Storage management in GridPP

Greig A. Cowan

University of Edinburgh

Outline

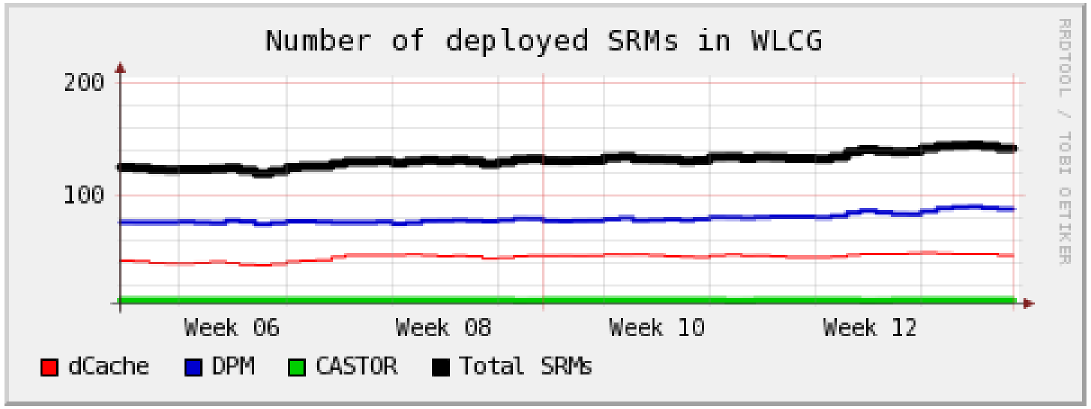
1. Deployed SRMs in WLCG

2. GridPP storage “community”

3. Storage accounting

4. Summary

Deployed SRMs in WLCG



	DPM	dCache	CASTOR	Total
WLCG	88	46	7	141
UK	12	7	1	20

- Query BDII for `/dpm`, `/pnfs` and `/castor` in the `GlueSARoot` field.
- Some sites may not expose this or may use an alternative SRM (StoRM. . .).

GridPP

UK Computing for Particle Physics

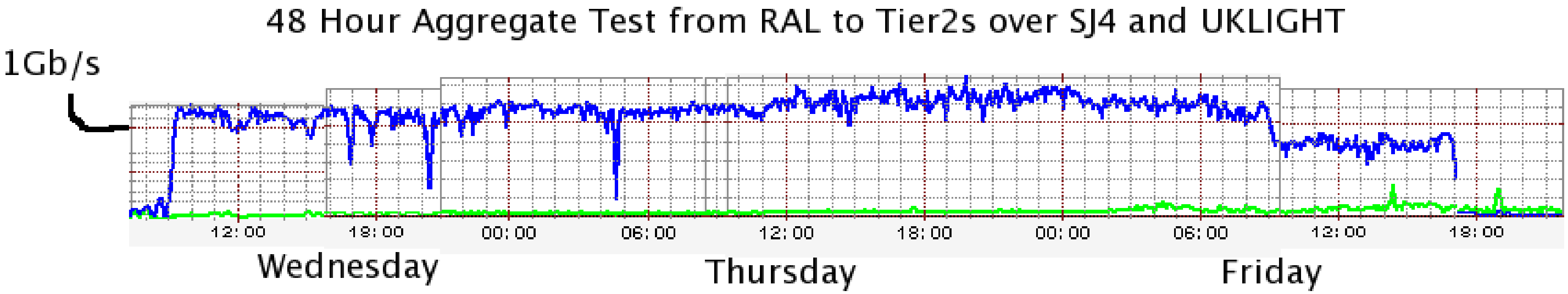
GridPP storage group

Background

- Q1 2003 All GridPP sites were running with a Classic SE, but there was little understanding of SRM middleware.
- Q4 2004 Positions were created to support Tier-2 deployment and operations of SRM middleware.
  - J. Jensen (RAL) to coordinate and manage the group.
  - G. Cowan at Edinburgh (2005).
  - 2 positions at RAL were filled until recently.
  - Other interested parties (G. Stewart, D. Ross. . .).
- Q2 2006 All 20 GridPP sites have operational SRMs.

