University of Sussex

Sussex Site Report

Matt Rásó-Barnett
Linux Systems Administrator - School of Mathematical and Physical Sciences

Sussex HPC

- Started out in 2010 as collaboration between Physics Department and central IT services
- Now a number of departments have bought compute resources, although Physics is still the largest contributor.
- Have a very mixed user base with a number of heavy MPI users, and have a corresponding mix of hardware to suit the groups - very much a ‘patchwork’ cluster
- We run a small Grid Tier-2 site that supports ATLAS and SNO+. Integrated as part of the larger Tier-3 cluster, utilising the same batch system and storage but Grid jobs run on dedicated compute nodes

Server Room

- Housed in University's main data centre, built in late 2010.
- Centrally managed / monitored. Power, environment, exterior networking all taken care of.
- HPC occupies 5 racks with plans to hopefully take a 6th with new ITS investment at the end of 2015

Hardware

- 109 compute nodes, 3280 cores, nearly all Dell.
  - ~50% Dell C6145
    2x 64-core AMD Opteron 6378
  - Rest are mix of 1U 12-core Intel R410, 2U C6100 and 2U 4x 16-core C6220
  - Tend to have 4GB/per core with a few small islands of 2GB/core
- 4 GPU nodes with Tesla K40s and K60s, mostly used by Informatics

Network

- All storage (Lustre + NFS) and compute on QDR (40Gb) Infiniband (Qlogic/Intel)
- Until recently (Dec 2014) has worked well without troubles using in-kernel IB drivers
- RedHat 6.6 kernels included some recent patches which have broken IPoIB. Patches have also been backported into 6.5 kernels starting with -431.40.1.el6
- Currently running a kernel produced by NSC in Sweden based off -431.40.2.el6 which removes the bad IB patches and also backports some recent security fixes.
- Surprisingly little public discussion of the problem!?
- Following LKML discussion and think that 4.1 will include the fix, but unsure when RedHat will backport it.

Storage

- Persistent data (Home directories, research outputs) stored on multiple NFS servers often bought by departments. Well over 200TB now with more demanded
  - Backed up to tape by IT services every night
- Not scaling very well as usage of these grows.
  - Constantly monitoring for job I/O to these areas as can quickly saturate disks and ruin interactive sessions
  - Backups becoming prohibitively long
- Looking for alternatives: Ceph(FS)?

Lustre

- Scratch space is stored on Lustre
- 300TB currently on Lustre 1.8.9
  210TB currently on (pre-prod) Lustre 2.5.3
- **1.8.9 System -- Commissioned by Alces Software in 2010**
  - **Metadata server (MDS):** 2x Dell R510 + 8-disk MD3220 -- with automatic failover
  - **8x Object Servers (OSS):** each a 12-disk R510 + 1 or 2x 12-disk MD1200 -- mixture of 1TB to 3TB disks in RAID6
- **2.5.3 System**
  - **MDS:** 2x Dell R430 + 8-disk MD3220 -- haven't quite got failover working on this yet
  - **2x OSS:** each a 12-disk R730xd + 3x MD1200 with 3TB disks in RAID6
- Will aim to merge the 1.8.9 system into the 2.5.3

Lustre 1.8.9 -> 2.5.3 Migration

- Looking to move all storage onto 2.5.3 ~June/July
- Can mount a 2.5.3 Lustre filesystem using the 1.8.9 client software running on the compute nodes, so can copy between filesystems without requiring downtime
- Working on an opt-in migration service for users, so burden falls on them to request what they want to move over. Hopefully won't be copying all ~270TB currently in use
- Currently testing a few parallel copy tools:
  - pcp (https://github.com/wtsi-ssg/pcp)
  - dcp (https://github.com/hpc/dcp)
  - homegrown parallel rsync scripts
- Looking into Robinhood policy engine (https://github.com/cea-hpc/robinhood/wiki) as a reporting / analysing tool for the new filesystem

