Delivering Experiment Software to WLCG sites

A new approach using the CernVM Filesystem (cvmfs)

Ian Collier – RAL Tier 1

ian.collier@stfc.ac.uk

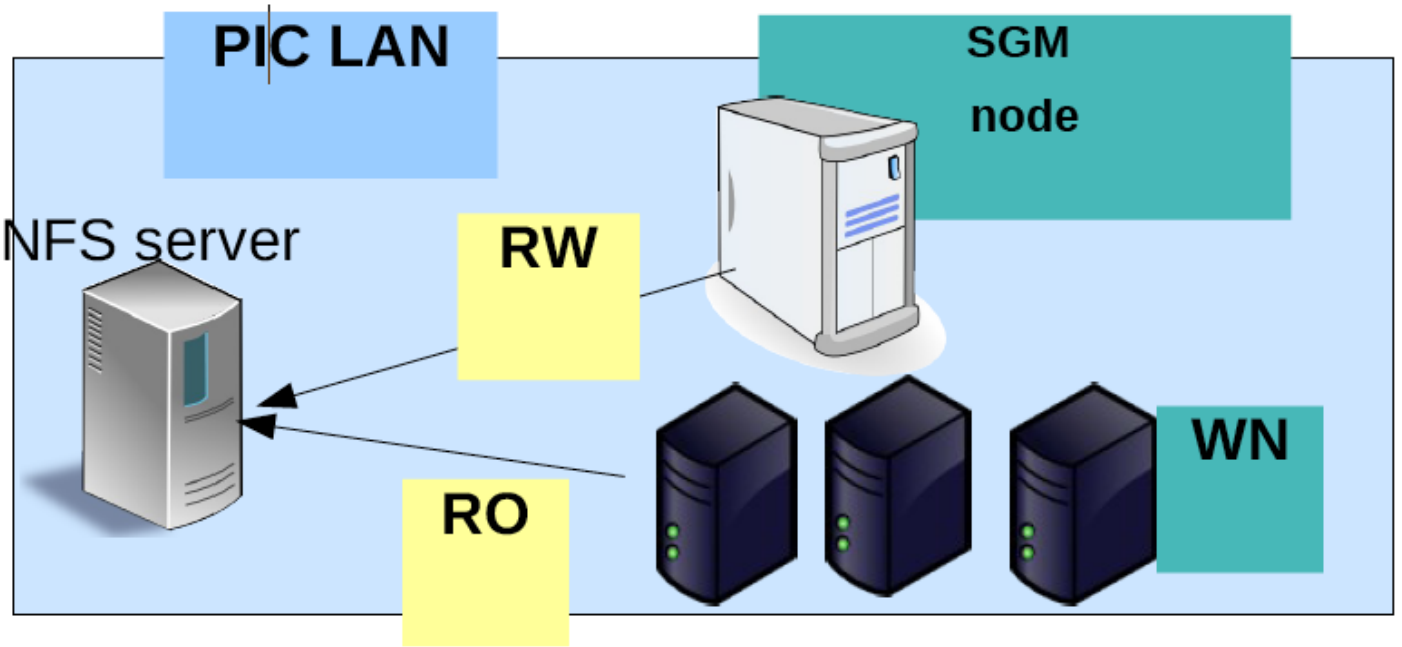
HEPSYSMAN November 22nd 2010, Birmingham

Contents

- What’s the problem?
- What is cvmfs - and why might it help?
- Experiences at PIC
- Experiences at RAL
- Future

So, what's the problem?

- Experiment application software (and conditions databases) is currently installed in a shared area of individual site (NFS, AFS…)
- The software is installed by jobs which run on a privileged node of the computing farm, these must be run at all sites



So, what's the problem? II

Issues observed with this model include:

NFS scalability issues

- at RAL we see BIG load issues, especially for Atlas
  - Upgrading the NFS server helped – but problem has not gone away
- at PIC they see huge loads for LHCb

Shared area sometimes not reachable (stale mounts on the WN, or the NFS server is too busy to respond)

Software installation in many Grid sites is a tough task (job failures, resubmission, tag publication...)

Space on performant NFS servers is expensive

- if VOs want to install new releases and keep the old ones they have to ask for quota increases

In the three months to September there were 33 GGUS tickets related to shared software area issues for LHCb

What might help?

- A caching file system would be nice – some sites use AFS but install jobs seem still to be very problematic
- A read only file system would be good
- A system optimised for many duplicated files would be nice

What, exactly, is cvmfs?

It is *not,* really, anything to do with virtualisation

A client-server file system

- Originally developed to deliver VO software distributions to (cernvm) virtual machines in a fast, scalable, and reliable way.

Implemented as a FUSE module

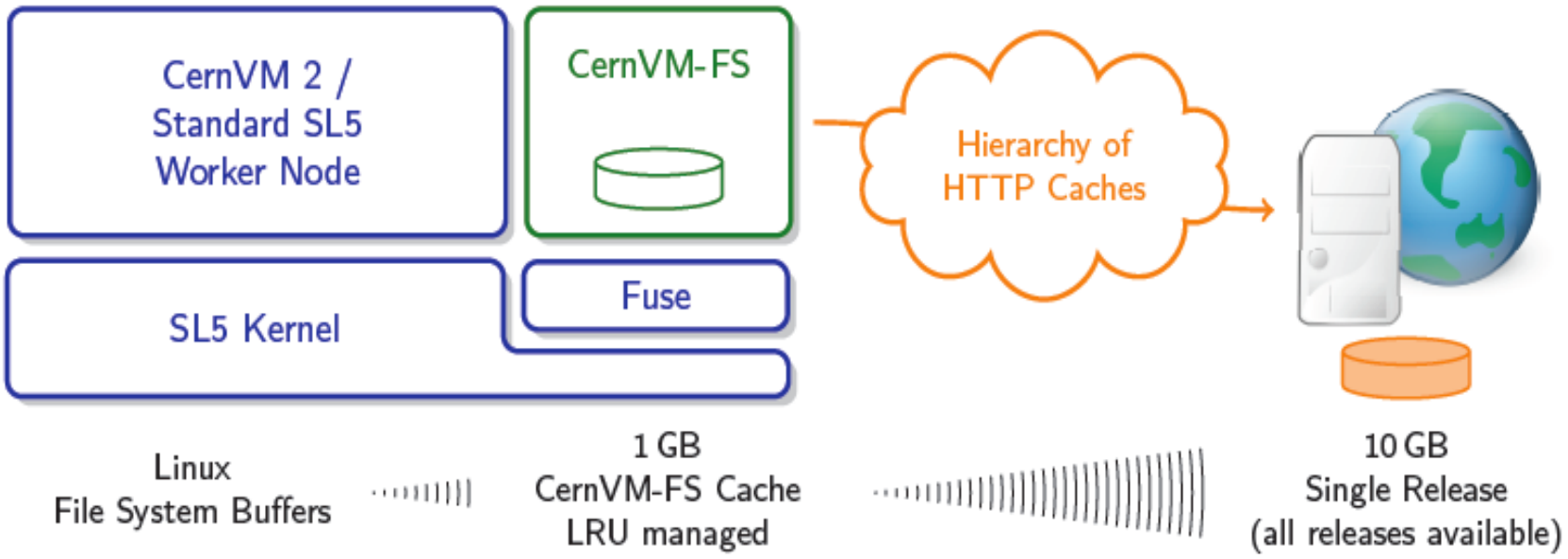
- Makes a specially prepared directory tree stored on a web server look like a local read-only file system on the local (virtual or physical) machine.

Uses only outgoing HTTP connections

- Avoids most of the firewall issues of other network file systems.

