GridPP

UK Computing for Particle Physics

CernVM-FS Production Service

Future Computing Workshop, NESC

16 June 2011

Ian Collier, STFC-RAL

ian.collier@stfc.ac.uk

Science & Technology Facilities Council

e-Science

Overview

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

CernVM-FS - Introduction

CVMFS Big Idea:

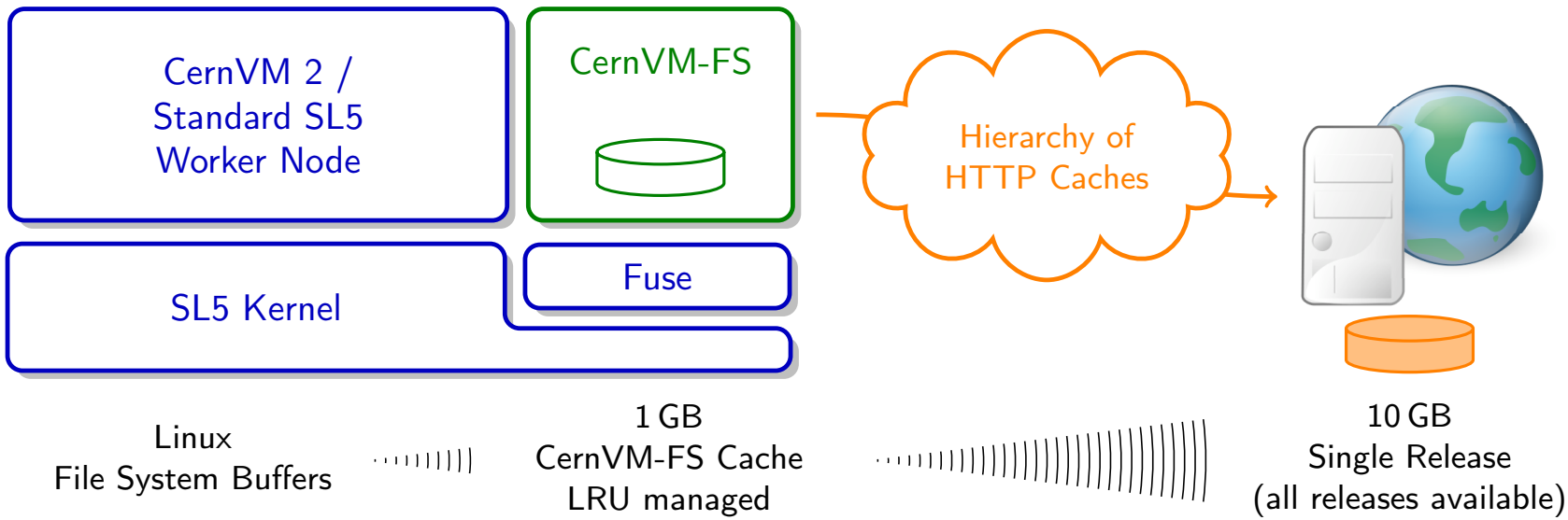
- Virtual software installation by means of an http based file system

- Based on standard components
  - http, fuse, sha1, squid, sqllite
- Caching
- Read-only
- Scalable

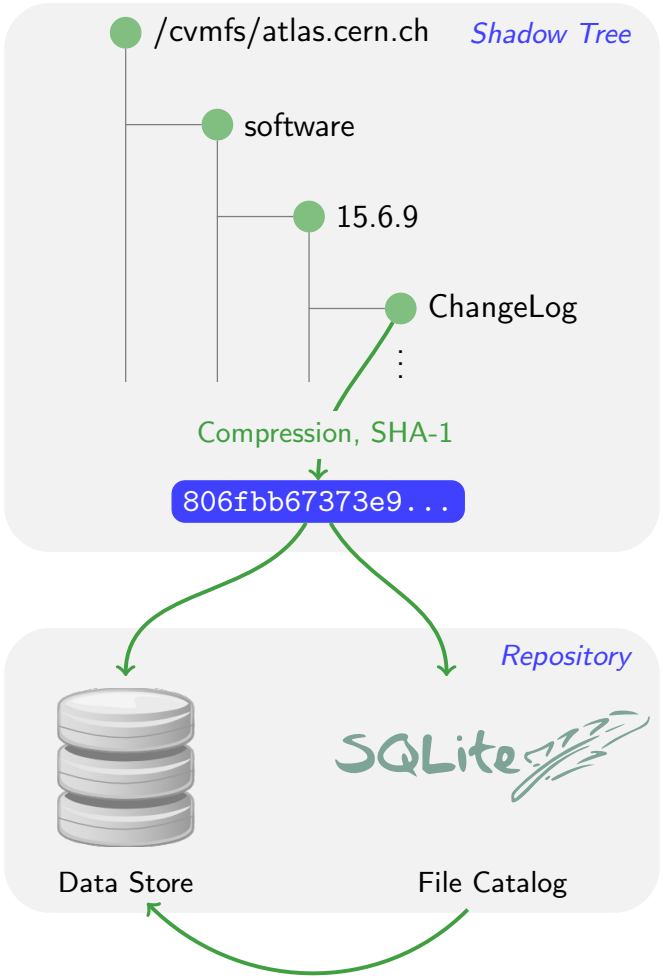
CernVM-FS - Introduction

CVMFS Principle:

