Best Practice on System Management and Configuration: The Ulisse Linux Cluster

Piero Calucci
SISSA

Scuola Internazionale Superiore
di Studi Avanzati

Outline

cluster management & infrastructure management:

- installation and configuration
- monitoring
- maintenance

xCAT

- we use xCAT for both node deployment and configuration management
- `http://xcat.sf.net`
- 100% free, developed by IBM
  - especially suited for medium-sized to large clusters, and for RH- or SUSE-based distributions (but can install also debian-based distros; and Windows too)
- everything is scriptable

xCAT /2

- can install nodes with a single command, sync files to nodes, run preconfigured scripts or any other command on nodes
- can work on single node, preconfigured sets or arbitrary list of nodes
  - (re)install a whole rack: `rinstall rack04`
  - run a command on all GPU nodes:
    ```
    psh gnode /path/to/my_command.sh
    ```
  - update custom config files on all nodes:
    ```
    updatenode compute -F
    ```
  - power on an entire rack:
    ```
    rpower rack01
    ```

xCAT /3

- needs some preliminary work
  - set up tables with node name / IP / mac
  - IPMI must work (at least power commands)
  - prepare software list (kickstart or similar), plus customization scripts and config files
- good if you have 100s of identical nodes
- not so good if you have a very small or highly heterogeneous cluster
  (but highly heterogeneous clusters are evil anyway, so…)

Monitoring: logs

- have a central log server
  - can be the master node, or a dedicated log server
- forward syslog from *everywhere* to log server
  - compute nodes and login nodes, obviously
  - service processors (iLO/IMM/whatever)
  - storage servers
  - switches
  - UPS, air conditioning, environmental monitoring, …

Monitoring: logs

- know how to analyze logs
  - our cluster generates ~200k log lines per day, on «good» days
  - can be several millions when you are in troubles
- logwatch provides a starting point for automated log analysis
  - several custom scripts plugged in
- never underestimate the power of one-line scripts!

Monitoring: logs

- example: you notice `/var/log/messages` is growing faster than usual. Why so?

```
# wc -l /var/log/messages
113624 /var/log/messages

# awk '{print $4}' </var/log/messages | sort | uniq -c |
sort -g | tail -1
   4767 cn06-08
```

a single node is generating 4% of total log volume (we have ~250 nodes, so you would expect 0.4%)

It turned out that a user was running benchmarks of his own and had 100s of processes killed by OOMk

Monitoring: logs

- sometimes log messages are so obscure that reading them doesn't help
  - `tNetScheduler[825a12a0]: Osa: arptnew failed on 0`
- however just knowing how many of them come from where is interesting
  - you have a problem when your usually silent IB switch spits out 10 messages per second
  - look into running jobs when compute nodes become too «noisy»
  - you probably need hardware maintenance when IPMI logs grow out of bound