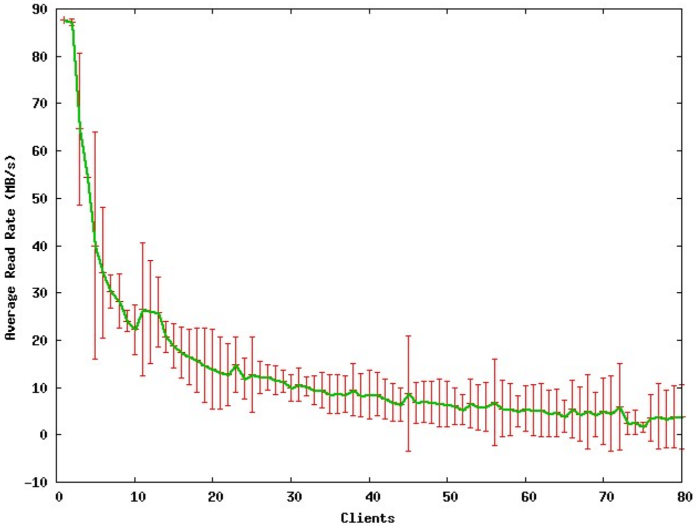
Optimisation on the WAN

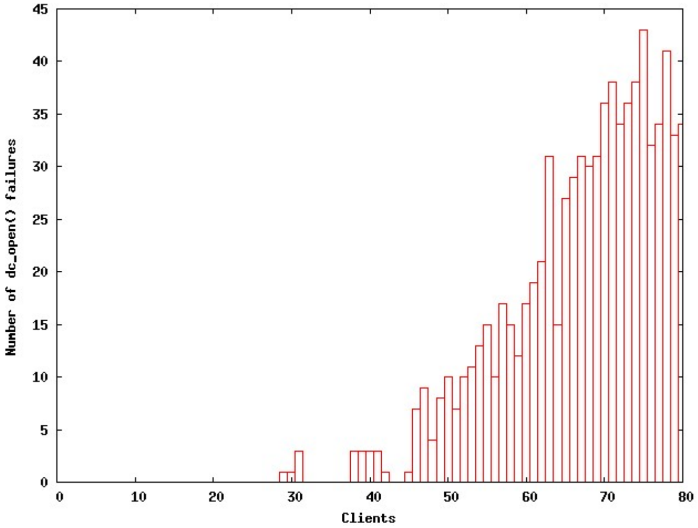
- The group has been actively testing storage infrastructure for $>$1year, e.g.,
  - Filesystems - XFS gives the greatest WAN transfer rate.
  - Kernel tunings.
- UK-wide transfer testing highlighted a number of bottlenecks, e.g.,
  - Regional and local networks.
  - Firewall problems.



Optimisation on the LAN

- GridPP identified need to study **local access** to the storage from WNs.
  - dCache known to handle 50 file opens/sec. Will there be any limits with DPM?
  - What rates are the VOs expecting? (∼2MB/s/job for Atlas)
- We created a test client to run on WNs that simultaneously read files from the SE.
- Problem already found with DPM ⟶ JPB has fixed in v1.6.3.





Storage availability with SAM

- SRM and SE tests use the `lcg-cr, lcg-cp` ...tools to probe storage.
  - Run from a machine at CERN.
  - Depend on SAM BDII and central catalog.

  ⇒ SAM tests do **not** give true measure of site storage availability.

