Managed Storage @ GRID

or

why NFSv4.1 is not enough

Tigran Mkrtchyan
for dCache Team
What the hell do physicists do?

• Physicists are **hackers** – they just want to know how things work.
• In modern physics, given cause does not produce the same effect.
• Statistics is used to describe behavior.
• Physics data is IMMUTABLE: you keep it forever or you removed it, but you never FIX it!
Large Hadron Collider:

Expected start July 2008
800 million collisions per second
(25 km long)
Data rate ~ 1.5 GB per second
~15PB per year
Multiple tier model

Tier Structure

Tier 0
Offline farm
CERN computer centre
Tier 1
Online system
Tier 1 National centres
RAL, UK
USA
Tier 2
Germany
Italy
France
Tier 2 Regional groups
NorthGrid
ScotGrid
SouthGrid
London
Tier 3
Sheffield
Lancaster
Manchester
Liverpool
Tier 3 Institutes
Tier 4
Workstations
GRID as core infrastructure

GRID middleware applied to solve two major goals:

• Physical
  • space, power, cooling, connectivity
• Political
  • let regional investors to spend many for regional centers
What is a GRID?

“The term Grid computing originated in the early 1990s as a metaphor for **making computer power as easy to access as an electric power grid.**”
“The term Grid as a metaphor for making computer power as easy to access as an electric power grid originated in the early 1990s.”
What is a GRID?

“The term Grid computing originated in the early 1990s as a metaphor for making computer power as easy to access as an electric power grid. While or most of the people GRID is a distributed CPU resources, it's all about distributed storage!”
To hide storage system implementation a top level management interface was defined - SRM.

SRM together with 'Information Provider', which allows to query storage system called 'Storage Element (SE)'
Storage Resource Managers (SRMs) are middleware components whose function is to provide dynamic space allocation and file management on shared storage components on the Grid.

SRM interface defines following functions:

- Data Transfer
- File Pining/UnPining
- Space Management
- Request Status queries
- Directory operations
- Permission management
SRM data transfer based on two concepts: SURL and TURL.

- SURL - is a “site URL” which consists of “srm://host.at.site/<path>”.
- TURL - is the “transfer URL” that an SRM returns to a client for the client to “get” or “put” a file in that location. It consists of “protocol://TFN”, where the protocol must be a specific transfer protocol selected by SRM from the list of protocols provided by the client.

**SRM behaves as a load balancer and redirector**

**de facto**, GSI enabled FTP protocol is used for transfers
SRM PUT (ftp)

SRM

get TURL

data flow

ftp write

dCache

Dest.

Client

ftp
SRM GET (ftp)

data flow

get TURL

ftp READ

Client

Dest.

dCache
SRM for pNFS people

- **Client**
  - PUT/GET
  - TURL
  - READ/WRITE
  - DONE
  - SUCCESS/FAIL

- **Server**
  - LOCK + LAYOUTGET on bunch of files
  - READ/WRITE
  - UNLOCK
  - LAYOUTRETUN
  - DEAD
  - LAYOUTRETUN
  - DESTROY_SESSION
  - LAYOUTRETUN

- **Data Server**
  - READ/WRITE
SRM COPY-PULL

GET

SRM

Need It!

Client

DESY
SRM Space Management

- allows to reserve space prior the transfer
  - Quota system, where you never get “file system full”
- has three space descriptions and allows transitions between them:
  - CUSTODIAL, ONLINE (Tape1Disk1)
  - CUSTODIAL, NEARLINE (Tape1Disk0)
  - REPLICA, ONLINE (Tape0Disk1)
SRM Space Management (use case)

1. Reserve space for data set in T0D1 space
2. Put data set into space
3. Analyze Dataset
4. Is Data OK?
   - Yes: Move dataset into T1Dx space
   - No: Remove dataset
5. Send dataset to remote site
SRM Space Management (use case)

1. Reserve space for data set in T0D1 space
2. Put data set into space
3. Analyze Dataset
4. Is Data OK?
   - No: Remove dataset
   - Yes: Move dataset into T1Dx space
5. Send dataset to remote site
• X.509 based certificates
• extensions for Virtual Organizations (VO) support
• no trusted hosts

subject : /O=GermanGrid/OU=DESY/CN=Tigran Mkrtchyan/CN=proxy
issuer : /O=GermanGrid/OU=DESY/CN=Tigran Mkrtchyan
identity : /O=GermanGrid/OU=DESY/CN=Tigran Mkrtchyan
type : proxy
strength : 512 bits
timeleft : 11:59:40

