BNL dCache Status and Plan

dCache Workshop: January 18-19, 2007

Carlos Gamboa, Zhenping (Jane) Liu, Ofer Rind, Jason Smith & Yingzi (Iris) Wu

BROOKHAVEN NATIONAL LABORATORY

Content

- BNL dCache Architecture.
  - Network configuration.
  - SC results and lessons learned.
- Recent Activities
  - Performance tests & hardware upgrades.
  - SUN Thumper (X4500) test results.
  - dCache 1.7 test experiences.
    - Central Flushing System & gPlazma (SRM 2.2).
- Phenix (RHIC) dCache System

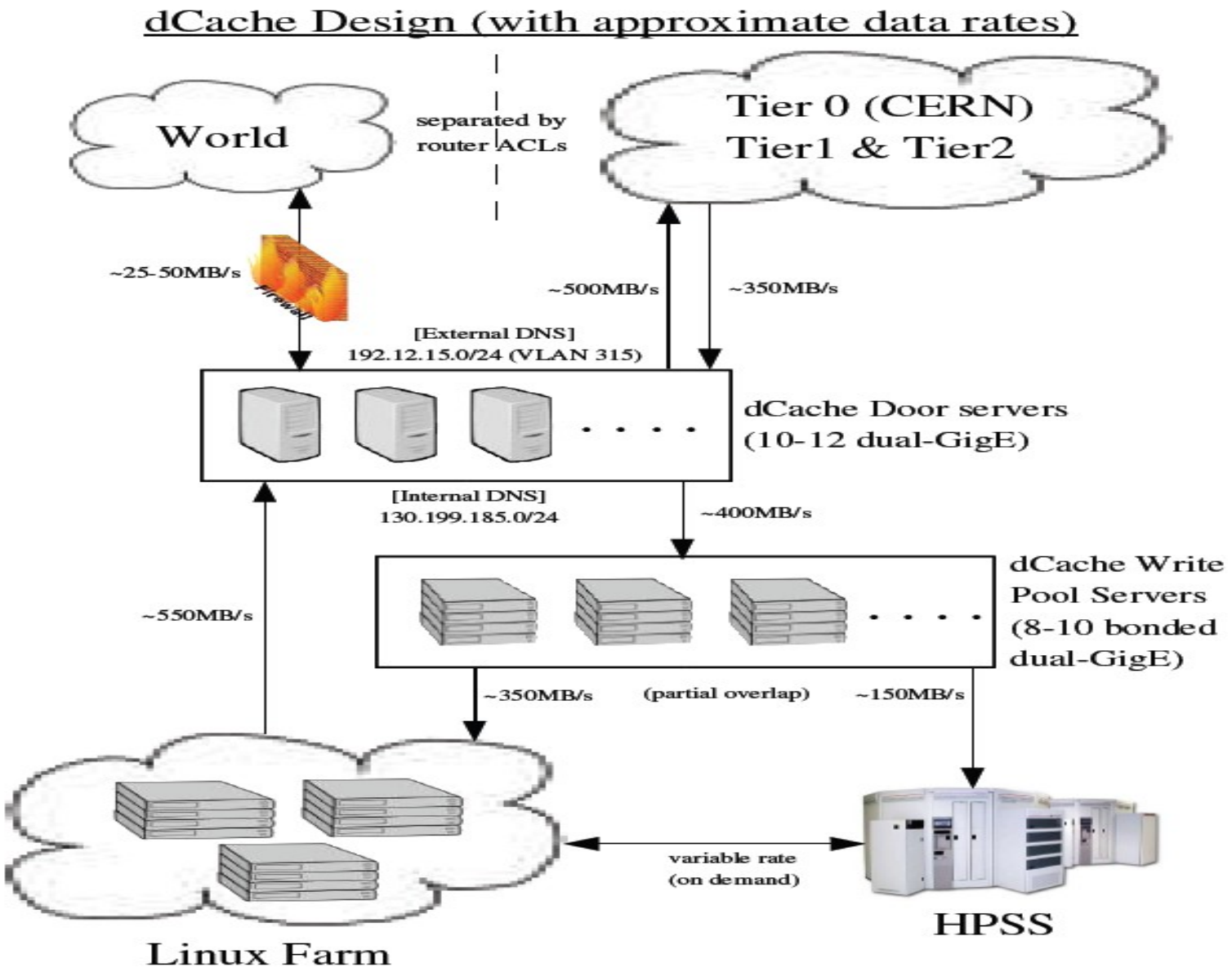
BNL dCache Architecture

- USATLAS Tier1 site, dCache used in production
- 10 dedicated write pools, total size: 2TB
- ~479 read pools, total size: 416TB
- HPSS is used as tape backend.
- 1 admin + 1 PNFS server + 4 dCap + 1 SRM
- 5 GFTP nodes with 2GFTP doors each with 2x1Gb/s
- File system: xfs on write pools, ext3 everywhere else.

Network Configuration

- Most of the dCache servers (including read/write pools, pnfs, admin, SRM, dcap doors) are protected by two levels of firewalls, first BNL firewall then RCF/ACF firewall, especially for farm nodes (dCache read nodes)
  - We may lose high network performance caused by firewall limitation.

- Only a small group of gridftp door servers are outside of BNL firewall for Atlas data transfers (to ensure high performance)
  - Each gridftp door has two interfaces, one uses LHC-OPN subnet IP (which is outside of BNL firewall), the other one uses 185 subnet (which is inside RCF/ACF firewall).


dCache Design (with approximate data rates)
World
separated by router ACLs
Tier 0 (CERN) Tier1 & Tier2
~25-50MB/s
Firewall
~500MB/s
~350MB/s
[External DNS] 192.12.15.0/24 (VLAN 315)
dCache Door servers (10-12 dual-GigE)