- e.g., Summary of recent SAM failures at UKI-SCOTGRID-GLASGOW:

		Reason for failure			Availability	
SAM test	Total failures	SAM BDII	Site	Unknown	SAM	True
SRM	16/650	**14**	1	1	**94.5%**	**99.5%**
SE	20/650	**18**	1	1		

Examples of errors

- SE-lcg-del

```
+ lcg-del -v --vo ops -a lfn:SE-lcg-cr-srm.epcc.ed.ac.uk-1175392410
BDII ERROR: sam-bdii.cern.ch:2170 Success
lcg_del:  Invalid argument
```

- SRM-put

```
+ lcg-cr -v --vo ops file:/home/samops/.same/SRM/testFile.txt
-l lfn:SRM-put-srm.epcc.ed.ac.uk-1175394118 -d srm.epcc.ed.ac.uk
BDII Connection Timeout:  sam-bdii.cern.ch:2170
lcg_cr:  Connection timed out
Using grid catalog type:  lfc
Using grid catalog :  prod-lfc-shared-central.cern.ch
```

Suggestions for improvements

- Need to **correlate** failures with SAM BDII errors/timeouts.
- Or **filter** out the SAM BDII failures (error messages are clear).
- SRM and SE critical tests essentially do the same thing!
- SRM test should only probe the **lower level** functionality.
  - `srmPut, srmGet, srmCopy...`
  - No interaction with catalogs, information system.
- GridPP already has such a test so could contribute to SAM.

https://savannah.cern.ch/bugs/index.php?25249

Monitoring LAN access to storage

- New SAM test: `CE-sft-posix`.
- Use GFAL to read a file from close SE using suitable protocol (rfio, gsidcap...)
- Initial results in GridPP were promising:

	Passed	Failed
Num. of sites	14	6

- Majority of problems appear to be with firewalls.
- Test is **non-critical** at the moment.

Communication

- Storage group has good relationship with all stakeholders:
  - Site administrators.
  - Storage middleware developers.
  - WLCG (through the GDB).
- 50 subscribed to mailing list. Not just UK.
- Weekly (30 min!) meetings to discuss latest storage developments and assign work.
- Up-to-date documentation about dCache and DPM.

http://www.gridpp.ac.uk/wiki/Grid_Storage

- Storage blog:

http://gridpp-storage.blogspot.com

Storage Accounting

Background

- Extensive levels of CPU accounting available. What about storage?
- Different user communities have different questions:
  - Which sites are meeting their MoU targets?
  - Which VOs (and VO groups) are using the storage at my site?
  - Are these grid or non-grid users?
- GridPP started prototype system for accounting (Dave Kant and myself).

Details

- Every SE runs a generic information provider (GIP) which publishes information according to the GLUE schema.

- Concept of storage areas (often 1 SA per VO).

