Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Evolution of Database Replication Technologies for WLCG

To cite this article: Zbigniew Baranowski *et al* 2015 *J. Phys.: Conf. Ser.* **664** 042032

View the article online for updates and enhancements.

Evolution of Database Replication Technologies for WLCG

Zbigniew Baranowski, Lorena Lobato Pardavila, Marcin Blaszczyk, Gancho Dimitrov, Luca Canali

European Organisation for Nuclear Research (CERN), CH-1211 Geneva 23, Switzerland

Abstract. In this article we summarize several years of experience on database replication technologies used at WLCG and we provide a short review of the available Oracle technologies and their key characteristics. One of the notable changes and improvement in this area in recent past has been the introduction of Oracle GoldenGate as a replacement of Oracle Streams. We report in this article on the preparation and later upgrades for remote replication done in collaboration with ATLAS and Tier 1 database administrators, including the experience from running Oracle GoldenGate in production. Moreover, we report on another key technology in this area: Oracle Active Data Guard which has been adopted in several of the mission critical use cases for database replication between online and offline databases for the LHC experiments.

1. INTRODUCTION

The Worldwide LHC Computing Grid (WLCG [1]) project is a global collaboration of more than 170 computing centres around the world, linking up national and international grid infrastructures. The project goal is to provide global computing resource to store, distribute and analyse tens of Petabytes of data annually generated by the experiments at the LHC.

Several grid applications and services rely in large extent on relational databases for transactional processing. In particular conditions and controls data from the LHC experiments are typically stored in relational databases and used by analysis jobs in the grid. Deployment of database services in such complex distributed environment is challenging and often has to handle parallel workloads originated in different part of the world by the grid applications. In order to provide better scalability and lower latency, certain databases have been deployed in multiple centres within the grid. Each database system installation in such environment must have a consistent copy of the data originated at Tier 0. This requirement translates into the need for a reliable database replication solution in the deployment model for WLCG. In 2006 the first prototype of a service providing database replication was deployed using native database replication technologies provided by Oracle, the database vendor. This has since provided a consistent way of accessing database services at CERN and selected Tier 1 sites to achieve more scalable and available access to non-event data (e.g. conditions, alignment, geometry, environmental parameters or bookkeeping) [2].

This paper describes the evolution of database replication technologies since its first production deployment in 2007 until 2015, when LHC was restarted for Run 2. In particular, each database replication solution used within WLCG project will be covered with a description of their features and their role for CERN's use cases.

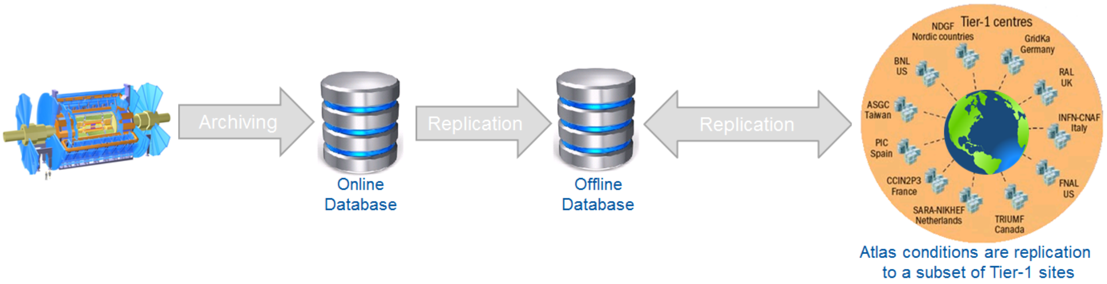


Figure 1: Database Replication Service for ATLAS experiment

2. DATABASE REPLICATION FOR HEP

Data replication is a key component of online-offline database model deployed for all LHC experiments' database services (see Fig. 1). It implements data synchronization between the online database installations, used by the experiments control systems, and offline databases, available to a larger community of users. In particular, there are two data sets that are required to be consistent between online and offline database: the detectors conditions data and the controls system archives (generated by the WinCC OpenArchitecture, formerly called PVSS). Both data sets are initially archived on the online databases and subsequently all changes to the data are propagated with low latency to the offline databases by the database replication solutions.

