SRM support in GridPP3

J Jensen (RAL), G Cowan (Edinburgh)

The GridPP Storage Group deploys and supports storage middleware, to interface storage systems to the Grid, to enable the UK to successfully participate in Grid collaborations in general and WLCG in particular. We have deployed storage middleware to all Tier 2 sites in the UK and taken a very proactive approach to testing and support, and thanks to the group's efforts, the UK is at the forefront of storage deployment, with the largest deployment (21 sites) in WLCG. We continue to work closely with upstream[1] software maintainers and we know that also other countries find our work useful, because we have members from some countries, and because we have had positive feedback from others.

The overall *strategy*, into which this fits, is:

The GridPP storage group will provide and support storage middleware necessary and sufficient to Grid-enable storage at every GridPP site.

Recall that "Grid interfaces" means *Storage Elements*, or SEs. An SE has a control protocol, usually a version of the SRM protocol, data transfer protocols (usually GridFTP plus one LAN protocol, but experiments have started asking for xrootd), and information publishing protocols. To implement the strategy, we thus have to meet the following high-level goals:

- Provide SRM interfaces to disk and tape storage systems at all sites in GridPP (tape at Tier 1 only).
- Provide data transfer protocols with emphasis on performance, robustness, and security.
- Provide information publishers for service discovery (to enable higher level Grid services to locate an SE), and accounting (giving a high level overview of resources).
- Ensure the storage middleware can be easily deployed and redeployed at Tier 2 sites, configured, monitored, maintained, upgraded, and debugged. Additional investigations and recommendations are needed to improve the reliability of installations, for example by deploying redundant services, and to provide advice on optimisations, essential for high speed and volume transfers.
- Ensure that all components of storage middleware is standards compliant and interoperable with other storage middleware used throughout WLCG (and, eventually, the National Grid Service).

In deciding how to meet the goals over the long term, the following points have been considered:

- Middleware will and must adapt to changing requirements and new version of protocols;
- Features, quality, and stability of the implementations, and the size of the knowledge base within GridPP and in WLCG as a whole. Sites have enough local variations in their infrastructure that no single implementation meets the needs of all sites.
- We save effort by supporting the *minimal* set of implementations that meet the goals.
- We save further effort by ensuring that the implementations are *coordinated* across GridPP: if every site with implementation X runs the same version with the same setup (subject to local hardware variations, policies, and restrictions), the effort to support implementation X is much less than if all sites had free reign to run whatever they like.
- It is precisely the need to maintain coordination across GridPP that is the principal reason why GridPP needs its own storage support instead of relying wholly on upstream support. Additional reasons are that upstream support is often limited, provided by the developers only, and that it helps to be "local" to the users (i.e., in the same country—it is feasible to send a support person by train for the cases when remote access won't do).

Ongoing tasks for GridPP support staff include

- Help sites maintain, upgrade, and, when necessary, migrate the existing infrastructure;
- Provide QA and upstream feedback;
- Maintain the GridPP storage knowledge base.

[1] Upstream, i.e., developers' support staff (which is usually the developers themselves).

Site infrastructure is sufficiently diverse that no single solution meets the needs of all sites. Storage Elements vary in size from a single machine (a few hundred gigabytes) to Manchester's 1000 nodes, and in quality from bare disk to high-availability RAID arrays. RAL has a petabyte tapestore, Edinburgh has 10TB SAN that may be better managed like tape. Some sites have 64 bit or dual core processors, and the software is developed and largely run on 32 bit single core processors. Lancaster and Tier 1 have optical networks which permit special transfer protocol optimisations. Larger sites have dedicated storage administrators with lots of experience, most smaller sites have a fraction of someone's time, and that someone usually doesn't have much storage experience.

GridPP chose DPM and dCache to Grid-enable storage at Tier 2 sites (for further background, please see Task 11-12 deliverables, an early draft of which is included below, "Status of SRM support"). We also need to support the SRM implementations for CASTOR for Tier 1. These three SRMs together almost meet the needs of all UK sites. Only dCache meets the needs of the largest sites, but does require dedicated administrators, particularly to optimise its performance. DPM was written for the smaller sites and to require relatively little support, and it fills that need very well. The principal remaining gap is support for distributed filesystems, which isn't covered well by resilient dCache, nor by DPM. It is possible that either solution may mature in this area within the next three years, but sites have started looking at dedicated distributed filesystems to bridge the gap. Bristol is currently looking at GPFS, QMUL at PoolFS, others at Lustre. It is only a matter of time before we may need to support one or more distributed filesystems. Further gaps were identified in deliverables for Level 1 Tasks 09 and 10.

1 Status of SRM support

[The following is a draft of the deliverable for Tasks 11 and 12. Authors: G Cowan, J Jensen.]

