Overview of cluster management tools

Open Science Grid

THE UNIVERSITY OF CHICAGO

Marco Mambelli – marco@hep.uchicago.edu

August 9 2011

OSG Summer Workshop

TTU - Lubbock, TX

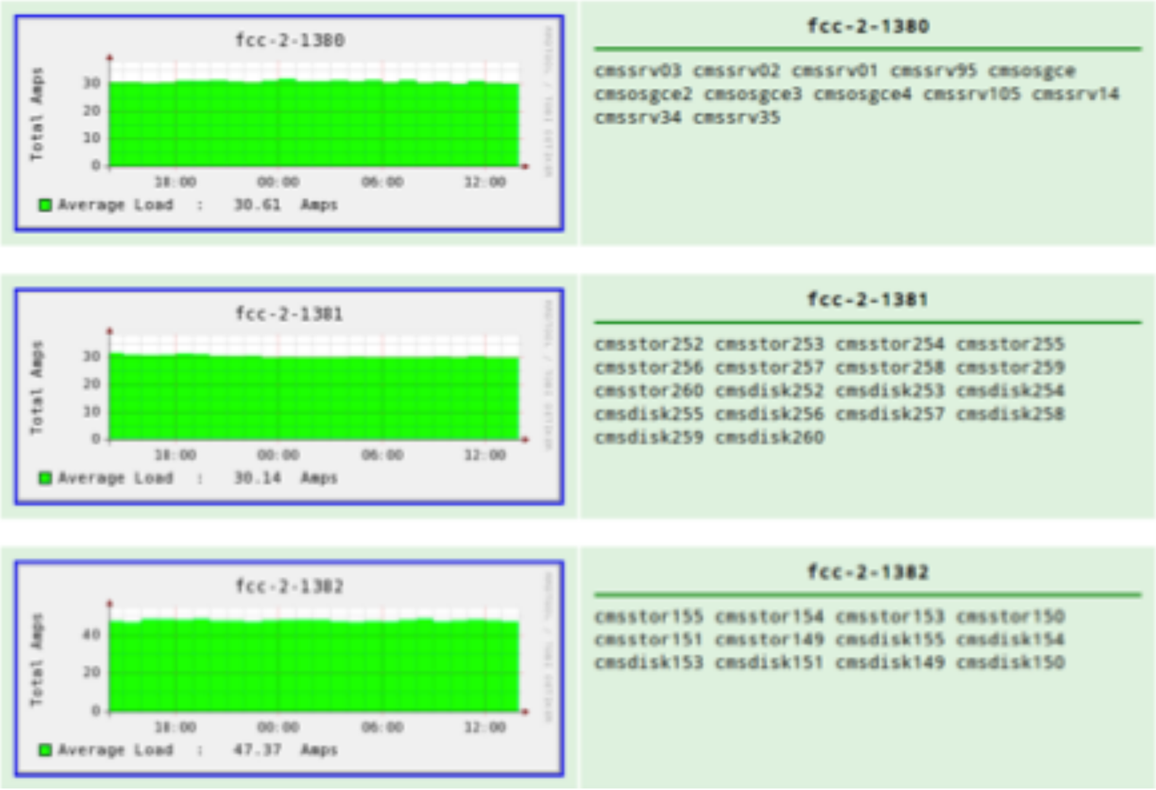
Cluster Management Overview

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

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

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

Provisioning Tools

- Provisioning Tools
  - Preparing a system for use; OS installation and initial configuration
- Tools
  - PXE/Kickstart
  - Rocks
  - Cobbler
  - Perceus

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 / nodecombinations
- Very flexible
- Kickstart files are created dynamically based on selections from the web GUI
- Planning to eval Cobbler later this year.