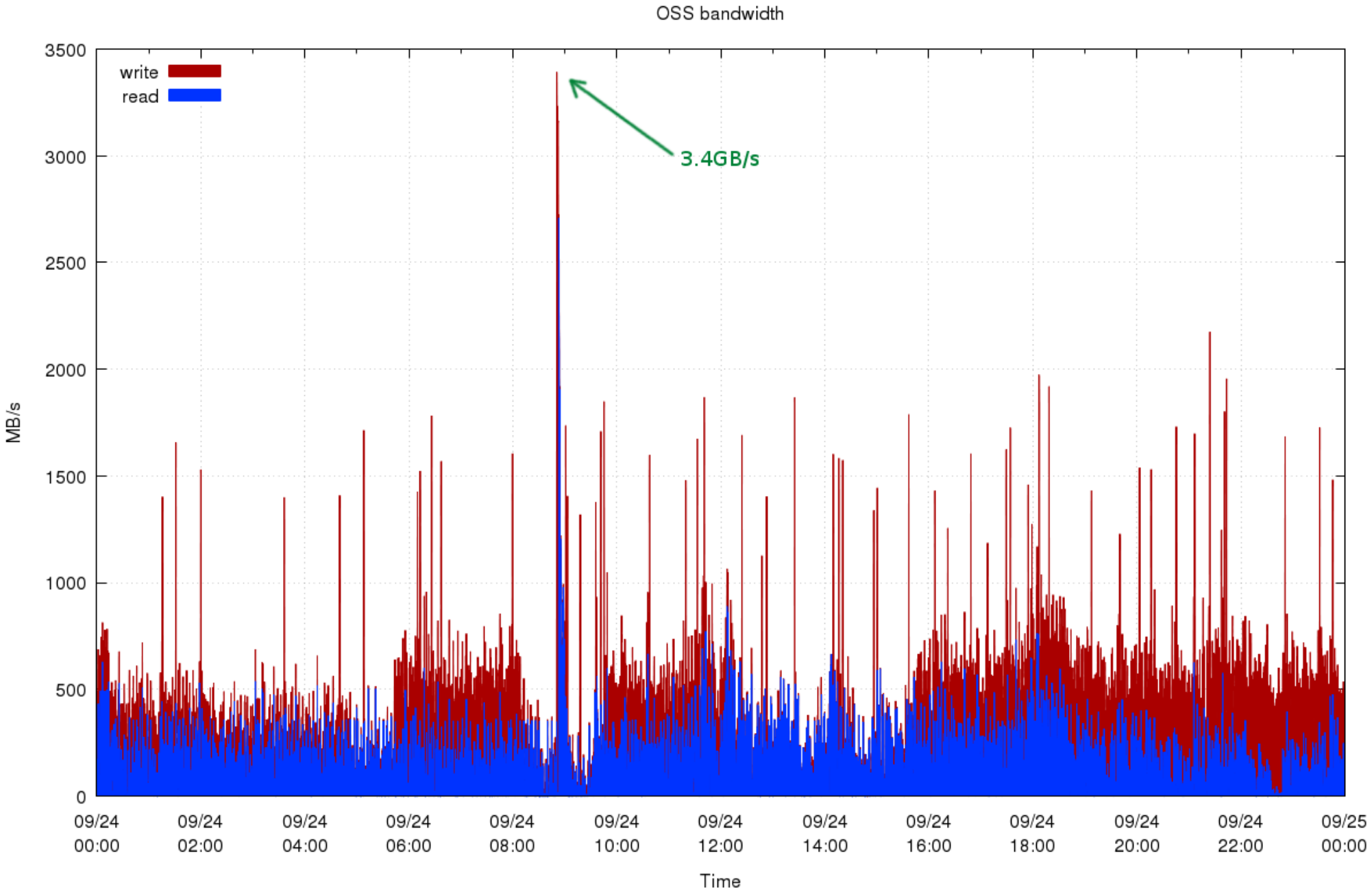
Monitoring: performance

- different methods
  - sysstat / PCP / collectl instead of syslog
  - queue system logs also provide performance data
- different goals
  - is the cluster performing «well»?
  - are people actually using the computing resource?
  - are they using it efficiently or are they wasting resources?