Virtual software installation by means of an HTTP File System



CernVM-FS - Introduction



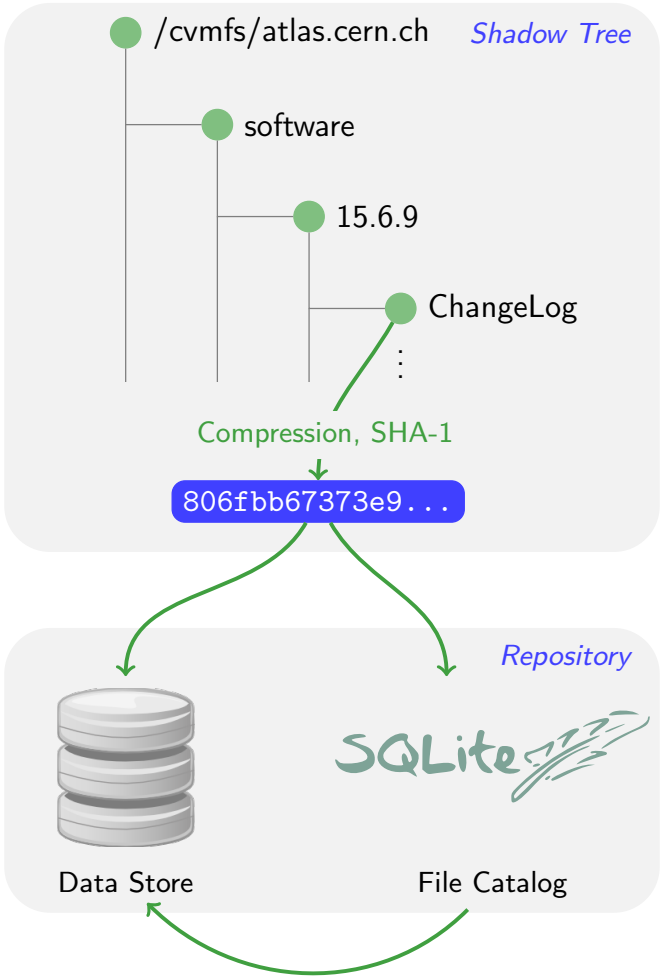
Data Store

- Compressed Chunks (Files)
- Eliminates Duplicates
  Never Deletes

File Catalog

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

CernVM-FS - Introduction



Data Store

- Compressed Chunks (Files)
- Eliminates Duplicates
  Never Deletes

File Catalog

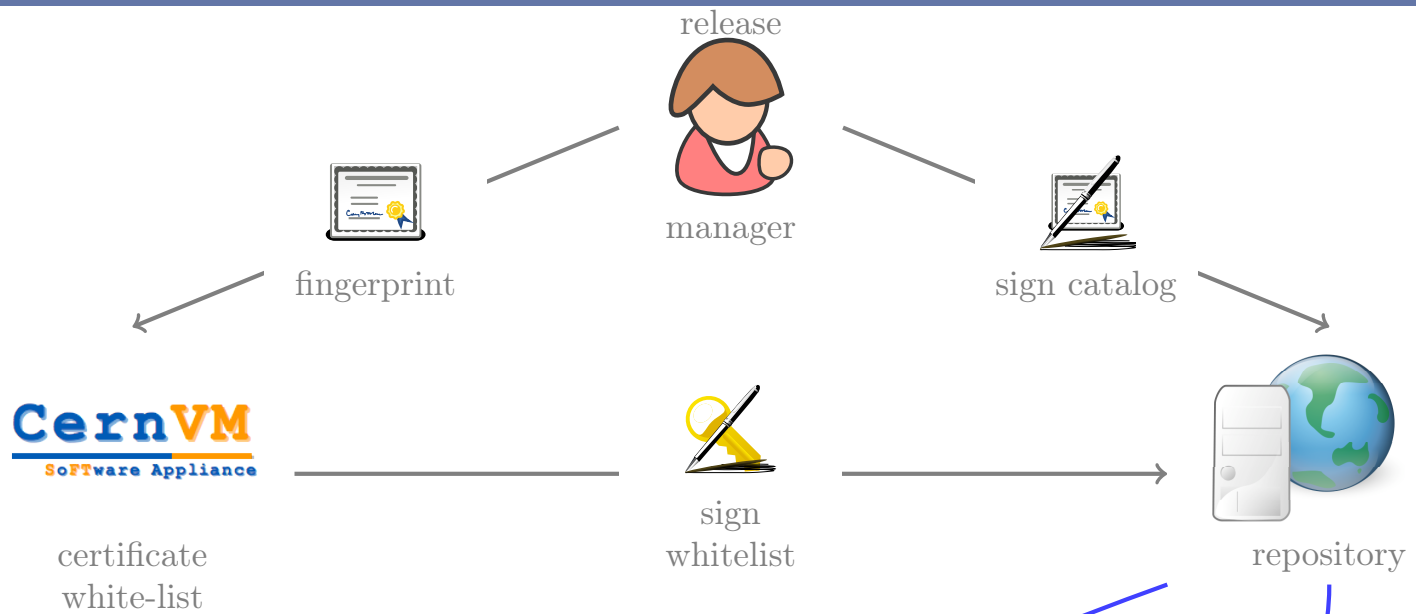
- Directory Structure
- Symlinks
- SHA1 of Regular Files
- Digitally Signed
- Time to Live
- Nested Catalogs

⇒ Immutable Files, trivial to check for corruption

CernVM-FS - Introduction

- 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

CernVM-FS - Release Process



CernVM-FS within WLCG

- Current 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.

