Introduction to dCache

Zhenping (Jane) Liu

ATLAS Computing Facility, Physics Department
Brookhaven National Lab

09/12 – 09/13, 2005 USATLAS Tier-1 & Tier-2 dCache Workshop

Agenda

- dCache System
- dCache Components
- Sample layouts

dCache project

- Developed by DESY and FERMI.
- In production since 2001.

dCache system

- Provides a system for transparent access to huge amounts of data, distributed among a large number of heterogeneous server nodes, or stored on tapes.
  - Providing the users with a single virtual filesystem tree.
- When Connected to a tertiary storage system, the cache simulates unlimited direct access storage space.
  - Significantly improving the efficiency of connected tape storage systems, through caching, i.e. gather & flush, and scheduled staging techniques.
  - Data exchanges to and from the underlying HSM are performed automatically and invisibly to the user.

dCache system (Cont.)

- Clever selection mechanism and flexible system tuning
  - Determining whether the file is already stored on one or more disks or on HSM.
  - Determining the source or destination dCache pool based on storage group and network mask of clients, also CPU load and disk space, configuration of the dCache pools.
- High performance and load balanced
  - Optimizing the throughput to and from data clients as well as smoothening the load of the connected disk storage nodes by dynamically replicating files upon the detection of hot spots.

dCache system (Cont.)

- Tolerant against failures of its servers.
    - Multiple pools, Multiple doors of each type
- Various access protocols, including GRIDFTP, SRM and DCAP.
    - Local: DCAP (e.g., dccp command line tool or dCap library)
    - Grid users: GridFTP, SRM
        - Provide SRM based storage element
- Cheap Linux farm solution to achieve high performance throughput.

dCache Components

Ftp Server (gsi, kerberos)

Pnfs

dCache Core

Cell Package

PNFS

- Used by dCache as metadata database for the file entries.
  - Not designed for storage of actual files.
  - Managing the filesystem hierarchy and standard metadata of a UNIX filesystem
- Serves as mountable filesystem presenting the file repository.
  - Implementing an NFS server.

Cell Package

- A framework for a distributed and scalable server system in Java.
  - All of **dCache** makes use of the **cell package**.
- The **dCache** system is divided into cells which communicate with each other via messages.

dCache core module

dCache Core ->

PoolManager

Finds best pool for each request.

Cleaner

Forwards 'rm'requests from pnfs filesystem to pools.

PnfsManager

Interface between dCache and filesystem.

Door (Launcher)

Starts appropriate door for incoming connections.

Pool

Does space management and launches 'movers'.

The I/O Doors

- The I/O Doors
  - Clients send requests for a datafile to a "door" of a dCache system.
  - A door is a network server which performs user authentication and forwards client requests to the pool managers.
  - There can be more than one type of door to a dCache system, each potentially handling a distinct authentication mechanism and each perhaps residing on a separate host.
  - The concept of Doors allows to have multiple instances of one same kind of door running on different hosts for load sharing and fail safeness.

The PnfsManager

- Interface between dCache and PNFS

The PoolManager

- Each space request either for PUT or GET is handled by the PoolManager.
- It performs a pre-selection of possible pools and queries the selected pools for more information to optimize the final decision.
- Each Pool has to register itself to the PoolManager together with information about its affinity to certain storage classes and possibly about its topology and performance.

