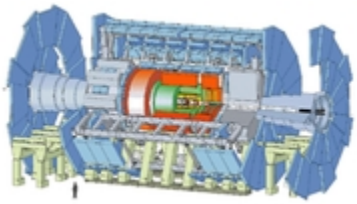
Big data processing and analysis workshop
Dubna, 30th Jan 2015

the ATLAS Experiment



Usage of ORACLE in ATLAS

Gancho Dimitrov (CERN)

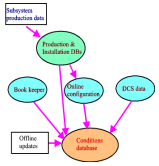
Outline

- ATLAS databases based on the ORACLE relational database management system

  1. Main roles
  2. Topology
  3. Data replication
  4. Applications, requirements, technical solutions

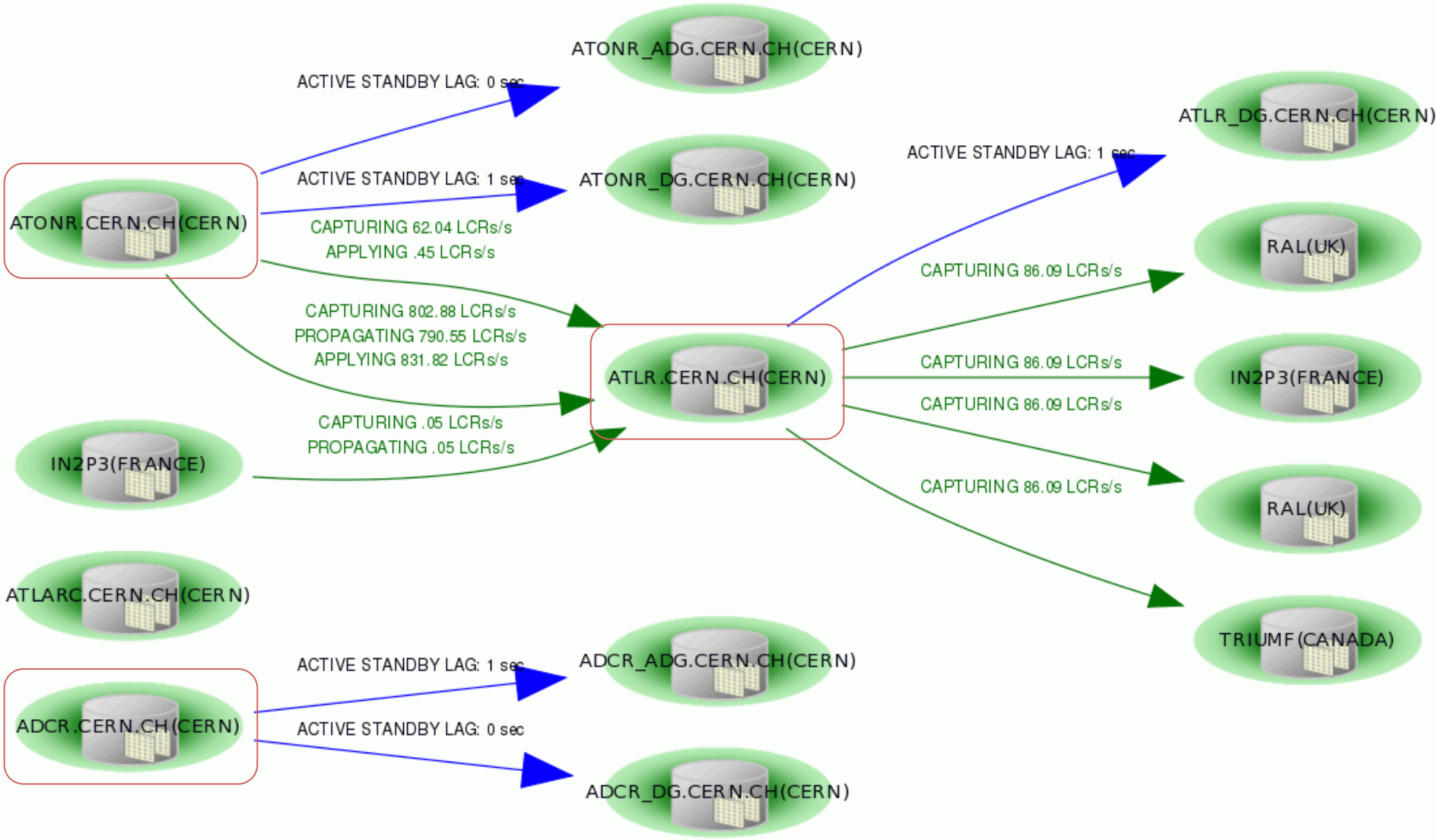
- Readiness for LHC Run 2

ATLAS production database clusters



Database / info	ATONR	ATLR	ADCR	ATLARC
Main database role	Online data taking	Post data taking analysis	Grid jobs and file management	Events metadata
Oracle version	11.2.0.4	11.2.0.4	11.2.0.4	11.2.0.4
# DB nodes	2	3	4	2
DB volume	10 TB	23.3 TB	25 TB	18 TB
