dCache with tape storage for High Energy Physics applications

A Agarwal[1], R Enge[2], K Fransham[1], E Kolb[2], C Leavett-Brown[2], D Leske[2], K Lewall[2], H Reitsma[2], E Rempel[2], R Sobie[1,3]

[1] Department of Physics and Astronomy, University of Victoria, Victoria BC, Canada
[2] University Systems, University of Victoria, Victoria BC, Canada
[3] Institute of Particle Physics, Canada

aka@uvic.ca

Abstract

AN interface between dCache and the local Tivoli Storage Manager (TSM) tape storage facility has been developed at the University of Victoria (UVic) for High Energy Physics (HEP) applications. The interface is responsible for transferring the data from disk pools to tape and retrieving data from tape to disk pools. It also checks the consistency between the PNFS filename space and the TSM database. The dCache system, consisting of a single admin node with two pool nodes, is configured to have two read pools and one write pool. The pools are attached to the TSM storage that has a capacity of about 100TB. This system is being used in production at UVic as part of a Tier A site for BaBar Tau analysis. An independent dCache system is also in production for the storage element (SE) of the ATLAS experiment as a part of Canadian Tier-2 sites. This system does not currently employ a tape storage facility, however, it can be added in the future.

1. Introduction

- The dCache software is developed as a joint venture between Deutsches Elektronen-Synchrotron, DESY [1], Fermi National Accelerator Laboratory, FNAL [2] and the Nordic DataGrid Facility, NDGF [3].
- dCache supports the disk caching system as a front end to the tertiary storage tape systems with one unique namespace for all the data files.
- Such a system with tape storage has been implemented by several groups ([4],[5],[6],[7] and [8]).
- A similar interface between dCache and the local tape storage facility (TSM) [9] has been developed at the University of Victoria (UVic) for High Energy Physics (HEP) applications, and its implementation for BaBar and ATLAS experiments is highlighted.

2. dCache-TSM Interface

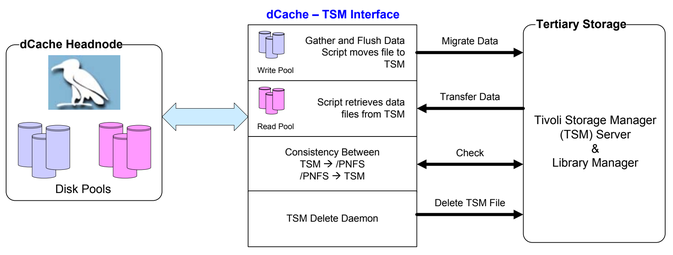

Figure 1: *Characteristic features of the interface between dCache and the TSM*

- *Gather and flush data* - Whenever a new file is written to the write pool, dCache runs the interface script which copies the file to tape storage. If the write pool is full, the oldest file is removed from the write pool only, while preserving the same file on tape storage.
- *Retrive data from tape to disk* - Whenever the user makes a request for the file that is not on the read and write pools, dCache runs the interface script which brings the file from tape to the read pool on disk.
- *Check consistency between the PNFS filename space and the TSM database* - A Cron job is set up to check the consistency of the files between dCache and TSM by checking their size, name and number in both directions: (i) TSM → dCache /pnfs filesystem and (ii) dCache /pnfs filesystem → TSM.
- *Monitor the TSM cleaner daemon* - Whenever any file is deleted in the dCache /pnfs filesystem, it is moved to the dCache trash directory (/opt/pnfsdb/pnfs/trash/1). If the daemon finds any file in this trash directory, it is cleaned from this directory and the corresponding file is also removed from the TSM (if any).

3. dCache Configuration for Tier A site for BaBar Tau analysis

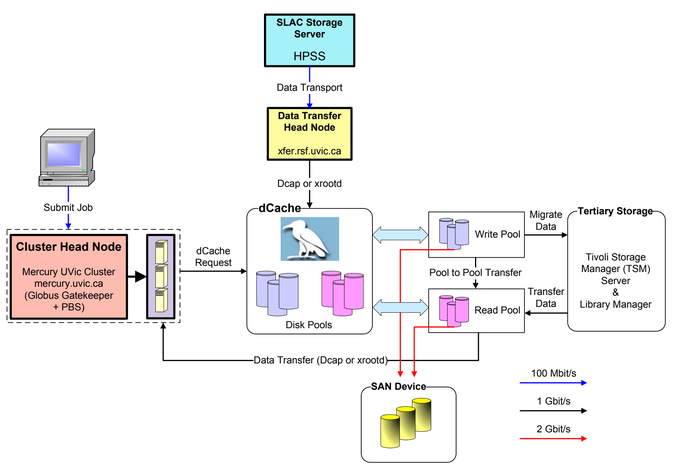

Figure 2: *Networking and key components of the dCache-TSM system used for the BaBar Tier A*

- This dCache tape system is in production and is configured to have two read pools of 5TB each and one write pool of 3TB.
- The disk pools are served by the Storage Area Network (SAN) box with 2GB fiber optic links.
- The pools are attached to the TSM storage system that has a capacity of more than 100TB.
- Data from the High Performance Storage System (HPSS) at SLAC is directly downloaded using the dCap library into the dCache system using a script called BbkImport, which uses bbcp or bbftp protocols and is provided by SLAC.
- The external network to SLAC from the dCache Head node is linked at 100 MB/s, which collects data at about 1-2 TB per day to the dCache tape storage system.

