SCOTGRID AND THE LCG

G. A. Cowan*, P. J. Clark, S. Thorn, University of Edinburgh, UK
J. Ferguson, D. Martin, F. Speirs, G. Stewart, University of Glasgow, UK
L. Heck, M. Nelson, University of Durham, UK

Abstract

ScotGrid is a distributed Tier-2 computing centre formed as a collaboration between the Universities of Durham, Edinburgh and Glasgow, as part of the UK's national particle physics grid, GridPP. This paper describes ScotGrid's current resources by institute and how these were configured to enable participation in the LCG service challenges. Particular emphasis will be placed on the deployment of storage resource management middleware. In addition, we outline future development plans for ScotGrid hardware resources and plans for the optimisation of the ScotGrid networking infrastructure. It is intended that this work will feed information into the GridPP community such that all sites can improve their current setups, thereby enhancing the quality of service that can be provided to Grid users.

Keywords: ScotGrid, Tier-2, GridPP, LCG.

INTRODUCTION

The UK's Grid for particle physics (GridPP) [1] started in 2001 with the aim of creating a computing grid that would meet the needs of particle physicists working on the next generation of particle physics experiments (i.e. the LHC). To meet this aim, participating institutions were organised into a set of Tier-2 centres according to their geographical location (see Figure 1). Here, we discuss the setup of ScotGrid [2] within GridPP and the wider LCG project, emphasising its role in deployment, monitoring and testing of storage resources as are suitable for a Grid environment.

Storage on the Grid

Due to the large volume of data that the LHC will produce, it is essential for the operation of the LCG that there is both sufficient storage capacity across the Grid and that the heterogeneous collection of storage resources is accessible to middleware applications via a common application protocol interface (API), the storage resource manager (SRM) [3]. GridPP is responsible for the deployment of SRM interfaces to storage at all UK Tier-2 centres involved in LCG. ScotGrid plays a key role in this deployment due to the work of personnel in the storage management and data management fields. GridPP has chosen to use two different middleware products to enable Tier-2s to manage their distributed collection of disk servers under a single namespace, with one method of accessing this space being through the SRM interface:

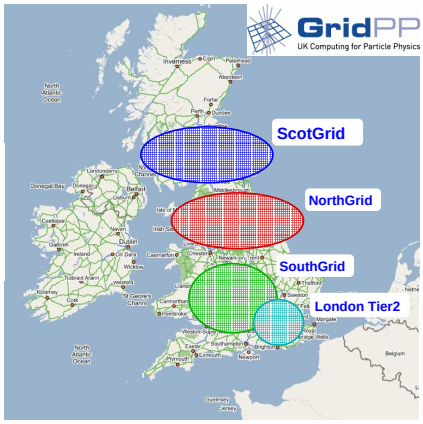

Figure 1: Federated Tier-2 centres and their compositional sites within the GridPP collaboration.

- **dCache -** Jointly developed by DESY and Fermilab to provide a highly configurable and scalable mechanism for managing a set of disk pools and tertiary storage. The SRM version 1 interface to dCache has been developed by Fermilab, enabling dCache to be used in a distributed Grid environment [4].

- **DPM** - Developed at CERN to provide a system for managed storage of disk servers, particularly at Tier-2 sites without a tape storage system and who require low system administration overhead. Scalable system since multiple disk servers are aggregated to provide a single namespace. Provides an SRM v1 interface to the storage, as well as some SRM v2 functionality [5].

ScotGrid Tier-2

We provide a summary of the available resources at each of the ScotGrid sites, paying attention to the setup of available storage resources.