# DB schemes	**74**	**165**	**16**	**55**
HW specs	CPU Intel E5-2650@ 2GHz - 16 cores per node RAM 128 or 256 GB (depends on the DB role) 10 GigE for storage and cluster access NetApp NAS storage with 1.5 TB SSD cache			

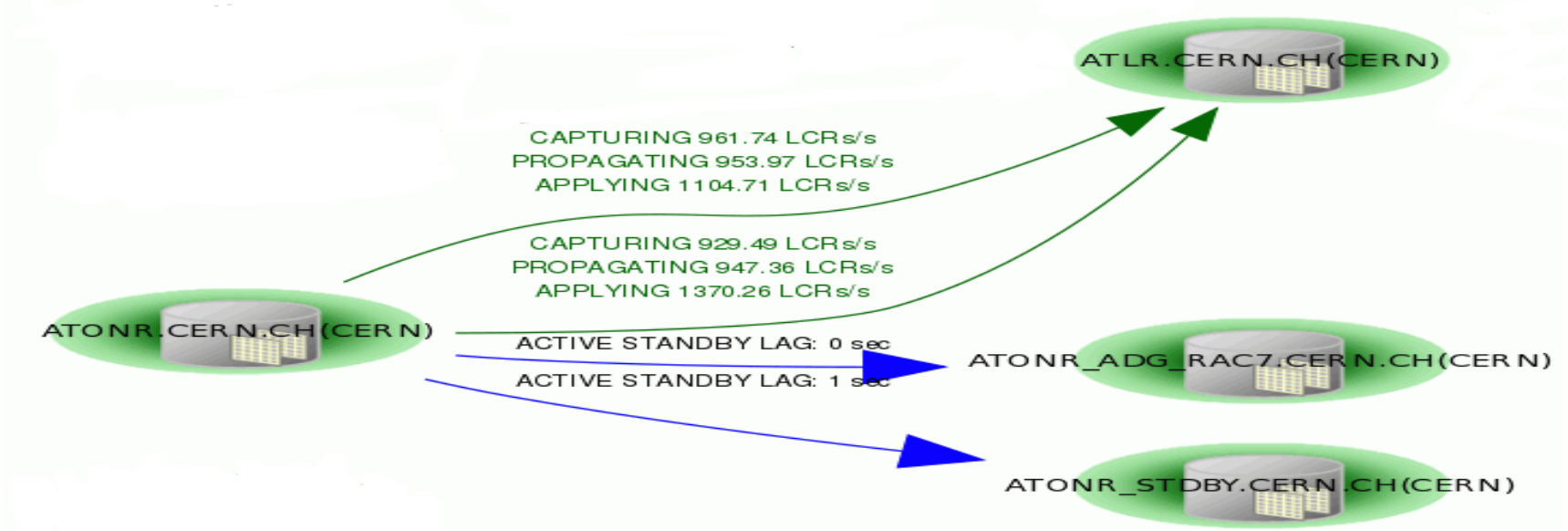
Oracle databases topology for ATLAS



ATONR database

ATONR database

- **Main role:** run configurations, trigger settings, detector control systems and conditions data, online data taking.
- **Workload type** : transactional
- **DB load:** typically low as the applications are well selected and with tuned SQL statements. Database is offloaded from heavy read activity as data is replicated to destination databases.
- **Data replication:** via Oracle Streams or GoldenGate technologies and use of a standby database Active Data Guard (ADG)