4. dCache System Usage for BaBar Analysis

- BaBar users submit jobs to the head node of the Mercury cluster (mixture of Intel and AMD nodes consisting of about 450 CPUs).
- Data access is provided through dCap, Xrootd, or GridFTP protocols.

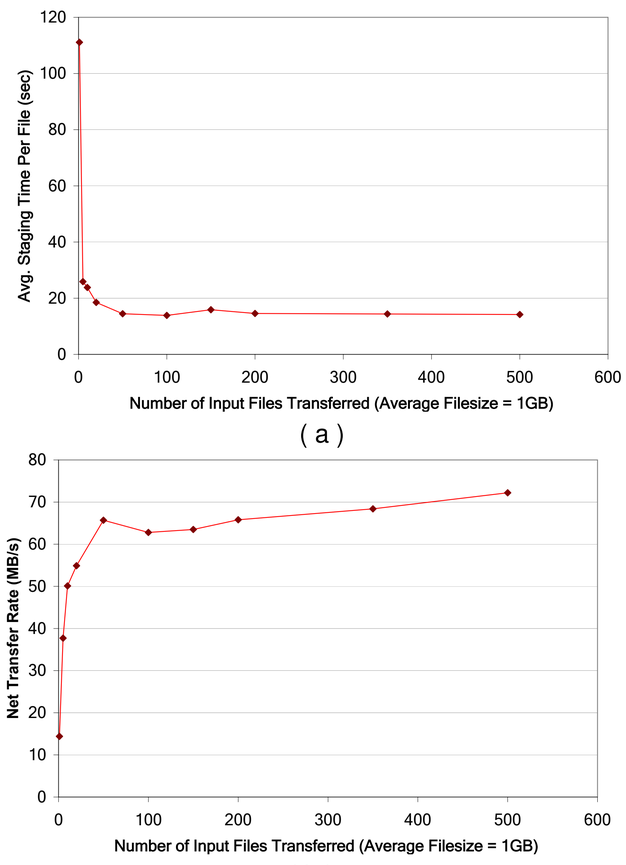

Figure 3: *Average staging time per file (top) and the net transfer rate (bottom) as a function of the number of input files transferred.*

- Figs. 3(a) and 3(b) show that when the number of input files transferred are more than 50, the average staging time is drastically reduced to about 15-20 sec per file and the net transfer rates are about 70 MB/s.
- The reduction in average staging time is attributed to less overhead incurred due to tape mounting and dismounting.

Table 1: *Performance tests.*

	BaBar Jobs Runtime $<$ 2hrs			BaBar Jobs Runtime $>$ 24hrs		
	Files on Disk	Files on Tape		Files on Disk	Files on Tape	
Number of Input Files	CPU Time (MM:SS)	CPU Time (MM:SS)	% Overhead	CPU Time (MM:SS)	CPU Time (MM:SS)	% Overhead
1	9:03	10:43	18	1450:20	1451:45	.09
5	8:59	16:51	87	1460:52	1465:05	.28
10	10:20	23:24	127	1470:36	1475:50	.35
50	11:58	38:36	222	1475:08	1495:22	1.37
100	14:46	66:29	350	1482:10	1517:18	2.37

- Table 1 shows a performance comparison of BaBar analysis jobs while running with the input files on disk or on tape.
- It is found that when the total run time of BaBar jobs is larger than the time taken to bring the input files from tape to disk ($<$ 1 or 2hrs), the latency is not significant from the user's point of view.

5. dCache System for ATLAS Tier-2 at UVic

- Another important dCache instance in production at UVic is the storage element (SE) for the ATLAS experiment as a part of Canadian Tier-2 sites which belongs to the Large Hadron Collider Computing Grid (LCG) project at CERN, Geneva [10].
- This SE uses dCache version 1.8 with SRM version 2.2. It consists of a single admin node with two pool nodes with a total of 14 TB disk space.
- Since the ATLAS Tier-2 sites require the data files to remain on disk pools, the dCache system presently used is kept tapeless to provide very fast access (ONLINE).
- To improve file retention on this system with ONLINE access, the "gather and flush" technique in the dCache-TSM interface will be changed to "gather, flush and pin." The pinning option will keep the data files on the disk, providing both very fast, ONLINE, access and safe retention.