Transfers data file by file on demand, verifying the content by SHA1 keys.

What is cvmfs? - Software Distribution



Essential Properties:

· Distribution of read-only binaries

· Files and file meta data are downloaded on demand and cached locally

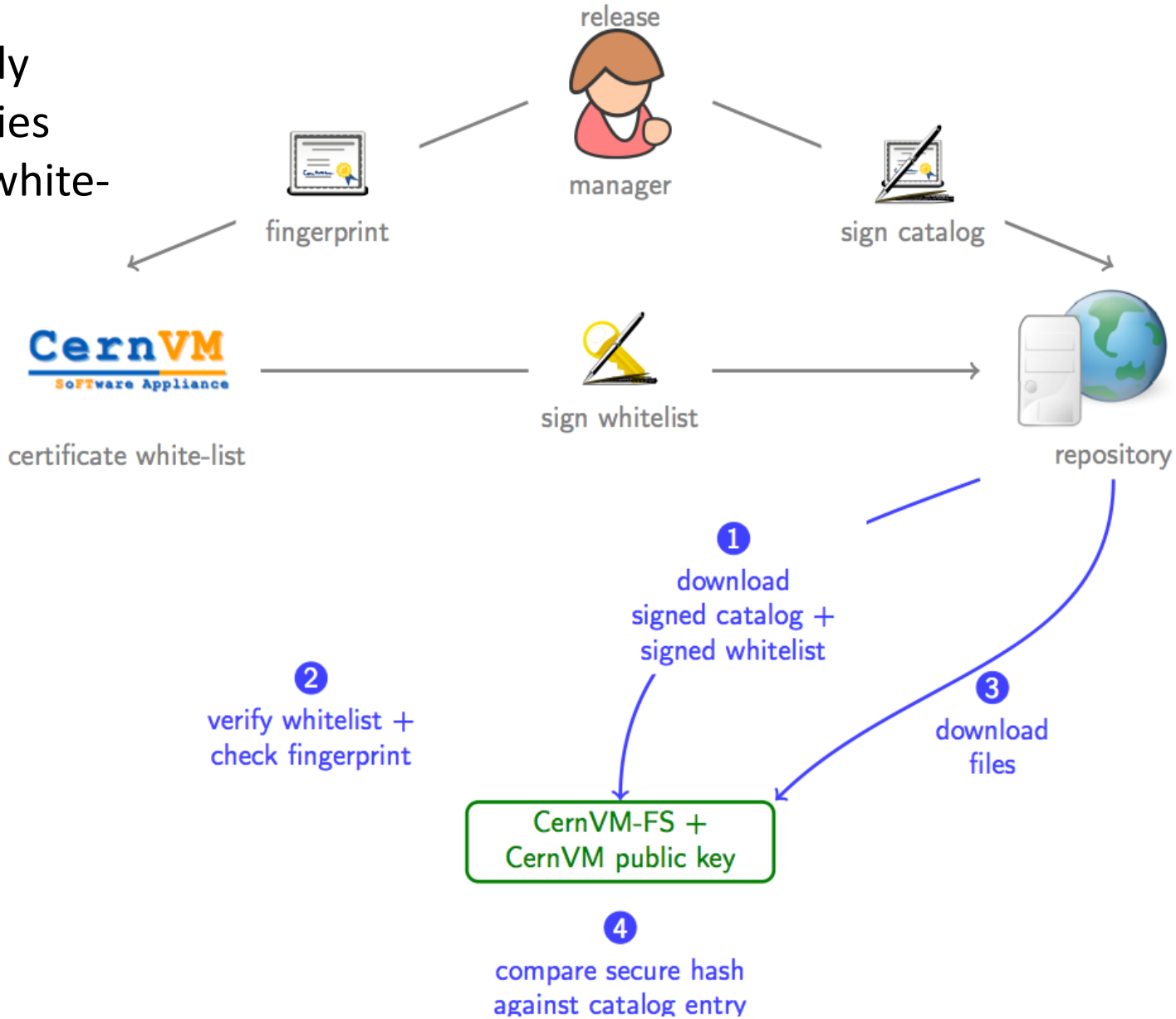
· Intermediate squids reduce network traffic further

· File based deduplication – a side effect of the hashing

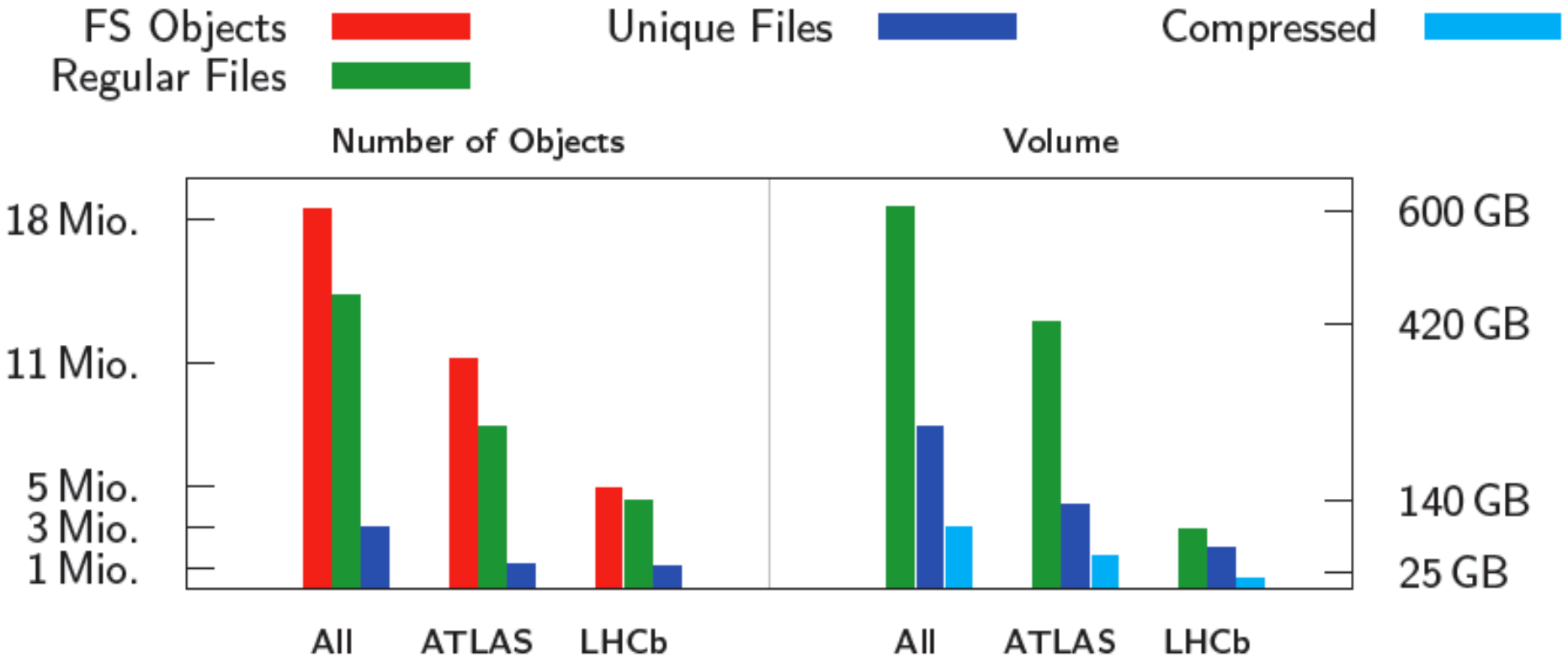
· Self-contained (e. g. /opt/atlas), does not interfere with base system