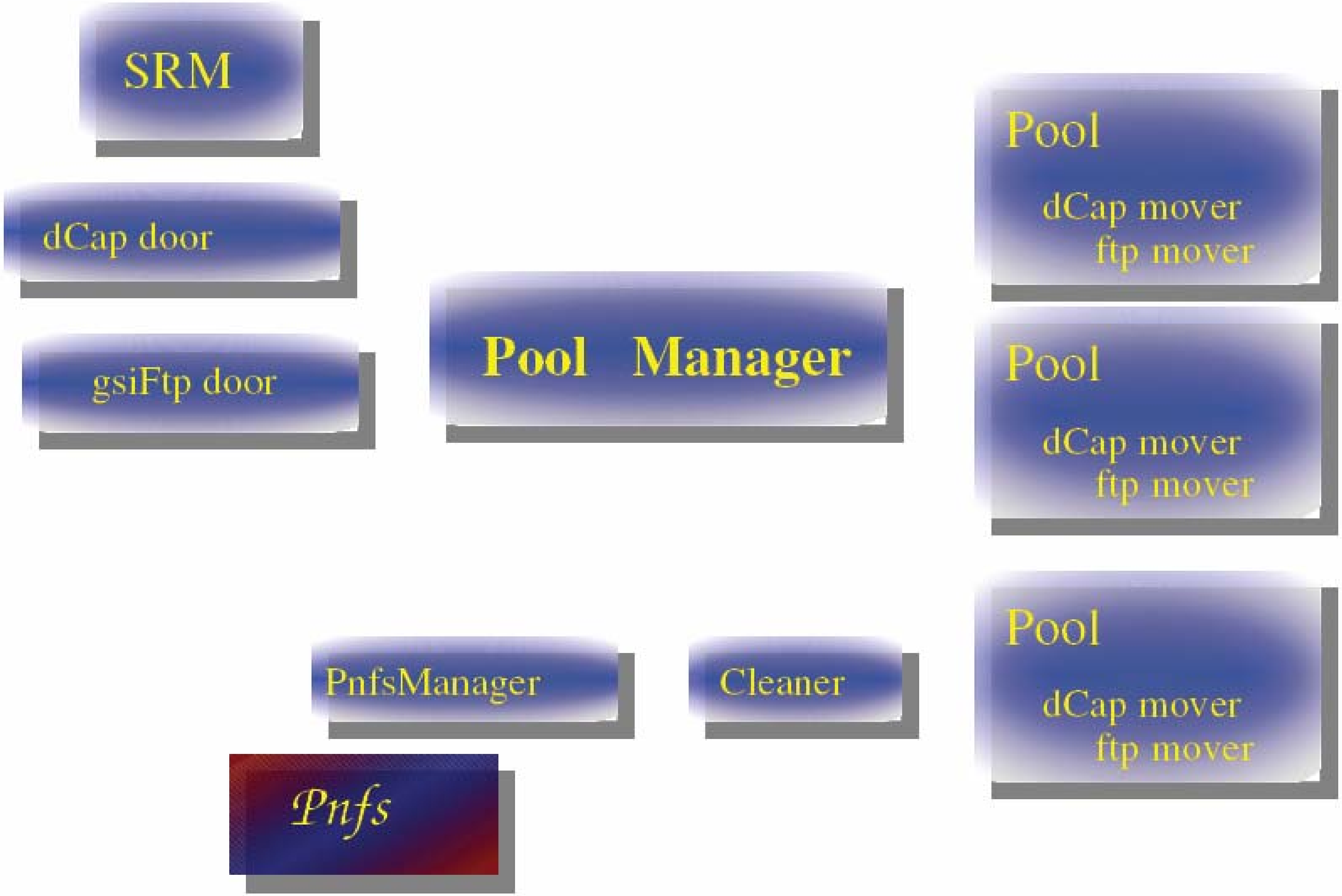
The Pool

- The pool is responsible for a contiguous disk area:
  - Monitoring disk space.
  - Holding a list of files, which are candidates for removal if disk space is running short.
  - Initiating the file copy process (Mover) to and from tertiary storage.
  - It connects to data clients for the data transfer.
  - It monitors the total bandwidth to and from the disk area.