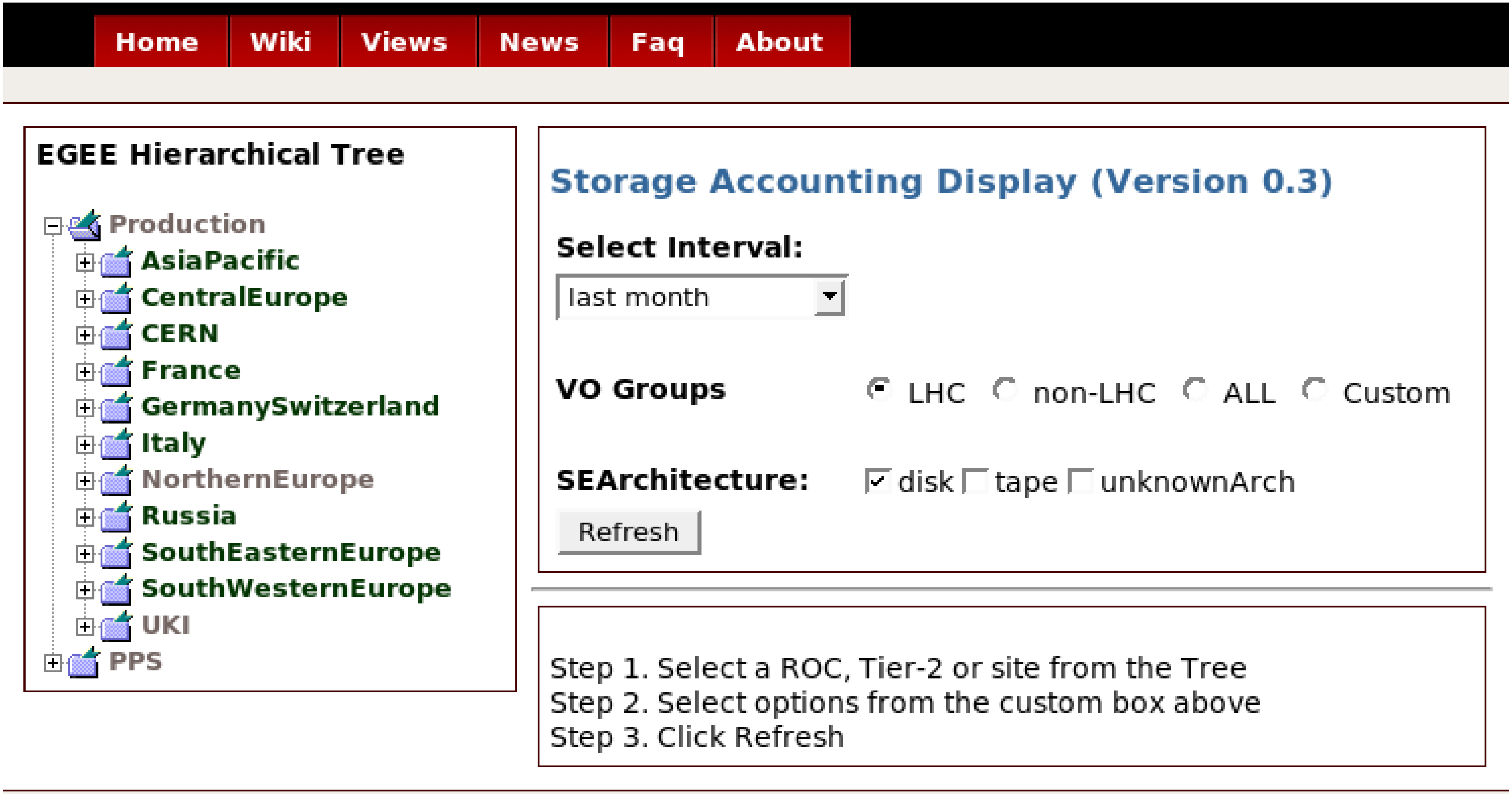
```
GlueSiteName GlueSEArchitecture GlueSAStateUsedSpace
GlueSAStateAvailableSpace GlueSAType GlueSAPath
```

- Harvest these attributes by querying a top level BDII → MySQL DB.

- User-friendly front-end created:
  - Allows users to query the DB.
  - Dynamically generates historical plots of the used space.

http://goc02.grid-support.ac.uk/storage-accounting/view.php

Front-end



EGEE Hierarchical Tree

- ⊟ Production
  - ⊞ AsiaPacific
  - ⊞ CentralEurope
  - ⊞ CERN
  - ⊞ France
  - ⊞ GermanySwitzerland
  - ⊞ Italy
  - ⊞ NorthernEurope
  - ⊞ Russia
  - ⊞ SouthEasternEurope
  - ⊞ SouthWesternEurope
  - ⊞ UKI
- ⊞ PPS

Storage Accounting Display (Version 0.3)

