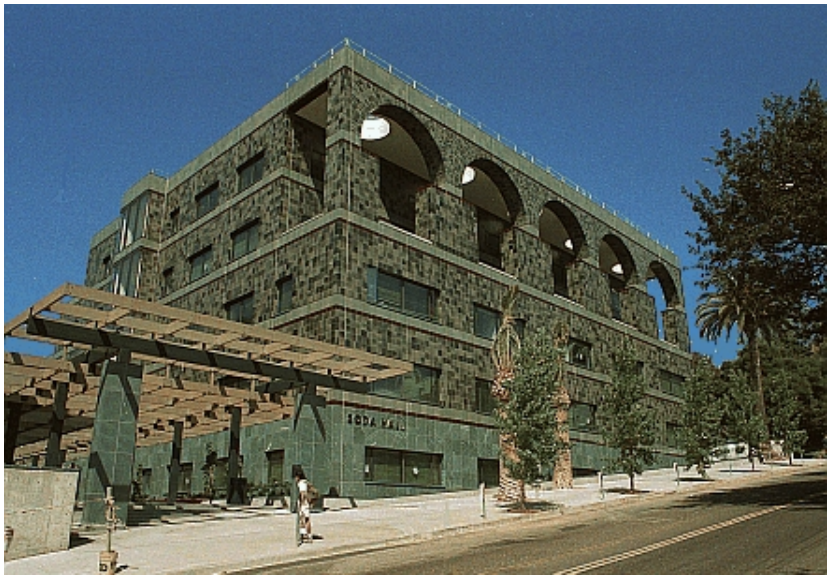
NPACI Rocks: Tools and Techniques for Easily Deploying Manageable Linux Clusters

Philip M. Papadopoulos, Mason J. Katz and Greg Bruno

IEEE Cluster 2001, Newport Beach, CA
October 10, 2001

NPACI Rocks Is Made Possible By ...

- UC Berkeley
  - David Culler
    - Co-Principal Investigator for the *Network of Workstations* and *Millennium* projects
  - And his talented staff:
    - Eric Frazer
    - Matt Massie
    - Albert Goto



- Compaq Computer Corporation
  - Especially our account representative Sally Patchen
  - Early access to Itanium and blade servers
  - IA-32 equipment donations
  - Testing of Rocks in corporate environment

- IBM
  - Equipment donations through Shared University Research (SUR) program

Motivation and Goals

- We Hate System Administration

- Enable Non-Cluster Experts to Run Clusters
  - Should be easy to deploy, expand/contract and manage
- Essential to Track Software Updates
  - Open source moves fast!
    * Red Hat 6.2: 191 updates
    * Red Hat 7.0: 176 updates
    * Red Hat 7.1: 91 updates
      · In 177 days, that's 3.5 updates a week!
- Essential to Track Red Hat Releases
  - NPACI Rocks built on top of a full Red Hat release
- Run on Heterogeneous, Standard High-Volume Components

Philosophy

- All nodes are 100% automatically installed
  - Zero hand configuration
    - All node-local configuration is automatically generated

- NPACI Rocks is an Entire Cluster-Aware Distribution
  - Included packages:
    - Full Red Hat release
    - De-facto standard cluster packages (e.g., MPI, PBS, Maui)
    - NPACI Rocks packages
  - Focus on ease of use for cluster lifecycle
    - Deployment, management, application development and execution
    - All services required to install compute nodes, develop and run parallel jobs are bundled in
    - Initial configuration via simple web page
    - One CD installs all servers and nodes in a cluster

More Philosophy - Common-Mode Mechanism: Install

- Software Install is the Common Action Performed When:
  - First bringing up a cluster

    ```
    # insert-ethers
    ```

  - Replacing a dead server

    ```
    # insert-ethers --replace=<dead-node>
    ```

  - Adding a new server to the cluster

    ```
    # insert-ethers --cabinet=1
    ```

- We Use the "Install" Mechanism For One More Function: Software Consistency
  - Question: "Is server X's software up-to-date?"
  - Question: "Is server X's configuration up-to-date?"
  - Question: "How do restore server X to the last known-good state?"
  - Answers: Reinstall. Wait 10 minutes. "Yes."

