LARGE SCALE, GRID-ENABLED, DISTRIBUTED DISK STORAGE SYSTEMS AT THE BROOKHAVEN NATIONAL LAB RHIC/ATLAS COMPUTING FACILITY

B. Gibbard, Z. Liu, R. Popescu, O. Rind, J. Smith, Y. Wu, D. Yu, X. Zhao
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. The increasing challenge of providing local and grid-based access to very large datasets in a reliable, cost-efficient and high-performance manner, is being addressed by a large-scale deployment of dCache, the distributed disk caching system developed by DESY/FNAL.

Currently in production for the PHENIX and ATLAS experiments, dCache is employing the same worker nodes utilized by the RHIC and ATLAS analysis clusters, making use of the large amount of low-cost, locally-mounted disk space available on the computing farm. Within the hybrid storage/computing model, the worker nodes function simultaneously as file servers and compute elements, providing for a cost-effective, high throughput data storage system. dCache also serves as a caching front-end to the HPSS Mass Storage System, where access to the data on tape is provided through an integrated optimizing layer that was developed at BNL.

BNL's dCache functions as SRM-based Storage Element in the context of OSG and LCG. It has been serving on a production scale at BNL since November 2004, exhibiting quality performance through a number of Service Challenges and production runs.

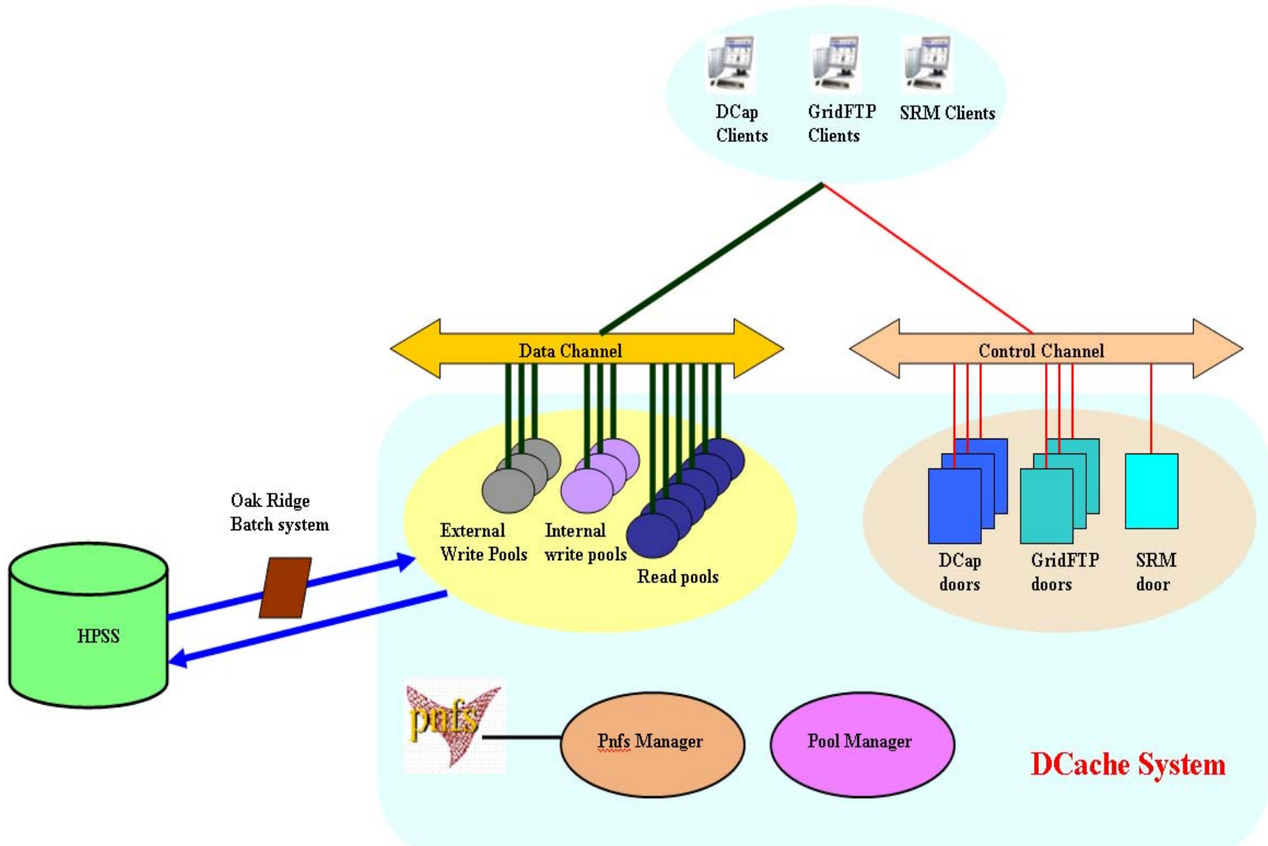


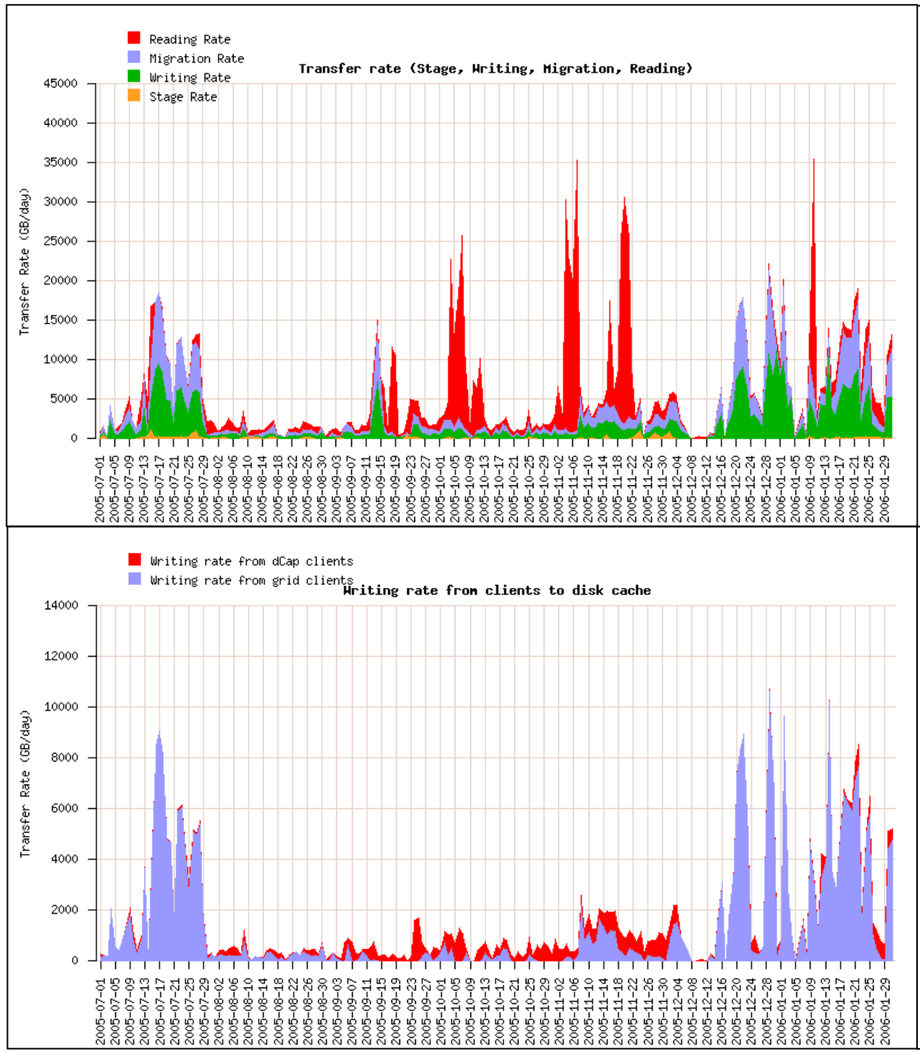
Fig. USATLAS dCache system architecture

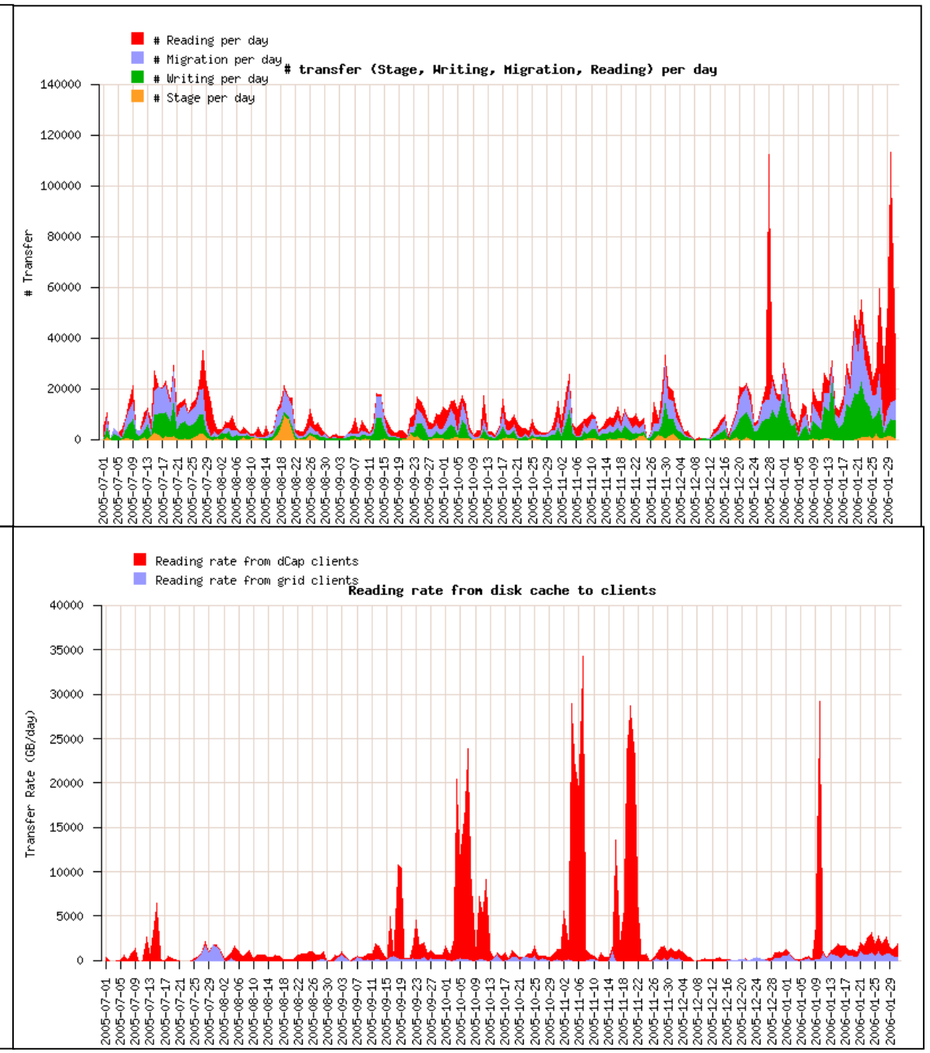
USATLAS dCache System

USATLAS dCache System:

- **Large scale, grid-enabled, distributed disk storage system**
  - 336 servers, 147 TB disk space (as of Jan 31st 2006)
  - 152TB ATLAS Production data have been written into dCache name space (as of Jan 31st 2006)
  - Grid-enabled (SRM, GSIFTP) Storage Element in the context of OSG and LCG
- **Cost-effective and high throughput**
  - Utilizing low-cost, locally-mounted disk space on the computing farm
  - Exhibiting high performance during a series of Service Challenges and US ATLAS production runs.
- **Load balanced and fault tolerant**
- **Highly scalable**
- **Caching front-end to the HPSS Mass Storage System**
- **Efficient and optimized tape data access**
- **User Guide Web Site:**
  - http://www.atlasgrid.bnl.gov/dcache/manuals/

Statistics of Data Transfer (Date: Jul. 2005 – Jan. 2006)





USATLAS dCache performance during LHC Service Challenge (SC)

DCache during SC:

- **Worked as BNL Grid-enabled Storage Element** during a series of LHC Service Challenge activities (SC3 throughput and Service phase, SC3 re-run)
- **Goals achieved**
  - SC3 throughput phase: 120MB/sec disk-to –disk from CERN to BNL; 80MB/sec disk-to-tape from CERN to BNL
  - SC3 re-run: 90 MB/sec disk-to –disk from CERN to BNL; 50~60 MB/sec disk-to-tape from CERN to BNL
- **Performance tuning experiences:**
  - ***RHEL 4 on write servers***
    - Better performance compared to RHEL3 (higher disk I/O with lower load)
  - ***XFS for write servers***
    - Higher disk I/O and lower load compared to EXT3
  - ***Increased Kernel Stack size on write pool***
    - To avoid the stack overflow problem upon high disk I/O on XFS system
  - ***Increased free memory***
    - To avoid dCache application Java page allocation error
  - ***TCP network tuning***
  - ***Bonnie disk I/O testing***
    - To get the optimized number for maximum parallel write threads on write server.
  - ***Hyperthreading disabled on door servers and write servers***
    - To utilize the CPU power better.
  - ***On write pool servers, allowing I/O in only one direction at a time shows better performance***

