Distributed database project ensures replication to LCG sites

Physics metadata stored in relational databases play a crucial role in the Large Hadron Collider (LHC) experiments and also in the operation of the LHC Computing Grid (LCG) services. A large proportion of non-event data such as detector conditions, calibration, geometry and production bookkeeping relies heavily on databases. Also, the core Grid services that catalogue and distribute LHC data cannot operate without a reliable database infrastructure at CERN and the LCG sites.

The Distributed Deployment of Databases (3D) project is a joint activity between the IT department's Physics Services Support (IT/PSS) group, the LHC experiments and LCG sites to coordinate database services that are coherent, scalable and highly available.

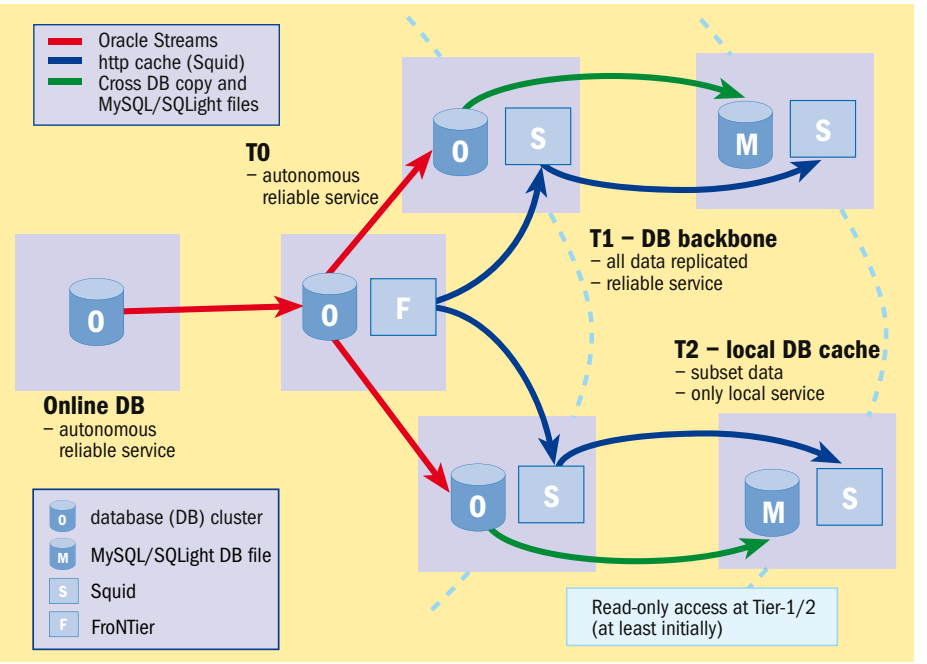

Fig. 1. Database clusters form the main building-blocks of the service architecture in the LCG Distributed Deployment of Databases project. Oracle clusters are used up to Tier-1.