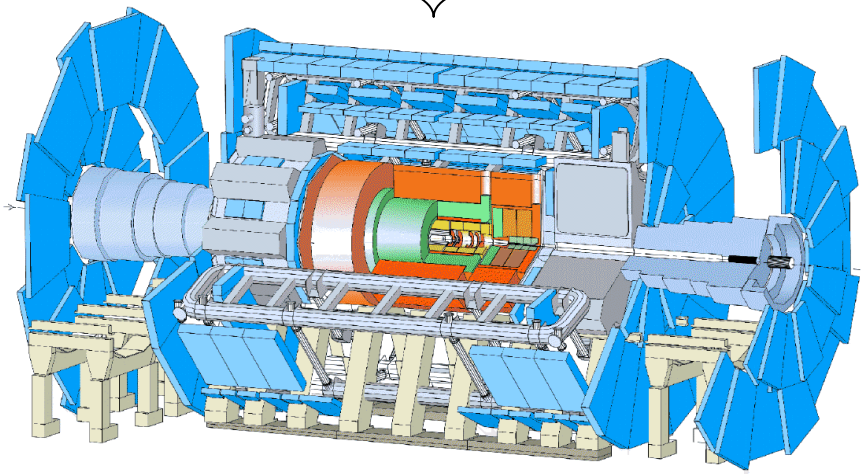


ATONR main application: PVSS

PVSS (Process Visualization and Steering System, now named WinCC Open Architecture) is a Control and Data Acquisition system, developed by the Austrian company ETM (now owned by Siemens AG).
Chosen in year 2000 as a control system for the LHC experiments.

PVSS by ETM

Thousands of data point elements



The ATLAS detector

PVSS Oracle archive -
keeps history of the detector status,
e.g. high voltages, temperatures

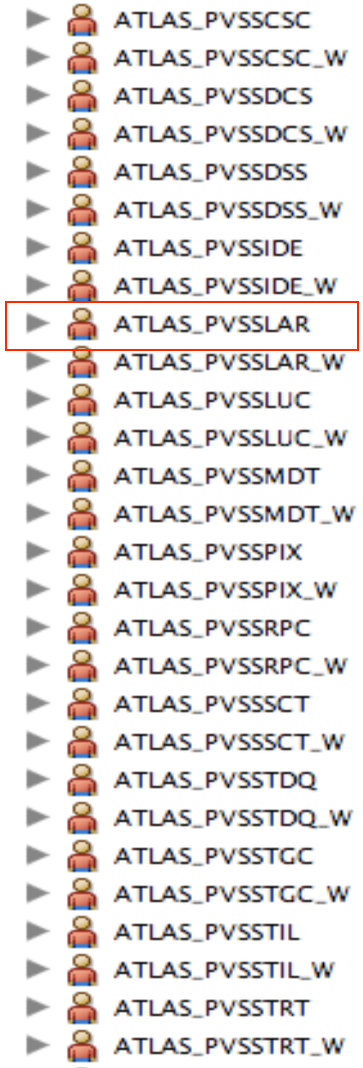
ATONR

ORACLE

The ATLAS 'online' Oracle DB

The ATLAS PVSS DB schemes and table descriptions

- A database schema per subdetector (as total 18)



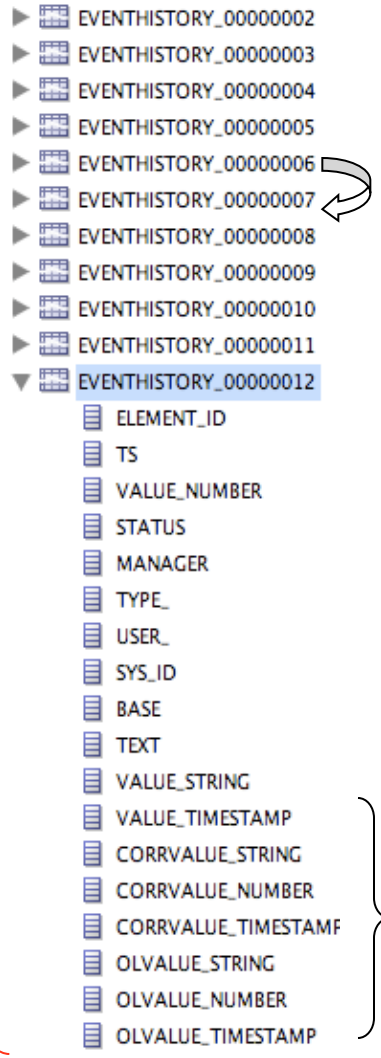


Table is 'switched' on defined time interval and a view object is updated to keep them together for the application to access the data (the EVENTHISTORY view)

Not used from ATLAS, get NULL values, thus do not take occupy space

The row length is in the range 55-60 bytes

PVSS insert rates and data volumes

- **Usual rows insert rate** is within the range 800-1200 rows/sec
- **On average used disk space per day** : 7 GB (table + index segments)
- **Current volumes:** 4TB data segments, 3.4 TB indices overhead
- **Policy on the PVSS data retention @ ATONR**:

  Keep the data for most recent 3 years and keep the replicated data on the ATLR database forever.