Data transfer rate during SC2, SC3 throughput, SC3 re-run

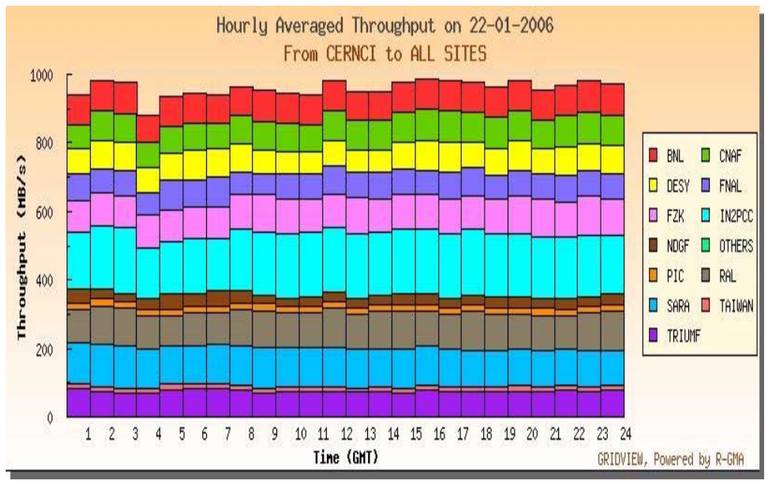


Fig. SC3 re-run: disk-to-disk transfer from CERN to Tier-1 sites
(The "Red" bar represents data transfer to BNL)

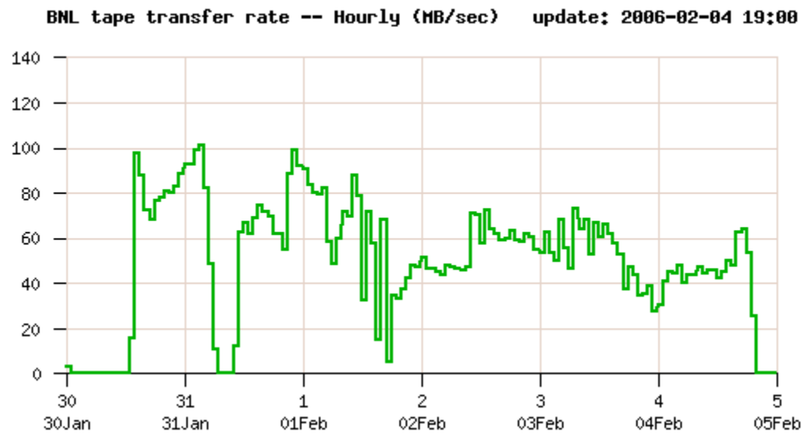


Fig. SC3 re-run: disk-to-tape transfer from CERN to BNL

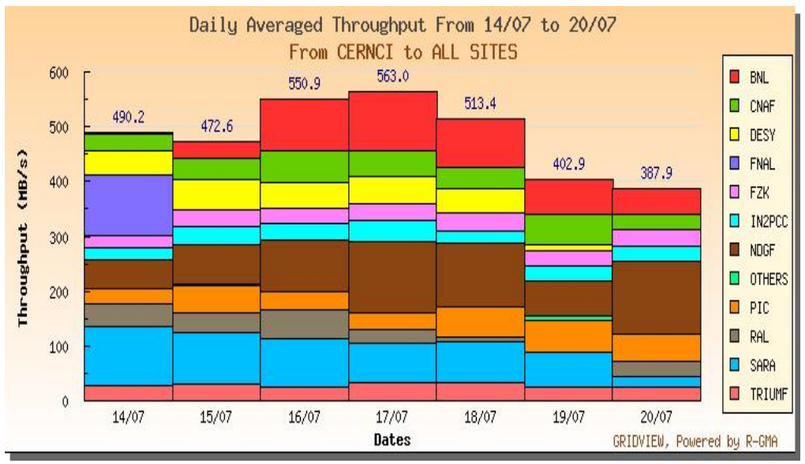


Fig. SC3 throughput phase: disk-to-disk transfer from CERN to Tier-1 sites ("Red" bar represents data transfer to BNL)

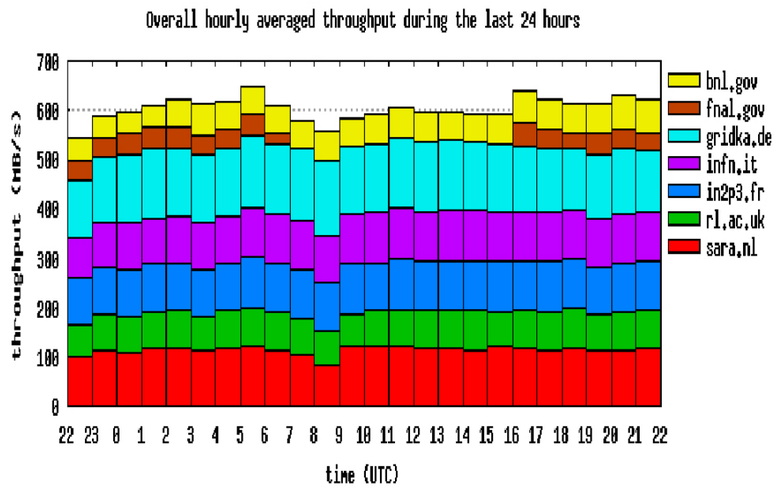


Fig. SC2: disk-to-disk transfer from CERN to Tier-1 sites
("Yellow" bar represents data transfer to BNL)