Rocks Clusters

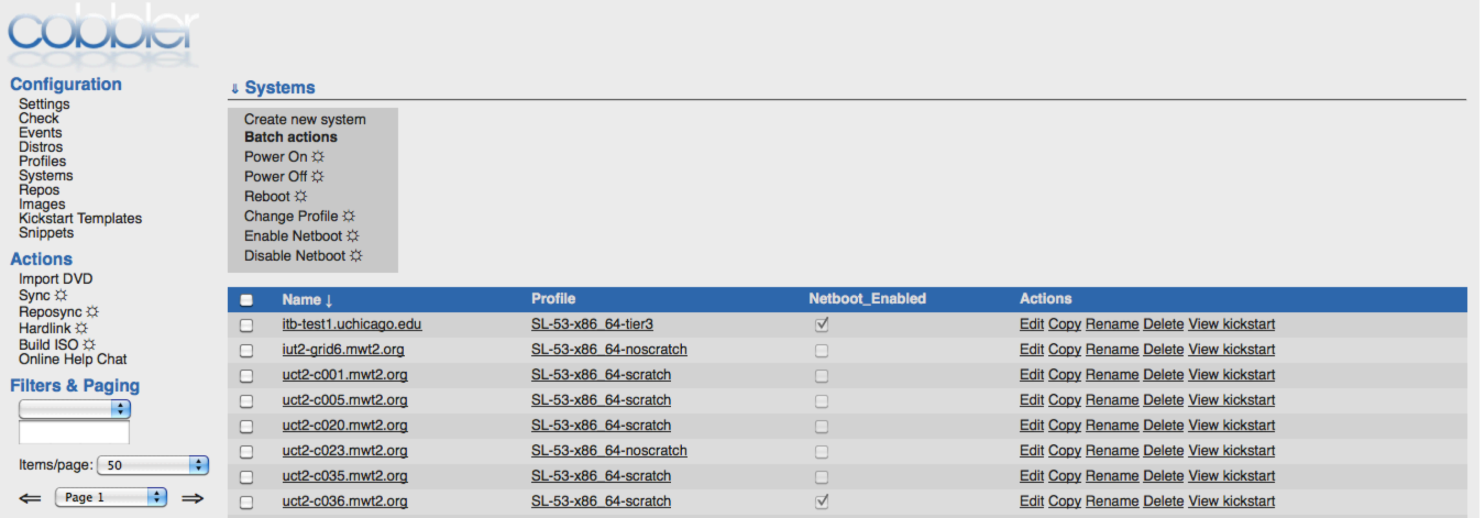
- Open source Linux distribution based on CentOS
- Created in 2000
- Created for easy deployment of large clusters
- Used by CMS Tier 1 at Fermilab for 2400 machines
- CMS can install ~500 systems in 1 hour
- Only one OS version per Rocks server may be a deal breaker for some

Cobbler

- Cobbler is infrastructure to provision a node
- Cobbler is RedHat specific, uses kickstart
- Cobbler holds OS, profiles and system data
- A new system requires a profile, a MAC address and a name, nothing more.
- Provisions new system via PXE boot, via settings for individual system.
- CLI and Web GUI

Cobler at MWT2

- Cobbler is implemented through multiple services on a single server. It acts as a TFTP server for PXE booting, controls the repositories used for install, and provides DHCP/DNS support as well.



□	Name ↓	Profile	Netboot_Enabled	Actions
□	itb-test1.uchicago.edu	SL-53-x86_64-tier3	☑	Edit Copy Rename Delete View kickstart
□	iut2-grid6.mwt2.org	SL-53-x86_64-noscratch	□	Edit Copy Rename Delete View kickstart
□	uct2-c001.mwt2.org	SL-53-x86_64-scratch	□	Edit Copy Rename Delete View kickstart
□	uct2-c005.mwt2.org	SL-53-x86_64-scratch	□	Edit Copy Rename Delete View kickstart
□	uct2-c020.mwt2.org	SL-53-x86_64-scratch	□	Edit Copy Rename Delete View kickstart
□	uct2-c023.mwt2.org	SL-53-x86_64-noscratch	□	Edit Copy Rename Delete View kickstart
□	uct2-c035.mwt2.org	SL-53-x86_64-scratch	□	Edit Copy Rename Delete View kickstart
□	uct2-c036.mwt2.org	SL-53-x86_64-scratch	☑	Edit Copy Rename Delete View kickstart

Configuration and Package Management

- Tools that help manage system configuration (files, dirs, permissions, etc.) and software packages
- Popular open source solutions
  - Cfengine
  - Puppet
  - Bcfg2
  - Warewulf

Cfengine

- First version released in 1993
- Written in C
- Fairly easy to understand syntax
- Relatively easy to find sysadmins with experience
- FEF and UWM used for many years