The Cleaner

- Responsible for deleting the actual files from the pools asyncronously

dCache tool box



SRM

dCap door

gsiFtp door

Pool Manager

Pool
dCap mover
ftp mover

Pool
dCap mover
ftp mover

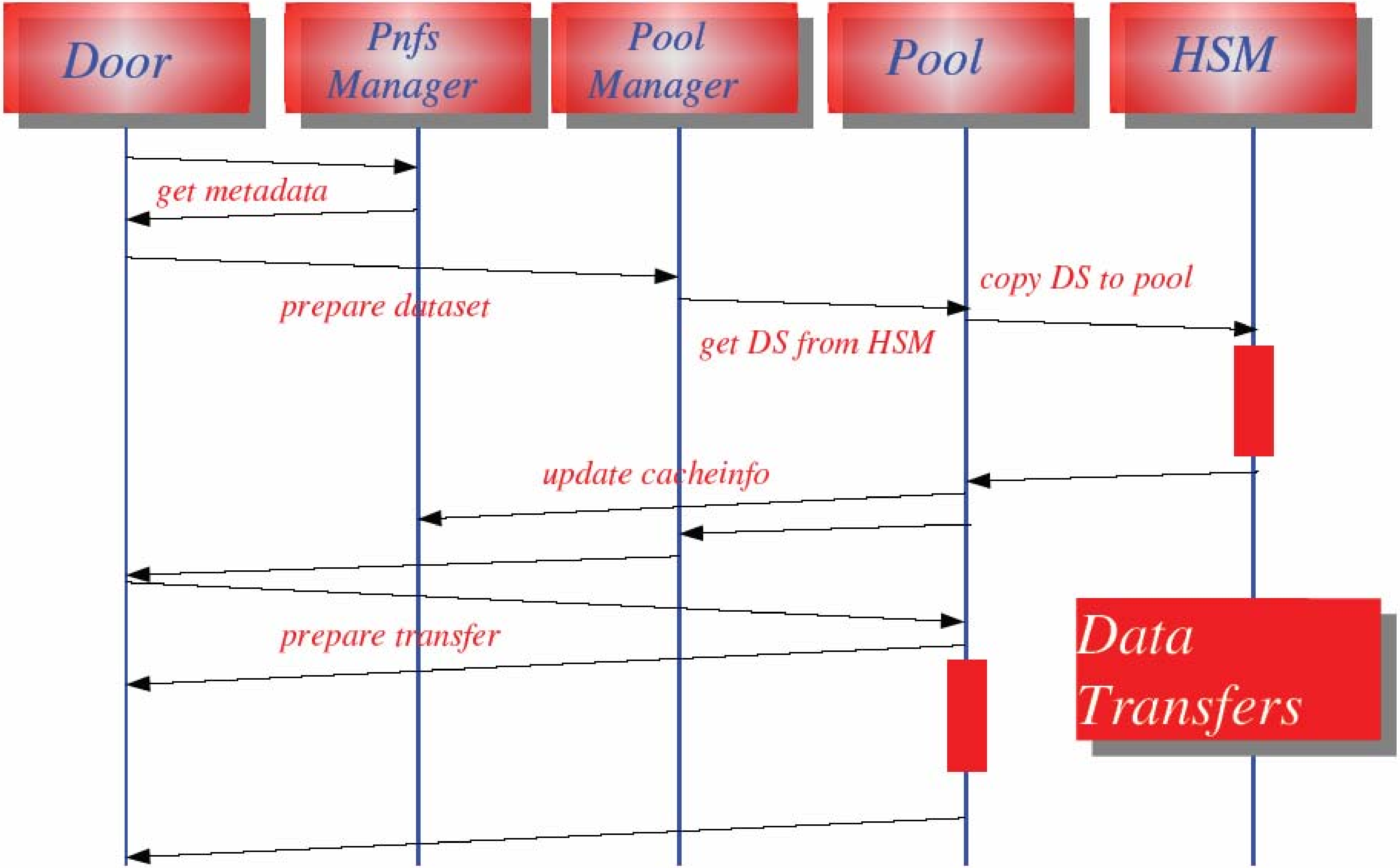
Pool
dCap mover
ftp mover

PnfsManager

Cleaner

Pnfs

dCache transfer



Extended Central Services

- Prestager
- HSM Flush Manager
- Resilient Manager
    - Trying to keep number of replicas available online for the each file in the predefined valid range (min, max).

