Relational databases for conditions data and event selection in ATLAS

F Viegas[1], R Hawkings[1] ,G Dimitrov[1,2]

[1]CERN, CH-1211 Genève 23, Switzerland
[2]LBL , Lawrence-Berkeley National Laboratory, Berkeley, CA 94720, USA

Abstract. The ATLAS experiment at LHC will make extensive use of relational databases in both online and offline contexts, running to O(TBytes) per year. Two of the most challenging applications in terms of data volume and access patterns are conditions data, making use of the LHC conditions database, COOL, and the TAG database, that stores summary event quantities allowing a rapid selection of interesting events. Both of these databases are being replicated to regional computing centres using Oracle Streams technology, in collaboration with the LCG 3D project. Database optimisation, performance tests and first user experience with these applications will be described, together with plans for first LHC data-taking and future prospects.

1. Introduction

In the ATLAS Computing Model, various applications require access to the data resident in relational databases. This data includes, but are not limited to: technical databases (detector production, installation and survey data), detector geometry, online/TDAQ databases, conditions databases (online and offline), event data, offline processing configuration and book-keeping, distributed data management, and distributed database and data management services.

ATLAS data processing will make use of a distributed infrastructure of relational databases in order to access various types of non-event data and event metadata. Distribution, replication, and synchronization of these databases, which employ more than one database technology, must be supported according to the needs of the various database service client applications.

This paper describes the work completed, under way and planned within the DB Project to meet the data-handling needs of ATLAS at startup and beyond. It focus specifically on two of the most challenging applications in the ATLAS software, Conditions and Event Metadata. These applications require state-of-the-art technology, complex software architecture and human resource coordination, in order to meet the demands of the experiment, in terms of size, throughput and reliability.

2. Data Types and Data Storage

In the ATLAS computing model, data comes in all shapes and sizes, and when taken as a whole, it is expected to amount to O(10PB) per year. So, in the specification of the computing capabilities, usage patterns and technology determine which data will be stored uniquely on file-based storage, and which data will be uploaded into relational-database storage. From the computing TDR we have the quote that identifies in a very pragmatic manner the data that is the best candidate for relational storage: "Database storage is used where concurrent writes and transactional consistency are required; where data handling is inherently distributed, typically with centralized writers and distributed readers; where indexing and rapid querying across moderate data volumes is required; and where structured archival storage and query-based retrieval is required."[1].

The conditions data and event metadata fall into several of these categories.

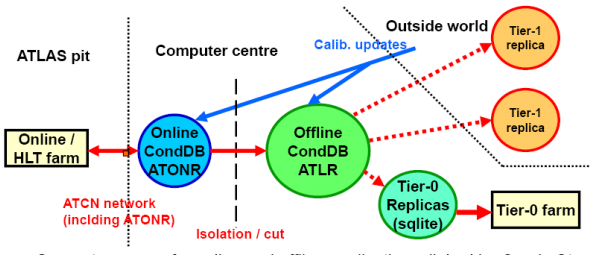

Figure 1. Conditions Data Flow – Insertion and Replication of Data

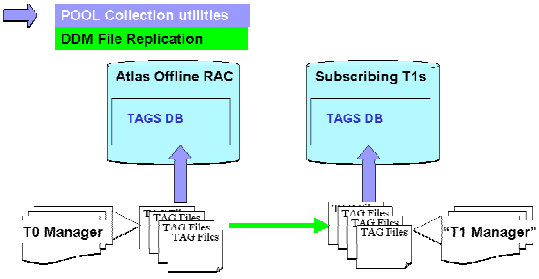

Figure 2. Event Metadata Dataflow

2.1. Conditions Data

Conditions data refers to nearly all the non-event data produced during the operation of the ATLAS detector, together with that required to perform reconstruction and analysis. Conditions data varies with time, and is usually characterized by an 'interval of validity' (IOV), i.e. a period of time for which it is valid, expressed as a range either of absolute times or run and event numbers. Conditions data includes data archived from the ATLAS detector control system (DCS), online book-keeping data, online and offline calibration and alignment data, and monitoring data characterising the performance of the detector and software during any particular period of time.

2.1.1. Data Flow. Conditions data will be input at the online and offline database level. Sources of data will be[2]:

- Online DAQ systems, inserting into the online database. These processes include: Run Control at the start of each run, Central and High level trigger with information about the luminosity and Monitoring services e.g. histogram information.
- Subdetector data, specific data will be input into COOL schemas, including configuration, monitoring and calibration data.
- ATLAS DCS data, coming from the PVSS online database archive, will be extracted and input into the offline database by the dedicated process PVSS2COOL.
- Calibration processing will be updated at the offline database and extracted to SQLite for updates at the online level.

