

**GridPP**

UK Computing for Particle Physics

# Cern VM-FS Production Service

Hepsysman, RAL

30 June 2011

Ian Collier, STFC-RAL

[ian.collier@stfc.ac.uk](mailto:ian.collier@stfc.ac.uk)



Science & Technology Facilities Council  
**e-Science**



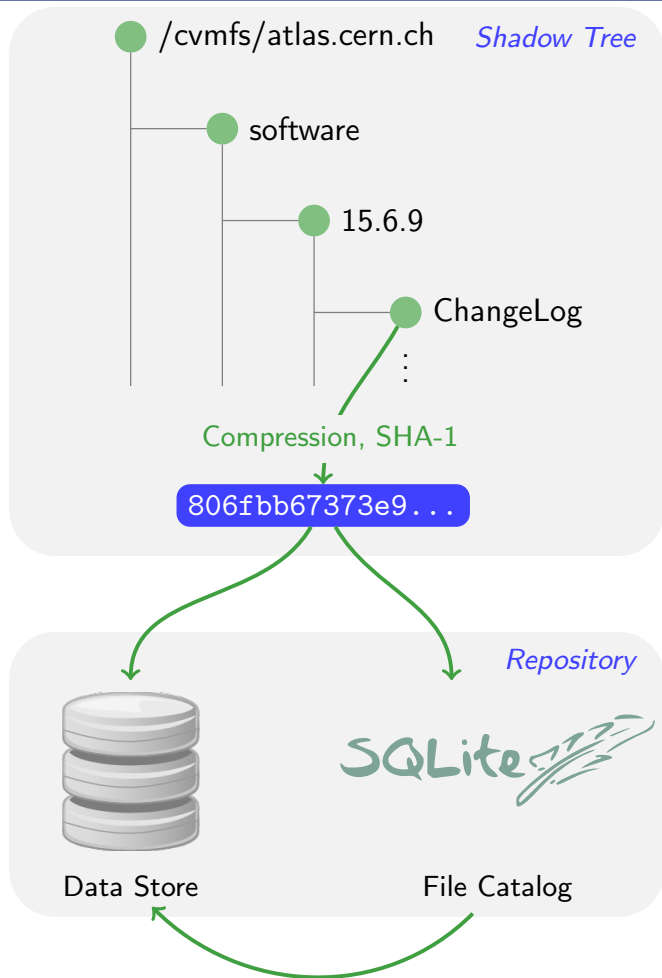
- CernVM-FS Introduction
  - Concepts
  - Performance
- CernVM-FS Service Status
- CernVM-FS Site deployment
  - Client Config
- CernVM-FS In Use
- Summary

30th June 2011



## CVMFS - What's the Big Idea?

- Virtual software installation by means of an http based file system
- Based on standard components
  - http, fuse, sha1, squid, sqllite
- Caching
- Read-only
- Very scalable



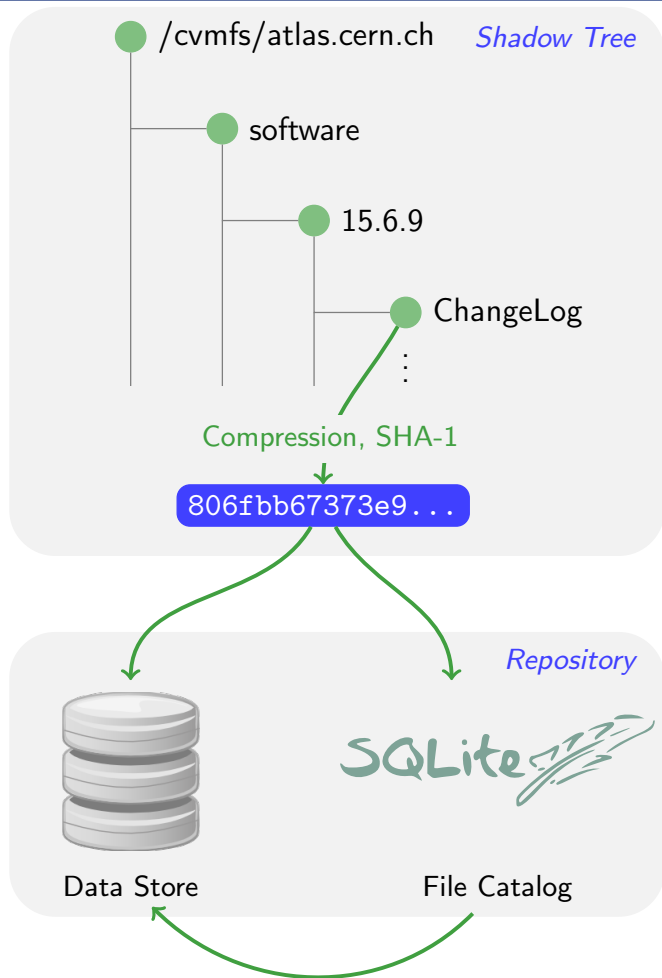
## Data Store

- Compressed Chunks (Files)
- Eliminates Duplicates
- Never Deletes

## File Catalog

- Directory Structure
- Symlinks
- SHA1 of Regular Files
- Digitally Signed
- Time to Live
- Nested Catalogs

