DDM development

Architecture and developers workplan

Distributed Data Management

Miguel Branco
on behalf of the DDM group

Outline

- 0.2 status
- Development
  - Recent changes
- A few words on architecture

0.2 status

- Current version is 0.2.12-5
- Recent changes:
  - 0.2 production catalogs near ‘meltdown’ during XMAS holidays
  - For 2 reasons:
    - number of files in a dataset had grown past the default POOL File Catalog interface container
      - therefore multiple queries (LIMIT N,M) were being sent per query to the server (bringing DB near collapse)
      - found and fixed during holidays
    - .. but still found an enormous amount of ‘queryFilesInDataset’ queries’
      - often multiple per second (some for 20K datasets)
      - we assumed it was site services queries
      - but no, it was due to **BROKEN** implementation of getNumberOfFiles in a dataset !!!

0.2 status

- 0.2.12 in use on LCG, still to be deployed on OSG (AFAIK)
  - Mixed success
    - When it works, it works
      - (and seems to work on maintained installations)
    - When it doesn't, it really doesn't
      - (and rarely says why…)
- Why then?
  - Still a fair amount of confusion using subscription options (HOLD)
  - Proxy certificates expired ?! (more later)
  - Some HOLD conditions not recoverable automatically
  - One known bug in the internal multi-process handling (thanks to Hiro for helping debugging - only reported in BNL though…)
    - racing condition: O/S spawns process but only gives it CPU much time later; DQ2 site services assume process died and attempt to start more; after a while all processes (O(10) start serving the same request…)

0.2 deployment

- 0.2.12 easier to install on LCG
  - but no or little site configuration has been done
    - not enough documentation?
    - not enough knowledge of configuration options?
- One example is proxy certificates:
  - We stated the need for a reserved DN (mine so far) for Tier-0->Tier-1 transfers
    - this is the only requirement and this proxy is only used for Tier-0 transfers during SC4 exercises
  - The site services require anyone's proxy to run (must be production role)
    - We assumed site contacts would use their own proxy
      - Installations on DDMSiteInstallation wiki for FTS-specifics
        - » **Using myproxy-* or vobox-* commands?**
    - If Tier-0 transfer, DQ2 overrides 'default' proxy with the Tier-0 one
      - just for Tier-0 transfers!!

Recent changes

- From last s/w week:
  - ARDA joined the DDM project
    - with responsibilities on helping the deployment and communication with Grid m/w providers
    - but also effectively joined the development team
- Resumed regular “DDM development” meetings:
  - Wednesdays at 16h (Geneva)
    - Meeting rooms bld 40 + Phone
  - Will send announcements to
    - atlas-dq2-dev@cern.ch
  - Agendas under GT&S category on Indico:
    - http://indico.cern.ch/categoryDisplay.py?categId=649

Recent changes

- 0.3 preparations
  - Agreed on 0.2 to 0.3 migration policy
    - From 0.2 MySQL directly to 0.3 ORACLE
      - inline with discussions during DDM review
    - Migration will be done incrementally:
      - data in 0.2 will be incrementally added onto the 0.3 catalogs, making both instances consistent
    - While more complex migration procedure, ensures softer migration, more 0.3 testing
      - and real data on 0.3 from the early days
    - When confident with 0.3, we can open access to catalogs and these will already be “preloaded” with all 0.2 datasets
      - until that time, 0.3 access is restricted for testers

Recent changes

- Deployment:
  - Still the primary issue from the developers PoV
    - too slow deploying bug fixes and/or optimizations to sites
      - common Grid m/w problem
      - c.f deployment of 0.2.12
    - no real validation step in between except unit testing
  - Therefore decided for automatic build system and real testbed at CERN:
    - nightlies, stable, unstable release
    - generating tarballs / RPMs
      - default is set to be APT repository
        - » “apt-get update dq2-site-services”
    - Build system is a contribution from the ARDA project
  - Confident will improve greatly DQ2 deployment

Recent changes

- Service management at the sites:
  - Evolving from the ‘cron job’ based solution to managed site services
    - single start/stop command for DQ2 site services
    - remote monitoring of service status
    - (yet another contribution from ARDA)

Recent changes

- Client interfaces:
  - splitting single ./dq2 into multiple tools
    - so that different user communities have different set of tools
      - introduction of dq2-admin-* tools
    - (considering including DASHBOARD client interface to monitor directly subscriptions from command line)

Recent changes

- 0.3 catalogs
  - Expensive queries have changed
    - queryFilesInDataset
      - Now can query per modification dates
        - » Important optimization for site services and when querying for open datasets
  - Open datasets
    - Our initial design assumptions for DQ2 assumed a majority of closed datasets
      - In fact, ~50% are open
    - Adapted the queries to reflect this reality

Still need to do

- Real testbed
  - machines ordered (ARDA), waiting for delivery
- Integrating newer callbacks to ARDA monitoring framework
  - site services “heartbeat”
  - cancellation notification
  - fetching new subscriptions
  - …
- Still a bit of cleaning to do on 0.3 agents

Still need to do

- External clients
  - Need to migrate to new 0.3 API
    - For those using DQ2.py (python API) migration will likely be transparent
      - … new methods added
    - For those using HTTP directly, expect changes
  - Client/server HTTP protocol is now easily extensible
    - If you do not have a command that does your particular datasets-based query, let us know and we can probably add it easily!

