Development, Deployment and Operations of ATLAS Databases

A V Vaniachine[1], **J G von der Schmitt**[2]

[1] Argonne National Laboratory, 9700 S Cass Ave, Argonne, IL, 60439, USA
[2] Max Planck Institute for Physics, Föhringer Ring 6, 80805 Munich, Germany

E-mail: vanachine@anl.gov

Abstract. In preparation for ATLAS data taking, a coordinated shift from development towards operations has occurred in ATLAS database activities. In addition to development and commissioning activities in databases, ATLAS is active in the development and deployment (in collaboration with the WLCG 3D project) of the tools that allow the worldwide distribution and installation of databases and related datasets, as well as the actual operation of this system on ATLAS multi-grid infrastructure. We describe development and commissioning of major ATLAS database applications for online and offline. We present the first scalability test results and ramp-up schedule over the initial LHC years of operations towards the nominal year of ATLAS running, when the database storage volumes are expected to reach 6.1 TB for the Tag DB and 1.0 TB for the Conditions DB. ATLAS database applications require robust operational infrastructure for data replication between online and offline at Tier-0, and for the distribution of the offline data to Tier-1 and Tier-2 computing centers. We describe ATLAS experience with Oracle Streams and other technologies for coordinated replication of databases in the framework of the WLCG 3D services.

1. Introduction

Since the previous CHEP in ATLAS database activities a coordinated shift from development towards operations has occurred, with Oracle technology was validated as a baseline for ATLAS Distributed Database operations. Figure 1 shows three tiers of ATLAS database architecture deployed in production operations: the centralized online and offline Oracle RAC clusters followed by the distributed clusters at all ten ATLAS Tier-1 computing centers worldwide.

The online Oracle RAC cluster is hosting databases that are used in the ATLAS online environment. Online databases store Trigger/DAQ and subdetector configuration data, calibration and alignment data, conditions data recorded by the Detector Control System (DCS), and bookkeeping data. For security reasons there are restrictions on the network access to the online Oracle cluster. To alleviate the impact of these restrictions (e.g. to enable remote access to real-time detector DCS data for experts at their home institutions) we replicate most of the online data to the offline server in real-time. Since Oracle was chosen as a baseline technology for online databases, uniform technology choice avoids translating from one technology to another and benefit from the support for Oracle from CERN IT and WLCG 3D project on Distributed deployment of databases (3D).

Technical details on ATLAS database applications are presented at this conference in contributions [1-12].

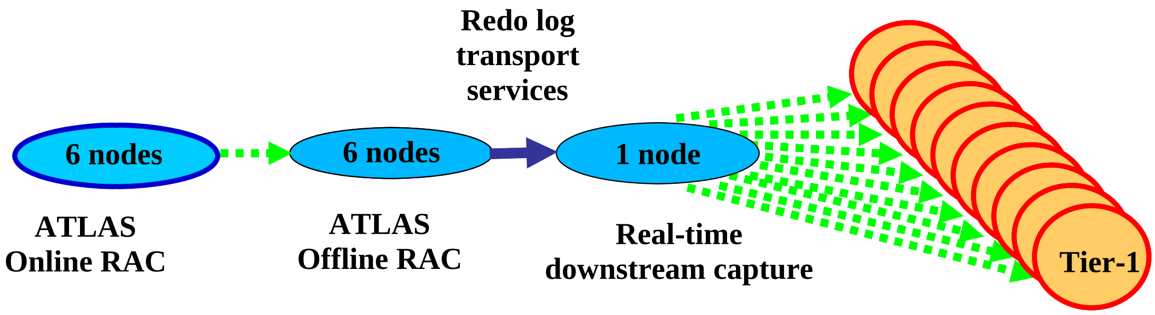


Figure 1. ATLAS database deployment architecture.

2. ATLAS database applications

In operations, ATLAS databases fall under two categories: distributed and centralized. To leverage the power of grid computing a subset of critical database applications (Geometry DB, Conditions DB and Tag DB) have to be distributed world-wide to eliminate potential bottleneck in database access,

The Geometry DB application was developed by ATLAS and was the first ATLAS database application deployed worldwide. That early operational experience was useful in providing a basis to determine ATLAS requirements for database capacities [13]. By nature, the Geometry DB data are mostly static and are now distributed world-wide in SQLite files (as a part of ATLAS Database Release).

