Cluster Management

Jason Allen

OSG All Hands Meeting, Boston

March 8, 2011

About Me

- Fermilab for 10 years
- Head of Fermilab Experiments Facilities (FEF) Department in the Computing Division.
- FEF manages approx 3500 physical Linux systems: farms, workstations, online servers, offline servers, and more!
- FEF is also responsible for the development and support of Scientific Linux. **SL 6 now available!**

Fermilab

For this talk

- Primarily discuss Free and Open Source Software (FOSS) tools
- Tools in production at Fermilab or ones that we have recently evaluated
- “Cluster” defined as any large group of Linux systems
- Skimming the surface of cluster management

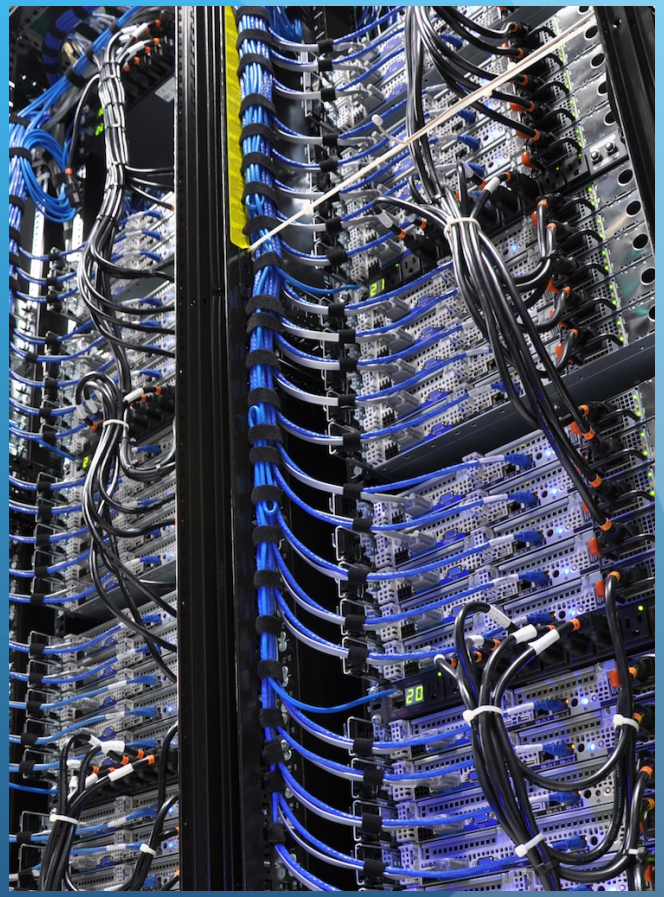
Cluster Management Basics

- Management infrastructure
- Provisioning
- Configuration and software package management
- Monitoring

Fermilab

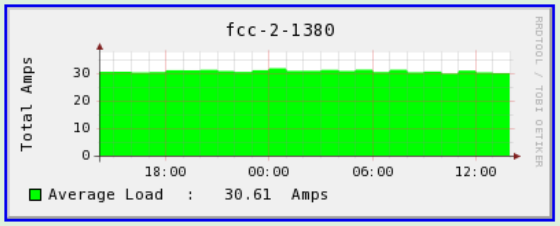
Management Infrastructure

- Remote power-cycling and serial console access
- FEF has standardized on Avocent ACS serial console servers and Avocent PM line of PDU
- Other depts use IPMI or a mix of Avocent and APC products
- All depts have scripts to control and configure remote power-cycling and serial console access



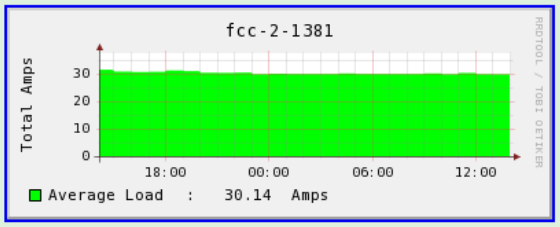
Fermilab

CMS Tier 1 Power Usage Plots



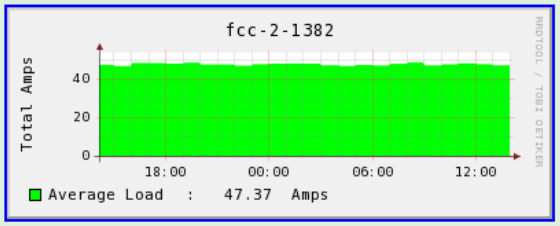
fcc-2-1380

cmssrv03 cmssrv02 cmssrv01 cmssrv95 cmsosgce
cmsosgce2 cmsosgce3 cmsosgce4 cmssrv105 cmssrv14
cmssrv34 cmssrv35