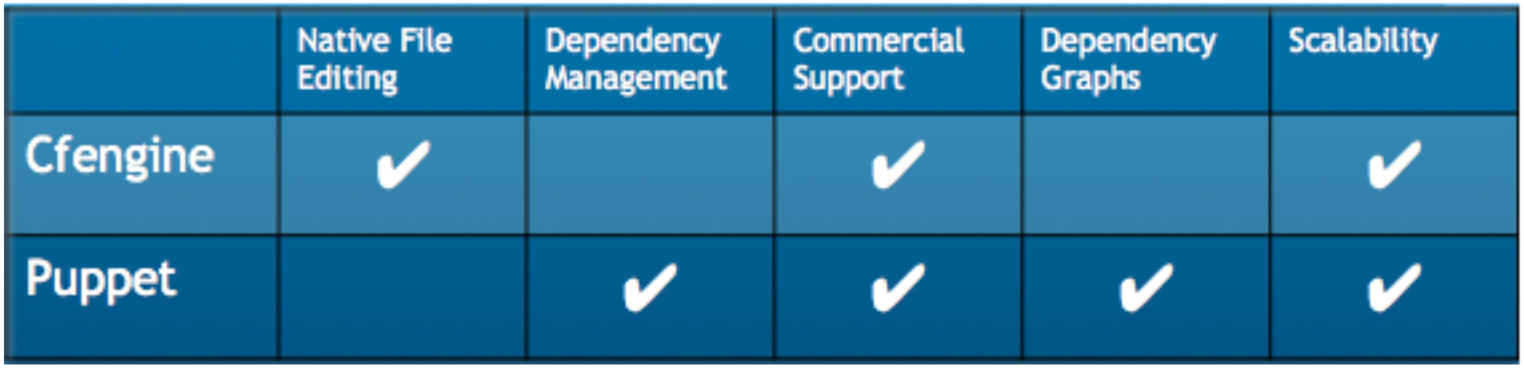
Puppet

- New generation of configuration management system
- Extensible, declarative language
- Understands dependencies (huge benefit)
- Better reporting than Cfengine
- Auto generation of documentation (think Javadoc)

FEF Puppet Usage

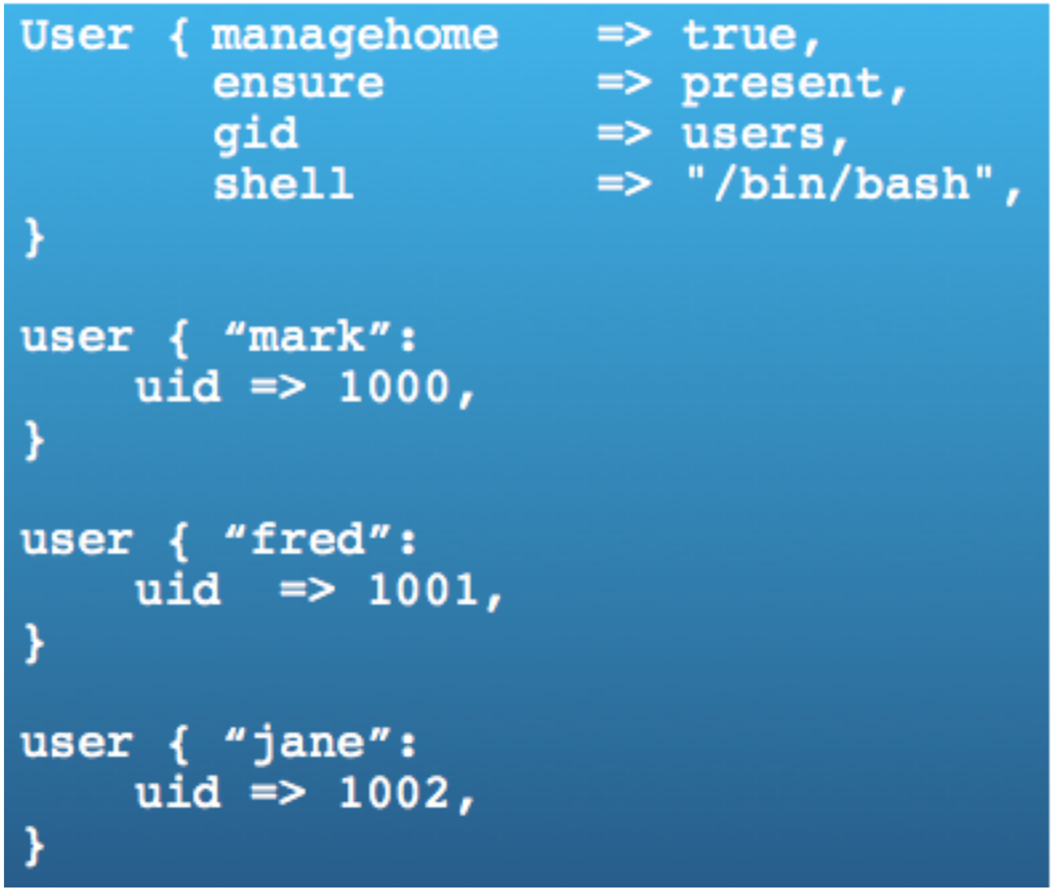
- Management of all external mounts
- Kerberos files -- keytab files, .k5login, etc
- Package management (RPM sets grouped by cluster)
- NIC bonding configs
- Group quotas
- Grid host certs
- FEF_backup
- FEF avg is 325 actions per node every Puppet run

Cfengine vs. Puppet (High Level)



