GridPP Experiences with the Cern Virtual Machine File System (CVMFS)

Christopher J. Walker

Queen Mary, University of London

Overview

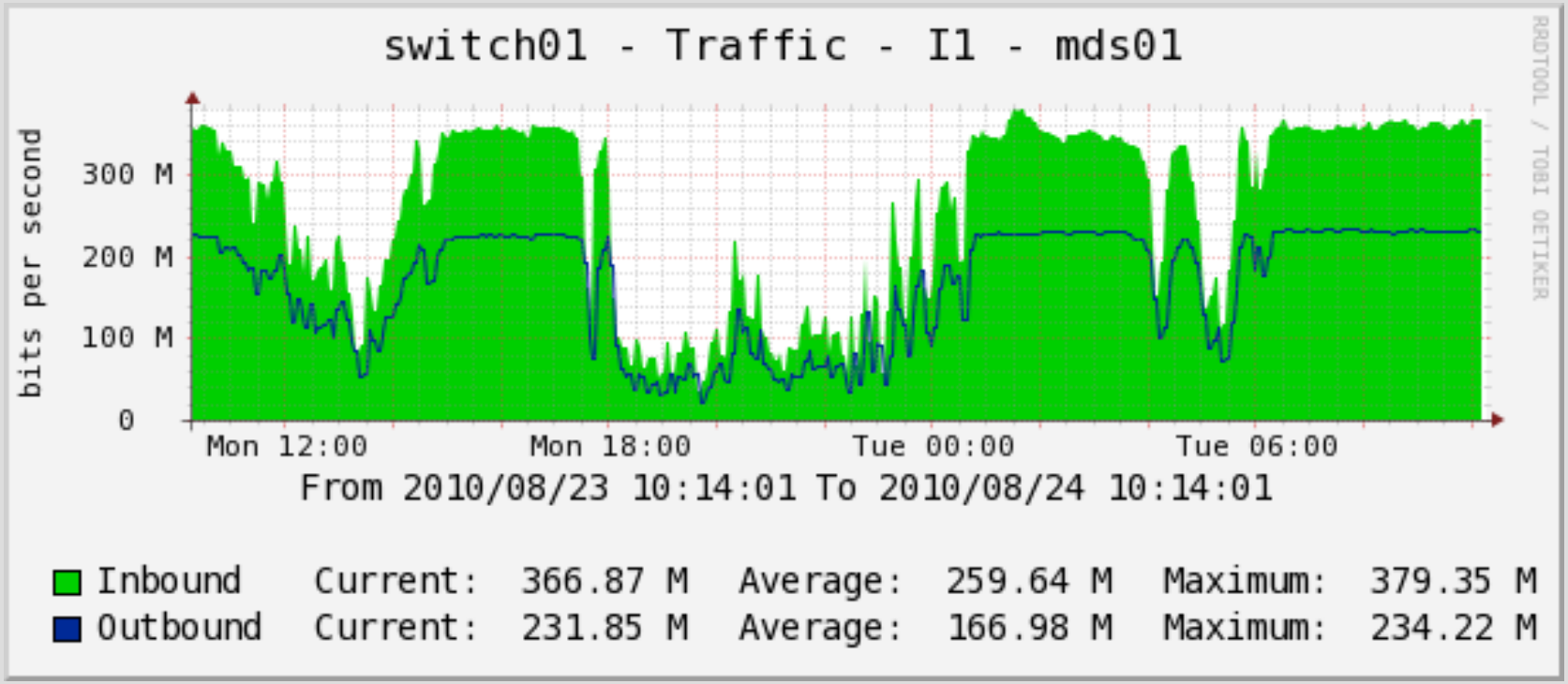
- What is CVMFS
  - Not just for virtual machines
- QMUL's Motivation
  - NFS locking issues
  - Lustre - MDS load
  - Software install hassles
- Installation
- Performance
- Conclusions

What is CVMFS

- Developed for the CERN VM
  - Standalone software – VM not needed
- FUSE module
  - Web directory tree looks like a filesystem
    - Updates automatically
  - Local cache
  - Deduplication
  - SQUID Proxy

