Operational matters: Documentation, Monitoring, Troubleshooting, Support

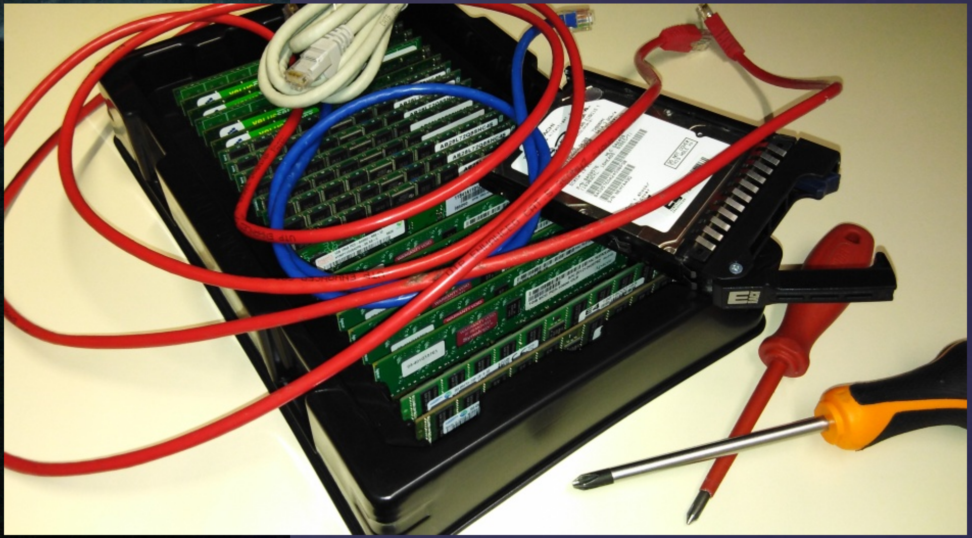
{ HPC cluster life in Production

Maria Verina& Marco Ratosa

Abdus Salam International Center for Theoretical
Physics (ICTP)
Trieste, Italy

The Information and Communication Technology
Section (**ICTS**)

Operational Maters



Operational Maters

- Day-to-day monitoring
- Documentation of procedures and processes
  - labeling of equipment
- Maintenance & Troubleshooting
- Supports contracts
- Spare parts/inventory

Documentation

- public: for users, short and clear
- internal: for managers (architecture, labels, best practices) e.g. create new user
- communication: mailing list to inform users (scheduled down-time)

ICTP Argo cluster documentation

- http://argo-doc.ictp.it/

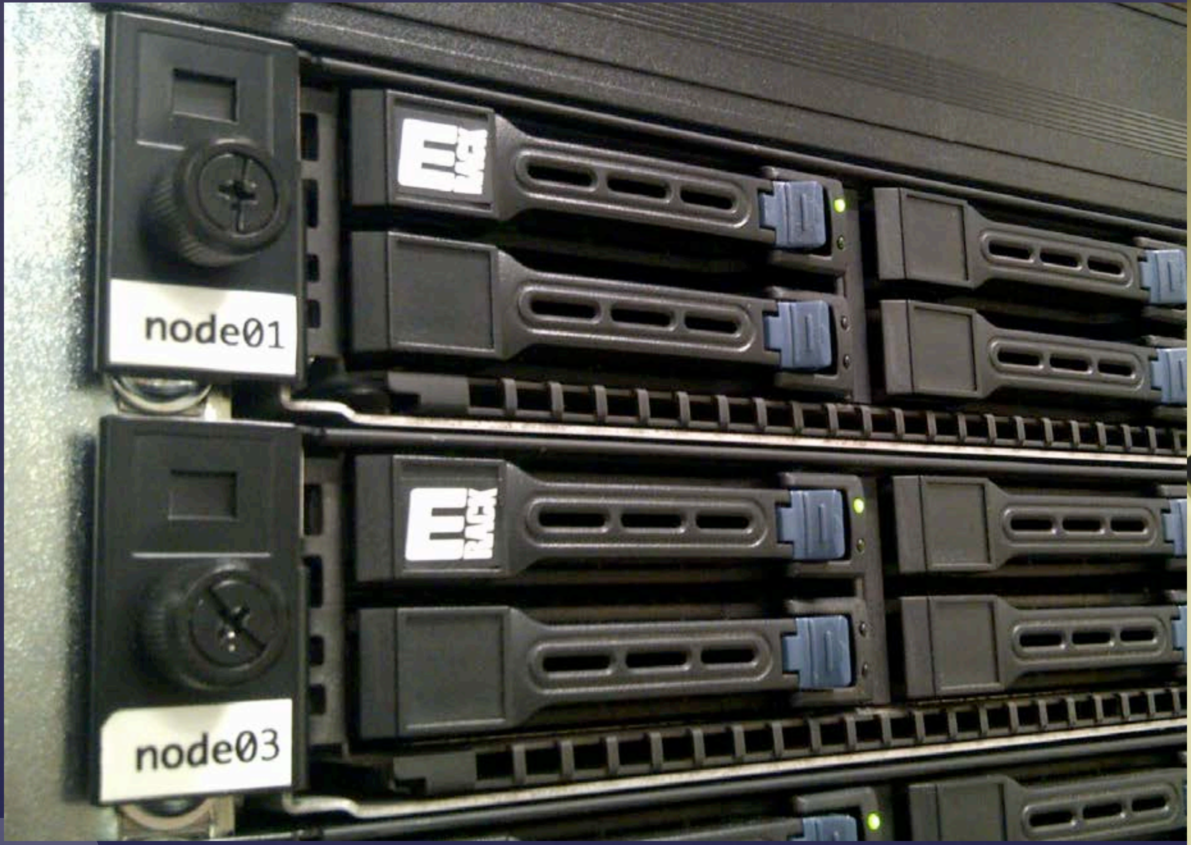
  1. **Overview, Table of available queues/partitions**
  2. Software Overview
  3. **Storage Overview**
  4. **How to use the queue manager**
  5. **Using “module” command**

- Infopack for new users.