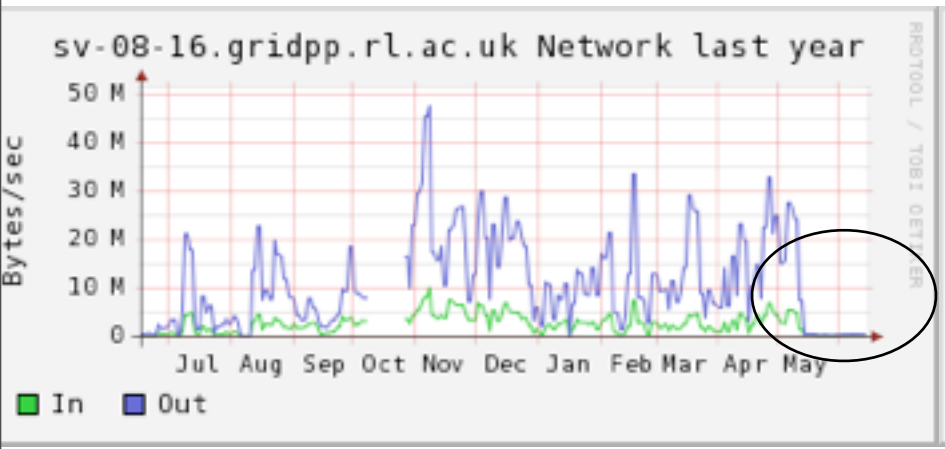
CernVM-FS within WLCG

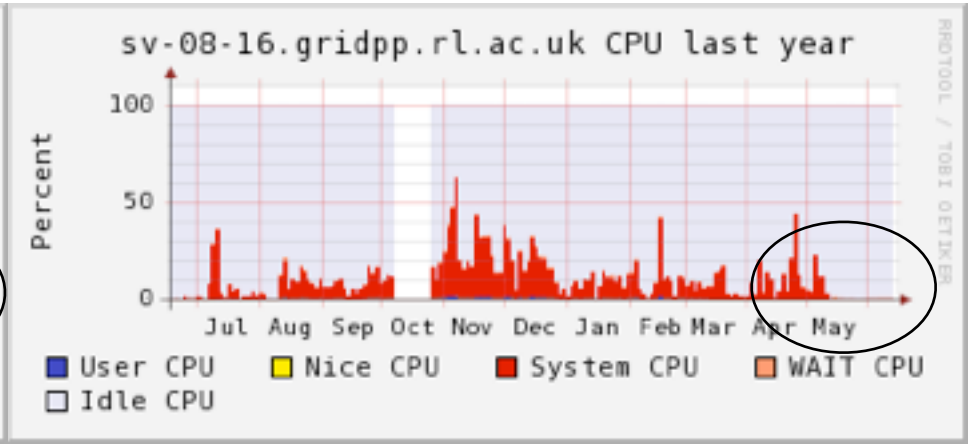
New method:

- Install once (at stratum zero)
  - Files appear everywhere across WLCG.
- This can be many days faster.
- Hopefully less variation across sites.
  - Common path `/cvmfs/`... (cf `/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

CernVM-FS Performance

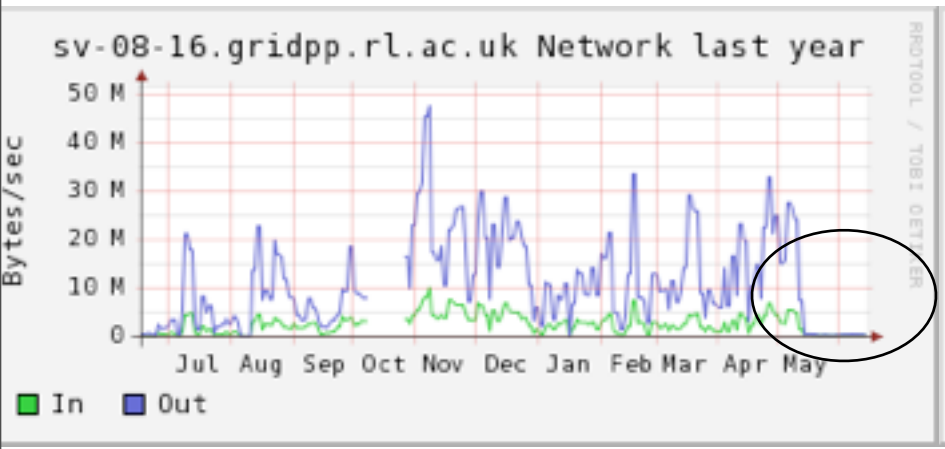
- NFS Atlas SW Server Loads - switched Atlas to CVMFS in May