Select Interval: last month ▾

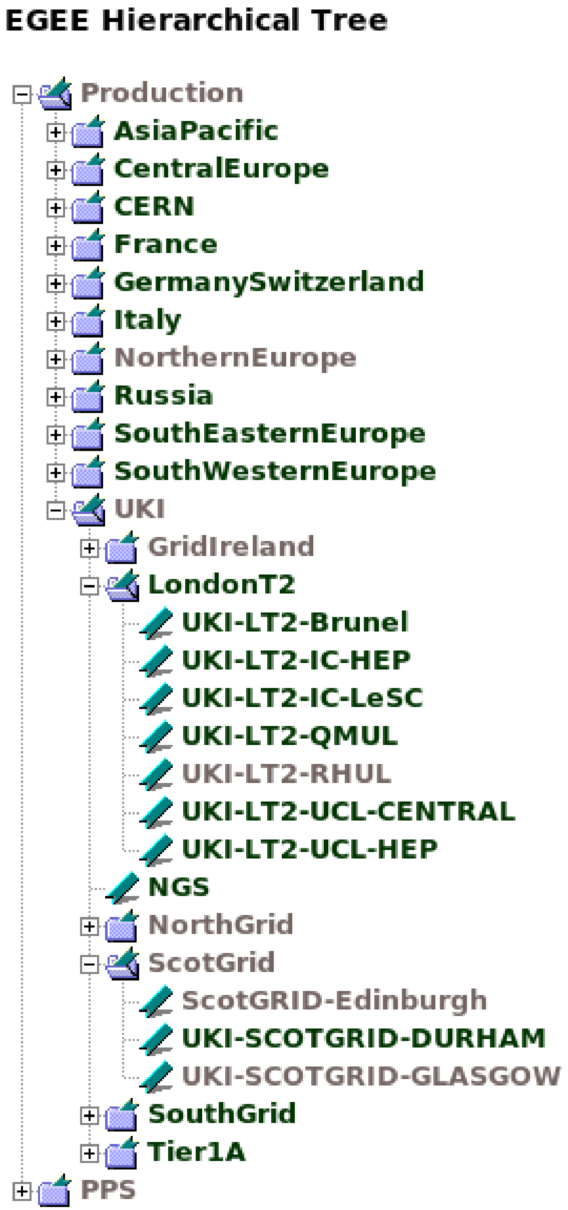
VO Groups ● LHC ○ non-LHC ○ ALL ○ Custom

SEArchitecture: ☑ disk □ tape □ unknownArch

Refresh

Step 1. Select a ROC, Tier-2 or site from the Tree
Step 2. Select options from the custom box above
Step 3. Click Refresh

