LCG-2 Data Management Planning

Ian Bird

LHCC Referees Meeting
28th June 2004

Background

- ➢ Continue to develop LCG-2 service to deploy and validate basic underlying infrastructure services essential to have in place

- ➢ Cannot wait for new EGEE developments – but ensure we are aligned
  - What we do now may/will be replaced but there is still much to learn and understand
  - What we propose is consistent with EGEE developments
  - Underlying system-level issues (firewalls, security, network behaviour, error handling, …) need to be addressed now
  - Much is learned in the DC’s – need to validate solutions to those problems
  - Intend to deploy/validate EGEE solutions in parallel (on pre-production service)

Data Management – 3 areas

- Reliable Data Transfer Service
  - Essential to have in place and verify by end 2004 that we are able to reliably distribute data at significant fraction of data rate expected at LHC start-up
    - Series of service challenges associated with this
- File Catalogue
  - Based on lessons learned in DC’s in last few months – address some fundamental issues of performance and scalability
  - Valuable input for EGEE developments –
    - with appropriate interfaces could also be alternative implementation of EGEE model (?)
- Lightweight disk pool manager
  - Recognised as necessary – (LCG, Grid3, EGEE) – will be a collaborative effort

Reliable Data Transfer service

- Two areas of work:
  - Basic underlying infrastructure for service challenges
  - Management software

- Underlying infrastructure:
  - Load-balanced gridftp service at each end point
    - (500 MB/s would require several gridftp servers ~ >5?)
  - Disk pools in place
    - Disk management policies – garbage collection, etc.
  - Routing for data transfers around firewalls vs control channel

  - This is being set up now by ADC/CS together with FNAL and Nikhef

Reliable Data Transfer – management

- Implementation:
  - Currently investigating/testing 3 possibilities:
    - TMDB (from CMS) – together with EGEE and CMS
      - We could use “as-is”, EGEE want to adapt to new architecture
    - Stork (from VDT)
    - pyRFT (python implementation of Globus RFT)
  - Decide within a week – TMDB looks a good candidate solution

- All of these could be used with little adaptation, allowing us to focus on system-level issues
  - Optimising performance, security issues, etc

- Effort:
  - 1-2 people in GD team, together with CMS and EGEE – assuming TMDB
  - Work in testing has started, set up test framework to FNAL and Nikhef
    - Already being done in context of basic network infrastructure testing

File catalogues

- Proposal: -

  Key is to simplify, concentrate on functionality and performance

  - Single central file catalogue:
    - GUID → PFN mappings – no attributes on PFNs
    - LFN → GUID mappings – no user-definable attributes (they are in metadata catalogue)
    - System attributes on GUID – file size, checksum, etc
    - Hierarchical LFN namespace
    - Multiple LFNs for a GUID – compatible implementation with EGEE & Alien
    - Bulk inserts of LFN→GUID→PFN
    - Bulk queries, and cursors for large queries
    - Transactions, Control of transaction exposed to user
  - Metadata catalogue:
    - Assume most metadata is in experiment catalogues
    - For VO that need it – simple catalogue of “name-value” pair on GUID – separate from file catalogue

File catalogues – 2

- Other issues to be addressed:
  - Fix naming scheme (has been source of problems)
  - Transactions
  - Cursors for efficient and consistent large queries
  - Collections – in file catalogue – seen as directories/symlinks (or as GUID)
  - GSI authentication …
  - … simple C clients (extend existing C clients)
  - Management tools – logging, accounting, browsing (web based)

- Availability
  - Short term: assume fail-over between instances on several sites
    - Use Oracle tools, db clients look in IS for current primary
    - Oracle DataGuard makes them consistent
  - Longer term: multi-master database would fit this logic also – using IS

File catalogues – 3

- Has been discussed with POOL team
- Will be discussed in PEB tomorrow
- Effort is identified in Deployment team
  - Estimate 1 month for basic efficient implementation using existing catalogue
  - Begins now if PEB agrees – some up-front work has been done to investigate potential solutions
  - Prototype in mid-August

- Not addressed directly:
  - Replication –
    - Consider conflict resolution in implementation
    - Expect replication to use DB tools (Oracle) – subject of separate project
  - WAN interaction –
    - Several ideas (RRS, DB proxy from SAM)
    - Needed to provide connection re-use, timeouts, retries

Lightweight disk pool manager

- Recent experience and current thinking gives following strategy for storage access:
  - LCG-2, EGEE, Grid3 all see a need for a lightweight dpm
  - SRM is common interface to storage; 3 cases:

  1) Integration of large (tape) MSS (at Tier 1 etc) –
     - Responsibility of site to make the integration – this is the case
  2) Large Tier 2's – sites with large disk pools (10's Terabytes, many fileservers), need a flexible system
     - dCache provides a good solution, but needs effort to integrate and manage
  3) Sites with smaller disk pools, less available management effort
     - Need a lightweight (install, manage) solution

- We propose to develop 3)

Lightweight DPM

- Implementation
  - Re-use same catalogue infrastructure/name server as file catalogues
  - SRM interface leveraging what exists now
  - Re-use Globus gridftp server if possible
  - Local I/O using rfio

- Effort
  - in GDA/EGEE at CERN and Orsay
  - Interest from Grid3/OSG also

- Timescale:
  - Catalogue infrastructure in August
  - SRM implementation can start in parallel

Summary

- Propose to address 3 data management issues:
  - Reliable data delivery
  - File catalogues
  - Lightweight disk pool manager

- Focus on basic essential services – leave higher levels to experiments

- All important to have in place to understand basic system
  - Data transfer and file catalogues have priority …
  - … but a simple dpm is missing

- Work can start now