Storage Classes requirements for LHCb

Input provided by **Philippe Charpentier** and **Nick Brook**.

1. **Storage Classes needed at various sites**
   - For Tier0 and Tier1: **Tape1Disk0**, **Tape1Disk1** and **Tape0Disk1**
   - For Tier2: no special storage requirements. Only simulation production running at Tier2, output is stored on the worker nodes and transferred. For Tier2's large enough to run analysis: **Tape0Disk1**. The size would be discussed locally by LHCb physicists and selected datasets would be replicated there.
2. **Data flow between Tier0, Tier1, Tier2 and transfer rates**
   - Mainly available in presentations by N. Brook at pre-GDB Storage Classes meetings in October and December
3. **Space reservation requirement**
   - Static space reservation.
   - It would be great to have the possibility however to be able to know how much space is left free (e.g. before bulk transfers).
4. **Space token descriptions per VO**
   - Raw data: Tape1Disk0 (at Tier0 and Tier1s) - **LHCb_RAW**
   - Reconstructed data (RDST): Tape1Disk0 at reconstruction Tier0/1; the space could be shared with RAW - **LHCb_RDST**
   - Stripped data and MC data (DST): Tape1Disk1 at production site (or closest Tier1 for MC), Tape0Disk1 at all other Tier1s (one other Tier1 for MC) - **LHCb_M-DST** (M for Master) Tape1Disk1 ; **LHCb_DST** Tape0Disk1; **LHCb_MC_M-DST** Tape1Disk1 (even if shared with M-DST); **LHCb_MC_DST** Tape0Disk1 (even if shared with DST).
   - User files, calibration files etc...: Tape1Disk1 (no replication, most probably each user using mainly a single SE for convenience?). Note: files might be small, hence not necessarily convenient for e.g. Castor. These are micro-DST, Ntuples, private format files, temporary alignment DB (SQLite files) etc... - **LHCb_USER**
5. **Special requirements**
   - If xrootd demonstrates it can be "the" solution for generic file access, replacing gfal (option to be seriously considered), of course LHCb would be interested on it replacing existing access protocols.
6. **Data access patterns**
   - LHCb RAW:
     - Written (from the DAQ at Tier0, from Tier0 at Tier1) in (almost) real time with data taking. Access from both WAN (for distribution) and LAN (for reconstruction) . Question: should there be distinct pools depending on the access with automatic disk-to-disk copy?
     - Files pinned on disk-cache for a few days, allowing reconstruction to take place.
     - Files processed within a few days and unpinned by the reconstruction job (still unclear: when to unpin files at Tier0 that are reconstructed at Tier1s?)
     - For Re-reconstruction: files staged from tape before launching reconstruction jobs (not clear yet how to synchronize job s with staging), pinned and unpinned by reconstruction jobs
   - LHCb RDST:
     - Written by the reconstruction job, pinned on disk for further stripping (unpinned by stripping job)
     - For Re-stripping: same procedure as for Re-reconstruction
   - LHCb (MC) M-DST:
     - Written by the stripping job at local Tier1
     - WAN access for distribution to other Tier1s
     - Frequent and chaotic local access by analysis jobs
   - LHCb (MC) DST:

        - ◊ Distributed over WAN from M-DST
        - ◊ Frequent and chaotic local access by anlysis jobs
    - LHCb USER:
        - ◊ Fully chaotic usage
        - ◊ Files written by analysis jobs running at other Tier1s (over WAN), presumably using the public network (primary storage)
        - ◊ Files might be small, frequently accessed locally even from non-grid nodes (e.g. copy to desktop/laptop)
7. **Plans from 1st April 2007 till end of the year**
    - Analysis of "DC06" stripped data (for the LHCb physics book). Using "LHCb MC (M-)DST" at Tier1s
    - Alignment/Calibration challenge (at Tier0?): production of mis-aligned data (small sample), running alignment jobs, feeding into the Conditions-DB, streaming at Tier1s, reconstruction of control samples at Tier1s. All this doesn't involve large datasets)
    - Computing Model exercise (so-called "dressed rehearsal"): repeat the DC06 computing exercise: ship data from CERN at nominal rate (80 MB/s), reconstruct and stored RDST, strip and distribute DST * (not fully discussed yet): new simulation round using measured detector position, reconstruction and analysis, in order to be prepared for data (possibly same data at 900 **GeV** ?).
8. **How much disk should Tier-1s and Tier-2s provide ?** Derived from Megatable.
    - What LHCb has always quoted is their estimate of disk needed **_excluding_** the disk-cache in front of the MSS. All sites have not interpreted this the same way, e.g. CERN considers this includes the cache while PIC doesn't. Anyway it was estimated at the time that the cache was almost in the background of the rest, as we probably need of the order of 20-40 TB at each site. The Megatable also makes no assumption on whether separate disks are needed for import/export (WAN pools). Total for 2008 (from Nick's slide 8): ~1400 TB.
9. **Of that amount of disk, how much must be set up for each of the storage classes T1D0, T1D1 and T0D1?**
    - This can be extracted from Nick's slides at the Dec pre-GDB (slide 8) with a bit of gymnastics. Note: the "RAW" and "rDST" on disk in this table do correspond to a small fraction of data that we want to keep on disk, not the cache. It is not yet clear how LHCb handles that data. A possibility would be to have a real replication of part of the LHCb_RAW on a storage class Tape0Disk1, and not a fraction of the RAW being on a Tape1Disk1. The result is the same but the handling is different and there is no need for class migration. This would add another Tape0Disk1 storage class: **LHCb_BUFFER** (Tape0Disk1).
    - **LHCb_BUFFER** : 120 TB
    - **LHCb_M-DST** (1/7 of the "Stripped" data) : 110 TB
    - **LHCb_DST** : 665 TB
    - **LHCb_MC_M-DST** (1/2 of the Simulation on disk): 190 TB
    - **LHCb_MC_DST** : 190 TB
    - **LHCb_USER** : 115 TB
    - Note that the amount of disk for Tape1Disk0 is not included (see caveat at the beginning). At CERN for example it very much depends on the strategy for CDR, data shipping to Tier1s and local reconstruction (one pool, two pools...). Order of 10-15% of the total?
    - A prelimianry breakdown for "real data" of the 2008 dataflows at all Tier-1 centres & CERN can be found in Dataflows.pdf (corrected 28th June 2007)
10. **What transitions exist between those instances, i.e. how much data will be moved between which instances?**
    - Currently the only move foreseen is Tape1Disk1 -> Tape1Disk0 (for "archiving" datasets of DSTs when they become ~obsolete)
11. **What network access is needed per instance?**
    - This is complex and concerns mainly the disk-cache buffers. LHCb expects the "permanent" disk storage to be highly optimised for local access from the WNs. This means that if they have to be exported, a disk-to-disk copy would be needed, as well as when they are imported.

However if sites want to implement them as WAN pools, they should be careful that this doesn't decrease the performance for LAN access (this is how it is set up currently at CERN).

12. **Use of namespace to group files**
    - The LHCb data namespace can be found here

-- Flavia Donno - 18 Jan 2007

This topic: LCG > GSSDLHCB
Topic revision: r13 - 2008-10-01 - AndrewCSmith