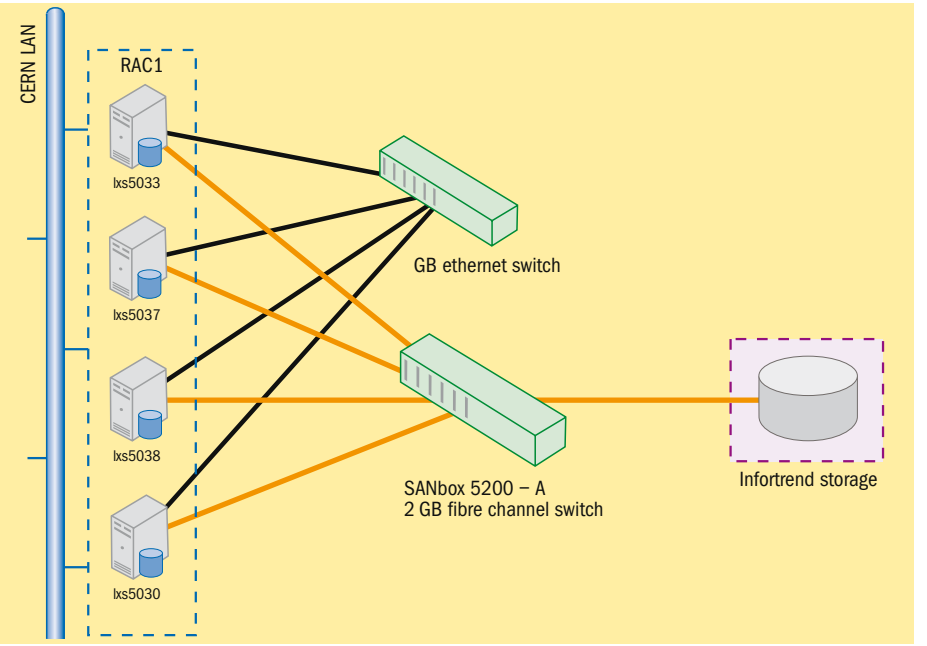

Fig. 2. An example of a database service provided by an Oracle Real Application Cluster (RAC). Several processing nodes access data shared in a storage area network (SAN).

The 3D project and service architecture

Most of the LHC data can be stored and distributed as read-only files. Nevertheless, a significant proportion of data from the central experiment and the Grid requires database services such as:

- consistent and highly available storage for data that is simultaneously accessed or updated;
- recovery to a consistent state after hardware, software or human failures;
- support for efficient *ad hoc* queries.

The 3D project started in mid-2004. Database representatives from 11 Tier-1 sites and three LHC experiments (ATLAS, CMS and LHCb) are profiting from a close cooperation with Oracle as part of CERN openlab. The project also involves two teams from IT/PSS group: physics database services and LCG persistency framework, which also provide the main LCG database services and applications. From the start the project has focused on using database clusters as the main building-blocks in the database service architecture (figure 1).

Building-blocks for a distributed service

Oracle Real Application Clusters (RACs) are used to implement reliable database services at the different stages of the main data flow, from online to offline up to Tier-1 sites. Each database cluster (figure 2) consists of several processing nodes that access data shared in a storage area network (SAN). Today for each experiment there are typically eight nodes at Tier-0 and two to three nodes at Tier-1. The cluster nodes use a dedicated network to share cached data blocks to minimize the number of disk operations. A public network connects the database cluster to client applications, which may execute queries in parallel on several nodes. The set-up provides important flexibility to expand the database server resources (CPU and storage independently) according

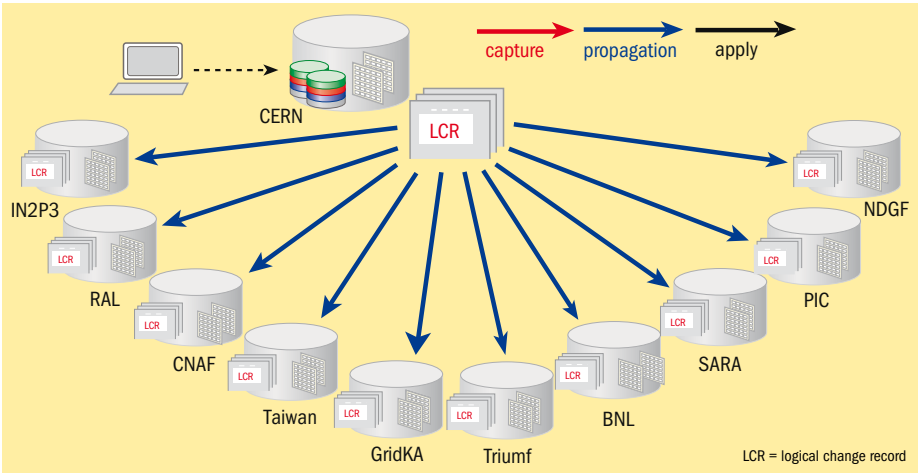

Fig. 3. Oracle Streams forms a replication backbone between Tier-0 and Tier-1 sites.

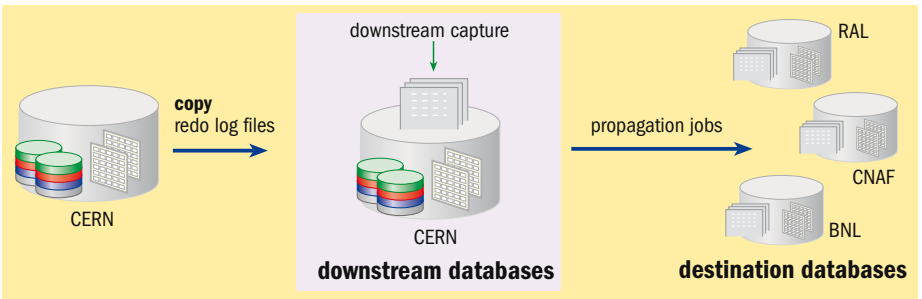

Fig. 4. Data is captured downstream at Tier-0 before being replicated to Tier-1 databases.

to users' needs. This is particularly important during the early phases of the LHC operation, since several applications are still under development and data volume and access patterns may change.

