USATLAS dCache System at BNL

Zhenping (Jane) Liu, Razvan Popescu

ATLAS Computing Facility, Physics Department

Brookhaven National Lab

08/30/2005 DESY dCache Workshop

Agenda

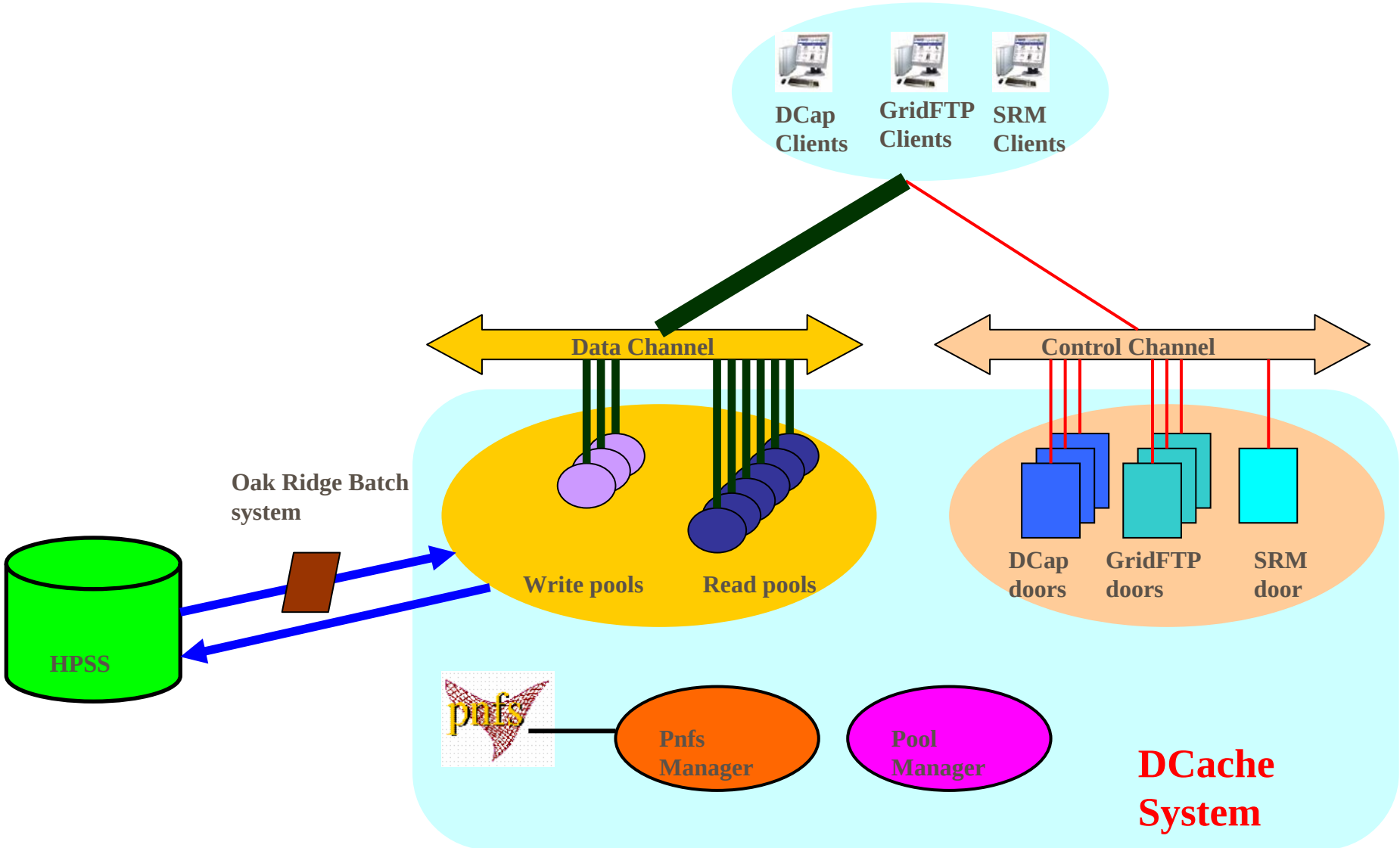
- BNL dCache system
- Plans
- Experiences and issues
- Suggestions

BNL dCache system

- In production service from November 2004.
- Works as a distributed disk caching system as a frontend for Mass Storage System - HPSS system.

BNL dCache system (Cont.)

- Hybrid model for read pool servers
  - Each node in Linux farm acts as both storage and computing unit.
- Dedicated core servers and write servers
  - Dedicated PNFS node, door nodes, write pool nodes.
  - More critical.
- Optimized backend tape prestage batch system.
  - Oak Ridge Batch System
- System Architecture (see the next slide)


DCap Clients
GridFTP Clients
SRM Clients
Data Channel
Control Channel
Oak Ridge Batch system
HPSS
Write pools
Read pools
DCap doors
GridFTP doors
SRM door
pnfs
Pnfs Manager
Pool Manager
DCache System

Size of the current system

Server Type	Numbers of servers	Disk cache space
PNFS Core server node	1 (dedicated)	N/A
SRM server node	1 (dedicated)	N/A
GridFTP and DCAP Core server nodes	4 (dedicated)	N/A
Internal/External Read pool nodes	322 (shared)	145 TB
Internal/External write pool nodes	8 (dedicated)	1 TB
Total	**336**	**146 TB**

Usage of the system

- Total amount of datasets (only production data counted)
  - 82.3TB as of 08/23/2005
- Used by Rome production grid jobs as data source.
  - Positive feedback.
  - Will use dCache as data source and destination, and also repository of intermediate data in the next version.
- Used in SC3 testing phase.