Installation Performance

Nodes	Total Reinstall Time (minutes)
1	10.3
2	9.8
4	10.1
8	10.4
16	11.1
32	13.7

- Setup:
  - HTTP server: dual 733 MHz PIII, 100 Mbit Ethernet
  - Compute nodes: 733 MHz - 1 GHz with Myrinet
  - Each node transfers approximately 150 MB of Red Hat packages

Hardware Configuration

- Minimum Components
  - Server (x86 or IA-64) with a local hard disk
  - Ethernet
  - Power
- Optional
  - High-Performance Network (e.g., Myrinet)
  - Network-Addressable Power Distribution Unit
- Evil Keyboard/Video/Mouse Network Not Required
  - Pros:
    * Works on all standard high-volume hardware
    * Don't have to manage yet another (low volume and/or proprietary!) network
  - Cons:
    * Can't interact with BIOS remotely
    * Blind until kernel brings up network
    * Can't interact with installations remotely. Or, can you ...

eKV – Ethernet Keyboard and Video

- Developed eKV to monitor and interact with installations

- After Red Hat's Kickstart brings up the network, one can interact with the installation via `telnet`
  - Telnet server disabled on normal operation


```
$ telnet compute-1-2 8000
```

Managing Your Software With `rocks-dist`

- Tool to Manage and Customize Your Rocks Distribution
  - Used to keep your distribution up-to-date
  - Used to collect all packages (Red Hat + NPACI Rocks + your own) into a Red Hat++ distribution
  - All the software components that *could be* installed

- Step 1: Mirror

```
$ rocks-dist mirror
```

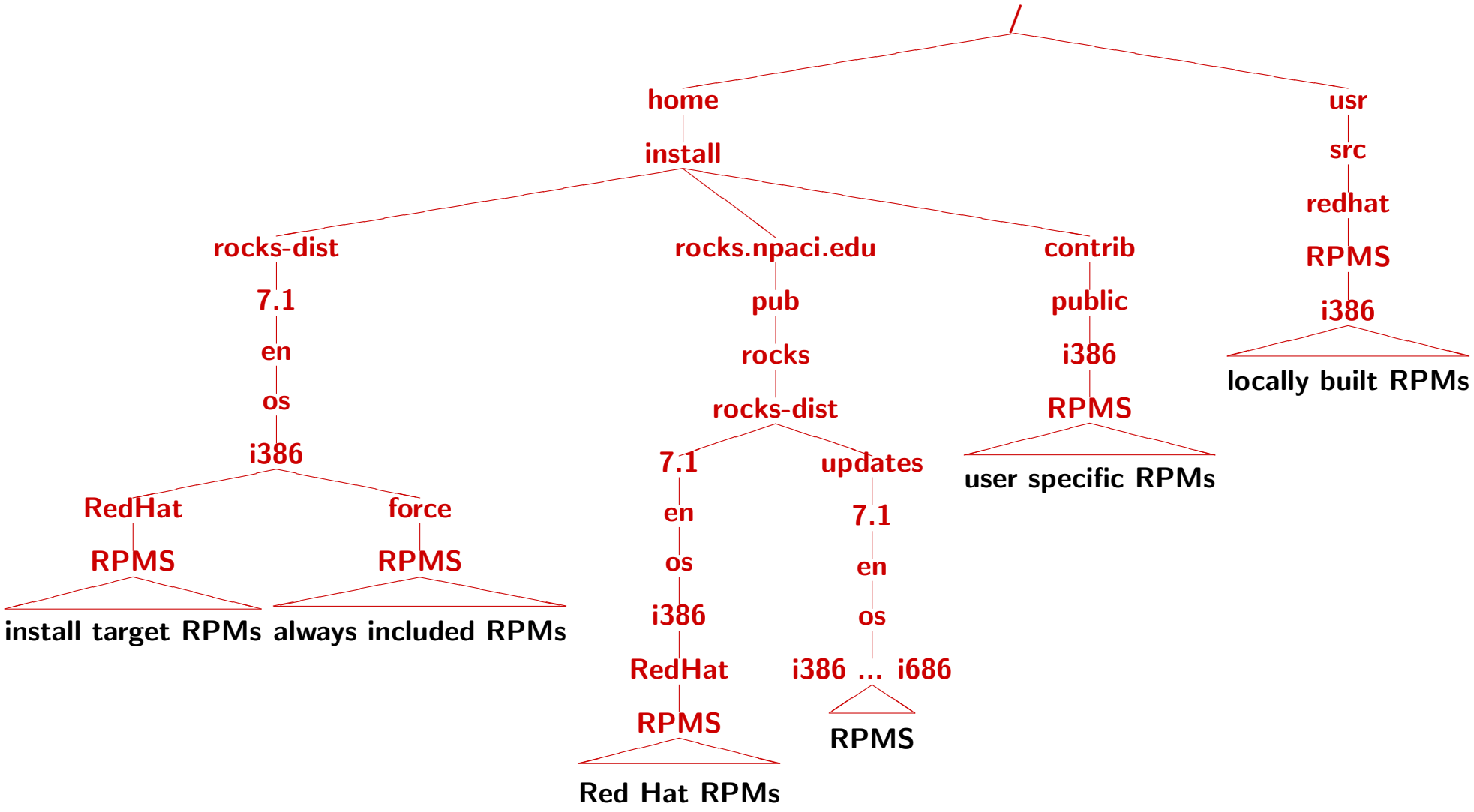
  - This mirrors the entire Rocks distribution from SDSC

