



LCG 3D Project

Dirk Düllmann, IT-ADC
Service Challenge Meeting at CERN
24 February, 2005



Distributed Deployment of Databases

- LCG provides infrastructure for storage, distribution and replication of file based data
- Physics applications (and grid m/w) require a similar services for data hosted in relational databases
 - Several applications and grid services use RDBMS - and more are coming this year
 - Several sites have already experience in providing RDBMS services
- Goals for the 3D project as part of LCG
 - provide LCG apps with a consistent, location independent **access to database services**
 - **increase the availability and scalability** of database applications via replication
 - arrange for **shared deployment and administration** of this infrastructure during 24 x 7 operation
- Joint project between LCG sites, experiments and s/w projects
 - Time frame for first service: **deployment in autumn this year**
 - Evolve together with computing models and requirements towards LHC startup
- Held Distributed Database workshop at CERN last December
 - <http://agenda.cern.ch/fullAgenda.php?ida=a044341>

LCG 3D Non-Goals



- 👎 Store all database data
 - Experiments are free to deploy databases and distribute data under their responsibility.
- 👎 Setup a single monolithic distributed database system
 - Given constraints like WAN connections one can not assume that a single synchronously updated database could work and provide sufficient availability.
- 👎 Setup a single vendor system
 - Technology independence and a multi-vendor implementation will be required to minimize the long term risks and to adapt to the different requirements/constraints on different tiers.
- 👎 Impose a CERN centric infrastructure to participating sites
 - CERN is one equal partner of other LCG sites on each tier
- 👎 Decide on an architecture, implementation, new services, policies
 - Produce a technical proposal for all of those to LCG PEB/GDB



WP1 -Data Inventory and Distribution Requirements

- Members are s/w developers from experiments and grid services that use RDBMS data
- Gather data properties (volume, ownership) and requirements
- Integrate access to 3D services into their software

WP2 - Database Service Definition and Implementation

- Members are technology and deployment experts from LCG sites
- Propose a deployment implementation and common deployment procedures

WP3 - Evaluation Tasks

- Short, well defined technology evaluations against the requirements delivered by WP1
- Evaluation are proposed by WP2 (evaluation plan) and typically executed by the people proposing a technology for the service implementation and result in a short evaluation report

3D Data Inventory



- Collect and maintain a catalog of main RDBMS data types
- Experiments and grid s/w providers fill a table for each data type which is candidate for storage and replication via the 3D service
 - Basic storage properties
 - Data description, expected volume on T0/1/2 in 2005 (and evolution)
 - Ownership model: read-only, single user update, single site update, concurrent update
 - Replication/Caching properties
 - Replication model: site local, all t1, sliced t1, all t2, sliced t2 ...
 - Consistency/Latency: how quickly do changes need to reach other sites/tiers
 - Application constraints: DB vendor and DB version constraints
 - Reliability and Availability requirements
 - Essential for whole grid operation, for site operation, for experiment production,
 - Backup and Recovery policy
 - acceptable time to recover, location of backup(s)

Requirement Gathering



- All participating experiments have signed up their representatives
 - Rather 3 than 2 names
 - Tough job, as survey inside experiment crosses many boundaries!
- Started with a simple spreadsheet capturing the various applications
 - Grouped by application
 - One line per replication and Tier (distribution step)
 - Two main patterns
 - Experiment data: data fan-out - $T_0 \rightarrow T_n$
 - Grid services: data consolidation - $T_n \rightarrow T_0$
 - But some exceptions which need to be documented...
- Aim to collect complete set of requirements for database services
 - Eg also online data or data which is stored locally but never leaves a tier
 - Needed to properly size the h/w at each site

Preliminary Summary for 2005



- Applications mentioned so far
 - FileCatalog, Conditions, Geometry, Bookkeeping, Physics Meta Data, Collections, Grid Monitoring, TransferDB
 - Suspect that several smaller(?) applications are still missing
- Total volume per experiment: 50-500 GB
 - Error bars likely as big as the current spread
 - Significant flexibility/headroom required for service side!
- Number of applications to be supported: O(5)
 - Some still not existing or bound to MySQL at the moment
 - Distributed applications use either RAL or ODBC based implementation
- Distributed Data becomes read-only down from T0
 - Conservative approach for first service deployment
 - Current model **does not require multi-master replication**
- **Please make sure your experiment representatives know about your application requirements!**

Service Integration with Application Software



- Distributed database service to be coupled to the user application
 - Plan to use POOL RAL as 3D reference implementation
- How does an application find a suitable database?
 - DB vendor may vary depending on site where job is scheduled
- Database Service Catalog
 - Avoid embedding physical connection strings in code
 - ..or spreading them as part of configuration file copies
 - Allow a local service to be transparently relocated to another machine
- 3D prototype based on POOL file catalog
 - supports network disconnected lookup based on XML files
 - Final implementation may still change
 - Eg if a suitable grid-service has been identified

More Application Integration..



