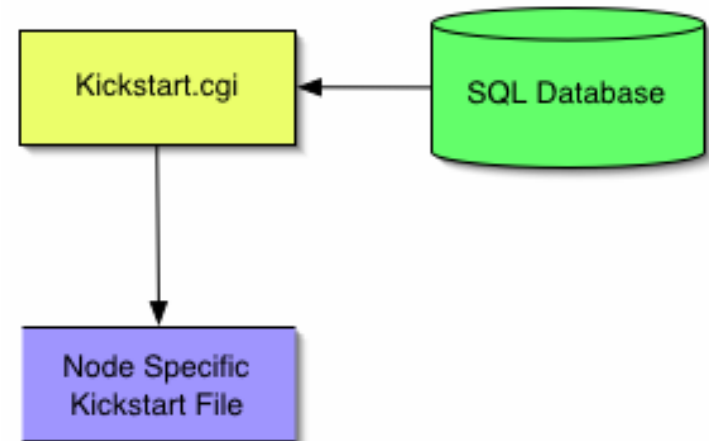
- RPMs
 - **Standard Red Hat (desktop) packaged software**
 - **Or your own addons**
- Rocks-dist
 - **Manages the RPM repository**
 - **This is the distribution**

Software Configuration

- Tuning RPMs
 - **For clusters**
 - **For your site**
 - **Other customization**
- XML Kickstart
 - **Programmatic System Building**
 - **Scalable**

Kickstart

- Red Hat's Kickstart
 - **Monolithic flat ASCII file**
 - **No macro language**
 - **Requires forking based on site information and node type.**
- Rocks XML Kickstart
 - **Decompose a kickstart file into nodes and a graph**
 - Graph specifies OO framework
 - Each node specifies a service and its configuration
 - **Macros and SQL for site configuration**
 - **Driven from web cgi script**



Sample Node File

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE kickstart SYSTEM "@KICKSTART_DTD@" [!ENTITY ssh "openssh">]>
<kickstart>
  <description>
    Enable SSH
  </description>

  <package>&ssh;</package>
  <package>&ssh;-clients</package>
  <package>&ssh;-server</package>
  <package>&ssh;-askpass</package>

</post>

cat &gt; /etc/ssh/ssh_config &lt;&lt; 'EOF' <!-- default client setup -->
Host *
  ForwardX11 yes
  ForwardAgent yes
EOF

