GridPP
UK Computing for Particle Physics

Managing Tier-2 Storage

Greig A. Cowan

University of Edinburgh

Outline

1. What is a Tier-2?
2. GridPP experiences
   (a) Configuration, administration, monitoring
3. Open questions
4. Summary

What is a Tier-2?

In terms of storage, they can typically be characterised by:

- No tape backend.
- Relatively small amount of RAID5 disk (∼10-100TB).
- ∼3 high end servers.
- 1GbE external connectivity (most sites).
- Resources may have to be shared with non-WLCG users.
- Limited manpower (∼1 FTE).
  - Ease of configuration, management and monitoring are essential to maximise availability.

Deployed SRMs

	DPM	dCache	CASTOR	Total
WLCG	71	35	7	113
UK	12	7	1	20

- Obtained by querying BDII for instances of `/dpm`, `/pnfs` and `/castor` in the `GlueSARoot` field.
- Some sites may not expose this or may be using an alternative SRM (StoRM. . .).
- Discussion of Tier-2 storage must include DPM and dCache.

What is a Tier-2 becoming?

- In the UK, many Tier-2s have bought large batch farms and large quantities of disk.
- Some sites have decided that instead of buying dedicated disk and servers, they will use the “free” disk that comes with each WN.
  - i.e., $>$300TB (usable space) spread across $\sim$500 WNs.
- dCache has been shown to operate at this scale. Are there DPM sites with this amount of disk?
- Do sites understand what hardware to purchase?

Configuration

- YAIM used for initial basic installation.
- Admin typically performs final tweaks by hand (i.e., adding extra pools/filesystems, pool groups).
- Integration of dCache with YAIM has improved greatly over the past 9 months.
  - Sites using DESY repository.
- A large site has started trying to use cfengine with dCache. So far proving difficult.

Pool Management

- Pool draining essential.
  - dCache 1.7.0 comes with improved mecahnism for moving groups of files among pools.
  - DPM 1.5.10 has `dpm-drain`.
- Tools for checking/fixing namespace - disk pool synchronisation.
- Disk quotas
  - As VO disk allocations change (increase) at the Tier-1, they add new disk servers.
  - Tier-2s do not have this luxury. Ability to dynamically resize the disk pools would be very useful. Take storage away from VOs that underuse.

What happens when data is lost?

- Hardware or human error *will* result in data loss.
- Is this really a problem?
  - TDRs indicate that much of the data is either backed up at the Tier-1 or can be (fairly) easily regenerated.
  - What about user analysis data?
- In GridPP we recommend that sites:
  - obtain a list of SURLs of the lost files.
  - notify the VO managers using the broadcast tool, providing a link to the SURLs.

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

http://www.gridpp.ac.uk/wiki/DCache_Problems_and_Workarounds#PNFSid_to_SURL_mapping

Minimising data loss

Some common sense procedures:

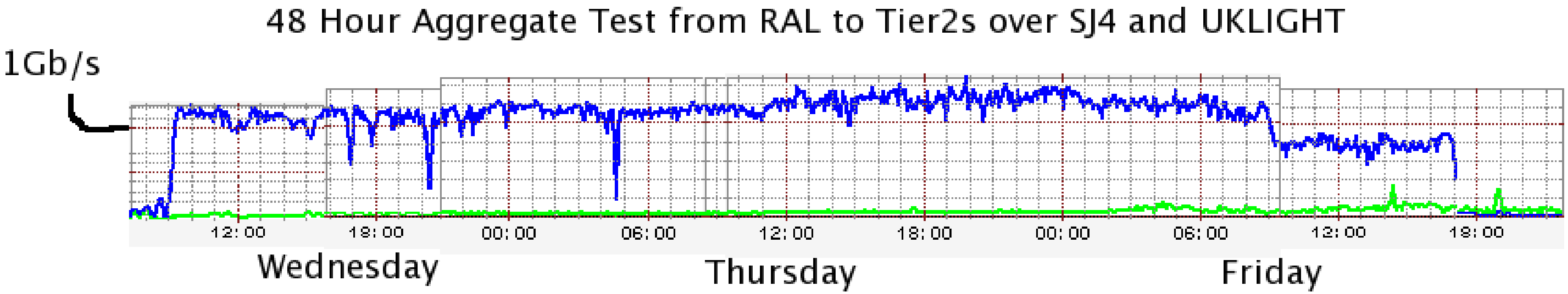
- Configure disks to use RAID5 or 6.
- If using WN disk then build in redundancy (resilient dCache).
- Dual power supplies.
- Dual fibre channel connections between server and disk.
- Backup namespace databases.