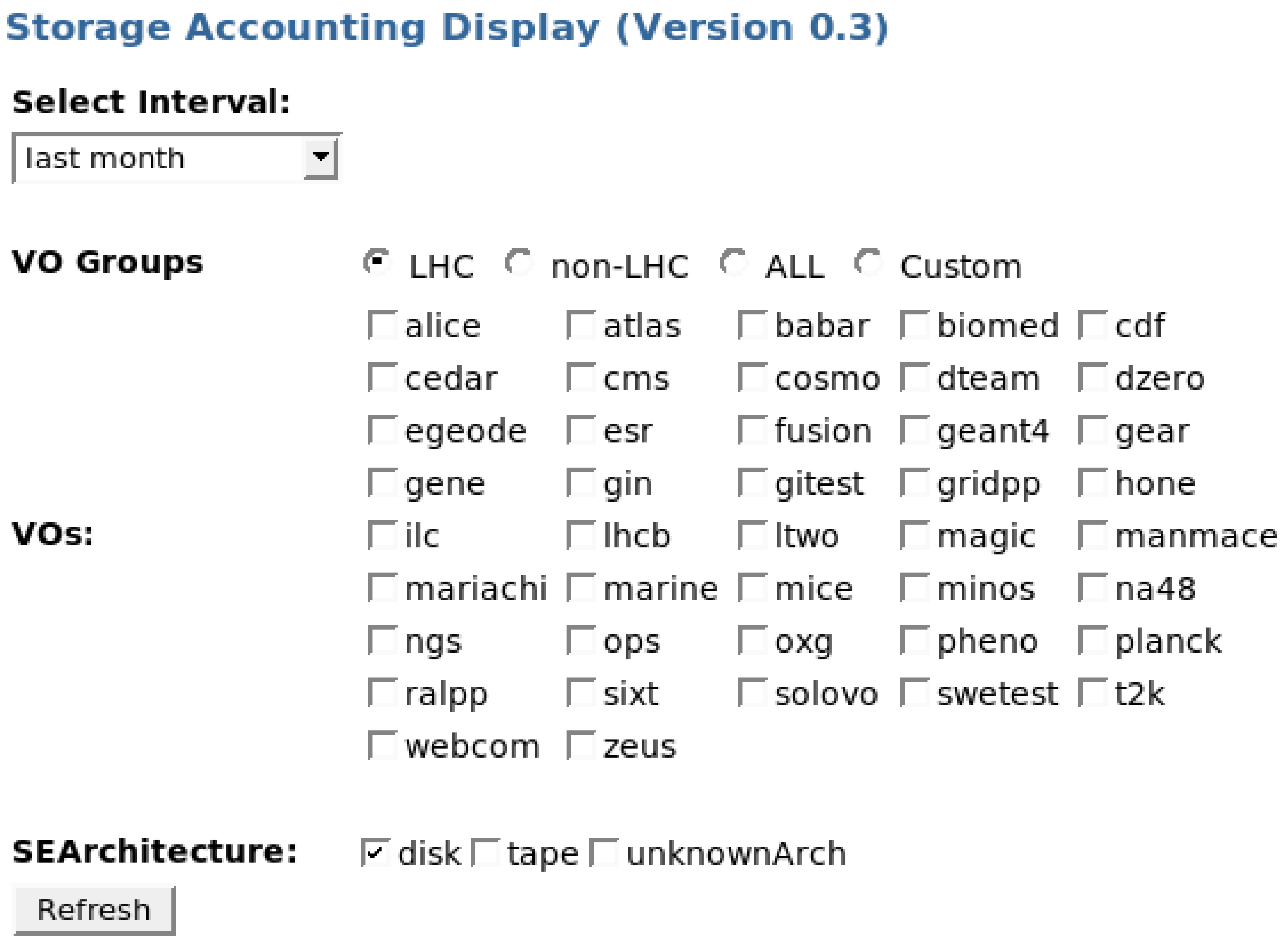
EGEE tree



EGEE Hierarchical Tree

- Production
  - AsiaPacific
  - CentralEurope
  - CERN
  - France
  - GermanySwitzerland
  - Italy
  - NorthernEurope
  - Russia
  - SouthEasternEurope
  - SouthWesternEurope
  - UKI
    - GridIreland
    - LondonT2
      - UKI-LT2-Brunel
      - UKI-LT2-IC-HEP
      - UKI-LT2-IC-LeSC
      - UKI-LT2-QMUL
      - UKI-LT2-RHUL
      - UKI-LT2-UCL-CENTRAL
      - UKI-LT2-UCL-HEP
    - NGS
    - NorthGrid
    - ScotGrid
      - ScotGRID-Edinburgh
      - UKI-SCOTGRID-DURHAM
      - UKI-SCOTGRID-GLASGOW
    - SouthGrid
    - Tier1A
- PPS

VO selection



Storage Accounting Display (Version 0.3)

Select Interval:

last month ▾

VO Groups ● LHC ○ non-LHC ○ ALL ○ Custom

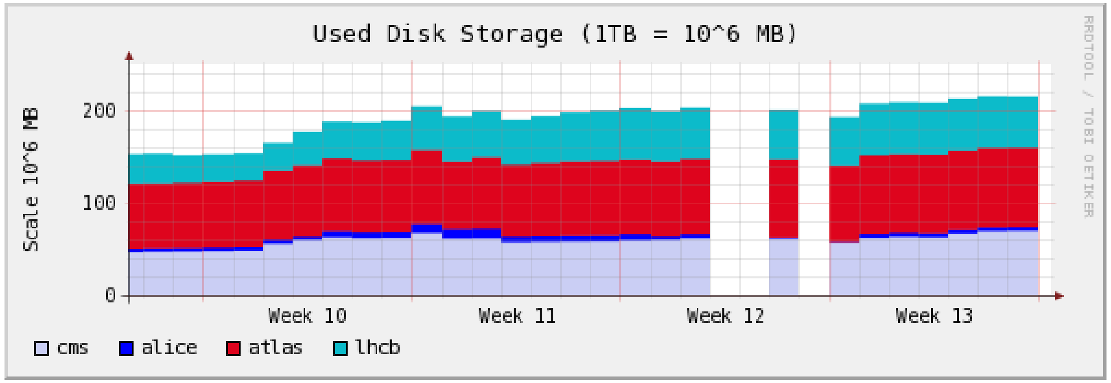
VOs:

□ alice □ atlas □ babar □ biomed □ cdf
□ cedar □ cms □ cosmo □ dteam □ dzero
□ egeode □ esr □ fusion □ geant4 □ gear
□ gene □ gin □ gitest □ gridpp □ hone
□ ilc □ lhcb □ ltwo □ magic □ manmace
□ mariachi □ marine □ mice □ minos □ na48
□ ngs □ ops □ oxg □ pheno □ planck
□ ralpp □ sixt □ solovo □ swetest □ t2k
□ webcom □ zeus

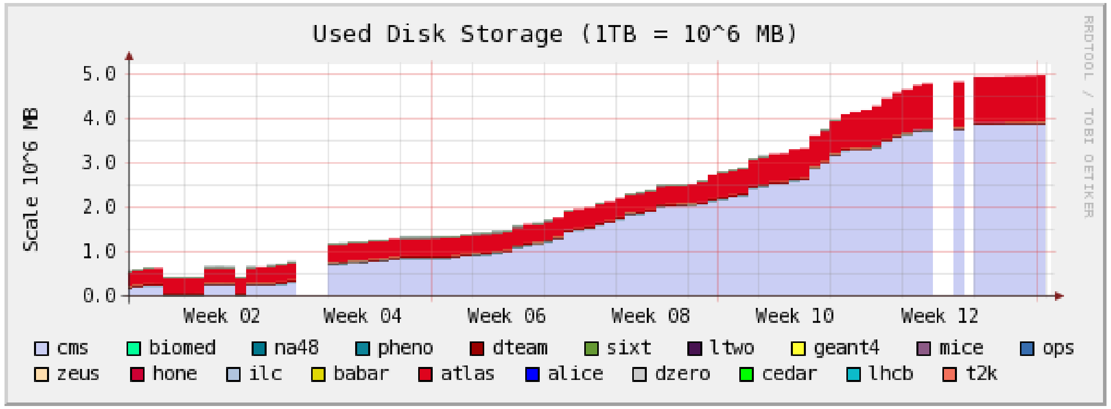
SEArchitecture: ☑ disk □ tape □ unknownArch

Historical usage plots

GridPP disk usage over past month for LHC VOs (>200TB)



UKI-LT2-RHUL disk usage over past 3 months for all VOs (∼5TB)



Project organisation

- Implementation details:

http://www.gridpp.ac.uk/wiki/Storage_Accounting

- Bug tracker:

https://savannah.cern.ch/projects/storage-account/

Double/triple/. . . counting

- If VOs **share** DPM pools or dCache pool groups then the default GIPs do **not** correctly report the available and used space per VO.

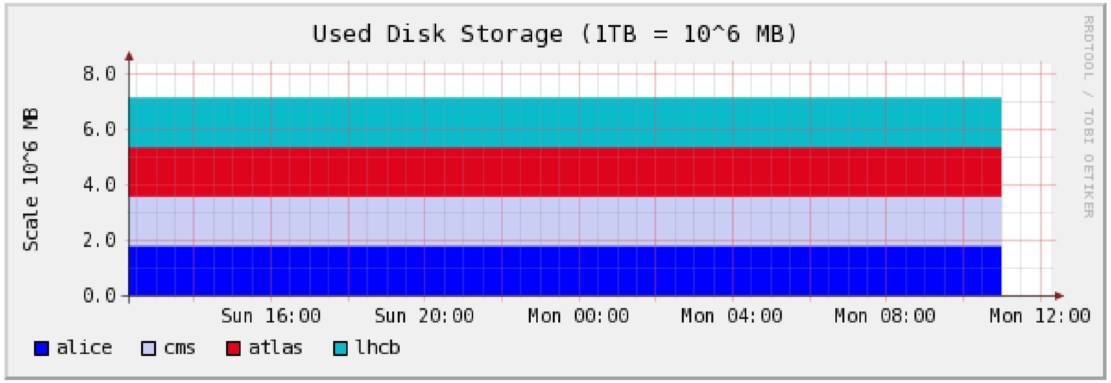
DPM:

* GridPP have created a custom GIP which correctly reports the **used** space. Available space unknown from shared pools.