What is cvmfs? – Integrity & authenticity

Principle: Digitally signed repositories with certificate white-list



What is cvmfs? – Repository statistics



Note how small the bars for unique files are

- both in numbers and by volume

The hashing process identifies duplicate files and once they are cached they are never transferred twice

Tests at PIC

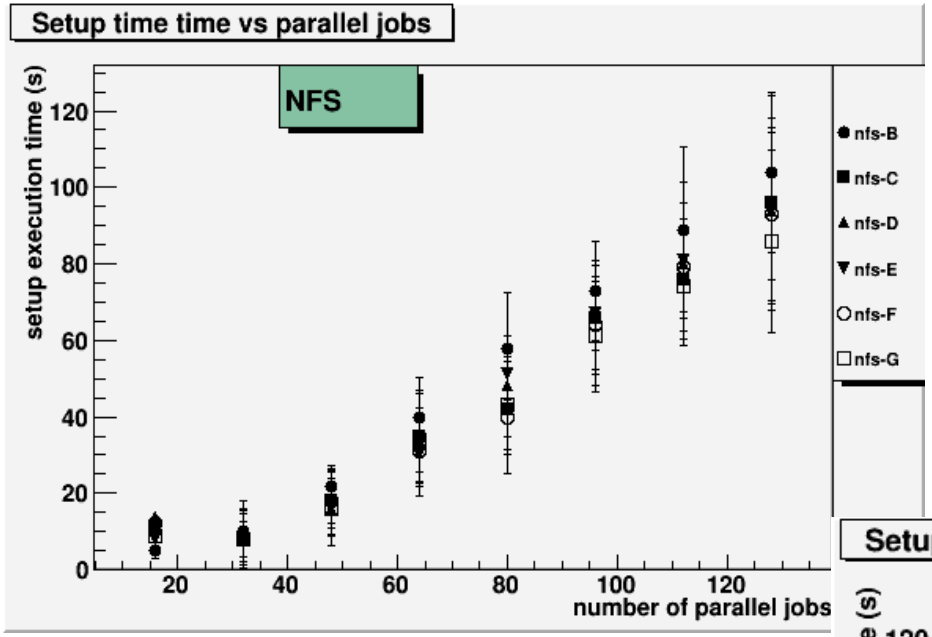
- Set out to compare performance at WN in detail
- Metrics measured

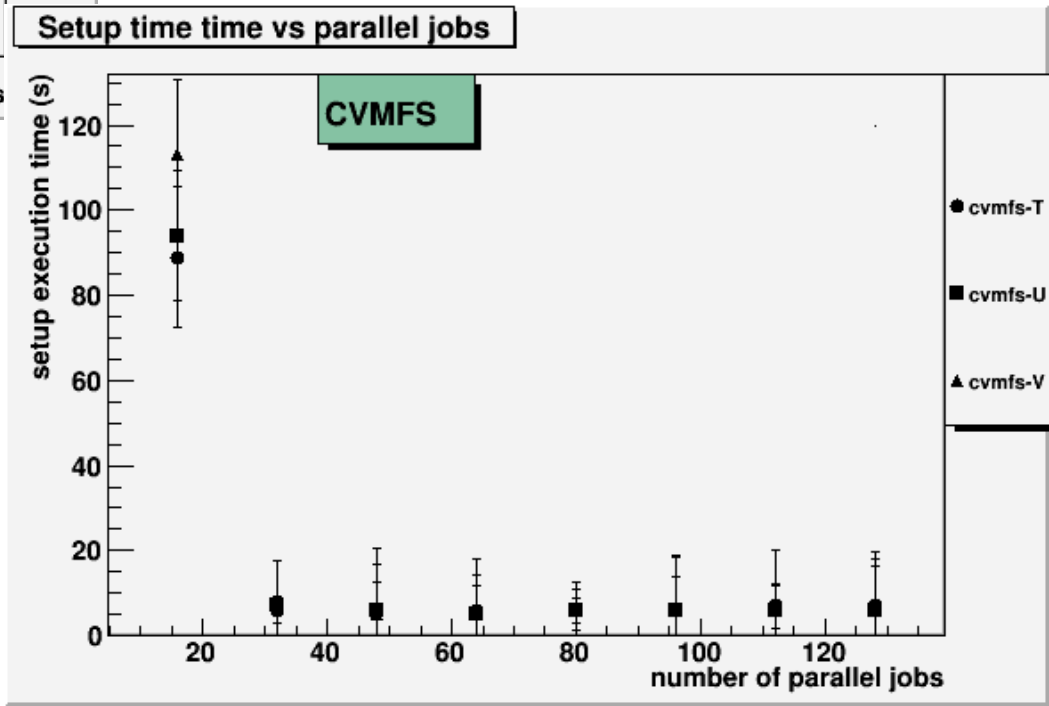
  Execution time for SetupProject - the most demanding phase of the job for the software area (huge amount of stat() and open() calls)

  Execution time for DaVinci

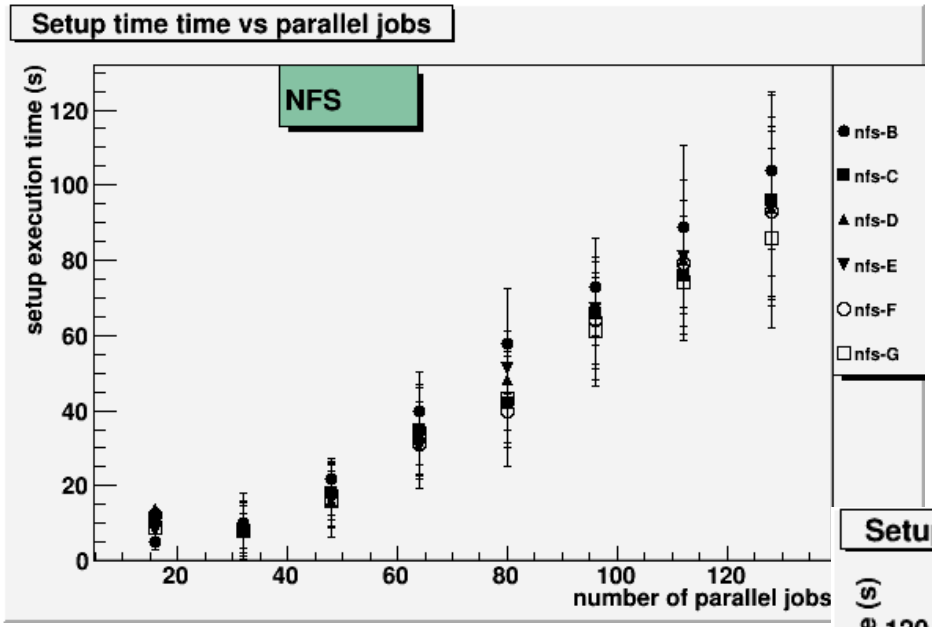
  Dependence on the number of concurrent jobs