Edinburgh: Edinburgh provides the second largest Tier-2 disk resource within GridPP, serving 31TB for use by LCG VOs. The storage is split such that 28TB is managed by dCache and 3TB managed by DPM, both providing SRM v1 interfaces for the VOs to use. The use of two storage resource managers to control access to the disk is key to Edinburgh being able to act as a knowledge base regarding storage within GridPP. The dCache

* g.cowan@ed.ac.uk

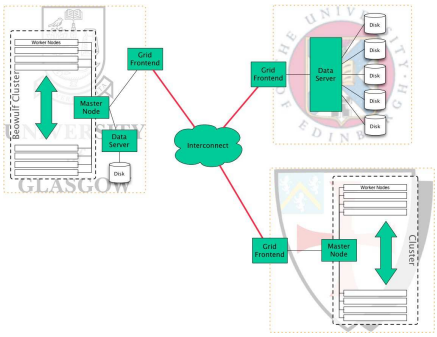

Figure 2: Schematic layout of ScotGrid Tier-2.

setup can be seen in Figure 3. A single dCache *administration* node hosts the main dCache services (PNFS and billing databases, SRM) while the dCache pools are hosted on a single *pool* node. Using fibre channel connections, the pool node is attached to a IBM Dual FAStT900 22TB RAID (level 5) disk array which provides a level of resiliency against common malfunctions, protecting against data loss. An additional 8 TB of storage is NFS mounted on the pool node from the University Storage Area Network (SAN) which uses a RAID level 5 configuration.

Edinburgh operates a second production SRM, running DPM over two nodes, one of which was previously the GridFTP [6] server for the site, but has now been migrated into our DPM instance. Additional storage is NFS mounted from the SAN onto the node running the core DPM services (DPM namespace, SRM interfaces). Running two SRMs places Edinburgh in a good position within GridPP, allowing us to study the interaction of the two SRM services.

In addition to the storage, Edinburgh has 7 CPUs for processing LCG jobs as well as hosts providing LCG front end services (CE, MON, UI) and test platforms.

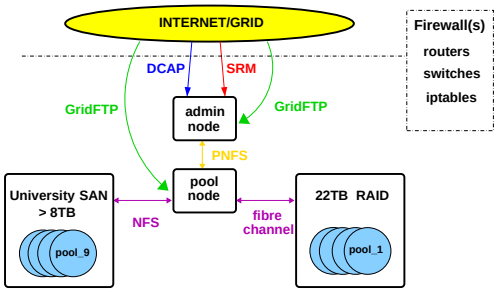

Figure 3: Schematic layout of ScotGrid dCache at Edinburgh, showing the available nodes and protocols that can be used to access the data stored within the dCache.

Glasgow: Currently providing 200 CPUs for use by LCG VOs and additional hosts for testing purposes. DPM manages the 5TB of RAID level 5 storage, spread across a two pool nodes, each allowing access to the disk via GridFTP (see Figure 4). A third pool node with 3 filesystems is planned for deployment, providing additional storage to supported VOs.

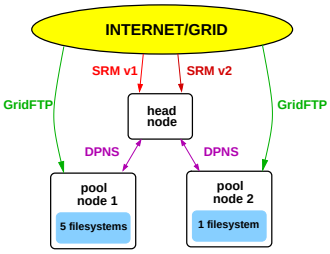

Figure 4: Schematic layout of the DPM install in Glasgow.

Durham: Durham provides 100 hosts, each with dual-2.2GHz Pentium 4 processors, 2GB of memory and 30GB local disk. Single 2TB filesystem (RAID level 5 disk) is accessible via the SRM interface provided by DPM, running on a single node.

STORAGE DEPLOYMENT

ScotGrid has played a key role in the deployment and testing of SRMs within the GridPP project. The understanding that has been gained during this process has been extensively documented in the GridPP wiki, and has resulted in a effective and complete rollout of dCache and DPM to all LCG sites within GridPP.

Monitoring

It is GridPP's aim to build a production level Grid that can be used in the next generation of particle physics experiments. Monitoring plays an essential part of this deployment as it allows a record to be kept of how sites are progressing and indicates when sites experience technical problems. This is particularly important in the era of SRMs since it is a relatively new technology of which Tier-2 sites have little experience. Through the work done at Edinburgh, ScotGrid has taken on the task of developing and operating a monitoring framework, based upon the information that sites already publish via the GLUE schema [7] in the LCG information system. By combining Perl, HTML, ROOT and C++ we are able to show the current status of SRM deployment in the UK while maintaining historical records which are useful for tracking long term progress. Figures 5 and 6 show the total storage capacity per Tier-2 site within GridPP on a daily basis [8]. Figure 6 clearly shows the gains that have been made in terms of available storage since the monitoring began. In particular, it should be noted that the GridPP Tier-2s have achieved greater than 150TB accessible via the SRM interface. What the Figure also shows are the fluctuations that occur in the availability to access sites storage, observed most frequently during upgrades to site SRMs.

