Grid Interfaces to dCache

Timur Perelmutov

for the dCache team

Joint EGGE and OSG Workshop on
Data Handling in Production Grids
HPDC 2007, Monterey, CA

SRM V1.1 interface

- SRM V1.1 has been a part of dCache for over 4 years
- Used in production by US-CMS for over 2 years
- Solid protocol **but**
- Did not include
  - Explicit Space Reservation and Management
  - Directory functions
  - File Access Permission management
  - Abstractions to describe type and quality of service
- Weak Error and status reporting
- SRM 2 addressed many of the issues

LHC needs

- Common interface to T0, T1 and T2 storage
- Guarantee of space availability
  - Space Reservation
- Storage Class differentiation
  - Access Latency and Retention Policy
- Flexible Namespace management
  - Directory Functions
- ACL Support
  - Permission management functionality
- SRM V2.2 is the answer

dCache SRM v2.2 history

- Prototype of SRM 2.0 interface demonstrated at SC 2003
- Work on dCache SRM 2.1 since late 2004
- LHC experiments input led to SRM 2.2 definition
- in May 2006 WLCG chose a subset of SRM v2.2 which became a dCache project target
- Beta of dCache 1.8 with SRM 2.2 released in April 2007

dCache Services that support SRM

- SpaceManager
  - dCache introduced a new way to partition total space according to their support RetentionPolicy, AccessLatency and VO Groups/Roles.
  - Support for a streaming to HSM model
- LoginBroker – a service for the discovery of all dCache Doors (a transfer protocol incarnation deployed on a given host:port)
- PinManager –
  - a service for staging and pinning files (Control of online state)
  - Unifies pin and bringOnline requests
  - pin lifetime management
- Pool Repository and Namespace are modified to better support “pin in Cache” operation and “Online” file parameters

Grid Access Control

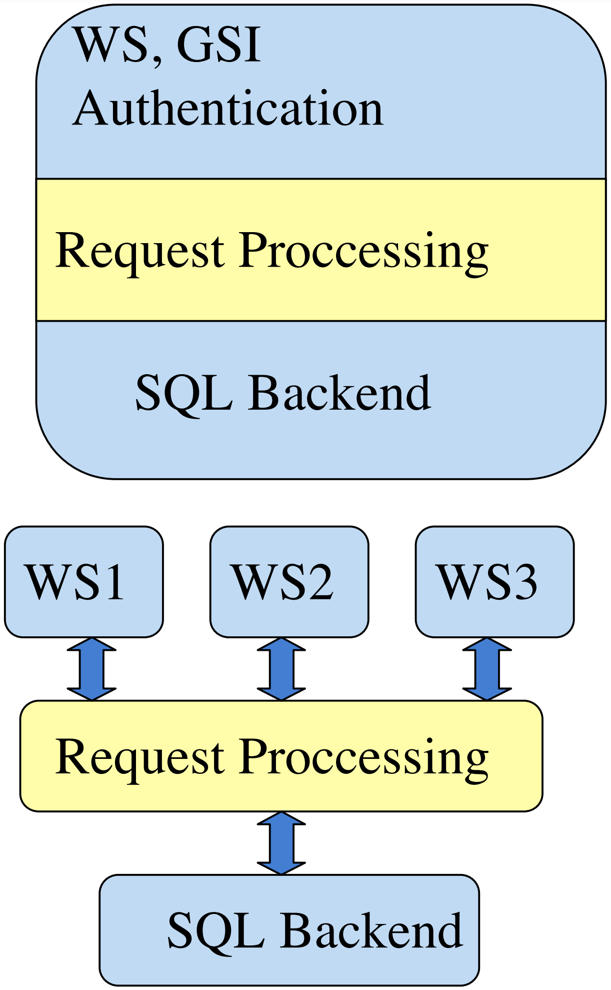
- Dcache Authorization (gPlazma)
  - Supports VO Certificate proxies
  - Multiple VO Memberships
- PNFS Namespace Service
  - Files are owned by a particular User and Group.
  - No ACL Support
- Chimera Namespace (currently in Beta testing)
  - Full ACL Support by Fall 2007
- SRM permission management functions
  - need both VO Authorization System and ACL capable Namespace Service
  - Full support of SRM Permission Management will follow

GridFTP Door

- Gridftp V1 - dCache 1.6, 1.7
  - GSI Authentication
  - Stream and Extended Block (multi-stream) modes of transfer
  - Protocol makes penetrating firewalls and accessing private network data difficult
  - In production for the last 5 years
- GridFTP V2 -dCache 1.8
  - Get/Put for data transfer
  - X Block transfer mode
  - Data Integrity Verification

To Do: Horizontal Scaling

- SRM Interface dCache
  - WEB Service deployed in Tomcat/Axis
  - SQL database for Persistent State Storage
  - Monolith module
  - GSI Authentication – 90% CPU load
  - Does not scale to multi-nodes
- Future work
  - Decuple Web Service interface from Business Logic
  - Allow multiple WS endpoints for a single system
  - This will enable usage of DNS Load Balancing