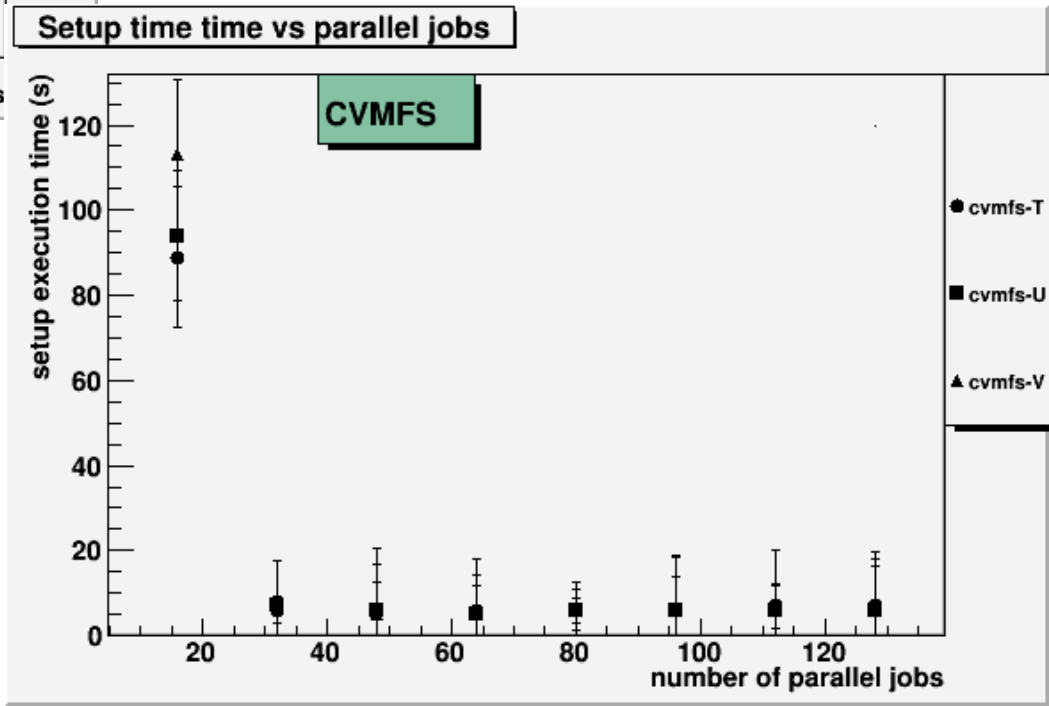
Tests at PIC – setup time





Tests at PIC – setup time





Tests at PIC – local cache size

LHCb:

One version of DaVinci (analysis package): the software takes 300 MB
+ CernVMFS catalogue overhead, total space: 900 MB
The catalog contains file metadata for all LHCb releases
Download once (for every new release) and then keep in cache
Each additional version of DaVinci exceuted adds 100 MB

ATLAS:

One release of Athena: 224MB of SW + catalog files: 375MB
The overhead is less since the catalog has been more optimised for ATLAS software structure

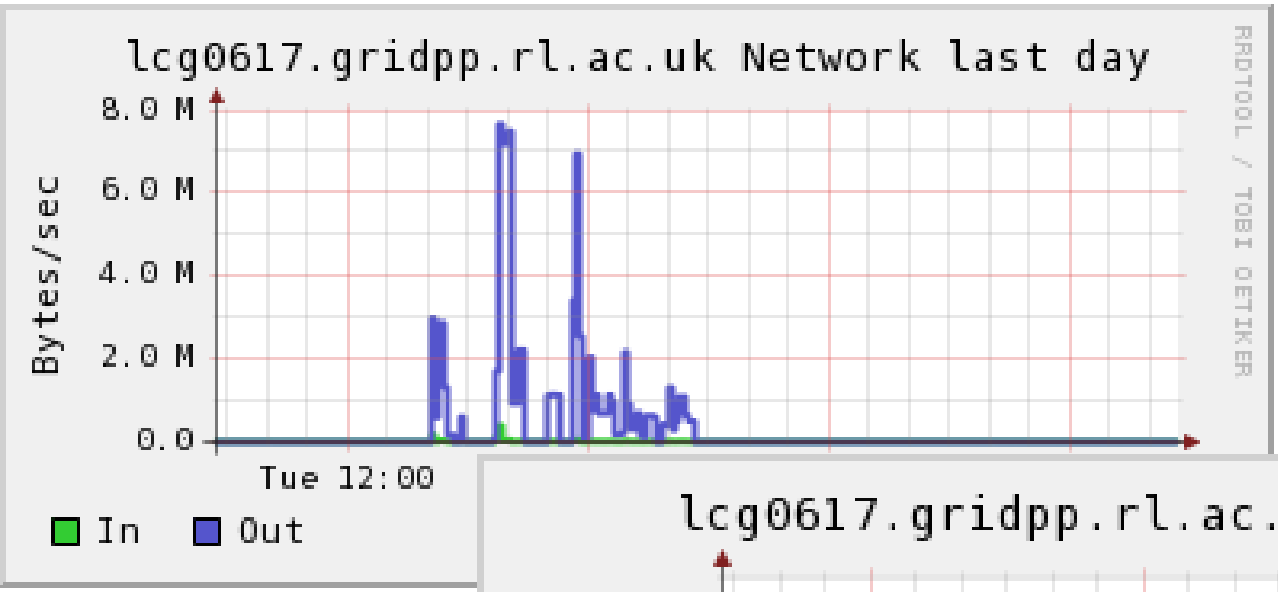
Tests at RAL

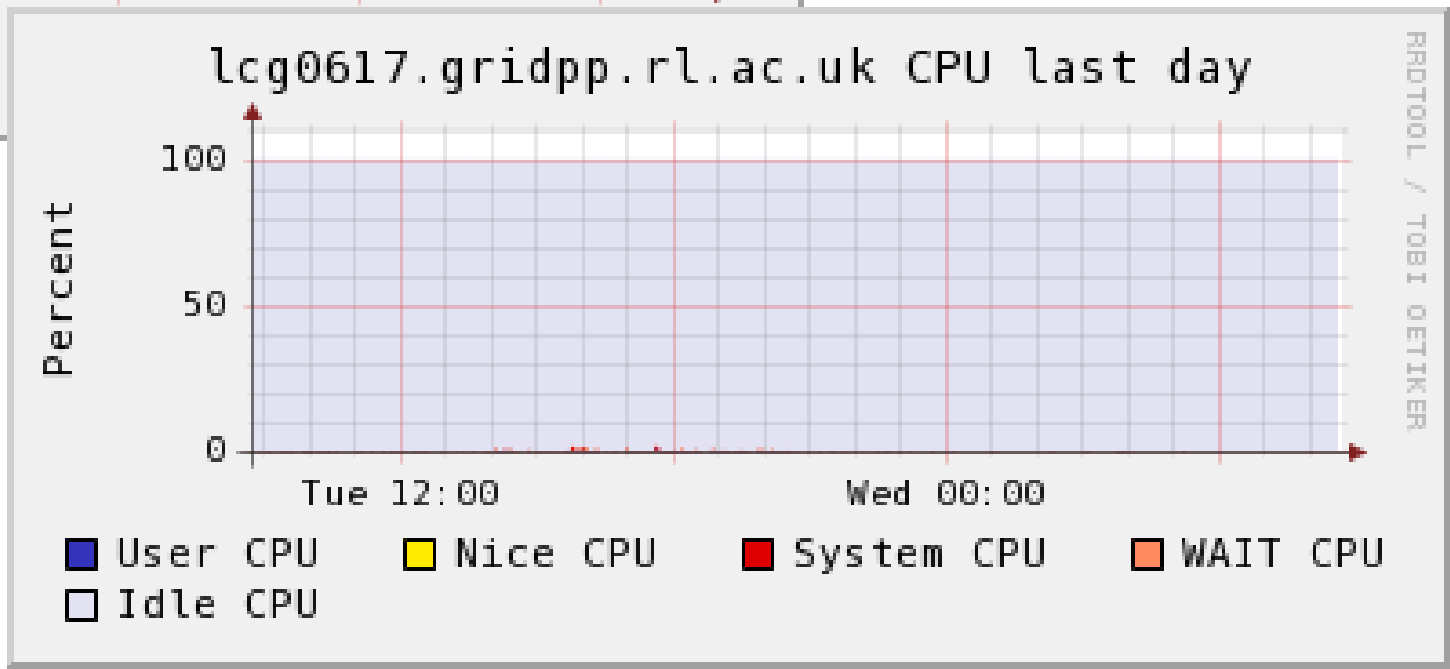
More qualitative

- Interested in scalability for servers
  - Focus on how well it might replace overloaded NFS server
- Have not examined in such detail at what happens at the client
- Have confirmed that we can run through 10-20000 jobs with no problems