- **Connection Pooling and Error Handling**
 - ATLAS developments being integrated into RAL
- **Authentication and Authorization**
 - Mapping between grid certificate and database role
 - Need to support roles in experiments computing model
 - Plan to evaluate/integrate with package developed by G.Ganis for ROOT and xrootd
- **Client Diagnostics**
 - Collect timing of top queries of the current application in RAL
 - Evaluate package developed at FNAL to aggregate diagnostics across many clients
- **Most of the above are useful for local or centralized database applications too**

3D Site Contacts



- **Established contact to several Tier 1 and Tier 2 sites**
 - Tier 1: ASCC, BNL, CERN, CNAF, FNAL, GridKa, IN2P3, RAL
 - Tier 2: ANL, U Chicago
- **Regular 3D meetings**
 - Bi-weekly phone meeting back-to-back for
 - Requirement WG
 - Service definition WG
 - Usual difficulties of distributed
 - Organised a 3day 3D workshop at CERN
- **Visited RAL, FNAL and BNL**
 - Very useful discussions with experiment developers and database service teams there
 - Plan to visit CNAF beginning of March together with SC meeting there
- **All above sites have expressed interest in project participation**
 - BNL has started to setup Oracle setup
 - Most sites have allocated and installed h/w for participation in the 3D test bed
 - U Chicago agreed to act as Tier 2 in the testbed
- **Will contact DB teams at PIC, NHIKEF and Bari**

Database Service Policies



- Several sites have deployment policies in place
 - E.g. FNAL:
 - Staged service levels
 - Development -> Integration -> Production systems
 - Well defined move of new code / DB schema during development process
 - Apps developers and DB experts review and optimize schema before production deployment
- Similar policy proposal prepared for CERN physics database services
 - To avoid recent interference between key production applications of different experiments on shared resources
 - Caused by missing indices, inefficient queries, inadequate hardware resources
 - Storage volume alone is not a sufficient metric to define a database service
- Need a complete list of applications and a **reference workload** for each key application to define and optimize the service
 - How many requests (queries) from how many clients on how much data are required to meet the performance goals?
- Especially for distributed DB service this will be essential to avoid surprises on either side

Tier 1 Service Split



- Discussion only just starting
 - Draft proposal based on input received from FNAL (Anil Kumar)
- Local Services
 - Server installation, bug & security fixes
 - OS patches/upgrades
 - Backup/recovery support
 - Data migration (between db servers)
- Shared Services
 - Db and OS accounts & privileges
 - Storage support (adding more space)
 - Monitoring
 - DB alerts, “killer queries cron job output
 - Host system load, space thresholds
 - Performance problems & optimization
- Site Communication
 - Proposal to setup a shared (web based) Log-Book, mailing lists
 - Need to establish regular DBA meeting
 - eg as part of weekly/bi-weekly 3D meetings