- Step 2: Customize Packages
  - Put in the packages you want

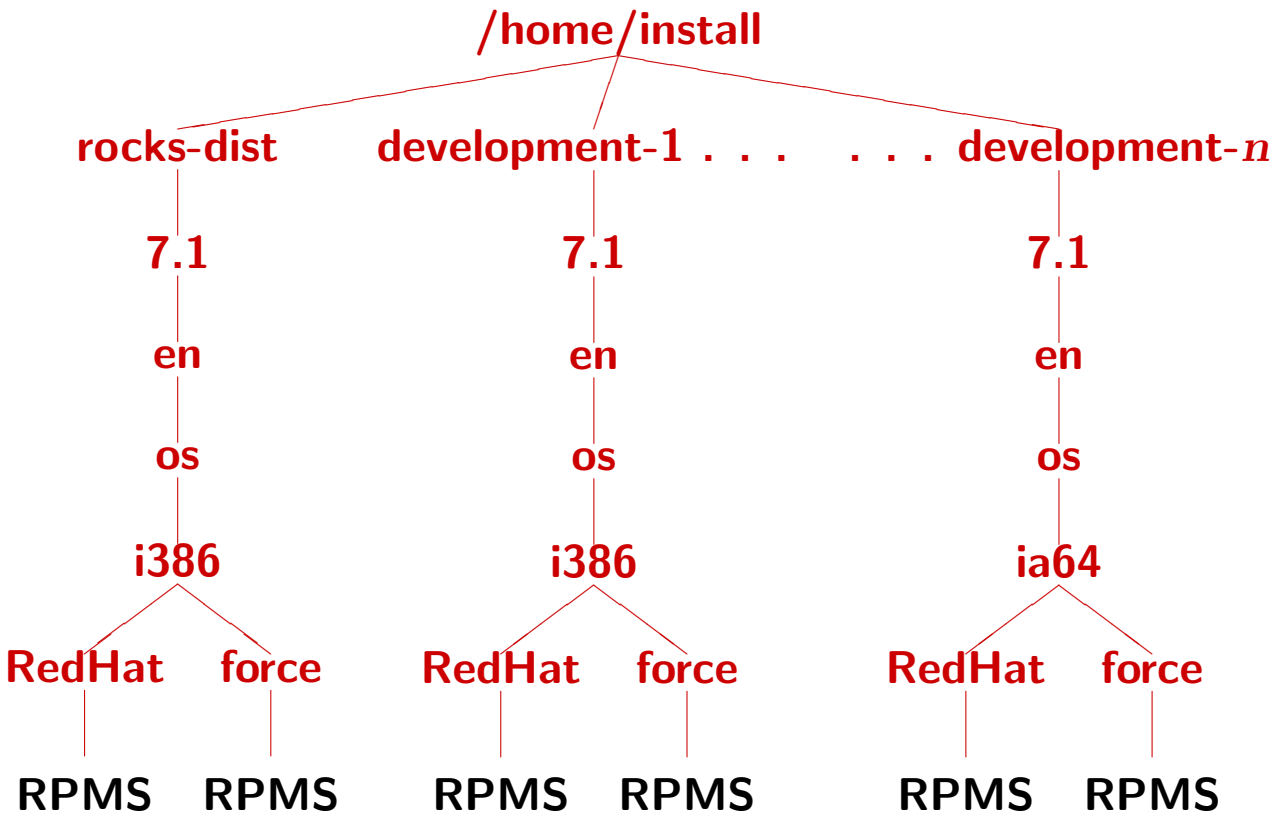
- Step 3: Rebuild Distribution

```
$ rocks-dist dist
$ rocks-dist --dist=development dist
```