The information system is structured such that each service at a site published information about itself to the site-central BDII (Berkeley Database Information Index). The information that each service provides is defined by the GLUE schema. Examples for storage elements would be

items such as the transfer protocols it uses for file transfer, paths to the storage areas on the host and dynamic information about the used and available storage space. This information can be obtained from the site BDII via an LDAP query. This functionality of the information system is exploited by our monitoring software which consists of a Perl script that queries the site BDII for the dynamic storage information and the flavour of SRM that is being used (determined from the storage paths). After extracting this information, script outputs to an HTML file, the result of which seen in Figure 5. Using C++ code integrated with the ROOT analysis package, we are able to create daily plots of the available storage at each Tier-2 site and aggregated to show the storage available at each of the 4 federated Tier-2s.

Maximum Capacities for LCG (TB) Current Published Capacities via LCG (TB)

Location	2005	2007	Capacity	Available	Used	Used (%)	SRM	SRM File Type
ScotGrid								
Durham	5	5	1.88	1.32	0.56	29	dpm	permanent
Edinburgh	70.5	70.5	28.543	15.92	12.623	44	dpm, dcache	permanent, permanent
Glasgow	14.8	14.8	4.505	2.739	1.766	39	?, dpm	permanent, permanent
ScotGrid Totals	**90.3**	**90.3**	**34.928**	**19.979**	**14.949**	**42**		
NorthGrid								
Lancaster	86.7	86.7	1.996	0.842	1.154	57	Classic SE	permanent
Liverpool	80.3	80.3	2.385	2.309	0.076	3	dcache	permanent
Manchester	372.6	372.6	45.507	45.414	0.093	0	dcache	permanent
Sheffield	3	3	4.247	4.076	0.171	4	dcache	permanent, permanent
NorthGrid Totals	**542.6**	**542.6**	**54.135**	**52.641**	**1.494**	**2**		
SouthGrid								
Birmingham	9.3	9.3	1.76	1.58	0.18	10	dpm	permanent
Bristol	1.9	1.9	0.183	0.167	0.016	8	dpm	permanent
Cambridge	4.4	4.4	3.22	3.06	0.16	4	dpm	permanent
Oxford	18.5	24.5	3.16	2.41	0.75	23	dpm	permanent
RAL-PPD	11.8	24.4	6.828	6.356	0.472	6	dcache	permanent
SouthGrid Totals	**45.9**	**64.5**	**15.151**	**13.573**	**1.578**	**10**		
London								
Brunel	21	21	0.879	0.814	0.065	7	dpm	permanent
IC-HEP	28.3	93.3	0.31	0.283	0.027	8	?, dcache	permanent, permanent
IC-LeSC	??	??	0	0	0	0	dcache	permanent
QMUL	28.5	58.5	15.92	14.61	1.31	8	dpm	volatile
RHUL	23.2	23.2	2.73	2.47	0.26	9	dpm	permanent
UCL-CCC	0.7	0.7	1.072	0.889	0.183	17	?	volatile
UCL-HEP	0.7	0.7	0.006	0	0.006	100	dpm	permanent
London Totals	**101.7**	**196.7**	**20.917**	**19.066**	**1.851**	**8**		
GridPP Tier-2 Totals	**780.5**	**894.1**	**125.131**	**105.259**	**19.872**	**15**		
GridPP Tier-1 RAL	**191**	**1100**	**88.43**	**36.073**	**52.357**	**59**	**dcache, dcache**	**permanent, permanent**
GridPP Totals	**971.5**	**1994.1**	**213.561**	**141.332**	**72.229**	**33**		

Figure 5: Snapshot of the current status of storage resources at each Tier-2 site within GridPP. Automatic updates occur every 10 mins.