Retrieval of data will occur at every point of the chain.

2.1.2. Data volume. The data volume estimated for the conditions data resident in the database is around 1TB a year, in a nominal year. Recent testing activities and best-guess estimates have given a ballpark figure of 2GB/day, and this figure has been used for testing the different pieces of technology envolved, and shaping the database organization. This figure breaks down into 1.7Gb of data, and 300M of indices, where 1.4 GB are Online DCS data. [2]

2.2. Event Metadata

2.2.1. Data Flow. Event Metadata will arrive at the CASTOR storage from processes deriving the data from ESD and AOD files. As the TAG files are produced the Tier-0 Manager (aka T0MS) will upload the TAG files into database relational collections, using the POOL Collection utilities. These files will be also replicated to the Tier-1 centers, where they may be loaded into the database, should the Tier-1 wish so. The data flow activity is mainly a heavy bulk insert activity into the database, and a heavy querying activity from the applications detailed in 2.2.3.[3]

2.2.2. Data volume. The data volume estimated for the TAGS database is 6 TB per year, based on a rate of insertion of TAG Events of 200Hz, and a record size averaging 1kB. It is the largest database application running on ATLAS databases. It is very resource and storage intensive, so it is not part of the mandatory resource requirements to Tier-1s. Outside of CERN, Institutes have TAGS in their database on a voluntary basis.

2.2.3. Retrieval Use Cases. There are three main use cases for TAGS data selection[4]:

- Pool Collection Utilities, retrieve from database Pool Collections and create ROOT files that the user can access locally.
- GANGA TNT plug-in, indirect use of the POOL collection utilities to retrieve a selection of AOD and/or ESD file pointers, into a ROOT file, and schedule the grid jobs to process them.
- Web query and retrieval, there is a website created with Stream Test data, that puts in practice a web interface for direct querying of the TAGS database, with controlled SQL statements, but full use of the fields for event selection. This will evolve into a full production web querying tool .

3. Storage Technology

3.1. Choice of Database Software

Both the conditions data and TAGS data are stored and retrieved primarily via a database-neutral interface named Relational Access Layer (CORAL). CORAL implements a performance-optimized, vendor-neutral, C++-based DB access. CORAL-created schemas can also be accessed through conventional DB tools. ATLAS utilizes CORAL both directly and through layered services for DB storage via C++ with either Oracle, MySQL or SQLite back-end engines.

Although CORAL provides this level of flexibility, it quickly became apparent that some of the data processing stages need more robustness, throughput and scalability than others. So ATLAS chose the Oracle Database Software to be the large scale DB servicing the online,T0 and T1 levels, while MySQL can be used at installations outside of institutional IT departments (T2 and T3) and SQLite for local file-based storage with SQL relational capacities to be used in processing subsets of data on local machines and grid jobs.

The Oracle Software administration is being provided by CERN IT's Physics Database Service, with high availability and data protection support, complemented by the ATLAS DBAs for developer support and application architecture steering.

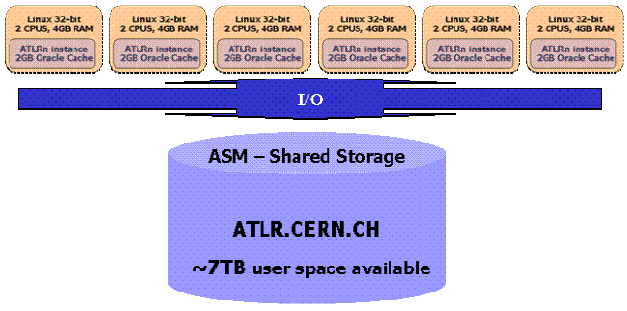

Figure 3. Oracle Database Architecture for the ATLAS production T0 database

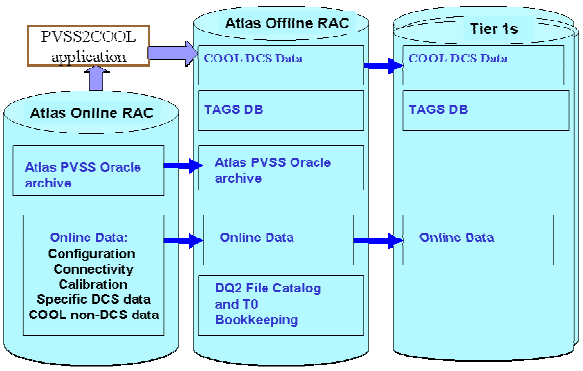

Figure 4. ATLAS relational distributed database, comprising all tiers.

