Database operations in CMS

Distributed Database Operations Workshop 2010

19 November 2010

Giacomo Govi (CERN)

On behalf of the CMS experiment

Outline

- CMS Data into RDBMS
- Database infrastructure
- Condition Database
- Luminosity data management
- Operations
- Planning
- Summary

CMS Data stored in RDBMS

- CMS operations involves a relevant number of services/activities
  - requiring data management of various nature
  - with different volumes, performance or scaling issues
- RDBMS is a key component for several applications
  - Online systems, offline reconstruction and analysis, workflow management, configuration management, authentication, monitoring, logging
  - strategic data: accessible by many applications
  - private data: internally used by a single application

Online data

Critical for the detector operation and the data taking:

- Detector configuration XDAQ, TStore
- Trigger configuration TStore
- DAQ
- Slow control PVSS+RDB
- Monitoring
- Storage manager

Most of the clients of the applications/services are devices (or humans) located at P5

- Data should be accessible to a reduced community

Special requirements:

- The system must be able to fully operate with no external network connection
- Private DB storage (cannot be shared with other systems)

Offline data

- Wide category, includes the data involved in the offline production activities
  - Calibration, detector condition
    - Varying with time and frequently updated
  - Detector description
    - Static (or quasi static)
  - Beam and luminosity information
  - Run information
- Critical for the physics data analysis chain:
  - Data are exposed to a large community
    - Many institutions of the collaboration involved
    - Potentially little control on volumes expected, technologies, standards, practices, access patterns

Other data

Related to some general services, critical for many activities of the experiment:

- File transfer management FedEx
  - Basically ‘tactical’ data, but highly transactional
- File bookkeeping DBS
  - Large volumes, write once…
- Jobs and file processing T0AST
  - Transactional
- Authentication/Authorization SiteDB
  - Quasi static read-only data

Infrastructure

2 production clusters:

1) CMSONR, 6 nodes Oracle RAC located at P5

- only 'visible' within the P5 network
- two logical databases:
  OMDS stores data for sub-detectors, trigger,
  slow control, luminosity, configuration, monitoring
  ORCON stores calibrations and other condition data.
- h/w from CMS, s/w + maintenance from CERN/IT

2) CMSR, 6 nodes Oracle RAC located at the IT center

- visible within the CERN network
- storage for condition, run, luminosity (ORCOFF)
- storage for workflow management data
- fully maintained by IT