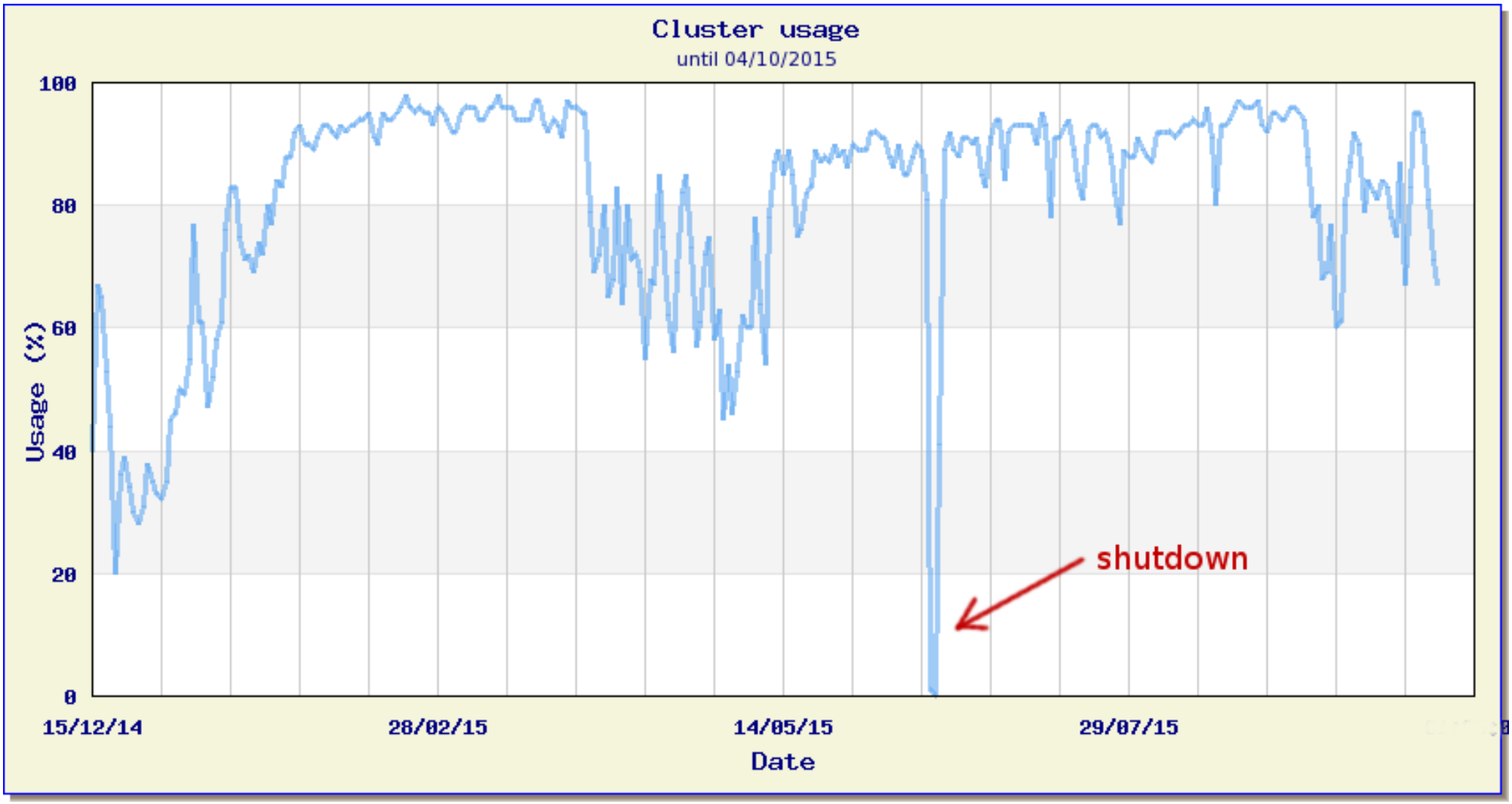
Monitoring: performance

- different goals (continued)
  - does that shiny new 300k€ storage system deliver what it promised?
  - is there some bottleneck that slows down the entire cluster?
  - shall we spend some more money on GPUs? or to buy more memory? or faster CPUs?
  - how much are we going to pay in utility bills if we run like that for the next 6 months? and if we install 50% more nodes? (and do we really need those more nodes?)