Internal mgmt doc (HOW-TOs, architecture)

- add user
- increase storage quota
- node reinstall
- push config change to all nodes
- run command on all nodes
- list of offline/down nodes
- details about one job
- list jobs NOT running
- powe-cycle node via IPMI
- see node console (via remote console)
- …

Labeling the equipment



No labels on the back, yes in switch port descriptions

Monitoring

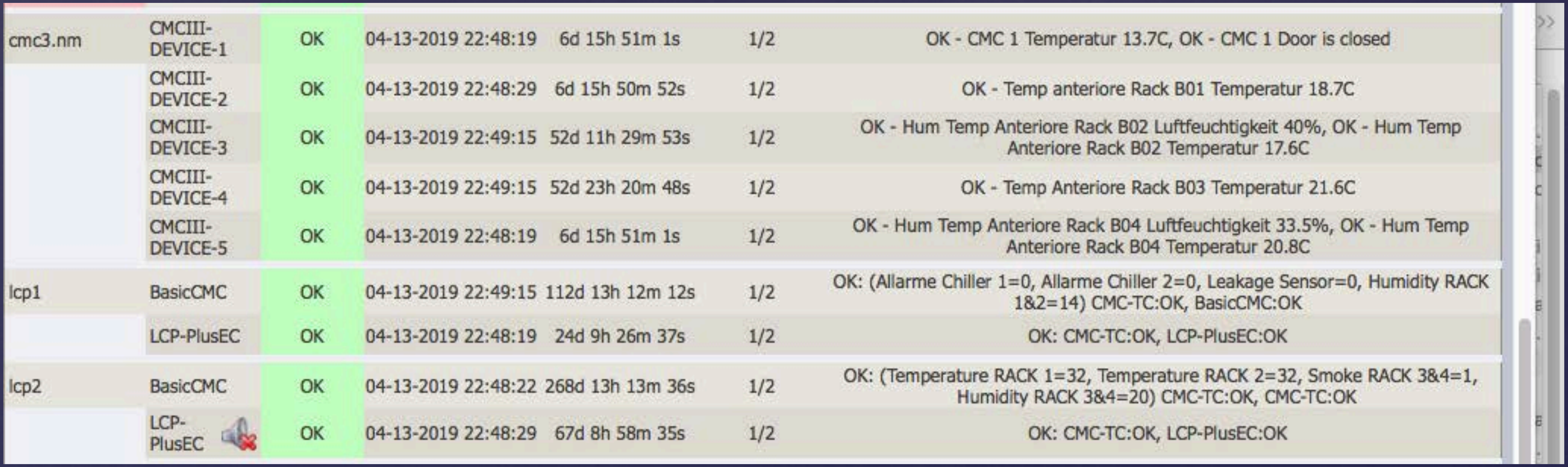
- We want to know before the users!
- Nagios is our "Team member"
- automate health checks

What can go wrong?

- Power (main!, one PDU, one power supply)
- Cooling (temperatures become hot)
- HPC Cluster itself (nodes down, jobs can not run)

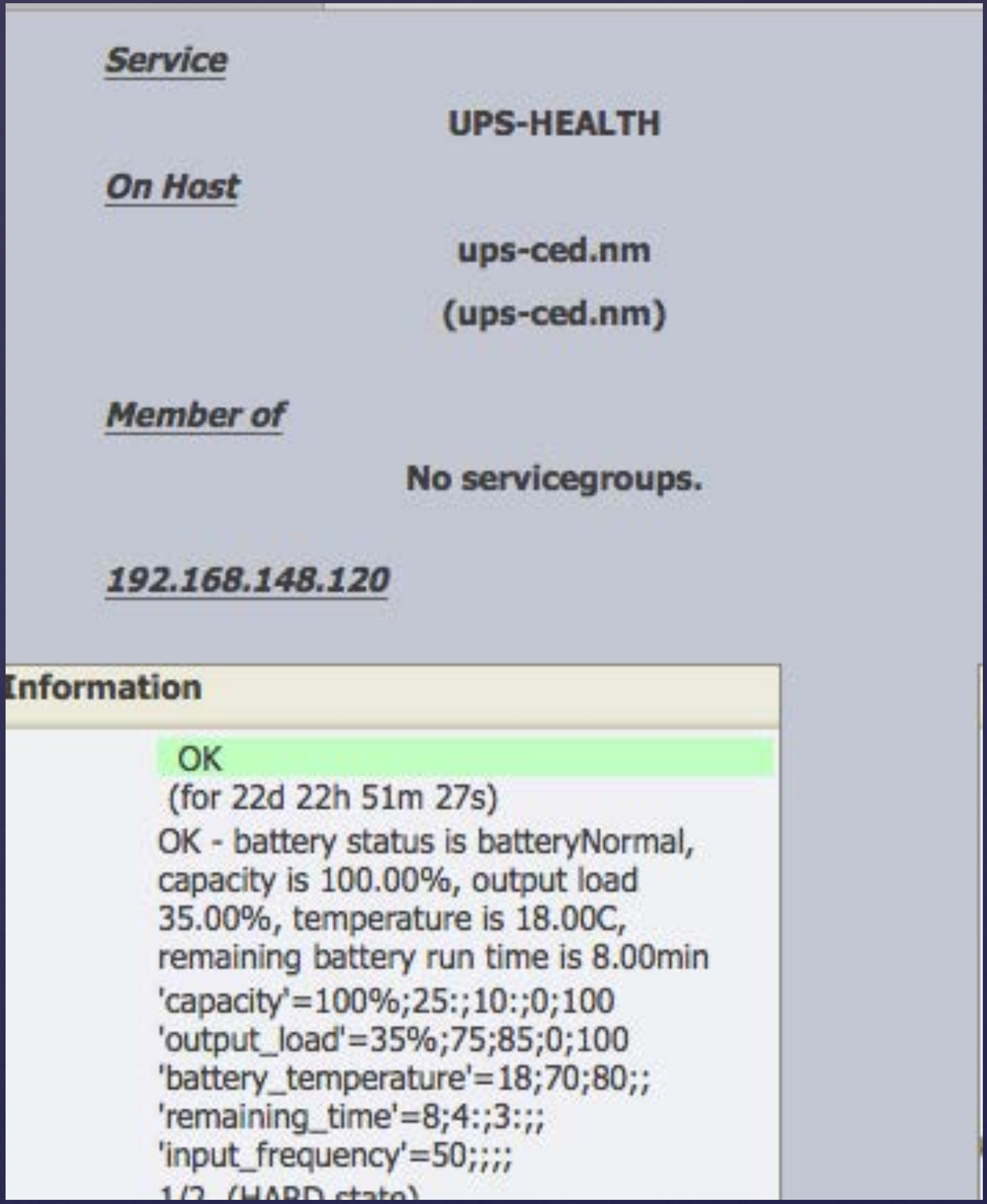
- Strategies:
  - preventive actions: monitor all (known weak points)
  - Monitor user community: RT (trouble tickets)
  - corrective actions