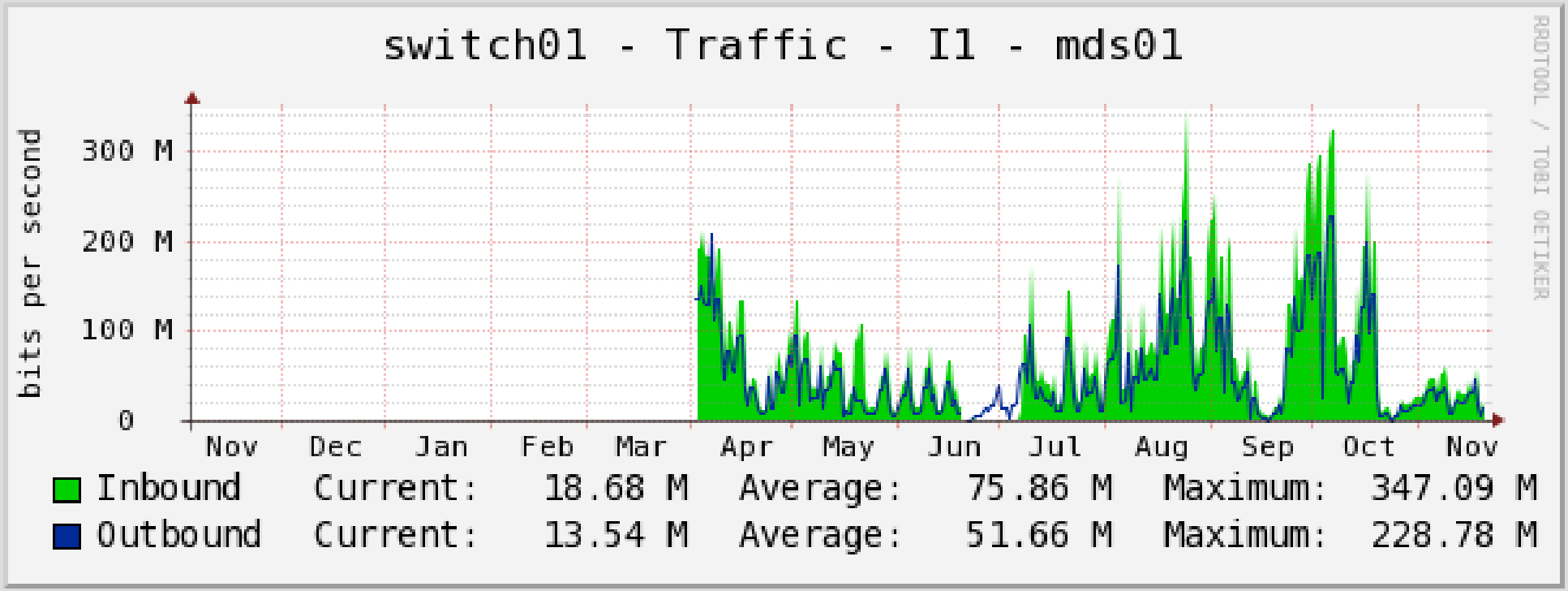
QMUL's Motivation

- NFS: locking issues
- 4 900 000 files in atlas software area
- Lustre: High MDS Load (>100)
- Which releases to provide