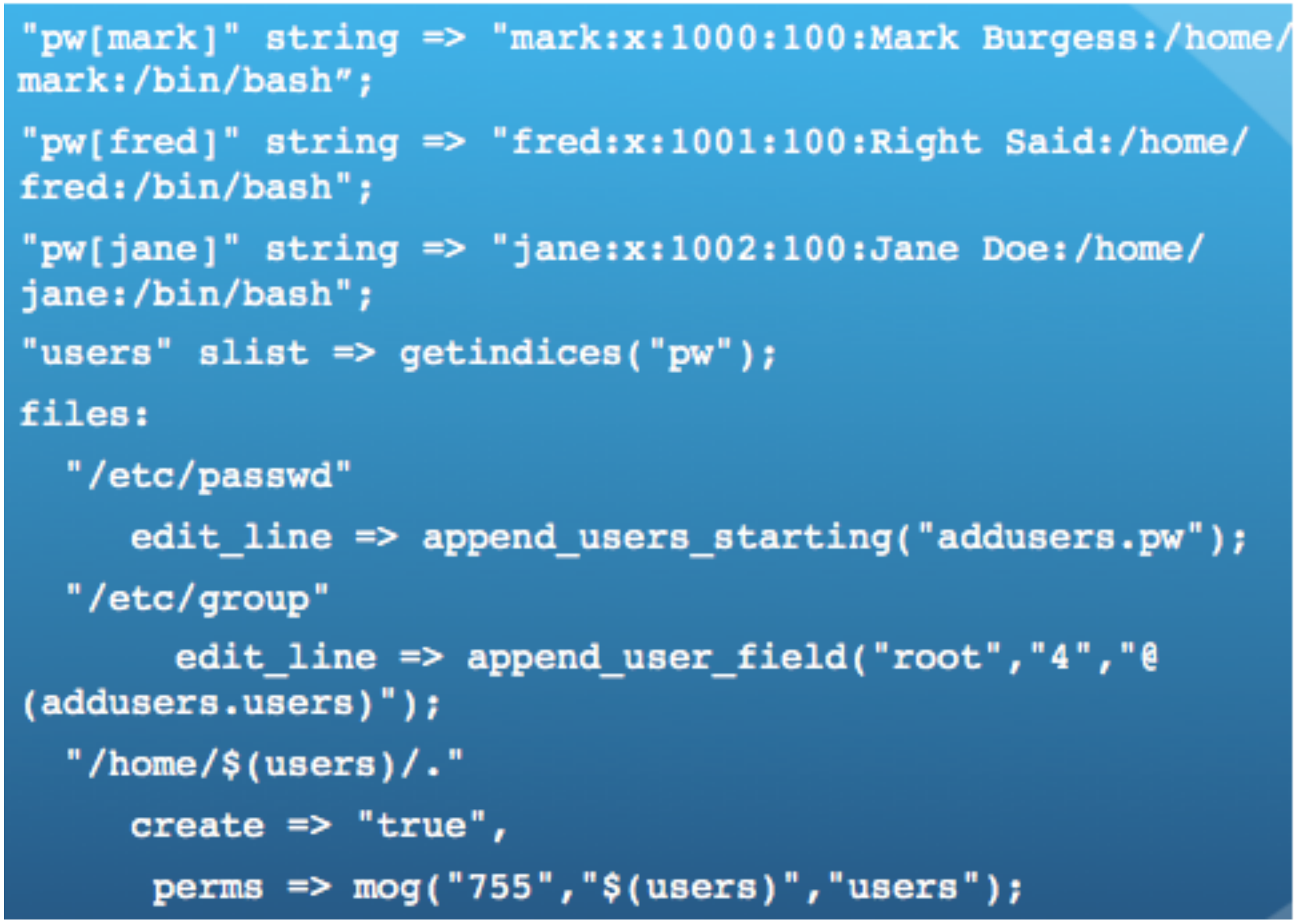
	Native File Editing	Dependency Management	Commercial Support	Dependency Graphs	Scalability
Cfengine	✓		✓		✓
Puppet		✓	✓	✓	✓

