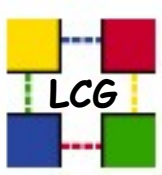


LHC Computing Grid Project

Oracle-based Production Services for LCG 1

Jamie Shiers
IT Division, CERN
Jamie.Shiers@cern.ch

Overview

- What are we trying to do?
- Current status
- Open questions
- Future possibilities

Goals

- To offer **production quality services** for LCG 1 to meet the requirements of forthcoming (and current!) data challenges
  - e.g. CMS PCP/DC04, ALICE PDC-3, ATLAS DC2, LHCb CDC'04
- To provide distribution kits, scripts and documentation to **assist** other sites in offering production services
- To leverage the many years' experience in running such services at CERN and other institutes
  - Monitoring, backup & recovery, tuning, capacity planning, …
- To understand experiments' requirements in how these services should be established, extended and clarify **current limitations**
- **Not** targeting small-medium scale DB apps that need to be run and administered locally (to user)

What Services?

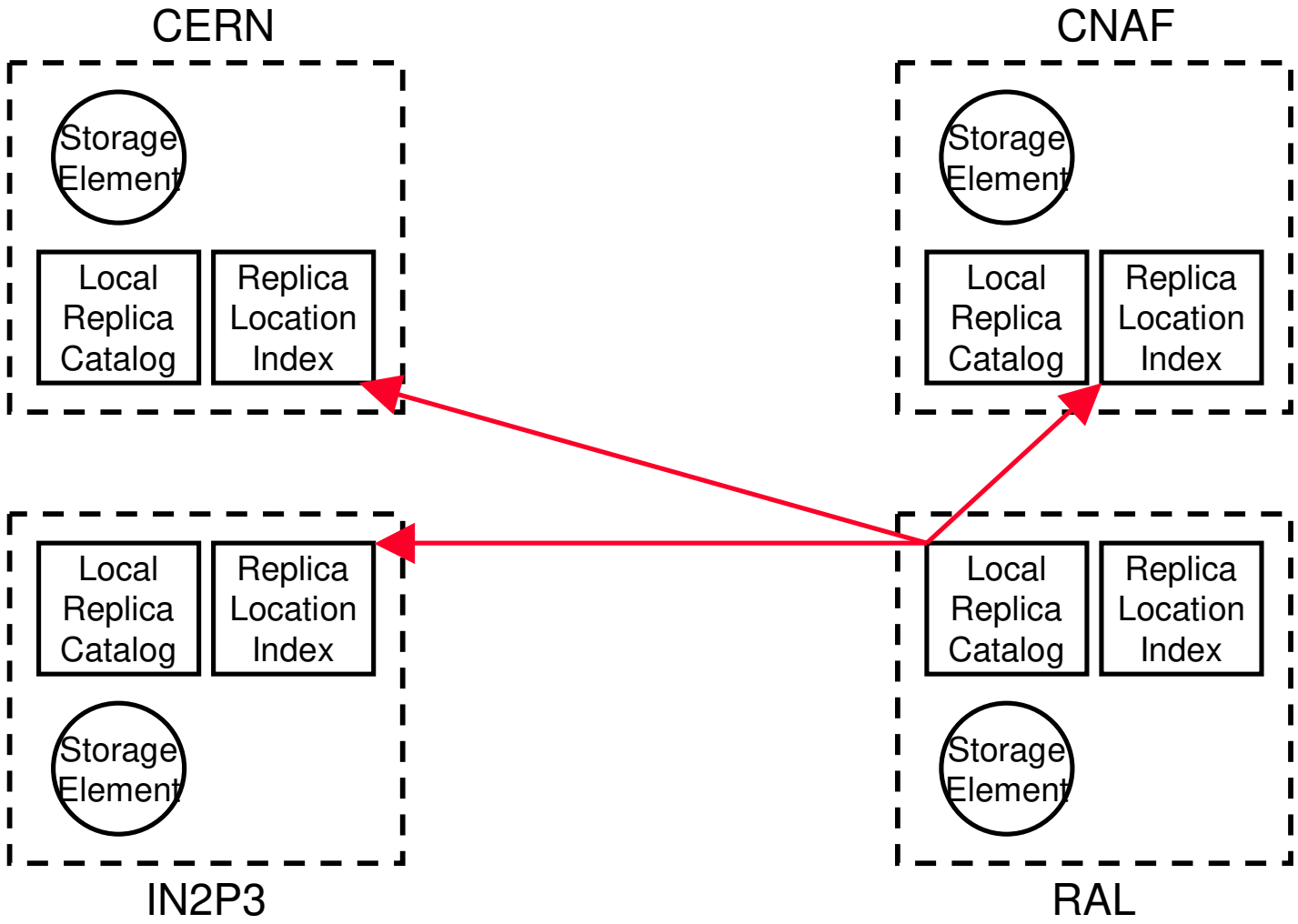
- POOL file catalogue using EDG-RLS (also non-POOL!)
  - LRC + RLI services + client APIs
  - For GUID <-> PFN mappings
- and EDG-RMC
  - For file-level meta-data: POOL currently stores:
    - filetype (e.g. ROOT file), fully registered, job status
  - Expect also ~10 items from CMS DC04: others?
- plus (service behind) EDG Replica Manager client tools
- Need to provide robustness, recovery, scalability, performance, …
- File catalogue is a critical component of the Grid!
  - Job scheduling, data access, …