Lustre 2.X and StoRM

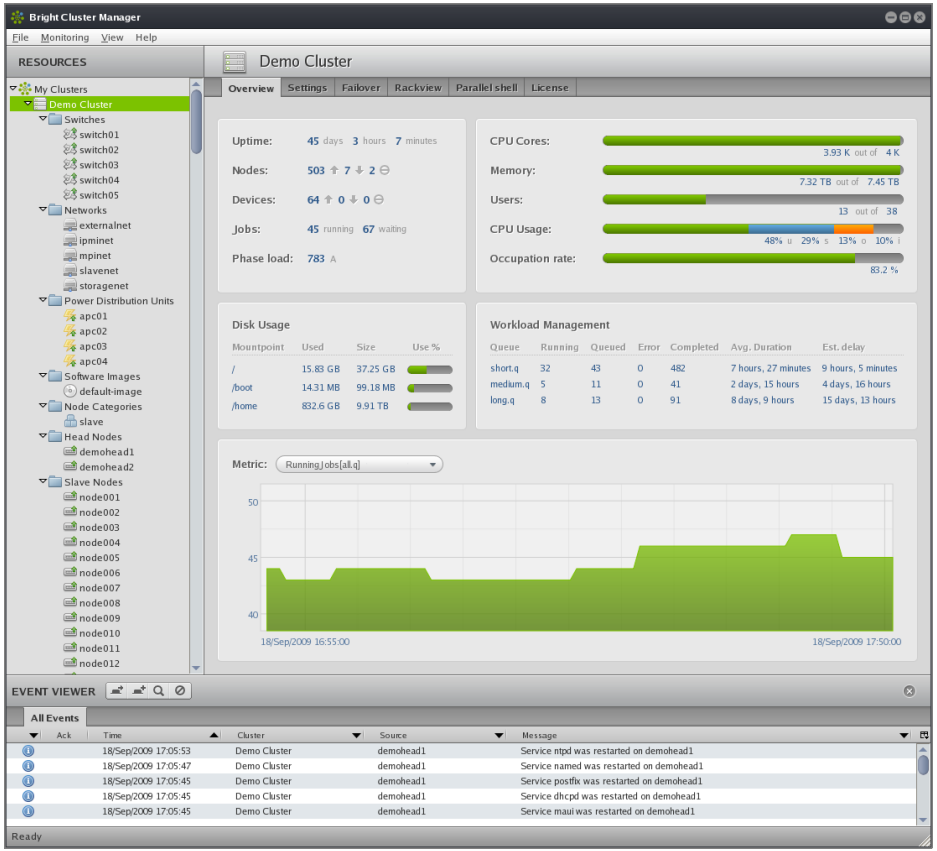
- Lustre bug **[LU-1482]** currently prevents StoRM from working with Lustre 2.X. No fix upstream yet although hoped for 2.8.0 release (~Q4 2015/Q1 2016?)
- We've patched 2.5.3 with (non-public) community-supplied fix for the bug -- it does work, but haven't tested with StoRM yet.
- Plan to build test StoRM server to try it out in next couple of weeks.
- Will keep Lustre 1.8.9 system running until we have some confidence in the patch as a Grid-only filesystem.
- Worried about long-term future of Lustre as Intel have recently announced they will stop producing maintenance releases. Even if fix makes it upstream, as it stands will have to backport and manage patches ourselves.

Univa Grid Engine

- Currently run a single batch system UGE 8.2.1 across the whole site (Tier 3 + Tier 2), so in theory would be a great candidate for an integrated site
- In practice resources are still partitioned into smaller queues, depending on which department bought the nodes
- Constantly working to reduce this but a lot of work still needs to be done:
  - Cluster-wide share-tree
  - Improving resource isolation (cpu,memory cgroups supported since 8.1.7, default resource requests via server side-JSV)
  - Better system for accounting of usage
- Tier 3 usage is very bursty, with frequent periods of little activity which would be perfect for backfilling Grid usage

Config Management / Puppet

- Using Bright Cluster Manager 6 for management of compute/login nodes
  - Image-based deployment
  - All-in-one, integrated tool, with monitoring, IPMI, dhcp etc
- Not a proper configuration management tool, difficult to manage changes to images amongst a distributed admin team
- Slowly moving state out of the Bright images and into Puppet
- Puppet+PXE also for all of Lustre, Storage, Grid and other services nodes



Puppet + R10K

- R10K is a tool to deploy modules into Puppet environments dynamically
- Solves two key problems for me:
  - Using, modifying and ***contributing back*** to 3rd party Puppet modules
  - Testing code changes before deploying to production

- Allows you to become very versatile. Built around using the full power of popular git development workflows.