The analysis and reconstruction of LHC events within WLCG requires conditions data, therefore data access has to be available worldwide. Each experiment has an individual approach for enabling conditions data access within the grid. For example, ATLAS and LHCb decided to distribute conditions data from their offline systems to database installations in specific Tier 1 data centres. CMS instead deployed a single replica of the conditions data at CERN on the offline database, and installed on top of it a distributed cache service, called Frontier [3]. Finally, during Run 1, LHCb decided to replace database replication with a file based solution: CERN Virtual Machine File System (CVMFS). However, the database replication between the online and the offline systems was also preserved.

In addition to distributing conditions data, LHCb has used in Run 1 database replication solutions to propagate the LHC File Catalog (LHC) data to the replicas at Tier 1s. More recently this has been decommissioned and replaced by the Dirac File Catalog with a migration performed at the end of Run 1.

ALICE has chosen a custom ROOT file based solution for distribution of the detector conditions data. However other data sets like the data for the ALICE Detector Control System (DCS) have to be replicated between the online and the offline database with conventional database replication solutions.

There are two additional use cases where data are originated at Tier 1 or Tier 2 sites and are later consolidated at CERN. Both are a part of ATLAS replication topology (see Fig. 2). The ATLAS Metadata Interface (AMI) data are replicated from the IN2P3 centre in Lyon to the ATLAS offline database. Muon calibration data originated at Munich, Michigan and Rome are also replicated to the ATLAS offline database.

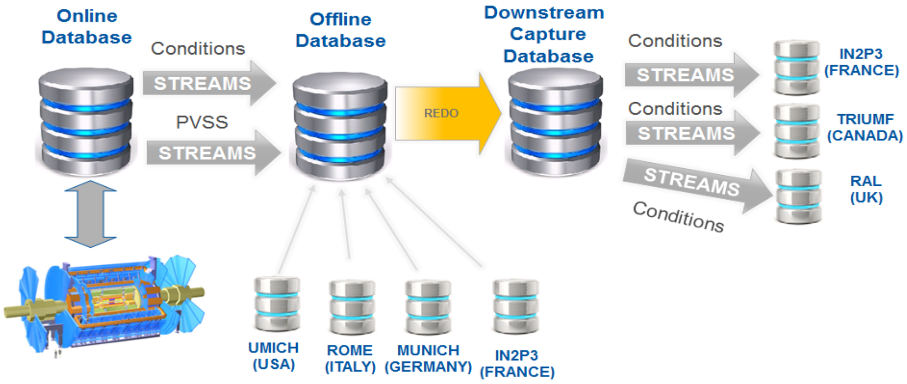


Figure 2: ATLAS database replication for WLCG using Oracle Streams

In 2004, Oracle Streams [4] was evaluated as the most suitable solution to implement database replication for all use cases of interest for WLCG. The main characteristic of the technology is to provide “logical” replication which means that changes in the source database are replicated to the target database as SQL operations. Because Oracle Streams has a modular architecture (see Fig. 3) where each of the three components can work independently, it provides a great flexibility in designing data flow topologies. Moreover, administrators have full control on replicated content and can apply data filtering at any stage of the replication flow. The details of the technology architecture are not covered in this paper, they can be found in references [2][4].

After a pilot phase and comprehensive evaluation of the Oracle Streams functionality, the first implementation of database replication was deployed in production in 2007 for online-to-offline and offline-to-Tier 1s data propagation (the latter only for ATLAS and LHCb experiments).