+ Integration RAC: INT2R – visible from P5

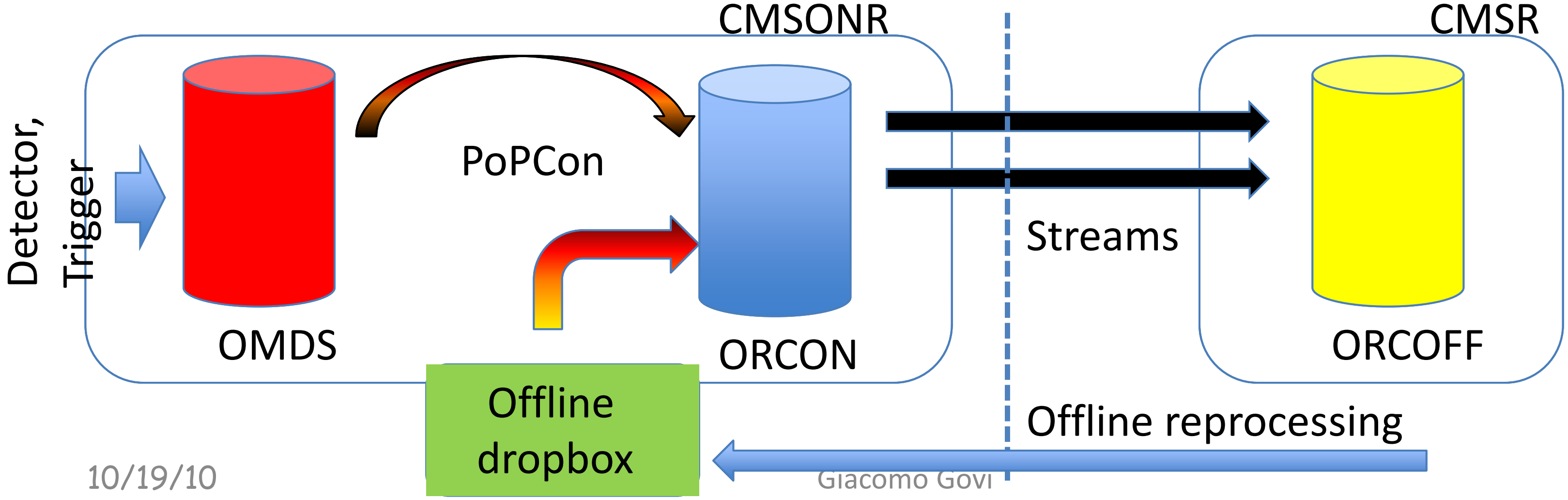
Data flow

A subset (summary) of condition data is read from OMDS, reformatted and stored in ORCOFF

- processes running in dedicated nodes at P5
- performed by a single application (PopCon)

2 Oracle streams populate ORCOFF with data from OMDS and ORCON:

1. ORCON + Luminosity + Storage Manager data
2. PVSS accounts and monitoring data from OMDS



Condition Database

Manages a large set of the categories listed before:

- required by the HLT, DQM and Offline reconstruction
- ~200 data sources feeding their tables
- managed by several institutions/groups within CMS
- with wide range of frequency and data volume
- accessed for reading by more than 15000 jobs/day (only from Tier0/1!)

Stability/performance require to limit the access patterns

- By policy: no delete, no update
- By the architecture:
  - write with a single application in ORCON (PopCon application)
  - stream the ORCON into ORCOFF, read ORCOFF

Severs overload in reading kept low with FronTier caches

- See Dave talk in next session

Population

PopCon (Populator of Condition object) application

- Package integrated in the CMSSW framework
- Stores/Retrieve data as C++ objects supporting more RDBMS technologies (Oracle, SQLite, FronTier)
- Depends on some externals from LCG/AA: Root, Coral
- Dependency on POOL recently dropped
- Writes in ORCON (O2O jobs and Offline dropbox)

Offline Dropbox

- Export automatically condition data processed offline into the ORCON database
  - Accessing the Oracle DB within the P5 private network
  - Files are pulled into the P5 network every 10 minutes
- Python application based on Http proxy

Condition data reading/distribution

The Offline Reconstruction jobs running in the Tier0/1 are potentially creating a massive load on ORCOFF.

- jobs from Tier0 and Tier1s (~15000)+ a variable subset of jobs from Tier2s (~30000?)
- 200 condition objects to read with 3-5 tables => ~800 queries
- volumes involved are modest: ~40-50 Gb for the entire DB
- data is read-only at large extent

- FronTier caches allow to minimize the direct access to ORACLE servers
  - At the price of a possible latency implied by the refreshing policy
  - 2 Frontier services implemented (ORCON to P5 and ORCOFF to Tier0/1)
- Snapshot from Oracle DB heavily used for reprocessing
  - Data set are exported in a dedicated server
- SQLite files provide the additional, simple way to ship data through the network
  - Used by the Offline DropBox to export calibration data into ORCON
  - Can be also used to ship MC data to Tier1's

Luminosity data management

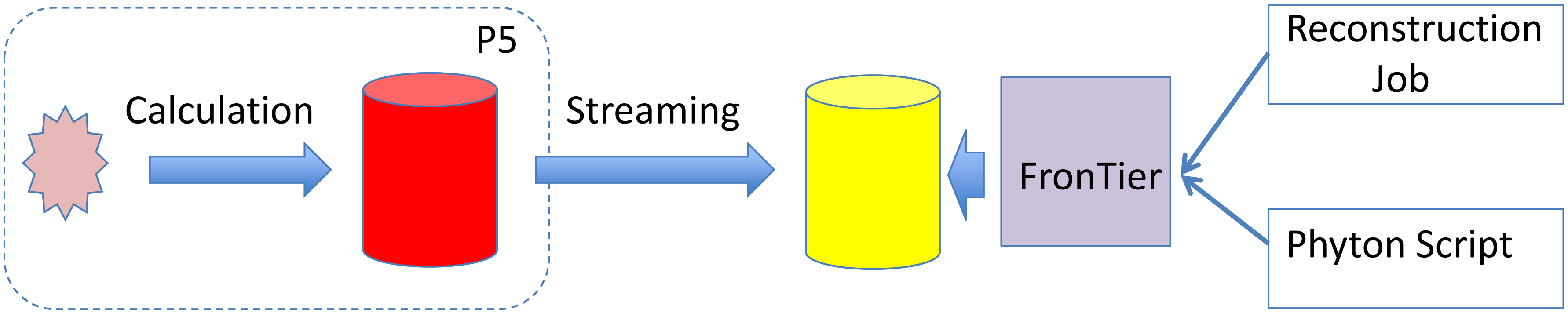
Luminosity data have been recently added in the CMS DBs.