Two other applications - Conditions DB and Tag DB - were developed as Common LHC components that have a longer development cycle and late deployment schedule. The Conditions DB was developed by LCG with ATLAS contributions and is now distributed world-wide via Oracle streams. The Conditions DB is now in production operations with real data from commissioning and data from simulations. The Tag DB – the event-level metadata for physics – was developed by LCG with ATLAS contributions. The Tag DB is to be distributed world-wide in files. The Tag DB is now in central operations for large-scale trials of Event Data streaming models. Technical details on these database applications are covered in CHEP contributions [3], [6], and [7].

Those database applications that are not accessed by many computers on the grid are only deployed centrally (at CERN or select Tier-1 centers). The following database applications are accessed interactively by people or by limited number of computers:

- Technical Coordination databases: detector construction information, etc.
- Online databases: DCS data stored in the PVSS DB, Trigger DB, and online databases for subsystems;
- Computing Operations databases: task request database (physics tasks definitions for job submission), production system DB (job configuration and job completion), Distributed Data Management (DDM) databases: (Dashboard DB, etc.);
- AMI (ATLAS Metadata Information) database [10].

However, centralized databases must also be replicated (for reasons other than scalability). These applications present different challenges for distributed operations, where the database replication technology of choice is Oracle streams.

One example of such challenge was presented by the PVSS DB. The value of the PVSS data is in their use for detector diagnostics and data quality assessment. But the online server must not be overloaded with such diagnostic queries. Also, it is behind the firewall - the sub-detector experts at their home institutions have no access to it. That is why we now replicate the PVSS and other subdetector databases to the offline server in real-time. In case of PVSS database it has been challenging. And not only due to the large volume of data involved - up to 6 GB/day have to be replicated. Unlike the ATLAS sub-detector databases, PVSS is a commercial product that has not been designed with support for replication. ATLAS database administrators invested considerable efforts to put the PVSS replication into production. Since this is not an ATLAS-specific problem we will share ATLAS PVSS replication solution with other LHC experiments.

3. ATLAS Conditions database replication

3.1. Conditions database challenges

A critical application for ATLAS database operations is the Conditions database. The Conditions DB manages conditions and calibration data needed for reconstruction and analysis. The Conditions DB is also a challenging database application because of the unprecedented complexity of LHC detectors (orders of magnitude more channels than in previous experiments). To address these challenge ATLAS adopted Common LHC technology for Conditions DB called COOL [14]. In COOL architecture the conditions data are usually characterized by the Interval-of-Validity (IOV) metadata and a data payload, with an optional version tag. Separation of the IOV and payload allows storage of the select "payload" data in files outside of the database server. An example of such data is calorimeter calibration constants that are not really suited to relational databases because of their size and access patterns.

The COOL technology is successfully deployed in production for ATLAS detector commissioning operations, such as major detector commissioning exercise – the M4 Cosmics Run in August of 2007, when the number of COOL accesses varied between 8K and 34K sessions per day.

A scale of ATLAS Conditions DB production operations can be represented by their current snapshot: more than 70 GB of data are stored. Each day up to 1 GB of COOL data is added and replicated. The data growth rates are increasing as more subdetector elements are being instrumented during ATLAS detector commissioning.

Access to Conditions DB data is critical for event data reconstruction. To achieve scalability in Tier-0 operations slices of the corresponding conditions/calibrations data will be delivered to Tier-0 farm via files on afs. Beyond Tier-0 we need a different technology for data distribution.

3.2. Distributed database deployment

For the replication of the slow-varying data in files, such as 'static' Database Releases (Geometry DB, etc.) and for replication of large files with Conditions DB payload we use ATLAS distributed data management technology DQ2 for file-based database data replication. For the replication of fast-varying data such as Conditions DB data ATLAS is using 3D Oracle streams to Tier-1 sites. The critical operational infrastructure for these technologies is delivered by the ATLAS DDM operations [2] and the WLCG Project on Distributed Deployment of Databases (3D) [15].

Leveraging the 3D Project infrastructure, ATLAS Conditions DB worldwide replication is now in production with real data (from detector commissioning) and data from simulations. Figure 2 shows the Snapshot of real-time monitoring of 3D operations on EGEE Dashboard.