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

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

Optimisation

- XFS shown to be the FS that gives the greatest WAN transfer rates.
  - Most sites continue to use ext2/3.
- Higher transfer rates obtained with 2.6 kernel.
- UK wide transfer testing highlighted a number of bottlenecks.
  - Regional and local networks.
  - Firewall problems.

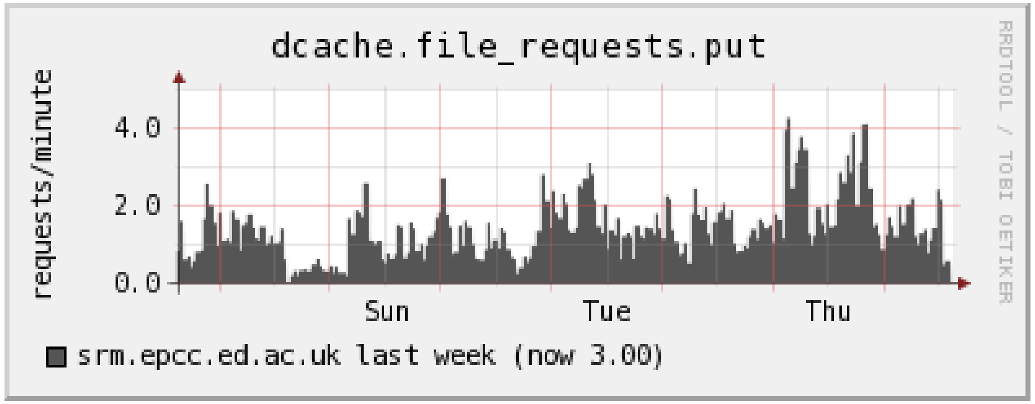


Optimisation

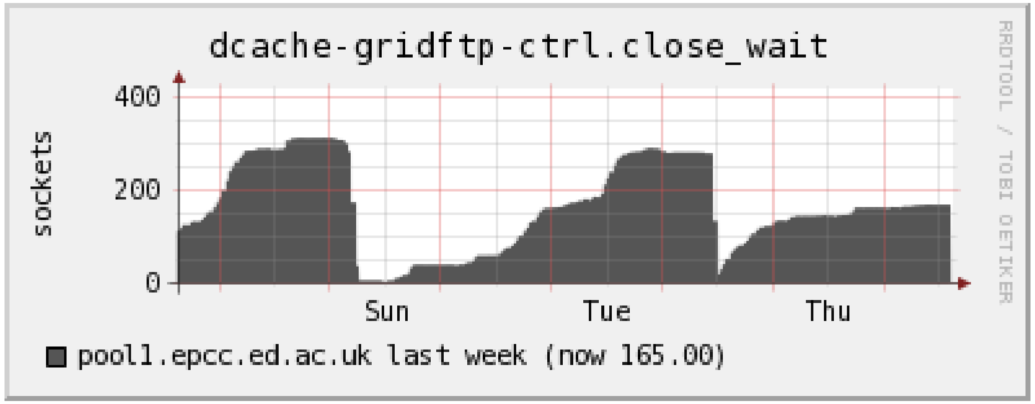
- Need to study local access to the storage from batch farm.
  - dCache shown to handle 50 file opens/sec. Will there be any limits with DPM?
  - What rates are the VOs expecting?
    - ∗ Initial tests suggest we can easily achieve 65MB/s (single write).

Local monitoring

- dCache has a status webpage. Difficult to use if pool number large.
- Specialised dCache and DPM monitoring using MonAMI.
  - Integrates with existing site tools (ganglia and nagios).
  - Can monitor individual processes; publish via RGMA.



srmPuts (last week)



CLOSE_WAITs (last week)

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

Remote monitoring

- SAM tests use the `lcg-cr, lcg-cp`...tools to probe the SRM.
- These tests also depend on BDII, LFC, RB.
- SAM tests do not give true measure of site availability.
- Suggest that new tests be developed that only depend on SRM/gridftp and dcap/rfio.
- What if an atlas pool fails?
  - From the cms viewpoint, the site is still 100% available.
  - Let ops write to all pools.