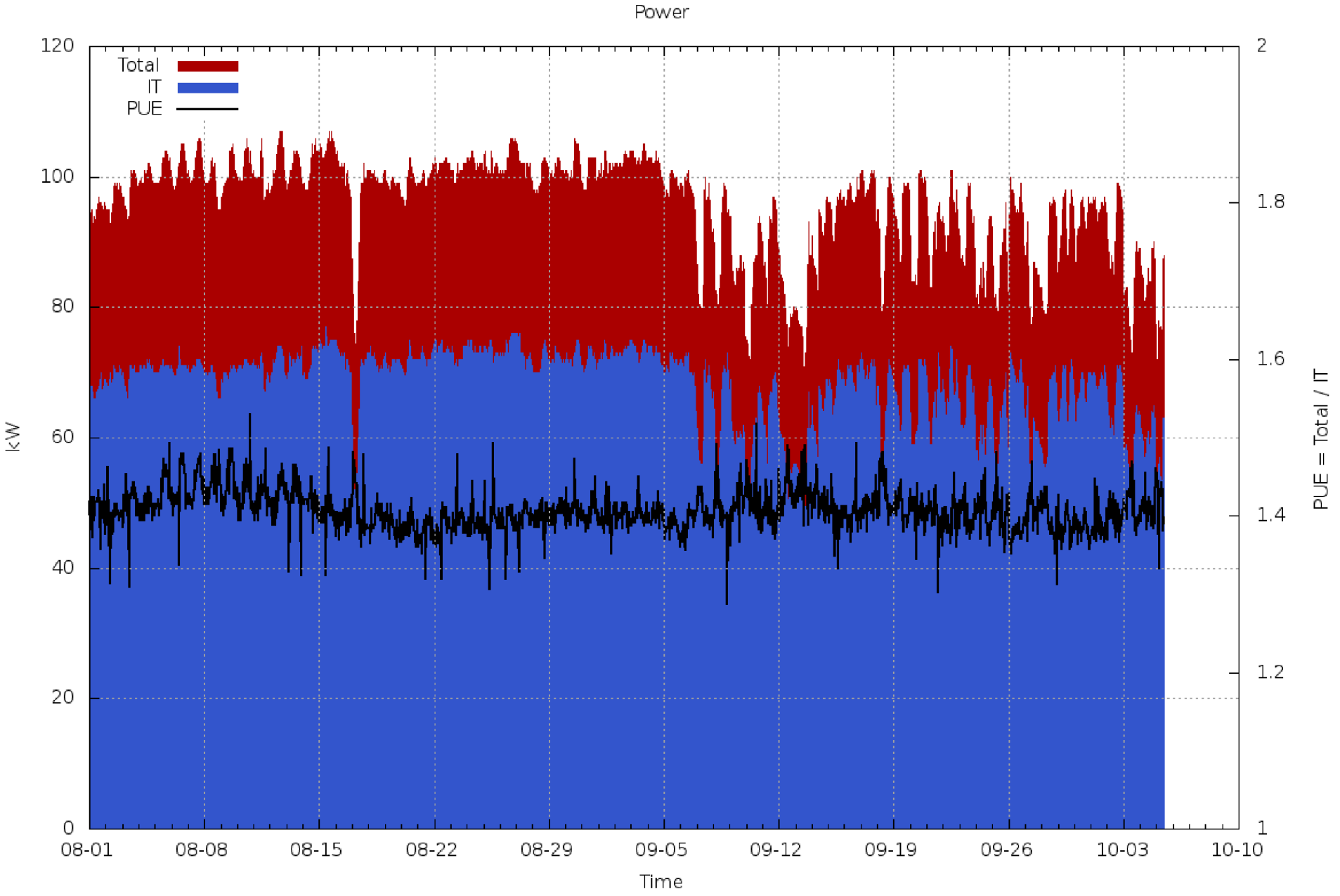
Performance example: filesystem



Performance example: overall cluster usage



Performance example: energy consumption



Hardware Maintenance

- reactive
  - be ready to replace broken disks / memory / power supplies / …
  - (so far, we have replaced more memory modules than all other hw components combined)
- preventive
  - almost mandatory for the non-IT part: UPS, air conditioning, switchboards, fire extinguishing system, …

Hardware Maintenance

- can you reliably detect when a piece of hw is failing?
  - disks → SMART, native RAID utilities
  - memory → EDAC / mcelog
  - CPU, mb, fans, power supply → IPMI
  - network → ethtool, ping, ibcheckerr
  - all of them → degraded performance, system is unstable, unexpected reboots

Monitoring with Nagios

not only hosts up / down!

node allocated to some job, but
no user process running

Host	Service		Status	Last	Status Information
cn02-11	job processes		WARNING	2015-	WARNING: missing job process?
cn02-15	MCE_dimm-error	?	WARNING	2015	corrected DIMM memory error count exceeded threshold: 40 in 24h
	MCE_page-error	?	WARNING	2015-1	Corrected memory errors on page 83cad8000 exceed threshold 10 in 24h: 10 in 24h
	MCE_socket-error	?	WARNING	2015-	corrected Socket memory error count exceeded threshold: 40 in 24h
cn02-27	temperature		UNKNOWN	2015	No temperatures

failing
memory

temperature sensor is
not reporting current
value

Questions?

<calucci at sissa dot it>