fcc-2-1381

cmsstor252 cmsstor253 cmsstor254 cmsstor255
cmsstor256 cmsstor257 cmsstor258 cmsstor259
cmsstor260 cmsdisk252 cmsdisk253 cmsdisk254
cmsdisk255 cmsdisk256 cmsdisk257 cmsdisk258
cmsdisk259 cmsdisk260



fcc-2-1382

cmsstor155 cmsstor154 cmsstor153 cmsstor150
cmsstor151 cmsstor149 cmsdisk155 cmsdisk154
cmsdisk153 cmsdisk151 cmsdisk149 cmsdisk150

CMS plots power utilization by querying PDUs using SNMP. This data can be particularly useful to datacenter managers.

Fermilab

Provisioning Tools

Preparing a system for use; OS installation and initial configuration.

Fermilab

Provisioning Tools

At Fermilab

- PXE/Kickstart
- Rocks

Fermilab

FEF's PXE/Kickstart Setup

- FEF uses custom tool built on top of MySQL and Perl DCHP server modules.
- No dhcpd restart required.
- Web front-end for specifying kickstart / node combinations.
- Very flexible
- Kickstart files are created dynamically based on selections from the web GUI
- Planning to eval Cobbler later this year.

Rocks Clusters

- Open source Linux distribution based on CentOS.
- Created in 2000
- Created for easy deployment of large clusters.
- Used by CMS Tier 1 at Fermilab for 2400 machines
- CMS can install ~500 systems in 1 hour.
- Only one OS version per Rocks server may be a deal breaker for some.

Fermilab

Configuration and Package Management

Tools that help manage system configuration (files, dirs, permissions, etc.) and software packages.

Fermilab

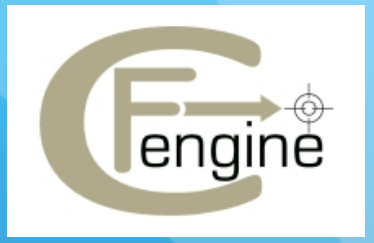
Configuration and Package Management Tools

Popular open source solutions:

- Cfengine
- Puppet
- Bcfg2

Fermilab

Cfengine



- First version released in 1993
- Still the de facto standard configuration management tool, though Puppet is quickly catching-up
- Written in C
- Fairly easy to understand syntax
- Used by D0 workstation cluster, ClueD0
- FEF used for many years
- Relatively easy to find sysadmins with experience

Puppet

- New generation of configuration management system
- Extensible, declarative language
- Understands dependencies (huge benefit)
- Better reporting than Cfengine
- Auto generation of documentation (think Javadoc)
- FEF avg is 325 actions per node every Puppet run

Fermilab

FEF Puppet Usage

- Management of all external mounts
- Kerberos files -- keytab files, .k5login, etc
- Package management (RPM sets grouped by cluster)
- NIC bonding configs
- Group quotas
- Grid host certs
- FEF_backup

Fermilab

Cfengine vs. Puppet (High Level)

	Native File Editing	Dependency Management	Commercial Support	Dependency Graphs	Scalability
Cfengine	✔		✔		✔
Puppet		✔	✔	✔	✔

Fermilab

Puppet Add User Example

```
User { managehome    => true,
       ensure        => present,
       gid           => users,
       shell         => "/bin/bash",
}

user { “mark”:
    uid => 1000,
}

user { “fred”:
    uid  => 1001,
}

user { “jane”:
    uid => 1002,
}
```