- Requires Puppet 3.6+ for the ‘directory environments’ feature
- Requires Git currently (not sure of status of support for other DVCSs)

R10K workflow (1)

Control Repository:

- environments/
  - hiera/…
  - manifests/…
  - site/profile/…
    - /role/…
  - **Puppetfile** ---------->

All modules in their own Git repo:

- modules/apache/…
- modules/postgres/…

…

```
  :tag => '1.0.0'
mod 'permute',
  :git => 'https://github.com/xaque208/puppet-permute',
  :tag => '0.0.4'
mod 'group',
  :git => 'https://github.com/pdxcat/puppet-module-group',
  :branch => 'master'
mod 'puppetboard',
  :git => 'https://github.com/puppet-community/puppetboard',
  :branch => 'master'
mo 'python',
  :git => 'https://github.com/stankevich/puppet-python',
  :tag => '1.7.16'
mod 'vcsrepo',
  :git => 'https://github.com/puppetlabs/puppetlabs-vcsrepo',
  :tag => '1.2.0'
mod 'xinetd',
  :git => 'https://github.com/puppetlabs/puppetlabs-xinetd',
  :tag => '1.3.1'
mod 'tftp',
  :git => 'https://github.com/puppetlabs/puppetlabs-tftp',
  :tag => '0.2.2'
mod 'file_concat',
  :git => 'https://github.com/electrical/puppet-lib-file_concat.git',
  :tag => '0.1.0'
mod 'logstash',
  :git => 'https://github.com/elasticsearch/puppet-logstash.git',
  :tag => '0.5.1'
mod 'elasticsearch',
  :git => 'https://github.com/elasticsearch/puppet-elasticsearch.git',
  :tag => '0.4.0'
mod 'logstashforwarder',
  :git => 'https://github.com/elasticsearch/puppet-logstashforwarder.git',
```

R10K workflow (2)

- **R10K: Branch in Git repository == Environment in Puppet**
- Nodes only see code from **their** environment (by default ‘production’ environment, but can be changed at run-time or set in puppet.conf)
- **Example:** want to rewrite how you currently manage something (eg: ssh)
  1. Make branch in module(s) you want to change. Commit changes on branch.
  2. Make branch in **control** repo for this feature (eg: ssh-config-update)
  3. Change Puppetfile so R10K knows to use new branch on updated module(s)
  4. Run R10K on puppet master (this can be automated with a git hook)
  5. Run a test node in this new environment (ssh-config-update) to check your changes.
  6. Iterate until happy…
  7. Merge into production branch/environment once tested