rocks-dist – RPM Locations

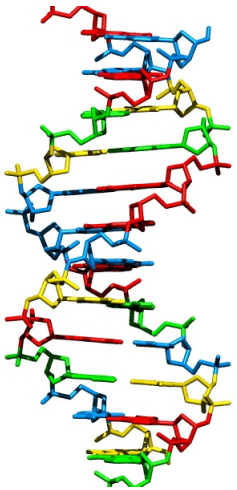


`rocks-dist` – Default and Development Trees



Description-Based Software Configuration

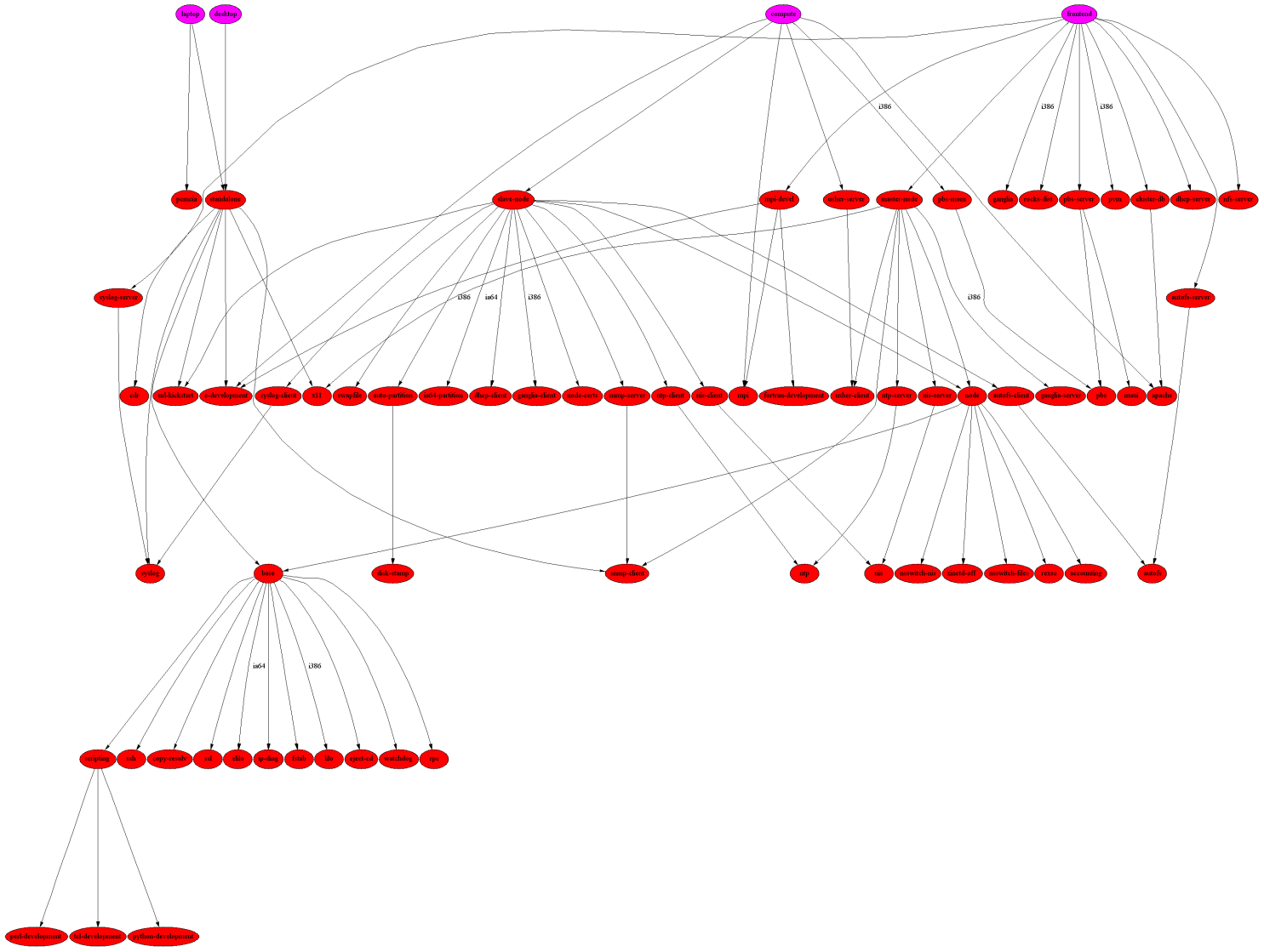
- Built an infrastructure that "describes" the roles of cluster nodes
  - Nodes are installed using Red Hat's *kickstart*
  - Kickstart file: ASCII file with names of packages to install and "post processing" commands
  - NPACI Rocks builds kickstart files on-the-fly tailored for each node



