HEPiX Spring 2006 Storage Report

11 Apr 2006

Document Type	GridPP Storage Level 2 deliverable
Document Identifier	gridpp-storage-D10-2-????
Associated Level 1	Storage Task 10
10-11.04.06	First draft. J Jensen
12.04.06	Updated dCache section with comments from Greig. Also expanded SRM2 plans a bit.
13.04.06	Added some more non-storage stuff from my notes.

This document is a report on the HEPiX Spring conference, hosted by CASPUR, Rome. It is a Level 2 deliverable; it's associated Level 1 deliverable is Task 10 (report on Local Storage Management).

1 Attendees

From GridPP storage, the following people attended[1]:

- Greig Cowan, Edinburgh;
- Graeme Stewart, Glasgow (OK, he's technically not storage, but we like to think of him as one of us :-);
- Jiří Mencák, RAL Storage;
- Jens Jensen, RAL Storage.

2 Issues identified or discussed during workshop

2.1 General storage issues

2.1.1 SRM protocol

- The workshop at FNAL. CERN will try to push their changes through, e.g., functions for prestaging files. Likely to lead to SRM version 2.2.
- JLAB reported that they are participating in the "SRM in SciDAC2 proposal"

[1] Paul Millar and Martin Bly also attended the whole thing but they are of course not storage either... and there were other GridPP folks but some of them were hiding in the GDB some of the time, so they don't count :-)

2.1.2 CASTOR

Olof and Sébastien attended the meeting. Sébastien gave a talk on CASTOR.

2.1.3 XFS

From the work done at Glasgow and Edinburgh running both dCache and DPM pools on XFS (and presented at HEPiX), we know that the SE's performance improves. Indeed, this may be necessary to meet the target rates for the Tier 2s. However, XFS is unlikely to be supported in SL4, which will be bad for Tier 2s with XFS storage pools when they upgrade it to SL4. Thus, it is uncertain whether we can recommend using XFS for storage pools.

2.2 DPM issues

Jean-Philippe did not attend the workshop, although we had expected him to.

2.2.1 SL4

We need to test building DPM on SL4.

2.3 dCache issues

Both Patrick Fuhrmann and Michael Ernst attended the storage day. We discussed all the items below with Patrick.

2.3.1 `du`'ing pnfs

The PNFS database cannot be used to determine the amount of storage used per VO. To do this, it is neccesary to run a normal du on the filesystem, and this command needs to recurse through the filesystem (as on any other filesystem).

`du`'ing the filesystem is relatively inefficient because it needs to recurse down through the file tree. It would be better to say something like

```
select sum(size) from files where vo = 'atlas';
```

or somesuch, but we can't do that with the current database.

Greig points out that for the space used, at least `du`ing provides the right result, because each VO has its own directory.

For space *available*, it's slightly more subtle because VOs may or may not share pools. Of course it is good if VOs *can* share pools because we have a lot of them, and they are supposed to be dynamic. Which in turn leads to the questions of quotaing, but we didn't discuss that much.

Related to that, the PNFS database is not human readable. This will change when Chimera is deployed, expected by the end of the year.

2.3.2 GridFTP Logging/monitoring

We talked to Patrick about the logging problem, that data needed for GridView is split across two not necessarily consecutive lines. He said it should be possible to provide all the required data in one of the lines — and then we could ignore the other one, or it would go away.

Greig asked for access to the statistics module.

2.3.3 Performance

Patrick gave a talk about dCache in general and perfomance in particular. It is expected to scale well in most directions (e.g., number of files), but the main remaining issue is with the number of requests it can handle per second; he reported 50 Hz (on, presumably, high spec. hardware?).

2.3.4 SRM 2.1

The 2.1 interface (promised for "mid January") for dCache is still being developed. Patrick reported that reservations do not work yet, but Jens asked if GridPP could have it anyway - we only really need upload/download for now, and the sooner we start testing interoperability with CASTOR and DPM 2.1s, the better. Patrick would speak to FNAL (Timur). We were told that FNAL had started testing interoperability, but it's always better to test this ourselves.

The Cunning Plan™ is that if and when we get the interface for dCache, we can start doing transfer tests using the 2.1 interface between all the GridPP SRMs: dCache, DPM, and RAL's CASTOR. Only remaining problem is that FTS doesn't support 2.1.

2006-04-12 Update: had a note back from Patrick saying he'd spoken to Jon Bakken and they (FNAL) would prefer to wait two weeks before releasing it for testing, and that he (Patrick) thought that was OK.

2.3.5 VOMS support

Patrick mentioned that VOMS support is being worked on, and there are some slightly problems right now with the userid mapping, but otherwise a release is expected soon.

2.3.6 YAIM integration

DESY asked if Owen could go to DESY for two weeks to work with them on improving the YAIM-to-dCache integration. GridPP collectively agreed that this was a good idea and will fund this; DESY will pay his accommodation. Date tbd, but soon.

2.3.7 Using HTTP(S) for data transfers

dCache supports HTTP but apparently nobody uses it (Graeme reported this in his talk). We should test it, and test if it downgrades the encryption (SSL renegotiation to NULL cipher) when doing the data transfer – or whether it uses plain HTTP for the transfers.

2.3.8 Matthias

Matthias de Riese has left, and they are now interviewing for a replacement. This is likely to affect the support we are getting from DESY because Matthias was a good egg and interfaced frequently with Owen (or perhaps the other way around).

2.4 Miscellaneous non-storage

...that will affect storage middleware. Or not.

2.4.1 SL support

2.4.2 IHEPCCC

Randall Sobie from International High Energy Physics Computing Coordination Committee (IHEPCCC) asked about non-HEP usage of HEP resources. Jens told him about, and sent him some further info on, biomed usage of GridPP resources.

2.4.3 64 bit machines

From work done at FNAL, we know running 32 bit apps on 64 bit OS can improve performance, and performance is improved further if the software is recompiled for 64 bit (assuming it works). It may be useful to investigate performance improvements of DPM on 64 bit hardware. At the moment only RHUL has 64 bit machines running storage middleware - and RHUL is a DPM site.

SL4 is likely to support x86_64 but not IA64.

2.4.4 Networking

The GARR talk (Italian NREN) promised the ability to reserve bandwidth which would be a nice thing to do — for example, Don Petravick wanted SEs to be able to reserve bandwidth between each other for 3rd party transfers. In discussion with most networking people, they say "nice idea, but not something we can practically do."

2.4.5 Authentication

There seems to be enough interest in authentication technologies to keep the track going, and to create an area in the HEPiX wiki/plone.

2.4.6 Skype

There was talk about using Skype which violates (some?) NREN AUP. Jens said he'd been told U Oxford had made a deal preventing Oxford nodes becoming supernodes. People expressed interest. Would investigate.

3 Recommendations

TODO!

4 Conclusion

HEPiX was useful; apart from numerous GridPP talks (storage and metadata middleware, plus ScotGrid and Tier 1 reports), there were plenty of highly relevant issues raised throughout the workshop.