Production flow

- Computed by processes running at P5
- Written in a dedicated account in OMDS
- Streamed to ORCOFF (Condition Stream)

Two main reading use cases

- Selected via FronTier and attached to reconstructed data in the jobs running at Tier0
- Selected via FronTier for user queries of specific runs (python script)



Space usage

CMSONR

OMDS

- ~300 accounts R/W, ~200 RO
- 2,5 Tb + 30 Gb/week
- Monitoring + PVSS largest contribution

ORCON

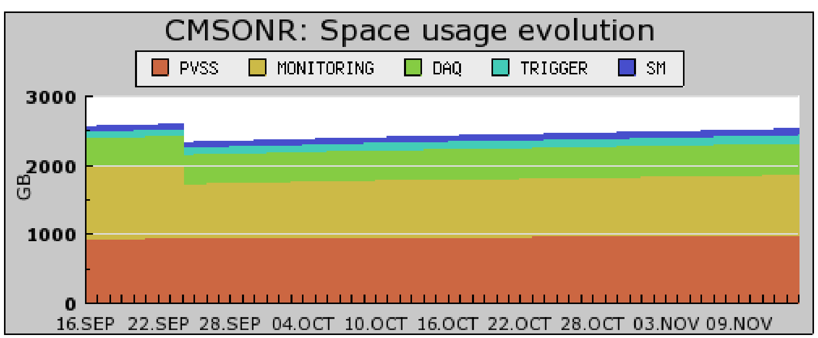
- ~ 200 accounts RW
- 50 Gb growing 1Gb/week