Monitor Cooling and Temperatures



cmc3.nm	CMCIII-DEVICE-1	OK	04-13-2019 22:48:19	6d 15h 51m 1s	1/2	OK - CMC 1 Temperatur 13.7C, OK - CMC 1 Door is closed
	CMCIII-DEVICE-2	OK	04-13-2019 22:48:29	6d 15h 50m 52s	1/2	OK - Temp anteriore Rack B01 Temperatur 18.7C
	CMCIII-DEVICE-3	OK	04-13-2019 22:49:15	52d 11h 29m 53s	1/2	OK - Hum Temp Anteriore Rack B02 Luftfeuchtigkeit 40%, OK - Hum Temp Anteriore Rack B02 Temperatur 17.6C
	CMCIII-DEVICE-4	OK	04-13-2019 22:49:15	52d 23h 20m 48s	1/2	OK - Temp Anteriore Rack B03 Temperatur 21.6C
	CMCIII-DEVICE-5	OK	04-13-2019 22:48:19	6d 15h 51m 1s	1/2	OK - Hum Temp Anteriore Rack B04 Luftfeuchtigkeit 33.5%, OK - Hum Temp Anteriore Rack B04 Temperatur 20.8C
lcp1	BasicCMC	OK	04-13-2019 22:49:15	112d 13h 12m 12s	1/2	OK: (Allarme Chiller 1=0, Allarme Chiller 2=0, Leakage Sensor=0, Humidity RACK 1&2=14) CMC-TC:OK, BasicCMC:OK
	LCP-PlusEC	OK	04-13-2019 22:48:19	24d 9h 26m 37s	1/2	OK: CMC-TC:OK, LCP-PlusEC:OK
lcp2	BasicCMC	OK	04-13-2019 22:48:22	268d 13h 13m 36s	1/2	OK: (Temperature RACK 1=32, Temperature RACK 2=32, Smoke RACK 3&4=1, Humidity RACK 3&4=20) CMC-TC:OK, CMC-TC:OK
	LCP-PlusEC	OK	04-13-2019 22:48:29	67d 8h 58m 35s	1/2	OK: CMC-TC:OK, LCP-PlusEC:OK

Monitor UPS and Generator



Service

UPS-HEALTH

On Host

ups-ced.nm

(ups-ced.nm)

Member of

No servicegroups.

192.168.148.120

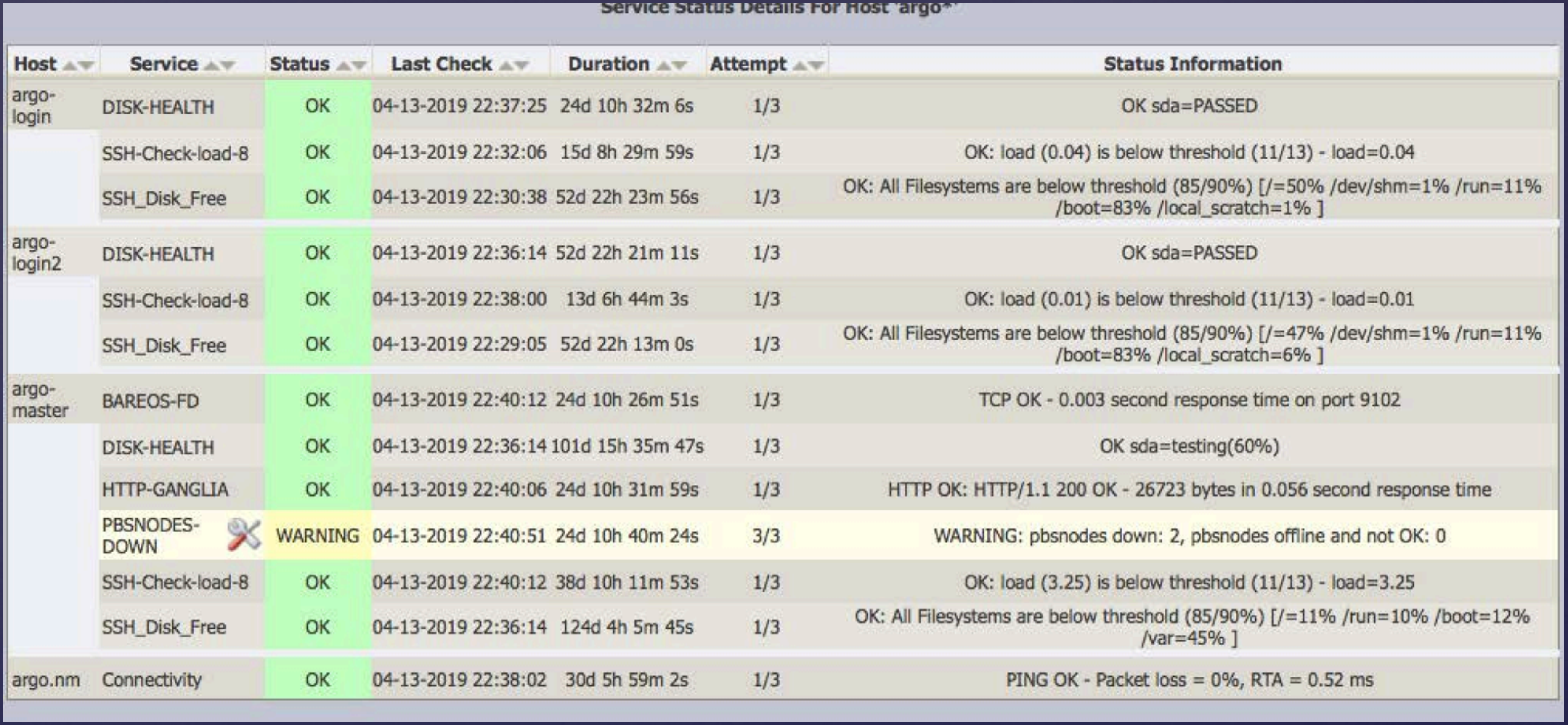
Information

```
OK
(for 22d 22h 51m 27s)
OK - battery status is batteryNormal,
capacity is 100.00%, output load
35.00%, temperature is 18.00C,
remaining battery run time is 8.00min
'capacity'=100%;25:;10:;0;100
'output_load'=35%;75;85;0;100
'battery_temperature'=18;70;80;;
'remaining_time'=8;4:;3:;;
'input_frequency'=50;;;;
1/2 (HARD state)
```

What can go wrong? (inside the Cluster)

- hw (mem, HD, net cable)
- sw (kernel oops) – reboot, opt reinstall
- sw: queue manager problems: Can you run a short job?
- user reports a problem (inspect job script, job output, log files)
- several users report similar problem (oops!)