Puppet Add User Example



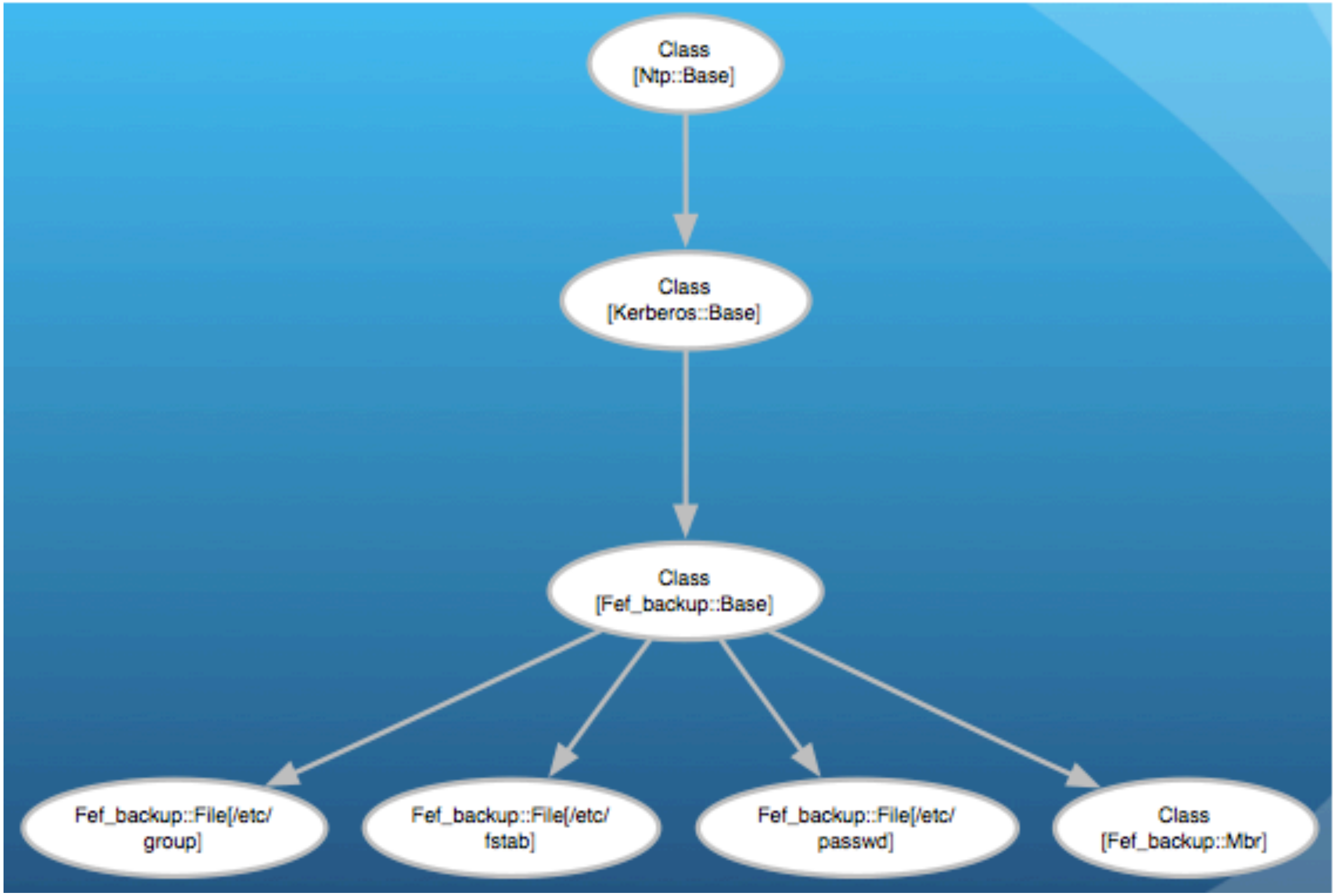
```
User { managehome    => true,
       ensure        => present,
       gid           => users,
       shell         => "/bin/bash",
}

user { “mark”:
    uid => 1000,
}

user { “fred”:
    uid  => 1001,
}

user { “jane”:
    uid => 1002,
}
```

Cfengine Add User Example

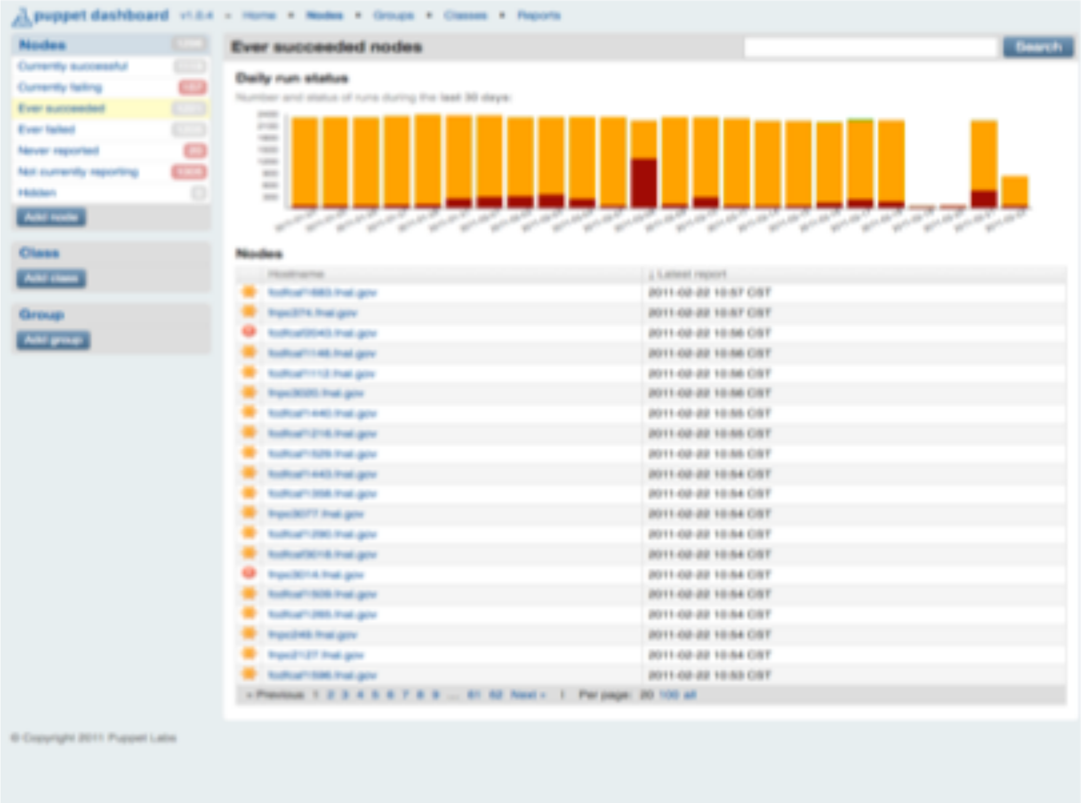


```
"pw[mark]" string => "mark:x:1000:100:Mark Burgess:/home/
mark:/bin/bash";
"pw[fred]" string => "fred:x:1001:100:Right Said:/home/
fred:/bin/bash";
"pw[jane]" string => "jane:x:1002:100:Jane Doe:/home/
jane:/bin/bash";
"users" slist => getindices("pw");
files:
  "/etc/passwd"
     edit_line => append_users_starting("addusers.pw");
  "/etc/group"
       edit_line => append_user_field("root","4","@
(addusers.users)");
  "/home/$(users)/."
     create => "true",
      perms => mog("755","$(users)","users");
```