[Internal DNS] 130.199.185.0/24
~400MB/s
~550MB/s
dCache Write Pool Servers (8-10 bonded dual-GigE)
~350MB/s
(partial overlap)
~150MB/s
variable rate (on demand)
HPSS
Linux Farm

Network Configuration (Cont.)

- We tried but failed to set up adapter on GridFTP doors to force all traffic go through GridFTP doors as suggested in the page "GridFTP with pools in a Private Subnet" at:
  http://www.dcache.org/manuals/Book/cb-net-pool-priv.shtml
- Current situation:
  - All grid two-party outbound, and third-party inbound/outbound data traffic go directly from pools to the destination.
  - All grid two-party inbound data traffic go over Grid FTP nodes.
    - During service challenges, FTS data traffic go over GridFTP doors since FTS version used didn't do real third-party transfer.

dCache Throughput Performance

- **During Service Challenge Throughput, we observed our system reach 250M byte/second for one day. (intensive write into BNL disk only)**
  - Intensive monitoring and system tuning.
  - Many manual interventions and coordination between CERN, BNL, and other Tier1s.
- **When coupled Service Challenge, data migration to HPSS and farm nodes, along with USATLAS production, the dCache system performance can sustain 120 M Byte/second (intensive reads+writes).**
- **Problems:**
  - Pnfs & SRM performance bottlenecks.
  - Linux SCSI driver/buffer cache cannot efficiently handle parallel read stream and write stream.
    - Exclusive read or write performance is good, but we see a 50% ~ 80% performance degradation when mixing read and write streams.
  - Software Raid, Linux Volume Manager, and file system affect disk I/O performance too, but are relatively minor compared to Linux kernel buffer cache problem.

