Experience with CMS Offline Condition Database and prospects for the future

Database Futures Workshop

CERN, 6 June 2011

Giacomo Govi (CERN)

On behalf of the CMS experiment

Outline

- Condition Data
- Strategy and Choices
- Infrastructure
- Data flow
- What we use/need

Condition Data

Wide category, most entirely involved in the offline production activities

- Calibration, detector condition
  - Varying with time and frequently updated
- Configuration
- Hardware management and description
  - Static (or quasi static)
- Beam and luminosity information
- Run information

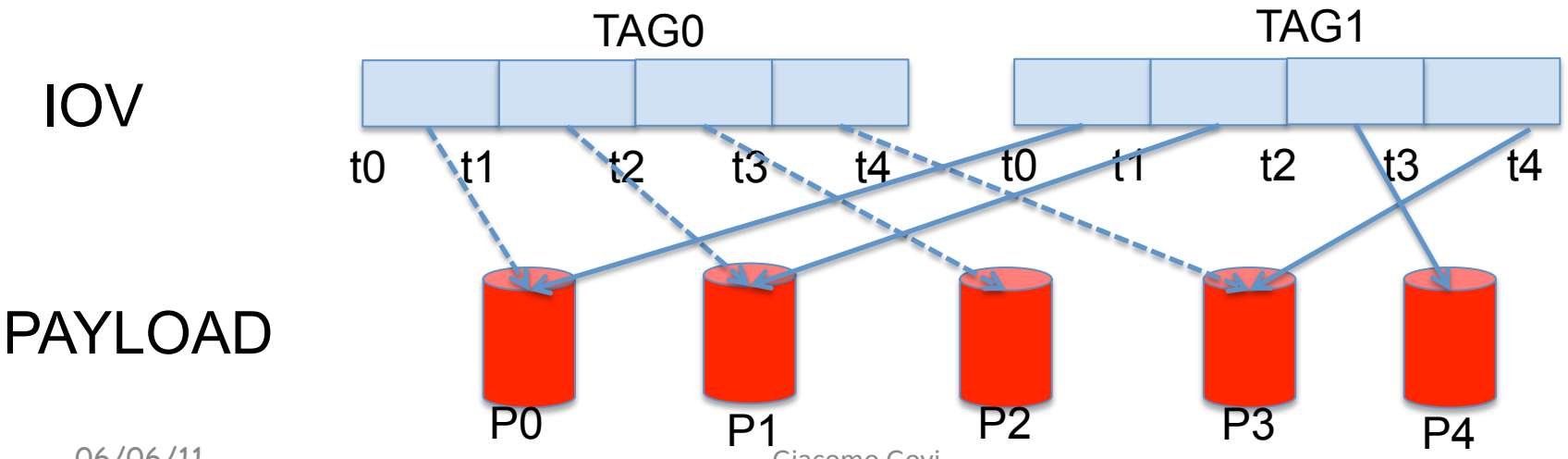
Critical for the physics data analysis chain:

- Data are exposed to a large community
  - Many institutions of the collaboration involved
  - Potentially little control on volumes expected, technologies, standards, practices, access patterns

Data model

Data structure is defined in terms of objects (from C++)

- Payload : designed according to the detector/task needs. Typically: header + param container(s)
- IOV (Interval Of Validity): array of intervals (time or run number) containing a reference to a Payload
- Tag: label identifying/categorizing a specific IOV. A catalogue for the interesting IOVs…
- Global Tag: A consistent set of Tags including all the condition required for a given processing scope



Application goals

- Enforce the DB access via a common software
  - Implementing transparently the mapping objects-tables (Object Relational Access)
- Support of a well defined set of use cases
  - Allow to control the volumes and access patterns
  - Queries are predefined and can be tuned a priori
  - No support for arbitrary query on the IOV or Payloads
- Focus on data integrity
  - IOV created and updated, never deleted
  - Limiting the manipulation of tags

Database model

- Payloads + their IOV data categorized by source (Detector or Task)
  - Individual schema (ORACLE account)

- Object based approach reduces the 'relational' complexity of the schema
  - Object instances are mapped to records in their corresponding tables
  - Only the instance ids are referenced in relations
  - Large arrays (>200 elements) are stored as Blobs for performance reason

- Queries are simple and well established
  - Cursors contain most of the time one row only!