In addition to its intrinsic scalability, the cluster also increases significantly the availability of the database. Should individual nodes fail, applications can failover to one of the remaining cluster nodes and many regular service interventions can be performed without database downtime node by node.

During the last year the physics database service run by IT/PSS has undertaken major preparations for the start-up of the LHC and is now fully based on Oracle clusters on Intel/Linux. More than 100 database server nodes are deployed in some 15 clusters serving almost 2 million database sessions per week. The positive experience with this new architecture at CERN and other sites has led to the setting up of similar database installations at the Tier-1 partner sites worldwide, forming one of the largest Oracle RAC installations.

Connecting database servers via replication

To enable LHC data to flow through this distributed infrastructure, Oracle Streams, an asynchronous replication tool, is used to form a database backbone between online and offline and between Tier-0 and Tier-1 sites. New or updated data from the online or offline database systems are detected from database logs and then queued for transmission to all configured destination databases. Only once data has been successfully applied at all destination databases is it removed from message queues at the source (figure 3). A dedicated downstream capture set-up is used for replication via the wide-area network to further insulate the source database in case of problems in connecting to Tier-1 replicas (figure 4).

The replication process can be applied either to individual tables or to whole application schemata and takes care of transactional dependencies between several related data changes. In the case of network outages or overloading, the destination databases may experience a time lag to the source database at Tier-0 until the connection is re-established. Data consistency, however, is preserved at each site and at any point in time. This property greatly simplifies the deployment of applications against replica databases, as the semantic of state changes remains intact.

Tests performed by the application developers have confirmed that the experiment and Grid database applications (e.g. the LFC catalog and LCG COOL) work correctly against replicated databases without significant changes in code.

Even though the Oracle Streams replication mechanism supports more complex configurations (e.g. updates at replica databases), its initial deployment will be based on the model that all data changes happen at the replication source at CERN, and replica databases offer read-only access.

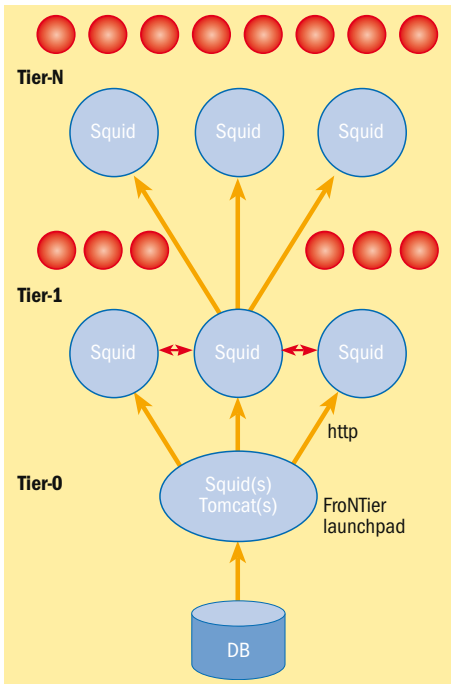

Fig. 5. CMS is using a FroNTier server and connected Squid servers to cache data.

Distribution and caching of database results

Complementary to database replication, the CMS experiment is deploying a second technique for its Tier-1 and Tier-2 sites, based on distributing and caching database data via a web protocol. This approach, which is based on the FroNTier package developed at Fermilab, encodes database queries as url requests and transfers the corresponding query results as xml documents. Using http and html enables standard web cache servers such as Squid to enhance the scalability of the distribution system by caching frequently requested data close to the client location at each Tier-1 or Tier-2 site. Figure 5 shows the FroNTier server and a hierarchy of connected Squid cache servers. At CERN a load-balanced set-up with three FroNTier servers was used successfully as part of the CMS challenges in 2006.

A key advantage of FroNTier/Squid distribution is that it requires less effort to install and operate the cache servers, compared with deploying a replica database server. Since all cached data are still available at Tier-0 there is no need to recover the cache state after a hardware failure. Today some 30 Tier-1 and Tier-2 sites have been successfully integrated into the CMS distribution set-up.

To fully benefit from caching, care must be taken when designing the application to avoid subtle inconsistencies that may be caused by stale cache content