Tests at RAL – Atlas Monte Carlo

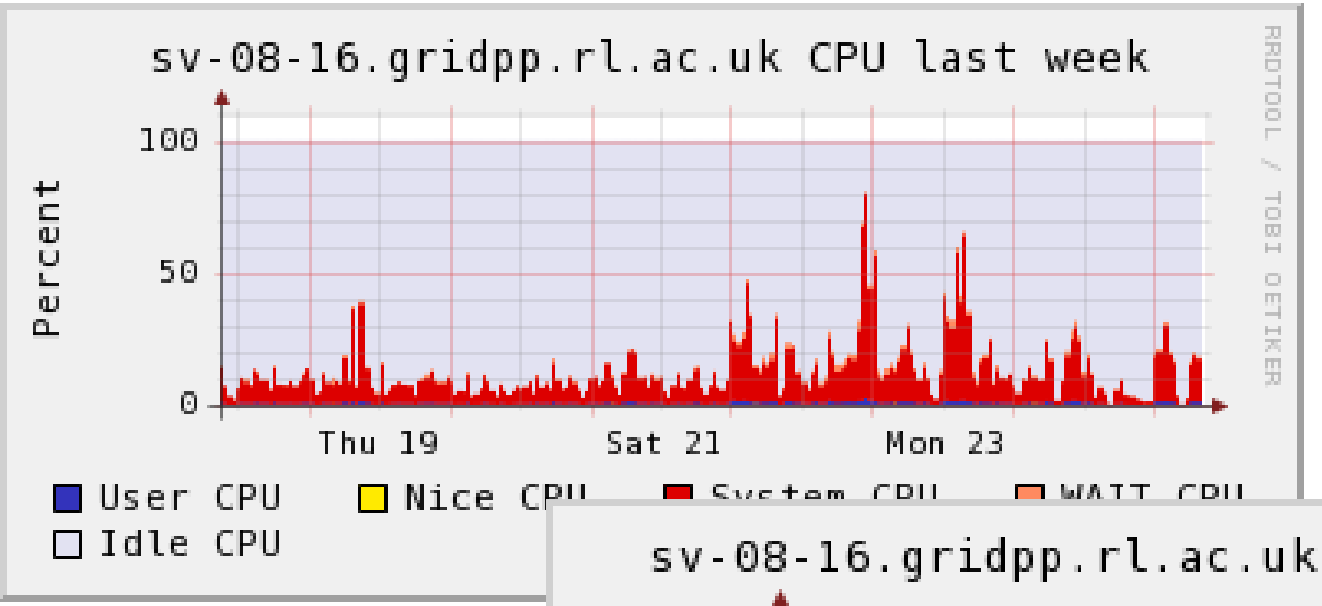
Early test with 800 or so jobs – the squid barely missed a beat

