Ruđer Bošković Institute

Computing and Informatics Center

High Performance Cluster Distribution Design

Authors:

Nikola Pavković

Valentin Vidić

Karolj Skala

CUC 2004

Cluster Classification

- High Availability Clusters
  - Redudancy, fail-proof design
- High Performance Clusters
  - Distributed processing capabilities

- Single System Image
- Non-SSI

Cluster Classification

- High Availability Clusters
  - Redudancy, fail-proof design
- High Performance Clusters
  - Distributed processing capabilities

- Single System Image
- Non-SSI

Facing the problems...

- Automatic node installation
- Filesystem sharing
- Centralized user-account management
- Resource management
- System health monitoring
- Node configuration model

Existing solutions…

- Complete solutions
  - NPACI ROCKS, OSCAR
  - RedHat (RPM)
- Stand-alone tools
  - System Installation Suite, FAI
  - Prepackaged HPC software (MPI, PVM…)
- No complete Debian-based solution!

The Idea with Debian GNU/Linux

- System Installation Suite
- Prepackaged software
  - mpi, pvm, lapack…
- Additional software
  - Torque queuing system
  - C3 suite
  - Ganglia-webfrontend
  - …

SIS suite

- Automatic installation of nodes
- Automatic image building
- Comfortable image administration model
- Database backend
- Issues
  - Manual database administration
  - Bugs in the SIS suite
    - Patched ☺

The SIS database

- Hostnames, MACs, Network settings…
- DHCP server configuration
- The main problem…
  - Discovery of new nodes
  - New nodes are added manually to the database
  - After that the automatic installation can begin
- The solution…
  - discover_node