30th June 2011



## Data Store

- Compressed Chunks (Files)
- Eliminates Duplicates
- Never Deletes

## File Catalog

- Directory Structure
- Symlinks
- SHA1 of Regular Files
- Digitally Signed
- Time to Live
- Nested Catalogs

⇒ Immutable Files, trivial to check for corruption

- CVMFS client:
  - Provides a `/cvmfs/` filesystem area.
  - Files served from a web-server.
  - File accesses are intercepted by fuse.
  - Metadata operations (`ls`, `cd`, ...) work on user space fs (fuse) built with signed sqlite database - downloaded from web-server.
  - File operations (`cat`, `open`, ...) trigger download of file from web server.
- Lots of caching:
  - On batch workers between jobs.
  - On intermediate squid servers.
- Everything transparent to user

- **Traditional method:**

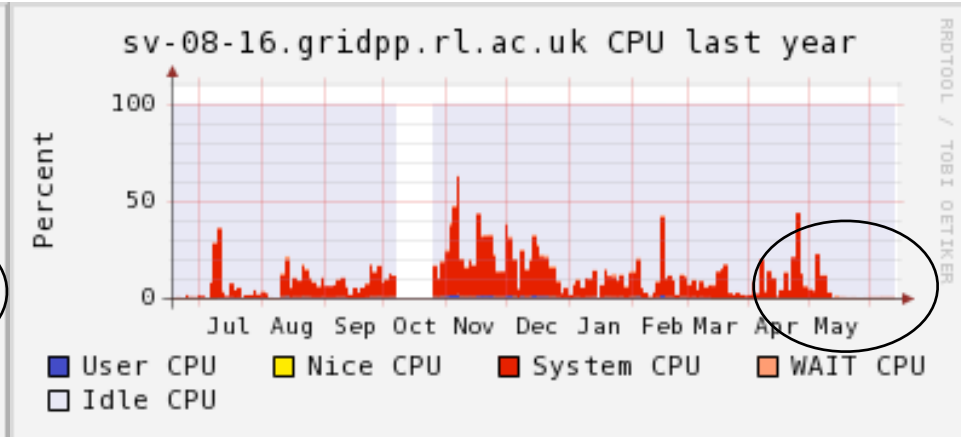
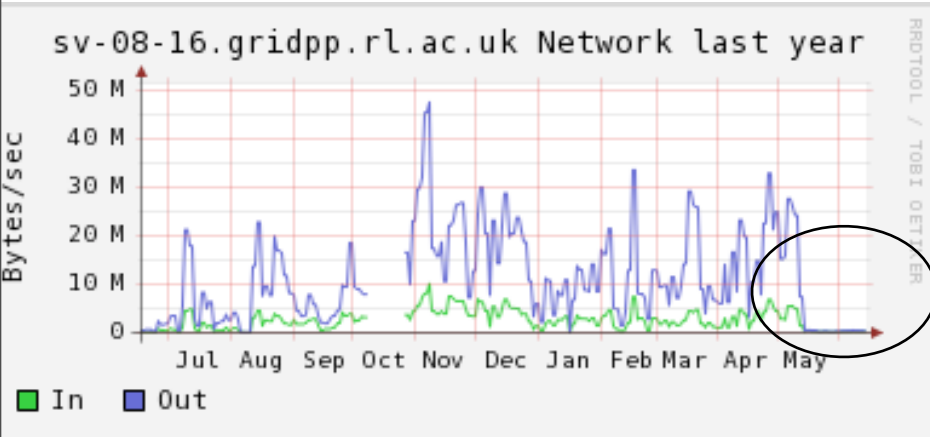
- Run <VO>sgm job via grid
- Write files within job to some shared storage.
- Validate software.
- Publish tag in BDII.
- Process has to be repeated at every site.
  - Process has to be debugged at every site.