Table 1. ATLAS Conditions DB workload used in the replications and scalability tests.

Schema	Folders	Channels	Channel payload	N/Run	Total (GB)	Used in re-construction
INDET	2	32	160 char	1	0.21	yes
CALO	17	32	160 char	1	1.8	yes
MDT	1+1	1174	CLOB: 3kB+4.5kB	0.1	17.0	yes
GLOBAL	1	50	3 x float	6	0.25	yes
TDAQ/DCS	10+5	200+1000	25 x float	12	80.0	yes
TRIGGER	1	1000	25 x float	12	8.0	no

Using 3D infrastructure ATLAS is running a 'mini Calibration Data Challenge' with regular conditions data updates on the Online RAC, testing propagation to the Offline RAC and further to ten

Tier-1s. Since April more than 2500 runs and 110 GB of COOL data were replicated to Tier-1s sites at rates of 1.9 GB/day. Table 1 presents the realistic Conditions DB workload using multiple COOL schemas with mixed amount and types of data that was used in ATLAS tests.

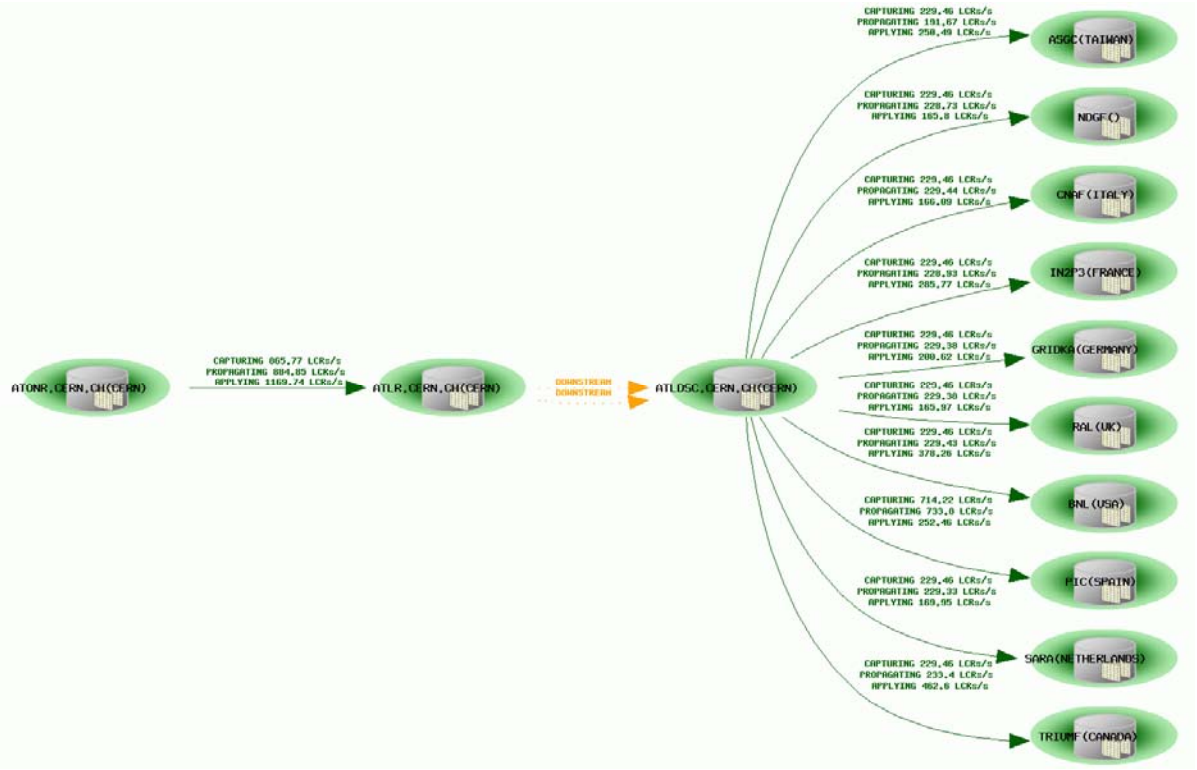


Figure 2. All ten ATLAS Tier-1 sites in production operations. .

