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 centers around the world, linking up national and international grid infrastructures. The project goal is to provide global computing resource to store, distribute and analyze the many tens of Petabytes of data annually generated by the LHC experiments. Several grid applications and services rely in large extent on relational databases for transactional processing. Moreover conditions and controls data from the LHC experiments are typically stored in relational databases and later retrieved by analysis jobs in the grid. Deployment of database services in such complex distributed environment is challenging and has to handle a parallel workload originated in different part of the world by grid applications. In order to provide better scalability certain databases have been deployed in multiple centers within the grid. Each database system set up in such environment should have a consistent copy of the data originated at Tier 0 with respect to that there was a need for data synchronization solution. This requirement was implemented with native database replication technologies provided by the database vendor. In 2006, significant preparation related to the replication technologies was mentioned in other contributions to this conference [2]. In view of this, data replication is used as a consistent way of accessing database services at CERN Tier 0 and specific collaborating Tier 1 sites to achieve a more scalable and available access to non-even data (e.g. conditions, alignment, geometry, environmental parameters or bookkeeping).

This paper describes the evolution of database replication technologies since its first deployment in 2007 until 2015 when LHC was restarted for Run2. In particular each replication solution used within WLCG project will be covered with a description of their features and their role for CERN’s use cases.
2. DATABASE REPLICATION FOR HEP

Data replication is a key component of online-offline database model deployed for all LHC experiments’ database services (Figure 1). It implements data synchronization between 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 and acquisition system archives (generated by WinCC OpenArchitecture, formerly called PVSS). Both data sets are initially archived on the online databases and subsequently all changes to the data are propagated to the offline databases by database replication solutions with low latency.

The analysis and reconstruction of LHC events within WLCG requires the 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 centers. 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 Run1, LHCb decided to replace database replication with a file based solution, CERN Virtual Machine File System. However database replication between online and offline systems was also preserved.

In addition to the distribution of the conditions data LHCb was using database replication solutions to propagate data of the LHC File Catalog (LHC) to the replicas at Tier 1s. Also this use case was not needed anymore when the functionality of LFC was migrated to the Dirac File Catalog.

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 part of ATLAS replication topology. ATLAS metadata interface data is replicated from IN2P3 center in Lyon to the ATLAS offline database. Muon calibration data originated at Munich, Michigan and Rome is also replicated to the ATLAS offline database.
In 2004, Oracle Streams [4] evaluation started, where the first implementation of database replication was configured for online-to-offline and offline-to-Tier 1s replication. This technology was used for sharing information between databases, where changes were captured from the database transactional change records (Oracle redo log and archive log files) and propagated asynchronously as Logical Change Records (LCRs). The architecture of this technology is not covered here since it was already covered in other contributions [2][4].

Oracle Streams was used for several years, allowing asynchronous data flow, automatic blacklisting and unavailable replica recovery among other characteristics. Between versions 10g and 11g, there was a notable improvement on this technology, having better performance, stability and, since data filtering was enabled during extraction, no high bandwidth network was needed.

In contrast, one of the weak points was the lack of guarantee of data coherency by this type of logical data replication. The main reason was data changes flow was based on messages propagation) that were translated into single DML statements. Afterwards, they were applied on target system without transaction integrity validation. Therefore, if for any reason a message or set of messages would have been lost, the corresponding changes would have been lost too and consequently data divergence between primary and replica systems would be introduced.

Another aspect which could lead to inconsistency master-slave system was related to the fact of replica database needed to be in read-write mode. This could lead the users to break the data consistency by performing data modification on replica database.

In addition, real cases confirmed about high overhead grew in proportion to number of replicas in a database. For this reason, among other details mentioned above, there was a need of deploying additional system called Downstream Capture, to offload all primary databases which have multiple replicas.

3. TECHNOLOGY EVOLUTION

Motivation

Oracle Streams has been a key solution for the distribution of experiment metadata to the Tier 1 sites as well as within CERN, and has been the focus of significant development effort within CERN openlab program [5]. Thanks to the joined efforts, significant improvements in performance and stability of the software have been deployed. This was instrumental for successful operations during first run of LHC (2008-2013). Meanwhile, two new options for database replication, Oracle Active Data Guard and Oracle GoldenGate became available and presented advantages for CERN’s use cases by offering higher performance and lower maintenance efforts than Oracle Streams. At the same time Oracle announced that Streams are not the preferred data replication solution anymore in favor of OGG. Therefore, the study and evaluation of both new technologies begun.
Active Data Guard (ADG)

Active Data Guard is a technology providing real time data block level replication, available since database version 11.2. The main difference regarding to Oracle Streams, is that replica database is exact a mirror copy of source system, being constantly updated with the latest data changes and being accessible in read-only mode. This feature, among additional block checksum mechanism, ensures data coherency with ADG.