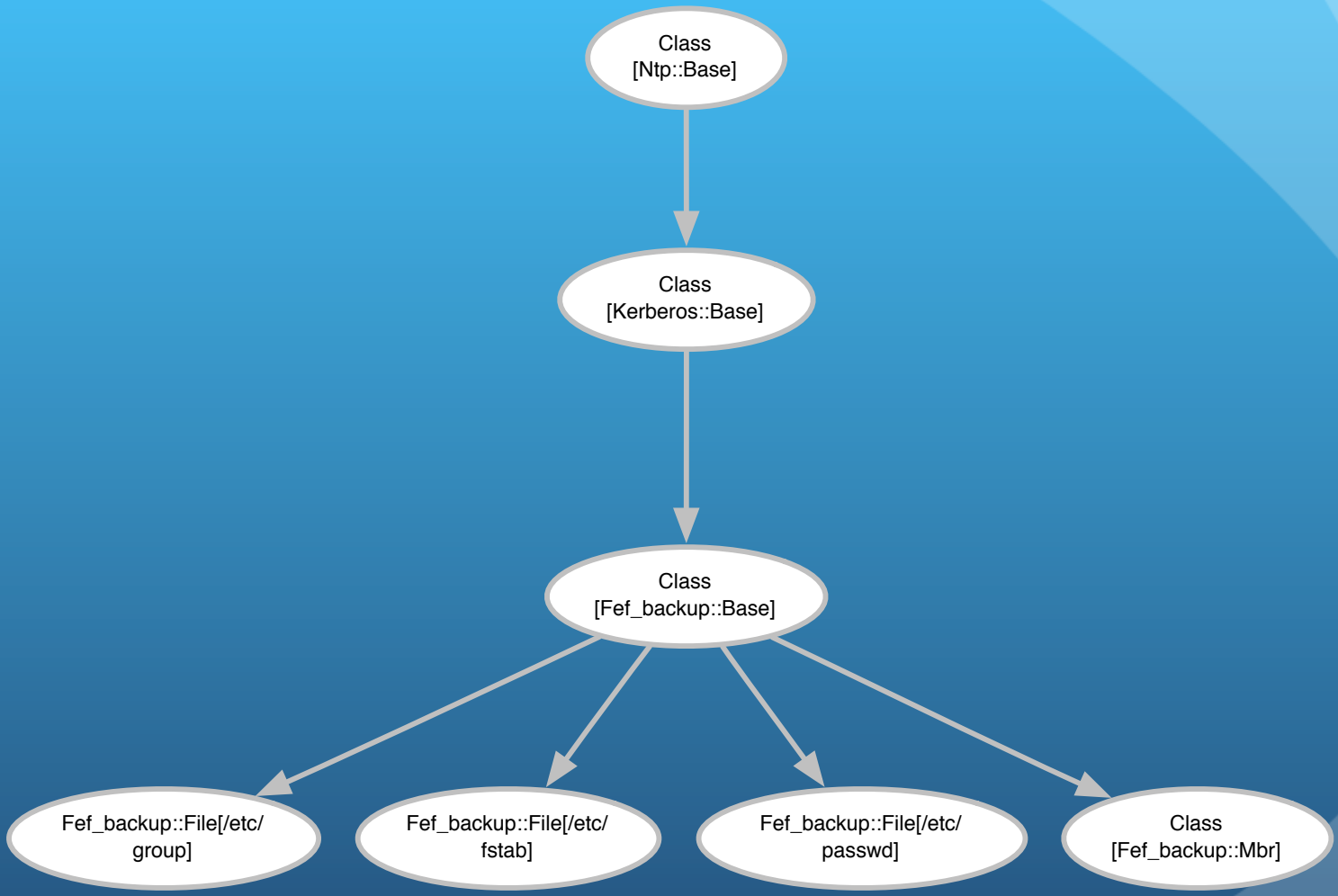
Fermilab

Cfengine Add User Example

```
"pw[mark]" string => "mark:x:1000:100:Mark Burgess:/home/
mark:/bin/bash”;

"pw[fred]" string => "fred:x:1001:100:Right Said:/home/
fred:/bin/bash";

"pw[jane]" string => "jane:x:1002:100:Jane Doe:/home/
jane:/bin/bash";

"users" slist => getindices("pw");

files:

  "/etc/passwd"

     edit_line => append_users_starting("addusers.pw");

  "/etc/group"

       edit_line => append_user_field("root","4","@
(addusers.users)");

  "/home/$(users)/."

     create => "true",

      perms => mog("755","$(users)","users");
```