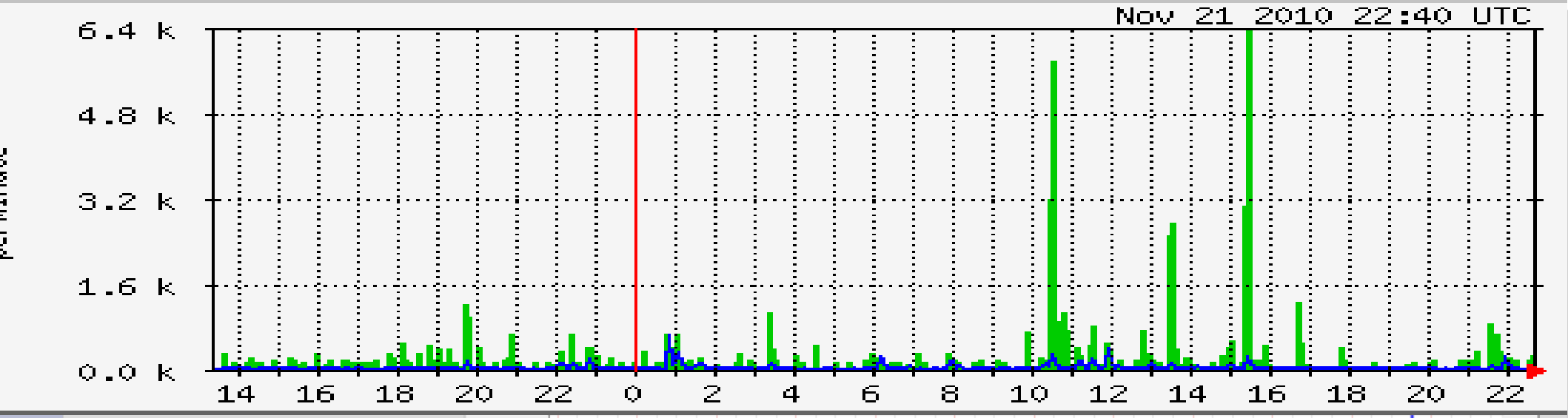


QMUL's Results

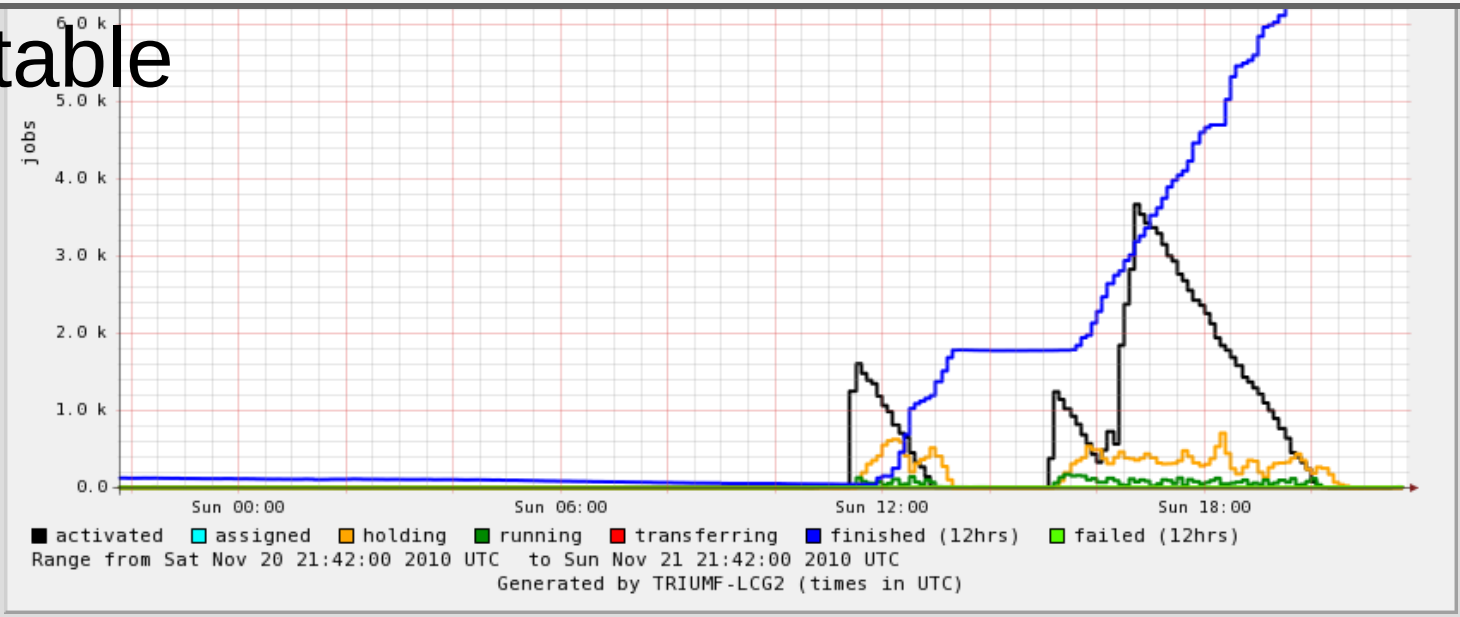
- Installed 21 Oct 2010
- Metadata traffic reduced
- Job failures reduced