These successful ATLAS operations contributed to a major WLCG milestone to deploy Conditions DB in operations and to test it by the LHC experiments. ATLAS data replicated by 3D project operations are used for reading back conditions data at the Tier-1s. This challenging deployment was done in a close collaboration with WLCG project on Distributed Deployment of Databases (3D). Together, we build one of the largest distributed database systems worldwide.

4. Oracle scalability tests

Reasons for building such large distributed database system are outlined in ATLAS Computing Model [16]. The Model provides following requirements at Tier-1 with respect to Conditions DB:

- running reconstruction re-processing: O(100) jobs in parallel;
- catering for other ‘live’ Conditions DB usage at the Tier-1 (Calibration and Analysis), and perhaps for the associated Tier-2/3s.

A critical issue in database operations is provisioning of the capacities. We deployed Oracle world-wide, but will the deployed hardware capacities be sufficient to support massive re-processing (the second-pass reconstruction) at the Tier-1s?

To assure enough capacities, at Tier-0 ATLAS uses file-based Conditions DB slice, but at Tier-1 DB access differs. Because of rate considerations we may have to stage and process all files grouped by physical tapes, rather than datasets. Such data access mode may reduce the value of the database data caching during the Tier-1 second-pass reconstruction. That is why ATLAS scalability tests do not rely on data caching and test the random data access pattern.

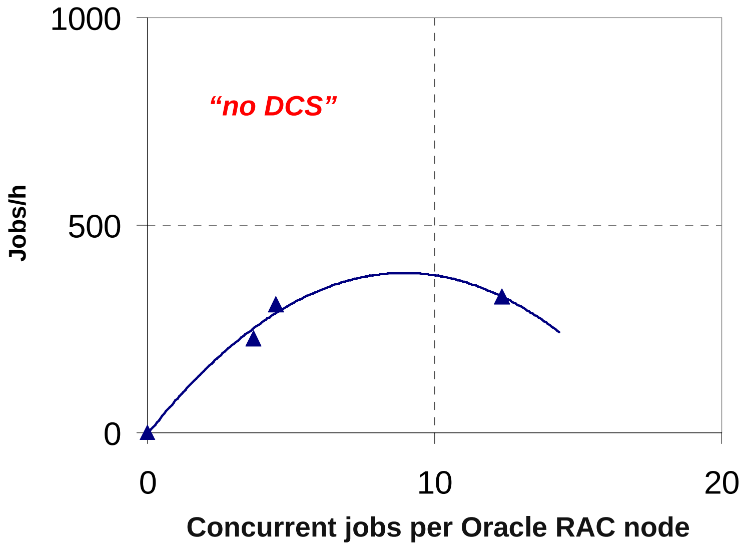


Figure 3. Calculated reconstructions jobs throughput vs. the number of concurrent jobs accessing Oracle database. In that particular plot the overload was caused by lack of optimization in the COOL 2.1 version that was used in the very first test. It is a first demonstration that our testing approach works.

To provide input to future hardware purchases for Tier-1s (how many servers required, balance between CPU, memory and disk, etc.) we performed Oracle scalability tests with the following considerations:

- although reconstruction jobs last for hours, most conditions data is read at initialization;
- staggered submission of jobs eliminates the need to initialize O(100) jobs at once.

The Conditions DB data replicated in 3D operations provided read-back data for ATLAS scalability tests. The top five rows in table 1 represents the 'best guess' of the Conditions DB data needed for initialization of ATLAS reconstruction jobs. The total data volume is dominated by the DCS data. To probe the range of realistic conditions data workload, three workload combinations were used in the tests:

- **"no DCS" -** Data from 19 folders of 32 channels each, POOL reference (string) payload, plus 2 large folders each with 1174 channels, one with 3k string per channel, one with 4.5k string per channel, which gets treated by Oracle as a CLOB, plus 1 folder with 50 channels simulating detector status information. This is meant to represent a reconstruction job running reading calibration but no DCS data. The data is read once per run.
- **"with DCS"** - As above, but an additional 10 folders with 200 channels each containing 25 floats, and 5 folders with 1000 channels of 25 floats, representing some DCS data, again read once per run.
- **"10xDCS" -** As above, but processing 10 events spaced in time so that all the DCS data is read again for each event. This represents a situation where the DCS data varies over the course of a run, so each job has to read in 10 separate sets of DCS data.