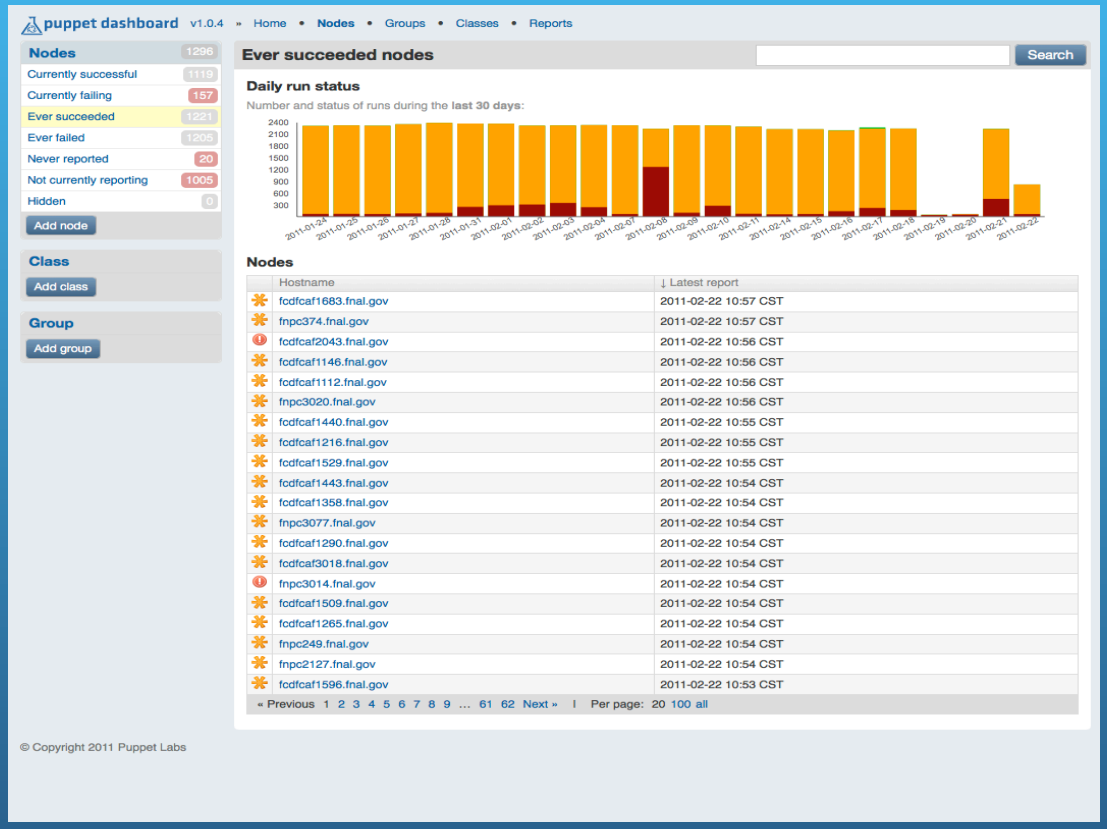
Fermilab

Example Puppet Dependency Graph



Fermilab

Puppet Dashboard



Puppet dashboard is a web interface for quickly viewing puppet run status and the state of individual system configurations.

Fermilab

Bcfg2

- BCFG2 is an xml-based configuration management system
- Developed by Argonne National Lab
- Being used by CMS Tier 1 at Fermilab for 2 years to manage a limited number of configuration items. Not RPMS
- Developers are very responsive; provide support via mailing list and IRC
- Complex file manipulation can be tricky
- Does simple pre/post dependencies