Mapping & queries

```
struct Pedestals {
   int m_status;
   std::vector<float> m_peds;
};
```

OID	M_STATUS	M_PEDS
0	0x30	blob0
1	0x31	blob1
2	0x40	blob2

BLOB

OID	POS	M_PEDS
0	0	1.34
0	1	1.43
0	2	1.36
1	0	1.29
1	1	1.32
1	2	1.40

NO BLOB

SELECT OID FROM NAME_SERVICE WHERE TAG=?MYTAG
SELECT IOV_SEQUENCE,POS FROM IOV_SEQ_A0 WHERE OID=?OID

SELECT M_STATUS, M_PEDS FROM PEDESTALS WHERE OID=?OID
SELECT M_PEDS, POS FROM PEDESTALS_A0 WHERE OID=?OID

Hardware

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.

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

- visible within the CERN network
- ORCOFF: storage for condition, run, luminosity
- Shared with other production applications

+ Integration RAC: INT2R – visible from P5

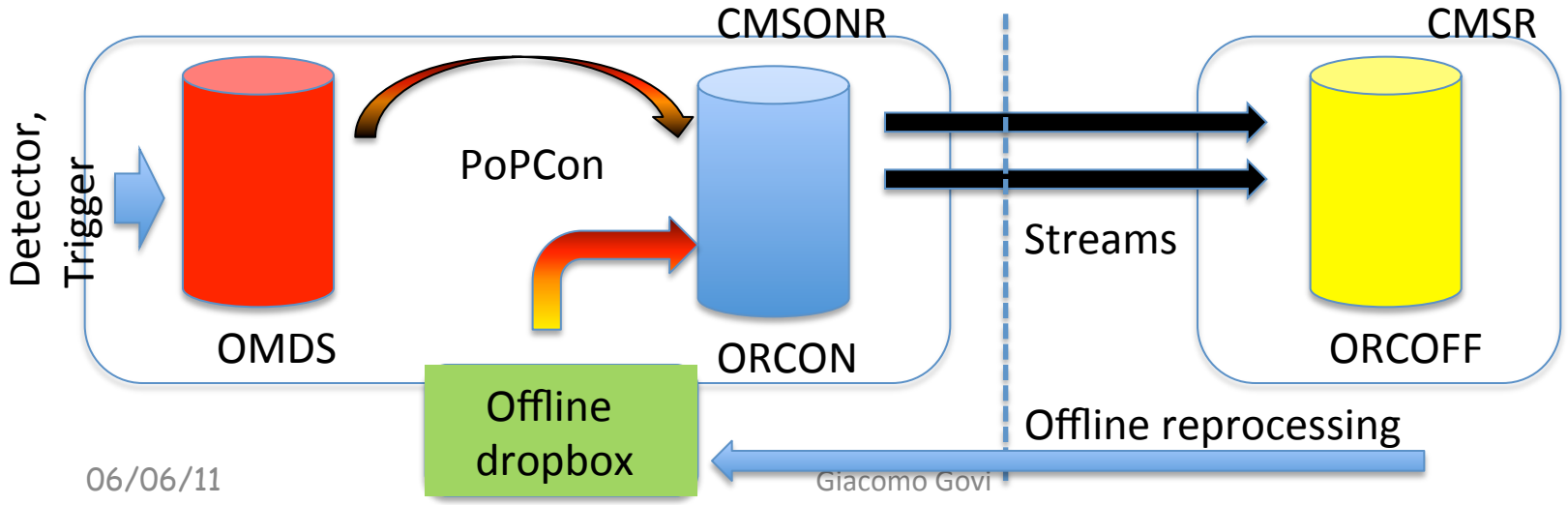
Data flow

The application only writes in ORCON

- A subset of data is read from OMDS and stored by processes running in dedicated nodes at P5
- Condition processed offline are stored by the DropBox

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



Population

- Time critical conditions
  - Include L1Trigger parameters, Run summary data
  - Data taking relies in their prompt availability (HLT, DQM)
  - A set of jobs (O2O) based on a common application (PopCon) write in ORCON
  - deployed in dedicated nodes within the online network