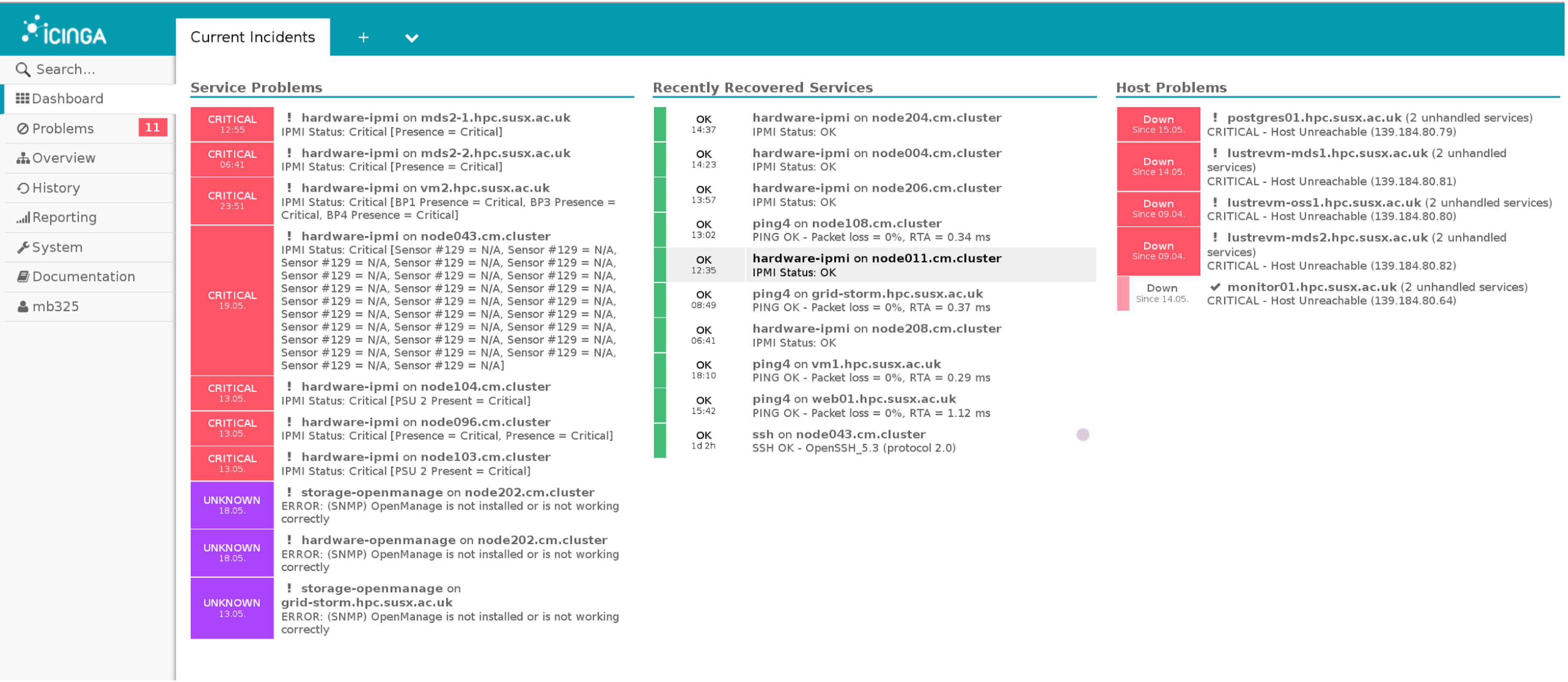
Monitoring

- Starting from scratch essentially, building up around new tools -- been observing the many great GridPP & HEPSYSMAN talks on these topics!
- **Alerting:** Icinga2 + Icinga Web 2
- **Basic Host Metrics:** Ganglia
- **Other Metrics:** Graphite + Grafana 2 for everything else
  Currently feeding it with collectd and Diamond (https://github.com/BrightcoveOS/Diamond) for Grid Engine related stats but very early days. Goal is to get Lustre Server/Client stats feeding in from CollectL.
- **Logs:** ELK -- not much yet beyond deploying it and feeding in syslog. Looking for inspiration, especially for Grid Engine accounting-related uses?

Icinga 2

- Used Nagios Core in previous jobs -- works but clunky interface, clunky to puppetise, clunky to scale…
- Icinga 1 was a fork of Nagios back in 2009
- Icinga 2 is a ground-up rewrite of this aiming at improving:
  - Performance
  - Clusterable/Scalability
  - Modern syntax
  - Remains compatible with Nagios Plugin Architecture. Can still use all the nagios plugins developed over the years, NRPE checks, NSCA etc if you desire.
- So far I'm very pleased with it, using it for primarily hardware, disk checks so far. Feels familiar but nicer to experiences with Nagios.

