The Brookhaven National Lab dCache Status and Plan

G. Carcassi, M. Ernst, C. Gamboa, C. Hollowell, J. Hover, H. Ito, D. Katramatos, Z. Liu, J. Packard, R. Petkus, R. Popescu, O. Rind, J. Smith, Y. Wu, D. Yu, X. Zhao
Physics Department, Brookhaven National Laboratory, Upton, NY 11973, USA

ABSTRACT

The Brookhaven RHIC/ATLAS Computing Facility serves as both the tier-0 computing center for RHIC and the tier-1 computing center for ATLAS in the United States. Two large-scale deployment instances of dCache, the distributed disk caching system developed by DESY/FNAL, are deployed in BNL to meet the increasing challenge of providing local and grid-based access to large datasets in a reliable, cost-efficient and high-performance manner. This paper describes the structure, usage and monitoring methods of dCache instances. Challenges and issues facing dCache RHIC/ATLAS systems and the deployment of SRMv2.2 are also discussed.

1. INTRODUCTION

dCache provides a system for users to store and retrieve huge amounts of data, distributed among a large number of server nodes or stored in a Hierarchical Storage Manager (HSM), under a single virtual file system tree with a variety of standard access methods. In addition, it significantly improves the efficiency of connected tape storage systems through caching, i.e. gather & flush, and scheduled staging techniques [1].

Currently at BNL, there are two large dCache instances, one for the ATLAS experiment and one for the PHENIX experiment. BNL's dCache implementation utilizes local disks on worker nodes to provide peta-scale on-line storage. Both USATLAS dCache and PHENIX dCache have been deployed based on the "hybrid model", meaning the worker nodes function simultaneously as file servers and compute elements, providing a cost-effective, high throughput data storage system.

The BNL dCache systems also serve as caching front-ends to the HPSS Mass Storage System (HPSS), which providing archival and redundancy. The dCache backend interface to HPSS was developed locally using primary data transfer protocol, Parallel FTP (PFTP) with a Hierarchical Storage Interface (HSI)[2]. For data retrieval, USATLAS dCache is integrated with a backend tape prestaging batch system, the Oak Ridge Batch System. PHENIX dCache uses a PHENIX software layer, known as "data carouse"[3], to manager the restore requests.

BNL USATLAS Computing Facility needs to provide a Grid-based storage system with these requirements: a total of one gigabyte per second of incoming and outgoing data rate between BNL and ATLAS T0, T1 and T2 sites, thousands of reconstruction/analysis jobs accessing locally stored data objects, three petabytes of disk/tape storage in 2007 scaling up to 25 petabytes by 2011, and a cost-effective storage solution. The paper described the solutions under testing or already deployed to fulfill these requirements.

dCache1.8 is discussed in the paper. The estimated deployment time of dCache 1.8 in ATLAS dCache is end of 2007.

2. USATLAS DCACHE SYSTEM

2.1 System Deployment and Usage

USATLAS Tier1 dCache is a large scale, grid-enabled and distributed disk storage system. It is deployed for production usage since Oct 2004. It also participated in a series of LHC Service Challenges activities since then. As of end of May 2007, total production space is 904 TB and Service Challenge space is about 321 TB. USATLAS Tier1 dCache provided 762 TB disk pool space and back-end HPSS tape system to store these data. There are about 582 nodes in total, including 15 Core servers, 555 Read servers and 12 Write Servers. Figure 1 shows the architecture of the USATLAS dCache system.