Example Puppet Dependency Graph


Class [Ntp::Base]
Class [Kerberos::Base]
Class [Fef_backup::Base]
Fef_backup::File[/etc/ group]
Fef_backup::File[/etc/ fstab]
Fef_backup::File[/etc/ passwd]
Class [Fef_backup::Mbr]

Puppet Dashboard



Ever succeeded nodes

Nodes

Hostname	↓ Latest report
[illegible]	2011-02-22 10:57 CST
[illegible]	2011-02-22 10:57 CST
[illegible]	2011-02-22 10:56 CST
[illegible]	2011-02-22 10:56 CST
[illegible]	2011-02-22 10:56 CST
[illegible]	2011-02-22 10:56 CST
[illegible]	2011-02-22 10:55 CST
[illegible]	2011-02-22 10:55 CST
[illegible]	2011-02-22 10:55 CST
[illegible]	2011-02-22 10:54 CST
[illegible]	2011-02-22 10:54 CST
[illegible]	2011-02-22 10:54 CST
[illegible]	2011-02-22 10:54 CST
[illegible]	2011-02-22 10:54 CST
[illegible]	2011-02-22 10:54 CST
[illegible]	2011-02-22 10:54 CST
[illegible]	2011-02-22 10:54 CST
[illegible]	2011-02-22 10:54 CST
[illegible]	2011-02-22 10:54 CST
[illegible]	2011-02-22 10:53 CST

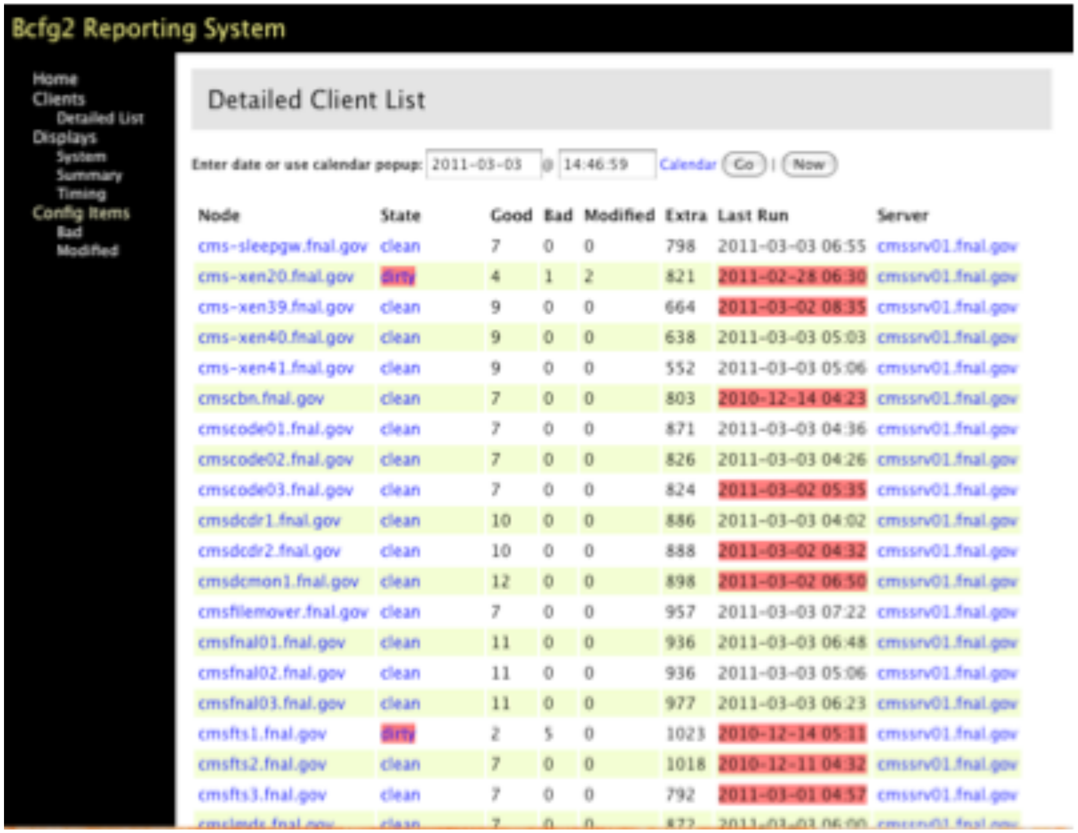
« Previous 1 2 3 4 5 6 7 8 9 … 61 62 Next » | Per page: 20 100 all

