Introduction to SRM v2.2

2nd hands-on dCache Workshop - University of Cologne
29-30 April 2008

Flavia Donno
CERN

Outline

- Storage Services in WLCG
- The classic SE
- The requirements for a Storage protocol
- The Storage Resource Manager v2.2: concepts and main methods
- Basic Use Cases
- The GLUE Schema
- The S2 testing framework for SRM v2.2
- The Storage coordination bodies

Storage Services in WLCG

- ***Storage Services*** are ***crucial*** components of the ***Worldwide LHC Computing Grid*** (WLCG) infrastructure spanning more than 200 sites and serving computing and storage resources to the High Energy Physics LHC communities.
- Up to ***tens of Petabytes of data*** are collected every year by the 4 LHC experiments at CERN.
- It is crucial to ***efficiently transfer*** data to ***Tier-1s*** that contribute with their storage and computing power to the reconstruction step.
- An important role is also covered by the ***Tier-2s*** that provide experiments with the ***results of the simulation***. Such results need to be transferred to Tier-1s and safely stored on permanent media.

Storage Services in WLCG: the Classic SE

- ***The Classic SE*** : an optimized FTP server with Grid authentication and authorization.
  - The first Storage Server in the Grid based on Globus GridFTP
  - Very simple solution that included simple and complex tape-based systems
- What are the capabilities of such a service ?
  - No possibility to query the service itself about its status, space available, etc. (one has to rely on the Information System)
- How are data accessed on such a storage server ?
  - Protocols supported: NFS/file, rfio, root, gsiftp
  - Discovery of related information
    - Different root directory for GridFTP and NFS or rfio
- What about growing file systems according to needs ? Or more in general managing space ?
  - Sometimes very hard
  - No explicit support for tape backend (pre-staging, pool selection, etc.)

Requirements definition by dates

- In ***June 2005*** the ***Baseline Service Working Group*** published a report:
  - http://lcg.web.cern.ch/LCG/peb/bs/BSReport-v1.0.pdf
  - A Storage Element Service is mandatory and high priority.
  - The experiment requirements for a Grid storage service are defined
  - The full set of recommended feature available by ***February 2006***
  - Experiments agree to use only high-level tools as interface to SRM
- The report was based on the early experience acquired with SRM v1.1 and v2.1 (never deployed).
- ***Mumbai workshop*** (***CHEP2006***): the experiments had learned more about what was needed and changed their requirements.
- In ***May 2006*** at FNAL the WLCG SRM Memorandum of Understanding (MoU) was agreed on:
  - http://cd-docdb.fnal.gov/0015/001583/001/SRMLCG-MoU-day2%5B1%5D.pdf

Basic requirements

- Support for Permanent files (and volatile copies)
- Support for Permanent Space
- Space Reservation : static only per VO.
  - If dynamic space reservation is available, allow for the possibility of releasing the allocated space
- Permission Functions only on directories based on VOMS group/roles
- Directory Management Functions
- Data Transfer and File Removal Functions
- File access protocol negotiation
- VO-specific relative paths

The Storage Resource Manager SRM v2.2

- The *Storage Resource Manager* (SRM) is a ***middleware component*** whose function is to provide ***dynamic space allocation*** and ***file management*** on shared storage components ***on the Grid***.
- More precisely, the SRM is a ***Grid service*** with several different implementations. Its main specification documents are:
  - A. Sim, A. Shoshani (eds.), The Storage Resource Manager Interface Specication, v. 2.2, available at http://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.pdf.
  - F. Donno et al., Storage Element Model for SRM 2.2 and GLUE schema description, v3.5 available at: http://glueschema.forge.cnaf.infn.it/uploads/Spec/V13/SE-Model-3.5.pdf