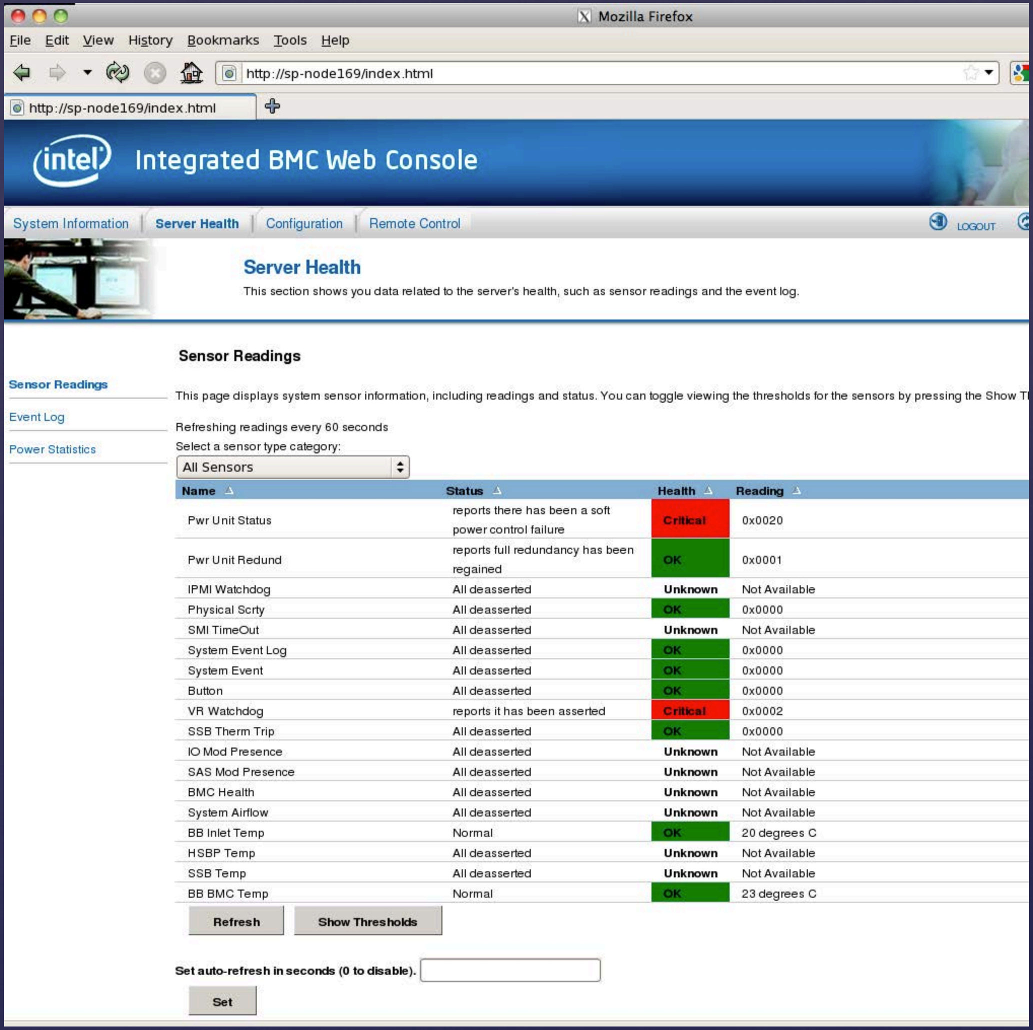
Cluster checks from Nagios



Service Status Details For Host 'argo*'

Host	Service	Status	Last Check	Duration	Attempt	Status Information
argo-login	DISK-HEALTH	OK	04-13-2019 22:37:25	24d 10h 32m 6s	1/3	OK sda=PASSED
	SSH-Check-load-8	OK	04-13-2019 22:32:06	15d 8h 29m 59s	1/3	OK: load (0.04) is below threshold (11/13) - load=0.04
	SSH_Disk_Free	OK	04-13-2019 22:30:38	52d 22h 23m 56s	1/3	OK: All Filesystems are below threshold (85/90%) [/=50% /dev/shm=1% /run=11% /boot=83% /local_scratch=1%]
argo-login2	DISK-HEALTH	OK	04-13-2019 22:36:14	52d 22h 21m 11s	1/3	OK sda=PASSED
	SSH-Check-load-8	OK	04-13-2019 22:38:00	13d 6h 44m 3s	1/3	OK: load (0.01) is below threshold (11/13) - load=0.01
	SSH_Disk_Free	OK	04-13-2019 22:29:05	52d 22h 13m 0s	1/3	OK: All Filesystems are below threshold (85/90%) [/=47% /dev/shm=1% /run=11% /boot=83% /local_scratch=6%]
argo-master	BAREOS-FD	OK	04-13-2019 22:40:12	24d 10h 26m 51s	1/3	TCP OK - 0.003 second response time on port 9102
	DISK-HEALTH	OK	04-13-2019 22:36:14	101d 15h 35m 47s	1/3	OK sda=testing(60%)
	HTTP-GANGLIA	OK	04-13-2019 22:40:06	24d 10h 31m 59s	1/3	HTTP OK: HTTP/1.1 200 OK - 26723 bytes in 0.056 second response time
	PBSNODES-DOWN	WARNING	04-13-2019 22:40:51	24d 10h 40m 24s	3/3	WARNING: pbsnodes down: 2, pbsnodes offline and not OK: 0
	SSH-Check-load-8	OK	04-13-2019 22:40:12	38d 10h 11m 53s	1/3	OK: load (3.25) is below threshold (11/13) - load=3.25
	SSH_Disk_Free	OK	04-13-2019 22:36:14	124d 4h 5m 45s	1/3	OK: All Filesystems are below threshold (85/90%) [/=11% /run=10% /boot=12% /var=45%]
argo.nm	Connectivity	OK	04-13-2019 22:38:02	30d 5h 59m 2s	1/3	PING OK - Packet loss = 0%, RTA = 0.52 ms

PSU problem



Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://sp-node169/index.html
http://sp-node169/index.html
Integrated BMC Web Console
System Information
Server Health
Configuration
Remote Control
LOGOUT

Server Health

This section shows you data related to the server's health, such as sensor readings and the event log.