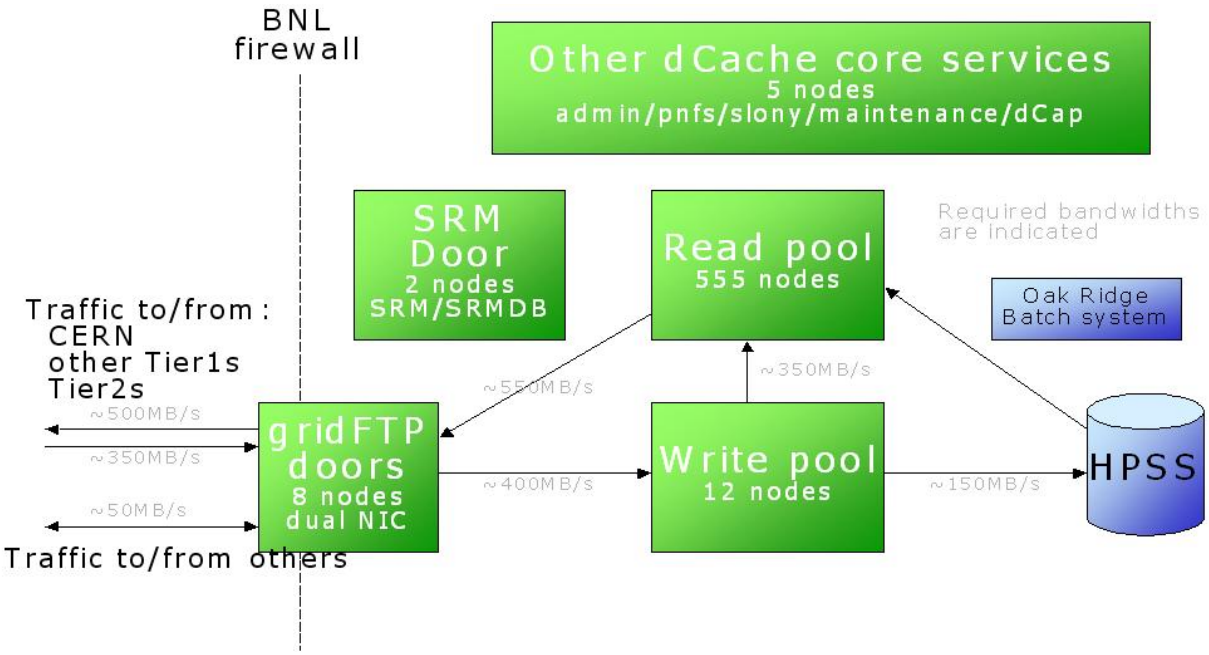

Figure 1: USATLAS dCache Architecture

Read pool disk space is low-cost, locally mounted disk space on the computing farm. Write pools are dedicated servers. In Aug 2007, hard ware of core servers, such as PNFS, Slony, admin, maintenance, SRM and SRM DB nodes were upgraded to four core CPU 8GB memory servers. Database servers, like PNFS, Slony, SRM Database, Billing database are using SAS disk and critical servers without Database are using SATA database. For better performance, except PNFS server is running on 32-bit application. SRM, SRM DB, Admin and Maintains servers are running on 64-bit OS and with 64-bit application. Dual home GridFTP door nodes bypassing the BNL firewall work as adapters for grid traffic. This design allows any internal pool node to send and receive data to and from remote users without exposing these nodes to the Internet. Although not yet really work this way for all transfer during the ATLAS data export exercise, BNL achieve 300MB/s between CERN and Tier1.

The system has exhibited quality performance through a series of Service Challenges and US ATLAS production runs. Two critical data transfer tasks were started exercising to validate the readiness of BNL USATLAS data storage in terms of stability and performance, namely: 1) for WAN data transfer, running basic transfer (e.g. SRMCP w/ and w/o FTS) and data replications based on ATLAS DDM between BNL and USATLAS Tier 2 sites, and 2) exercising the LAN based dCache Posix I/O functionalities (e.g. dcap, TDCacheFile, and Tfile) and measuring the performance of concurrent read access to the same data set by a large number of analysis jobs. The following two figures, figure 2 and figure 3 show the statistics of data transfer in USATLAS dCache from January 2007 to Jun 2007. Please note, reading represents transfer from dCache disks to clients, writing represents transfer from clients to dCache disks, stage represents transfer from HPSS to dCache disks, and migration represents transfer from dCache disks to HPSS [1]. Figure 4 is the ATLAS data volume at BNL RACF as Aug 2007. Almost all of these data are in dCache.