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

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

Bcfg2 Reporting System



Detailed Client List

Enter date or use calendar popup: 2011-03-03 @ 14:46:59

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
cmslmdx.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

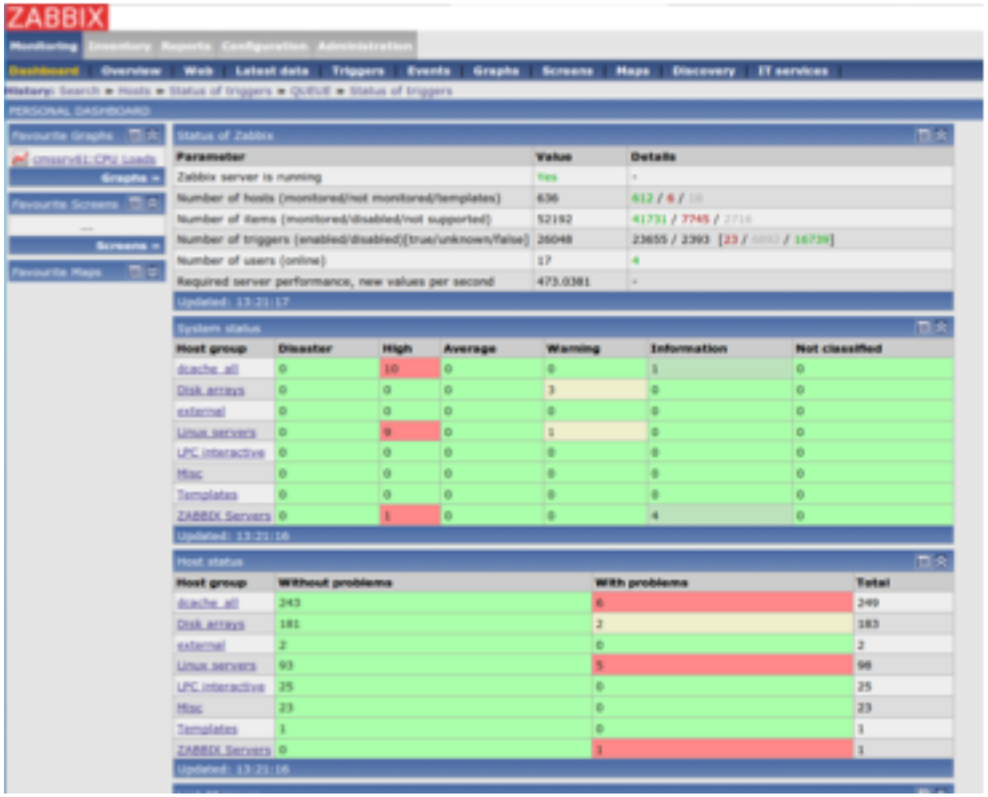
Monitoring

- Tools to help monitor system state and performance
- In use at Fermilab:
  - Zabbix
  - Nagios
  - Ganglia
  - Many custom solutions using MRTG, RRDtool, etc

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



Status of Zabbix

Parameter	Value	Details
Zabbix server is running	Yes	-
Number of hosts (monitored/not monitored/templates)	636	612 / 6 / 18
Number of items (monitored/disabled/not supported)	52192	41731 / 7745 / 2716
Number of triggers (enabled/disabled)[true/unknown/false]	26048	23655 / 2393 [23 / [illegible] / 16739]
Number of users (online)	17	4
Required server performance, new values per second	473.0381	-