Sensor Readings
Event Log
Power Statistics

Sensor Readings

This page displays system sensor information, including readings and status. You can toggle viewing the thresholds for the sensors by pressing the Show T

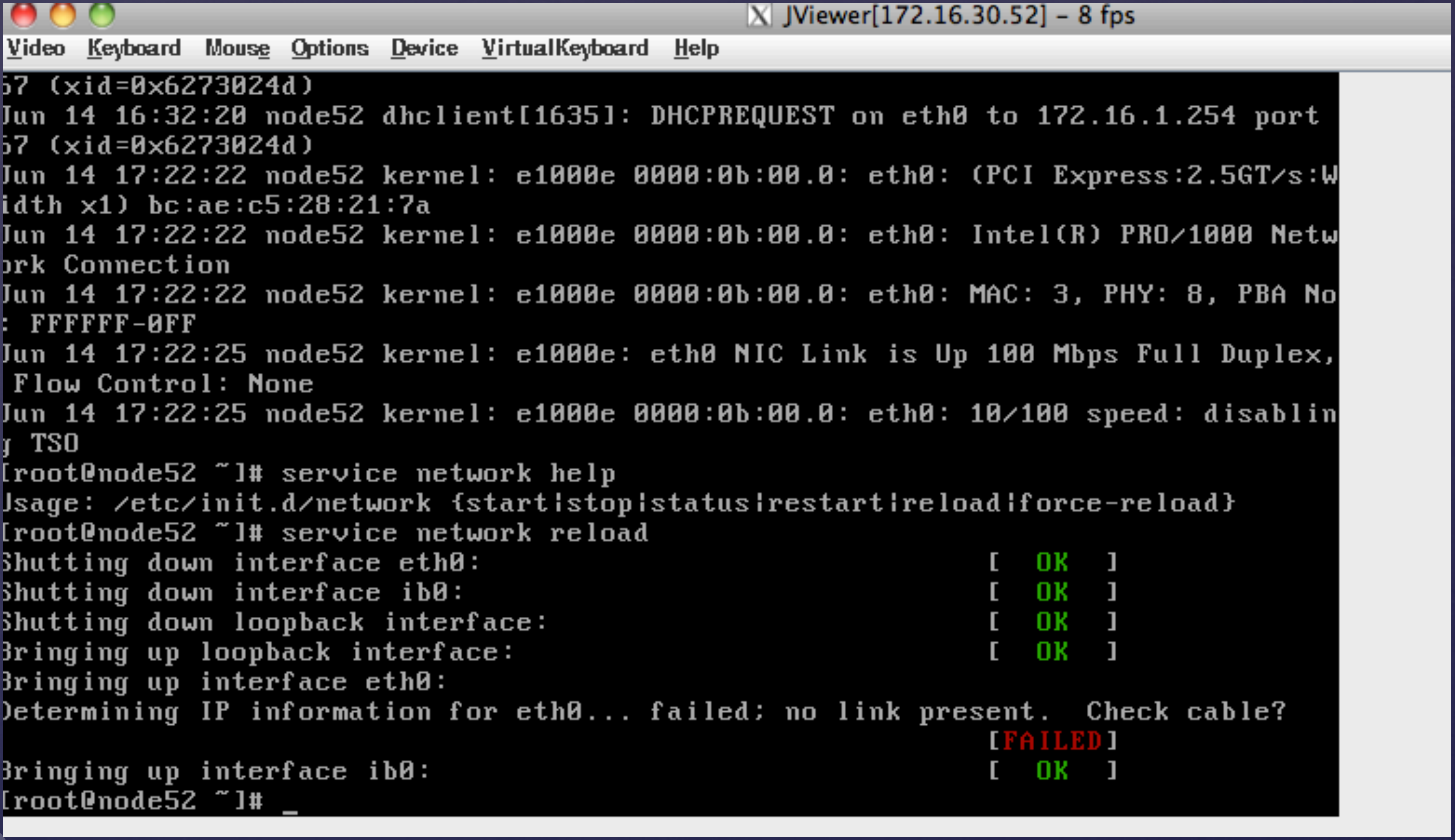
Refreshing readings every 60 seconds

Select a sensor type category: All Sensors

Name	Status	Health	Reading
Pwr Unit Status	reports there has been a soft power control failure	Critical	0x0020
Pwr Unit Redund	reports full redundancy has been regained	OK	0x0001
IPMI Watchdog	All deasserted	Unknown	Not Available
Physical Scrty	All deasserted	OK	0x0000
SMI TimeOut	All deasserted	Unknown	Not Available
System Event Log	All deasserted	OK	0x0000
System Event	All deasserted	OK	0x0000
Button	All deasserted	OK	0x0000
VR Watchdog	reports it has been asserted	Critical	0x0002
SSB Therm Trip	All deasserted	OK	0x0000
IO Mod Presence	All deasserted	Unknown	Not Available
SAS Mod Presence	All deasserted	Unknown	Not Available
BMC Health	All deasserted	Unknown	Not Available
System Airflow	All deasserted	Unknown	Not Available
BB Inlet Temp	Normal	OK	20 degrees C
HSBP Temp	All deasserted	Unknown	Not Available
SSB Temp	All deasserted	Unknown	Not Available
BB BMC Temp	Normal	OK	23 degrees C

Set auto-refresh in seconds (0 to disable).