Fermilab

Bcfg2 Reporting System

Node	State	Good	Bad	Modified	Extra	Last Run	Server
cms-sleepgw.fnal.gov	clean	7	0	0	798	2011-03-03 06:55	cmssrv01.fnal.gov
cms-xen20.fnal.gov	dirty	4	1	2	821	2011-02-28 06:30	cmssrv01.fnal.gov
cms-xen39.fnal.gov	clean	9	0	0	664	2011-03-02 08:35	cmssrv01.fnal.gov
cms-xen40.fnal.gov	clean	9	0	0	638	2011-03-03 05:03	cmssrv01.fnal.gov
cms-xen41.fnal.gov	clean	9	0	0	552	2011-03-03 05:06	cmssrv01.fnal.gov
cmscbn.fnal.gov	clean	7	0	0	803	2010-12-14 04:23	cmssrv01.fnal.gov
cmscode01.fnal.gov	clean	7	0	0	871	2011-03-03 04:36	cmssrv01.fnal.gov
cmscode02.fnal.gov	clean	7	0	0	826	2011-03-03 04:26	cmssrv01.fnal.gov
cmscode03.fnal.gov	clean	7	0	0	824	2011-03-02 05:35	cmssrv01.fnal.gov
cmsdcdr1.fnal.gov	clean	10	0	0	886	2011-03-03 04:02	cmssrv01.fnal.gov
cmsdcdr2.fnal.gov	clean	10	0	0	888	2011-03-02 04:32	cmssrv01.fnal.gov
cmsdcmon1.fnal.gov	clean	12	0	0	898	2011-03-02 06:50	cmssrv01.fnal.gov
cmsfilemover.fnal.gov	clean	7	0	0	957	2011-03-03 07:22	cmssrv01.fnal.gov
cmsfnal01.fnal.gov	clean	11	0	0	936	2011-03-03 06:48	cmssrv01.fnal.gov
cmsfnal02.fnal.gov	clean	11	0	0	936	2011-03-03 05:06	cmssrv01.fnal.gov
cmsfnal03.fnal.gov	clean	11	0	0	977	2011-03-03 06:23	cmssrv01.fnal.gov
cmsfts1.fnal.gov	dirty	2	5	0	1023	2010-12-14 05:11	cmssrv01.fnal.gov
cmsfts2.fnal.gov	clean	7	0	0	1018	2010-12-11 04:32	cmssrv01.fnal.gov
cmsfts3.fnal.gov	clean	7	0	0	792	2011-03-01 04:57	cmssrv01.fnal.gov
cmslmds.fnal.gov	clean	7	0	0	872	2011-03-03 06:00	cmssrv01.fnal.gov

Node status and last run times are viewable from the Bcfg2 web interface.

Fermilab

Monitoring

Tools to help monitor system state and performance.

Fermilab

Monitoring

In use at Fermilab:

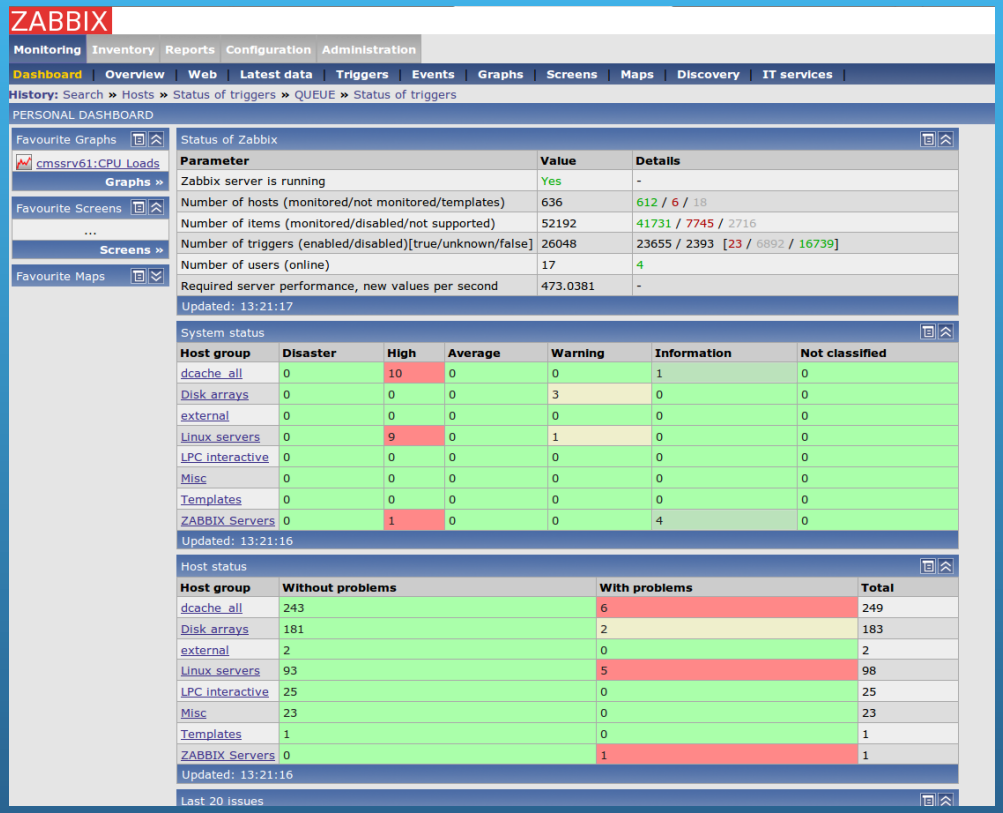
- Zabbix
- Nagios
- Ganglia
- Many custom solutions using MRTG, RRDtool, etc