- Calibration calculated with offline analysis
  - possibly completed after several iterations
  - The DropBox performs the automatic exportation in ORCON
  - Handles the firewall issue accessing the Oracle DB within P5
  - Synchronize more data set fragments produced by multiple jobs
  - Updates existing Tag or creates new Tag according to the user instructions

Condition data reading/distribution

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

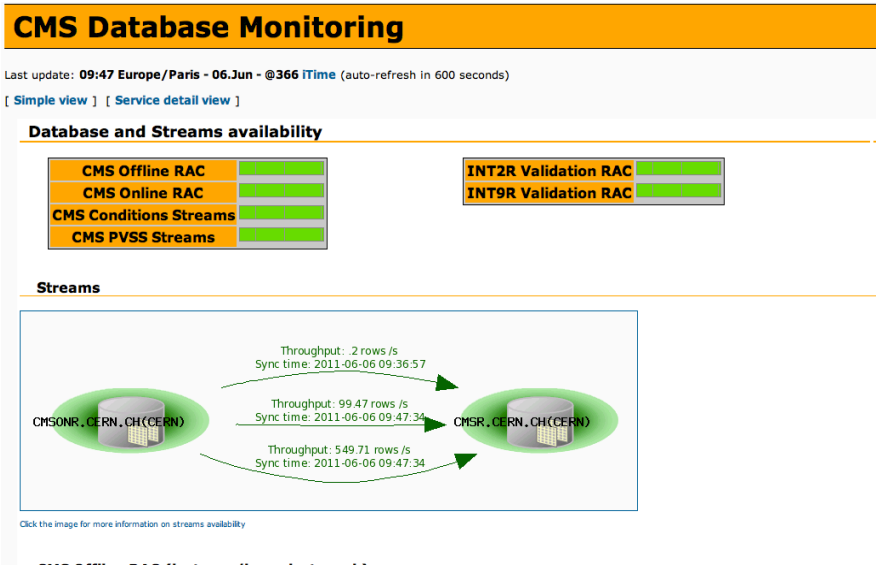
- jobs from Tier0 and Tier1s (~15000)+ subset of jobs from Tier2s (~50000?)
- 200 condition objects to read with 3-5 tables => ~800 queries
- data is read-only at large extent

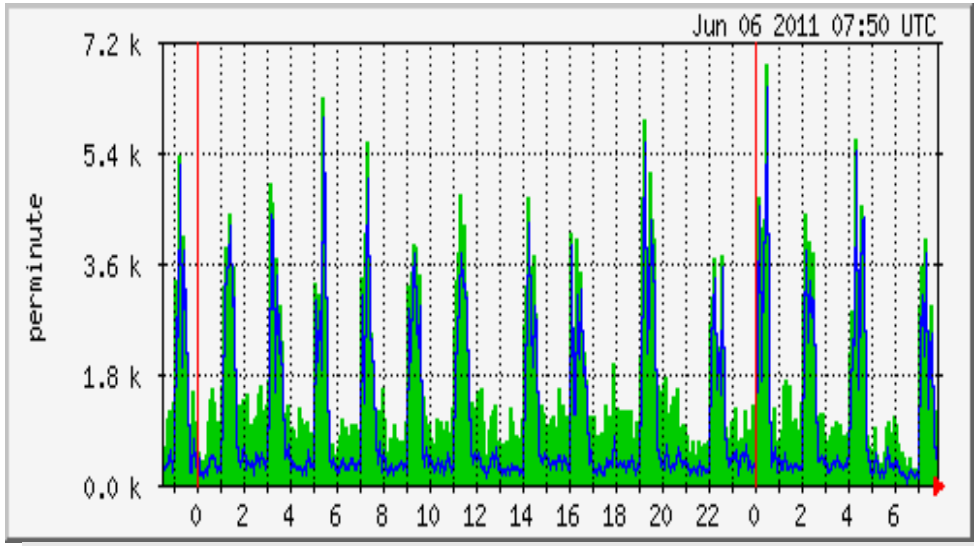
- FronTier caches allow to minimize the direct access to ORACLE servers
  - At the price of a possible latency implied by the refreshing policy
  - 2 Frontier services implemented (ORCON to P5 and ORCOFF to Tier0/1)
- Snapshot from Oracle DB are exported in a dedicate server
  - Guarantee full reproducibility and robustness against delete/insert mistakes
- 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

