Storage Classes
Input from the experiments

18th January 2007

Flavia Donno, Maarten Litmaath
IT/GD, CERN

Overview

- What input are the experiments asked to provide
- ATLAS
- LHCb
- Summary

Input requested per VO

- See presentation given at WLCG GDB
  http://indico.cern.ch/getFile.py/access?contribId=8&resId=1&materialId=slides&confId=8468

  - Storage classes needed at Tier-1s and Tier-2s
  - Data flows between Tier-0,Tier-1s,Tier-2s
  - Static or dynamic space reservation ?
  - Which space token descriptions per storage class ?
  - How to divide the available disk space over those spaces ?
  - Transitions (data flows) between spaces?
  - Data access patterns
  - Network connectivity per space
  - Special requirements: xrootd, …
  - Plans from 1st of April 2007 till the end of the year.

Input requested

- Furthermore, what about Tier-2s ?
  - Is data loss an issue at Tier-2s ?
  - How many file open/sec, read-write/sec, etc. ?
  - Is it important to publish the real size of a space (available vs. used) ?
  - Which storage classes ? Are Tier-2s required to make available reliable space (CUSTODIAL) ?
  - Dynamic reservation ?
  - Do transfers happen between Tier-2s ? Interoperability tests?
  - What to do in case of unused/corrupted data sets? What about full disk pools ? Empty pools ?

ATLAS

- Storage classes needed at T1
  - T1D0
  - T0D1
  - T1D1
    - Used for reprocessing data
    - For now can be emulated by T1D0 + srmBringOnline
- Data flows per site given by mega table
  - Site to be able to buffer at least 2 days of data taking
  - Plus at least as much for reprocessing
- Space reservation static for now

ATLAS space token descriptions

- Are space tokens related to the way files are migrated to tape ?
  - Might consider ATLAS_RAW to ensure all raw is put together on tape
- Splitting per data type complicates FTS handling

- ATLAS_PROD_ONLINE (or _DISK)
- ATLAS_PROD_ARCHIVE (or _TAPE)
- ATLAS_PROD_REPROCESS ??
  - Should have a larger disk buffer

- ATLAS_USERS

ATLAS

- Data access for processing only from local site
- Currently each file copied to WN
  - Does not scale
- Analysis job will access 100-1000 files via POSIX-like I/O
  - rfio/dcap/GFAL/xrootd under study
    - ROOT/POOL version compatibility issues
  - Typical read rate per job 2 MB/s

ATLAS plans until end 2007

- 3 major commissioning activities before summer
  - Tier-0 internal tests
  - T0-T1-T2 data distribution
  - Calibration data challenge
- Continuous simulation production
  - Increasing up to 8 M events/week in Dec.
- July-October
  - Integration test for Final Dress Rehearsal
  - Ready for data taking by end of October

LHCb

- **See presentations given during the Storage Classes Working Group meetings:**
  - http://indico.cern.ch/getFile.py/access?contribId=s1t0&resId=5&materialId=0&confId=a058490
  - http://indico.cern.ch/getFile.py/access?contribId=1&resId=0&materialId=slides&confId=a058492
- **Storage Classes needed**
  - Tier-0 and Tier-1s: Tape1Disk0, Tape1Disk1, Tape0Disk1
  - Tier-2: Tape0Disk1
- **Data Flow between Tier-0, Tier-1s and Tier-2s**
  - Numbers given by 2nd presentation
- **Space reservation**
  - Static
  - It is needed to know how much space is free before bulk transfers.

LHCb

- **Space Token Description**
  - Not fully discussed in LHCb
  - Raw data at Tier-1s and Tier-0 on Tape1Disk0 - **LHCb_RAW**
  - Reconstructed data (RDST): Tape1Disk0 at reconstruction Tier0/1 - **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** and **LHCb_DST**, **LHCb_MC_M-DST** and **LHCb_MC_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**
- **Special Requirements**
  - Seriously interested in the xrootd tests as possible replacement for GFAL.

LHCb

- **Data Access Pattern**
  - 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) . Q: 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

- **Data Access Pattern**
  - 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 analysis 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)

LHCb

- **Plans from 1st of April 2007 till the 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 misaligned 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?).

Summary

- ATLAS and LHCb have started identifying the properties of the storage class instances (i.e. spaces) they need
  - Some numbers, access patterns etc. already available
- ALICE and CMS to follow
- We need more details, per VO, per site
  - How to split the available disk over the various spaces ?
  - Decide per space which network connectivity is needed
  - What are the expected I/O rates per space ?
- If we have those numbers for a few example T1 and T2, the other sites can copy and scale the recipe, per VO