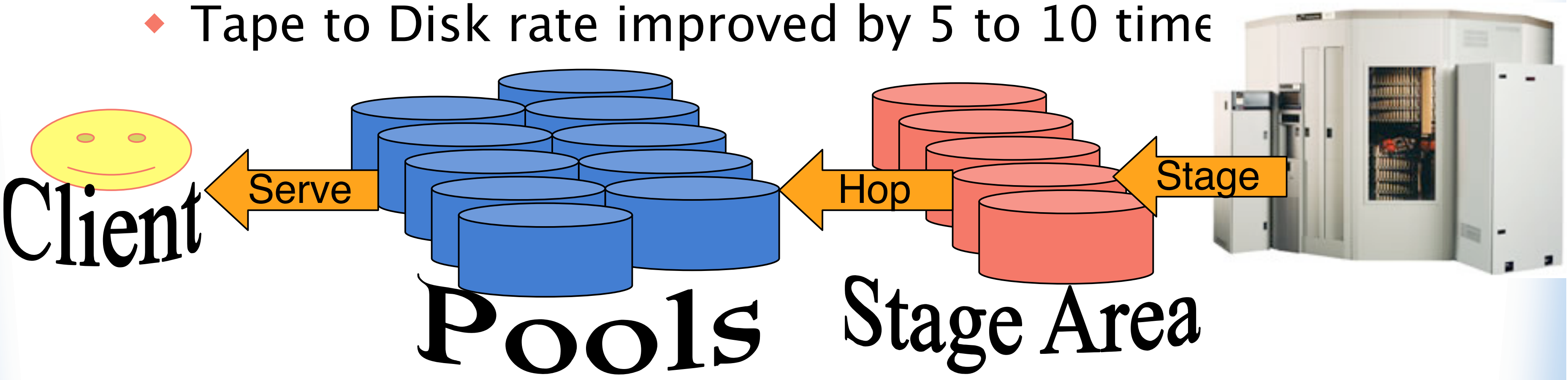


To Do: Automatic Space Recovery

- Open Science Grid storage is open to opportunistic use by multiple experiments and organizations
- Requires ability to Guarantee that upon the expiration of the lease on disk space, it will be automatically cleaned up
  - Files in the expired Space will be automatically garbage collected
- OSG Contribution will help add support for volatile files with managed lifetimes
- SpaceManager will be used for enabling this functionality

dCache installation example US-CMS T1 (1)

- Stage Area -11 nodes-10TB
  - Pools for staging files from tapes managed by dCache File Hopping
  - Pool-to-Pool copy to read pools
  - Limited resource tape drives running at full rate
  - Tape to Disk rate improved by 5 to 10 time