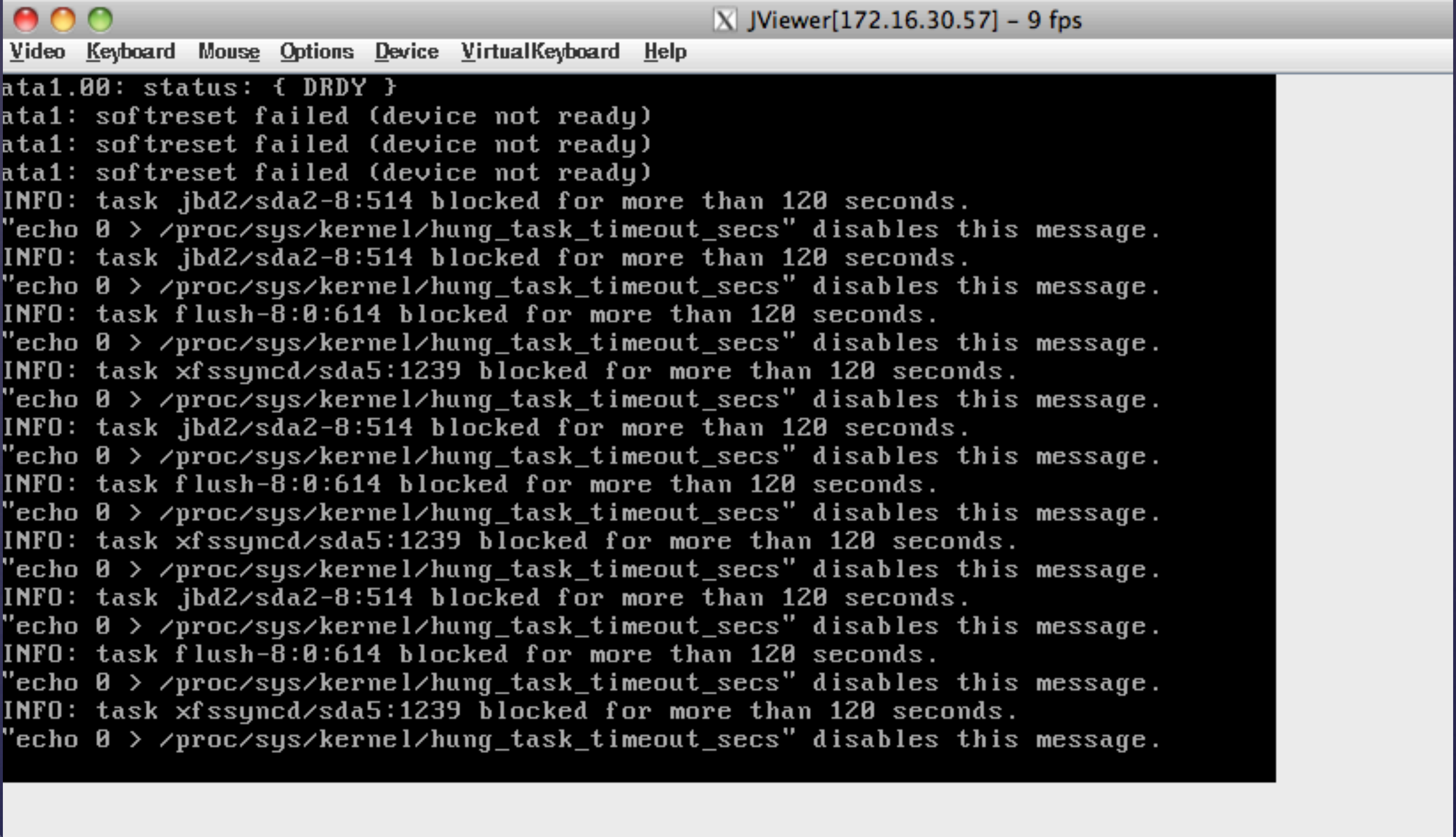
Eth cable problem



JViewer[172.16.30.52] - 8 fps
Video Keyboard Mouse Options Device VirtualKeyboard Help

```
57 (xid=0x6273024d)
Jun 14 16:32:20 node52 dhclient[1635]: DHCPREQUEST on eth0 to 172.16.1.254 port
57 (xid=0x6273024d)
Jun 14 17:22:22 node52 kernel: e1000e 0000:0b:00.0: eth0: (PCI Express:2.5GT/s:W
idth x1) bc:ae:c5:28:21:7a
Jun 14 17:22:22 node52 kernel: e1000e 0000:0b:00.0: eth0: Intel(R) PRO/1000 Netw
ork Connection
Jun 14 17:22:22 node52 kernel: e1000e 0000:0b:00.0: eth0: MAC: 3, PHY: 8, PBA No
: FFFFFF-0FF
Jun 14 17:22:25 node52 kernel: e1000e: eth0 NIC Link is Up 100 Mbps Full Duplex,
 Flow Control: None
Jun 14 17:22:25 node52 kernel: e1000e 0000:0b:00.0: eth0: 10/100 speed: disablin
g TSO
[root@node52 ~]# service network help
Usage: /etc/init.d/network {start|stop|status|restart|reload|force-reload}
[root@node52 ~]# service network reload
Shutting down interface eth0:                              [  OK  ]
Shutting down interface ib0:                               [  OK  ]
Shutting down loopback interface:                          [  OK  ]
Bringing up loopback interface:                            [  OK  ]
Bringing up interface eth0:
Determining IP information for eth0... failed; no link present.  Check cable?
                                                           [FAILED]
Bringing up interface ib0:                                 [  OK  ]
[root@node52 ~]# _
```