ADG does not provide data filtering and even though requires continuous shipping of database redo logs (transaction logs) between systems, it fulfils diverse use cases in CERN data distribution area. Naturally, as a high availability solution ADG is a preferred technology for critical database systems as online and offline databases. Also, compared with the other technologies, a maintenance effort is notable decreased, since provides robustness and relative simplicity of the replication. Since 2012, ADG is used by CMS, ALICE online databases and more recently by ATLAS for controls data (PVSS).

On the other hand, the default data copying granularity is a full database replication. Among absence of data filtering (entire data stream has to be transferred), can be problematic for WAN with high latencies. Also, since ADG requires exactly the same version of database binaries version running on source and slave systems, is specially a constraint when both systems are in different locations.

Oracle GoldenGate (OGG)

In 2009, Oracle acquired GoldenGate Software Inc. as a leading provider of heterogeneous database replication software. In context of Oracle-to-Oracle database replication, the main components of OGG and functionalities were close to what Streams offered. It became the strategic solution for SQL-level database replication (also called logical replication) for Oracle, concurrently with Oracle Streams being a deprecated feature.

OGG became quickly an improved version of Streams, with more features as logical replication over heterogeneous database systems, which supports more common data types than Oracle Streams. Also, replication granularity is at schema level, which is advantage over Active Data Guard when limited chunk of database is replicated.
A schematic description of main components of OGG is shown in Figure 5. As in the case of Oracle Streams, OGG relies on redo logs to capture data or schema changes on the source database. These changes are temporarily stored on disk, within series of files called trail files. On the source side, a second extraction process called DataPump, extracts the DML (data manipulation language, such as SQL insert, update and delete statements) and the DDL (data definition language, such as SQL create statements) operations from the local trails linked up to the primary extract process. Afterwards, this process performs further processing if needed, and transfers the data to the target database through the network.

On the target side there is a third process, called Replicat, reading the trail files associated and applies replicated DML and DDL operations to the target database.

The main difference in OGG setup comparing to Oracle Streams is the usage of trail files instead of buffered queues as a staging container. Relatively unlimited file buffers allows decoupling extraction process from data application processes, since not need to communicate and control each other (flow control). This improves the overall performance and reduces negative footprint on the source system.

4. TECHNOLOGIES ASSESSMENT

Whenever a new technology becomes available on the market, is need the examination to verify if the product satisfies functions, stability and performance requirements, before being deployed on production systems. This rule was applied to all replication technologies related to CERN. Consequently, an intensive evaluation of both Active Data Guard and OGG, have been performed before considering them as potential replacement of Oracle Streams. The intense work with production data subsets, was driven by the CERN openlab project. This paper will not cover the details of methodologies applied during above testing.

Initial performance tests with the default software configurations confirmed that OGG performance was similar but inferior to what could be achieved with Oracle Streams 11g or Active Data Guard 11g. One of the reasons was not supporting properly parallel data application from OGG. Moreover, single-threaded replication in first tests was slower than Oracle Streams running in the same mode, due to latencies caused by accessing to the trail files. As Active Data Guard, this is based on block-level replication which gives performance advantage over all the solutions based on mining redo logs and performing SQL-level replication (such as OGG and Streams).

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

In 2013, Oracle released OGG version 12c, including significant changes in the architecture, as the introduction of parallelism coordinator which was suggested by CERN. Exploiting the new OGG version, a better performance to Streams 11g was demonstrated, either synthetic data or real data (ATLAS conditions). In addition, an improvement in data delivery parallelism was revealed through last OGG new versions, resulting with higher overall throughput, as illustrated in Figure 7.

Also, there was a better integration with the database software which 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 CERN database workloads (e.g., native handling of Data Definition Language) have also been deployed.

A long term stability test performed with OGG and ADG provided remarkable results, as well as the team achieved over four months of continuous data propagation, without any negative impact on the source database.
Oracle GoldenGate and Oracle Active Data Guard are being developed actively with new features, whereas Streams Project was in maintenance mode, and therefore its functionalities will not be extended anymore. For that reason and having Streams configuration on production databases, the preparation to the migration begun, in order to improve the replication technologies environment.

5. DEPLOYMENT

5.1. Planning.

Oracle Active Data Guard appeared as a very reliable and performing solution since validation tests. But, lack of data filtering, relatively high network throughput imposed (dependency of primary database activity) and requirement of same software versions on multiple installations, have influenced on the decision of do not use ADG product for cross-tiers data replication. Nevertheless, ADG would be implemented for all replication use cases within CERN between online and offline databases. This set-up provided an additional real-time synchronized replica which significantly increased protection and availability of the online data.