0.3

- 0.3 is now under testing on the CERN ORACLE Integration
  - ORACLE Production service requested
- As soon as testbed machines arrive we start larger scale testing
  - Aiming to have a stable version mid-Feb
- Will run Tier-0 exercise (export Tier-0 to Tier-1s) with 0.3 catalogs and site services
  - likely move into production only after Tier-0 tests
    - and after external clients migrated to new HTTP API
  - do sites need to deploy an extra VO BOX and install 0.3?
    - No; we have ordered a set of “reserved” VO BOXes at CERN to be used for this exercise (part of the ‘testbed’)
    - … but having a stable 0.3, we can be more aggressive and have initial deployment in parallel with 0.2 at some sites

Workplan for ~2 months

- 0.2 -> 0.3
- Improvements to subscription performance
- Monitoring
- “SC4” Tier-0 exercises
- New location catalogue
  - Requirements gathering and design

Other issues

- SRM v2.2
  - Space reservation
- LFC
- FTS
- Security Model
- Federated sites
- End-user tools

SRM v2.2

- This week we learned SRM v2.2 may be usable by the experiments ~April
  - We should aim at having a handful of Tier-1s with SRM v2.2 front-ends in re-run of “Tier-0” by .. May?
- SRM v2.2 brings static space reservation, access latency, retention policy
  - Experiments can allocate space on disk, ..

quasi-non SRM v2.2 terminology

- Tape0, Disk1
  - data on disk only, garbage collection managed by the experiment
- Tape1, Disk1
  - data on disk with copy on tape
    - behaviour here is less clear to me and I assume it may slightly vary between dCache, CASTOR (I may very well be wrong) - *e.g. release file from disk means it is automatically recalled from tape?*
- Tape1, Disk0
  - data is put on disk “front-end” and migrated to tape by the storage

SRM v2.2 usage

- Data taking:
  - Understood: place data on T1D0, T0D1
    - similar to today's XXDISK, XXTAPE sites on DQ2
  - May (or not) pin files on T0D1 while export to Tier-2s is going on
    - depends on how large disk-cache is (we may not need to bother)
- Reprocessing:
  - When data to be reprocessed is on T0D1
    - **At the moment we expect to use:**
      - srmBringOnline (soft pinning, pin may not be enforced by all backends)
      - and srmReleaseFiles (to clean files from disk when done)
    - This needs to be tested
- *There may be other ways to do reprocessing (explore T1D1? ask to change space from T1D0 to T0D1? - not requested)*
- Discussions on how to use SRM v2.2 and how to help sites dimension storages (import vs export buffer, read vs write caches) are about to start

LFC

- New LFC server will be available shortly
  - bulk GUID lookups, optimized client/server protocol
    - *Please deploy this ASAP at your site*
    - We will need this LFC version during Tier-0 rerun

FTS

- Changes to submit options requested to FTS
  - Do [not] overwrite file on destination
  - Do [not] recall file if on tape
- Notification service being discussed
  - avoiding need for constant polling by site services
    - (generating increased load, irregular responses, occasional hangs, slower)

Security Model

- Baseline plan is:
  - have DQ2 place all data at the storages under /dq2/
  - protect this area so that it is only by ProdSys and DQ2
- Entries in this area should match exactly the corresponding catalog namespace entry
- Rest of storage is available to users:
  - while DQ2 may read from it, it will never write to it
- DQ2 owns all data under /dq2/
  - All data is readable by the ATLAS VO
  - No user can delete data except DQ2 (+production)
    - That is to guarantee data is never deleted “accidentally” by one user, affecting someone else

Federated sites

- A single replica catalog should exist per ‘federation’
- No problem reading data from federated sites
  - We read its catalog entries
- Still unsure how to proceed with writing:
  - Ideally single entry point
  - Minimum requirement is different entry points per ‘dataset allocation’
    - e.g. RAW goes to FEDERATED1_SITEA, AOD goes to FEDERATED1_SITEB

End-user tools

- Given that an analysis job may open ~100-1000 files must move to non-GridFTP interface
  - Distributed Analysis group evaluating multiple protocols
- Statement so far was that sites need to provide a POSIX-like interface
  - we **did not say which one** and assumed it was in the site's interest - in serving their physics community - to decide the best one for their backend
- End-user tools should nonetheless use SRM where available (prepareToGet) and return a POSIX-like URL directly
  - Can we PLEASE get a common environment on OSG, LCG, NDGF and my user desktop? e.g to find out the nearest storage?
    - do we need a HEPIX for this :-) ?
    - How are we with GLUE schema on OSG sites?