1.1 Multiple SRM solutions

SRM deployment at GridPP Tier 2 sites started over 18 months ago. At that time the Tier 1 had already deployed dCache [1] as their SRM solution; the decision to move to CASTOR was not taken till early '06 (dCache is not able to interface to CASTOR, so CASTOR needs is own SRM implementation). Tier 1 had already put a lot of effort into getting dCache experience with about 1 man-year of effort prior to the production deployment. The Tier 1 experience with dCache was influential in the decision to deploy dCache at the Tier 2s, and has also been useful since then, because Tier 1 has worked closely with the storage group on dCache and has often helped Tier 2s. More importantly, at that time dCache was the only middleware product that could be deployed in production at Tier 2s. CASTOR [4] was too heavyweight and does not manage disk-only storage well (the current version has problems when the disk cache gets full). At that time the only other product that provided an SRM interface to storage was DPM [2]. However, at that time DPM was not ready for deployment on production systems.

Initially SRM deployment within the UK was driven by a need for some Tier 2s to become involved in LCG Service Challenge (SC) 3 work. A requirement for this was that these sites had an operational SRM interface to their storage which the participating experiments could use. Therefore early adopter sites were chosen who would deploy an SRM. At this time dCache was the only viable solution. Edinburgh, Lancaster and IC-HEP were all chosen to deploy dCache.

Once DPM became a viable solution for sites to deploy, the GridPP storage group wrote up some guidelines regarding how sites should choose which product to deploy [5].

In the end, 7 of the 20 GridPP Tier 2 sites chose to deploy dCache. These can be broken down as follows:

- Early adopter sites (Edinburgh, Lancaster, IC-HEP)
- Local expertise was in dCache (RAL-PPD)
- Federated Tier 2 wanted to use dCache to have a shared experience of running single system (NorthGrid).

The remaining sites chose to deploy DPM after successful trial deployments had occurred at Edinburgh and Glasgow. These tended to be smaller (in terms of level of disk provided) sites.

1.2 Consolidation of SRM solutions

It is often asked whether the SRM implementations can be consolidated. This is not practical, for several reasons:

- There is no migration solution for moving from DPM to dCache, or vice versa.
- We may even need to support additional storage solutions to meet all the needs of the Tier 2 sites; e.g., INFN's StoRM, xrootd+SRM ...; possibly distributed filesystems.
- We have shown over the past two years that we are able to support the two solutions. Evidence can be seen for this in the fact that all sites now have an operational SRM and all sites have had this tested during the ongoing GridPP service challenge work.

- Generally, not too much support is needed for the day-to-day running of site SRMs. Where problems do occur is during times when new releases are made and there is pressure to test out fresh installs/upgrades of dCache/DPM within a limited period. A measure of the support level can be seen from the

- Vendor lock in: at the moment we have 1/3 of sites depending on dCache, and 2/3 on DPM. Relying wholly on a single provider can be unhealthy: if a problem is discovered with one of the solutions in production then the whole country goes red on the map. It is already a high risk to rely on only two solutions to support GridPP, but other countries do this as well. Of course the implementations are carefully tested before they are deployed, but occasionally problems are discovered on the implementations that are running in production.

- We depend on the interoperability between DPM and dCache for the successful operation of storage services in GridPP, for transfers to and from sites outside the UK. Having both in the same country where we can easily coordinate testing and debugging helps us identify and locate problems. For example, GridPP was the first to discover a serious performance problem when transferring data from dCache to DPM; discovering this early and working with the developers to find a solution was extremely useful. If we had not discovered it during testing, it would have been a serious problem during the service challenges.

1.3 Other considerations

GridPP has the largest deployment of storage middleware, and during GridPP2 we have gathered a lot of experience with both solutions, and with interoperability between them, and between the two and CASTOR. This information is also helpful for other countries, which also helps us interoperate with other countries. For example, Estonia is a Tier 2 site for UKI, and it is helpful that they, too, can access our information. It is not all one way: we conversely learn from other countries: for example, INFN CNAF have just discovered problems with transferring files larger than 2GB from CASTOR to DPM, and this problem does not seem to occur with dCache, nor the other way around. This problem wasn't discovered earlier because most tests are done with 1GB files.

Our knowledge base has also been helpful for the software developers. dCache used a lot of our information when they updated their documentation.

2 Storage support outlook

- Storage support has changed over the past 9 months or so, as the focus has moved from deployment to maintenance, and as the community has gradually gained more experience. Systems appear to have moved into a relatively stable state, into a "background noise" of regular support problems.

- A certain number of people *is* required to support 20 Tier 2 sites with 200TB disk storage, and a Tier 1 with a 5PB tape robot. Currently, much support continues to come from unfunded effort (Graeme (GridPP data mgmt), Owen (DESY), Tier 1 dCache experts, and RAL's tapestore group). This is a concern and is brought into acute focus with Jiri also leaving. A measure of the UK support level can be gleaned from the mailing list archives [11] and from the minutes of the weekly storage phone conferences [12].

- Another concern is the storage shortfall. Tier 2 sites do not meet, and may not be required to meet, their MoU storage targets, but they are still short of providing enough storage to meet the experiments' storage to CPU ratios.

- We still have stability problems with site's storage elements, as can be seen from the current storage monitoring plots [6] but this is partly to be due to the odd site having problems for a day or two them coming back online again (troughs in the plot are also due to problems with the site BDII that is being queried as well as problems in the monitoring that have now been addressed).

