egee

Enabling Grids for E-sciencE

Storage Classes
Summary of the last meeting

Flavia Donno

Pre-GDB - CERN, 7 November 2006

Enabling Grids for E-sciencE

- **Storage Classes**

- **Storage Class Transitions**

- **WLCG Storage Element Model**

- **Miscellaneous items**

- A **Storage Class** *determines the properties that a storage system needs to provide in order to store data.*

- The LHC experiments have asked for the availability of combinations of the following storage devices: **Tapes** (or reliable storage system always referred to as tape) and **Disks**.

- TapeN and DiskN:
  - If a file resides on Tape then we say that the file is in Storage Class Tape1.
  - If a file resides on an experiment-managed disk, we say that the file is in Storage Class Disk1.
  - Tape0 means that the file does not have a copy stored on a reliable storage system.
  - Disk0 means that the disk where the copy of the file resides is managed by the system: if such a copy is not pinned or it is not being used, the system can delete it.

Enabling Grids for E-sciencE

- Supported Storage Classes *for next implementation of SRM v2.2 are:*
    - Custodial-Nearline: this is the so-called Tape1Disk0 class.
    - Custodial-Online: this is the so-called Tape1Disk1 class
    - Replica-Online: this is the so-called Tape0Disk1 class

    - Tape0Disk0 is not implemented. It is pure scratch space that could be emulated using one of the available classes and removing the data explicitly once done. However, it could be handy for LHC VOs to have such a type of space actually implemented.

Enabling Grids for E-sciencE

- *Custodial-Nearline*: data is stored on tape. Access to data may imply certain latency. When a user accesses a file, the file is recalled in a cache that is managed by the system (Disk0). The file can be "*pinned*" for the time the application needs the file. However, the treatment of a pinned file on a system-managed disk is implementation dependent, some implementations choosing to honor pins and preventing additional requests, others removing unused on-line copies of files to make space for new requests.

- *Custodial-Online*: data is always available on disk. A copy of the data resides permanently on tape. The space owner (the virtual organization) manages the space available on disk. If no space is available in the disk area for a new file, the file creation operation fails. This storage class guarantees that a file is never removed by the system.

- *Replica-Online:* it is implemented through the use of disk-based solutions not necessarily of high quality. The data resides on disk space managed by the virtual organization.

Enabling Grids for E-sciencE

- Through the SRM call srmChangeSpaceForFiles it is possible to schedule Storage Class Transitions for a list of files.
  - **Tape1Disk1 -> Tape1Disk0**.
  - **Tape1Disk0 -> Tape1Disk1**. This transition would be implemented with some restrictions: the srmChangeSpaceForFiles call will complete successfully but the files will remain on tape. The files will be actually recalled from tape to disk only after a BringOnline operation is executed. This is done in order to avoid that a big set of files is unnecessarily scheduled for staging and therefore to smoothen operations in particular for those Mass Storage Systems that do not have a scheduler (namely TSM).
  - **Tape0<->Tape1** transitions are **not supported** at the start of LHC (if ever). For physics validation operations, since the amount of data to transfer to tape after the validation is not big (only 1-2% of total data) a change class operation from Tape0Disk1 to Tape1DiskN can be approximated by copying the files to another part of the space, specifying Tape1DiskN as the new storage class, and then removing the original entries.

Enabling Grids for E-sciencE

- A **Storage Element (SE)** *is an aggregate of Grid services that allows Grid users to store and manage files together with the space assigned to them.*

- The **SE Implementation** is the software system used to manage the storage devices and servers. Examples of this are: CASTOR, dCache, DPM, StoRM, etc.

- An SE exposes **Total Sizes**: an **Online** for space on disks and **Nearline** for space on tape or slow devices.

- An SE can have multiple **Storage Areas**.

- A **Storage Area (SA)** *is a view on a portion of the total space:*
  - *It is created by the System or VO Administrators*
  - *It can span different kinds of storage devices within a Storage Element*
  - *It exposes a single retention policy and a single access latency (which the underlying storage devices together can support)*
  - *In case of WLCG it implements a Storage Class instance.*

- An **SA** may be dedicated or shared between certain Vos/groups/roles
  - **For WLCG the default SA is typically shared** (it is Tape1Disk0 for CASTOR at CERN)