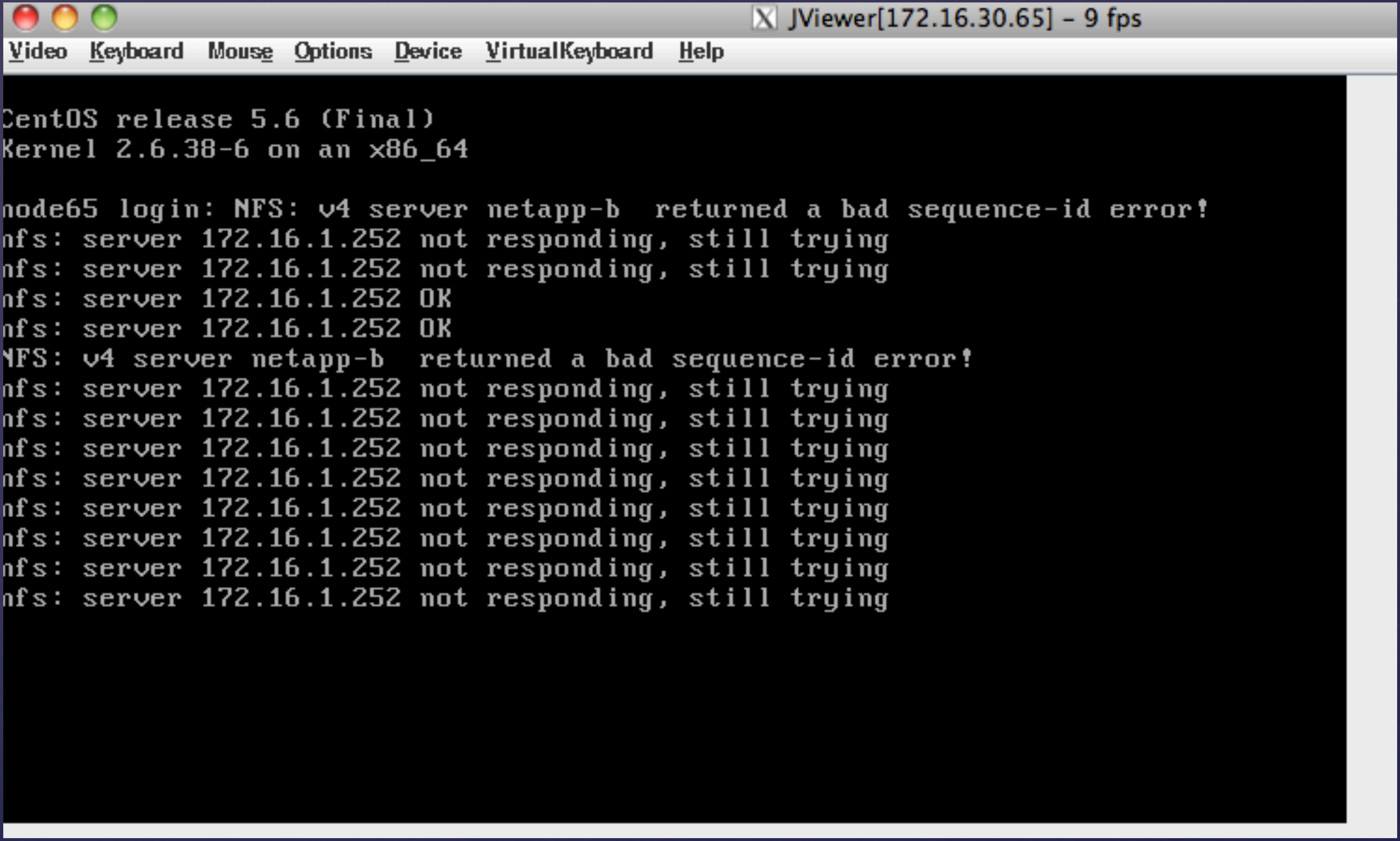
Hw: HD broken (DRDY)



JViewer[172.16.30.57] – 9 fps
Video Keyboard Mouse Options Device VirtualKeyboard Help

```
ata1.00: status: { DRDY }
ata1: softreset failed (device not ready)
ata1: softreset failed (device not ready)
ata1: softreset failed (device not ready)
INFO: task jbd2/sda2-8:514 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task jbd2/sda2-8:514 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task flush-8:0:614 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task xfssyncd/sda5:1239 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task jbd2/sda2-8:514 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task flush-8:0:614 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task xfssyncd/sda5:1239 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task jbd2/sda2-8:514 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task flush-8:0:614 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
INFO: task xfssyncd/sda5:1239 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
```

Network or NAS problem



JViewer[172.16.30.65] - 9 fps
Video Keyboard Mouse Options Device VirtualKeyboard Help

```
CentOS release 5.6 (Final)
Kernel 2.6.38-6 on an x86_64

node65 login: NFS: v4 server netapp-b  returned a bad sequence-id error!
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 OK
nfs: server 172.16.1.252 OK
NFS: v4 server netapp-b  returned a bad sequence-id error!
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
nfs: server 172.16.1.252 not responding, still trying
```

Health checks within the cluster

- TORQUE, SLURM, and other resource managers provide for a periodic "node health check"
- "unhealthy" nodes are marked as drained/ offline (prevent jobs from being run on them)
- Drained node can then undergo maintenance actions

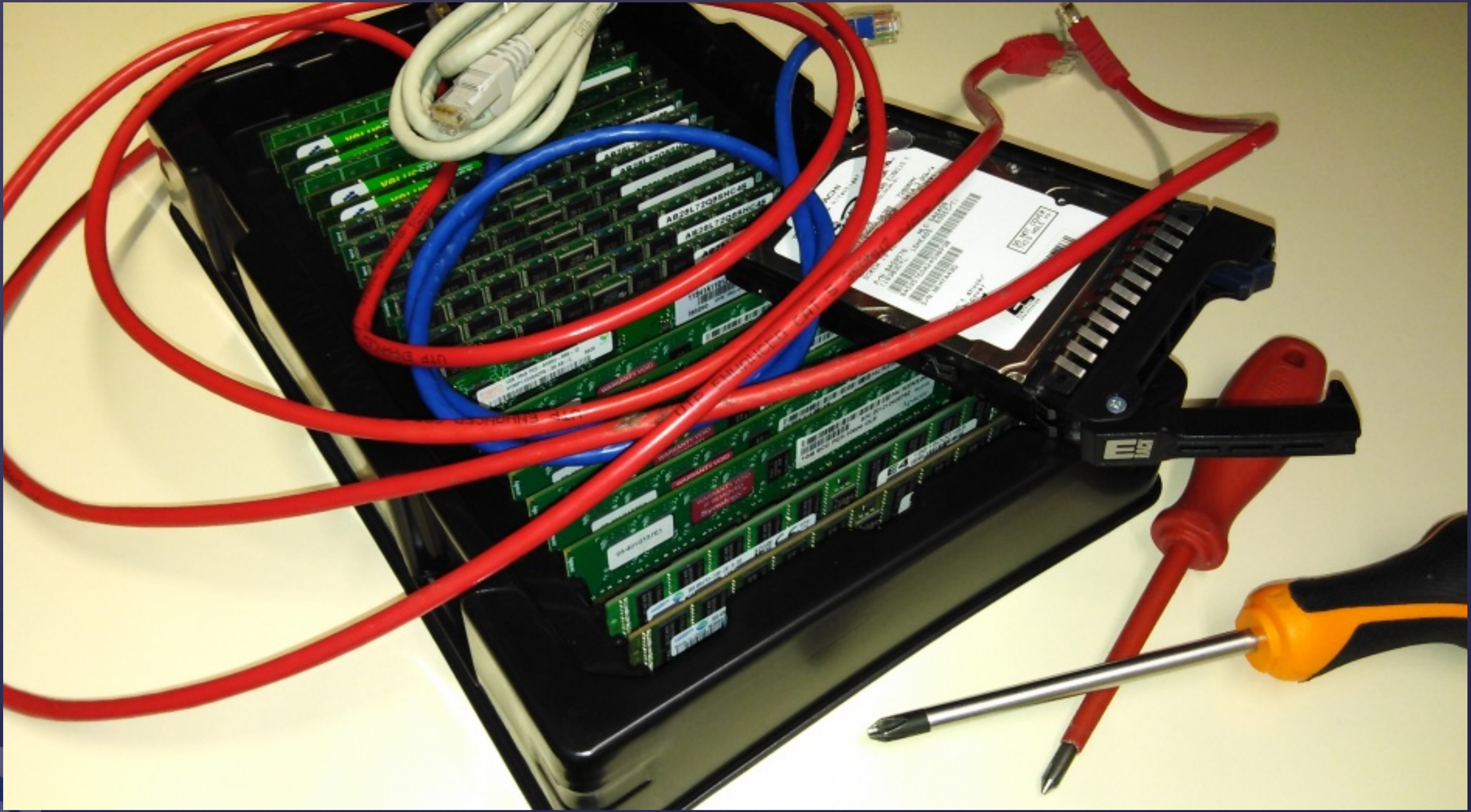
Health check examples

```
Nov 17 14:09:28 node52 pbs_node_health: ERROR IN CRASH STATE ETHERNET LINK (eth0) BAD SPEED of 100M

Jun  2 04:19:49 node33 pbs_node_health: ERROR IN CRASH STATE INFINIBAND LINK DOWN

[2018-07-12T08:42:24.775] error: Node node111 has low real_memory size (31903 < 64156)

Sep 27 10:33:03 node48 nhc[88875]: Health check failed:  check_ps_loadavg:  1-minute load average too high:  24 >= 15

Jun  1 14:32:20 node142 nhc[35064]: Health check failed:  Script timed out while executing "check_fs_free /local_scratch 3%".

Apr  8 12:18:37 node57 pbs_node_health: ERROR IN CRASH STATE UPTIME - BUT NODE IS BUSY
```