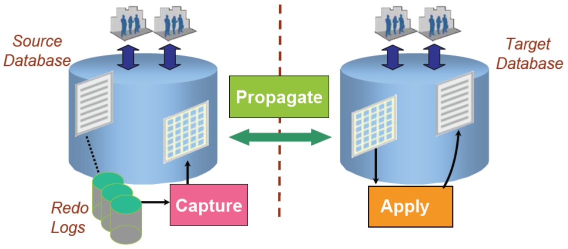


Figure 3: Main components of Oracle Streams

Oracle Streams have been used successfully for several years, allowing asynchronous data flow among complex distributed database environment. In contrast, one of the weak points of using logical replication is the lack of guarantee of data consistency between master and slave databases. The main reason of that is the implementation of the data changes and flow management. It is based on an internal message propagation framework which does not have lost message detection. If for any reason a message or a set of messages is lost, the corresponding changes would be lost too and this would possibly happen unnoticed. Consequently, data divergence between primary and replica systems may be introduced in the system.

Another aspect which could lead to a master-slave inconsistency is the need of having a replica database in read-write mode. This allows users to break the data consistency by performing data modifications on a replica database.

In addition, it has been found that the replication into a large number of targets (Tier 1 centres) has potentially an important overhead on the source production database. This has been solved by deploying an additional component: Oracle Streams Downstream Capture. This has the effected of offloading all the primary databases which have multiple replicas and move the bulk of Oracle Streams workload into the dedicated system for Downstream Capture.

3. TECHNOLOGY EVOLUTION

Motivation

Oracle Streams has been a key solution for the distribution of experiments' metadata to Tier 1 sites as well as within CERN and has been the focus of significant development and integration efforts within CERN openlab program [5]. Thanks to the joined efforts, essential improvements in performance and stability of the software have been deployed. This has been instrumental for successful operations during first run of LHC (2008-2013). Meanwhile, two new options for database replication, Oracle Active Data Guard [6] and Oracle GoldenGate [7] have been released by Oracle. Both have advantages for CERN's use cases by offering higher performance and lower maintenance efforts than Oracle Streams. With the appearance of this new technology Oracle has also deprecated Oracle Streams as a replication solution in favour of Oracle GoldenGate. Therefore, a study and evaluation of the new technologies has been performed during Run 1.

Oracle Active Data Guard (ADG)

Oracle Active Data Guard (see Fig. 4) is a technology providing real-time data replication at the physical level, available since Oracle database version 11.2. The main difference between Oracle Active Data Guard and Oracle Streams, is that an ADG database is a mirror copy (at the data block level) of the source system, being constantly updated with the latest data changes and being accessible in read-only mode. While Oracle Streams propagates the changes as SQL statements, Oracle ADG propagates the changes via the transactional logs of Oracle (redo log and/or archived log). This different architecture makes ADG more resilient against data divergence and more performant than Oracle Streams.

Naturally ADG also plays an important role as high availability solution, providing fall back in case of disaster recovery scenarios. Also, compared with the other technologies, ADG as lower maintenance effort, since it provides robustness thanks to a relative simple replication architecture. Since 2012, ADG is used by CMS and ALICE online databases. More recently ATLAS is also using ADG for controls data (PVSS) replication.

One important limitation of ADG is that it does not provide data filtering for the replicated content, all contents in transaction logs are shipped from the source to the destination database. The default data copying granularity for ADG is full database replication. The absence of data filtering capabilities (the entire data stream has to be transferred) can be problematic for replication over WAN with high latencies. Also, since ADG requires exactly the same version of a database binaries running on master and slave systems, is an essential constraint when the two systems are at a different locations and are administered by different teams.