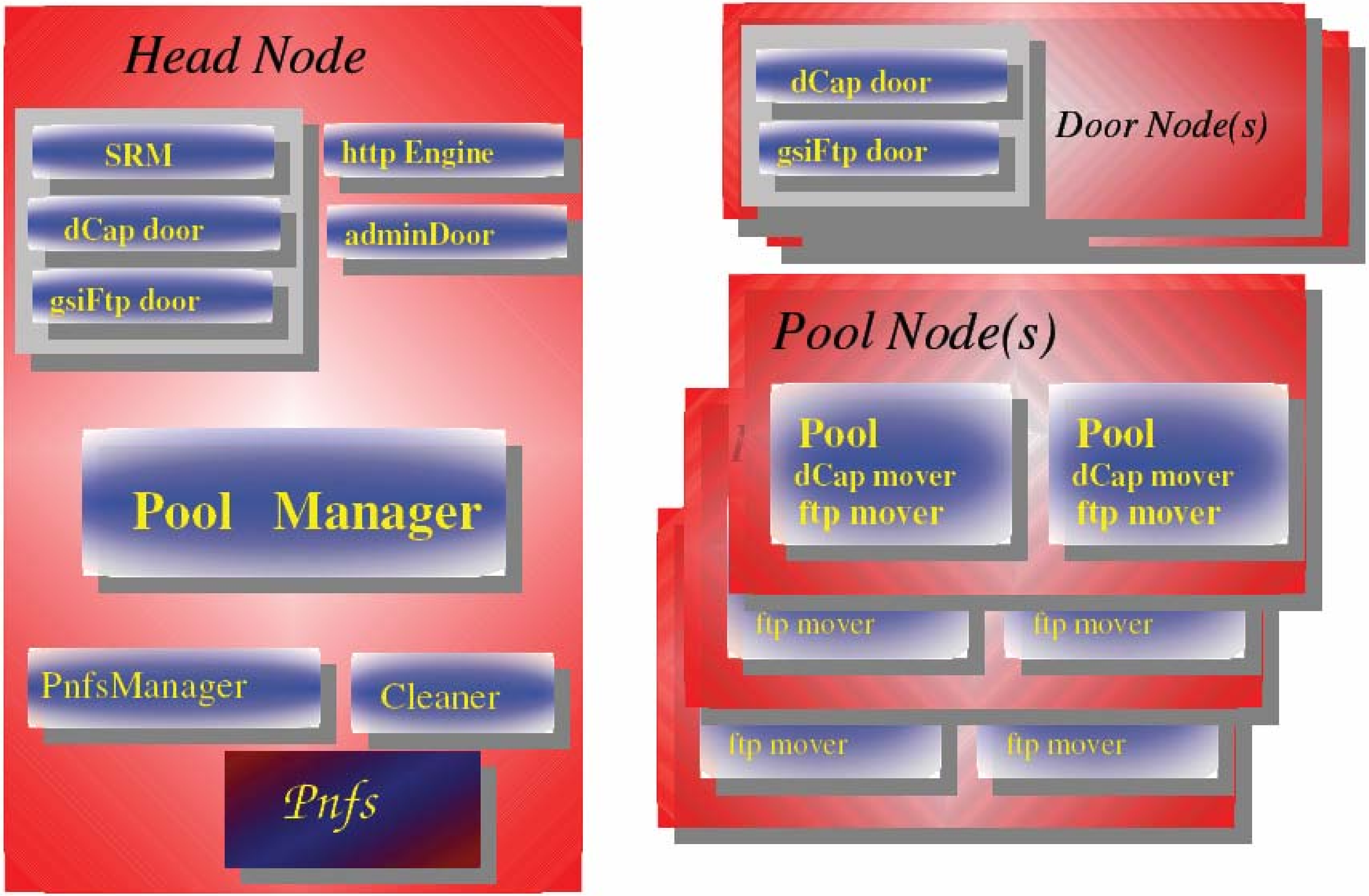
Other Modules

- Admin Door
    - A powerful administration interface.
    - Accessed with the SSH protocol
- HTTP Engine
    - dCache monitoring page