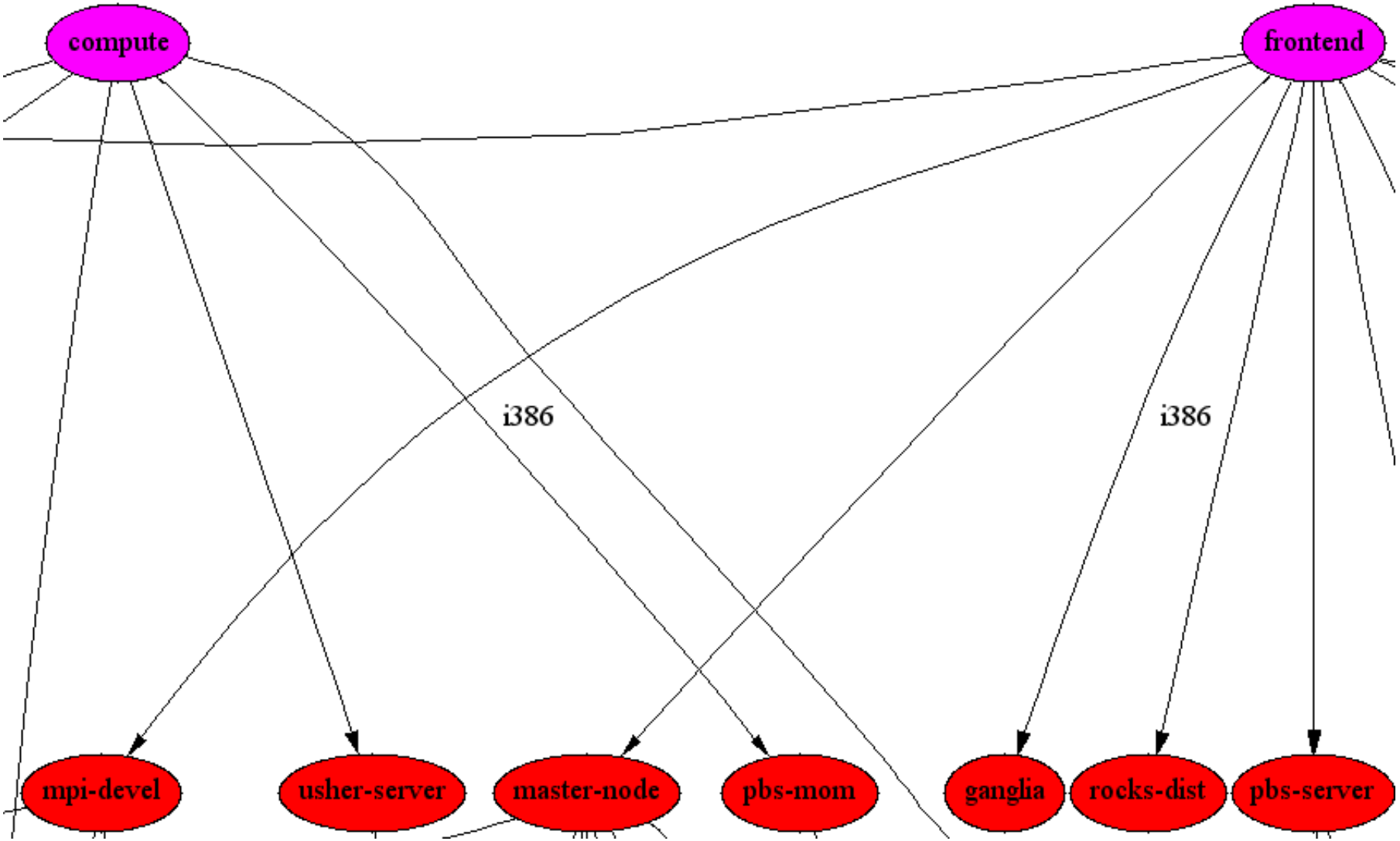
- NPACI Rocks kickstart file is general configuration + local node configuration
  - General configuration is described by modules linked in a configuration graph
  - Local node configuration (applied during post processing) is stored in a MySQL database