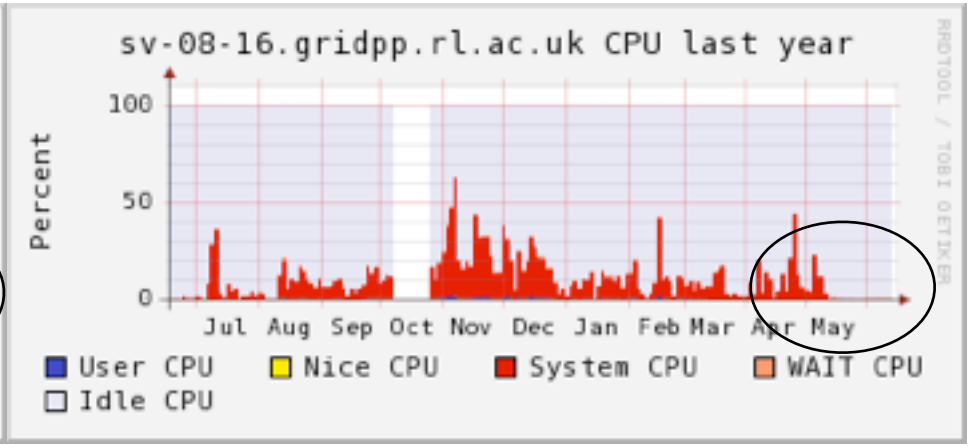




CernVM-FS Performance

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

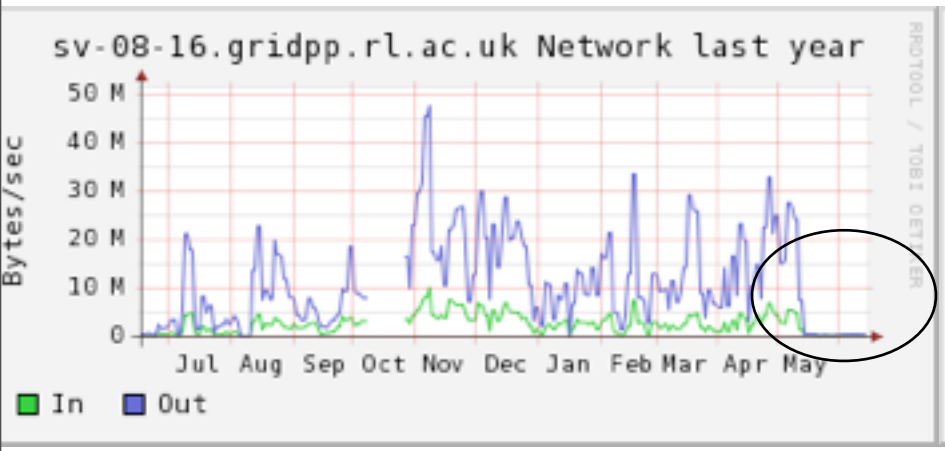


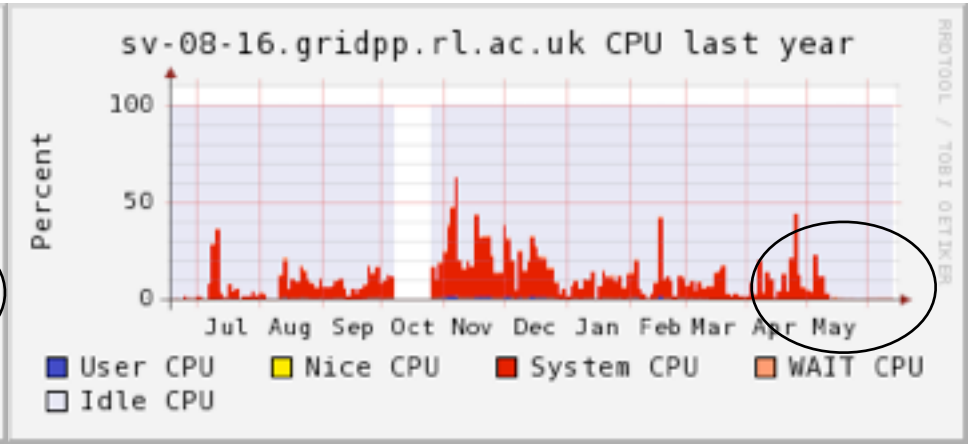


- Site (cluster) Squid Server loads - this is just one of the two

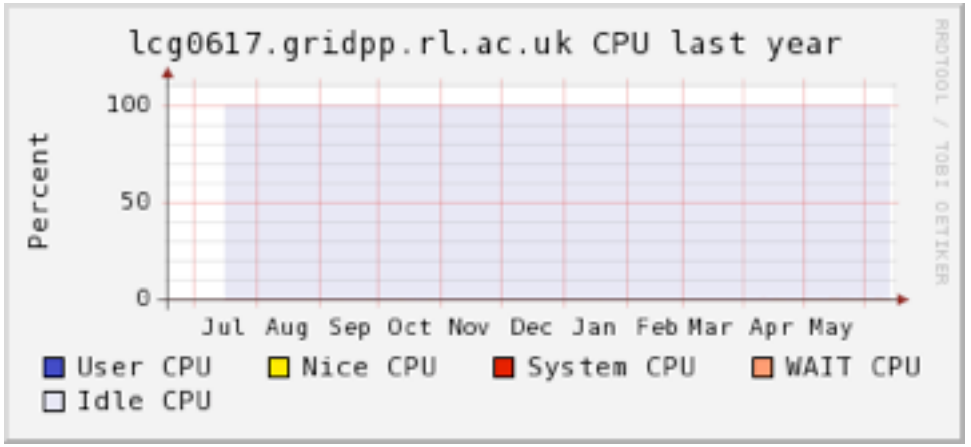
CernVM-FS Performance

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