In 2012, first deployment of online-offline replication based on ADG became possible with the DB upgrades from version 10g to 11g, starting the migration planning between Oracle Streams and ADG. This planning involved CMS online – offline for control and conditions data replication in one hand and, on the other hand, ALICE online – offline for control data.

Simultaneously, cross-tiers replication case needed to keep being implemented by a SQL-level technology like Oracle Streams or OGG. The goal was to optimize resources in network, storage, or maintenance operations, either CERN or Tier 1 sites. But, by that time, OGG was not yet ready for CERN use cases. Therefore, the replication between CERN and Tier 1 sites continued with the configuration of Oracle Streams. The same case was for ATLAS conditions data replication between online – offline but with cascading replication model (data were replicated to offline database and afterwards, were captured and propagated to corresponding Tier 1 sites).

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

In 2013, the new version 12c of OGG targeting all CERN requirements, was released. After successful validation, OGG would be deployed in production replacing Oracle Streams. The plan was established in two stages: online-offline ATLAS and LHCb replication would be completed in the third quarter of 2014, and offline-Tier 1 sites ATLAS replication would be in the fourth quarter.
5.2. *New infrastructure*

Deployment of Oracle Active Data Guard did not differ much from installation of a new database system. The hardware and software had to be installed identically. The main difference is the duplication of the primary data (by using a backup or direct data file copying over a network) instead 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, Oracle GoldenGate deployment is a complex procedure which greatly differs from configuration of Oracle Streams. Mainly, because it requires installation and configuration of additional software when Streams are embedded in the database binaries.

In order to ease installation, configuration and maintenance of OGG, a central OGG cluster has been deployed for hosting all replication installations (CERN and Tier 1s). In such configuration, a cluster of two machines has storage for all trail files and hosts all OGG processes (Extract and Replicat) except the DataPump. This process was avoided since the Replicat is running on the source database and connected remotely to the target. Therefore there is no need for an intermediate storage for the trail files. Also, multiple databases and OGG homes have been installed in order to cover the different source and target versions.

In such topology, OGG installation is not needed on any of database servers, as well as opening manager ports. With this configuration, all OGG processes and trail files are located in a single place, being able to monitor whole environments from CERN.

![Centralised configuration at CERN](image)

5.3. *Migration.*

*Active Data Guard (ADG)*

The process of migration between Streams and Active Data Guard is straightforward. Firstly, a read-only replica system has to be set up for a primary database, following the procedure briefly discussed in the previous paragraph. 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 connection can be done in different ways: at the network level can be done by modifying the IP aliases from the target system at DNS level. Another method is to update the connection string definition used by the Oracle clients in the file tnsnames.ora. A third solution would be to update the client software in order to use the connection string and credentials of the new ADG system. Finally, when all clients are successfully connected to the target system, being able to read the data, the old replica can be decommissioned.

*Oracle GoldenGate (OGG)*

Migration from Streams to OGG is more sophisticated than in the case of Active Data Guard due to the need of replication engine to be replaced (source and destination databases are unchanged). Therefore, such technology transition has to be done atomically. Both replication technologies cannot
run in parallel, being necessary to replicate a continuous stream of data changes in parallel. Failing both requirements, will lead to break data consistency. For that reason, before starting the migration, different tests have been done with production copies.

For setting up Oracle GoldenGate processes configuration, a migration script provided by Oracle was used. The main point is the conversion of processes from Oracle Streams into OGG processes by this script. Once all the processes are converted, parameter files should be configured to prepare the environment.

Finally, once all the OGG process are in place, the sequence of actions used for making the transition between Oracle Streams and Oracle GoldenGate looks as the following listing:

1. Start Extract process to capture new data changes, while Oracle Streams are still in operation
2. After 10 minutes all Oracle Streams processes need to be stopped
3. Start Replicat (data changes application process) with “handle collisions” mode. All overlapped changes will be detected and ignored.
4. Once Replicat is up to date (no lag), the “handle collisions” mode need 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 conditions data from CERN to selected Tier 1 centers. The 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 from 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 centers.

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

7. Acknowledgment

We thank colleagues from the LHC experiments and Oracle employees who have, in the CERN openlab context, enabled part of this work to be a success.

References
[1] Worldwide LHC Computing Grid project http://wlcg.web.cern.ch/
[2] D.Duellmann & al: LCG 3D Project Status and Production Plans in Proceedings of Computing in
Frontier project: http://frontier.cern.ch/
Oracle Streams documentation:
   http://docs.oracle.com/cd/B28359_01/server.111/b28321/strms_over.htm
CERN Openlab project http://openlab.web.cern.ch