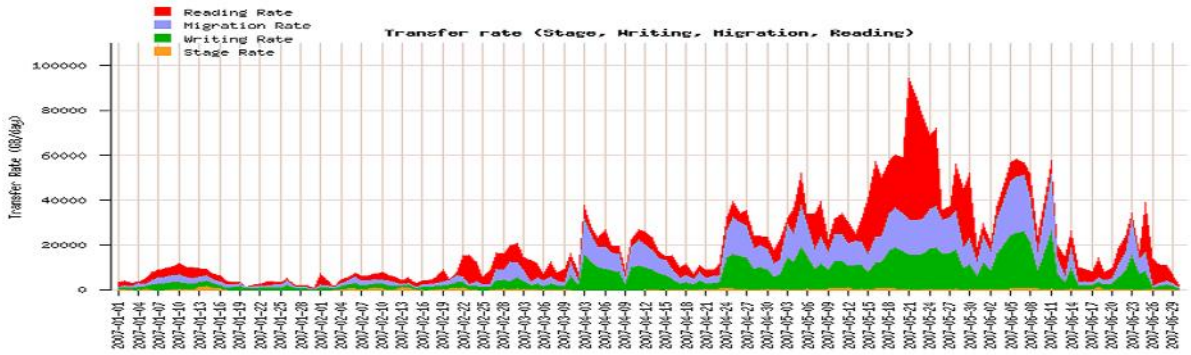

Figure 2: Transfer rate of reading, migration, writing and Stage

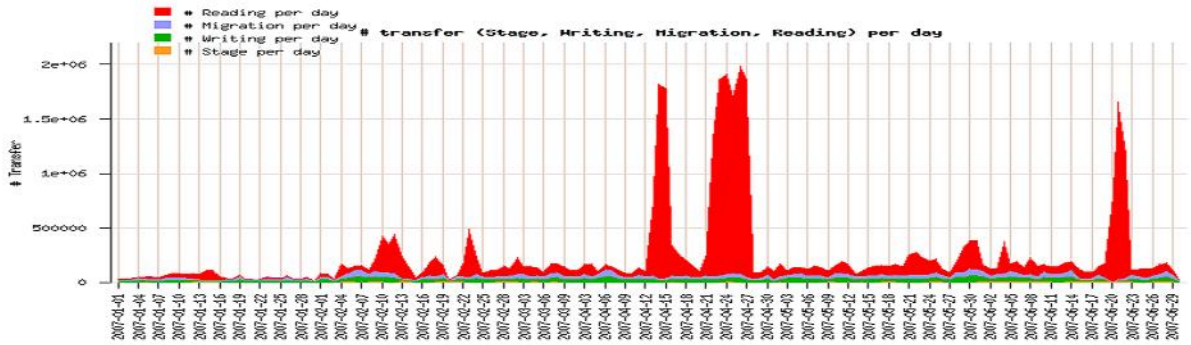

Figure 3: The number of transfers per day for reading, migration, writing and stage

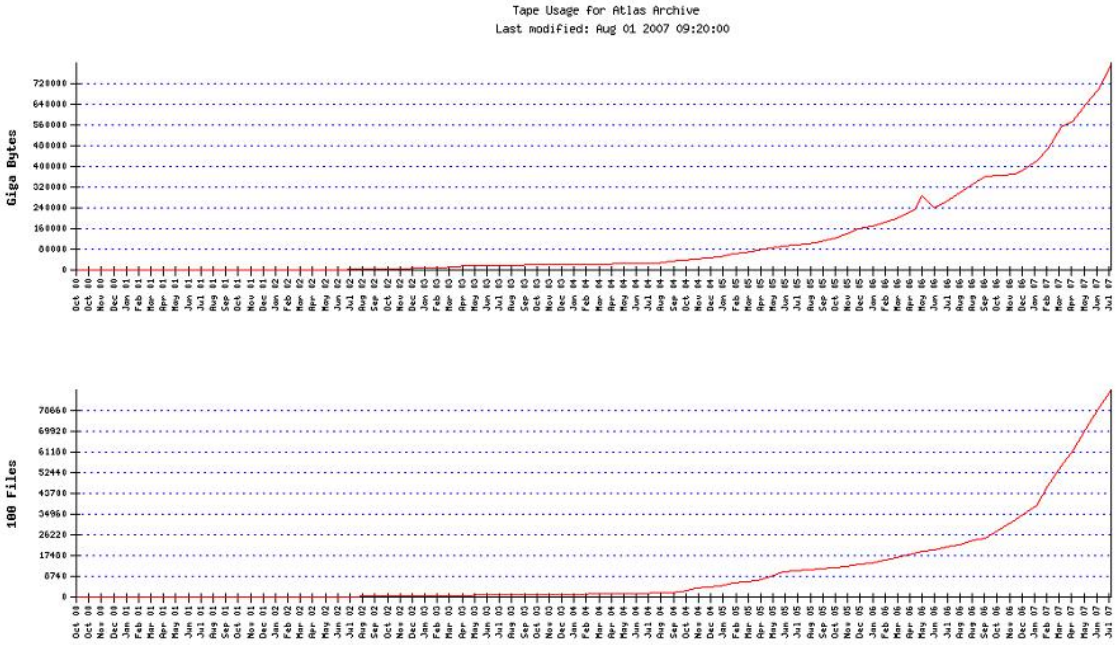

Figure 4: ATLAS data volume at BNL RACF

2.2 dCache Monitoring

To guarantee reliability and high performance for production and service challenges, various monitoring methods are deployed. These methods include Ganglia, Nagios and monitoring scripts created by system administrators.