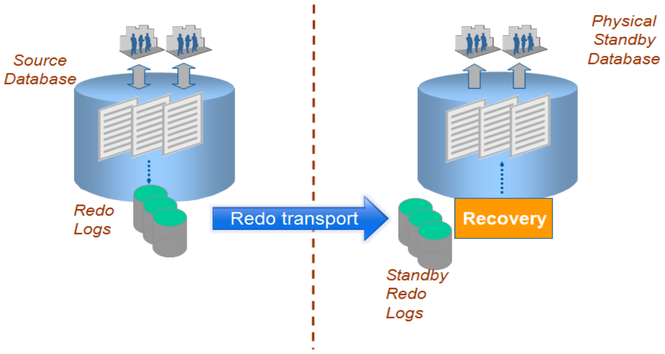


Figure 4: Main components of Oracle Active Data Guard

Oracle GoldenGate (OGG)

In 2009 Oracle acquired GoldenGate Software Inc., a leading provider of a heterogeneous database replication software. Since then Oracle has declared Oracle GoldenGate (OGG) the strategic solution for SQL-level database replication (also called logical replication) and has declared Oracle Streams a deprecated feature. However, in context of Oracle-to-Oracle database replication, the main components of OGG and functionalities are close to what Oracle Streams had offered. In this context we can consider OGG an improved version of Oracle Streams, with more features as logical replication over heterogeneous database systems, which supports more common data types than Oracle Streams. Also, in OGG the replication granularity can be defined at a schema level, which is an advantage, for example, over Active Data Guard when a limited part of the database is needed to be replicated.

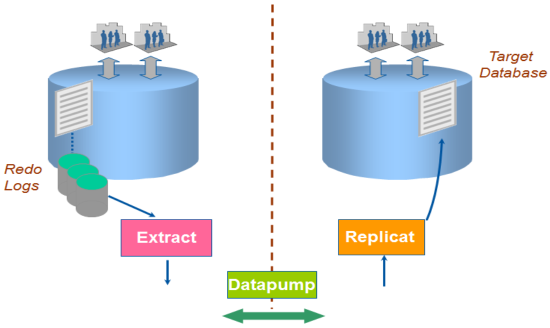


Figure 5: Main components of Oracle GoldenGate

The schematic description of the main components of OGG is shown in Fig. 5. Similarly, to Oracle Streams, OGG relies on a processing of database redo logs to capture data changes on the source database. These changes are temporarily stored on disk, within a series of files called trail files. On the source side, the second extraction process called DataPump, extracts data change operations from the local trail files generated by the primary extract process and transfers them through the network to another set of trail files located at the target system(s). On the target system(s) there is the third process, called Replicat, which has the task of reading the trail files and applying data modification operations to a target database.

The main difference in the OGG architecture comparing to Oracle Streams is the usage of trail files in OGG as a staging container instead of buffered queues as in the case of Oracle Streams. This removes an important dependency between source and target system: by using files the extraction process at the master database is decoupled from the data application processes at replica database. This improves the stability and overall performance of the replication and reduces potential negative impacts of the replication on the workload of the master system.

4. TECHNOLOGY ASSESSMENT

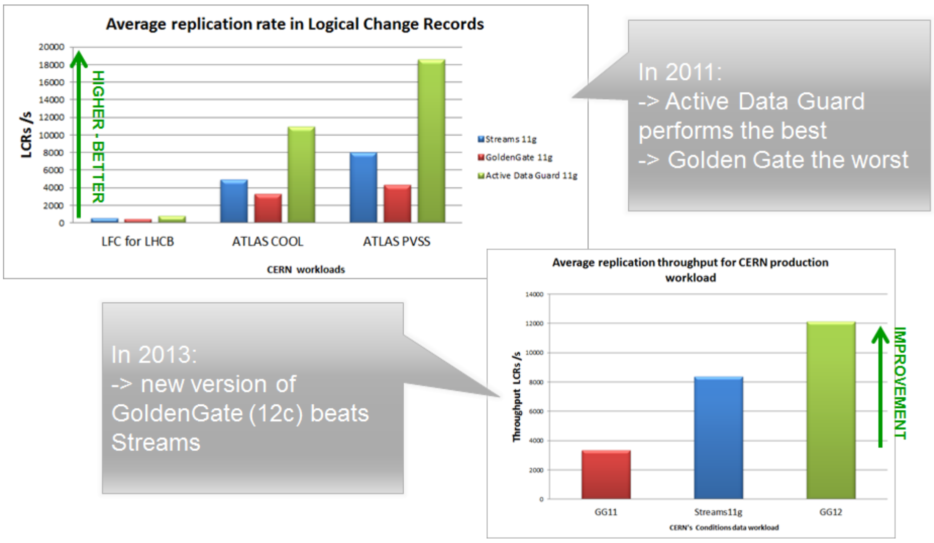
A comprehensive evaluation of both Active Data Guard and OGG, have been performed before considering them as potential replacements for Oracle Streams. Most of the work done in this area was driven by the CERN openlab project [5].