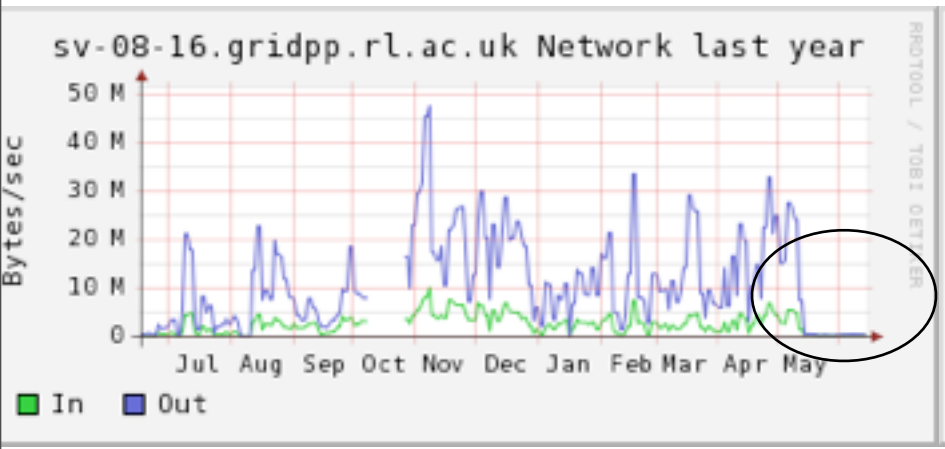


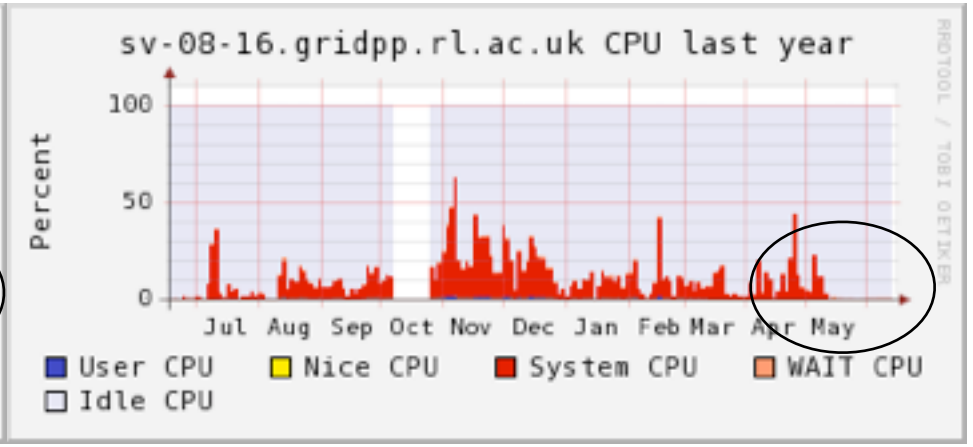
- Site (cluster) Squid Server loads - this is just one of the two



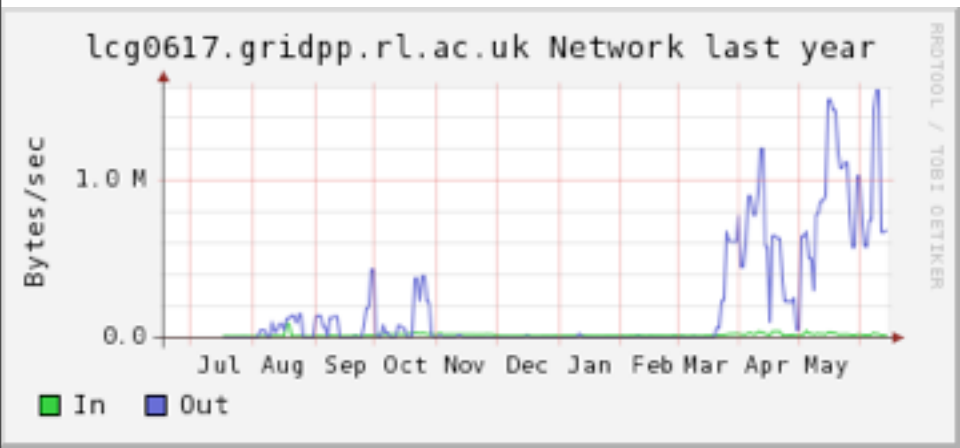
CernVM-FS Performance

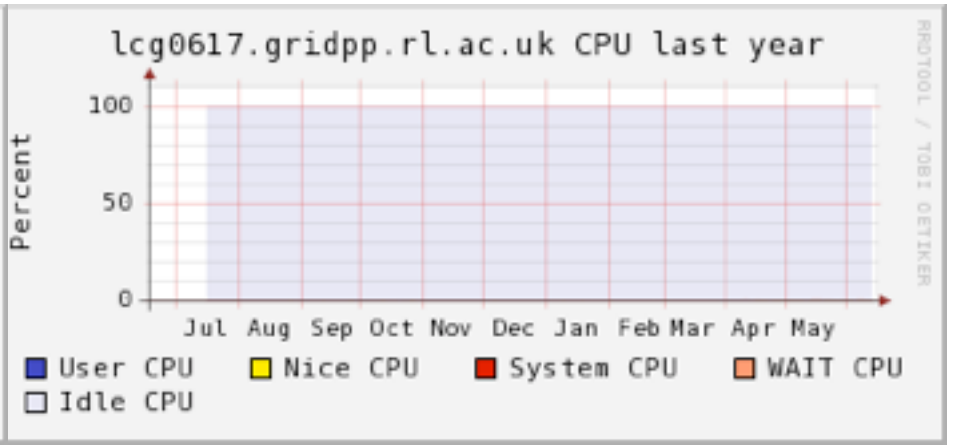
- NFS Atlas SW Server Loads - switched Atlas to CVMFS in May





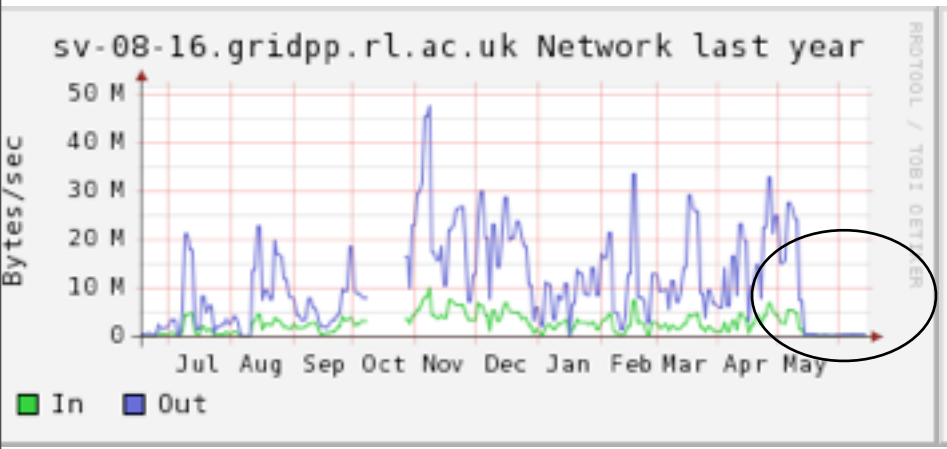
- Site (cluster) Squid Server loads - this is just one of the two