Storage providers involvement: the available implementations

- SRM v2.2 implementations are available today for the following Storage Services:
  - ***CASTOR2*** : Hierarchical Storage Server (HSS). Developed by CERN and RAL. SRM v2.2 support in versions >= 2.1.4/1.2
  - ***dCache*** : HSS developed by DESY and FNAL. SRM 2.2 support in v1.8.
  - ***DPM*** : disk-only developed by CERN. SRM v2.2 support in versions >= 1.6.5 in production.
  - ***StoRM*** : disk-only developed by INFN and ICTP. SRM v2.2 interface for many filesystems: GPFS, Lustre, XFS and POSIX generic filesystem. SRM v2.2 support in versions >= 1.3.15.
  - ***BeStMan*** : disk-based developed by LBNL. SRM v2.2 support in v2.2.0.0.

The SRM concepts: storage classes

- In SRM v2.2 it is possible to select the quality of storage
- A storage class is a quality of storage defined by the Retention Policy and Access Latency
  - Retention Policy: Custodial or Replica
  - Access Latency: Nearline or Online
- The WLCG SRM v2.2 MoU defines 3 cases:
  - Custodial x Nearline → "Tape1Disk0"
  - Custodial x Online → "Tape1Disk1"
  - Replica x Online → "Tape0Disk1"
- TapeN → N copies guaranteed on tape
  - Or other high-quality media
    - Tape1/Custodial → "Do not lose this data!"
    - Tape0/Replica → "No disaster if this data is lost." (a custodial copy may be elsewhere)
- DiskM → M copies guaranteed on disk
  - Disk0 managed by system, Disk1 managed by VO

The SRM concepts: spaces and tokens

- In SRM v2.2 it is possible to reserve space of a given quality. The space reserved always refers to the space on *disk*.
- The WLCG SRM v2.2 MoU establishes that spaces can be reserved statically, even though dynamic space reservation can be supported by a storage system.
- A "space token description" is a tag that identifies a "chunk of space" with given characteristics (such as its storage class, size, protocols supported, etc.)
- Nothing is specified concerning permissions and allowed operations on spaces.
- The space token description is used whenever files are created. For read operations the token is not needed.

The SRM concepts: files and copies

- In SRM v2.2 files are permanent: only the user can remove them from the system
- The copy on disk on the file in Tape1Disk0 space can be temporarily removed by the system if space is needed.
  - Copies can be “pinned” to prevent the system from deleting them from disk while not in use
  - Copies can be “released” when no longer needed. The garbage collector can then delete the copy on disk in order to make space for other copies
- A file in Tape1Disk0 space for which a copy does not exist on disk can be pre-staged by the user from tape to disk before real access.

The SRM concepts: SURL and TURL

- A Site URL (SURL) allows a user to contact a Storage Service at a site asking for file access
  - srm://pcrd24.cern.ch:8443/srm/managerv2?SFN=/flatfiles/cms/output10
  - srm://pcrd24.cern.ch:8443/srm/managerv1?SFN=/flatfiles/cms/output10
  - srm://pcrd24.cern.ch:8443/flatfiles/cms/output10
  - **srm – control protocol for the storage service**
  - **Fully specified SURL**
- A Transport URL (TURL) is temporary locator of a replica accessible via a specified access protocol understood by the storage service
  - rfio://lxshare0209.cern.ch/data/alice/ntuples.dat
- A Site File Name (SFN) is the file location as understood by a local storage system
  - /castor/cern.ch/user/n/nobody/file?svcClass=custorpublic&castorVersion=2
- A Physical File Name (PFN) is the physical entry in the storage name space:
  - /data/atlas/raw/run29340.dat

The SRM concepts: File access protocols

- SRM v2.2 allows for the negotiation of the file access protocols
  - The application can contact the Storage Server asking for a list of possible file access protocols. The server responds providing the file handle for the supported protocol
- Supported file access protocols in WLCG are:
  - [gsi]dcap
  - Gsiftp
  - [gsi]rfio
  - file

The SRM Interface in a nutshell

- The SRM Interface Specification lists the ***service requests***, along with the ***data types*** for their arguments.