Troubleshooting

- automatic monitoring (external and within the cluster)
- when problem happens, quickly determine the scope (one user?, one node? vs blanket-problem)
- useful commands (offline nodes and reason)
- inspect job's output and error files
- inspect logs for more details about the reason (OS, queue manager logs)
- inspect node's Console, for last messages reported
- cross-check with list/memory of Known problems
- involve HPC team, cluster architect, application specialist
- inform user(s)
- resolved problem = lesson learned: document Known Issue and it's resolution

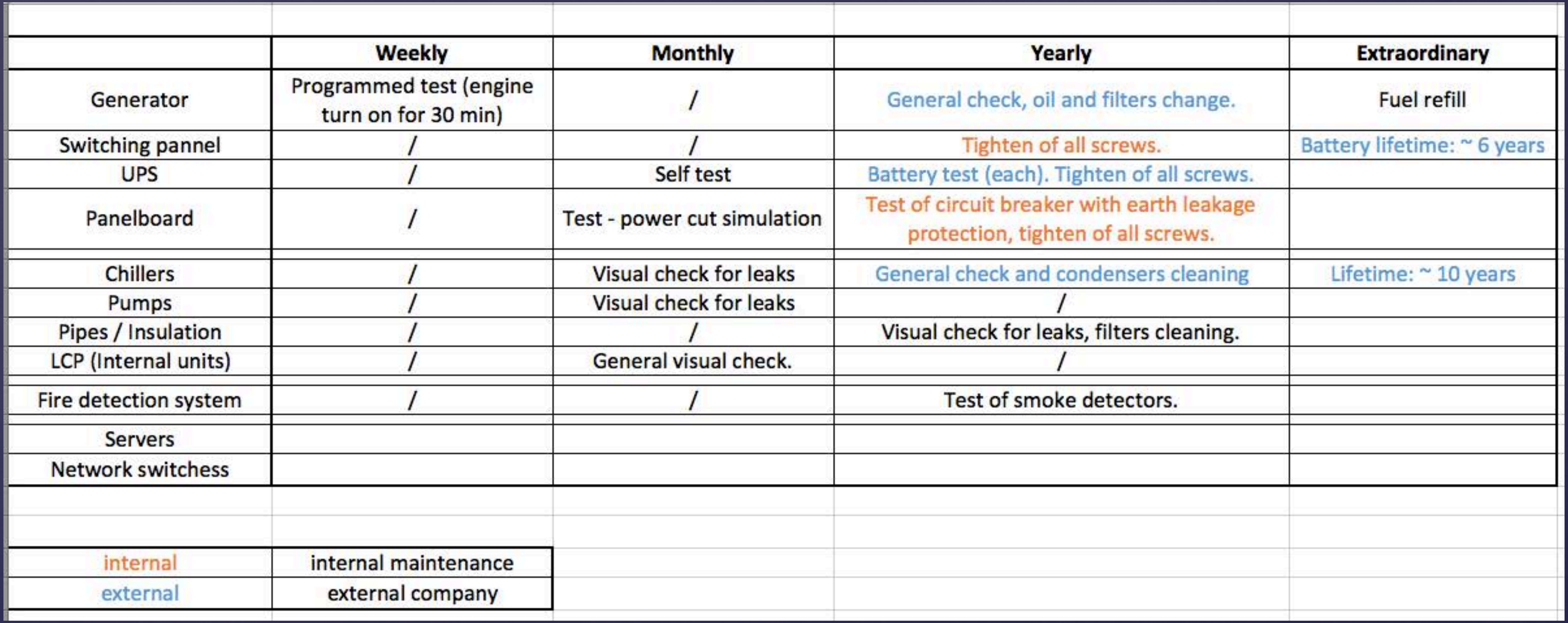
Spare parts



Support

- internal HPC team
- external support contracts

Maintenance schedule



	Weekly	Monthly	Yearly	Extraordinary
Generator	Programmed test (engine turn on for 30 min)	/	General check, oil and filters change.	Fuel refill
Switching pannel	/	/	Tighten of all screws.	Battery lifetime: ~ 6 years
UPS	/	Self test	Battery test (each). Tighten of all screws.	
Panelboard	/	Test - power cut simulation	Test of circuit breaker with earth leakage protection, tighten of all screws.	
Chillers	/	Visual check for leaks	General check and condensers cleaning	Lifetime: ~ 10 years
Pumps	/	Visual check for leaks	/	
Pipes / Insulation	/	/	Visual check for leaks, filters cleaning.	
LCP (Internal units)	/	General visual check.	/	
Fire detection system	/	/	Test of smoke detectors.	
Servers				
Network switchess				

internal	internal maintenance
external	external company

Conclusions:

- Monitor HPC cluster and it’s environment
- build HPC support team (including external suppliers)
- document all (for user, for the team)

Thank You!

Maria & Marco

Questions ?