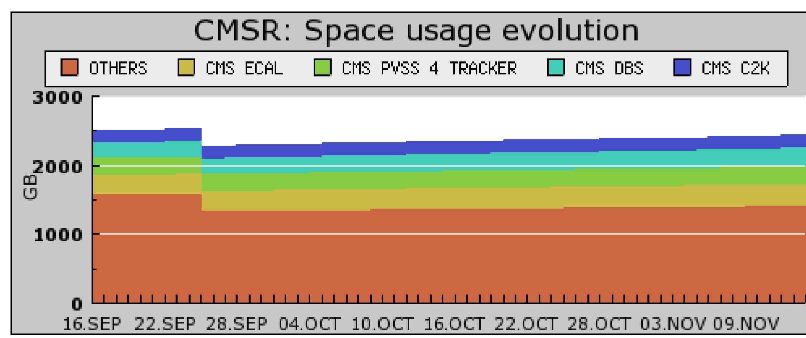
CMSR

- 2,4 Tb + 30Gb/week

ORCOFF

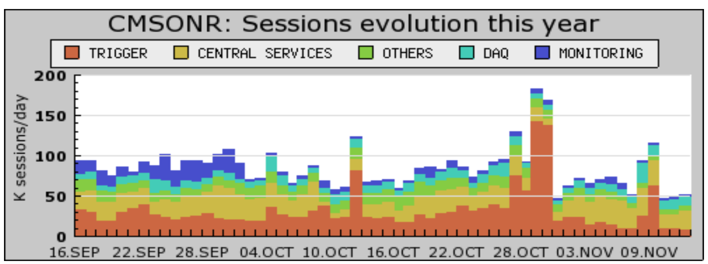
- Copy of ORCON

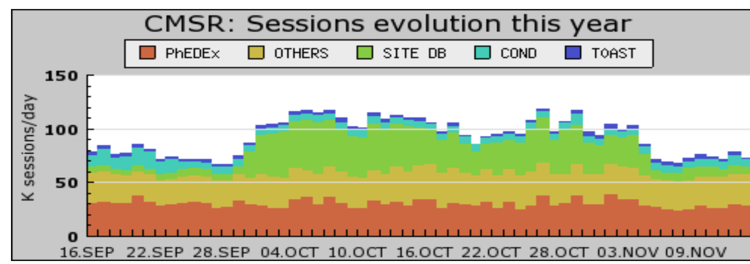


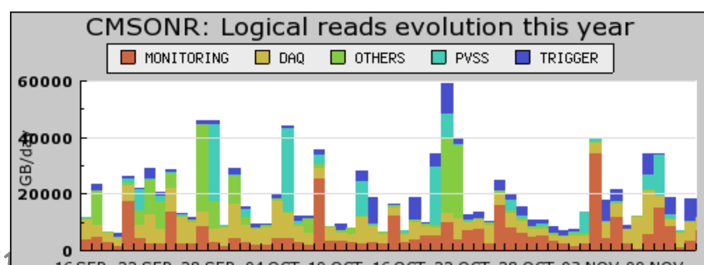


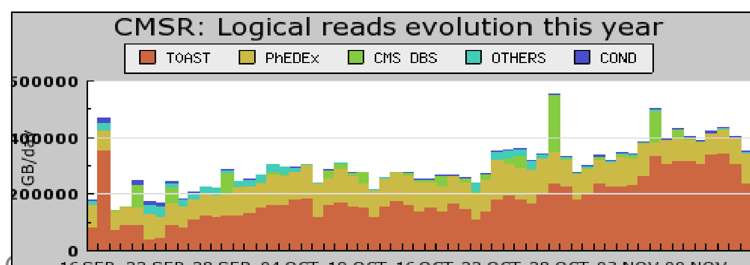
Server load

- Several applications running
  - With a variety of access patterns and data volumes
- Prior develop/deployment process in Test and Integration clusters (INT2R, INT9R) mandatory
  - tuning before production: schema layout, queries, workflow
- Complex systems require iteration process with the DBA’s
  - detailed feedback is required to obtain a good application
- Result is a load rather stable in terms of sessions and queries
  - CPU and operating system load shows some spike from time to time…









Streams

Both streams are performing well

- Condition stream highly stable
  - little or none direct DB access left to the users
  - narrow access pattern allowed by the common software
- PVSS stream had few short breaks in the past
  - the cause was some particular schema change or delete operations
  - expert from the concerned applications have wide access to the DB

In general, DB operation which might affect the streaming are discussed/planned with the DBAs.

Planning I

Move to Oracle 10.2.0.5

- after Christmas break
- need to test all applications beforehand
- already upgraded test DBs (cmsdevr, int2r)
- Online:
  - validated: TStore, Storage Manager
  - To be done: PVSS (6th Dec)
- Offline:
  - migrate ARCHDB + stress test offline sw (December)

Improve alarming and monitoring for Online DB

- Reviewed list of contacts for Online accounts (at least one email)
- IT will provide automatic tools for alarms and warning (pwd expiring, login failure, account locked, anomalous account growth,…)
- Same will be done for offline DB (after applications review)

Planning II

A couple of episodes of offline DB instability due to high load in the last two months

=> Offline/Computing applications review

Aim:

- analize the current use of the DB
- estimate the load increase for next years
- estimate the applications safety for the DB/ optimize

- Phedex presentation done 9th November
- Next: DBS, T0AST, SiteDB

Other plans:

- Conditions schema improvements
  - Will require to migrate (copy) a subset of data into the DB
- Various improvements in the Condition software

Summary

- CMS computing relies on RDBMS for many critical production applications.
- The variety of requirements in terms of access patterns, data volumes and workflow represents a relevant complexity.
- Overall DB architecture is focused in few principles:
  - Use RDBMS for quasi static and transactional data
  - Online systems safety
  - Distribute R/O data with web caches/files
- Operation during data taking brought more experience
  - Ensure stability with a well defined software process for DB applications
  - Monitoring systems essential for preventing performance degration