Ganglia is a scalable distributed monitoring system for high-performance computing systems such as clusters and Grids [4]. BNL dCache systems use Ganglia to monitor CPU, load, memory, usage for every dCache nodes and network performance for gridftp doors and pool nodes. There are more than 20 PNFS databases in USATLAS dCache. Abnormal activities of one database will affect the whole system, for example, pushing PNFS server to very high load. To better understand activities of each database, especially identify the problem database when system becomes unstable, statistics information, such as database blk_hit, xact_comment and xact_roolback are displayed in Ganglia. Snapshots of Ganglia plots were captured per hour for core servers and doors for investigation purposes.

Nagios is a free monitoring system. It became a de facto industry standard for monitoring of computing resources. Nagios monitors resources by executing monitoring code - so-called plugins - and reporting the result (OK, WARNING, CRITICAL, UNKNOWN) via web interface [5]. Nagios is used by dCache system to monitor the availability and disk usage of core and pool nodes. An internal/external dccp/globus-url-copy/srmcp request is sent by Nagios system to ensure dCache process are listening on the correct ports and system is under normal operation status. Also Nagios is used to check host certificate identification and expiration. Since Nagios is integrated with RT ticket system in BNL. The Nagios probes not only send email notification but also open RT ticket about the status of the services being monitored.

Other than Ganglia and Nagios monitoring tools, many monitoring scripts were used in the system to release heavy maintenance workload. For example, Oak Ridge Batch System monitoring tool ensures smooth retrieval process of files from tape, which is extremely useful when large pre-stage requests were submitted. dCache log files contain rich information of system activities. Some information is sign of unhealthy system. Log analysis scripts capture this information before the system goes to critical status or crashes.

2.3 Challenges and Issues

The ATLAS experiment will generate data volumes each year on the Petabyte scale starting in the near future, as grid-enabled tier 1 Storage Element, USATLAS dCache faces new challenges and provide solutions for issues. This section only discusses two issues USATLAS dCache faces: read pool data pinning and limited disk I/O in write pool.

Production team requires important data to stay on disk, which is a must for the smooth production data flow. But this is not always the case since in-active file will be purged out from disk by active file. Pining in dCache is a tedious administrative task and not reliable. For example, re-start of one pool will not keep pining status of file in the pool. To fulfill request from production and avoid manual involvement of pining, HoppingManager is deployed in the dCache system. Figure 5 shows the solution of deploying HoppingManager and Transfer pool to "pin" important production data in read pool disk. The solution is tested through in test bed.