## CVMFS method:

- Install once (at Stratum 0)
  - Files appear everywhere across WLCG.
- This can be many days faster.
- Hopefully less variation across sites.
  - Common path `/cvmfs/...` (c.f. `/afs/`)
    - ... some legacy use of `/opt/atlas` etc. being ironed out now
- Same install bugs everywhere - fix once.
  - e.g LHCb have had problems with CMT usage on `/cvmfs/` but at least it's everywhere.
- Some sites are struggling to provide scaled NFS/AFS.
  - squid scales out easily - and cheaply

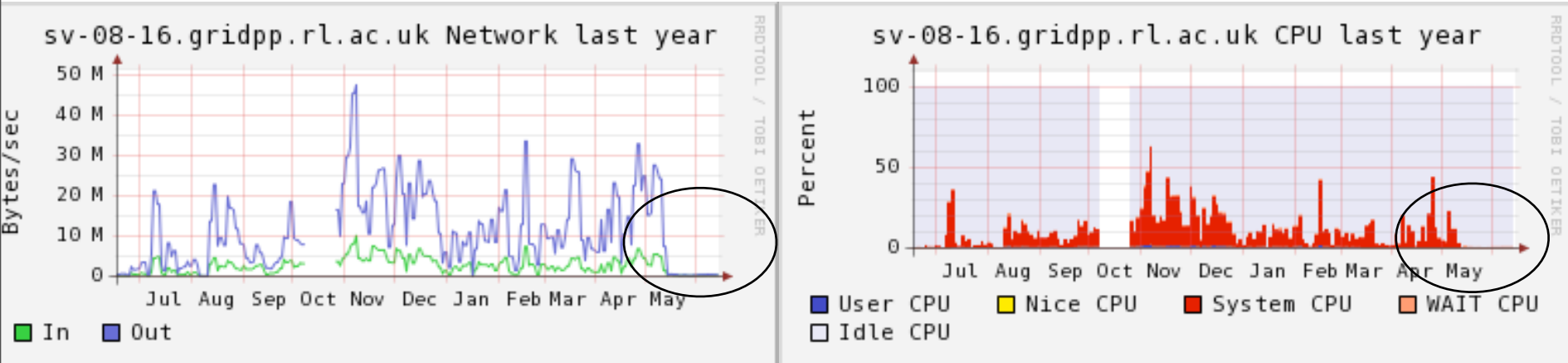


- NFS Atlas SW Server Loads - switched Atlas to CVMFS in May



30th June 2011

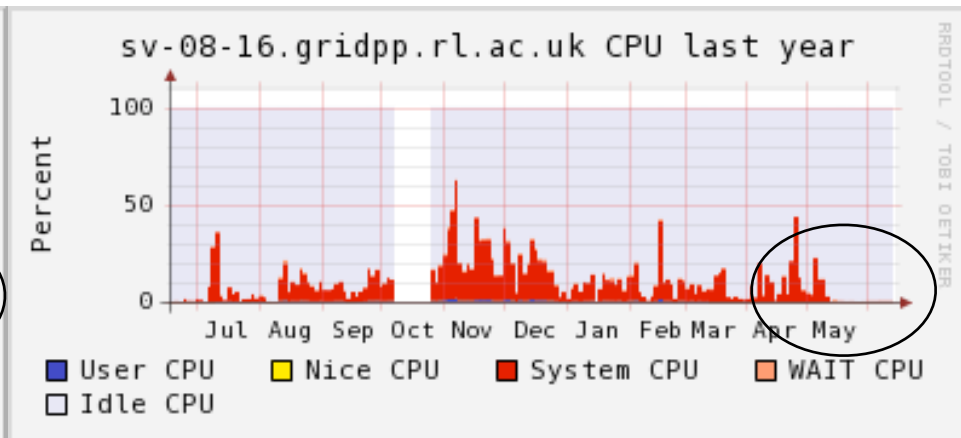
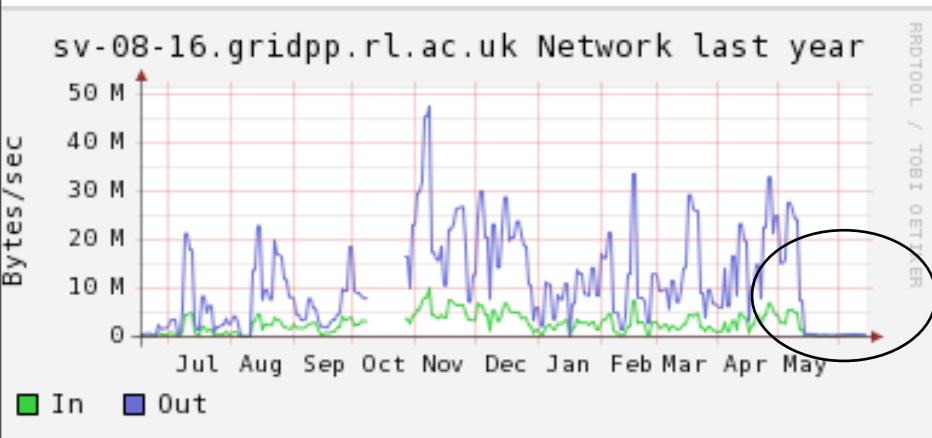
- NFS Atlas SW Server Loads - switched Atlas to CVMFS in May