Initial performance tests confirmed that OGG performance was similar but inferior to what could be achieved with Oracle Streams 11g or with Active Data Guard 11g. One of the reasons of such state was a suboptimal support of parallel data application by OGG. Moreover, due to latencies caused by accessing trail files, a single-threaded replication OGG was slower than Oracle Streams in first tests when running in comparable configurations. At the same time Active Data Guard showed to be the solution with highest performance out of all the three. This is because block-level replication gives performance advantage over all the solutions based on mining redo logs and performing SQL-level replication (such as OGG and Oracle Streams).

All the mentioned results were confirmed with tests performed with synthetic data, as well as with production data copies from LFC, ATLAS Conditions and Controls Archives (PVSS). As an outcome of the tests, a detailed feedback was provided to Oracle, including ideas for potential improvements which could be applied to OGG product.

In 2013, Oracle released the OGG version 12c, including significant changes in the architecture, such as the introduction of the parallelism coordinator, which incidentally was suggested by CERN as a result of the tests performed in the context of openlab. Overall OGG demonstrated better performance than Oracle Streams 11g in all tests: both with synthetic data and with real data (ATLAS conditions). Fig. 6 illustrates the performance comparison of OGG 11, Oracle Streams 11g and OGG 12c. OGG 12c has the best results mostly due to the improvements for parallel processing of replicated data.

Also, better integration with the database software was introduced in OGG 12c that allowed to profit from features and tools available before only for Oracle Streams, as in-database replication monitoring, diagnostic and troubleshooting tools. Some essential features required by the CERN database workloads (e.g., native handling of Data Definition Language) were also present in this new version.



In 2011:
-> Active Data Guard performs the best
-> Golden Gate the worst

In 2013:
-> new version of GoldenGate (12c) beats Streams

Figure 6: Replication Technologies Evaluation Performance

During validation tests, Oracle Active Data Guard appeared the solution with highest robustness and performance. In particular, it appeared suitable for online-offline database replication flows within CERN. However, lack of data filtering and consequently and the requirement of having the same software versions at source and destination, made it unsuitable for cross-tiers data replication. OGG however showed to be able to fulfil those use cases. The outcome of the evaluation of the new replication technologies was to move forward with the preparation for migration from Oracle Streams to Oracle GoldenGate and Oracle Active Data Guard.

5. DEPLOYMENT

5.1. *Planning.*

In 2012, the first deployment of online-offline replication based on ADG was performed together with the DB upgrades from version 10g to 11g. This change was applied to CMS online – offline replication for controls and conditions data and for ALICE online – offline replication for controls data.

At the time of Oracle upgrades from 10g to 11g and first deployments of Oracle ADG, Oracle GoldenGate was not yet ready for CERN use cases, for the reasons mentioned in the paragraph above on testing. Therefore, the replication between CERN and Tier 1 sites continued to be handled by Oracle Streams. The same conclusions applied to ATLAS conditions data replication between online – offline because of the constraint of cascading the replication to Tier 1 sites (data were replicated from online to offline and later captured and propagated to selected Tier 1 sites).

In the case of LHCb, a small portion of conditions data were being replicated between online and offline and afterwards, certain sets needed being replicated back to online. As it was mentioned before, having data to be copied partially, made using ADG not optimal.