3.2. Architecture

At present, the hardware assigned for the ATLAS databases is composed of Linux clusters with the symmetry of all nodes that access the database. The architecture of a typical cluster is shown in Figure 3.

The online and offline clusters at CERN are at present 2 clusters of 6 nodes each, with Intel PIV, running Linux. Each node has 1 dual-core processor and 4GB of RAM, of which 2GB are allocated to the Oracle Cache. The systems are connected to an ASM storage system with 7TB of user data space available for each production database. The names are ATONR for the online RAC (Real Application Clusters, Oracle nomenclature for parallel instances accessing the same storage database) and ATLR for the T0 offline database.

There are also two clusters with 2 nodes each for validation and testing of the applications before moving into production, the nodes are equivalent as described above, and the storage available totals 7TB for these systems.

Plans for upgrade of system are under the purchasing requisitions, and is estimated that at the beginning of 2008 each cluster will be substituted by 3 quad-core machines with 16GB of RAM each. This will increase throughput of the system by an estimated 300% CPU throughput and an increase of 60% for non-parallel operations [5].

3.3. Logical organization of data

Figure 4 depicts the internal Oracle schema organization for the ATLAS distributed database, comprising of online, offline and T1 systems . Great care had to be taken in separating different chunks of data because the Oracle Streams replication flow is unidirectional, and activated on a schema-level. For this, data had to be separated in different Oracle schema organizations depending on source and update points.

At present, there are 5 groups of data to manage:

- ATLAS PVSS archive – inserted at the online database and replicated to the T0 level
- ATLAS Online Data – comprising all subdetector data, including COOL non-DCS data, replicated to the T0 level and the T1 level.
- ATLAS Offline COOL DCS data – extracted from the online PVSS archive and inserted into the COOL Offline schemas, replicated to T1s.
- TAGS database – replicated to T1s that are interested, through file pushing through the Distributed Data Management facilities.
- Other T0 resource intensive applications, without replication to T1s planned so far.

3.4. Architectural challenges

3.4.1. Conditions database

The conditions database offers unique challenges in regard to management of size, management of replication and consistency, maintenance of efficiency and demands of users.

At the moment all tables are simple Oracle tables, for the online and offline data, but with the estimated size of data, especially from the DCS input, the size of the tables will become very large. This is quite challenging because the huge querying frequency of this data will make any inefficiency of SQL noticeable very quickly. To address this problem, large scale tests have been made and problems spotted immediately. The COOL developing team is also looking into the partitioning of the large tables for this matter.

This is especially critical for the replication of data, as the error recovery procedures may demand export of data to the T1s, and so modularization for minimizing downtime is critical.

Another aspect of a challenge in the replication is the filtering of data inside the schema, derived from the organization of the data in "databases" which cluster tables with similar prefixes. This has to be coded in the Oracle Streams replication, and is responsible for consuming CPU and slowing down replication. So far the extensive testing of the replication technology has offered solutions to workaround these issues, but surprises occur as testing is intensified. Hopefully technology is advancing at a fast rate, and the next hardware upgrade will overcome these issues.

3.4.2. TAGS database

The TAGS database is what we call internally our "data warehouse" for the event data. It is unique in the massive usage of resources it requires. The storage resource alone is enough to make Tier-1s shy away from hosting such a database, without seeing it work first, as the yearly amount of data is around 6 TB and this is on a conservative estimate, not allowing any data duplication, and the heart of the system which is the all-columns-indexed "Holy Grail".

For the TAGS computing model to work, great care has to be put in an architecture that allows very fast writing speed, of 200kB per second, that has to accommodate "any-field" queries, which can either get statistics or slice out a full month's worth of data into a ROOT file.

The effort put into reaching a solution for such extreme requirements is being put in several fronts:

- Schema architecture that is compliant with the POOL model, but that can accommodate fast writing (no indexes, bulk inserts, etc.) and at the same time make available data for reading, fully indexed and partitioned for performance.
- Querying use cases known, controlled as much as possible, sample queries and patterns taken from tests actual users analysing the data.
- Getting the right partitioning, and query optimizations in place to minimize resource usage as much as possible.
- Curbing the user activity to avoid overload of the systems, by prioritization and Oracle resource management, but still maintaining a level of service that makes TAGS a useful tool for analysis.
- With Oracle Streams out of the question as a replication tool to the Tier-1s, test other replication systems and the possibility of the uploading of the files at each subscribing site.