6. dCache Tape System for General User Storage

- Users can store useful data using either the dCap or gsidCap protocol on this tape system directly, rather than writing to NFS mounted disk file servers. The tape system provides better retention of the users data, by protecting against disk failures.

7. dCache Monitoring

- Various aspects of the system, such as the disk pools, cell services, pool to pool transfers, and transfers from the TSM to the pools are monitored through a web front-end utility provided by dCache.
- This is shown at http://swallow.rsf.uvic.ca:2288 for the BaBar dCache system and at http://lcg-se.rcf.uvic.ca:2288 for the ATLAS dCache system.

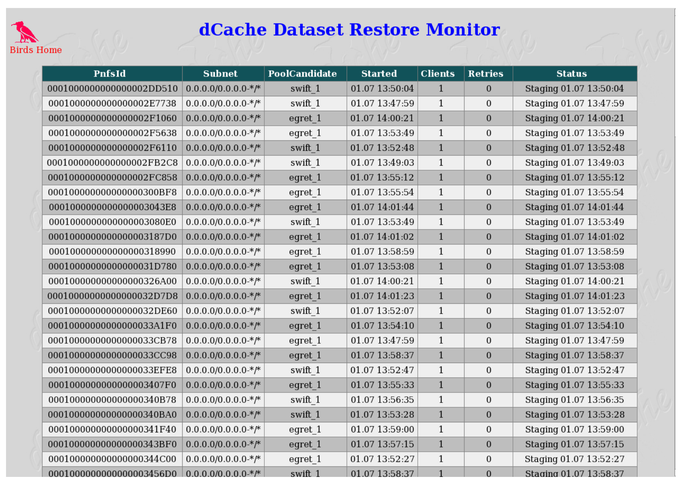

Figure 4: *Live monitoring of the dCache web interface to restore BaBar data input files from tape to the disk pool*

- Another monitoring system for dCache, developed by Vladimir Podstavkov (podstvkv@fnal.gov) at Fermilab [2] and used for their CDF and CMS experiments, is very useful.
- This software extracts information about the data transfer for the pools and the TSM directly from the dCache database and can generate daily, weekly and monthly plots (see Fig. 5).

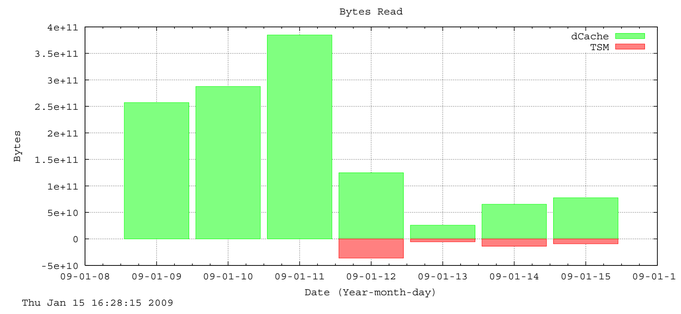

Figure 5: *A sample daily plot of the total bytes transferred, of BaBar data, through the dCache disk pools (green) and tape (red) over a one week period.*

8. Conclusions

- The BaBar dCache system allows transparent access to a large amount of data files distributed on disk pools and on TSM tapes.
- Whenever the run time of BaBar jobs is larger than 1 or 2 hrs, the increase in access latency for bringing the input files from tape to the read-pool is found to be insignificant.
- Using TSM storage provides reliability by making the system less vulnerable to disk failures.
- Data can be accessed by users from the dCache system through various protocols such as GridFTP, Xrootd, dCap and SRM.
- The disk pool nodes can be easily added or removed at any time without affecting system operation due to the separation of the PNFS filename space and the data pools.
- Attaching TSM storage to the Atlas dCache SE will provide both fast access and reliable retention of the data.
- The dCache tape system provides users the ability to safely store their data.

References

[1] http://www.dcache.org
[2] http://www.fnal.gov
[3] http://www.ndgf.org
[4] dCache, the OverView by Patrick Fuhrmann for the dCache team, Deutsches Elektronen Synchrotron Notkestrasse 85, 22607 Hamburg
[5] http://www.fnal.gov/docs/docs/products/enstore/
[6] BNL dCache Status and Plan by Carlos Gamboa et al., dCache Workshop : January 18-19, 2007
[7] http://www-306.ibm.com/software/Tivoli/products/storage-mgr)
[8] Data Migration Facility (DMF) at SARA/NIKHEF, Amsterdam, Netherlands NL - Site Report from SARA/NIKHEF T1 by Ron Trompert, Storage workshop, 2-3 July 2007.
[9] http://www.ibm.com/software/tivoli/products/storage-mgr/
[10] http://lcg.web.cern.ch/LCG
[11] http://lcg.web.cern.ch/LCG/peb/GTA/GTA-ES/Grid-File-AccessDesign-v1.0.doc