In 2013, Oracle GoldenGate 12c version was released, meeting all CERN requirements. After a successful validation of the new software it was approved for deployment in production replacing Oracle Streams. The technology transition plan included two stages: an upgrade of online-offline ATLAS and LHCb replication which was completed in the third quarter of 2014, and an upgrade of offline-Tier 1 sites replication as the second step which took place in the 4th quarter of 2014. The upgrade to Oracle GoldenGate has proved to be smooth and overall improve the robustness and performance of WLCG replication use cases, as expected from the results of the evaluation phase and tests.

5.2. *New infrastructure*

Deployment of Oracle Active Data Guard is similar to the installation of a new database system. The main difference is the duplication of the primary data (by using a backup or direct data file copying over a network) instead of the initialization of an empty system. Redo log shipment services between master and the slave and recovery process have to be additionally configured and started in ADG configuration.

In contrast to ADG deployment, Oracle GoldenGate deployment is a complex procedure which greatly differs from the configuration of Oracle Streams. Mainly, because it requires installation and configuration of an extra software layer while Streams are embedded in a database binaries.

In order to ease installation, configuration, maintenance and monitoring of OGG, a central OGG cluster (see Fig. 7) has been deployed for hosting all replication installations (CERN and Tier 1s). In such configuration, a cluster of two machines manages the storage for all trail files and hosts all OGG processes (Extract and Replicat). An OGG DataPump process in such central deployment is not needed as both Extract and Replicat processes are running on the same machine and can share a single copy of the trail files. In such deployment model (unlike in case of a classical OGG installation) the configuration of the servers hosting the databases stays unchanged. No additional software from OGG is needed there, nor extra TCP port opening. This has a great advantage as it allows to ease global configuration management by preserving a single configuration template for all the database servers.

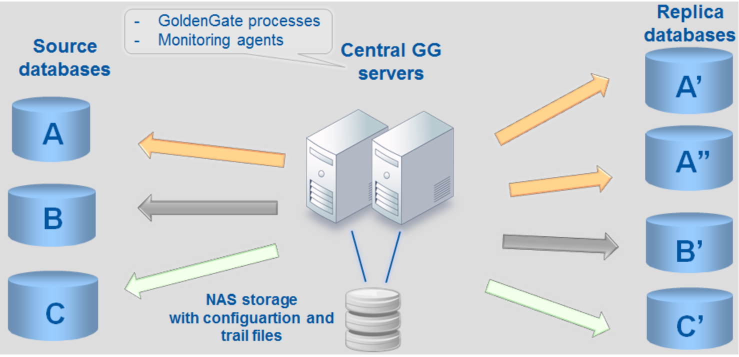


Figure 7: Centralised configuration at CERN

5.3. *Migration details.*

Active Data Guard (ADG)

The migration process between Streams and Active Data Guard is straightforward. Firstly, a read-only replica system has to be set up for the primary database. The next and most important step is to redirect all the users' sessions that are accessing a copy of the data from the current system implemented with Streams to the new one based on ADG solution. Such rerouting of the client connections can be done in different ways: at the network level can be done by modifying the IP aliases of the target database server (the replica) at a DNS level. Another method is to update the connection string definition used by Oracle clients. A third solution would be to update the client software in order to use the connection string and credentials of a new ADG system. Finally, when all clients are successfully connected to the target system and are able to read the data, the old replica (implemented with Oracle Streams) can be decommissioned.

Oracle GoldenGate (OGG)

Migration from Streams to OGG is more complex than in the case of Active Data Guard due to the replacement of the replication engines that has to be done in place. At a given time only one of the replication solutions can process the data changes between two end points, otherwise the data consistency will be broken. Therefore, the replication technology migration has to be done in an atomic fashion. In order to ensure the logical equivalence of the configuration by both replication solution, a migration script provided by Oracle was extensively tested and later used for setting up the Oracle GoldenGate processes.