Icinga Web 2


icinga
Current Incidents
Search...
Dashboard
Problems 11
Overview
History
Reporting
System
Documentation
mb325
Service Problems
CRITICAL 12:55 ! hardware-ipmi on mds2-1.hpc.susx.ac.uk
IPMI Status: Critical [Presence = Critical]
CRITICAL 06:41 ! hardware-ipmi on mds2-2.hpc.susx.ac.uk
IPMI Status: Critical [Presence = Critical]
CRITICAL 23:51 ! hardware-ipmi on vm2.hpc.susx.ac.uk
IPMI Status: Critical [BP1 Presence = Critical, BP3 Presence = Critical, BP4 Presence = Critical]
CRITICAL 19.05. ! hardware-ipmi on node043.cm.cluster
IPMI Status: Critical [Sensor #129 = N/A, Sensor #129 = N/A]
CRITICAL 13.05. ! hardware-ipmi on node104.cm.cluster
IPMI Status: Critical [PSU 2 Present = Critical]
CRITICAL 13.05. ! hardware-ipmi on node096.cm.cluster
IPMI Status: Critical [Presence = Critical, Presence = Critical]
CRITICAL 13.05. ! hardware-ipmi on node103.cm.cluster
IPMI Status: Critical [PSU 2 Present = Critical]
UNKNOWN 18.05. ! storage-openmanage on node202.cm.cluster
ERROR: (SNMP) OpenManage is not installed or is not working correctly
UNKNOWN 18.05. ! hardware-openmanage on node202.cm.cluster
ERROR: (SNMP) OpenManage is not installed or is not working correctly
UNKNOWN 13.05. ! storage-openmanage on grid-storm.hpc.susx.ac.uk
ERROR: (SNMP) OpenManage is not installed or is not working correctly
Recently Recovered Services
OK 14:37 hardware-ipmi on node204.cm.cluster
IPMI Status: OK
OK 14:23 hardware-ipmi on node004.cm.cluster
IPMI Status: OK
OK 13:57 hardware-ipmi on node206.cm.cluster
IPMI Status: OK
OK 13:02 ping4 on node108.cm.cluster
PING OK - Packet loss = 0%, RTA = 0.34 ms
OK 12:35 hardware-ipmi on node011.cm.cluster
IPMI Status: OK
OK 08:49 ping4 on grid-storm.hpc.susx.ac.uk
PING OK - Packet loss = 0%, RTA = 0.37 ms
OK 06:41 hardware-ipmi on node208.cm.cluster
IPMI Status: OK
OK 18:10 ping4 on vm1.hpc.susx.ac.uk
PING OK - Packet loss = 0%, RTA = 0.29 ms
OK 15:42 ping4 on web01.hpc.susx.ac.uk
PING OK - Packet loss = 0%, RTA = 1.12 ms
OK 1d 2h ssh on node043.cm.cluster
SSH OK - OpenSSH_5.3 (protocol 2.0)
Host Problems
Down Since 15.05. ! postgres01.hpc.susx.ac.uk (2 unhandled services)
CRITICAL - Host Unreachable (139.184.80.79)
Down Since 14.05. ! lustrevm-mds1.hpc.susx.ac.uk (2 unhandled services)
CRITICAL - Host Unreachable (139.184.80.81)
Down Since 09.04. ! lustrevm-oss1.hpc.susx.ac.uk (2 unhandled services)
CRITICAL - Host Unreachable (139.184.80.80)
Down Since 09.04. ! lustrevm-mds2.hpc.susx.ac.uk (2 unhandled services)
CRITICAL - Host Unreachable (139.184.80.82)
Down Since 14.05. ✔ monitor01.hpc.susx.ac.uk (2 unhandled services)
CRITICAL - Host Unreachable (139.184.80.64)

Current Incidents + X

Host | Service | Services | History X

Search...

- Dashboard
- Problems 11
- Overview
- History
- Reporting
- System
- Documentation
- mb325

Service Problems

CRITICAL 12:55	! hardware-ipmi on mds2-1.hpc.susx.ac.uk IPMI Status: Critical [Presence = Critical]
CRITICAL 06:41	! hardware-ipmi on mds2-2.hpc.susx.ac.uk IPMI Status: Critical [Presence = Critical]
CRITICAL 23:51	! hardware-ipmi on vm2.hpc.susx.ac.uk IPMI Status: Critical [BP1 Presence = Critical, BP3 Presence = Critical, BP4 Presence = Critical]
CRITICAL 19.05.	! hardware-ipmi on node043.cm.cluster IPMI Status: Critical [Sensor #129 = N/A, Sensor #129 = N/A]
CRITICAL 13.05.	! hardware-ipmi on node104.cm.cluster IPMI Status: Critical [PSU 2 Present = Critical]
CRITICAL 13.05.	! hardware-ipmi on node096.cm.cluster IPMI Status: Critical [Presence = Critical, Presence = Critical]
CRITICAL 13.05.	! hardware-ipmi on node103.cm.cluster IPMI Status: Critical [PSU 2 Present = Critical]
UNKNOWN 18.05.	! storage-openmanage on node202.cm.cluster ERROR: (SNMP) OpenManage is not installed or is not working correctly
UNKNOWN 18.05.	! hardware-openmanage on node202.cm.cluster ERROR: (SNMP) OpenManage is not installed or is not working correctly
UNKNOWN 13.05.	! storage-openmanage on grid-storm.hpc.susx.ac.uk ERROR: (SNMP) OpenManage is not installed or is not working correctly

Recently Recovered Services

OK 13m 14s	hardware-ipmi on node206.cm.cluster IPMI Status: OK
OK 14:37	hardware-ipmi on node204.cm.cluster IPMI Status: OK
OK 14:23	hardware-ipmi on node004.cm.cluster IPMI Status: OK
OK 13:02	ping4 on node108.cm.cluster PING OK - Packet loss = 0%, RTA = 0.43 ms
OK 12:35	hardware-ipmi on node011.cm.cluster IPMI Status: OK
OK	ping4 on grid-storm.hpc.susx.ac.uk