Statistics on transfer actions

Transfer Statistics (Daily Average)

	# Restore /day	Restore Rate (GB/day)	# Store/day	Store Rate (GB/day)	# Movers /day	Mover transfer rate(GB/day)
2005-Feb	236	59.0	1789	294.5	5403	1051.4
2005-Mar	311	84.5	2295	270.4	4111	461.2
2005-Apr	672	165.0	6891	442.9	14019	771.4
2005-May	450	96.8	5550	369.4	17950	972.6
2005-Jun	170	42.9	3218	166.6	9393	456.4
2005-Jul	564	173.1	5103	3174.1	8694	3853.1
2005-Aug	1272	48.9	2364	383.1	3801	1240.3

Note: SC3 testing Phase was run in July

Clients

- On-site users
  - Clients from Linux farm nodes (CONDOR jobs).
    - Local analysis application (using DCAP library or dccp)
    - Production grid jobs (submit to BNL)
  - Other users
- Off-site users
  - GridFTP clients
    - Production grid jobs from remote sites
    - Other grid users
  - SRM clients

Evaluation on dCache usage

- Pretty positive on the whole
  - Long-term solution for grid-enabled storage element.
  - USATLAS tier-2 centers will deploy dCache as storage elements soon.
- Nontrivial issues existed.

Long-term plan

- To build petabyte-scale grid-enabled storage system
  - Several Petabyte ATLAS data generated every year.
  - Petabyte-scale disk space on thousands of farm nodes to hold most data in disk.
  - HPSS as tape backup for all data.

Long-term plan (Cont.)

- DCache as distributed storage system solution
  - Advantages:
    - Unified namespace;
    - load balanced and fault tolerant
      - Multiple servers of same type, e.g., pools, all doors
      - Dynamically replicate files to avoid hot spot.
    - High performance
      - Direct data I/O from/to pool servers
      - Aggregated data throughput can be very high.
    - Clever selection mechanism and flexible system tuning;
    - Multiple access protocols (including standard grid interfaces);
    - Cheap Linux farm solution to achieve high performance throughput.

Long-term plan (Cont.)

- Issues: potential bottlenecks in dCache
  - Centralized metadata database currently.
  - Single metadata management component (PnfsManager).
- Many issues need to be investigated
  - Is dCache scalable to large cluster (thousands of nodes)?
    - Higher PNFS hit rate expected.
    - Many small dCache systems or one/several big dCache system(s)?
  - Will network I/O be a bottleneck for a large cluster in data-intensive computing environment?
    - How to avoid unnecessary data I/O and network I/O on Linux farm nodes?
  - Other issues not aware of yet?

Experiences and issues

- Read pool servers shares nodes with computing.
  - Utilizing idle disks on compute nodes.
  - Hybrid model works fine.
- Write pool servers
  - Much higher access rate.
  - Should run on dedicated servers.
    - Crashed frequently in the past when sharing node with computing.
    - Dedicated servers solved the problem.
  - XFS shows better performance than EXT3.

Experiences and issues (Cont.)

- SRM pinManager crashed a lot when SRM clients read from dCache to off-site even with mild rate.
  - FNAL provided a temporary fix and is also working on long-term solution.
- FTS doesn’t support srmcopy
  - All data traffic had to go over a limited number of GridFtp doors during SC3.
    - No direct data traffic to write pools; Contradiction with scalability.

Experiences and issues (Cont.)

- PNFS bottleneck problem.
  - Continuous write with the rate 1000 times/hour seemed causing very high load (>20) on PNFS core server.
- How to split an existed big directory into multiple database?

Experiences and issues (Cont.)

- No support for GridFTP 3rd party transfer
  - 3rd party transfer is very common in grid
  - SRM supports 3rd part transfer, however not deployed on all sites.
  - Next version of USATLAS production system will use srmcp for third party transfer.

Experiences and issues (Cont.)

- System administration
  - Not easy in early phase.
  - Much better later
    - Great help from DESY and FNAL dCache project team.
    - More documents
    - Bugs fixed in software.
    - Tools developed to avoid, detect and solve problems.

Experiences and issues (Cont.)

- Big size (>2G) log file caused the door off-line.
  - Solution: logroate daily
- 2GB limitation on PNFS gdbm database size
  - Solution:
    - Multiple databases
    - Use Postgres as PNFS database system (no 2GB limitation).
  - Issues: performance issue with large database.
- Client process hangs up when pool crashes in the middle of transfer.

Experiences and issues (Cont.)

- Sometimes, GridFTP connection couldn’t be closed properly.
- Other issues
  - A list was sent to dCache team.

Suggestion

- Build a forum for dCache administration discussion.
  - Consortium of developers and site administrators
  - Sharing issues, solutions and experiences.
  - Decreasing the burden on developers.
    - No redundant questions for developers.
    - Admin can help answer questions too.
  - New site admins can benefit a lot.

Suggestion (Cont.)

- System administration manual
  - Much better manual now compared to last year.
  - Still need more details, especially on system tuning.
  - Maybe experienced site admins can contribute too.

Suggestion (Cont.)

- Sharing system administration and monitoring tools
  - Additional monitoring tools at FNAL.
    - Into standard package?
  - Site admins can contribute useful self-made tools of common interests.


Thank You !