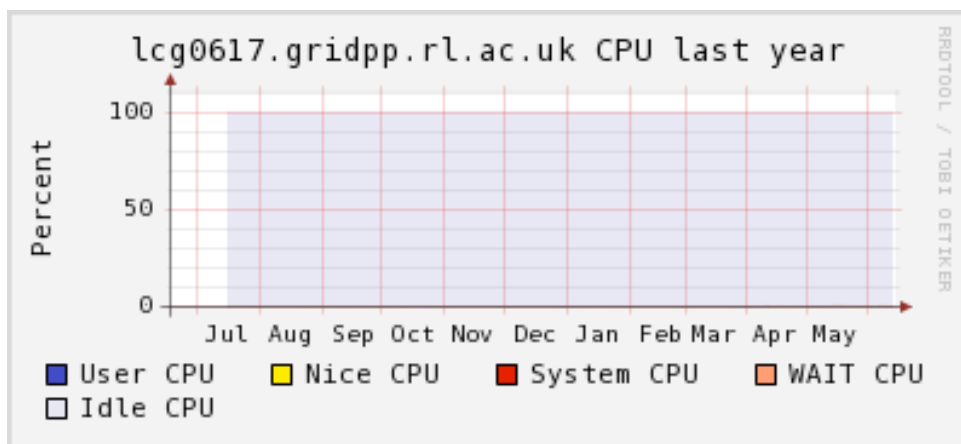
- Site (cluster) Squid Server loads - this is just one of the two

30th June 2011

- NFS Atlas SW Server Loads - switched Atlas to CVMFS in May

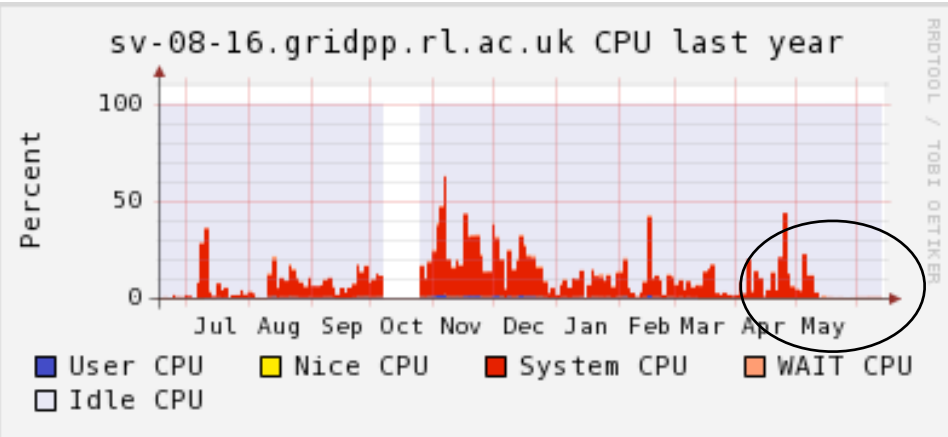
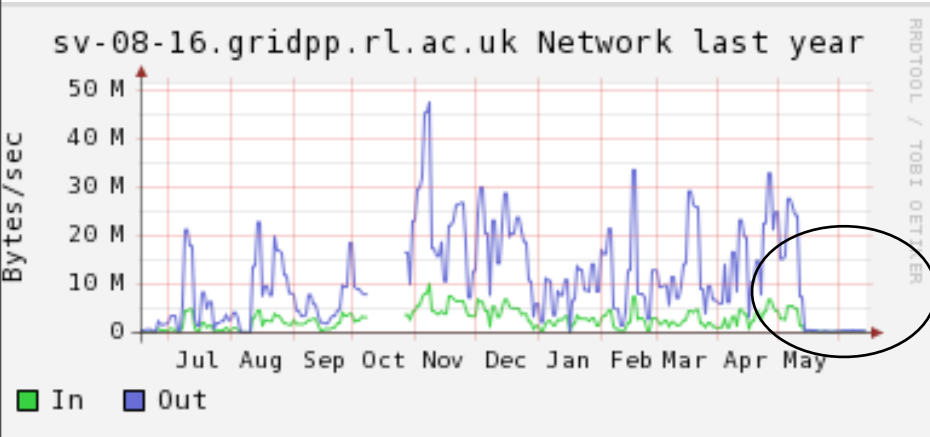