Fermilab

Zabbix

- Being used by CMS Tier 1 to monitor approx 2.4K nodes; performs 100K checks.
- Does status and performance monitoring.
- Relatively new compared to Nagios.
- Most configuration is done via the web interface.
- Easy to add custom checks and alerts.

Zabbix Dashboard



Zabbix provides a polished web interface that displays finely grained status and performance information.

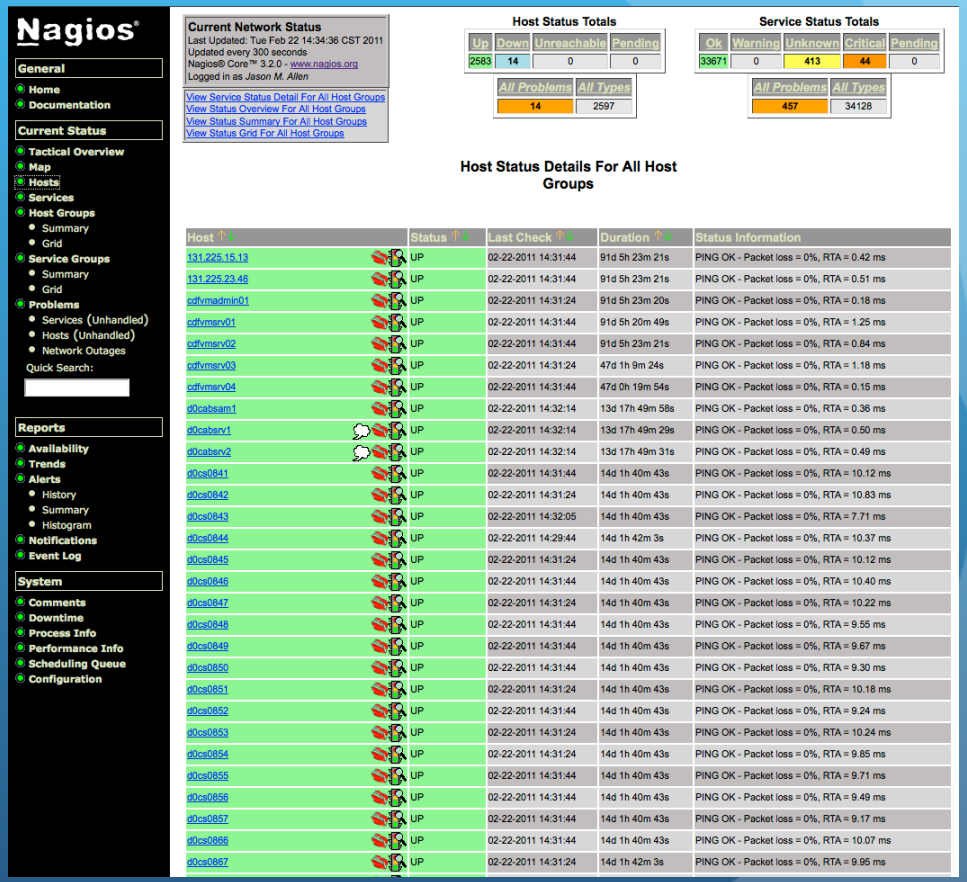
Fermilab

Nagios

- Used by FEF; 3.5K nodes, approx 30K checks on one server
- Around for many years.
- Create a new check by dropping shell script on node (check_mk plugin)
- Nagios support built-in to Puppet.
- Web interface can be slow and feels dated.

Nagios Web Interface

The Nagios web interface is functional but feels dated. Performance is an issue when monitoring many hosts.