Database Services at LCG Sites Today



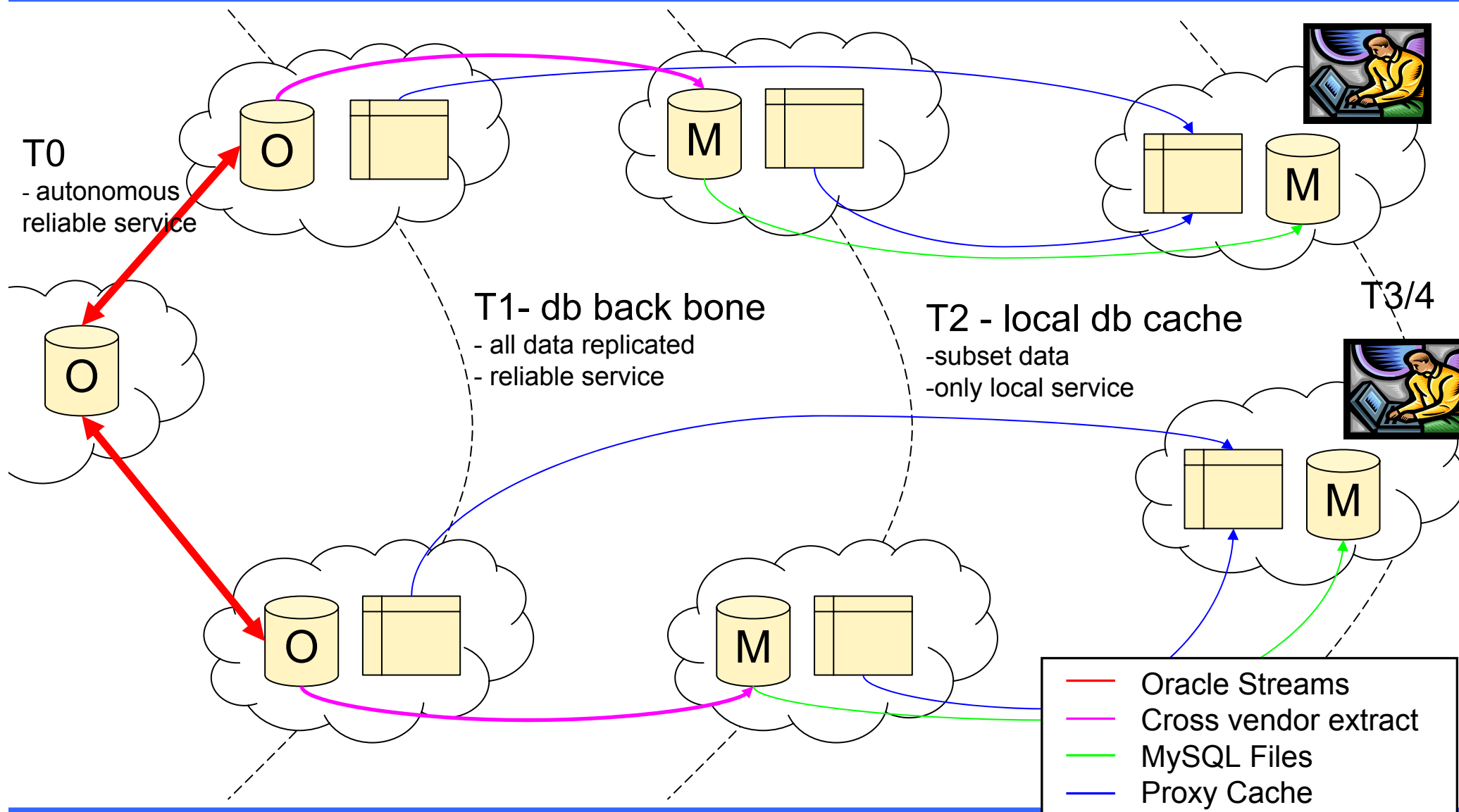
- Several tier 1 sites provide Oracle production services for HEP and non-HEP applications
 - Significant deployment experience and well established service exists...
 - ... but can not be changed easily without affecting other site activities
- Tier 2 sites can only provide very limited manpower for database service
 - Part time administration by the same people responsible for fabric
 - Only a simple, constrained database service should be assumed
- MySQL is very popular in the developer community
 - Several applications are bound to MySQL
 - Used for experiment production, though not at very large scale
 - Expected to be deployable with limited db administration resources
- Expect both database flavors to play a role implementing different parts of the LCG infrastructure

Local Database vs. Local Cache



- FNAL developed **FroNtier** system
 - a combination of **http based database access with proxy caches** on the way to the client
 - Performance gains
 - reduced real database access for largely read-only data
 - reduced transfer overhead compared to low level SOAP RPC based approaches
 - Deployment gains
 - Web caches (eg squid) are much simpler to deploy than databases and could remove the need for a local database deployment on some tiers
 - No vendor specific database libraries on the client side
 - “Firewall friendly” tunneling of requests through a single port
- **Expect cache technology to play a significant role**
 - towards the higher tiers which may not have the resources to run a reliable database service

Proposed 3D Service Architecture



LCG 3D Testbed



- Oracle 10g server
 - Install kits and documentation are provided for test bed sites
 - CERN can not offer offsite support though
 - At least 100GB storage
 - Application and reference load packaged by CERN / experiments
- FroNtier installation
 - FroNtier server at CERN and FNAL
 - plus squid installations at other sites
 - FNAL has prepared s/w packages and will install on the testbed
- { also expect some worker nodes to be available to run reference load }
- Oracle Enterprise Manager installation for test bed administration and server diagnostic
 - Client side monitoring being defined and integrated into POOL/RAL
 - Based on experience gained at RUN2 database deployment

Possible 3D use in SC3



- **Timescale for first 3D deployment only after SC3 start**
 - But will try to arrange for a preproduction service with the participating sites
 - Need a list of DB applications to be deployed in SC3
 - Eg a subset of the 3D spreadsheet with reduced volume (and lower number of sites)
 - Need to coordinate the software integration with the experiment development teams to be ready for SC3
 - Need to allocate database resources for SC3
- **Once the list of candidate apps has been discussed with the experiments:**
 - Either reduced database service at SC3 start
 - Or join SC3 later as 3D service / 3D database apps become available

Summary



- **A distributed database infrastructure promises to provide scalability and availability for database applications at LHC**
 - The LCG 3D project has been started as joint project between experiments and LCG sites to coordinate the definition of this service
 - Several distribution options are available for planning/design of new applications
- **DB Service Definition**
 - Very relevant deployment/development experience from RUN2 @ FNAL
 - Service task split and application validation policy proposals are firming up
 - Oracle 10g based replication test-bed expands to first T1
- **Several new DB applications will be developed this year**
 - Expect high support / consultancy load on database services (especially at Tier 0) for 2005
 - Early planning of concrete deployment and distribution models is key for successful production
- **SC3 support from 3D is only now being discussed**
 - Participation in SC3 would be a very useful first test of 3D services
 - Concrete time plan needs negotiation with experiments/sites