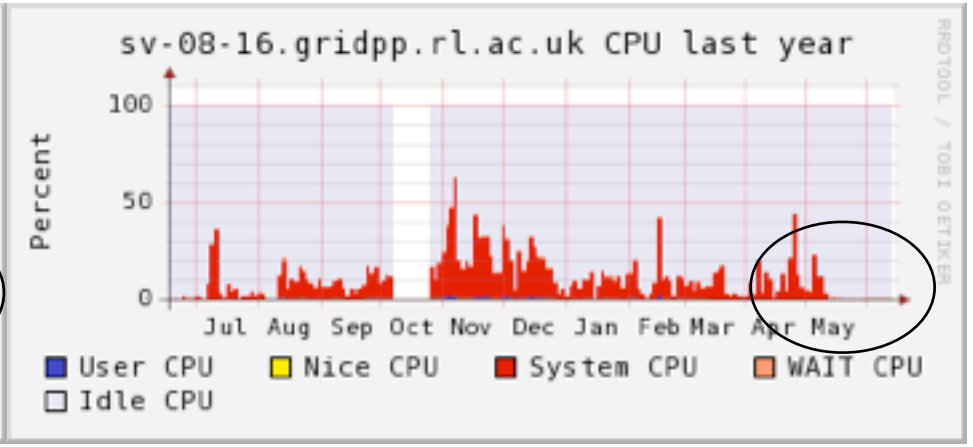




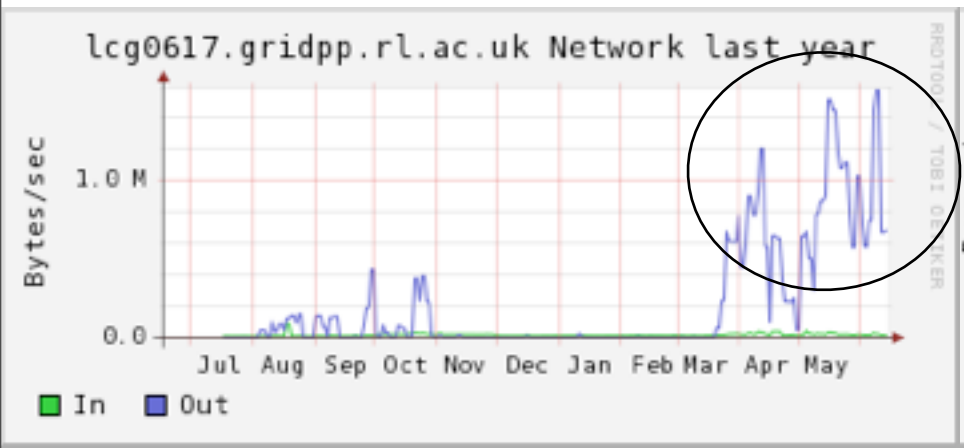
CernVM-FS Performance

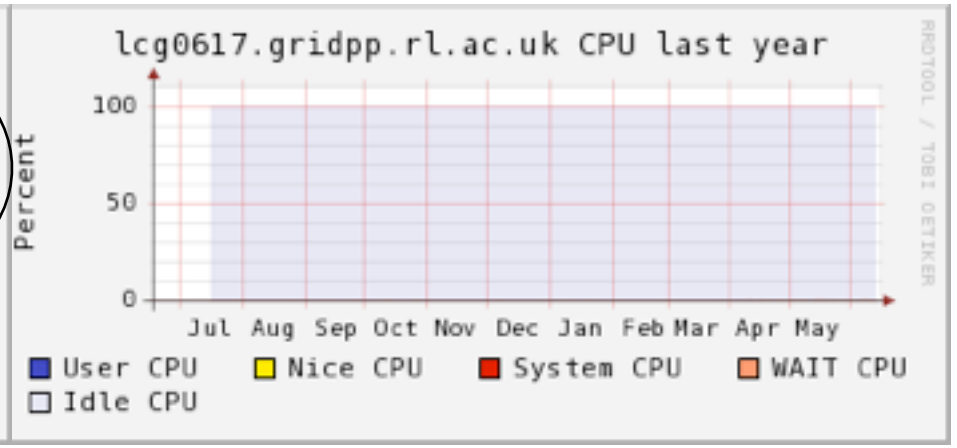
- NFS Atlas SW Server Loads - switched Atlas to CVMFS in May





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





CernVM-FS Performance

- 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?

CernVM-FS Performance

- 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....