UP Since 09.04. — **node104.cm.cluster** 10.141.0.104

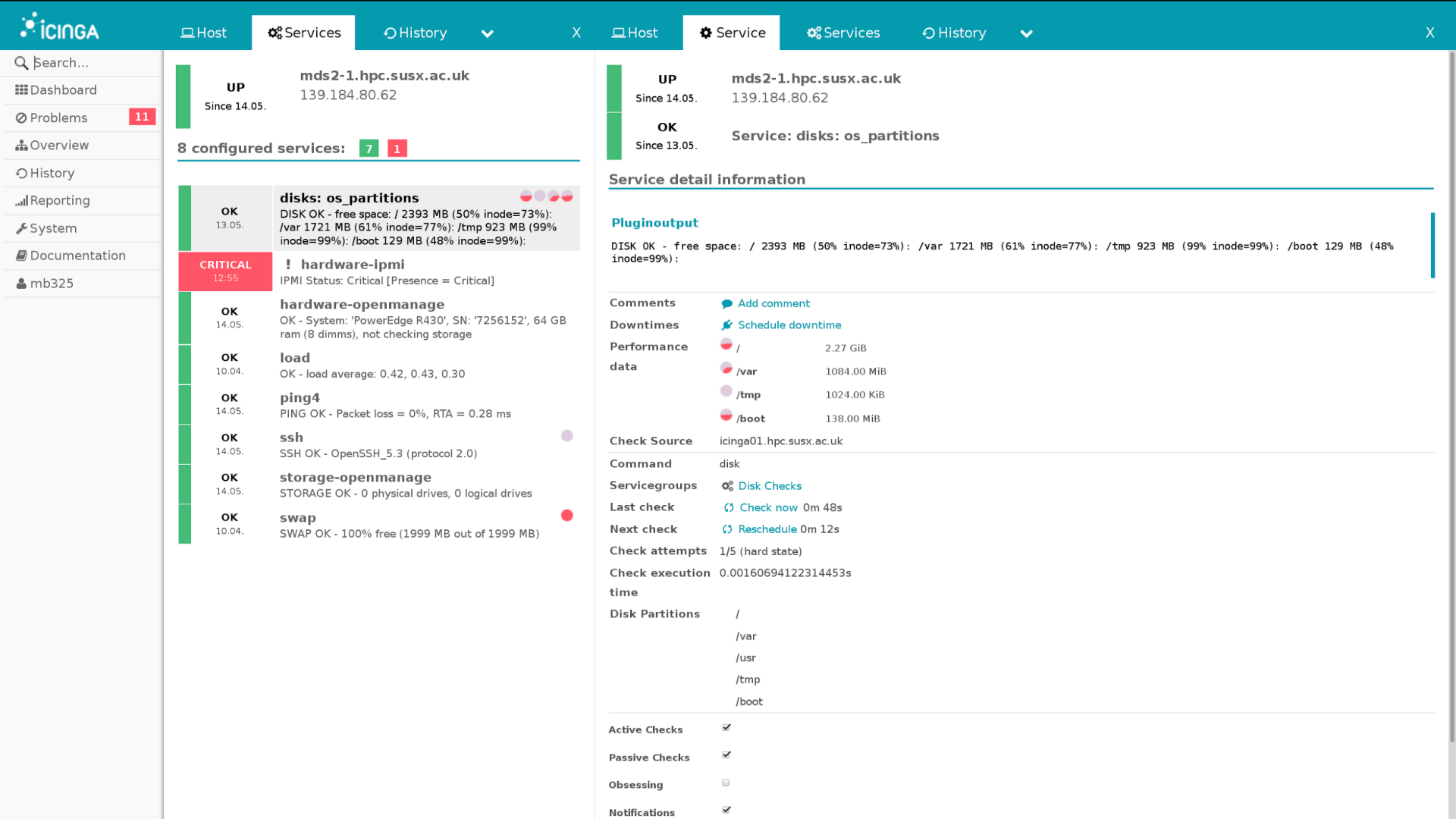
CRITICAL Since 13.05. — **Service: hardware-ipmiipmi !**