The Supported Configuration

All participating sites should run:

- A "Local Replica Catalogue" (LRC)
  - Contains GUID <-> PFN mapping for all local files
- A "Replica Location Index" (RLI) <-- independent of EDG deadlines
  - Allows files at other sites to be found
  - All LRCs are configured to publish to all remote RLIs
    - Scalability beyond O(10) sites??
    - Hierarchical and other configurations may come later…
- A "Replica Metadata Catalogue" (RMC)
  - ***Not*** proposing a single, central RMC
  - Jobs should use local RMC
  - Short-term: handle synchronisation across RMCs
    - In principle possible today "on the POOL-side" (to be tested)
  - Long-term: middleware re-engineering?

Component Overview



Where should these services be run?

- At sites that can provide supported h/w & O/S configurations (next slide)
- At sites with existing Oracle support team
- We do not yet know whether we can make Oracle-based services easy enough to setup (surely?) and run (should be for canned apps?) where existing Oracle experience is not available
  - Will learn a lot from current roll-out
  - Pros: can benefit from scripts / doc / tools etc.
  - Other sites: simply re-extract catalog subset from nearest Tier1 in case of problems?
  - Need to understand use-cases and service level

Requirements for Deployment

- A **farm node** running Red Hat Enterprise Linux and Oracle9iAS
  - Runs Java middleware for LRC, RLI etc.
  - One per VO
- A **disk server** running Red Hat Enterprise Linux and Oracle9i
  - Data volume for LCG 1 small (~10^5 – 10^6 entries, each < 1KB)
  - Query / lookup rate low (~1 every 3 seconds)
    - Projection to 2008: 100 – 1000Hz; 10^9 entries
  - Shared between all VOs at a given site
- Site responsible for acquiring and installing h/w and RHEL
  - $349 for 'basic edition' http://www.redhat.com/software/rhel/es/

What if?

- DB server dies
  - No access to catalog until new server configured & DB
  - 'Hot standby' or clustered solution offers prote… most common cases
  - Regular dump of full catalog into alternate… g. POOL XML?
- Application server dies
  - Stateless, hence relatively sim… new host
    - Could share with an…
  - Handled automatica… cation server clusters
- Data corrupted
  - Restore … alternate catalog
- Softwe… ms
  - …o predict and protect against
  - …d cause running jobs to fail and drain batch queues!
  - Very careful testing, including by experiments, before move to a new version of the middleware (weeks, including smallish production run?)
- Need to foresee all possible problems, establish recovery plan and test!

What happens during period when catalog is unavailable?

Backup & Recovery, Monitoring

- Backend DB included in standard backup scheme
  - Daily full, hourly incrementals + archive log – allows point in time recovery
  - Need additional logging plus agreement with experiments to understand 'point in time' to recover to – and testing!
- Monitoring: both at box-level (FIO) and DB/AS/middleware
- Need to ensure problems (inevitable, even if undesirable) are handled gracefully
- Recovery tested regularly, by several members of the team

- Need to understand expectations:
  - Catalog entries guaranteed for ever?
  - Granularity of recovery?