- Site (cluster) Squid Server loads - this is just one of the two

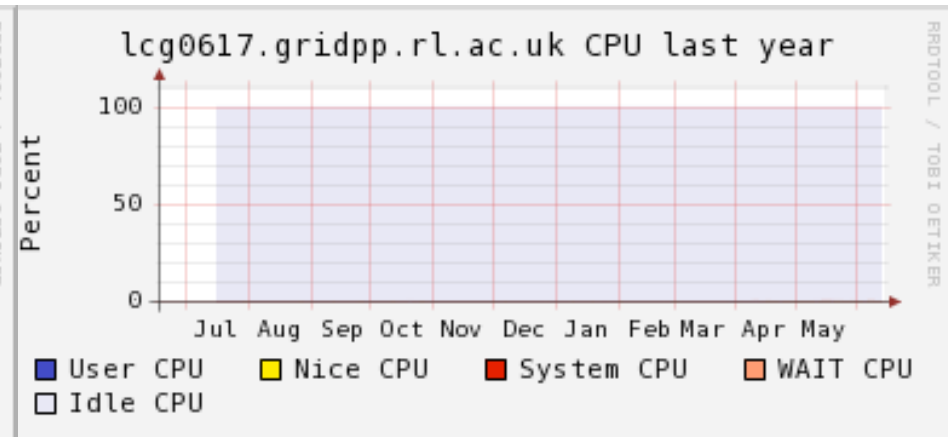
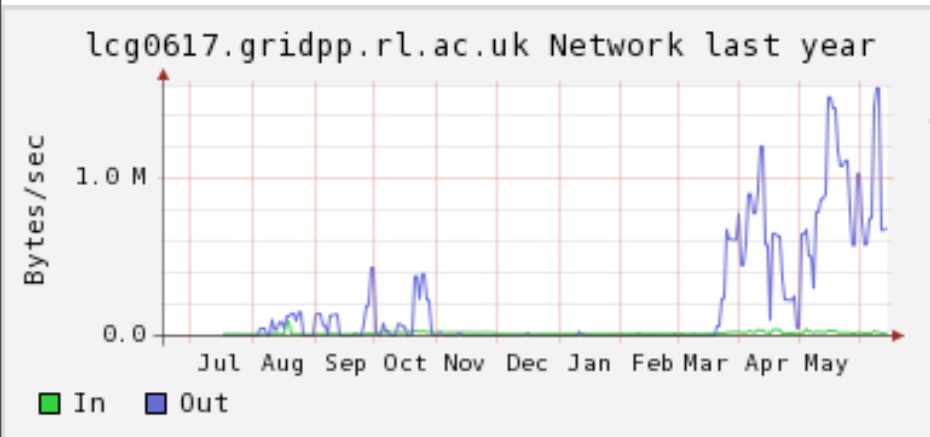


30th June 2011

- NFS Atlas SW Server Loads - switched Atlas to CVMFS in May

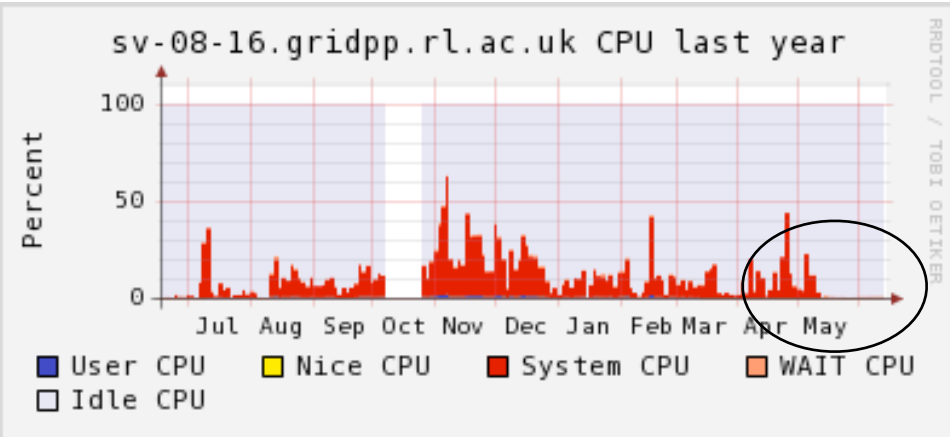
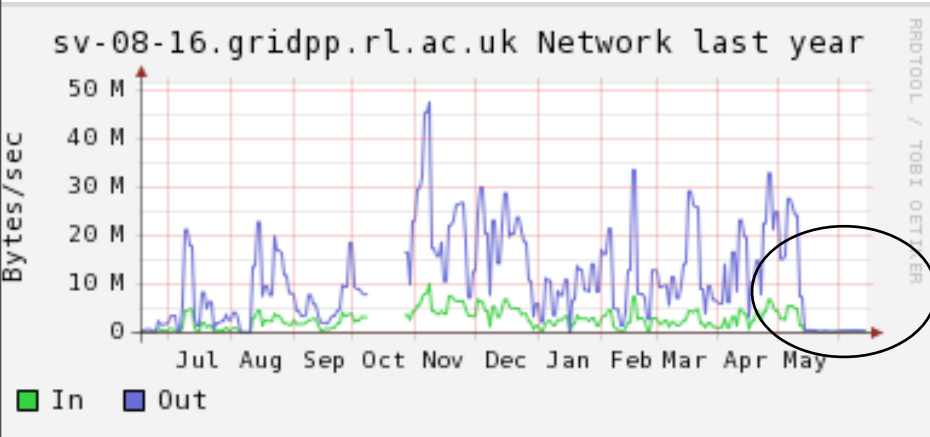