- Function signatures are given in an implementation-independent language and grouped by functionality:
  - ***Space management* functions** allow the client to reserve, release, and manage spaces, their types and lifetimes. Support for different qualities of storage space.
    - Reserve/Release Space, ChangeSpaceForFiles, ExtendFileLifeTimeInSpace, PurgeFilesFromSpace
  - ***Data transfer functions*** have the purpose of getting files into SRM spaces either from the client's space or from other remote storage systems on the Grid, and to retrieve them.
    - PrepareToPut/StatusOfPutRequest/PutDone
    - PrepareToGet/StatusOfGetRequest
    - BringOnline
    - Copy
    - ReleaseFiles, AbortRequest/Files, ExtendFileLifeTime
  - Other function classes are ***Directory*, *Permission*, and *Discovery* functions**.
    - Ping, Ls, Mkdir, Rm, Rmdir, SetPermission, CheckPermission, etc.

Basic use-cases

- Tier-0 data handling
- Tier-1 data handling
- Experiment data handling
- Reprocessing
- Recalling files from tape

Tier-0 data handling

- Central data recording into dedicated service class instances
  - Could be given space tokens if needed
- Operations in parallel:
  1. Write data to tape ASAP such that copy in experiment buffer can go
  2. Distribute data to Tier-1 centers
  3. Make data available to first-pass reconstruction
     - Output also to be written to tape and distributed to Tier-1 centers
- Step 2 may need data to be copied to other service class instance
  - (Better) connected to WAN
  - BringOnline / PrepareToGet would trigger disk-to-disk copy or tape recall
    - Note: the FTS currently expects the data to be available in 3 minutes!
- Step 3 may need data to be copied to yet another instance
  - Better matched to reconstruction program read patterns
  - Reduce interference with steps 2 and 3

Tier-1 data handling

- Receive data from Tier-0, other Tier-1 and Tier-2 sites
- Send data to other Tier-1 and Tier-2 sites
- CERN can act as Tier-1, any Tier-1 can act as Tier-2
- Tasks:
  1. Archive raw and reconstructed data received from Tier-0
  2. Make raw data available for second-pass reconstruction
     - Archive output and possibly copy it to other Tier-1 center(s)
  3. Regularly/occasionally recall data sets from tape for reprocessing
     - Archive output and possibly copy it to other Tier-1 center(s)
  4. Serve Tier-2 requests for data sets
  5. Archive Tier-2 Monte Carlo data and certain analysis results

Tier-1 data handling

- Disks receiving data from Tier-0 may not be suited for steps 1, 2, 3
  - Need d2d copies
- Data received from Tier-0 must remain pinned for step 2
  - Avoid tape recall
- Steps 2 and 3 need efficient staging of the necessary data
  - Reconstruction jobs should use batch system efficiently
- Large amount of disk space allows tapes to be fully read in one go
  - Allows many jobs to process data in parallel
  - Reduces need for merging small output files
- dCache honors “hard pinning”: a copy of the file will stay on disk for its entire lifetime, preventing concurrent activities if the disk become full.

Experiment data handling

- Small number of production managers
- Large number of unprivileged users, spread over physics groups
- Both categories desire a guaranteed Quality of Service
  - Dedicated disk spaces
  - Certain priorities in various request queues
- QoS handles available in CASTOR and/or dCache:
  - Space tokens
  - Name space
  - User identity/role
  - Client IP address, WAN/LAN flag
- Storage classes
  - Custodial-Nearline (T1D0) managed by system
    - Used for vast majority of the data
  - Custodial-Online (T1D1) managed by VO
  - Replica-Online (T0D1) ditto

Reprocessing

- Large T1D0 data sets will need to be reprocessed
  - Pre-stage from tape, controlled by production managers
  - Disk copies can be garbage-collected right after reprocessing
  - Large amount of disk space desired, ideally dedicated
    - Use BoL/PtG with desired pin time
- Alternative: temporarily change storage class to T1D1
  - T1D1 originally foreseen for data needed online for a long time, even when not used for a while
  - Not implemented