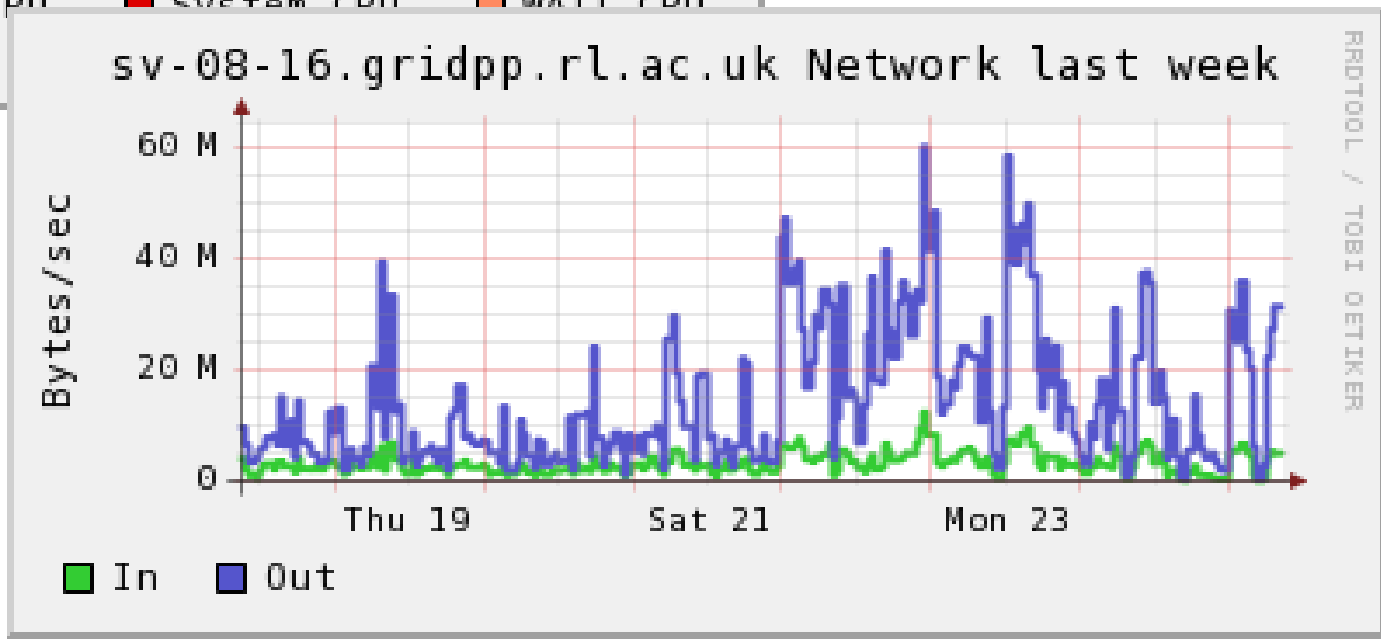




Tests at RAL – Atlas Monte Carlo

In the same period the NFS Atlas SW server – with 1500 or so jobs running





Tests at RAL – Atlas Hammercloud tests

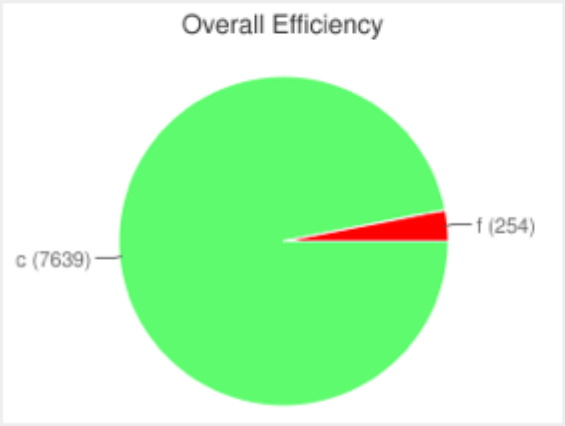
Average over several sites

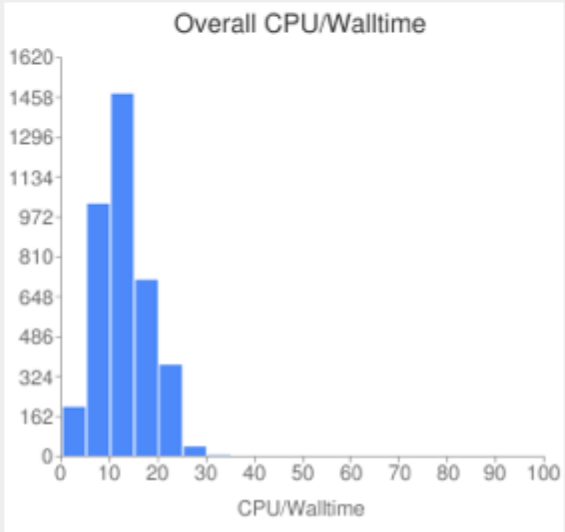
Summary

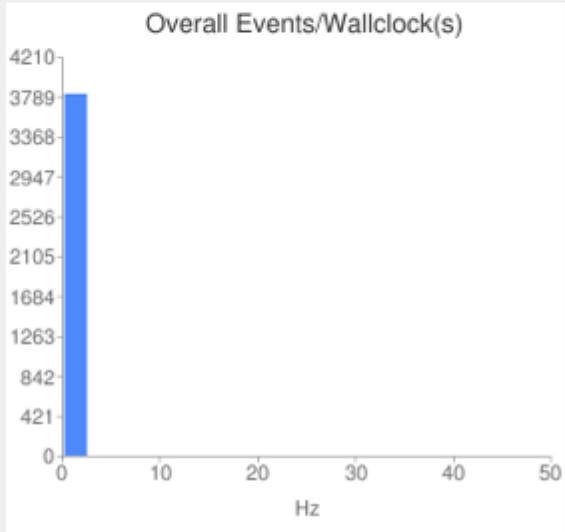
state	id	host	clouds	start time (CET)	end time (CET)	total jobs
completed	10001006	voatlas49	US,DE_PANDA,FR_PANDA,8 more...	2010-09-05 14:15:01	2010-09-06 14:15:11	8064

Input type: PANDA
Output DS: user.elmsheus.hc.10001006.*
Input DS Patterns: mc09_7TeV*merge.AOD*r1306*
Ganga Job Template: /data/hammercloud/atlas/inputfiles/15.6.9/1569_Panda.tpl
Athena User Area: /data/hammercloud/atlas/inputfiles/15.6.9/UserAnalysis_v1569.tar.gz
Athena Option file: /data/hammercloud/atlas/inputfiles/15.6.9/AnalysisSkeleton_topOptions_v1569.py
Test Template: 11 (functional) - UA 15.6.9 Panda

View Test Directory (for debugging)







Tests at RAL – Atlas Hammercloud tests

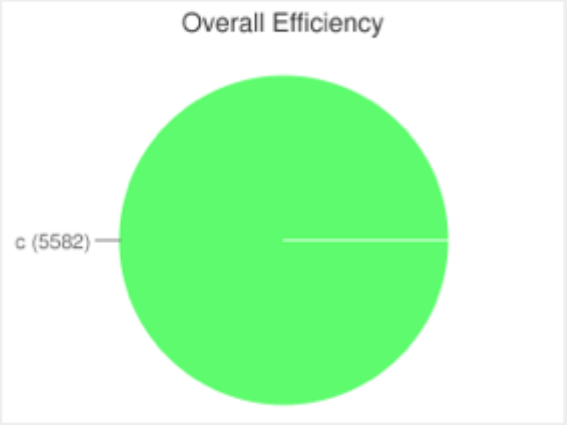
6150 jobs at RAL

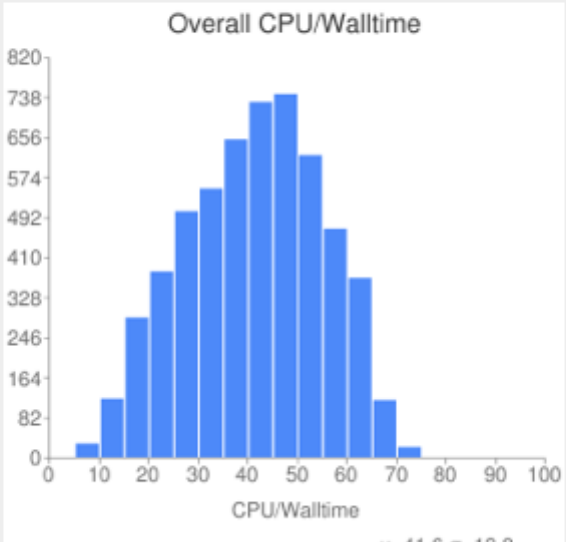
Summary

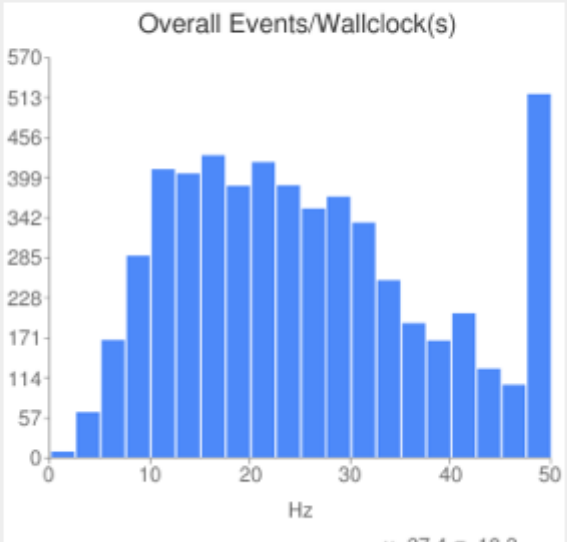
state	id	host	clouds	start time (CET)	end time (CET)	total jobs
completed	10001581	voatlas49	UK_PANDA	2010-10-25 18:00:00	2010-10-26 18:00:20	6150

Input type: PANDA
Output DS: user.elmsheus.hc.10001581.*
Input DS Patterns: mc09_7TeV*merge.AOD*r1306*
Ganga Job Template: /data/hammercloud/atlas/inputfiles/muon1590/muon1590_panda.tpl
Athena User Area: /data/hammercloud/atlas/inputfiles/muon1590/MuonTriggerAnalysis_v1590.tar.gz
Athena Option file: /data/hammercloud/atlas/inputfiles/muon1590/MuonTriggerAnalysis.py
Test Template: 50 (stress) - Muon 15.9.0 PANDA default data-access

View Test Directory (for debugging)