Finally, once all the OGG process were in place, the sequence of actions used for making the transition between Oracle Streams and Oracle GoldenGate was as follows:

1. Start Extract process in order to capture new data changes, while Oracle Streams are still in operation
2. After a while, when the Extract managed to consume all the backlog, all Oracle Streams processes can be stopped
3. Start Replicat (data changes application process) with 'handle collisions' mode. All overlapped changes already applied by Streams will be detected and ignored.
4. Once Replicat managed to consume all the backlog, the 'handle collisions' mode has to be disabled.

6. SUMMARY

Database replication is a key technology to enable distribution of conditions and controls data from online to offline databases for LHC experiments. It is also used to distribute the conditions data from CERN to selected Tier 1 centres. Replication technologies used to support these use cases have evolved since the beginning of the project in 2004. Oracle Streams has been used as the only technology in the first implementation of the replication services. Over the years Oracle has introduced new solutions with important advantages.

In particular Oracle Active Data Guard has been introduced with Oracle version 11g and has allowed to take advantage of the performance and robustness of block-based replication for the use cases of online-to-offline database replication. More recently the release of Oracle Golden Gate version 12c has provided an alternative and improvement to Oracle Streams for the use cases of schema-based replication such as ATLAS' conditions data replication from online to offline and from CERN to selected Tier 1 centres.

The evolution of the database technologies (see Fig. 8) deployed for WLCG database services have improved availability, performance and robustness of the replication service through the years.

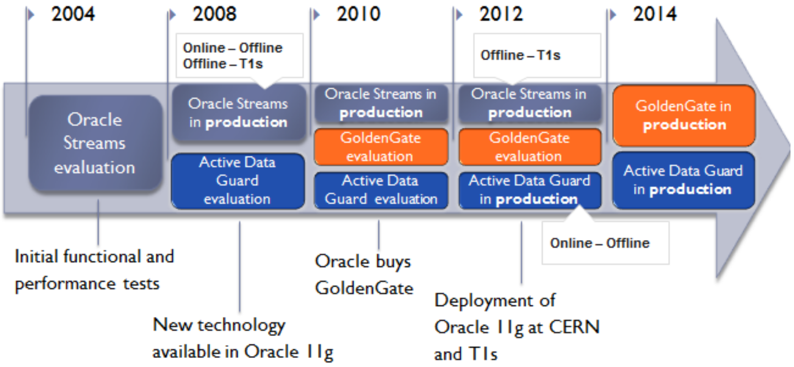

2004
2008
2010
2012
2014
Online – Offline
Offline – T1s
Offline – T1s
Oracle Streams evaluation
Oracle Streams in production
Active Data Guard evaluation
Oracle Streams in production
GoldenGate evaluation
Active Data Guard evaluation
Oracle Streams in production
GoldenGate evaluation
Active Data Guard in production
GoldenGate in production
Active Data Guard in production
Online – Offline
Initial functional and performance tests
New technology available in Oracle 11g
Oracle buys GoldenGate
Deployment of Oracle 11g at CERN and T1s

Figure 8: Timeline of the database replication technology for WLCG

7. Acknowledgment

We thank colleagues from the LHC experiments and Oracle employees who have, in the CERN openlab context, helped this project to succeed.

References

[1] Worldwide LHC Computing Grid project http://wlcg.web.cern.ch/

[2] D. Duellmann et al: LCG 3D Project Status and Production Plans in Proceedings of Computing in High Energy and Nuclear Physics (CHEP06), Mumbai, India, February 2006

[3] Frontier project: http://frontier.cern.ch

[4] Oracle Streams documentation: http://docs.oracle.com/cd/B28359_01/server.111/b28321/strms_over.htm

[5] CERN openlab project http://openlab.web.cern.ch

[6] Active Data Guard documentation http://docs.oracle.com/database/121/SBYDB/toc.htm

[7] Oracle GoldenGate documentation http://docs.oracle.com/goldengate/1212/gg-winux/index.html