Recommended Usage - Now

- POOL jobs: recommend extracting catalog sub-set prior to job and post-cataloging new entries as separate step
- Non-POOL jobs, e.g. EDG-RM client: minimum, test RC and implement simple retry + provide enough output in job log for manual recovery if necessary
  - Perpetual retry inappropriate if e.g. configuration error
- **In all cases, need to foresee hiccoughs in service e.g. 1 hour, particularly during ramp-up phase**
- Please provide us with examples of your usage so that we can ensure adequate coverage by test suite!
- Strict naming convention essential for any non-trivial catalogue maintenance

Status

- RLS/RLI/RMC services deployed at CERN for each experiment + DTEAM
  - RLSTEST service also available, but should not be used for production!
- Distribution mechanism, including kits, scripts and documentation available and 'well' debugged
- Only 1 outside site deployed so far (Taiwan) – others in the pipeline
  - FZK, RAL, FNAL, IN2P3, NIKHEF …
- We need help to define list and priorities!
- Actual installation rather fast (max a few hours)
- Lead time can be long
  - Assign resources etc – a few weeks!
- Plan is (still) to target first sites with Oracle experience to make scripts & doc as clear and smooth as possible
  - Then see if it makes sense to go further…

Registration for Access to Oracle Kits

- Well known method of account registration in dedicated group (OR)
- Names will be added to mailing list to announce e.g. new releases of Oracle s/w, patch sets etc.
- Foreseeing much more gentle roll-out than for previous packages
- Initially just DBAs supporting canned apps
  - RLS backend, later potential conditions DB if appropriate
- For simple, moderate-scale DB apps, consider use of central Sun cluster, already used by all LHC experiments
- Distribution kits, scripts etc in afs
  - /afs/cern.ch/project/oracle/export/
- Documentation also via Web
  - http://cern.ch/db/RLS/

Links

- http://cern.ch/wwwdb/grid-data-management.html

  High level overview of the various components; pointers to presentations on use-cases etc

- http://cern.ch/wwwdb/RLS/

  Detailed installation & configuration instructions

- http://pool.cern.ch/talksandpubl.html

  File catalog use-cases, DB requirements, many other talks…

Future Possibilities

- Investigating resilience against h/w failure using Application Server & Database clusters
- AS clusters also facilitate move of machines, addition of resources, optimal use of resources etc.
- DB clusters (RAC) can be combined with stand-by databases and other techniques for even greater robustness
- (Greatly?) simplified deployment, monitoring and recovery can be expected with Oracle10G

Summary

- Addressing production-quality DB services for LCG 1
- Clearly work in progress, but basic elements in place at CERN, deployment just starting outside
- Based on experience and knowledge of Oracle products, offering distribution kits, documentation and other tools to those sites that are interested
- Need more input on requirements and priorities of experiments regarding production plans

Licensing

- CERN has been an Oracle user since 1982
- Last major contract update in 1996
  - Based on "concurrent users" – a special agreement aimed at reducing overall cost
    - Concurrent users measured once per year using CERN developed tools (prior to accelerator start-up ☺)
    - # users growing at 15% compound (and hence maintenance)
  - List of software and platforms totally obsolete
- New contract from December 2002
  - Based on "named users" – an Oracle standard
    - (Per server process licensing also possible)
  - # users based on standard CERN figures
    - From Website, HR annual report… and also Oracle Web!
  - Products can be used anywhere in the world by any of the 'named users'
  - Products: Oracle Database and Application Server

Distribution

Extract from Collaborator Agreement:

It is understood that Oracle has granted to CERN a right and license to use and distribute the Products within the CERN research program. "Products" means:

- ***The Oracle computer products in object code form only.***
- ***User documentation.***

The undersigned Collaborator agrees not to engage in, cause or permit the reverse engineering, disassembly, decompilation, or any similar manipulation of the Products, make available the Products to any person or entity outside the collaboration, or copy the documentation.

Other Distribution Kits

- From Oracle 10*g*: ship client libraries e.g. with POOL?
  - Potential clients include conditionsDB, POOL with RDBMS backend etc.
- Server kit for other applications, e.g. local conditionsDB, local copy of COMPASS event metadata etc.
- Full distribution
  - Will require local experienced DBAs