First ATLAS scalability tests started at the CC IN2P3 Tier-1 site at Lyon, France. Lyon has a 3-node 64-bit Solaris RAC cluster which is shared with another LHC experiment (LHCb). In scalability tests our goal is to overload the database cluster by launching many jobs at parallel. Figure 3 shows that initially, the more concurrent jobs is running (horizontal axis) – the more processing throughput we will get (vertical axis), until the server became overloaded, when it takes more time to retrieve the data, which limits the throughput.

Figure 4 shows the results of the scalability tests with COOL 2.2 at the Tier-1 site CNAF in Bologna, Italy. In contract to Lyon, CNAF deployed a 2-node dual-CPU Linux Oracle RAC which is dedicated to ATLAS and is more representative of other Tier-1 sites. The baseline "with DCS" test achieved the throughput approaching 2,000 jobs per hour. Further COOL optimization is expected to provide some increase in performance, since queries for multi-version folders are not optimized in COOL 2.2.Our scalability tests with COOL 2.2 at the Tier-1 site CC IN2P3 in Lyon, France show the importance of doing tests at several sites, because the tests resulted in a lower throughput than in the CNAF tests (figure 4).

We estimate that during LHC operations ATLAS daily reconstruction and analysis jobs rates will be in the range from 100,000 to 1,000,000 jobs/day. (In current ATLAS production operations finishes up to 55,000 jobs/day.) For each of ten Tier-1 centers that corresponds to 400 to 4,000 jobs/hour. For many Tier-1s pledging ~5% capacities (vs. 1/10th of the capacities) that would correspond to the rates of 200 to 2,000 jobs/hour. Note that most of these will be analysis or simulation jobs which do not need so much Oracle Conditions DB access. Thus, our results from the initial scalability tests are promising. We got initial confirmation that ATLAS capacities request for deployment on WLCG (3-node clusters at all Tier-1s) is close to what will be needed for reprocessing in the first year of ATLAS operations.

We plan to continue scalability tests with new COOL releases. These tests will result in more precise determination of the actual ATLAS requirements for database capacities in operations. Having this information we will complete the next iteration on the ATLAS requirements for Oracle capacities at Tier-1 that will be aligned with other ATLAS computing resources at Tier-1, such as raw data storage fraction. Indeed, each Tier-1 will store 1/10th of the raw data on average. However, actual ATLAS computing resources at Tier-1 currently vary from 4% to 23%. These variations have to be matched in Oracle CPU count. Also, grouping files on tapes by the datasets (vs. in an arbitrary order) may reduce the Tier-1 requirements for Oracle database capacities during reprocessing. However, we will avoid building our system with a constraint that forces data to be processed linearly.