Fermilab

Misc Cluster Management Tools and Techniques

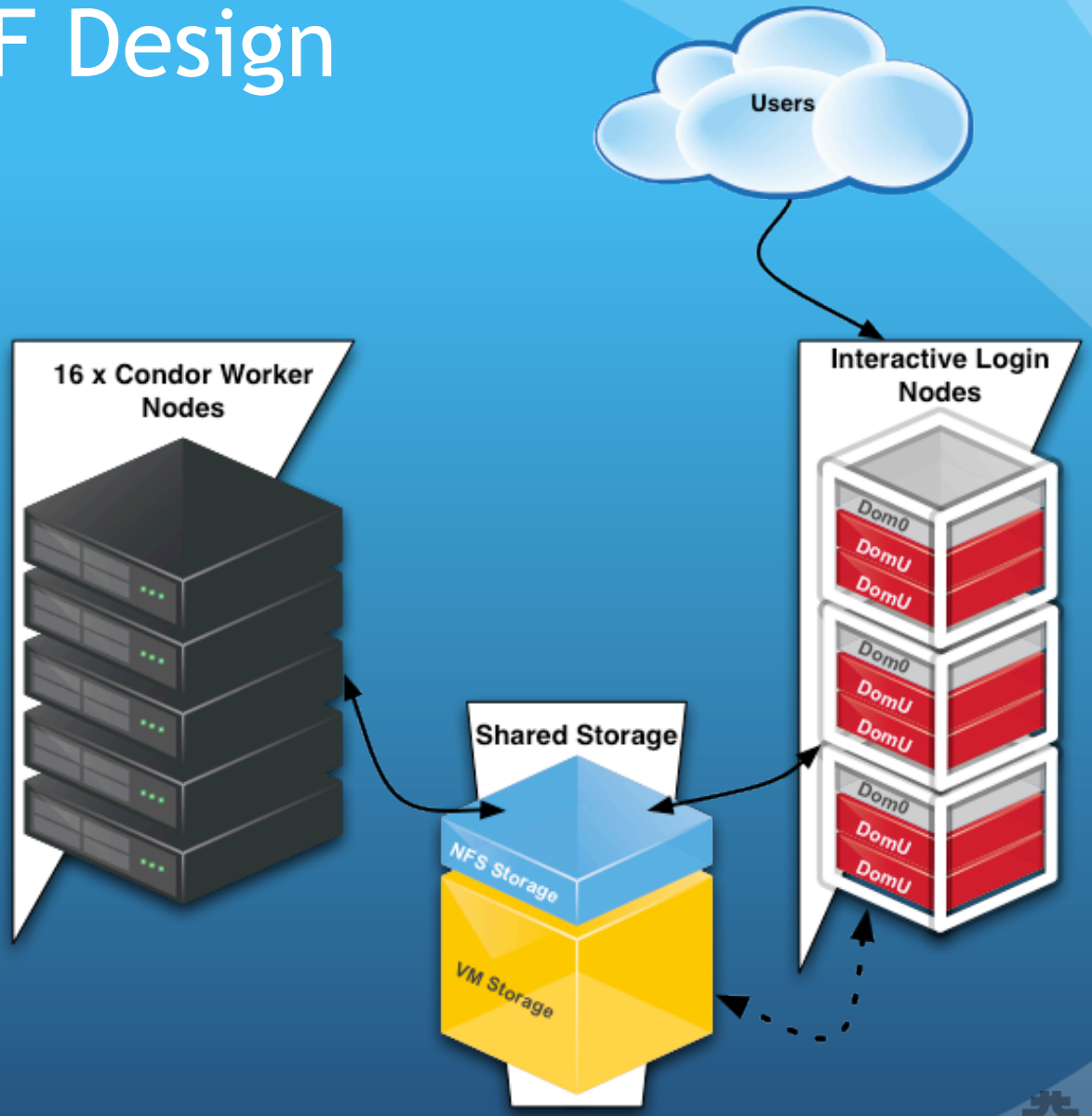
Other things you might find interesting

Fermilab

GPCF

- General Purpose Computing Facility
- Oracle VM cluster + batch cluster
- Quick provisioning for smaller or new experiments.
- Computing resources in days instead of weeks/months.
- Small batch cluster serves as training wheels for the Grid.