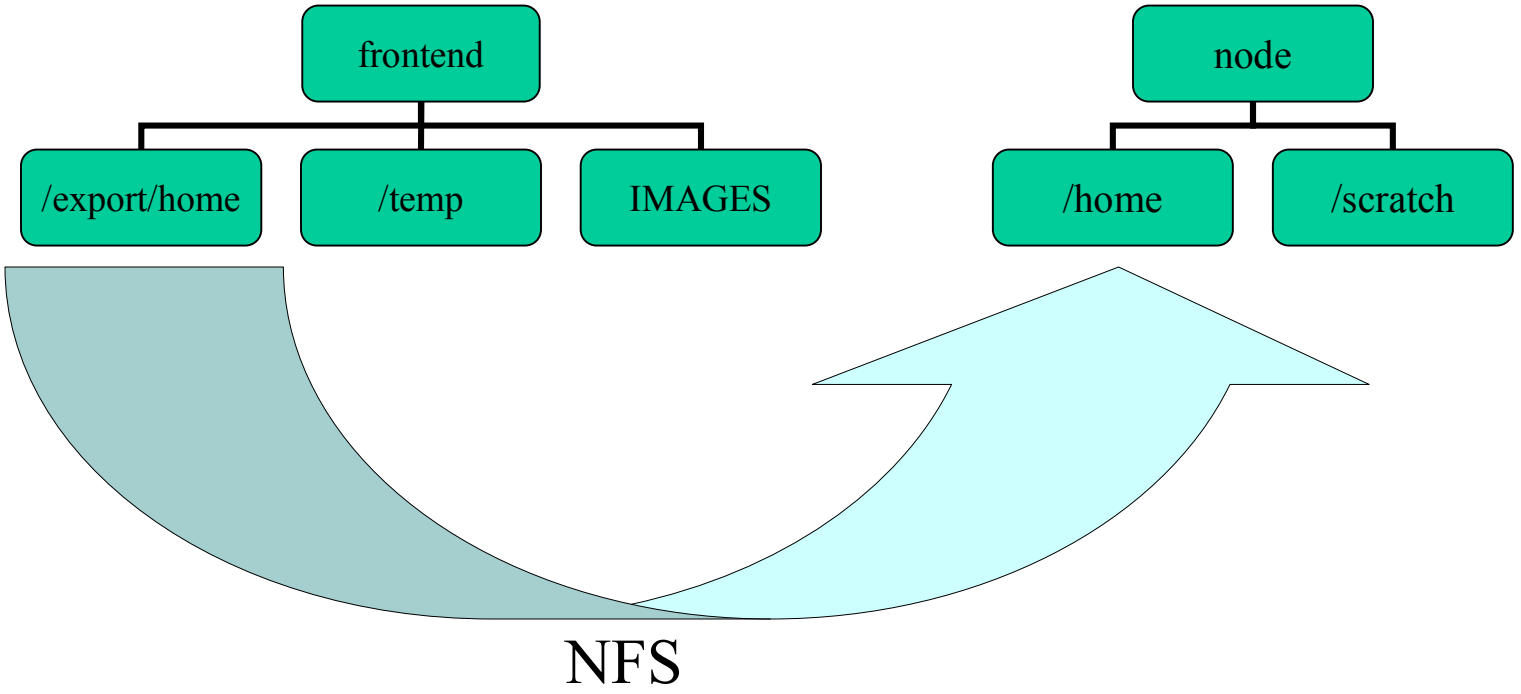
discover_node

- A typical DHCP request log entry:

```
Aug 28 14:15:02 fk-grozd dhcpd: DHCPDISCOVER from
00:02:b3:9c:40:ad via eth1: network 10.0.0.0/24: no free
leases
```

1. MAC address extraction
2. Query the SIS database
3. Synchronization of services
   - SIS database, DHCP, /etc/hosts, torque

Filesystems…



Centralized user account management

- User account information
  - /etc/passwd
  - /etc/shadow
  - /etc/group
- The need for centralization
- Solution: LDAP

LDAP integration

- LDAP as the authentication infrastructure
  - Server on the front-node
  - Scalability through replication
- Database administration
  - dcd_adduser
    - User account creation
    - Initial SSH key setup