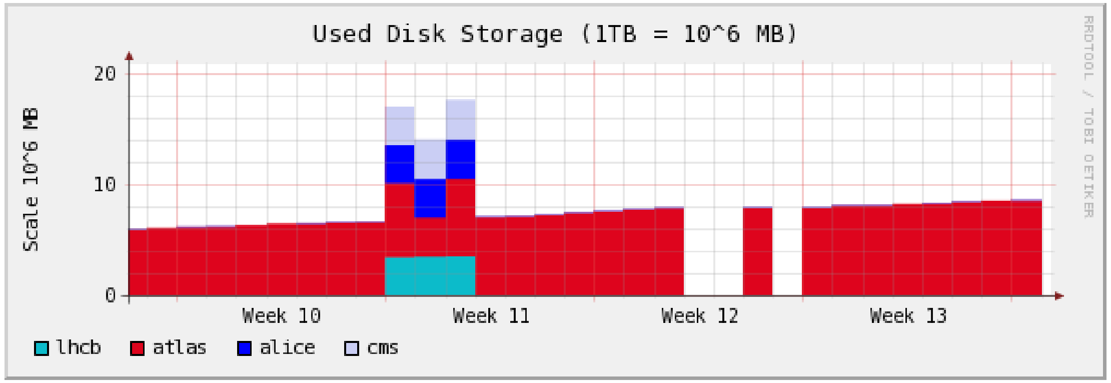
http://www.gridpp.ac.uk/wiki/DPM_Information_Publishing

DPM GIP plugin

NIKHEF-ELPROD



UKI-SCOTGRID-GLA



- Notice the blip after upgrade to DPM v1.6.3.
- New plugin will move into production gLite release - https://savannah.cern.ch/patch/index.php?1114, 1117

Double/triple/. . . counting

dCache:

- * Requires that site has pools for each VO (particularly LHC).
- * N.B. disk pool size $\leq$ partition size.

CASTOR:

- * Publishing static information via the top level BDII.

- Sites should check $\boxed{\textbf{consistency}}$ of the published data.
- Going through this process in GridPP.
- Report problems to me.

SRM 2.2 and GLUE 1.3

- With SRM 2.2, VOs will be able to **reserve spaces** on SRMs.
  - Static or dynamic depending on implementation.
  - Each SRM will advertise via space token descriptions, e.g., `ATLAS_AOD`.
  - It should be possible to account for storage at this level.
- Information still required by sites on how to set up these spaces.
  - Will old files automatically appear in new space?
- GIPs need to be **written** for all of the SRMs.
  - dCache v1.8.0 will not come with one.

Future plans

- **Additional views** of the data to be added, e.g.,
  - ROC level will show contributing sites, not individual VOs.
- Support the move to GLUE schema 1.3.
- Information system was designed for resource discovery, not accounting.
  - IS will **not scale** to publishing for each **user**.
  - The SRM knows the user level information.
    - ∗ The sensor could separate from the SRM, or part of the protocol (SRM v3?).

Summary

- **GridPP storage group** has extensive experience in deployment and operations of storage middleware at Tier-1 and Tier-2s.
  - Work continuing to understand the behaviour of these systems.
  - The group has greatly improved the grid accessibility to storage in the UK.
    - Communication has been key part of success.
- **Storage accounting** developed as a useful resource for different communities.
- Both the group and the accounting will have to evolve over time to deal with new features of the middleware and requirements from user groups.