Recalling files form tape

- If on disk, files are served from the “pool” where they reside, if the pool is accessible. Otherwise an automatic disk-to-disk copy is triggered.
- Files cannot be recalled into an SRM v2.2 space
- Sufficient disk space has to be left unassigned to any space token
- Pool selections allowed through static configurations
    - Pools can be associated with paths

Executive summary for dCache

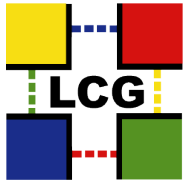
- Keep T1D0 spaces relatively small
  - Use them only as buffers for writing into the storage system
- Keep most of the T1D0 disk space unassigned to any space tokens
  - Can be used for restoring large data sets concurrently
- Possibly configure paths to allow for the selection of specific pools when recalling files from tape
  - Depending on name space layout per experiment

Future enhancements

- Protecting spaces
  - VOMS-enabled ACLs
  - Operation based (read,write,stage,query,modify-acl)
- Recalling files into specified spaces
  - Use of space tokens on Get/BringOnline/Copy operations
- Full support for T1D1->T1D0 transitions
  - srmPurgeFromSpace implementation

GLUE Schema

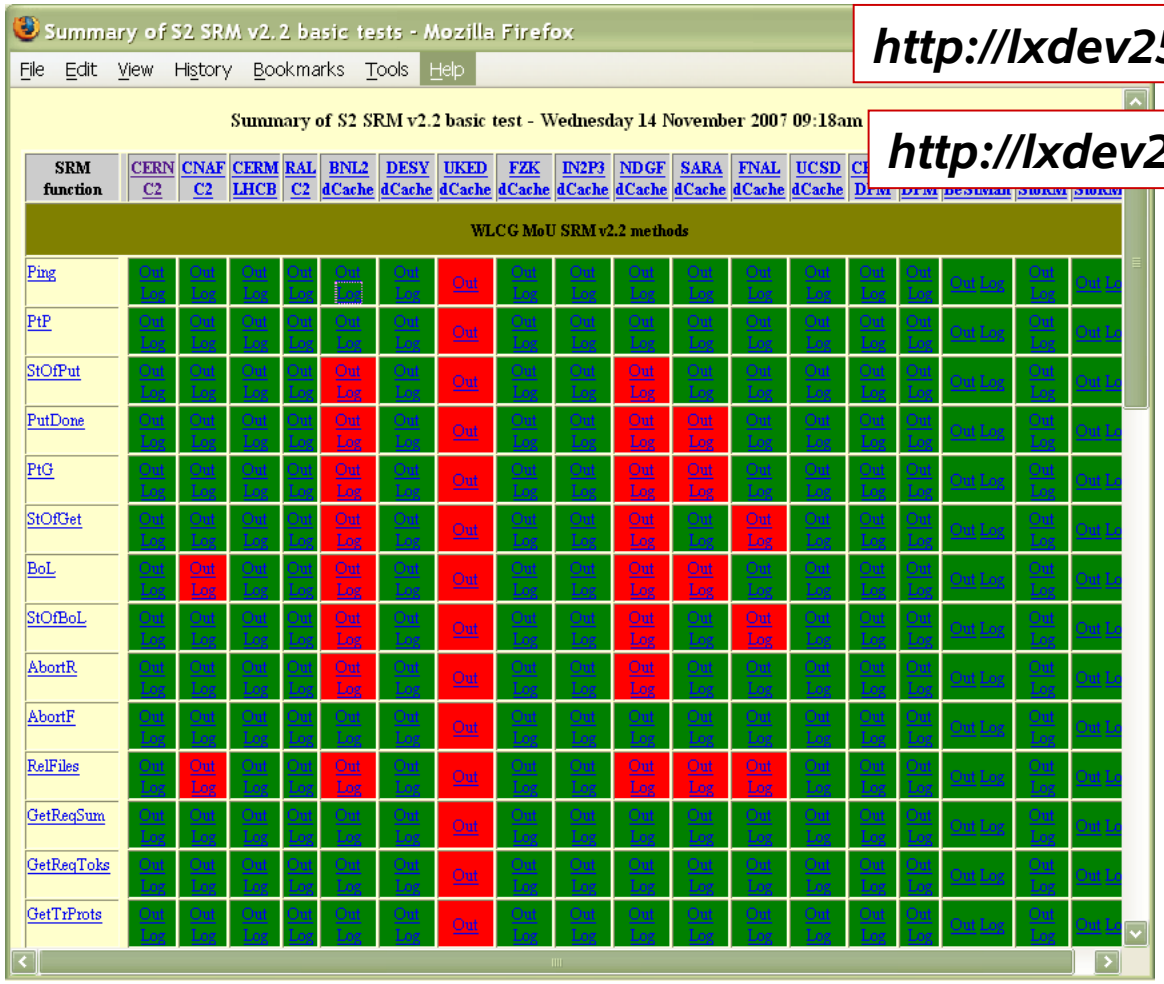
- The WLCG information system publishes details about grid services in a format that is specified by the GLUE schema.
- GLUE 1.3 includes information needed for SRM v2.2, such as the space information.
  - http://glueschema.forge.cnaf.infn.it/Spec/V13
- The current experience has also been taken as a base for the design of the new GLUE 2.0 model for Storage.