- This strategy is extremely flexible
  - Heterogeneous hardware is no harder than homogeneous
  - Straight-forward to customize

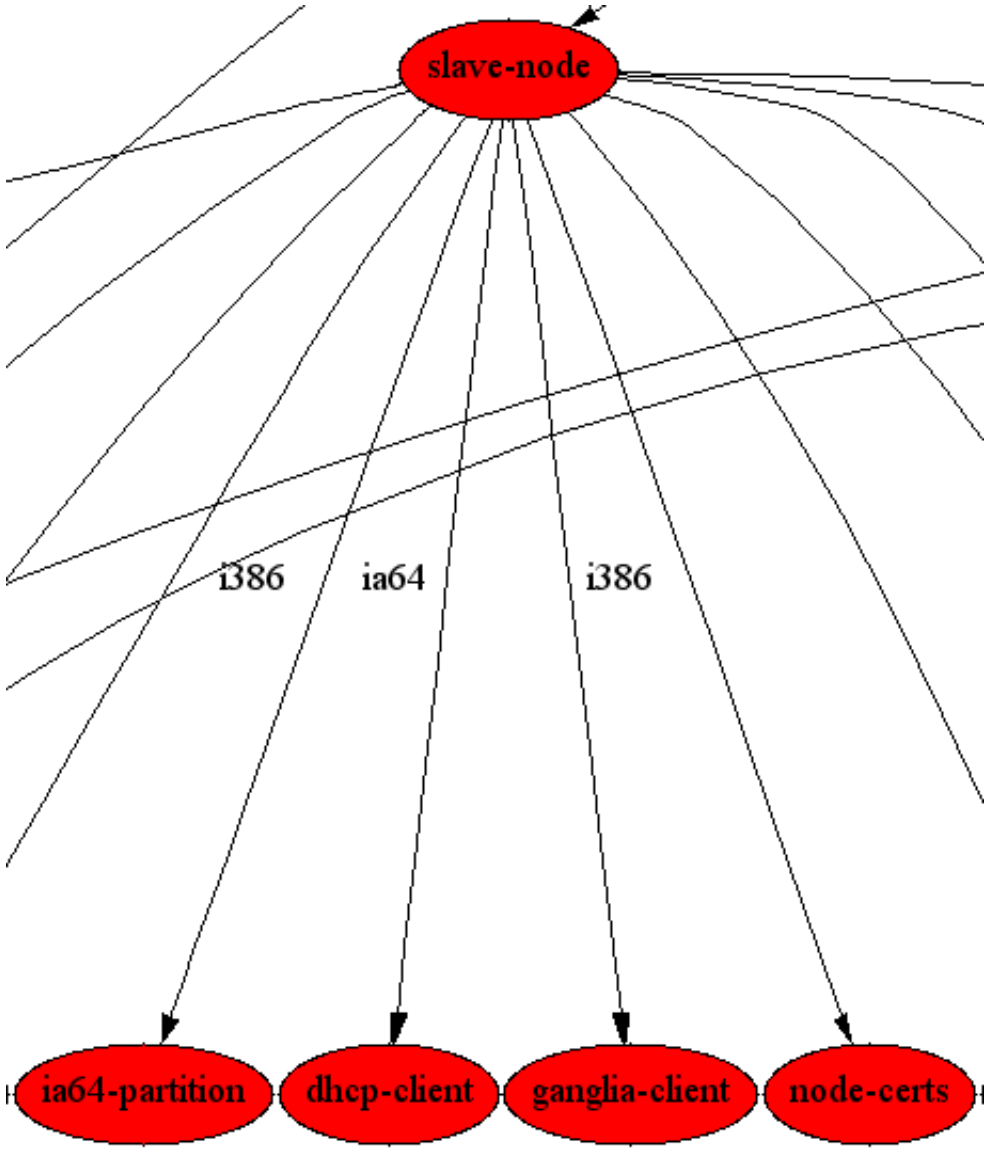
General Description Serves Every Cluster Component