- Site (cluster) Squid Server loads - this is just one of the two

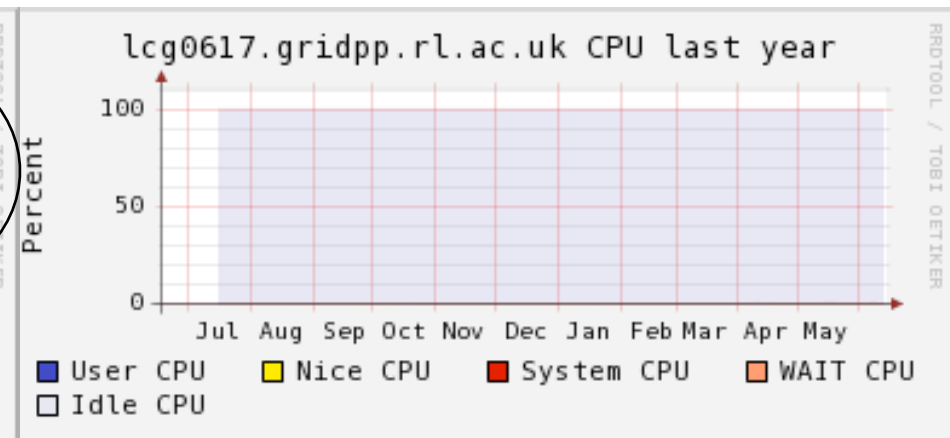
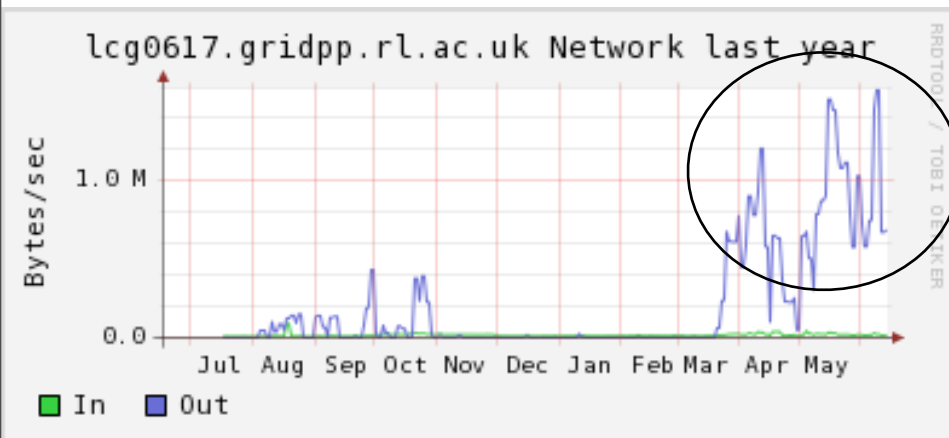


30th June 2011

- NFS Atlas SW Server Loads - switched Atlas to CVMFS in May



- Site (cluster) Squid Server loads - this is just one of the two



30th June 2011



- So, this all looks pretty good for the site admins, and not so bad for the VO release managers - surely there must be a downside - performance perhaps?

30th June 2011

- So, this all looks pretty good for the site admins, and not so bad for the VO release managers - surely there must be a downside - performance perhaps?
- Well, Victor Méndez at PIC tested just last month....