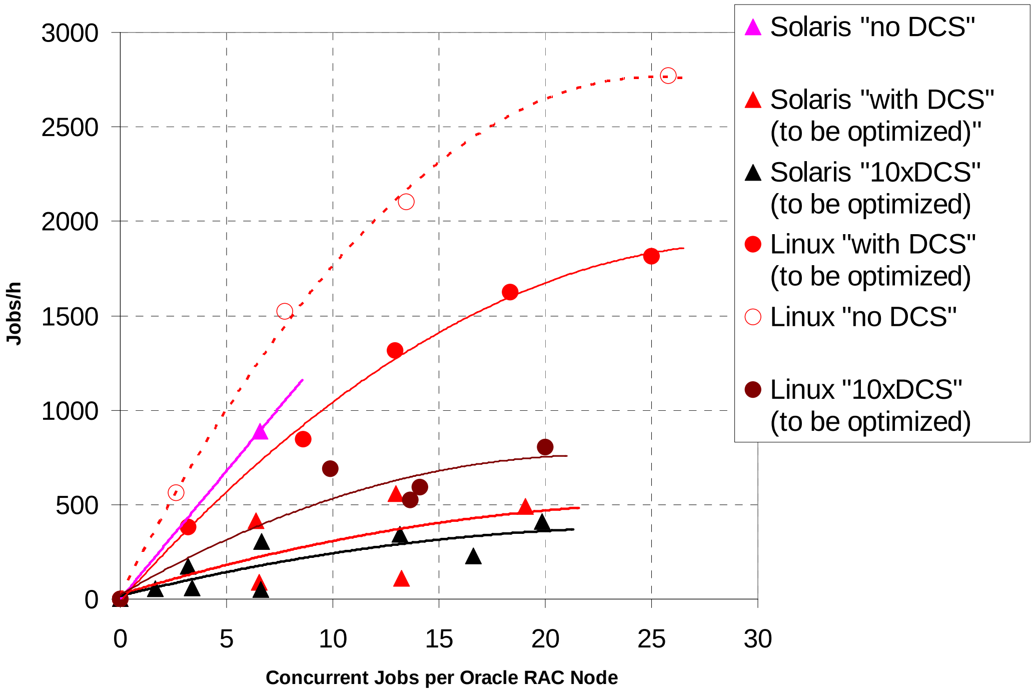


Figure 4. Calculated reconstructions jobs throughput vs. the number of concurrent jobs accessing Linux clusters and Solaris clusters.

5. Work-in-progress

5.1. *Ramping up database capacities for ATLAS*

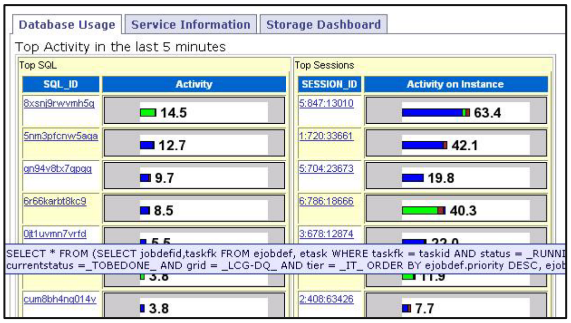
Table 2 presents ramp-up schedule over the initial LHC years of operations towards the nominal year of ATLAS running, when the database storage volumes are expected to reach 6.1 TB for the Tag DB and 1.0 TB for the Conditions DB. We are pleased that CERN IT is on track to deliver our request in advance of thee start of ATLAS data taking.

Table 2. Expected growth of ATLAS Oracle data volumes at Tier-0.

Year	Total (TB)	PVSS online (TB)	PVSS offline (TB)	TAG (TB)	COOL (TB)	DDM Dashboard (TB)
2007	6.3	2	2	1.0	0.3	1
2008	9.1	3	3	1.6	0.5	1
2009	14	3	3	6.1	1.0	1

5.2. *Replication of Conditions DB Payload Data to Tier-1s.*

Common LHC software used by ATLAS for data access is technology neutral. Currently, two common LHC interfaces are implemented. The first interface is POOL for access to ROOT files, which is used for ATLAS Event Store. The second interface is CORAL, which is used for access to Relational Databases. CORAL supports not only server-based data access to Oracle, MySQL, and squid/FroNTier servers, but also access to the file-based relations database SQLite. In ATLAS the file-based relational database technology is used to achieve scalability for slow-varying data, needed by every data processing job such as the Geometry DB. Such data are now packaged in the Database Release (decoupled from the Software Release) and distributed to all sites in an automated way by ATLAS DDM operations. Also, certain fast-varying Conditions DB payload data are stored in POOL/ROOT files. Automatic replication of Conditions DB payload files started by ATLAS DDM operations. Since the Conditions DB files should arrive before the event data files, the priority queues for the DDM file transfers has being implemented.



Top Activity in the last 5 minutes

Top SQL

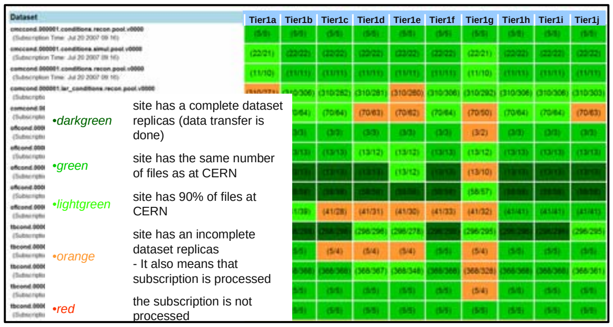
SQL_ID	Activity
8xsnj9rwvmh5q	14.5
5nm3pfcnw5aqa	12.7
gn94v8tx7qpqg	9.7
6r66karbt8kc9	8.5
0jt1uvmn7vrfd	5.5
	3.8
cum8bh4nq014v	3.8