4. Data Protection and Monitoring

4.1. Backup and Recovery Policy

Since the main critical database services (Online and T0) are provided and maintained by the IT PSS services, the backup and recovery policy of these databases was proposed by this service. The policy was presented to the ATLAS community and discussed in its specifics internally, and after some adjustments was agreed by the ATLAS database group.[6]

In essence, the databases have internal mechanisms that can be used by the user themselves to recover a dropped table, or to undo transactions, which are set on a time and space availability frame. After this frame has been exhausted, a 24-hour old image of the database is kept on disk, and the IT PHYDB support can recover any object within this time frame from this image. After this time data has to be recovered from the tape system which is the slowest method of recovery.

With this in mind, a recent recovery need from the disk image took 9 hours to restore, from a 1.1 Terabyte database, which is ATLR, or T0 database. This database has a typical transaction rate at present, of 16.6 Gb per day, which has great influence on the time taken to recover the data.

After this incident, another look was taken into the highly critical applications i.e. DQ2 and PVSS which cannot stop for 9 hours, and the database group is analysing these specific cases to ensure high availability.

4.2. Monitoring effort

A very intensive part of the human effort expended in maintaining these complex systems is spent in monitoring every aspect of the Oracle Database, to ensure proactivity and short response time.

The IT PSS group has in place the Oracle Enterprise Monitor, richly configured and the primary monitoring tool for all the CERN clusters. In addition CERN maintains a central OEM repository that monitors the Tier-1s activity.

In addition there are CERN developed monitoring tools that are extremely useful in diagnosing problems and extracting statistics and trends to be used in estimates for production. The ones used by the database group (as there are others used internally by IT PSS) are:

- All Oracle OEM tools to monitor CERN databases and Tier-1s.
- Streams monitoring tool. This tool was developed by Zbigniew Baranowski from the LCG 3D group, to overcome a very deficient area in the Oracle tools. It has become an invaluable component of the Oracle Streams tests, ATLAS dashboard and resource estimation needed in ATLAS.
- Weekly database reports, issued by the IT PSS group, give an overview of the database load for the week, schema organization, storage occupation, and several alerts on performance which are very useful to take back to the developers.
- Application metadata Monitoring – built by the ATLAS DBAs, it gathers information on storage history, database usage and overall database activity sectioned by subdetector, application group or Oracle service alias.

5. Tests and Results

5.1. Replication Tests

It is the scope of the LCG 3D project to provide a production-level infrastructure for replicating the data between the ATLAS databases, namely to the 10 Tier1s that receive the replicated data. In addition, it is the task of the ATLAS DBAs to setup the ATLAS applications, synchronize them where they are needed and monitor for failure and specific errors related to the nature of the data.

The Oracle Streams infrastructure has been setup last year and continually tested as the Tier-1s got their architecture and resources in place to receive the data. Several kinds of replication tests have been made and were crucial in determining limits, drawbacks and caveats. Overall the Oracle Streams technology has evolved greatly, as a work of collaboration between the Oracle Corporation, the IT LCG 3D and the experiments' DBAs. ATLAS is the most resource and feature demanding experiment to the Streams technology, compared with the other experiments.

The tests executed on this topic can be summarized as follows:

- LCG 3D simulated table replication for testing connections between nodes (Online, Offline and Tier-1s). These were always the first tests to be made when a new site was able to join the replication data flow.
- ATLAS TAGS tables, with the nominal rates, used as limit-testers of throughput to the Tier-1s as they became available. Used also as baseline for assessing differences between T1 connections and for pushing performance analysis and improvements to the LCG 3D team and Oracle Corporation developers. In the end it was decided that the amount of data was too high to use the Oracle Streams technology, especially on shipping to the American continent.
- ATLAS COOL tests, Full stream replication has been setup between online, offline and the 10 T1s. COOL activity is challenging for the technology because it has a high transaction consistency dependency. Oracle Streams was able to cope with the rate intended (2GB per day). Most problems have been solved, some open issues remain and the tests are ongoing.