Solutions to Problem

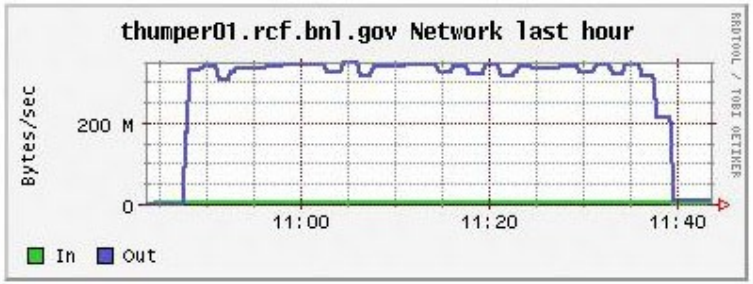
- Filesystem: write pools ext3 -> xfs (More important on RHEL3).
- Tune Postgres Database on PNFS and SRM to improve performance (Postgres shm buffers, DB and core services split, HW RAID disk).
- Linux Kernel Upgrade (RHEL3 -> RHEL4).
- Avoid Mixing Read and Write Operations to disks
  - dCache 1.7.0 has a central flushing system which alternates between data writing into dCache and data migrations into HPSS
- Put SRM database to memory to improve transaction rate since currently SRM DB is transient, no history needs to be kept.
- Multiple SRM servers (DNS RR or IPVS)

SRM Performance Issues

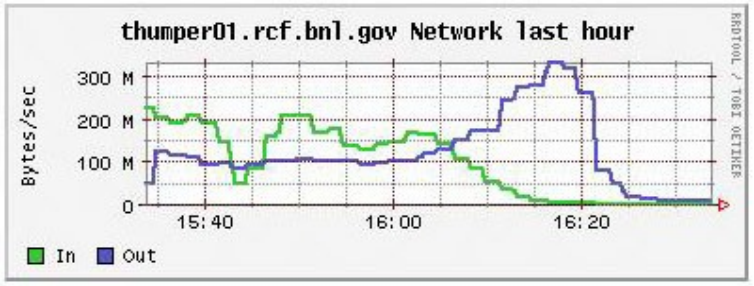
- Cleanup of SRM DB showed significant performance improvement:
  - Before cleanup, 40 simultaneous SRM operations, we observed large number of SRM errors, the system performance dramatically decreased.
  - After cleanup, 70 simultaneous SRM operations, dCache still sustains stable data transfers. Further intensive tests needed to show threshold.
- SRM Transaction rate is determined by SRM load. Copy 450 short files with different client concurrencies:
  - 10 users: 120 SRM transactions per minute.
  - 50 users: 30 SRM transactions per minute.
  - 70 users: 26 SRM transactions per minute.
- Tested new hardware & tried in memory DB (tmpfs):
  - 60 concurrent file transfers (disk): 46 transactions per minute.
  - 60 concurrent file transfers (tmpfs): 63 transactions per minute (40% better).

SUN Thumper Test Results

- **150 clients sequentially reading 5 random 1.4G files.**
  - Throughput is 350 MB/s for almost 1 hour:



- 75 clients sequentially writing 3x1.4G files and 75 clients sequentially reading 4x1.4G randomly selected files.
  - Throughput is 200 MB/s write & 100 MB/s read:



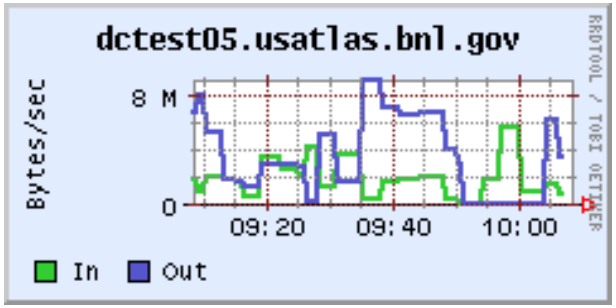
Recent dCache Activities at BNL

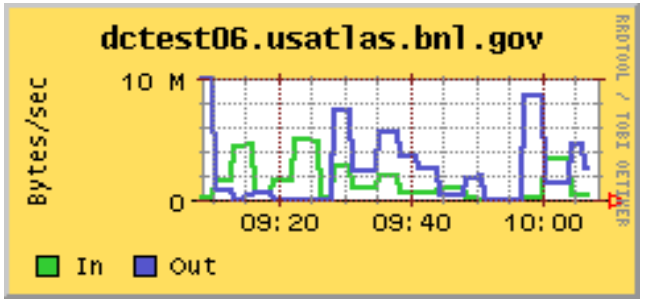
- Revised Atlas Data Rate Estimates: Based on BNL Data Storage Requirements. (Planned Rate 400MB/second)
  - 100 MB/s data writes and migration: Permanent (Disk0+Tape1): RAW data, allow 4 write pools, use dCache 1.7.0 central flush system to avoid mixing R+W. Each pool has 60 Mbyte/second.
  - 200 MB/s?: Disk Only (Disk1+Tape1): ESD and AOD data: directly write data into Linux Farm via GridFtp doors?
  - 100MB/s: Disk1+Tape1: Tier 2 simulations and User Analysis data: Allocate about 6 write pools with carefully tuned disk system. We already observed stable rates with only six write pools.
- dCache 1.7.0 was deployed on BNL testbed. New features, i.e. central flushing system, gPlazma/GUMS, were validated in the testbed.
- Plan to upgrade production system in ~2 weeks (need SRM overwrite).

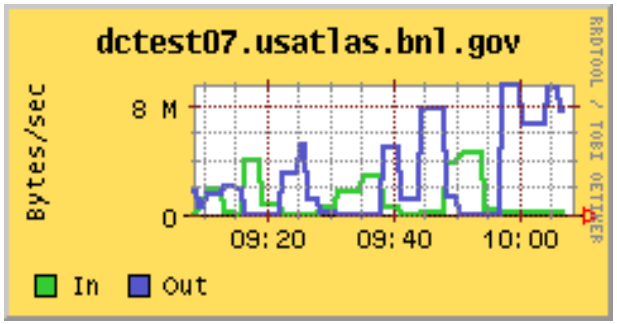
1.7 Tests: Central Flushing System

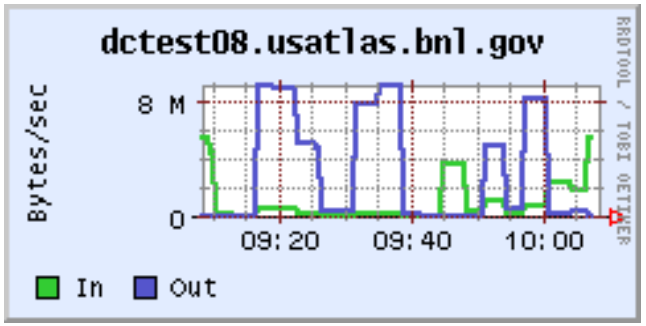
- Aim is to avoid simultaneous read/write operations on disk to improve the total throughput.
- The test consisted of copying files via dccp protocol from a client located in the same subnet to a directory on the dCache 1.7 test installation (PNFS).
- The AlternatingFlushSchedulerV1 driver was used. This driver was configured for a single Pool Group.
- There are two configurations considered on this test for the write pools.
  - Pool to pool connections allowed between write and read pools.
  - Trigger parameters
    - max.files=10
    - max.minutes=10
    - max.megabytes=200

Central flushing system with P2P









- Pool to Pool transfers allowed on this test.

Central Flush System & gPlazma

- A promising mechanism.
- More research and tests of this mechanism need to be done before we use it in our production environment:
  - Use two flush control managers to control flushing process for heavy and low loaded write pools per d-cache installation
  - To test flush control manager that controls a group of pools with P2P enabled and other that controls a group of pools without P2P transfers.
- **SRM gPlazma was tested using our GUMS server (also used by our globus gatekeepers).**
  - Initial results look good.

Phenix (RHIC) dCache

- Version: 1.6.6-5 (upgrade to 1.7 "soon")
- ~450 pools, 207 TB storage, 658K files on disk (173 TB)
- Aggregate throughput has exceeded 1.5 GBytes/s
- dCache is currently used as the end repository and archiving mechanism for the PHENIX data production stream.
- dCache is integrated into the PHENIX "Analysis Train" method, which aggregates user analysis jobs to run efficiently on common data subsets.