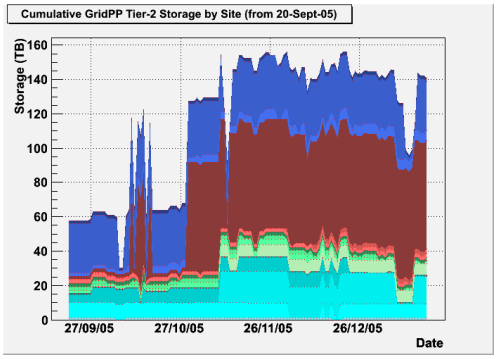

Figure 6: Archived daily status of storage resources at each Tier-2 site within GridPP. Blue colours correspond to the ScotGrid sites, red to NorthGrid sites, green to SouthGrid and cyan to London Tier-2 (cf. Figure 5)

With the release of LCG 2.7.0, both dCache and DPM come with plugins to the Grid Information Provider (GIP) [9], allowing the SRM to publish information about used and available storage on a per-VO basis. Previously, this was not possible and only total used and available storage space were published. The GridPP monitoring software will adapt to accommodate this change by showing these per-VO storage statistics. It is important to note that sites are free to chose how they allocate their storage resources to VOs and may setup disk pools (managed by DPM or dCache) to which multiple VOs can write. If this is the case, then with the current GIP plugins it will not be possible to calculate the total storage available at a site by adding the available space of each VO since this will lead to double counting. Alternative methods will have to be found.

SRM testing

As part of the work towards providing the reliable Grid service that LHC requires, ScotGrid has been actively participating in the LCG Service Challenges (SC). These aim to test particular aspects of site setups in order to observe the interaction of middleware components, networking and the individual site setups. The role of ScotGrid has become particularly important with respect to the testing of file transfers between deployed SRMs during SC3 and in preparation for SC4 due to the presence of GridPP personnel specialising in data and storage management at the Universities of Glasgow and Edinburgh respectively. Testing was initially carried out using the dCache SRM client, however, it is essential to study the inter-operation of SRMs using the File Transfer Service (FTS) component of the gLite suite of middleware [10], since this will be used to transfer files when the LCG goes into full production. An example of work done during the testing program can be seen in Figure 7, showing the achieved data transfer rate when using FTS to transfer 1TB of data from Edinburgh to the UK Tier-1 centre at Rutherford Appleton Laboratory. GridPP has set a target rate of 150Mb/s for such Tier-2 to Tier-1 transfers and the 440Mb/s observed here clearly exceeds this target.

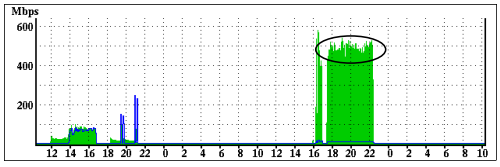

Figure 7: Network traffic of Edinburgh dCache node during a FTS transfers test of 1TB of data to the Tier-1 at Rutherford Appleton Laboratory. The final rate corresponds to $\sim$ 440Mb/s.

Figure 8 shows a similar 1TB data transfer in the opposite direction, from the RAL dCache to the Edinburgh dCache, writing to the directly attached disk pools (i.e. not to the NFS mounted pools). The final observed rate was $\sim$175Mb/s which is less than the GridPP defined target rate of 300-500Mb/s for Tier-1 to Tier-2 transfers. These tests have highlighted that differences exist between the read and write rates that are possible with the Edinburgh dCache. Understanding the dCache configuration will be the first step in characterising and optimising the current setup. Similar studies will be required to be performed at

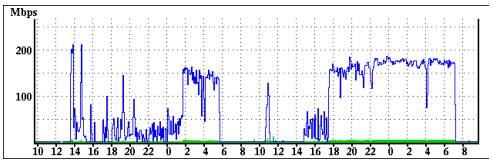

Figure 8: Network traffic of Edinburgh dCache node during a FTS transfers test of 1TB of data from the Tier-1 at Rutherford Appleton Laboratory. Transfer corresponds to the window 1800-0700 on the right hand side of the Figure. The final rate corresponds to $\sim$ 175Mb/s.

the other GridPP Tier-2 sites.