5.2. Conditions Data Tests

Tests on conditions data have been ongoing, as more "pieces of the puzzle" fall into place. For the conditions database the pieces are as follows:

- COOL replication to Tier-1s, described above in 5.1[8]
- Client Stress tests at Tier-1s. IN2P3 and CNAF have volunteered resources for staging these tests. The first round of tests uncovered some COOL software inefficiencies which are now tackled. Results were very encouraging, as the Tier-1s were able to sustain between 1500 to 2500 jobs/hour (the requirement is 400-4000 jobs/hour). New round is scheduled with optimized software, so better numbers are expected.[9]
- DCS data PVSS2COOL program, extracting data from online PVSS archive to offline. There are two replication paths from the online database to the offline database, and for PVSS2COOL this involves processes that periodically retrieve data from the online PVSS archive and transforms it in a COOL format for updating the offline DCS COOL schema. It is important that this is tested in a synchronous fashion with the Oracle Streams replication, to determine the optimal architecture in the Oracle cluster, for distributing load. This test will occur probably during the M4 run (Commissioning Cosmic Run) scheduled for the 23rd August.
- M4 test will input data from a subset of subdetectors, and test the full chain of the data flow, from the online database to the Tier-1 centres. It will be the first test with real data to test all the components together.

5.3. Event Metadata Tests

For the event metadata data, aka TAGS, testing has intensified last year, and will continue as the database schema and performance issues are refined. The tests executed so far are:

- Tag files loading into collections test, with a partitioned architecture and a first-attempt at an optimized reading environment. This was part of the Tier-0 tests last year, and it was a first step into having a TAGS database of 20Gb for testing. The test was very useful and triggered some improvements for the POOL utilities and food for thought for the schema architecture. The 200Hz loading rate was achieved.
- Replication tests to Tier-1s, detailed in 5.1
- Construction of a 10^9 rows database, with a simulation of TAG collections, with different architectures and different access modes tested. For this test a full dedicated cluster was assigned and measurements taken. It was extremely useful for discovering limits and shortcomings on access paths and schema organization. Rules for SQL structure were taken from this test, and are being studied as to their incorporation in the use cases for TAG collection retrieval.[10]
- Hardware resource stress tests. As a result of the above test, and using the same amount of data, scripts with the optimized SQL were run against one of our test clusters. The scripts were made to emulate the activity of a user, and executed different numbers of queries per minute. Precise measurements where taken on resouce (CPU and I/O) consumption that validated the resources requested for Tier-0 and Tier-1s that want to host this database.

6. Human Effort and Collaboration

6.1. Interaction between groups

The collaboration between the different projects involved in the database area has to be seamlessly integrated for such a complex computing environment to work. The key to success in building and maintaining such a system is upholding the following principles:

- Each of the groups has clear and defined tasks and boundaries of action.

- Each of the groups has defined contacts and protocols of interaction.
- When interests overlap, sharing of knowledge and working together to find solutions is the norm.
- Everyone is free to contact one another to seek advice and informal meeting are encouraged for prevention of future problems and for knowledge sharing.

In the database area, we have the following groups and modes of interaction defined:

- The IT PSS group hosts the databases and manages all aspects related to maintenance of hardware, software, backup and recovery of data. They are the primary DBAs of CERN's RDBMS online and offline.
- The IT LCG 3D project coordinates the database data replication effort, and all aspects related to the Oracle Streams software. They are the primary point of contact between the Tier-1s' teams and CERN PSS DBAs and experiment DBAs.
- The IT development and support teams, which develop COOL and POOL, and IT-controls, namely JCOP are points of contact for experiment developers and the database coordination project, for feedback on optimization and software support
- The experiment developers are organized in well defined groups of applications, and have direct contact with the ATLAS database coordination, and the ATLAS DBAs for all aspects related to the ATLAS databases.
- The ATLAS database coordination project oversees all the aspects related with ATLAS databases. The project steers development, resource allocation and milestone achievement within the ATLAS database scope.
- The ATLAS DBAs operate within the coordination project to implement management and monitoring procedures, assist the ATLAS developers on all matters related to the databases, coordinate efforts with the IT PSS group in user management and application architecture, work with IT LCG 3D in managing the Oracle Streams environment and the experiment replication needs, interact with the ATLAS Integration group for PVSS matters and feedback to subdetectors and JCOP, interact with IT development of COOL and POOL for monitoring tests and finding efficiency issues and solutions.

This mode of collaboration between interested parties has been extremely beneficial for ATLAS, for it is the experiment with the largest and most complex data handling issues. It is usually ATLAS who drives the technology limits and pushes the collaboration, described above, to the best use of technology and the improvement of many experimental features on the Oracle software.