Squid: load



- Load acceptable
- Proxy:
  - 5 years old
  - 2 Gig RAM
  - 2 cores



QMUL Tier-3

- Support Burden
  - Dear sysadmin, please install the latest ATLAS software...
  - Installs take time
  - Latency (sysadmins try not to work 24/7)
- CVMFS
  - Install once
  - Software available instantly
    - No latency
    - No sysadmin time

How do I install it?

Installation procedure

- *yum install fuse cvmfs cvmfs-init-scripts*

- Configuration:

```
cat /etc/cvmfs/default.local
CVMFS_REPOSITORIES=atlas,atlas-condb,lhcb,cms
CVMFS_CACHE_BASE=/scratch/lcg/cvmfs2
CVMFS_HTTP_PROXY="frontiercache.esc.qmul.ac.uk:3128"
CVMFS_QUOTA_LIMIT=15000
CVMFS_NFILES=65536
```

- Run
  - /sbin/chkconfig cvmfs on
  - /sbin/service cvmfs start
- https://twiki.cern.ch/twiki/bin/view/Atlas/Tier3CVMFS2SLC5

Old $VO_ATLAS_SW_DIR

```
$ ls -l /mnt/lustre_0/software/lcg_experimental_sw/sl5/atlas
total 44
drwxr-xr-x  3 atlassgm atlas 4096 Jan 25  2010 atlas-gcc
-rw-r--r--  1 atlassgm atlas  178 Feb  1  2010 AtlasSiteConfig.sh
-rw-r--r--  1 atlassgm atlas   94 Feb  1  2010 AtlasSiteConfig.sh.orig
drwxr-xr-x  4 atlassgm atlas 4096 Jul 15 12:56 cctools
lrwxrwxrwx  1 root     root    20 Oct 20 10:11 database ->
/opt/atlas/database/
drwxr-xr-x  6 atlassgm atlas 4096 Apr 23  2010 ddm
drwxr-xr-x  5 atlassgm atlas 4096 Nov 21 00:14 local
drwxr-xr-x  3 atlassgm atlas 4096 Jan 21  2010 prod
lrwxrwxrwx  1 root     root    39 Oct 20 10:40 software ->
/opt/atlas/software/i686-slc5-gcc43-opt
drwxr-xr-x 27 atlassgm atlas 4096 Oct 19 19:19 software.old
-rw-r--r--  1 atlassgm atlas 5314 Nov 21 01:47 tags
```

New VO_voname_SW_DIR

- VO_LHCB_SW_DIR=/cvmfs/lhcb.cern.ch
- VO_ATLAS_SW_DIR=/cvmfs/atlas.cern.ch
  - ATLAS_LOCAL_AREA=/path/to/shared_filesystem
  - Not easily set with YAIM

New VO_voname_SW_DIR

- VO_LHCB_SW_DIR=/cvmfs/lhcb.cern.ch
- VO_ATLAS_SW_DIR=/cvmfs/atlas.cern.ch
  - ATLAS_LOCAL_AREA=/path/to/shared_filesystem
  - Not easily set with YAIM

Manchester experience

- Installed last July to solve
  - Independent software areas problems in atlas
  - Load on storage caused by condb files
- LHCb also asked a number of times to run their reprocessing in Manchester
- Planned to take a week to migrate
  - Went through the docs for 2 days
  - LHCb worked straight away, after 12h Atlas was moved too with short validation
- Notes from the experience formed core docs followed by other UK sites

UK migration

- Pushed by Atlas last September (2011) we decided to move all sites before Xmas
  - Wrote a procedure for them to follow.
    - Included LHCb steps even though it was an atlas procedure
  - Opened tickets for sites to follow them one by one
  - Testing procedure:
    - 1 week of atlas analysis
    - short validation by the experiment
    - 3 sites were just tested with full validation by the experiment
  - Marked each of them in the status table at each stage
  - 13 sites were moved before Xmas ~1 a week
    - Last two T3 sites after they solved some local problems.