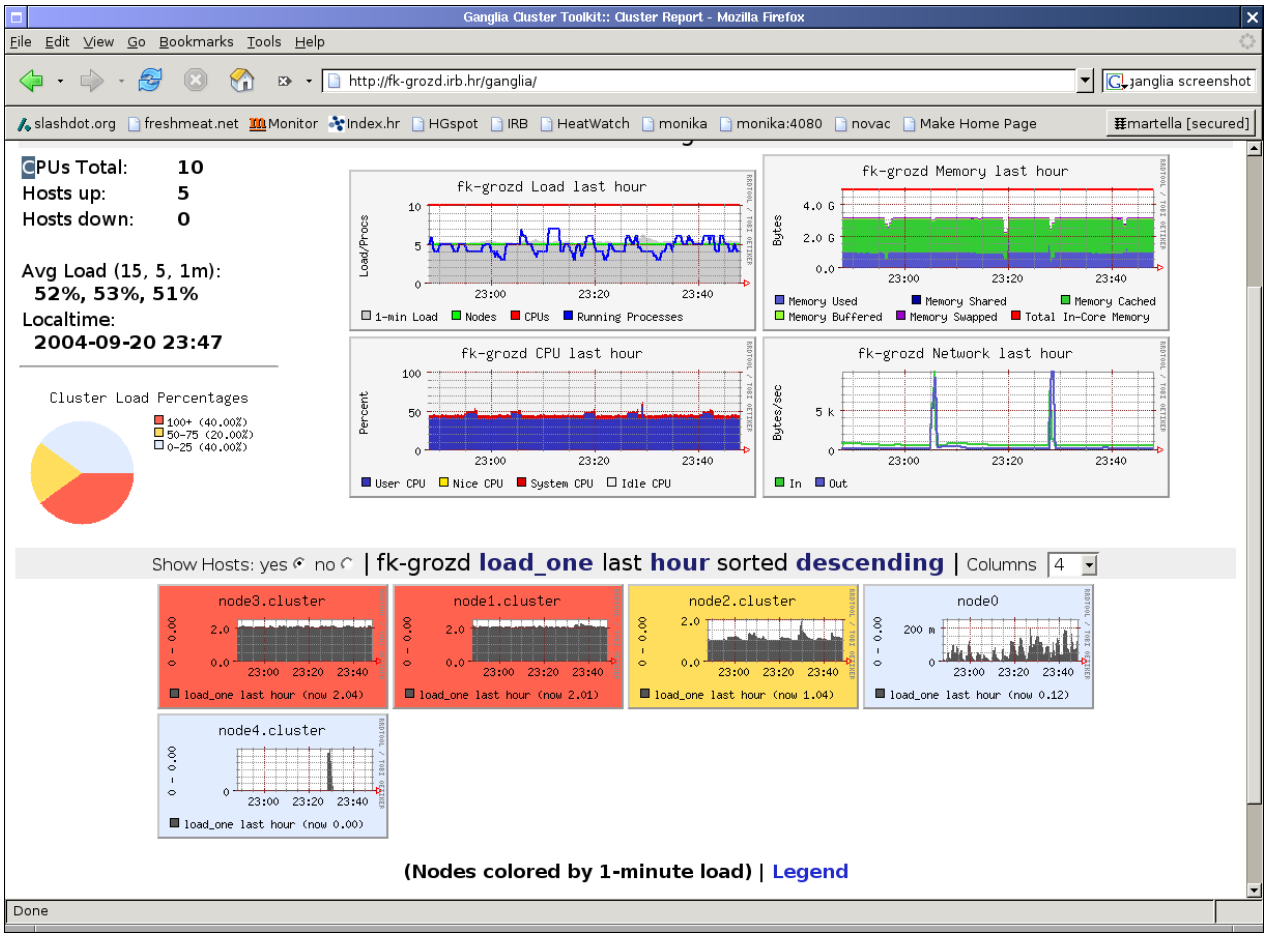
System Health Monitoring

- The scope of the problem
  - System load
  - Disk usage
  - Processor temperature
  - …
- Ganglia
  - Scalability, stability, flexibility

Ganglia

- Packages exist within the Debian tree
- Missing ganglia-web-frontend package
  - Packaged GWF into .deb
  - Automatic configuration
  - Automatic Apache configuration

Ganglia (2)


Ganglia Cluster Toolkit:: Cluster Report - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://fk-grozd.irb.hr/ganglia/
CPUs Total: 10
Hosts up: 5
Hosts down: 0
Avg Load (15, 5, 1m):
52%, 53%, 51%
Localtime:
2004-09-20 23:47
Cluster Load Percentages
fk-grozd Load last hour
fk-grozd Memory last hour
fk-grozd CPU last hour
fk-grozd Network last hour
Show Hosts: yes no | fk-grozd load_one last hour sorted descending | Columns 4
node3.cluster
node1.cluster
node2.cluster
node0
node4.cluster
(Nodes colored by 1-minute load) | Legend
Done

Node configuration model

- SIS images
  - Directories containing images
  - Natural environment for handling system configuration tasks
- chroot issues
  - Restarting the services while installing/upgrading a software package
  - Mounting the /proc filesystem

Editimage

- Safety measures
  - Lock file
  - Fake /sbin/start-stop-daemon
  - Opening the shell (chroot)
  - Mounting the /proc filesystem
  - $PS1 variable

Editimage

- Safety measures
  - Lock file
  - Fake /sbin/start-stop-daemon
  - Opening the shell (chroot)
  - Mounting the /proc filesystem
  - $PS1 variable

Additional software…

- C3 system
  - cexec, cpush, cpushimage
- Queuing system
  - torque, maui
- Commercial software
  - Gaussian, Mathematica, macromodel…

The results…

- Increased system administration performance over other solutions
  - Natural node-configuration environment
  - Easy system maintenance (APT)
  - SIS automation
- Increased security
  - Easy patch management model (APT)
  - Responsiveness of the Debian Security Team

The results… (2)

- Two production-grade clusters on the Rudjer Boskovic Institute (Zagreb)

The Results… (3)

- SystemConfigurator patches
  - Merged into version 2.0.10
- HOWTO document

Future work…

- debian-cluster meta-package
- LDAP
  - Autoconfiguration
  - Replication
  - Debconf, SIS integration…
- Predefined classes for work-nodes