GPCF Design

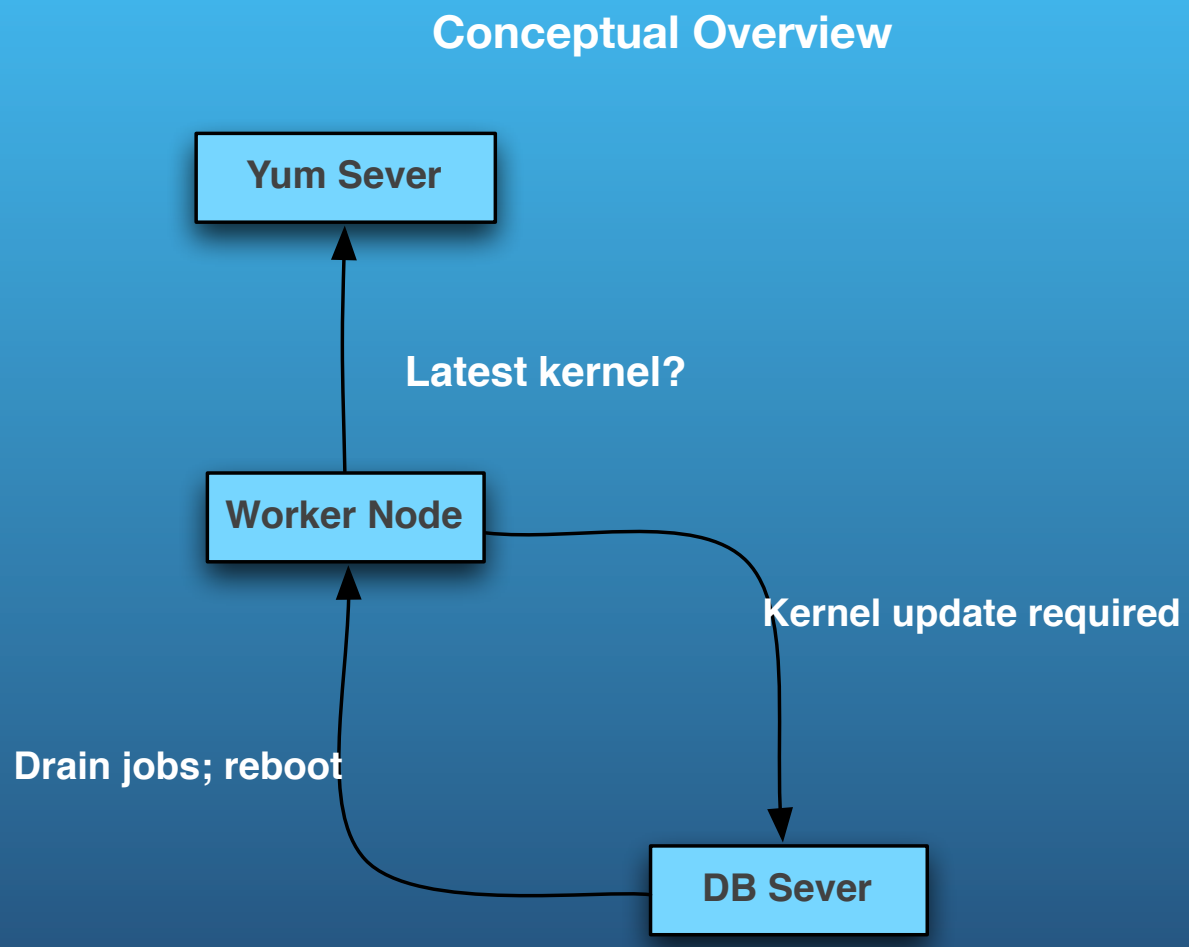


Fermilab

CDF Grid Rolling Upgrades

- Fermi security policy states that software patches to be applied within 60 days of release.
- Systems must be every rebooted every 2-3 months to load new Linux kernel.
- Tool automatically drains and reboots 50 nodes at a time.
- 1K node cluster is updated in appox 2 weeks.

Rolling Upgrades



Fermilab

Resources

Comparison of Puppet/Cfengine/Bcfg2
https://cd-docdb.fnal.gov:440/cgi-bin/ShowDocument?docid=3967

Evaluation and feature comparison of the Nagios and Zabbix monitoring systems

http://cd-docdb.fnal.gov/cgi-bin/ShowDocument?docid=3277

Download Scientific Linux 6

http://scientificlinux.org/

Fermilab