Cluster Description – "Appliances"



Cluster Description – Architecture Switches



The Payoff – Integrating Never Before Seen Hardware

- Dual-Athlon White Box, 20 GB IDE, 3Com Ethernet
  - 3:00 PM: In cardboard box
  - Shook out the loose screws
  - Dropped in a Myrinet card
  - Inserted it into cabinet 0
  - Cabled it up
  - 3:25 PM: Inserted the NPACI Rocks CD
  - Ran `insert-ethers` (assigned node name `compute-0-24`)
  - 3:40 PM: Ran Linpack

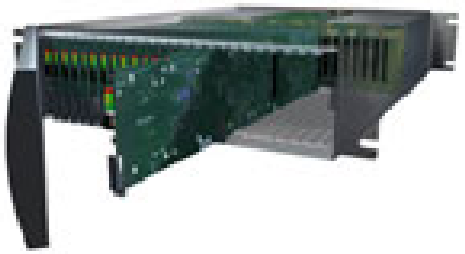
- Two IBM Dual-Itanium (IA-64), 18 GB SCSI, Intel Ethernet
  - 2:00 PM: In box
  - 3:40 PM: Debugged problem with nodes and 2.4.6 kernel
  - Downloaded 2.4.9 kernel RPMs from Red Hat's *rawhide* release
  - Rebuilt distribution with `rocks-dist`
  - 4:30 PM: Both integrated into cluster

Both machine types were installed from the same general description

Futures

- IA-64
  - Full IA-64 cluster support (frontend and compute nodes) to be released Nov '01
- Pre-Execution Environment (PXE) Boot
  - Nice for newer rack-mounted servers, but essential for blade servers
    - Blade servers: CPU + Disk + Ethernet + Proprietary Mgmt Network
      - Will look like any Rocks cluster, as all our tools run over Ethernet
  - Release: Nov '01

- Infiniband Interconnect
- Grid Tools (Development and Testing) - Rocks is one of many good targets for grid software
  - Globus
  - Grid research tools (APST)
  - Gridport toolkit



Status

- Growing User Base: academic, government and industrial sites around the world
  - We've installed 6 clusters at UCSD
    - Our cluster, "Meteor", is a 100-node cluster
    - Currently building out two 128-node clusters for the Scripps Institute of Oceanography
  - Pentium, Athlon, IDE, SCSI, Integrated RAID, Lots of Ethernet chips, Myrinet
- Freely Downloadable ISO Image
- All NPACI Rocks developed code is released in binary and source Red Hat packages

`http://rocks.npaci.edu`