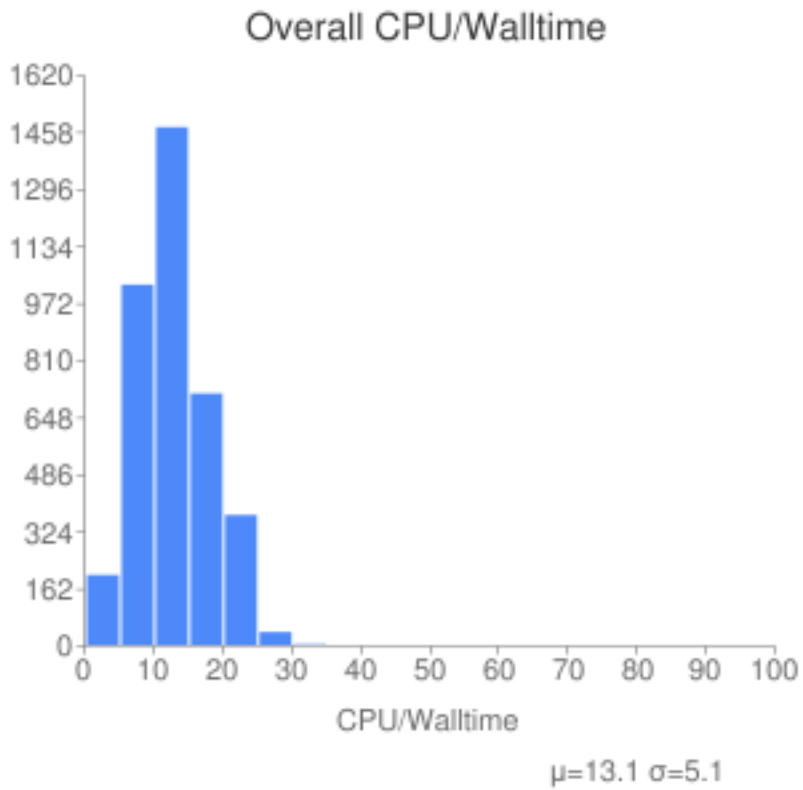




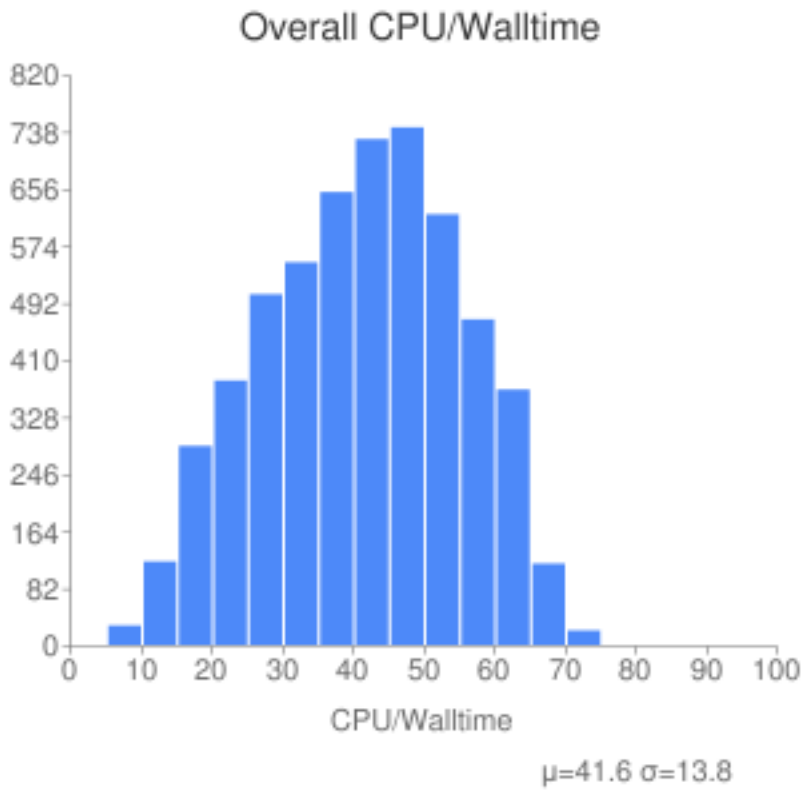
Tests at RAL – Atlas Hammercloud tests

Lets look in more detail –
CPU/Walltime

Average over several sites



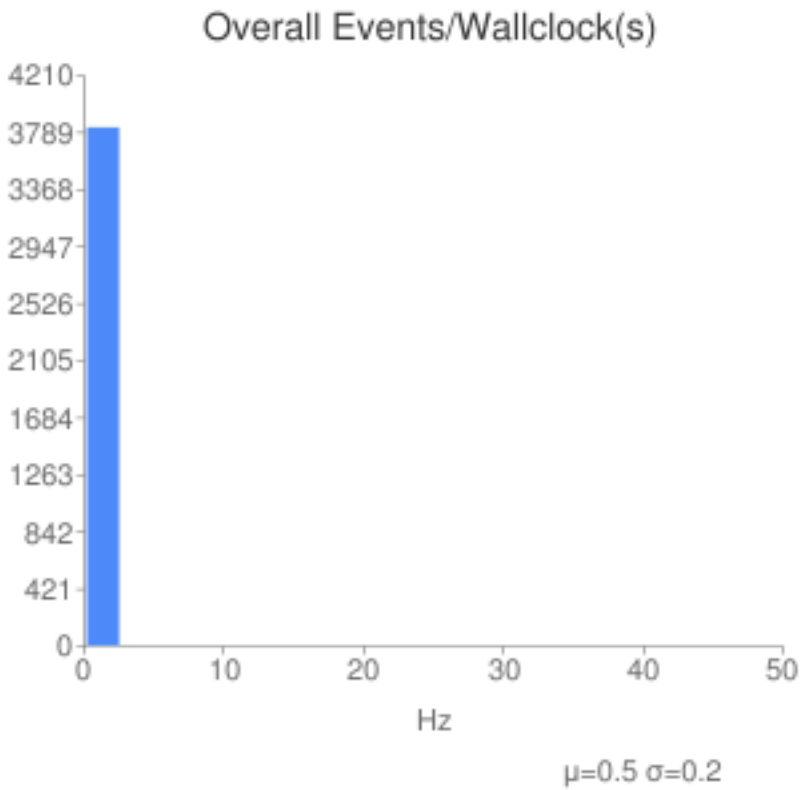
RAL



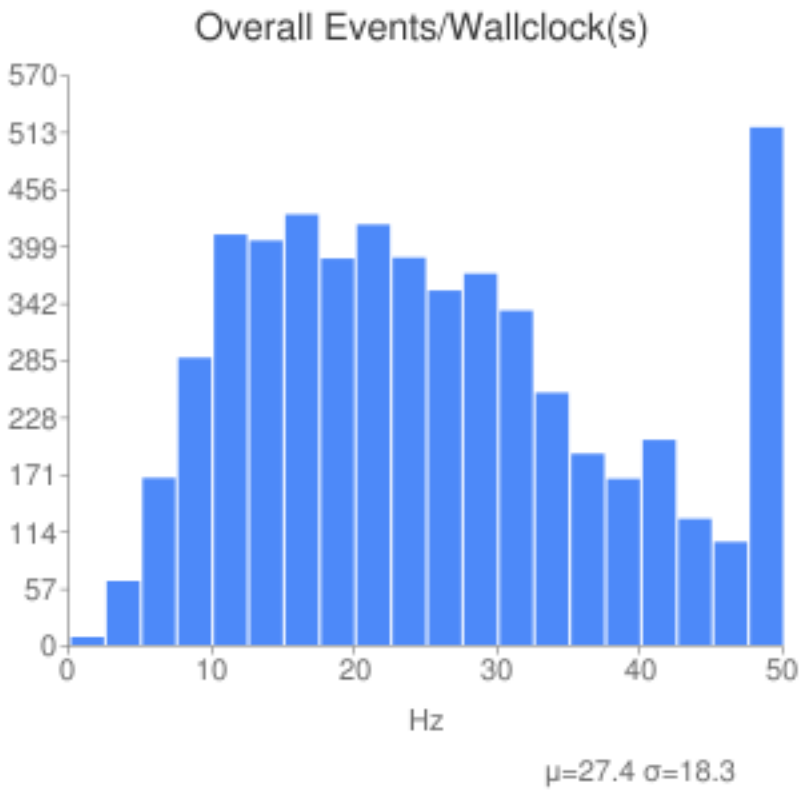
Tests at RAL – Atlas Hammercloud tests

Lets look in more detail –
Events/Wallclock

Average over several sites

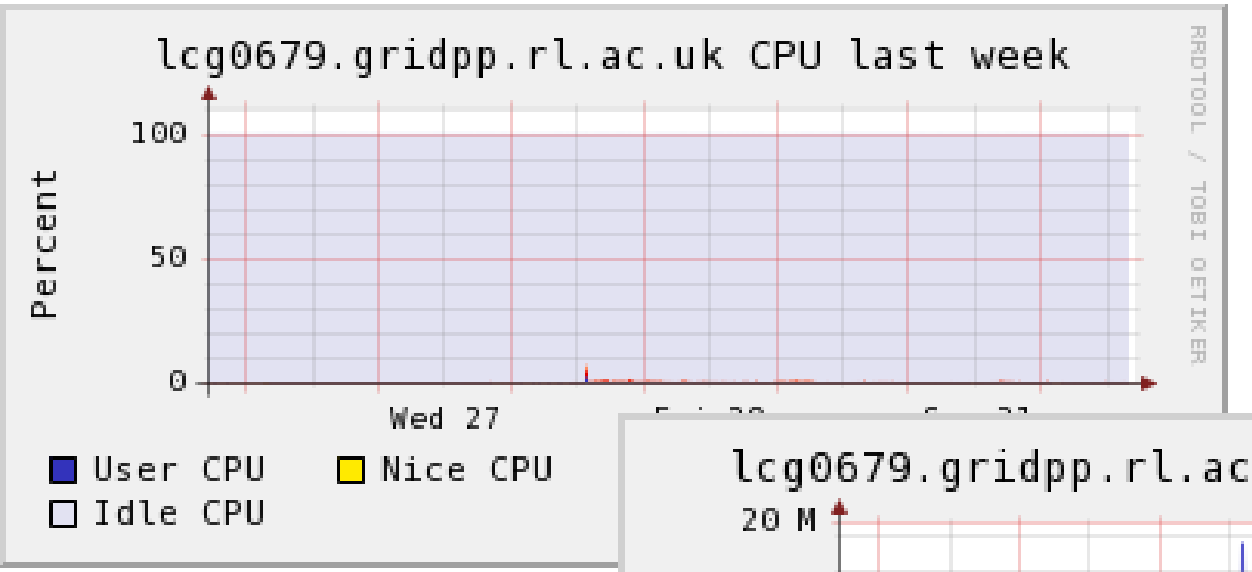


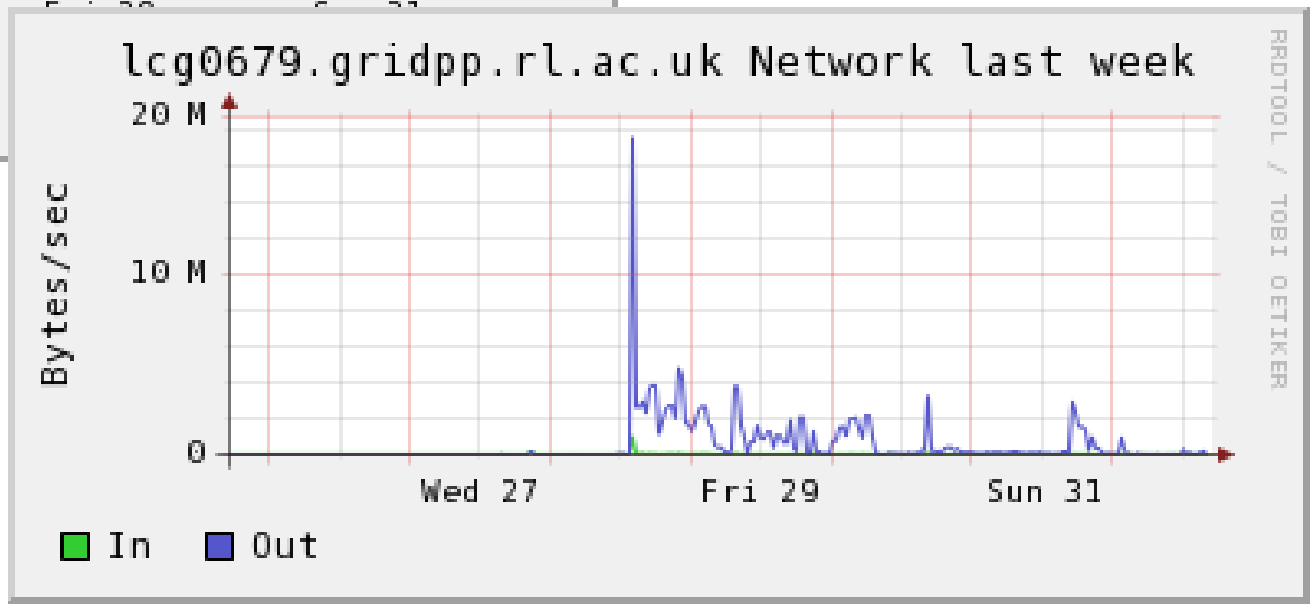
RAL



Tests at RAL – Atlas Monte Carlo

Since reconfiguring cvmfs and starting analysis jobs– the squid again happy
Worth noting that even at the point the network is busiest – cpu is not -





Tests at RAL – Load on Squids

Not quite negligible - but very manageable
For initial tests squid was running on the retired Atlas software server (5 year old WN)
    Separate from other frontier squids

For resilience (rather than load) we have added a second squid – client randomly mounts one or the other
    - failover appears transparent

Starting to accept Atlas user analysis jobs which will all use cvmfs - on an experimental basis
    Will learn much more
    Will use wider range of releases – should stress chaches more than current production and test jobs
Over last weekend removed limits – squids still happy

Current State of CernVM-FS

Network file system designed for software repositories
Serves 600 GB and 18.5 Million files and directories
Revision control based on catalog snapshots
Outperforms NFS and AFS
Warm cache speed comparable to local file system
Scalable infrastructure