Major problems (solved)

- While on Atlas shift looked at some failures happening at CVMFS sites
  - Cured by some sites running cvmfs fsck every day
  - Others preferred to delay installations
- Debugged the problem using info from atlas monitoring and local logs
  - http://savannah.cern.ch/support/?122564
- Debugged also mount slowness particularly visible on older machines
  - http://savannah.cern.ch/bugs/?86349

UK Deployment

- Rolled out gradually over 1 year
- Atlas
  - All sites using CVMFS
- LHCb
  - Most sites using CVMFS
- CMS
  - Expected soon at Tier-1
- Other VOs
  - Not asked for, but easy to provide
- RAL (UK Tier-1) provides stratum1

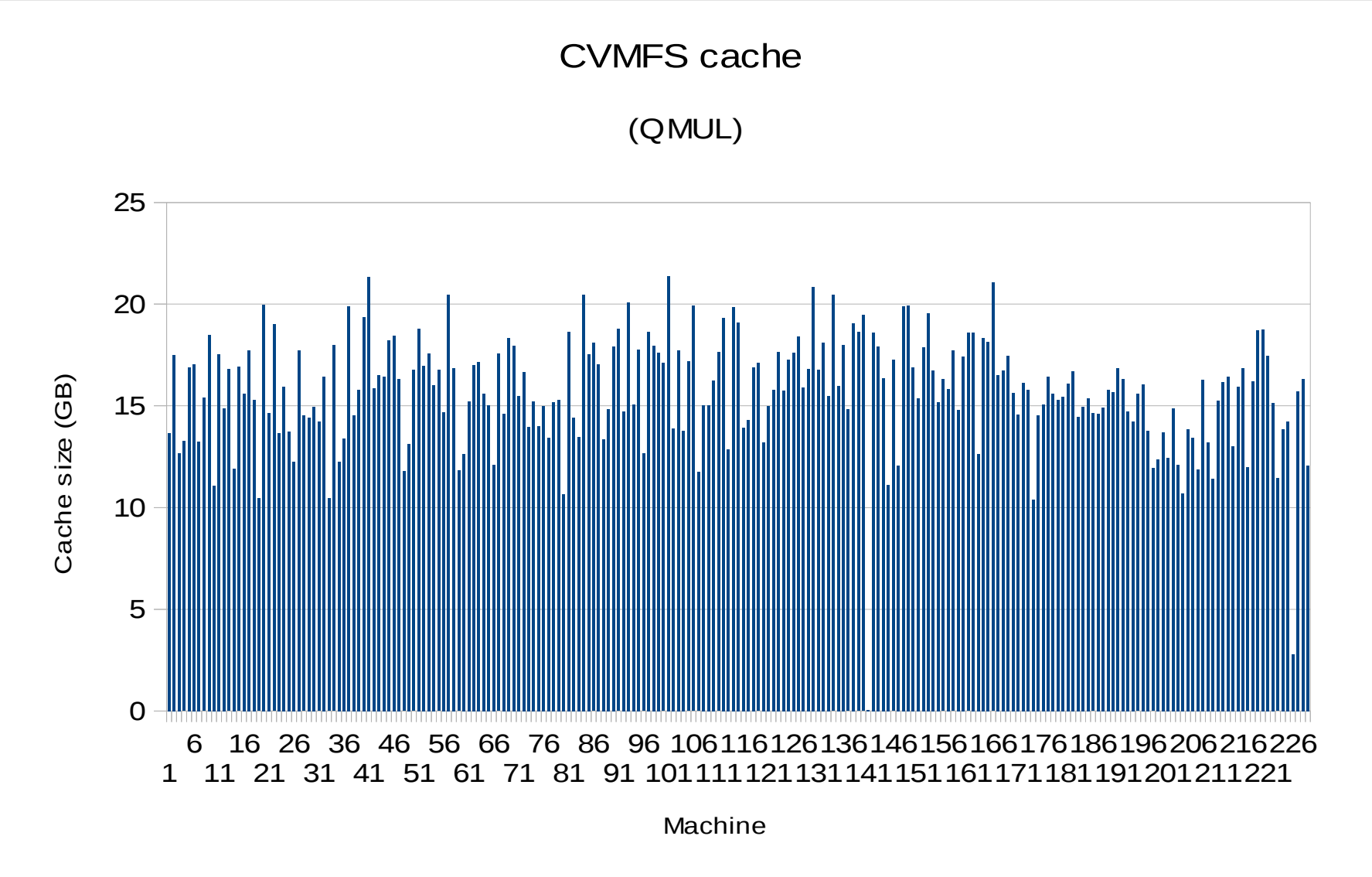
Sysadmin comments

- Good support
  - Fairly straightforward; after the initial slow start identifying the initial corruption issue, support has been responsive, rapid and effective.
  - Support has been very good; the few race conditions in the mounting process were found and removed.
- Easier to manage than NFS server at RAL Tier-1
- Requires local disk cache
  - One site concerned about interference with staging jobs

More sysadmin comments

- Monitoring
  - nagios monitoring is very useful to spot the problem and take some corrective actions
  - We found that previous versions failed fairly quietly. It left behind a stacktrace, that contained little that was of use to me, and sometimes hung the automount process
- Squid
  - Works fine with frontier squid
  - Failover recommended

Cache Size (GB)