- For WLCG the SA implements a Storage Class instance:
  - **It is identified by a Space Description (that can differ per VO).**
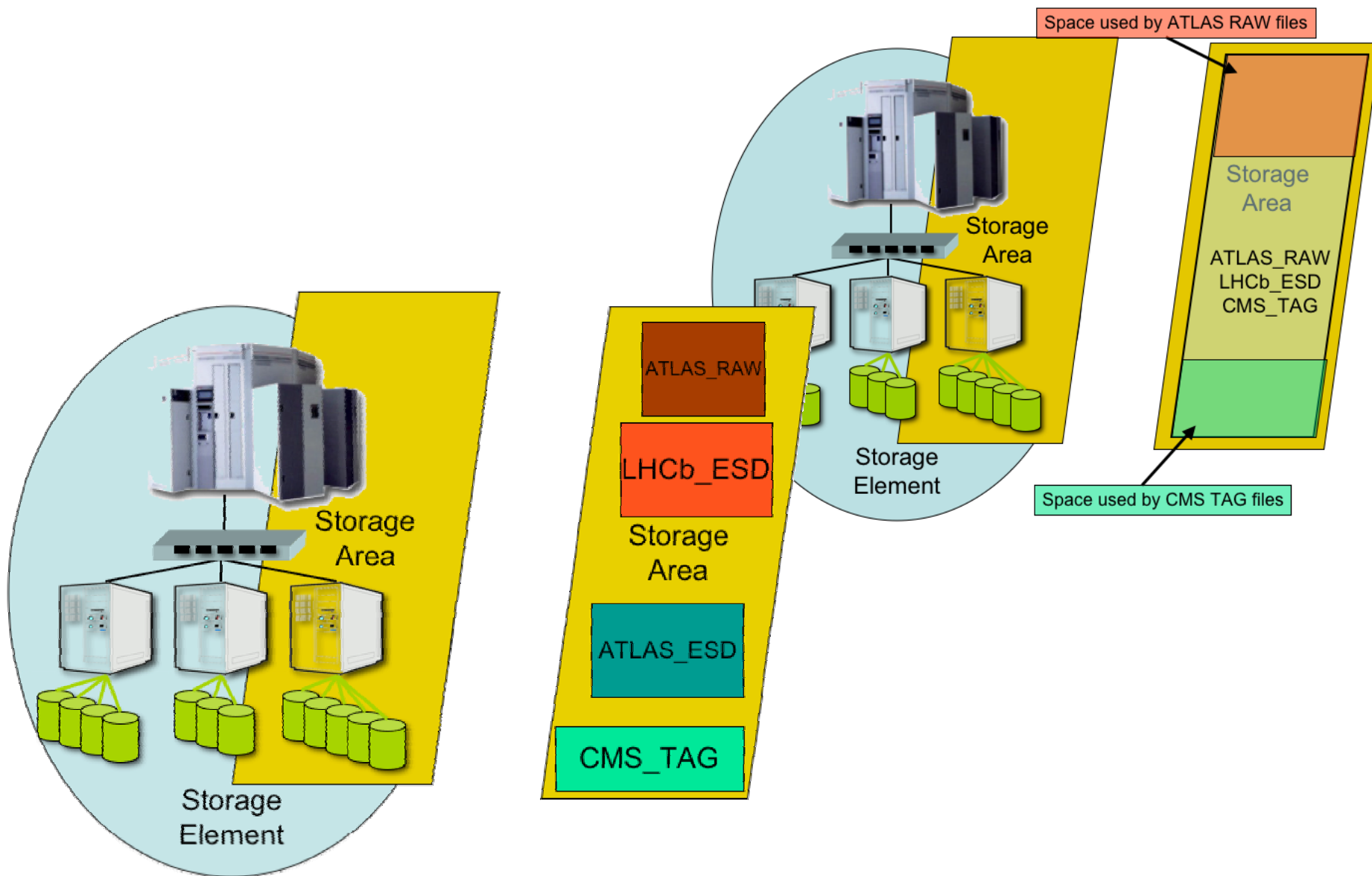
- Clients should *not use dynamic reservation* initially.
- *Storage Areas are created statically* by Site Administrators at sites. VO administrators can use srmReserveSpace to create Storage Areas.
- *Tools* such as the ones used *to publish* Experiment Software Tags will be available to publish *VO specific Space Descriptions* (ATLAS_RAW, CMS_TAG, etc.)
- *Namespaces* are meaningful for certain implementations such as dCache and StoRM.
    - dCache uses paths to specify dCache Storage Groups (SGs identifies tape sets or other space characteristics for VOs).
    - In dCache, experiments will be able to change Storage Classes without having to change the path for a given set of files.

- WLCG SRM v2.2 MoU:
  https://srm.fnal.gov/twiki/bin/view/WorkshopsAndConferences/GridStorageInterfacesWSAgenda

- WLCG Storage Element model for SRM v2.2 and Glue Schema Description: http://glueschema.forge.cnaf.infn.it/Spec/V13

- Minutes of the last pre-GDB Storage Class meeting:
  http://indico.cern.ch/materialDisplay.py?materialId=0&confId=a058490