Top Sessions

SESSION_ID	Activity on Instance
5:847:13010	63.4
1:720:33661	42.1
5:704:23673	19.8
6:786:18666	40.3
3:678:12874	22.0
	11.9
2:408:63426	7.7

```
SELECT * FROM (SELECT jobdefid,taskfk FROM ejobdef, etask WHERE taskfk = taskid AND status = _RUNNI
currentstatus =_TOBEDONE_ AND grid = _LCG-DQ_ AND tier = _IT_ ORDER BY ejobdef.priority DESC, ejob
```

Figure 5. Monitoring of ATLAS database services on Oracle RAC.



•darkgreen	site has a complete dataset replicas (data transfer is done)
•green	site has the same number of files as at CERN
•lightgreen	site has 90% of files at CERN
•orange	site has an incomplete dataset replicas - It also means that subscription is processed
•red	the subscription is not processed

Figure 6. Monitoring replications of Conditions DB datasets with 'payload' in POOL files.

5.3. *Database operations monitoring*

ATLAS database applications require robust operational infrastructure for data replication between online and offline at Tier-0, and for the distribution of the offline data to Tier-1 and Tier-2 computing centers. Monitoring is critical to accomplish that. Figures 5 and 6 show examples of ATLAS database operations monitoring in place.

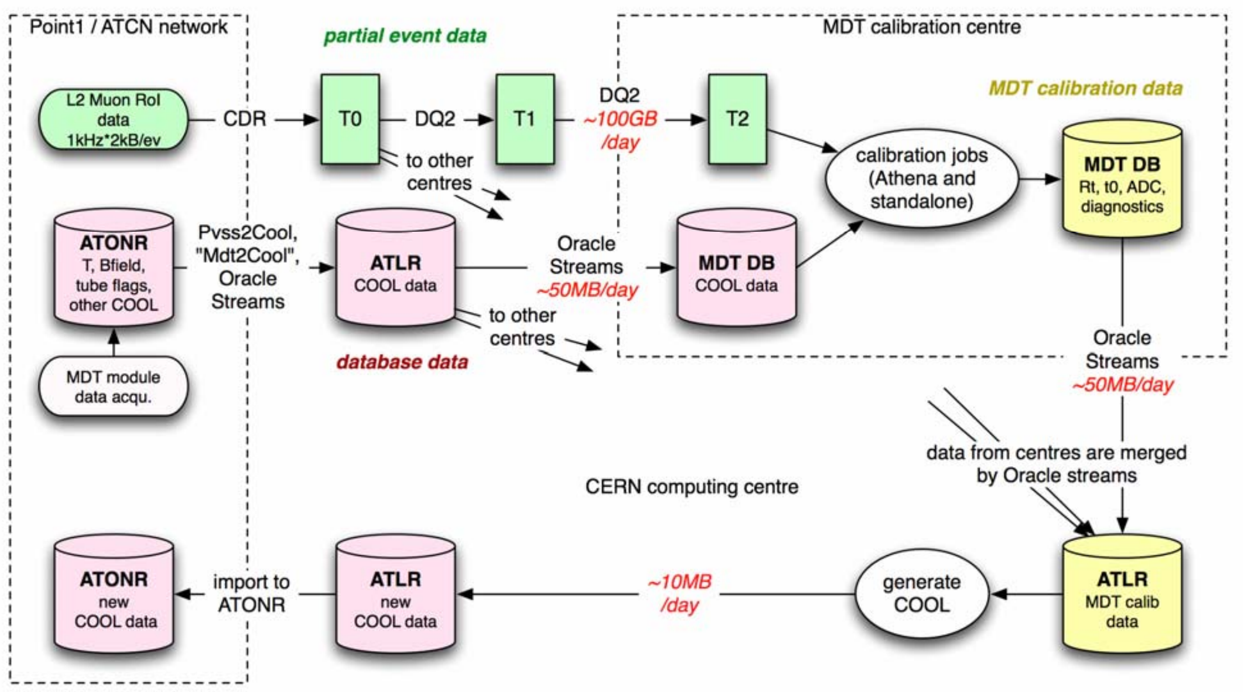


Figure 7. Database operations needed for muon calibrations.