In addition to testing the Tier-1/Tier-2 data transfer channels, Tier-2/Tier2 transfers within ScotGrid have also been studied. It was particularly important to continue with this work while the Tier-1 was occupied with the SC3 Tier-0/Tier-1 disk to disk re-run. Currently, Tier-2/Tier-2 transfers do not fall within the LCG experiments computing models [11], but it is useful from the point of view of the Tier-2s since optimisation of their interconnections could potentially allow for sites within the same federated Tier-2 using each other as close storage elements within LCG. For example, this would allow a job submitted to the computing cluster at Durham using the dCache at Edinburgh as its local storage space. This will only be possible if the networking infrastructure as well as the SRM setups are understood.

FUTURE WORK

The aim of SRM inter-operation testing during the SC4 was to allow sites to understand the interactions between the LCG middleware components and their hardware setups. ScotGrid plans to use the SC4 period to act as a testbed for studying these interactions, finding bottlenecks and optimising available performance. The results shown in Figures 7 and 8 are the first steps towards this. Other questions that have to be asked are:

- What is the most suitable pool filesystem to use with dCache and DPM?
- Which RAID configuration is best used for efficiency in reading/writing while also providing a suitable level of redundancy?
- What are the optimal kernel tuning parameters (sysctl) that can be used for dCache and DPM to maximise the file transfer rate?
- What alternative technologies are available to improve read and write to Tier-2 disk storage?
- How can we optimise the networks?

It is planned that ScotGrid will be able to characterise our own system configuration and develop general optimisation principles that can apply to all Tier-2 sites. The GridPP wiki [12] will play a key role in disseminating this information to the wider community, but will eventually allow GridPP to provide an optimised service to all Grid users.

CONCLUSIONS

The expertise of ScotGrid in the fields of data and storage management has allowed us to take a leading role within GridPP in the deployment and testing of LCG middleware products which allow grid-transparent access to storage resources, specifically storage resource managers. The use of a testing framework and monitoring has contributed to the successful deployment of an SRM interface to the storage at every GridPP Tier-2 site, while simultaneously feeding information back to the LCG middleware developers regarding future improvements. ScotGrid will continue in this role during the lifetime of the GridPP project, one of its aim now being the optimisation of the storage and data management middleware framework that it has helped deploy.

ACKNOWLEDGEMENTS

The author would like to thank the members of ScotGrid who have provided support and advice during the operation of the Tier-2. We would also like to thank the GridPP collaboration and the Edinburgh Parallel Computing Centre (EPCC) for assistance and support. ScotGrid hardware is funded by the Scottish Higher Education Funding Council.

REFERENCES

[1] GridPP
http://www.gridpp.ac.uk

[2] ScotGrid: Scottish Grid Service
http://www.scotgrid.ac.uk

[3] SRM collaboration
http://sdm.lbl.gov/srm-wg/index.html

[4] dCache collaboration
http://www.dcache.org

[5] Disk Pool Manager
http://uimon.cern.ch/twiki/LCG/DpmAdminGuide

[6] The Globus Alliance: GridFTP
http://www.globus.org/grid_software/data/gridftp.php

[7] GLUE Information Model
http://infnforge.cnaf.infn.it/glueinfomodel/

[8] GridPP storage monitoring
http://www.gridpp.ac.uk/storage/status/gridppDiscStatus.html

[9] Grid Information Provider
http://lfield.home.cern.ch/lfield/gip/documentation.html

[10] gLite middleware (EGEE)
http://glite.web.cern.ch/glite/

[11] LCG experiment computing models
http://uimon.cern.ch/twiki/bin/view/LCG/ComputingModels

[12] GridPP wiki
http://wiki.gridpp.ac.uk/wiki/Grid_Storage