CernVM-FS Performance

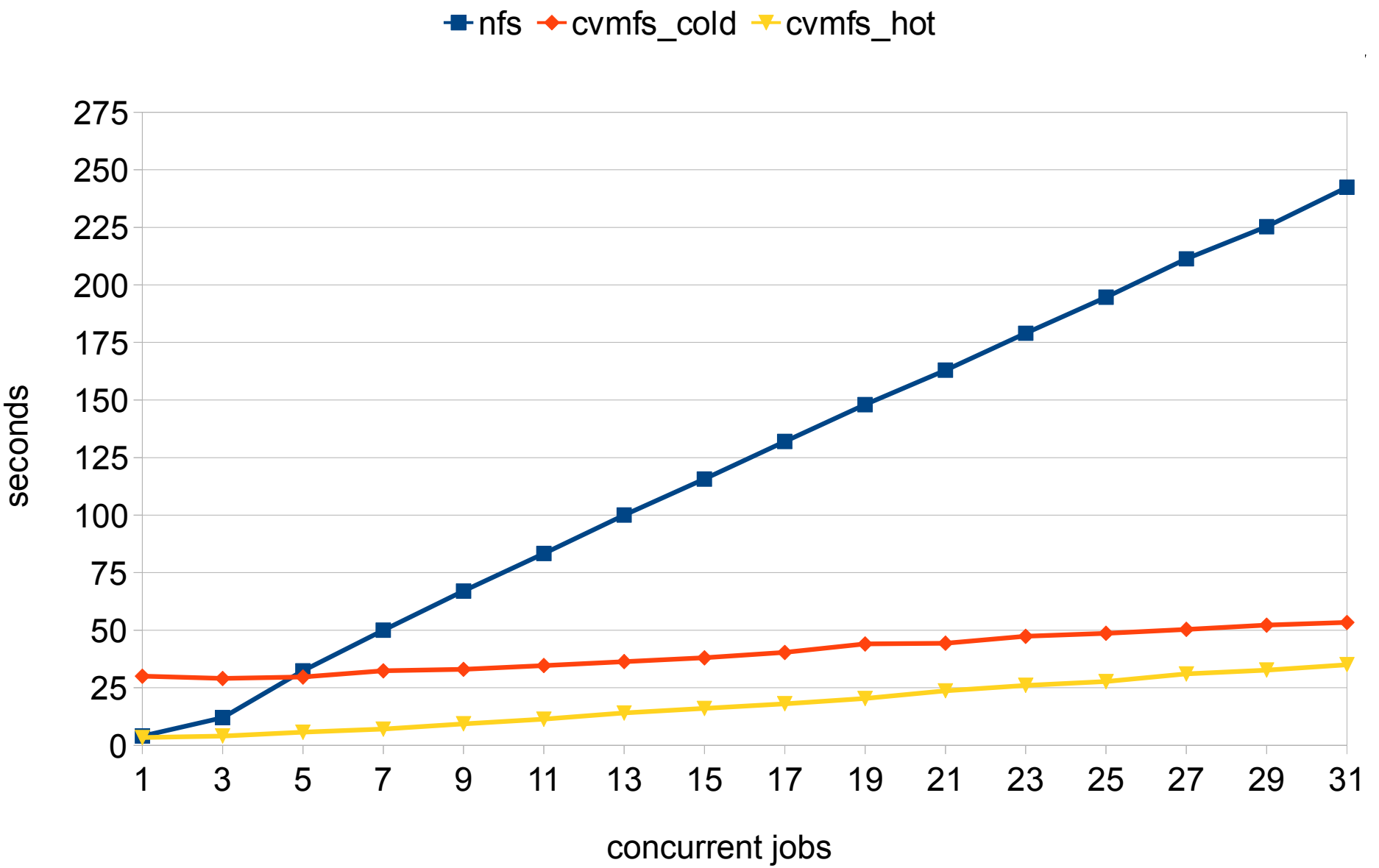
- So, this all looks pretty good for the site admins, and not so bad for the VO release managers - surely there must be a downside - performance perhaps?

- Well, Victor Méndez at PIC tested just last month....

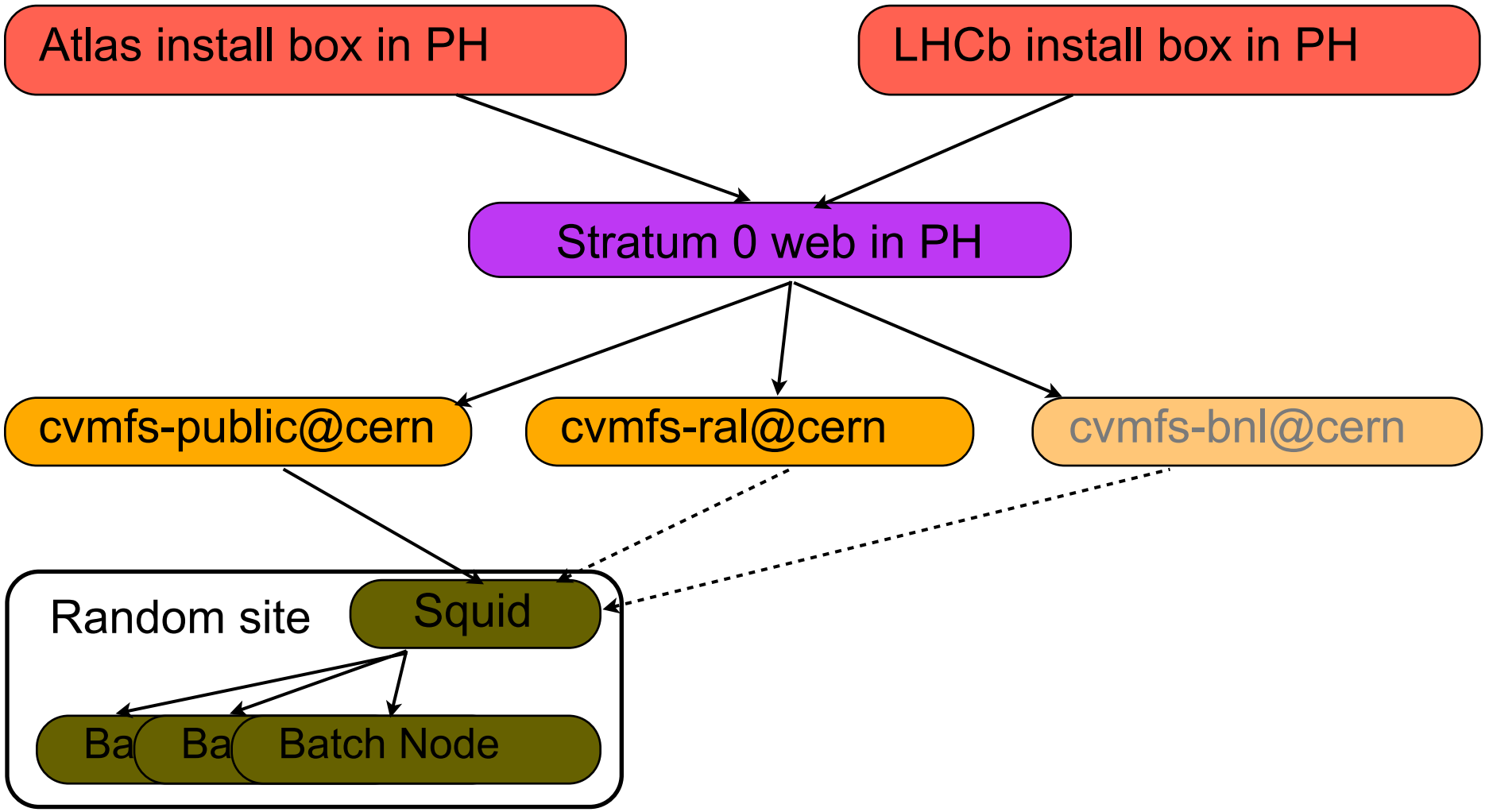
- Metrics measured:
  - Execution time for SetupProjec Gauss v38r9 - the most demanding phase of the job for the software area (huge amount of stat() and open () calls)
  - Dependence on the hot and cold local cache. Cache size 174MB (catalog 148MB), 1 job run: cold hit ratio = 0.54, another run with hot cache hit ratio = 0.99
  - Comparison with standard NFS shared area
  - Dependence on the number of concurrent jobs