chmod o+rx /root
mkdir /root/.ssh
chmod o+rx /root/.ssh

</post>
</kickstart>>
```

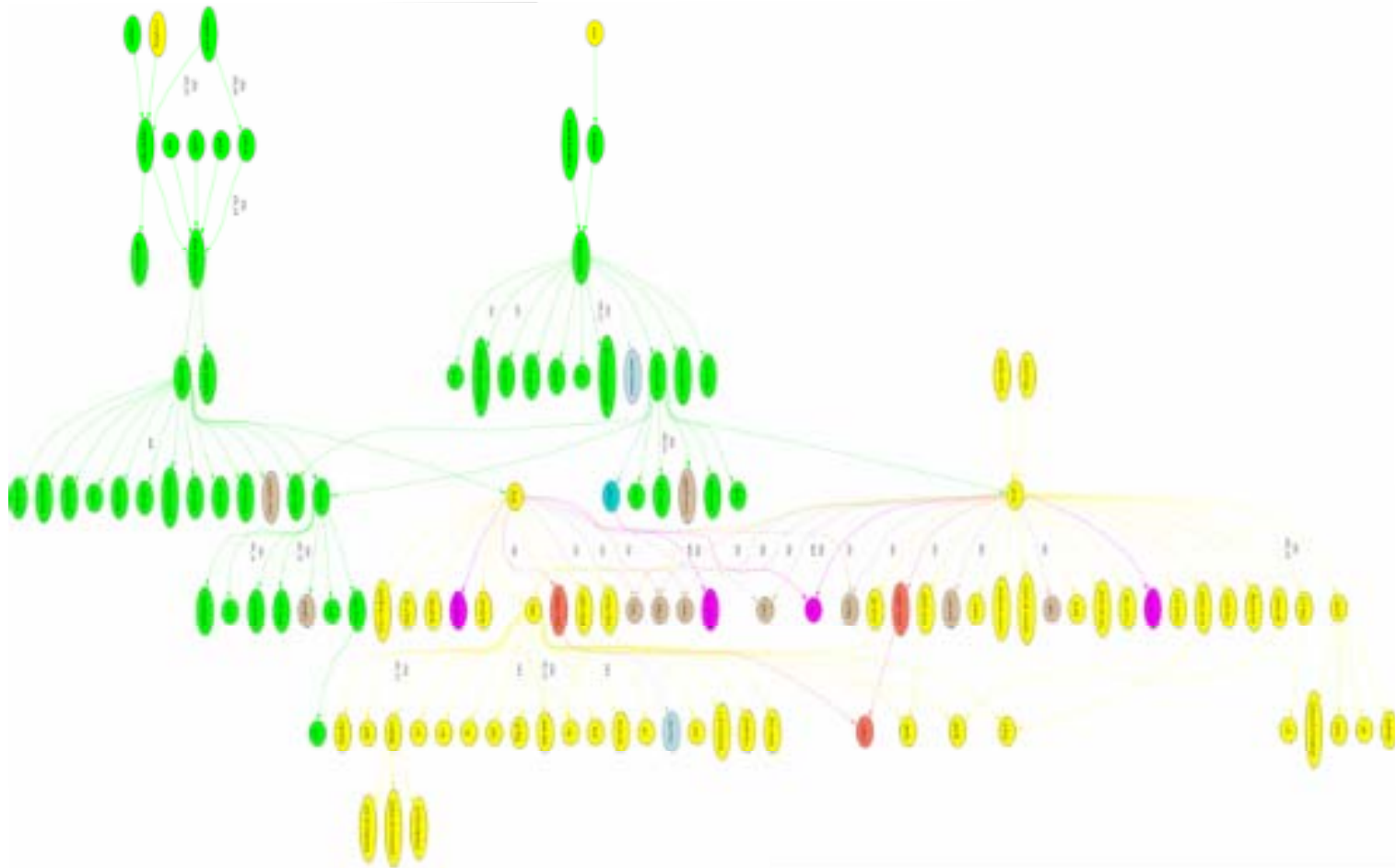
Sample Graph File

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE kickstart SYSTEM "@GRAPH_DTD@">

<graph>
  <description>
    Default Graph for NPACI Rocks.
  </description>

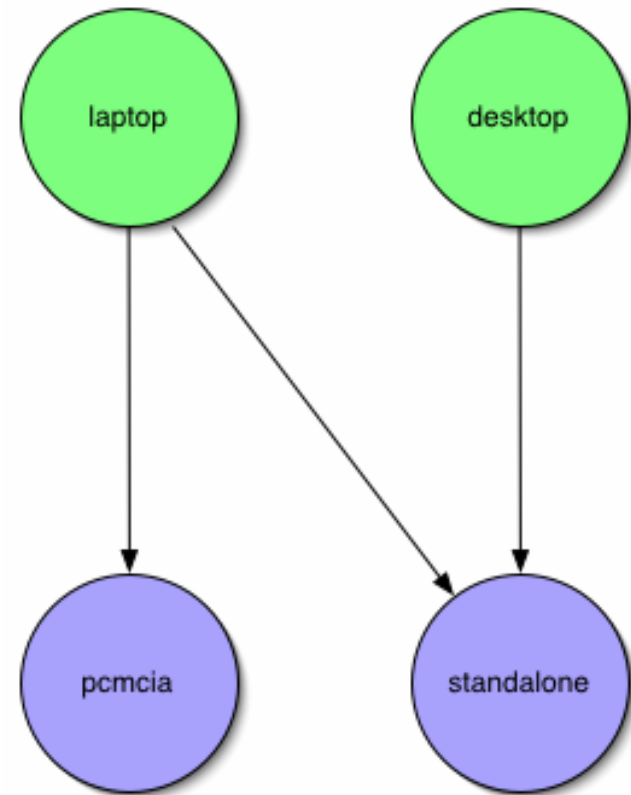
  <edge from="base" to="scripting"/>
  <edge from="base" to="ssh"/>
  <edge from="base" to="ssl"/>
  <edge from="base" to="lilo" arch="i386"/>
  <edge from="base" to="elilo" arch="ia64"/>
  ...
  <edge from="node" to="base" weight="80"/>
  <edge from="node" to="accounting"/>
  <edge from="slave-node" to="node"/>
  <edge from="slave-node" to="nis-client"/>
  <edge from="slave-node" to="autofs-client"/>
  <edge from="slave-node" to="dhcp-client"/>
  <edge from="slave-node" to="snmp-server"/>
  <edge from="slave-node" to="node-certs"/>
  <edge from="compute" to="slave-node"/>
  <edge from="compute" to="usher-server"/>
  <edge from="master-node" to="node"/>
  <edge from="master-node" to="x11"/>
  <edge from="master-node" to="usher-client"/>
</graph>
```