6.2. Error Prevention and Optimization Effort

There is a constant drive in the ATLAS database team to open all communication channels between the groups to anticipate and resolve as many problems as possible before the applications go into full production mode with real data.

For this the ATLAS experiment has two dedicated two-node clusters with the same type of Oracle database software, maintained in synch, as the online and offline databases. In these clusters the applications that are developed in-house at another development database, are installed for extensive testing with a large capacity storage, for performance measurements.

This is a very positive step for the application developers, as they have full access to the monitoring tools that the DBAs have, and so can adequately screen their applications for usage patterns, problem spots on the schemas and on the SQL used. At the same time lively informal discussion is encouraged with the ATLAS DBAs on finding solutions for architecture problems within the applications.

This effort has been a success within ATLAS as it optimizes the usage of the IT PSS service, at hardware and human resource levels. It also brings a quicker response-time for task-handling and knowledge-sharing within the ATLAS developers community.

6.3. Service Level and Response to Failure

The service level agreements are in place for the ATLAS databases, from the IT PSS service.[7] The databases require very high availability and response times and shift schedules are being detailed now by the experiment and the IT PSS service.

For the protection of the databases, backup and recovery policies were proposed by the IT PSS service and refined together with the ATLAS database coordination. Feedback was taken from the ATLAS community, and special cases are being addressed with extra tools, outside of the official policies.

There is also a backup and recovery policy in place for data failure in the replication path, involving all Tier-1s. At the recent LCG 3D DBA Workshop, the Tier-1 DBA teams got together and exercised the backup and recovery policies in the live system, together with the IT PSS DBAs. This way each party is sure to have the knowledge required and at the same time the procedures are standardized.

The ATLAS backup and recovery policy at Tier-0 level, has been tested recently with a live application and that some additional protective measures were found lacking for the application in question. This has triggered an effort to find more applications that might need additional data protection measures, and to sort out the best way to put them in place.

7. Conclusions

With less than a year to go for the startup of the LHC, with real data to analyze, the ATLAS database project team is confident that the requirements of the ATLAS computing model regarding databases will be met. Much effort has been put into making this a reality: application deployment is progressing at a fast rate, the technology is very well known by now and properly tested and all pieces are being put in their place to start the data taking activity.

The Conditions Database and the Event Metadata will take their place centre-stage in the upcoming startup and all of the human resources involved will be ready to maximize the data analysis throughput of these systems.

8. References

[1] Atlas Computing Group 2005 *Computing Technical Design Report - TDR* (CERN: LHCC-2005-022 ISBN 92-9083-250-9)

[2] Hawkings R 2006 *ATLAS Conditions database data volumes and workloads (*http://atlas.web.cern.ch/Atlas/GROUPS/DATABASE/project/conddb/)

[3] Assamagan, K A et al. 2006 *Report of the Event Tag Review and Recommendation Group* (CERN:ATL-SOFT-PUB-2006-002)

[4] CERN ATLAS Twiki *Event TAG Infrastructure and Testing* (https://twiki.cern.ch/twiki/bin/view/Atlas/EventTagInfrastructureAndTesting)

[5] CERN IT PSS *Performance testing of a Quad-Core Server for the Physics Database Services* (https://twiki.cern.ch/twiki/pub/PSSGroup/HAandPerf/PSS_quadcore_tests_April07.pdf)

[6] CERN IT PSS 2006 *ORACLE database backup and recovery policy at Tier 0* (https://twiki.cern.ch/twiki/pub/PSSGroup/PhysicsDatabasesSection/ORACLE_backup_policy.doc)

[7] CERN IT PSS 2006 *Oracle physics database services support levels* (https://twiki.cern.ch/twiki/pub/PSSGroup/PhysicsDatabasesSection/service_levels.doc)

[8] Hawkings R 2007 *Conditions database status and developments* Atlas Software Week CERN (http://indico.cern.ch/getFile.py/access?contribId=125&sessionId=11&resId=1&materialId=slides&confId=9026)

[9] Vaniachine A 2007 *Database deployment, distribution and operation*, sl.16 ATLAS Overview Week Glasgow

(http://indico.cern.ch/getFile.py/access?contribId=126&sessionId=11&resId=1&materialId=slides&confId=9026)

[10] Goosens L. 2007 *TAGS Scalability and Performance Testing –Preliminary results from the 1B TAG test* Atlas Computing Workshop Munich (http://indico.cern.ch/getFile.py/access?contribId=109&sessionId=8&resId=1&materialId=slides&confId=5060)