GridPP
UK Computing for Particle Physics

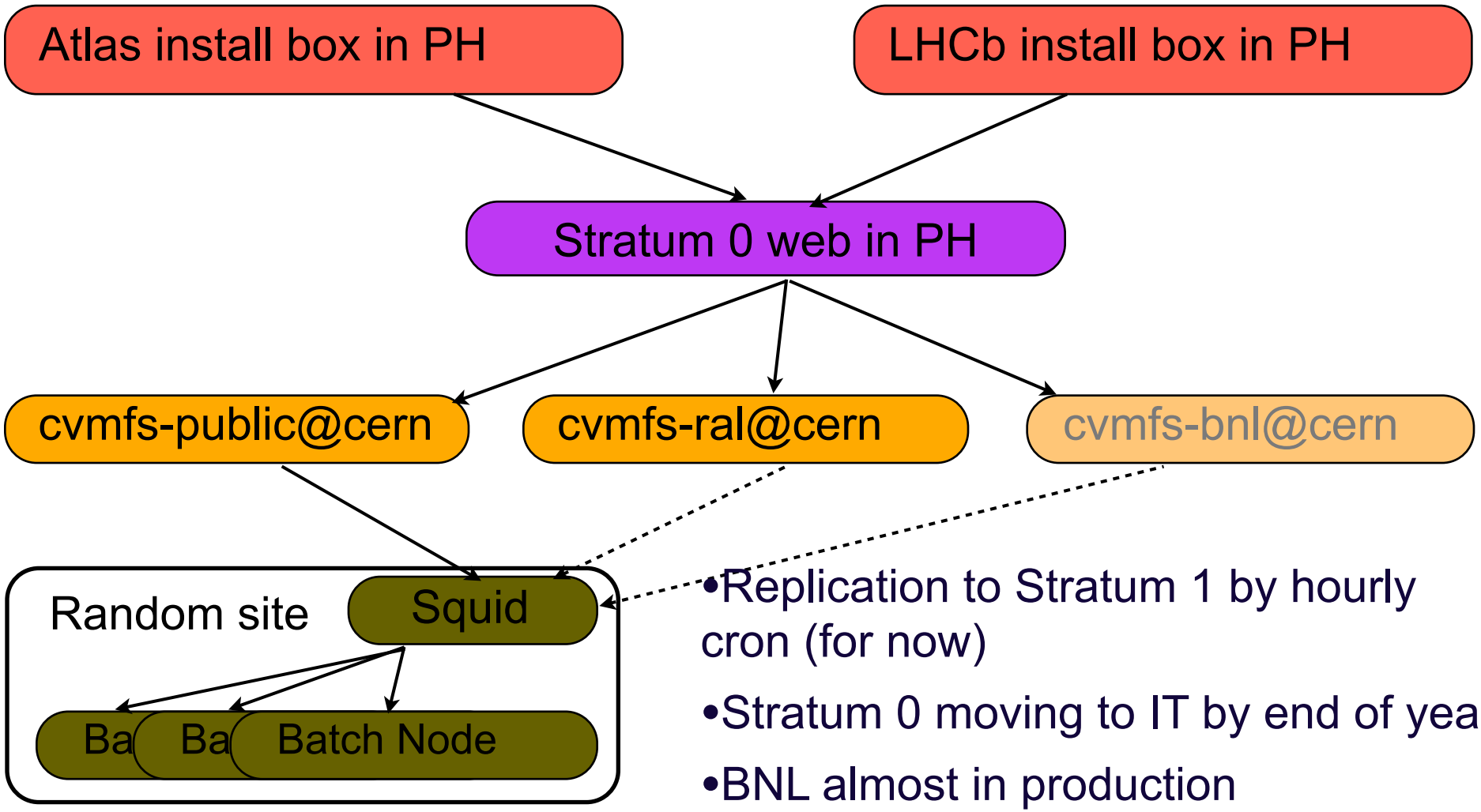
CernVM-FS Performance



CernVM-FS Service Status



CernVM-FS Service Status



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

CernVM-FS Site Deployment

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

CernVM-FS Site Deployment

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

CernVM-FS Site Deployment

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

CernVM-FS - In Use

- 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 the case with NFS/AFS)
  - So test/setup jobs at sites could concentrate on basic tests - is it there and basically working, and then trust the mechanism rather than exhaustive examination to test the presence of every file
  - 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.

CernVM-FS Links

Some Links	
Download	http://cernvm.cern.ch/portal/downloads
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
RAL Notes	https://www.gridpp.ac.uk/wiki/RAL_Tier1_CVMFS

Summary

- CernVM-FS is being used in production
  - LHCb at many Tier 1s (including RAL)
  - Atlas (at RAL)
    - also, Tier3s, PIC, NIKHEF, CERN, Wuppertal, QMUL, Munich, Lancester, Dortmund, JINR, . . .
  - 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