Kickstart framework



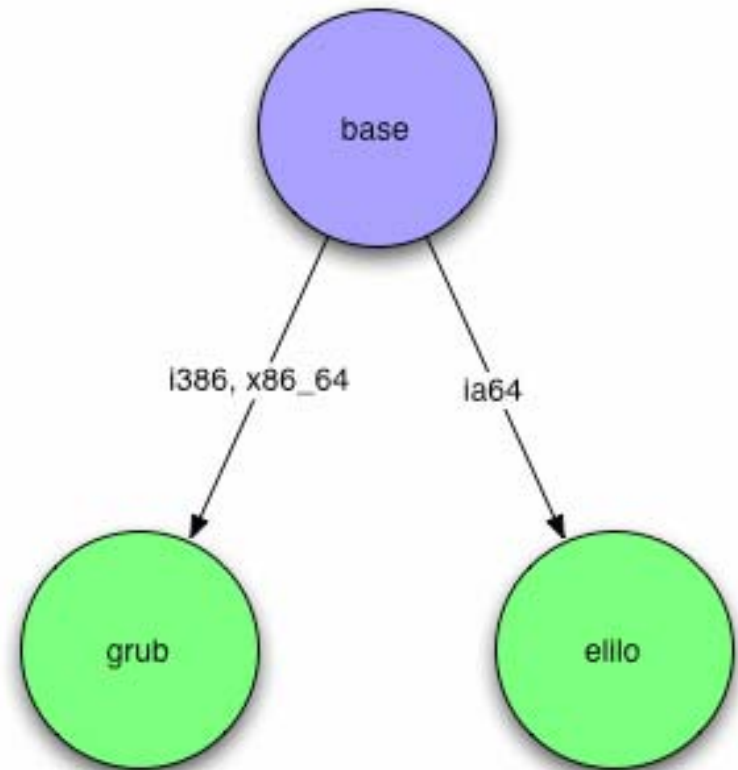
Appliances

- Laptop / Desktop
 - **Appliances**
 - **Final classes**
 - **Node types**
- Desktop IsA
 - **standalone**
- Laptop IsA
 - **standalone**
 - **pcmcia**
- Code re-use is good



Architecture Differences

- Conditional inheritance
- Annotate edges with target architectures
- if i386
 - **Base IsA grub**
- if ia64
 - **Base IsA elilo**
- One Graph, Many CPUs
 - **Heterogeneity is easy**
 - **Not for SSI or Imaging**



Your Cluster Software

- Currently Provided by
 - **Vendors**
 - Callident - BioBrew
 - Paracell - Sequencing
 - **Users**
 - Compile and run
- Future Rocks Rolls
 - **Bio Informatics**
 - **Visualization**
 - **Chemistry (GAMESS)**
 - **Ninf-G Applications (AIST)**



Rockstar Cluster (SC'03)

- Built live at Super Computer 2003
 - **From ground up (hw + sw)**
 - **In under 2 hours**
- Demonstrate
 - **We are now in the age of “personal supercomputing”**
 - **Highlight abilities of:**
 - Rocks
 - SGE
- Top500 list (#201)
- Hardware
 - **129 Intel Xeon servers**
 - 1 Frontend Node
 - 128 Compute Nodes
 - **Gigabit Ethernet**
 - \$13,000 (US)
 - 9 24-port switches
 - 8 4-gigabit trunk uplinks

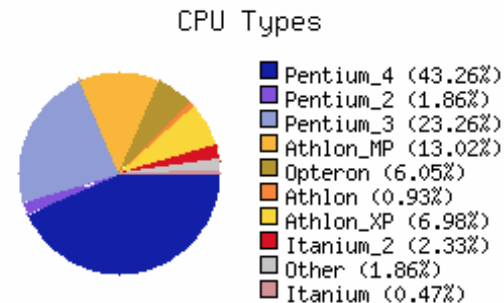


Building of Rockstar

QuickTime™ and a
MPEG-4 Video decompressor
are needed to see this picture.

Itanium Clusters?

- Linux is an x86 OS
 - **Slow to support others**
 - **Largest user base**
 - **Google debugging is x86**
- Rocks Clusters are x86
 - **75% x86**
 - **25% AMD**
 - **< 1% IA64**
- But we do support IA64



Challenges for IA64

- Pricing
 - **CPU prices are falling**
 - **Servers are still 3x Xeon pricing**
 - HP DL140 (dual 3.2 Xeon, 2GB RAM, 80 GB disk) \$3468
 - HP ZX6000 (dual 1.3 Itanium2, 2GB RAM, 73 GB disk) \$9869
 - **This is improving**
- FUD
 - **Opteron is 64bit (and also 32bit)**
 - **Nacona (Intel now has 2 64bit CPUs)**
 - **Everyone is 64 bit now**
- EFI
 - **Poor documentation**
 - **Difficult to manipulate BIOS boot order from Linux**
 - **Intel has announced all platforms are moving to EFI (BIOS is dead)**
- Seeding Labs and Researchers
 - **Very few people want to be CPU pioneers**
 - **Experience base for IA64 clusters is extremely small**

Opportunities for IA64

- Xeon has hit a “thermal wall”!
 - **Intel has announced Xeons will not get faster**
 - Tejas, Jayhawk lines are gone
 - **Switching to**
 - Multicore for servers
 - Mobile for desktops
 - **Where does x86 HPC go?**
 - Non-commodity multicore?
 - Low speed mobile?
 - Itanium?
- Large physical memory
 - **Opteron does not have enough DIMMS slots**
- TeraGrid, PNNL, others
 - **Large existence proofs for IA64 clusters / grids**
 - **Building the google expertise**



www.rocksclusters.org

Questions?