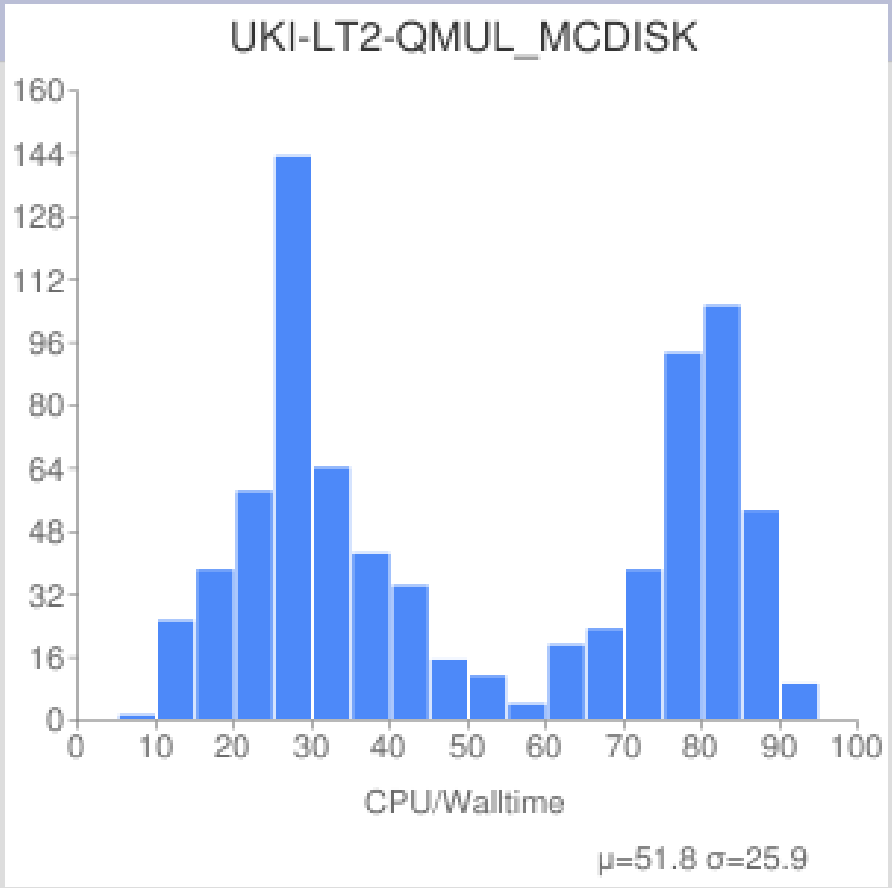


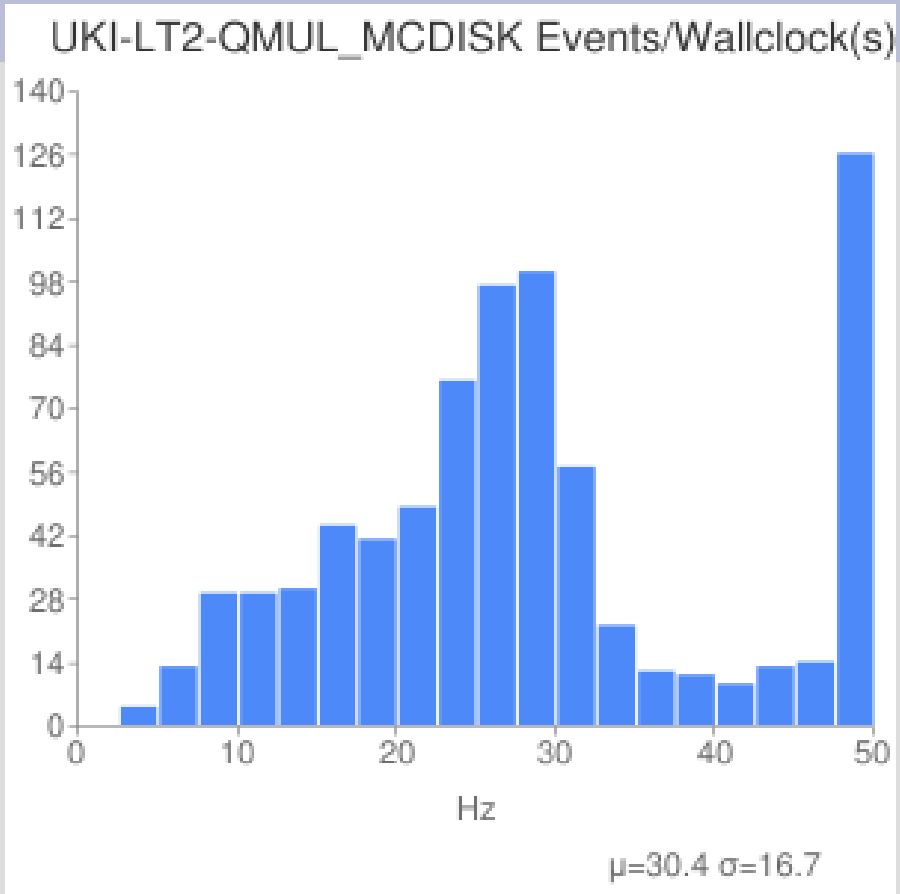
Conclusions

- Easy to install
- “Instant” software availability (Tier-3)
  - Validation job needed for Tier-2
- Performant
  - Dramatically improved QMUL's throughput
- Small disk cache (20 Gig)
- Requires reliable network connectivity
- Should sites upgrade?
  - Yes
- Non LHC Vos
  - May be a useful way to deploy software