Updated: 13:21:17

System status

Host group	Disaster	High	Average	Warning	Information	Not classified
dcache_all	0	10	0	0	1	0
Disk arrays	0	0	0	3	0	0
external	0	0	0	0	0	0
Linux servers	0	9	0	1	0	0
LPC interactive	0	0	0	0	0	0
Misc	0	0	0	0	0	0
Templates	0	0	0	0	0	0
ZABBIX Servers	0	1	0	0	4	0

Updated: 13:21:16

Host status

Host group	Without problems	With problems	Total
dcache_all	243	6	249
Disk arrays	181	2	183
external	2	0	2
Linux servers	93	5	98
LPC interactive	25	0	25
Misc	23	0	23
Templates	1	0	1
ZABBIX Servers	0	1	1

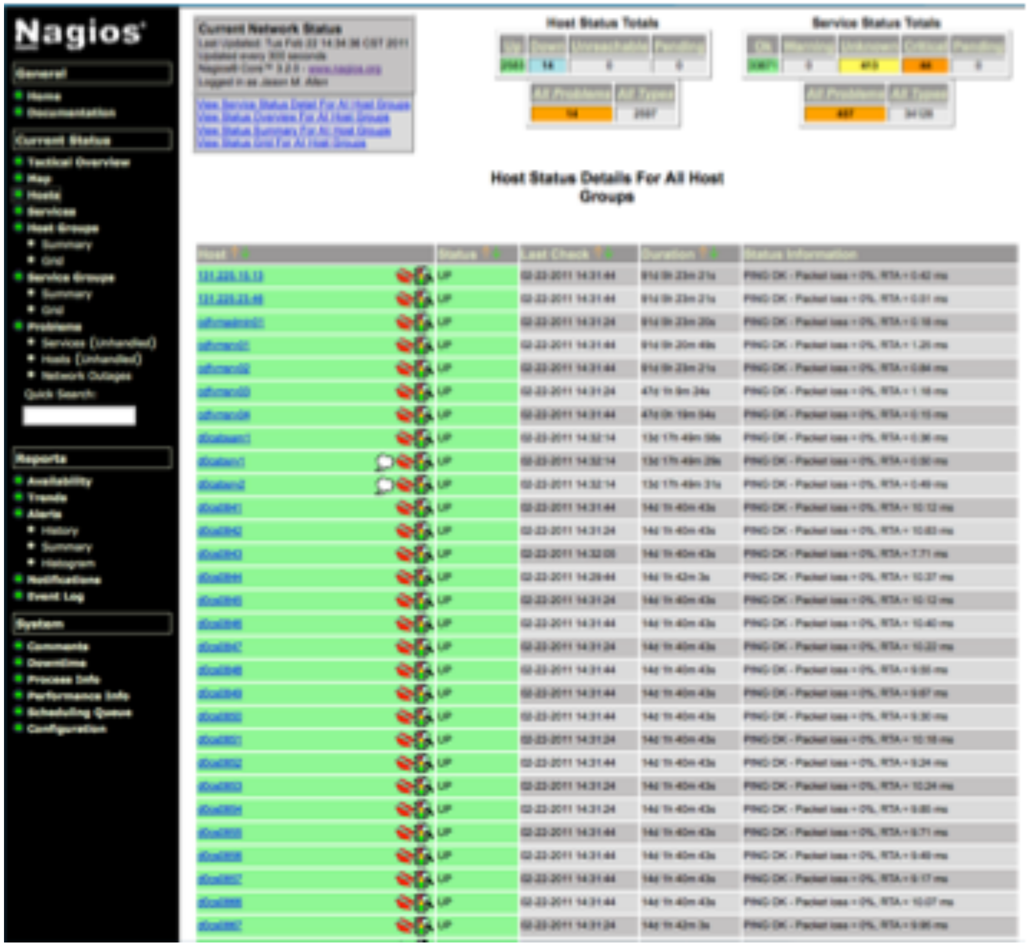
Updated: 13:21:16

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

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

OSG Survey

- The most used cluster management tool is Rocks, sometime customized
- Followed by Puppets, Cobbler and Cfengine.
- Users are happy with the performance of the tools they use, especially Rocks and Puppets.
- The difficulty of the first installation is average (sometime long or with some guesswork) to easy (works out of the box); same for the updates.

OSG Survey (cont)

- The operation is automatic or requires simple documented tasks.
- Rocks seem the easiest to operate
- Puppet is the easiest to install/update
- Available documentation is good for Rocks, good to average (there could be more or sometime is confusing) for the others.
- There are some long time users of Rocks and Cfengine while Puppet gained popularity in recent times.

- 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
- Survey about the setup of clusters in OSG
  - Available from OSG DOCDB

Credits

- Thank you to Jason Allen, Head of Fermilab Experiments Facilities (FEF) Department in the Computing Division – this talk is based heavily on his Cluster Management talk at the 2011 OSG all hands meeting

?

!