30th June 2011

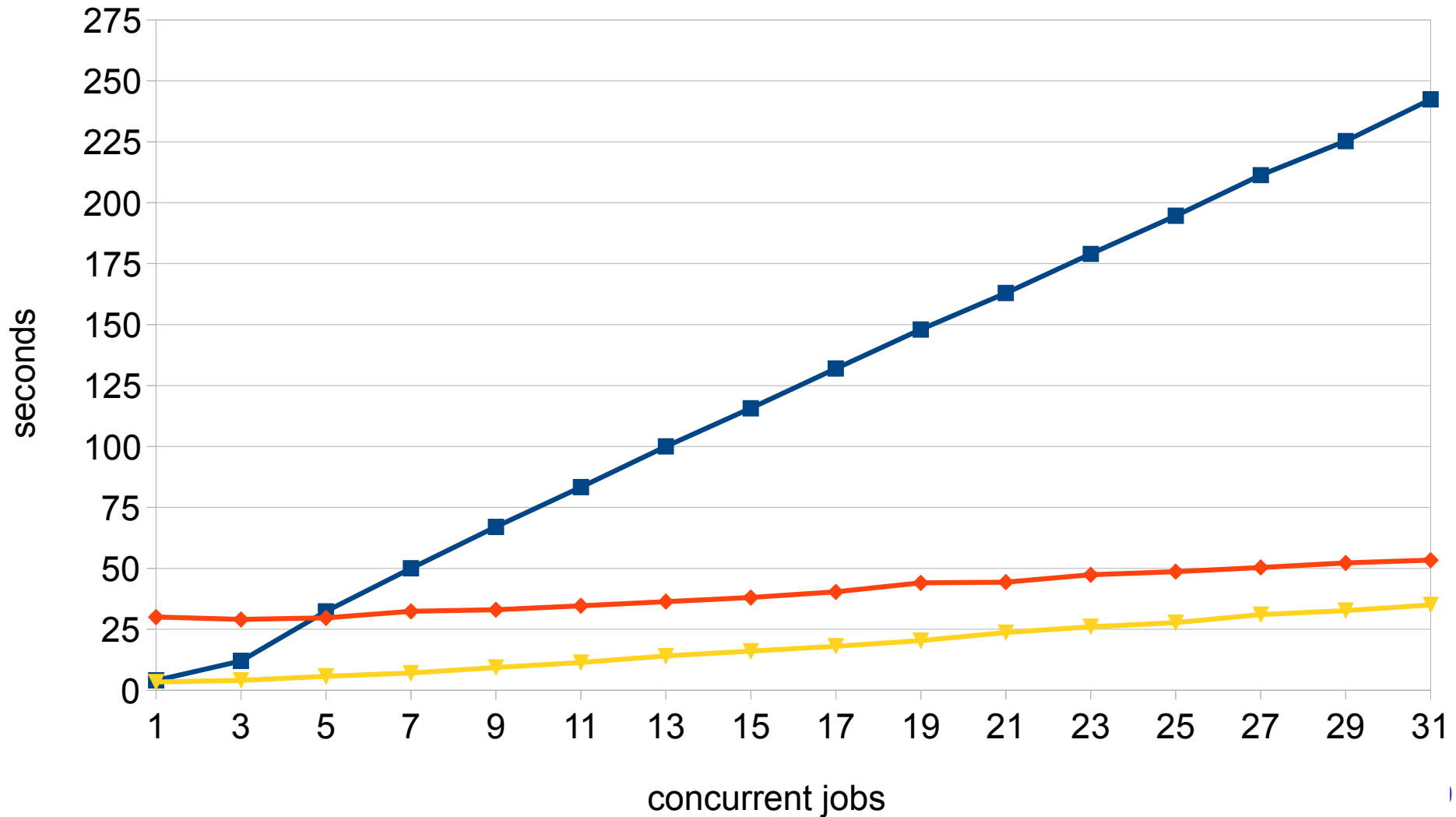
- So, this all looks pretty good for the site admins, and not so bad for the VO release managers - surely there must be a downside - performance perhaps?
- Well, Victor Méndez at PIC tested just last month....
- Metrics measured:
  - Execution time for SetupProjec Gauss v38r9 - the most demanding phase of the job for the software area (huge amount of stat() and open() calls )
  - Dependence on the hot and cold local cache. Cache size 174MB (catalog 148MB), 1 job run: cold hit ratio = 0.54, another run with hot cache hit ratio = 0.99
  - Comparison with standard NFS shared area
  - Dependence on the number of concurrent jobs