Storage GIPs

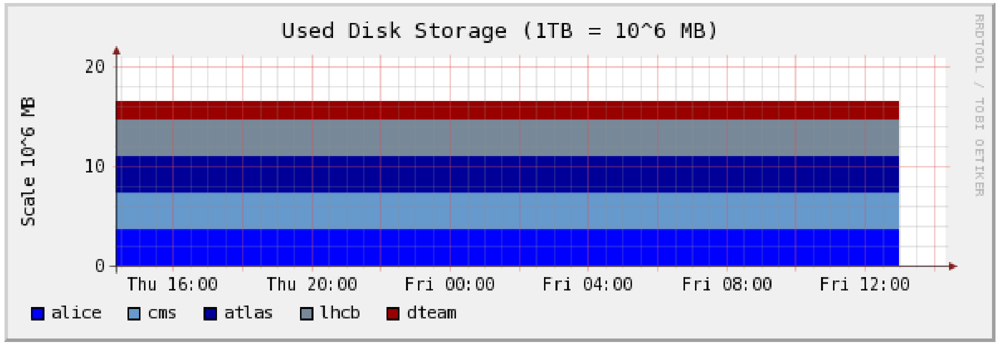
- If VOs share DPM pools or dCache pool groups then the standard GIP plugins do not correctly report the available and used space per VO.
- Where possible, sites should setup dedicated filesystems/pools for each LHC VO.
  - Easy for dCache since disk pool size $\leq$ partition size.
  - OS tools could be used to resize partitions.
- Where not possible, sites have deployed new GIPs in order to obtain used space information.

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

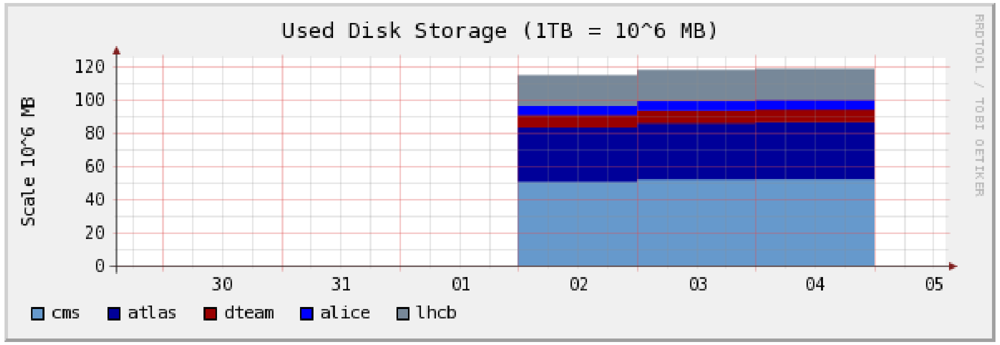
http://www.gridpp.ac.uk/wiki/DPM_Information_Publishing

Storage accounting

- Easy way for sites/ROCs to keep track of how their storage is being used.



NIKEF (last day)



UKI-ROC (last week)

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

SL4 and 64bit

- Recent hardware procurements have chosen 64bit machines.
- Sites are running dCache pool nodes on RHEL4 64bit (dCache written in Java).
- So far not been able to run the 32bit build of DPM on 64bit machines.

http://www.gridpp.ac.uk/wiki/Installing_SL3_build_of_DPM_on_SL4

SRM 2.2

- 18 GridPP sites now have a 2.2 endpoint.
- Storage Classes
  - Do we need to do anything to configure T0D1?
  - Will DPM offer more than just T0D1 storage?
- Space Reservation
  - Will there be some negotiation between the sites and VOs before data is written?
  - Does it happen dynamically?
- Interoperability testing required.
  - DPM ↔ FTS ↔ dCache (and CASTOR)

Other questions

How do Tier-2s manage. . .

1. datasets that are never used?
2. corrupted/incomplete datasets?
3. disk pools that are full?

} Are these site problems?

4. empty disk from VOs that do not write data? ← Quotas in the middleware?

5. What do Tier-2s do if all (or part) of their SRM fails? Clear procedure required.

Summary

- Good understanding within GridPP of how to setup basic Tier-2 SRM (see wiki).
  - Still gaining experience in setting up a large site ($>$100TB).
- Further investigation of batch farm access to the storage is needed.
- Clear technical and procedural instructions required in the event of data loss.
- SRM availability monitoring could be improved.
- Need to understand implications of SRM 2.2 on Tier-2s.
- Need to discuss how VOs and Tier-2s will interact when problems occur.