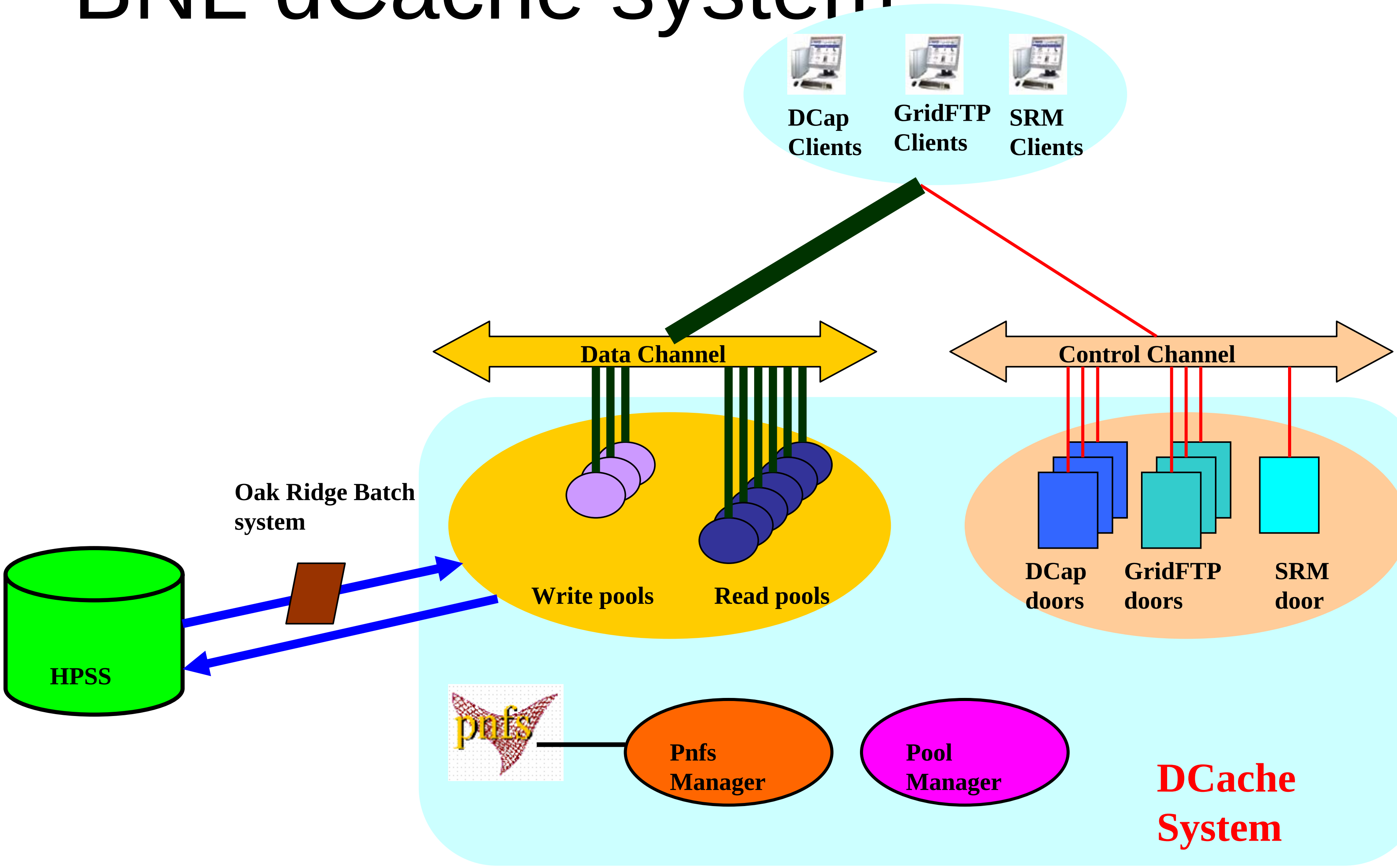
Sample systems

- Classical one from Patrick’s presentation
- BNL dCache system

dCache : classical



BNL dCache system


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