ATLAS T1/T2 Name Space Issue with Federated Storage

Hironori Ito

Brookhaven National Laboratory

Global name space

- ## Federated storage
  - ### Universal access to all available storages
    - #### Protocol
      - Xrootd, http, etc…
    - #### Name space
      - File name itself should uniquely determine the location of the file in the storage system.

GUID, physical file path in storage and logical file path in LFC

- GUID of a file is the unique identifier.
  - Unique key in LFC, eg.   lcg-la guid:XYZ
  - It maps to a particular path in LFC name space.
    - /grid/atlas/dq2/mc10_7TeV/NTUP_TOP/e737_s933_s946_r2215_r2260_p542/mc10_7TeV.107670.AlpgenJimmyZtautauNp0_pt20.merge.NTUP_TOP.e737_s933_s946_r2215_r2260_p542_tid326436_00/NTUP_TOP.326436._000527.root.1

- One guid/lfc path can have multiple physical files.
  - In the different SEs
  - In the same SE in different physical paths

Physical file path

- ATLAS physical file
  - No Global name space of physical files
    - Various space tokens in a single SRM end point
      - Identical files could be located at various SURL paths
    - Use of _sub by Panda system
    - Use of DQ2_XYZ extensions
      » Caused by the lack of "re-name" in SRM protocol.
        - SRM should do this internally!
          - Essential for the tape area since the external rename is a bit tricky
      » Without its use, due to asynchronous nature of write and delete, legitimate files could be deleted after the successful transfers (particularly after a few failed transfers)
        - Really painful to clean.

Logical file path

- ATLAS logical file
  - Logical file path in LFC is generally universal
    - No _DQ2_XZY extension unlike physical file
    - No storage dependent prefix or suffix
      - Does not depend on the space tokens.
      - Does not depend on the storage service name
        - » srm/managerv2
        - » srm/v2/server
      - Almost deterministic based on the dataset name.
  - It is the closest thing to be the global name space
    - Still has _sub directories at T1s
    - Still has some site dependence on the site configuration
      - Eg
        - » User datasets locations are defined in PANDA schedule configuration, which are different for each sites.

Not quite unique LFC path

- Example of dataset name and file
  - DSN: mc10_7TeV.107670.AlpgenJimmyZtautauNp0_pt20.merge.NTUP_TOP.e737_s933_s946_r2215_r2260_p542_tid326436_00
  - File name NTUP_TOP.326436._000527.root.1
  - Should be located in LFC as /grid/atlas/dq2/mc10_7TeV/NTUP_TOP/e737_s933_s946_r2215_r2260_p542/NTUP_TOP.326436._000527.root.1
  - Could be located at one of many /grid/atlas/dq2/mc10_7TeV/NTUP_TOP/e737_s933_s946_r2215_r2260_p542/mc10_7TeV.107670.AlpgenJimmyZtautauNp0_pt20.merge.NTUP_TOP.e737_s933_s946_r2215_r2260_p542_tid326436_00_subXYZ/NTUP_TOP.326436._000527.root.1
    - T1s only!!! For files produced at T2s, "_sub" are lost by DDM transfer.
  - ATLAS convention is not quite concrete
    - Forced by SE limitation
      - Ext3 file system: 64K sub directories.
      - HPSS file system: 64K sub directories.

Name space convention changing

- The convention has changed over the years.
  - Force by the storage file system limitation
    - 64K subdirectories limit in ext3
    - 64K subdirectories limit HPSS tape systems
  - Eg
    - DSN/file: data11_7TeV.00180309.physics_Egamma.merge.NTUP_TOP.f369_m812_p530_p577_tid367204_00/ NTUP_TOP.367204._000043.root.1
      - Current convention:
        » /atlas/dq2/data11_7TeV/NTUP_TOP/f369_m812_p530_p577/data11_7TeV.00180309.physics_Egamma.merge.NTUP_TOP.f369_m812_p530_p577_tid367204_00/ NTUP_TOP.367204._000043.root.1
      - Actual path
        » /atlas/dq2/data11_7TeV/NTUP_TOP/data11_7TeV.00180309.physics_Egamma.merge.NTUP_TOP.f369_m812_p530_p577_tid367204_00_sub021131151/ NTUP_TOP.367204._000043.roo.1

Different LFC path at different site

- Caused by the difference in the site configuration
  - Lfcpath=/grid/atlas/dq2
  - Lfcpath=/grid/atlas/users/pathena
    - Eg. File User.ABC.DDD.1 could be located at
      - /grid/atlas/dq2/user/ABC/DDD/user.ABC.DDD.1
      - /grid/atlas/users/pathena/user/ABC/DDD/user.ABC.DDD.1

Name-to-name module

- xRootd allows the use of external module to translate the externally requested path to the actual path in the storage
  - Name-to-name module

External request /a/b/c/d.1	Name-to-name →	Actual file /V/W/X/Y/Z.1

Current name-to-name module

- Current implementation of NtoN module by Charles
  - Search files in LFC name space
  - Leading /grid is removed from LFC path.
    - /grid/atlas/A/B/C -> /atlas/A/B/C
    - In US LFCs, there is a symlink in LFC. /atlas is a simlink for /grid/atlas
      - So, it is equivalent.
  - The request is expected to be correctly formatted in the current ATLAS convention.
    - /atlas/dq2/proj/..../filename
  - It can search parent _subXYZ directories (for T1s)
    - Eg. /atlas/dq2/A/B/C.1 -> /atlas/dq2/A/B_subXYZ/C.1

Deficiency of current NtoN

- Not 100% of files can be found.
  - Files are stored with the different conventions won't be found.
  - Search on _subXYZs might be problematic???
  - Inherent problem of assuming that the file path in LFC is unique.
    - Only GUID is unique.
      - If guid was the filename in the flat filename structure, we won't be having this discussion.
        » Can't be possible by the SE's limitation

Improving NtoN

- Modify the possible search path to accommodate the changing conventions and/or configuration

- Add correct symlink

- Use GUID:
  - DQ2 has 100% hit rate because it uses GUID to find a file.
    - Eg. To find /A/B/C/D.1, NtoN might accept /A/B/C/D.1+guid=XYZ

dCache xRootd door with NtoN

- dCache has the own, native xRootd service.
  - Can access files via xrootd protocol.
- Name-to-name is not quite modular or flexible.
  - Prefix can be changed with fixed path.
  - ATLAS needs dynamic changes of path.
    - Needs code change->doable.