ATLR database

ATLR database

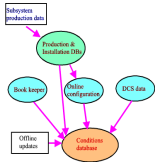
- **Main role:** post data taking analysis, detector conditions data based on intervals or validity or run numbers (COOL system), T0 express processing, Trigger data for the MC, ATLAS SW nightly build system, Detector descriptions (known as Geometry DB), Frontier.

- **Workload types** : transactional plus reporting and analysis

- **DB load:** typically low except on periods of T0 express processing. Other processing activity is served via the Frontier path thus offloading the database.

- **Replication:** conditions data replicated to IN2P3, RAL and TRIUMF T1 centers via Oracle's GoldenGate technology.

Insert rates and data volumes



- **Usual rows insert rate** is dominated by the PVSS data replication
- **On average used disk space per day** : 11 GB (table + index segments)
- **Current volumes:** 10 TB data segments, 13 TB indices segments
- **Policy on the PVSS data retention @ ATLR**:
  Keep complete historical data of the ATLAS detector conditions data

ADCR database

ADCR database

- **Main role:** database backend for the main ATLAS distributed computing applications: **PanDA** (Production and Distributed Analysis system) and **Rucio** (the ATLAS Data Management system),
  AGIS (ATLAS Grid Infrastructure Services)

- **Workload types** : transactional plus reports, analysis and data mining

- **DB load:** **moderate** except on periods of misbehaving application(s) or use of sub-optimal Oracle execution plans when serving user queries.

- **Replication:** **complete read-only copy of the ADCR database using a standby DB (Active Data Guard technology)**. Accounting, analytic queries with parallelism and data exports served by the ADG database cluster.

Specifics of the main ADC DB applications

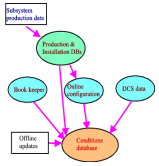
- PanDA system **generates two different workloads**: transactional (PanDA server) and data warehouse load (PanDA monitor).

- Data are logically split into **'operational' (**400 GB**)** and **'archive' (**8 TB**)** types and respectively stored into separate DB schemes.

- **Rucio's data are to high extent dynamic**. Large fraction of it is with short lifetime (transient data) due to the dynamic placement of files and datasets on the Grid.

  Sets a challenge for proper handling.

Useful DB monitoring metrics



- **Several Oracle metrics are used in our own alert system** for the DB instances where the applications run.
- **Defined thresholds** for sending alerts is useful for the DBAs and application owners knowing the specifics of the applications and the HW capabilities.

Metric_id = 2003, User Transaction Per Sec
Metric_id = 2030, Logical Reads Per Sec
Metric_id = 2106, SQL Service Response Time (in centiseconds)
Metric_id = 2058, Network Traffic Volume Per Sec (Bytes Per Second)
Metric_id = 2018, Logons Per Sec
Metric_id = 2147, Average Active Sessions
Metric_id = 2118, Process Limit %
Metric_id = 2119, Session Limit %
Metric_id = 2143, Session Count

- **Those proved to be handy for reacting quickly on applications misbehaviour or Oracle's sub-optimal query's execution plan**

Challenges in the ATLAS DB area

- **ATLAS databases have to serve two different workloads –** transactional (application data write activity) on one hand and data reporting and analysis on the other hand.
- **Spikes of workload have to be addressed in adequate manner.**
- **Emulate real Grid workload in test environment is very hard to achieve.**
- **Often DB tuning is done on a "moving target"** as user requirements are changing often.

Are we well positioned for LHC Run2?

- **We have better HW in place:**
  - 16 CPU cores per DB node
  - 128 or 256 RAM, larger data caches (data or index block = 8KB)
  - larger SSD cache on storage level (1.5 TB)

- **Significant changes into several DB schemes**
  - Use of partitioned on time index-organized tables for the PVSS systems. **Reduces twice the needed disk space** and orders the rows for faster and more efficient data retrieval.
  - New applications with **new tuned DB objects are introduced**
  - **Smarter way of data segmentation** and de-duplication on block level

- **Use of Active Data Guard databases for read-only access for a significant offload of the main databases** and a method of data replication (decommission Streams replication)

Still there is a room for improvements...

- **Proper use of connection pools on the middleware tier**
- **Rate of SQL statement executions to be justified, tuning based on the number or block reads (a block = 8KB) is the best approach.**
- **Avoid transactions serialization as much as possible.**
- **Distribute application modules on the nodes of the primary DBs or use the resources of the available Active Data Guards (ADG).**
- **Test, validate and use in production new DB features:**

  - Oracle 12c Database In-Memory Columnar store

  Dual format DB: Oracle supports row and column formats for the same table, simultaneously active and transactional consistent.

  Many orders of magnitude compression (depends on the data types and values). Analytics and reporting use new in-memory format.

Thank you for the interest!

Questions?