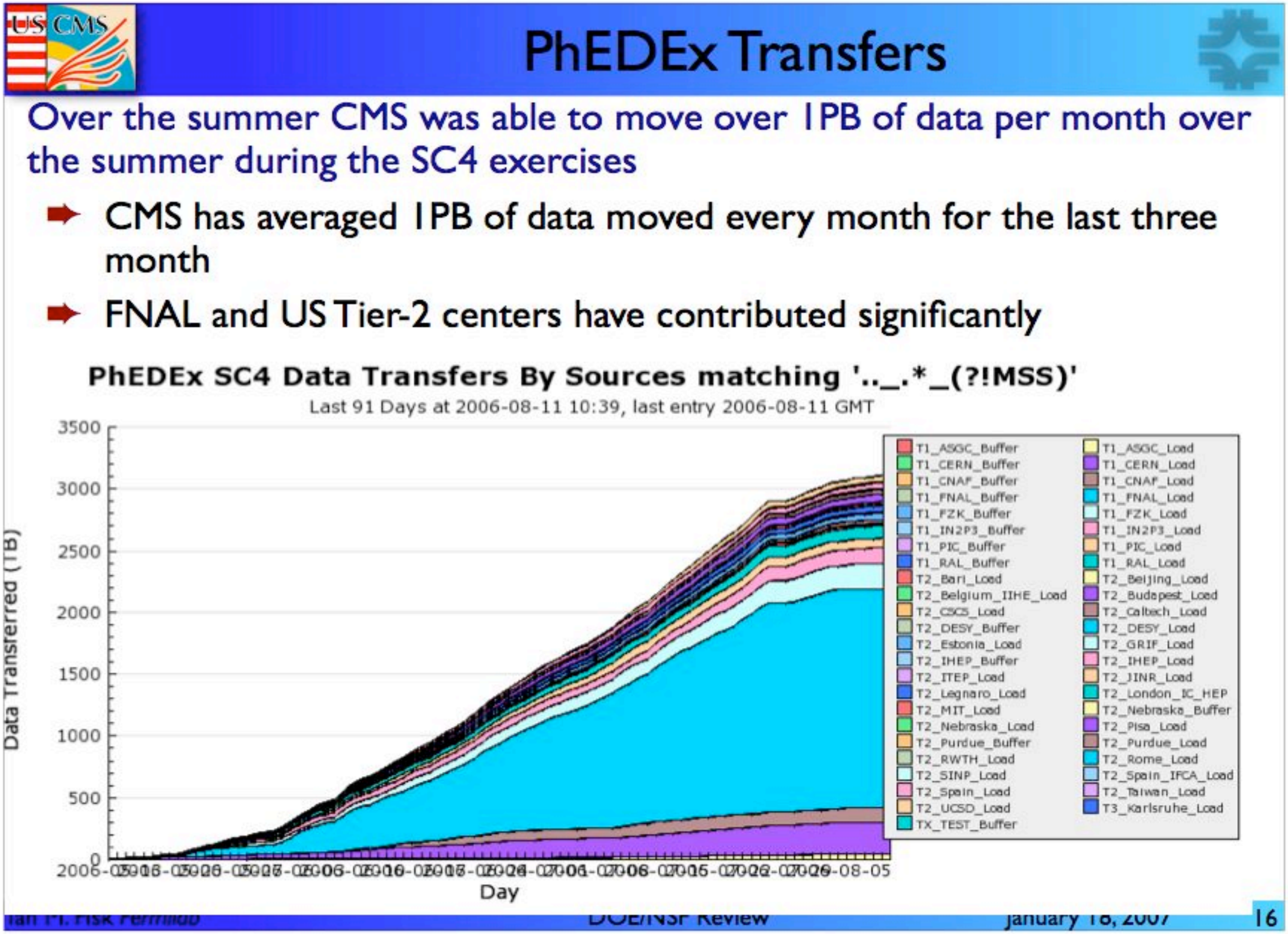


dCache installation example US-CMS T1 (2)

- Read/Write Area
  - >100 nodes
  - 700TB of Tape Backed pools
  - Will Grow to 1.5 PETABYTE By September 2007
  - One Gridftp server per node, used by SRM
  - All pools allow both WAN and LAN access
  - To improve reliability each pool has LAN and WAN queue
    - LAN Queue with 600 to 1800 active movers
    - WAN Queue with 5 to 15 active movers
    - Busy pool nodes saturate 2xGE for hours on end *each.*
    - Aggregate transfers exceed 40 Gb/s LAN+WAN.

dCache installation example US-CMS T1 (3)

- 2 Resilient Managers in the same dCache
- Worker Nodes Resilient Manager
  - PRECIOUS file
  - ~ 650 Worker nodes
  - More than 100TB
  - 3 copies of each file
- Precious Pools Resilient Manager
  - 55 TB of non-tape-backed PRECIOUS and RESILIENT pools for unmerged output
- Replica Monitoring is very useful



PhEDEx Transfers

Over the summer CMS was able to move over 1PB of data per month over the summer during the SC4 exercises

- CMS has averaged 1PB of data moved every month for the last three month
- FNAL and US Tier-2 centers have contributed significantly

PhEDEx SC4 Data Transfers By Sources matching '.._.*_(?!MSS)'

Last 91 Days at 2006-08-11 10:39, last entry 2006-08-11 GMT

Tier-2 Centers

Tier-2s also generally met the 50% milestone

- Sum of Tier-2 capacity is similar to the total Tier-1, as indicated in the model
- Tier-2 networking is in good shape

Site	CPU (kSI2K)	Disk (TB)	WAN (Gb/s)
Caltech	538	56	10
Florida	519	104	10
MIT	92	54	1
Nebraska	347	53	0.6
Purdue	743	184	10
UCSD	318	98	10
Wisconsin	547	119	10
TOTAL	3104	668	

References

- dCache www.dcache.org
- dCache SRM http://srm.fnal.gov
- SRM Working Group http://sdm.lbl.gov/srm-wg/
- SRM V2.2 spec http://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.html
- Organization: http://uscms.org
- US-CMS T1 dCache http://cmsdca.fnal.gov

Technical Details

- Follow

dCache support Types of Storage Services in SRM V2.2

- AccessLatency (Online, Nearline) – Speed of access to the data
- RetentionPolicy (Replica, Output, Custodial)*– quality of retention service
- dCache had to
  - Update PoolManager pool selection mechanism
  - New Pool repository code
  - SpaceManager – Space Reservation as vehicle for assignment of these attributes

* WLCG interpretation: Replica – Disk, Custodial –Tape, Online Output – not used.

dCache SpaceManager

- dCache PoolManager introduced a new way to group Pools according to their support RetentionPolicy, AccessLatency and VO Groups/Roles, such groups are called LinkGroups.
- SpaceManager makes reservations in one of such LinkGroups
  - Reservation can exceed size of the pool
  - LinkGroups can be used as VO based quotas
  - Late binding between transfer and a pool
- Support for a streaming to HSM model

Space Access Control

- LinkGroups in dCache can be assigned lists of VO Groups and VO Roles that are allowed to perform Space Reservations
- No Functionality for Restricting Access to the Space Reservations themselves