5.4. Towards muon calibration operations

ATLAS subdetectors such as Muon System will use the 3D Project database infrastructure for their own calibration databases, e.g. Monitored Drift Tubes (MDT). For that Oracle servers are installed at the MDT sites: Michigan, Rome, and Munich. The replication via Oracle streams tested from Michigan to CERN. Figure 7 shows databases involved in muon calibration operations.

6. Conclusions

In preparation for ATLAS data taking in ATLAS database activities a coordinated shift from development towards operations has occurred. In addition to development and commissioning activities in databases, ATLAS is active in the development and deployment of the tools that allow the worldwide distribution and installation of databases and related datasets, as well as the actual operation of this system on ATLAS multi-grid infrastructure. In collaboration with WLCG 3D Project a major milestone accomplished in database deployment: ATLAS Conditions DB is now in production operations with real data on one of the largest distributed database systems world-wide. First Oracle scalability tests indicate that WLCG 3D capacities in deployment for ATLAS are in the ballpark of what ATLAS requires. Future scalability tests will allow more precise determination of the actual ATLAS requirements for distributed database capacities

Acknowledgments

We wish to thank R. Hawkings, G. Dimitrov and F. Viegas and all our ATLAS collaborators who contributed to and facilitated ATLAS database development, deployment and operations.

Because of large allocation of dedicated resources, the scalability tests reported here required careful planning and coordination with Tier-1 sites, which volunteered to participate in these tests. We wish to thank to G. Rahal, J. R. Rouet, P. E. Macchi, E. Lancon and RD Schaffer for the Lyon tests and to B. Martelli, A. Italiano, L. dell'Agnello, L. Perini and D. Barberis for the Bologna tests.

This work supported in part by the U.S. Department of Energy, Division of High Energy Physics, under Contract DE-AC-02-06CH11357.

References

[1] Lassnig M et al. 2007 *Managing ATLAS data on a petabyte-scale with DQ2* CHEP’07 contribution id 64 (unpublished)

[2] Klimentov A et al. 2007 *ATLAS Distributed Data Management Operations. Experience and Projection* CHEP’07 contribution id 84 (unpublished)

[3] Nicholson C et al. *Integration of the ATLAS Tag Database with Data Management and Analysis Components* CHEP’07 contribution id 85 (unpublished)

[4] Verducci M *ATLAS Conditions Database Experience with the LCG COOL Conditions Database Project* CHEP’07 contribution id 90 (unpublished)

[5] Pommes K et al. 2007 *Glance Project: a database retrieval mechanism for the ATLAS detector* CHEP’07 contribution id 100 (unpublished)

[6] Viegas F et al. 2007 *Relational databases for conditions data and event selection in ATLAS* CHEP’07 contribution id 122 (unpublished)

[7] McGlone H et al. 2007 *Building a Scalable Event-Level Metadata System for ATLAS* CHEP’07 contribution id 161 (unpublished)

[8] Rocha R et al. 2007 *Monitoring the Atlas Distributed Data Management System* CHEP’07 contribution id 255 (unpublished)

[9] Amorim A et al. 2007 *Implementing a Modular Framework in a Conditions Database Explorer for ATLAS* CHEP’07 contribution id 333 (unpublished)

[10] Albrand S et al. 2007 *The ATLAS METADATA INTERFACE* CHEP’07 contribution id 430 (unpublished)

[11] Nevski P et al. 2007 *Steering of GRID production in ATLAS experiment* CHEP’07 contribution id 450 (unpublished)

[12] Cirilli M et al. 2007 *Database architecture for the calibration of ATLAS Monitored Drift Tube Chambers* CHEP’07 contribution id 462 (unpublished)

[13] Vaniachine A et al. 2006 *Database Access Patterns in ATLAS Computing Model* CHEP’06 contribution id 38 (unpublished)

[14] Valassi A 2007 *COOL Software Development and Service Deployment Status* CHEP’07 contribution id 204 (unpublished)

[15] Duellmann D 2007 *Production Experience with Distributed Deployment of Databases for the LHC Computing Grid* CHEP’07 contribution id 171 (unpublished)

[16] Jones R 2007 *The ATLAS Computing Model* CHEP’07 contribution id 200 (unpublished)