Service detail information

Pluginoutput

```
IPMI Status: Critical [PSU 2 Present = Critical]
```

Not acknowledged	✔ Acknowledge
Comments	Add comment
Notifications	No notification has been sent for this issue
Downtimes	Schedule downtime
Performance data	'FCB FAN1'=13100.00;;1500.00: 'FCB FAN2'=12900.00;;1500.00: 'FCB FAN3'=12900.00;;1500.00: 'FCB FAN4'=13100.00;;1500.00: 'PS 12V'=12.09;11.16:12.83;10.85:13.14 'PS 5V'=5.15;4.65:5.33;4.52:5.49 'Standby 3.3V'=3.33;3.04:3.48;2.94:3.59 'PS 3.3V'=3.33;0.00:6.63;0.00:6.63 'PS 1.2V'=1.18;1.15:1.27;1.10:1.31 'PS 1.1V'=1.09;0.00:2.22;0.00:2.22 'MLB TEMP 1'=39.00;~:75.00;~:80.00 'MLB TEMP 2'=51.00;~:75.00;~:80.00 'MLB TEMP 3'=56.00;~:75.00;~:80.00 'MLB TEMP 4'=50.00;~:75.00;~:80.00 NB1_TEMP=61.00;~:115.00;~:117.00 NB2_TEMP=62.00;~:115.00;~:117.00 'FCB Ambient1'=27.00;;~:50.00 'VCORE 1'=1.07;~:1.24;~:1.27 'VCORE 2'=1.15;~:1.28;~:1.32 'VCORE 3'=1.09;~:1.24;~:1.27


icinga
Host
Services
History
X
Host
Service
Services
History
X
Search...
Dashboard
Problems 11
Overview
History
Reporting
System
Documentation
mb325
UP
Since 14.05.
mds2-1.hpc.susx.ac.uk
139.184.80.62
8 configured services: 7 1
OK 13.05.
disks: os_partitions
DISK OK - free space: / 2393 MB (50% inode=73%): /var 1721 MB (61% inode=77%): /tmp 923 MB (99% inode=99%): /boot 129 MB (48% inode=99%):
CRITICAL 12:55
! hardware-ipmi
IPMI Status: Critical [Presence = Critical]
OK 14.05.
hardware-openmanage
OK - System: 'PowerEdge R430', SN: '7256152', 64 GB ram (8 dimms), not checking storage
OK 10.04.
load
OK - load average: 0.42, 0.43, 0.30
OK 14.05.
ping4
PING OK - Packet loss = 0%, RTA = 0.28 ms
OK 14.05.
ssh
SSH OK - OpenSSH_5.3 (protocol 2.0)
OK 14.05.
storage-openmanage
STORAGE OK - 0 physical drives, 0 logical drives
OK 10.04.
swap
SWAP OK - 100% free (1999 MB out of 1999 MB)
UP
Since 14.05.
mds2-1.hpc.susx.ac.uk
139.184.80.62
OK
Since 13.05.
Service: disks: os_partitions
Service detail information
Pluginoutput
DISK OK - free space: / 2393 MB (50% inode=73%): /var 1721 MB (61% inode=77%): /tmp 923 MB (99% inode=99%): /boot 129 MB (48% inode=99%):
Comments
Add comment
Downtimes
Schedule downtime
Performance data
/ 2.27 GiB
/var 1084.00 MiB
/tmp 1024.00 KiB
/boot 138.00 MiB
Check Source
icinga01.hpc.susx.ac.uk
Command
disk
Servicegroups
Disk Checks
Last check
Check now 0m 48s
Next check
Reschedule 0m 12s
Check attempts
1/5 (hard state)
Check execution time
0.00160694122314453s
Disk Partitions
/
/var
/usr
/tmp
/boot
Active Checks
Passive Checks
Obsessing
Notifications

Current Interests

- **Monitoring and Accounting**
  - Build up monitoring specifically around Grid Engine and Lustre
  - Keen to see if Logstash can fill our needs for cluster usage accounting

- **Batch system utilisation**
  - Eliminate ‘bad-neighbour’ problem through mandatory cgroup policy
  - Default resource requests through server-side JSV and job classes

- **Ceph**
  - Keen to try this out as block storage for VMs
  - Evaluate if it can serve as a new tier of storage for us -- Lustre remains as fast scratch, but CephFS as a scalable, long-term, replicated storage area for research output?