Integrated with automount/autofs
Delivered as rpm/yum package – and as part of CernVM
Atlas, LHCb, Alice, CMS… all supported

Active Developments

Failover-Mirror of the source repositories
    (CernVM-FS already supports automatic host failover)
Extend to conditions databases
Service to be supported by Cern IT (http://sls.cern.ch/sls/service.php?id=cvmfs)
Security audit – in progress
Client submitted for inclusion with SL

Summary

- Good for sites
  - Performance at client is better
  - Squid very easy to set up and maintain – and very scalable
  - Much less network traffic
- Good for VOs
  - VOs can install each release once – at CERN
    - No more local install jobs
  - Potentially useful for hot files to

Acknowledgements & Links

Thanks to

- Jakob Blomer (jakob.blomer@cern.ch) who developed cvmfs
- Elisa Lanciotti (elisa.lanciotti@cern.ch)who carried out the tests at PIC
- Alastair Dewhurst & Rod Walker who've been running the Atlas tests at RAL

Links

PIC Tests:
https://twiki.cern.ch/twiki/bin/view/Main/ElisaLanciottiWorkCVMFSTests
Elisa's talk at September 2010 GDB
RAL cvmfs squids:
http://ganglia.gridpp.rl.ac.uk/ganglia/?r=day&c=Services_Core&h=lcg0679.gridpp.rl.ac.uk
http://ganglia.gridpp.rl.ac.uk/ganglia/?r=day&c=Services_Core&h=lcg0617.gridpp.rl.ac.uk
CVMFS Downloads:
https://cernvm.cern.ch/project/trac/cernvm/downloads
CVMFS Technical Paper:
https://cernvm.cern.ch/project/trac/cernvm/export/1693/cvmfs-tr/cvmfstech.preview.pdf
Jakob's talk from CHEP:
http://117.103.105.177/MaKaC/contributionDisplay.py?contribId=39&sessionId=111&confId=3