=== VO desy extension information ===
VO : desy
subject : /O=GermanGrid/OU=DESY/CN=Tigran Mkrtchyan
issuer : /C=DE/O=GermanGrid/OU=DESY/CN=host/grid-voms.desy.de
attribute : /desy/Role=NULL/Capability=NULL
timeleft : 11:59:40
SRM – Uniform Data Access

SRM Client

SRM Interface

- dCache +MSS
- dCache
- CASTOR
- DPM
SRM – Uniform Data Access

SRM Client

SRM Interface

- dCache + MSS
- dCache
- CASTOR
- DPM

Need It!
We are doing well!
Mission Impossible

We are doing well!

dCache installations

Averaged Throughput during the last 24 hrs (10/04 - 11/04)
From CERN to ALL SITES

We are doing well!
dCache - Background

Access Time & Size

Price

CPU Cache

Disk Array

Tape Storage

Access Time & Size

Price

$10^{-9}$

$10^{-4}$

$10^3$
dCache - Background

- CPU Cache
- Disk Array
- Tape Storage

Access Time & Size

Price
The goal of the project is:

- to share and optimize access to non-sharable storage devices, like tape drives,
- make use of slower and cheaper drive technology without overall performance reduction,
- to provide a system for storing and retrieving huge amounts of data, distributed among a large number of heterogeneous server nodes, under a single virtual filesystem tree with a variety of standard access methods.
Requirement is:

to provide a system for storing and retrieving huge amounts of data, distributed among a large number of heterogeneous server nodes, under a single virtual filesystem tree with a variety of standard access methods.
dCache Design

- Name Space Provider
  - size, owner, acls, checksum, ...
- Pool Selection Unit
- Protocol Specific Doors
- Multiprotocol Pools
  - can talk several protocols simultaneously
dCache Design

Namespace operations

Namespace

Client

IO requests

Disk Pools

IO Door

Data server selection

IO requests
dCache Design

- Pools are grouped into PoolGroups
- PoolGroup selected by flow direction, 'path'(file set), protocol and client IP
- Pool selected by cost, where cost is

\[ n \times \text{<CPU cost>} + m \times \text{<space cost>} \]

- \( n=1, m=0 \) : fill network bandwidth first
- \( n=0, m=1 \) : fill empty servers first
Files arrives to a pool and declared as **Precious**

**Precious** files flushed according policy - time, size, number of files.
MSS connectivity

Cached files can be delivered immediately

Missing files retrieved from the MSS first
MSS connectivity

Write

Write Pools

Flush

Pool 2 Pool on request

Read

Read Pools

Restore
dCache let us build very large (capacity and bandwidth wise) storage system with small, independent building blocks. Building block need to provide:

- JVM $\geq 1.5$ (all components are Java based)
- local filesystem
- network Interface

No IO penalty while using Java.
Current Status

- Project started June of 2000 as a joint effort of DESY and FNAL
- First prototype April 2001
- In Production since March 2002
- Supported local access Protocols: dcap, xrootd
- Supported WAN access Protocols: ftp, http
- Deployed on AIX, Linux (x86, Power, x64), Solaris (Sparc, AMD)
- Run over country border
- Has an interface to OSM, Enstore, HPSS, TSM, DMF
  - easy to add any other MSS
- Largest Installation 2PB (FNAL)
  - ~1800 pools
  - ~1.2 GB/s WAN (Peak rate – 2.5 GB/s!)
  - 60 TB/day read (~100000 files!)
  - 2 TB/day write (8000 files)
Current Status (NorduGrid)
The dCache's **Namespace** provider called Pnfs:

*Perfectly Normal File System*

developed in 1997 and currently in replacement.
Uniform Data Access
Why new protocols

- There is a three 'popular' protocols used in High Energy Physics:
  - dCap – dCache Access Protocol
  - rfio – Remote File IO
  - xroot – eXtended ROOT IO

- all protocols was designed, while NFSv2/3 was not distributed
- existing distributed solutions not fit well
  - and expensive ( all of them )
  - and require special hardware
  - or require special OS/kernel versions
- fit well to dCache (and others) architecture
- Open Standard Protocol supported by industry NFSv4.1
- Client comes 'for free' with Operating System
The Vision:

Local Analysis Farm

Distributed Storage

NFSv4.1

SRM UP-Link

Need It!
References:

- www.dCache.ORG
Special Tanks to:

Andy Adamson (CITI)
Benny Halevy (Panasas)
Lisa Week (SUN)
Sam Falkner (SUN)
Robert Gordon (SUN)