Backup slides

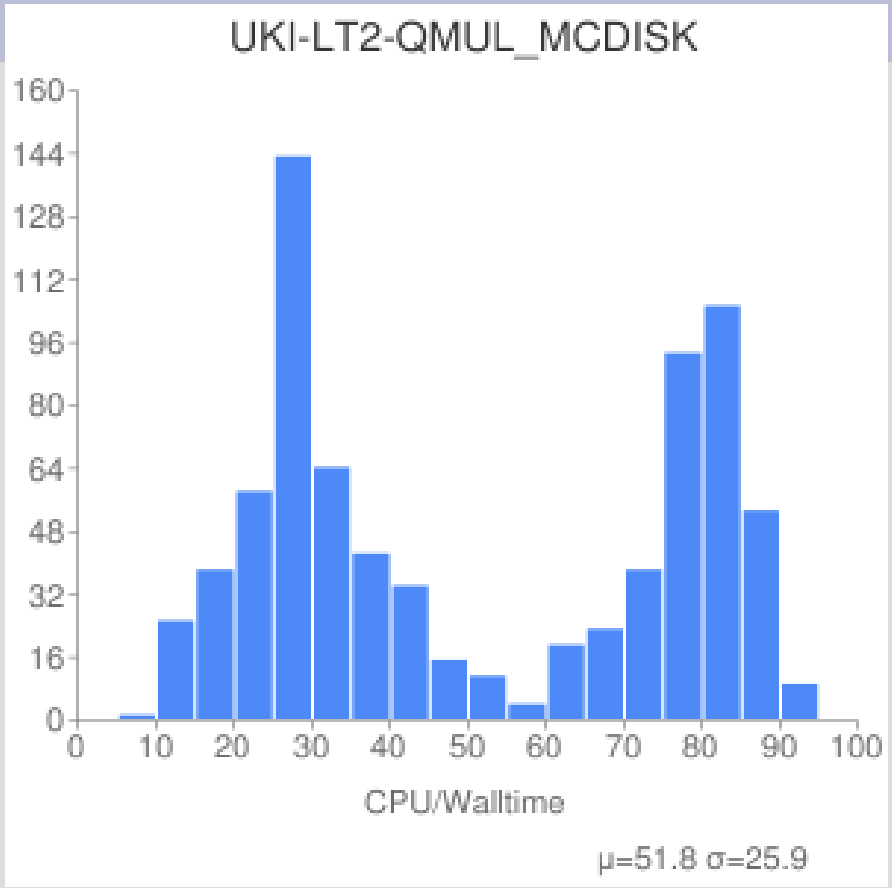
Results:Hammercloud: 10001578

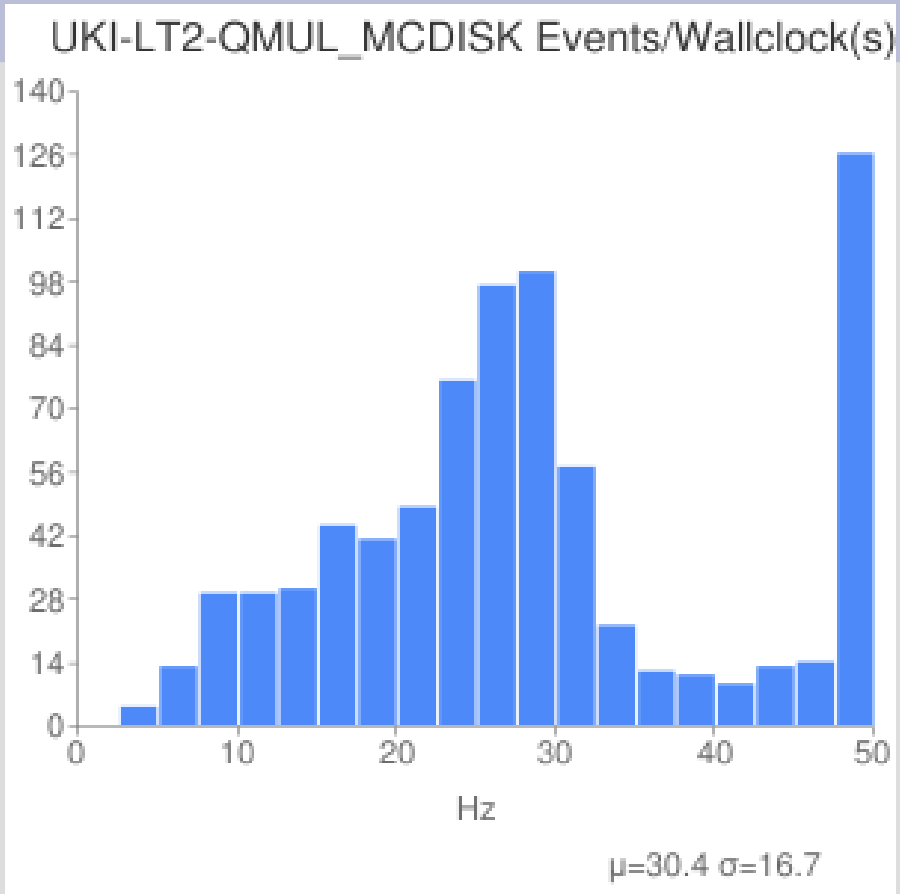




- 71 000 000 events (8h)
- Rack uplinks saturated – need bonded 10GigE.

Results:Hammercloud: 10001578





- 71 000 000 events (8h)
- Rack uplinks saturated – need bonded 10GigE.

Network Performance

- MDS load acceptable
- 10 GigE Storage node peaks at 5 Gig
- 10 Gig top of rack - bottleneck