30th June 2011





■ nfs    ◆ cvmfs\_cold    ▼ cvmfs\_hot



Atlas install box in PH

LHCb install box in PH

Stratum 0 web in PH

cvmfs-public@cern

cvmfs-ral@cern

cvmfs-bnl@cern

Random site

Squid

Ba

Ba

Batch Node

30th June 2011





Atlas install box in PH

LHCb install box in PH

Stratum 0 web in PH

cvmfs-public@cern

cvmfs-ral@cern

cvmfs-bnl@cern

Random site

Squid

Ba

Ba

Batch Node

- Replication to Stratum 1 by hourly cron (for now)
- Stratum 0 moving to IT by end of year
- BNL almost in production

30th June 2011

## You need:

- Some worker nodes
- Some rpms installed on your WNs
- and some local configuration

## You also **must** have:

- A squid cache at your site (ideally two or more for resilience)
  - Configured (at least) to accept traffic from your site to one or more cvmfs repository servers
  - You could use existing frontier-squids, or existing web caches
  - Just make sure it caches files larger than 4kb (1 GB is probably good limit)

## What you need to configure

- The repositories required at your site (atlas, cms, lhcb, etc.)
- The source repository URL(s) to use
  - <http://cernvmfs.gridpp.rl.ac.uk/opt/@org@> <http://cvmfs-stratum-one.cern.ch/opt/@org@>
  - Ideally set one primary and at least one secondary - failover is built in to the client
- The size of the cache(s) on your WNs
- You need an entry in the autofs master map and also fuse.conf
- And finally set the VO SW AREA variable to point to the cvmfs area
- If you're supporting Atlas you need some local directories & links
  - but that is going away really, really soon

Repository	Description	Dependencies
atlas.cern.ch	ATLAS experiment software	atlas-condb.cern.ch
atlas-condb.cern.ch	ATLAS conditions database	
cms.cern.ch	CMS experiment software	grid.cern.ch
alice.cern.ch	ALICE experiment software	
lhcb.cern.ch	LHCb experiment software	grid.cern.ch
na61.cern.ch	NA61 experiment software	sft.cern.ch
hone.cern.ch	H1 experiment software	sft.cern.ch grid.cern.ch
boss.cern.ch	BES experiment software	
lcd.cern.ch	Software of the Linear Collider Studies collaboration	sft.cern.ch grid.cern.ch
hepsoft.cern.ch	Software of CERN's theoretical physics group	sft.cern.ch
grid.cern.ch	Grid User Interface	
sft.cern.ch	LCG application's area software	
geant4.cern.ch	Geant4 software	sft.cern.ch grid.cern.ch

30th June 2011

## Tools

```
cvmfs_talk  
cvmfs_config chksetup  
cvmfs_config showconfig <repository>.cern.ch  
service cvmfs status  
service cvmfs probe  
service cvmfs restartclean  
service cvmfs restartautofs
```

cvmfs\_talk allows us to 'interrogate' the caches  
cvmfs\_config shows/verifies the configuration  
the service commands allow granular stopping/starting/restarting/probing of components



- Do you need to treat it differently?
  - Mostly no
  - All those metadata operations and filesystem inquiries are mostly handled better
  - But ls -R is kind of pathological, and it fills up your caches with file catalogs, and it is not as fast as it might be.
  - But it *is* all local to the client - pretty much does not matter what you do, you will not kill the squids (NOT always the case with NFS/AFS)
  - Sites do need to instrument the clients - work in progress
  - Somewhere you will want one job that exhaustively checks that what is on the cvmfs repository is complete and correct - but that is a job for the site hosting the repository.

## Some Links

Download <http://cernvm.cern.ch/portal/downloads>

Latest Docs <http://cernvm.cern.ch/portal/techinfo>

Yum <http://cvmrepo.web.cern.ch/cvmrepo/yum>

News <http://twitter.com/cvmfs>

Bug Tracker <https://savannah.cern.ch/bugs/?group=cernvm>

Mailing list [cvmfs-talk@cern.ch](mailto:cvmfs-talk@cern.ch)

RAL Notes [https://www.gridpp.ac.uk/wiki/RAL\\_Tier1\\_CVMFS](https://www.gridpp.ac.uk/wiki/RAL_Tier1_CVMFS)

- **CernVM-FS is being used in production**
  - **LHCb at many Tier 1s (including RAL)**
  - **Atlas (at RAL)**
    - and many others
    - very active work to make new layout independent of so many links
  - **Others, and some sites running their own servers**
- **Core service supported at CERN**
- **Replicas in place at CERN, RAL and BNL**
  - **for resilience not load**
- **It is now transforming WLCG VO software distribution**