GLUE Schema

- dCache information providers can be found here (R. Trompert author):
  - http://trac.dcache.org/trac.cgi/wiki/contributed/SpaceTokenInformationProvider
- An example of a static ldif file can be found here:
  - https://twiki.cern.ch/twiki/bin/view/LCG/GSSDGLUEProposal

The S2 testing framework for SRM v2.2

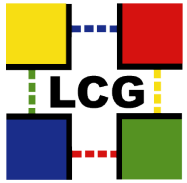


http://lxdev25.cern.ch/s2test/basic/s2_logs/

http://lxdev25.cern.ch/s2test/basic/history/

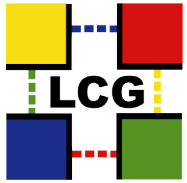
avail
usecase

This is not meant
to support the WLCG production infrastructure.
However, your endpoint can be tested on demand.
It will be soon part of SAM



WLCG Storage Coordination Bodies

- Grid Storage System Deployment (GSSD) Working Group
  - Work terminated at the end of 2007. Mailing list and twiki still in use
  - Help sites and experiments with the deployment of SRM v2.2
  - https://twiki.cern.ch/twiki/bin/view/LCG/GSSD
  - Mailing list: storage-class-wg@cern.ch
- Storage Solution working group (SSWG)
  - Provide a forum where storage issues are discussed and solved
  - Finalize the Addendum to the WLCG SRM v2.2 Usage Agreement
  - Mailing list: wlcg-ccrc08-storage-solutions@cern.ch
  - Indico: http://indico.cern.ch/categoryDisplay.py?categId=1613
- dCache deployment mailing list and weekly phone-confs with Tier-1s sites
  - srm-deployment@dcache.org



Summary

- The SRM specification definition and implementation process has evolved in a ***world-wide collaboration effort*** with developers, independent testers, experiments and site administrators.
- SRM v2.2 based services are now production ready. SRM v2.2 is the storage interface mostly used in WLCG.
- Many of the SRM v2.2 needed functionalities are in place. Further development is needed for meeting the requirements.
- The Information System schema supports SRM v2.2. The available information providers will become part of the standard EGEE gLite distribution.
- The S2 testing framework provides a powerful validation tool. It will be soon part of SAM.
- Storage coordination and support bodies have been setup to help experiments and sites.

Thank you!

Questions ?