- Even though sites themselves have got experience with their storage solutions, there are questions they cannot themselves resolve, and in particular, sites do now have the effort to do upgrade testing.

- We are now working on developing and improving the monitoring and accounting systems [7]; further improvments to local monitoring for dCache and DPM [8] will help sites keep track of their SRMs in operation.

- An important part of the support post has been in researching the optimal configurations for dCache and DPM [9, 10]. This has improved the reliability of these services and the performance of them during SC transfer tests. Such optimisation work is essential for the performance of the SRMs, and needs to be updated as the implementations are improved.

- We intend to get involved in development of test suite for dCache (will be done in collaboration with DESY). This should help speed up testing and certifcation and improve the quality of releases that are given to the sites which should further improve the upgrade successes.

- RAL's tapestore group has developed the SRM version 2 for CASTOR, partly with GridPP funded work. Ongoing support for this will be partly met by RAL, but the need for interoperability testing from CASTOR to DPM and dCache will have to be met by GridPP.

The form that future support will take will be somewhat determined by the answers to these (and other) questions:

- Changes in dCache/DPM
  - Licensing issues
  - New features in both
  - Changes to support different infrastructure (e.g., multi-core processors).
- New storage solution developments
  - StoRM
  - xrootd with SRM interface
  - Status of Andrew McNab's slashgrid project?
  - Should we evaluate these with the intention of deploying them? They do fill gaps that the existing solutions don't.
- Distributed storage filesystems (GPFS, Lustre ...)
- How do we deal with sites that want to utilise WN disk in their production storage environment? dCache is the only way to go at the moment, and it doesn't work well enough yet to put into production. dCache is also currently focusing on different areas, so the gap isn't likely to close in the next couple of years.
- NGS convergence. NGS have started investigating DPM as a common storage solution. If DPM is taken up by NGS, the DPM support may be spread across GridPP and NGS (but of course will have to support many more sites, with even more diverse infrastructures).

3 Additional information

3.1 International Standards

SRM The SRM protocol is essential for the success of GridPP — it is the accepted control protocol for interaction with storage services, and we must have implementations that support the version currently used in LCG. Thus, as LCG decides to move to the next version of the SRM protocol (currently from 1.1 to 2.2; but version 3.0 is close to being finalised and is a future option for LCG), there is a need to safely migrate existing files, and coordinate the update so everyone uses the same protocol.

GLUE

In the UK, thanks to the close collaborations within GridPP, we have gained a lot of operational experience with the GLUE information schema, and have often been able to contribute to the standardisation process.

3.2 WLCG Deployment

The following table shows the distributions of SRM implementations throughout WLCG. Apart from Lawrence Berkeley National Laboratory (LBNL) and Jefferson Lab (JLAB), no site is known to run a fourth implementation in production (the StoRM implementation is currently not in production, but we expect it to be used by INFN (Italy) once LCG moves to SRM version 2.2). Important points:

- The UK has the largest number of deployed SRMs (21).
- Belgium, Bulgaria, Canada, Spain, France, Greece, Italy, Holland, Russia, and the UK all use both dCache and DPM.
- Italy, Spain and the UK support all 3 SRMs (DPM, dCache, and CASTOR — Russia also has CASTOR but not in production).

Table 1: List of SRM implementations in WLCG

	dCache	DPM	CASTOR
.uk	**8**	**12**	**1**
.at	0	1	0
.au	0	1	0
.be	1	1	0

	dCache	DPM	CASTOR
.bg	1	1	0
.br	1	0	0
.ca	1	1	0
.ch	0	1	1
.cn	1	0	0
.com	0	1	0
.cz	0	1	0
.de	10	0	0
.edu	5	0	0
.ee	1	0	0
.es	2	2	3
.fr	1	4	0
.gov	1	0	0
.gr	1	3	0
.hr	0	2	0
.hu	0	3	0
.in	0	2	0
.it	3	8	1
.jp	0	2	0
.kr	1	0	0
.lv	0	1	0
.nl	1	1	0
.org	1	1	0
.pk	0	1	0
.pl	0	3	0
.pt	1	0	0
.ru	2	7	0
.su	0	1	0
.tw	0	5	1

References

[1] *http://www.dcache.org*

[2] *http://www.gridpp.ac.uk/wiki/DPM*

[3] *http://storage.esc.rl.ac.uk/documentation/html/dpm-eval/dpm-eval.html*

[4] *http://castor.web.cern.ch/castor/*

[5] *http://www.gridpp.ac.uk/wiki/Which_SRM*

[6] *http://www.gridpp.ac.uk/storage/status/gridppDiscStatus.html*

[7] *http://goc02.grid-support.ac.uk/accountingDisplay/view.php*

[8] *http://www.gridpp.ac.uk/wiki/MonAMI_DPM_plugin*
http://www.gridpp.ac.uk/wiki/MonAMI_dCache_plugin

[9] *http://hepix.caspur.it/spring2006/TALKS/5apr.cowan.opt.pdf*

[10] *http://*

[11] *http://www.jiscmail.ac.uk/lists/GRIDPP-STORAGE.html*

[12] *http://agenda.cern.ch/displayLevel.php?fid=338*