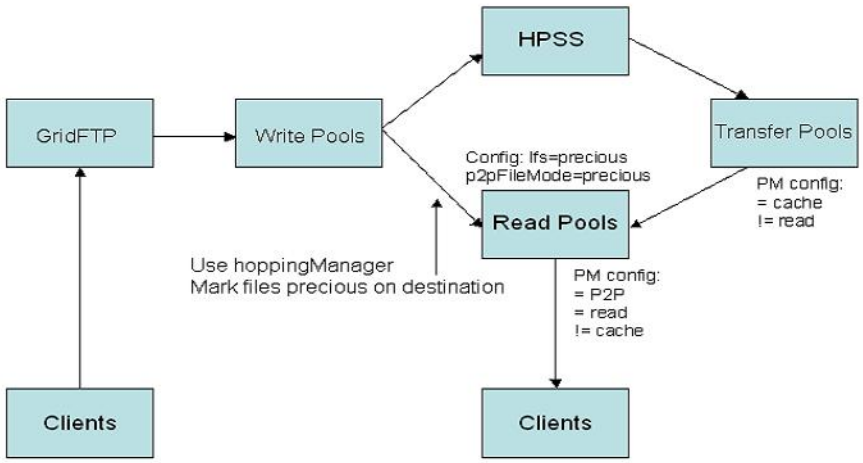

Figure 5: The workflow of "pin" data in disk using HoppingManager

Based on Bonnie disk I/O test, concurrent write threads are set at the optimized number. But since concurrent inbound and outbound traffic downgrade disk I/O performance, each dCache write pool could only sustain 15MB/s. Write pools became bottom neck of the system when capacity of other components, such as PNFS and Grdiftp door, were increased. SUN thumper server [6] was introduced as write pool to improve the disk I/O issue. Test result is very promising. The throughput is 350 MB/s for 150 clients sequentially reading 1 random 1.4G files. When using 75 clients to sequentially write 3x1.4G files and 75 clients to sequentially read 4x1.4G randomly selected files, the throughput is 200MB/s write and 100 MB/s read. In Oct 2007, five thumpers were added to production USATLAS dCache system as write pool. Based on the evaluation result, it is expected that the write I/O rate limit on each node to go from 15 MB/s to at least 100 MB/s with concurrent inbound and outbound traffic.

2.4 SRM 2.2 dCache and Chimera

The dCache team has finished the implementation of all SRM v2.2 elements required by the LHC experiments. The new functionality includes space reservation, more advanced data transfer, and new namespace and permissions management functions. SRM's "Bring Online" function required redevelopment of the Pin Manager service, responsible for staging files from the back-end tape storage system and keeping these files on disk for the duration of the Online state [7]. This feature supports ATLAS reconstruction jobs requiring pre-staged RAW data from tape to disk. dCache1.8 is deployed in BNL to test the reliable/high scalable storage element defined on SRMv2.2. The test bed is participating ATLAS specific use case test to produce the precise guidelines per implementation focusing on deploying the specific storage classes for a specific VO. An improved file system engine, Chimera is also introduced in dCache1.8. In addition to an improved performance profile, Chimera provides a wide set or enhanced functionalities. Evaluation of Chimera is underway in BNL RACF computing facility. Oracle cluster will be used in the Chimera evaluation to provide a robust, reliable database system.

2.5 Future Directions

USATLAS Tier-1 dCache-based, grid-enabled Storage Element is scale to the petabyte range. Estimated disk capacity in year 2010 and 2011 will be 17,262 TB and 24,427 TB. The future system will server as a greater and critical role in transferring data volumes generated by ATLAS experiments.

3. PHENIX DCACHE SYSTEM

The PHENIX experiment is one of two ongoing projects at BNL's Relativistic Heavy Ion Collider (RHIC). With data-taking rates well over 300 MB/s and nearly a Petabyte of data recorded during the last

run alone, PHENIX is constantly contending with storage and throughput resource limitations. As with ATLAS, dCache was initially deployed for PHENIX on the RCF to make use of the large amount of local disk storage available on the Linux farm computing nodes. Within the PHENIX computing model, dCache has since evolved into the primary storage resource for both data production and user analysis. It currently comprises over 280 TB of disk storage on nearly 400 dCache pools deployed across the Linux farm, but also including dedicated storage on SunFire X4500's and Aberdeen SCSI arrays. Almost all pools are interfaced to the central HPSS Mass Storage System on the backend, providing the main end repository and archiving mechanism for the PHENIX production stream. The system is also tightly integrated into the PHENIX "Analysis Train" process, which aggregates user analysis jobs to run efficiently on common data subsets through a centrally controlled mechanism. Data transfer proceeds principally on the LAN through the dCap protocol with aggregate rates reaching levels in excess of 1.5 GB/s. An increasing amount of offsite transfer between BNL and other collaborating PHENIX institutions (such as IN2P3) is effected via SRM, with a small dedicated set of externally accessible write pools available for incoming data.

REFERENCES

[1] Z.Liu, et al. “Large Scale, Grid-Enabled, Distributed Disk Storage Systems at the Brookhaven National Lab RHIC/ATLAS Computing Facility”, CHEP06, Mumbai, India, Feb. 2006

[2] HIS website: http://www.sdsc.edu/Storage/his/

[3] BNL PHENIX Data Carousel Website: http://www.phenix.bnl.gov/software/tutorials/datacarousel.html

[4] Ganglia website: http://ganglia.sourceforge.net/

[5] T. Wlodek, et al. “Integrated RT-Nagios System at BNL US Atlas Tier1 Computing Center”, CHEP07, Canada, Sep. 2007

[6] Sun Fire X4500 Server, http://www.sun.com/servers/x64/x4500/

[7] Timur Perelmutov, SRM 2.2 interface to dCache, http://osg-docdb.opensciencegrid.org/cgi-bin/ShowDocument?docid=645