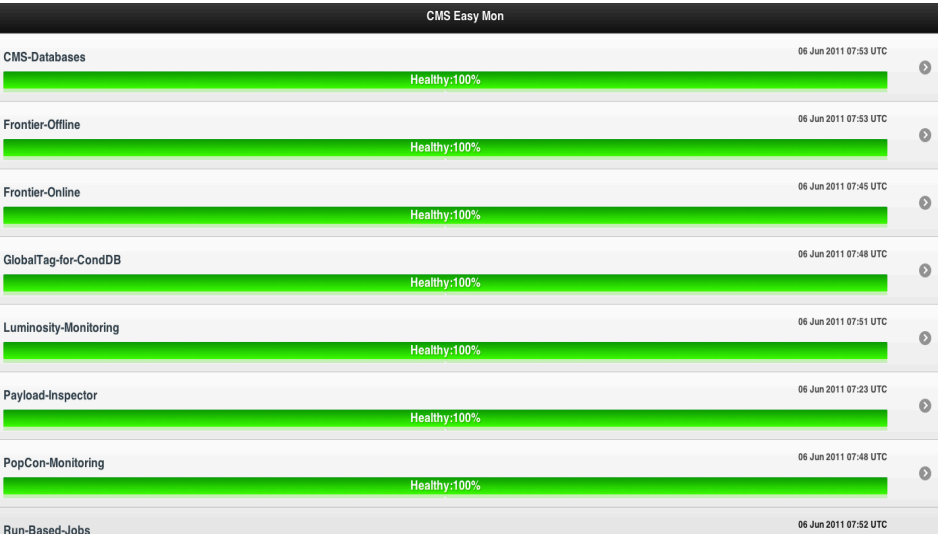
Monitoring

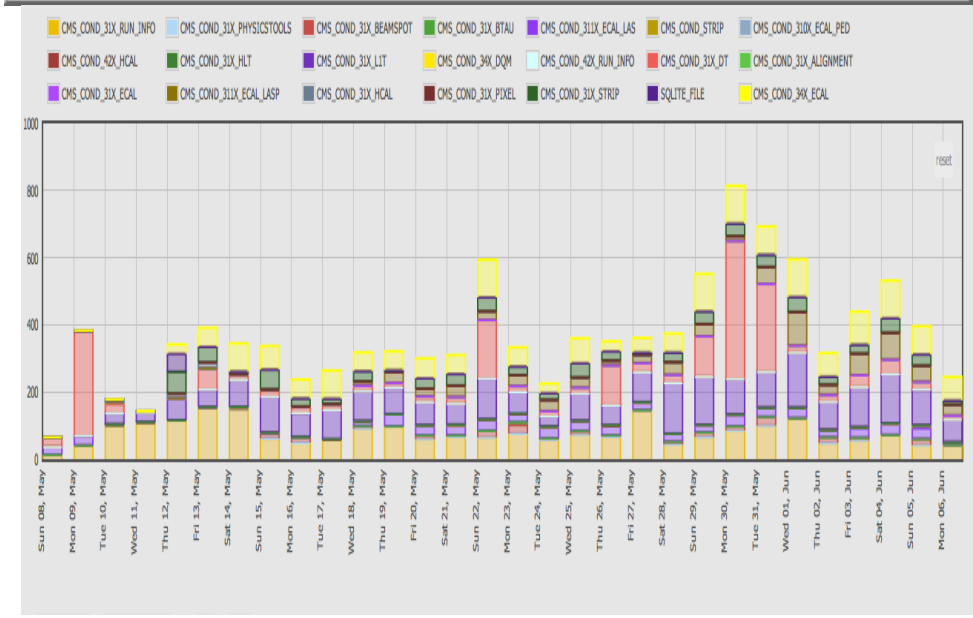
- Hardware and infrastructure
  - Disk I/O, CPU, network, streams, session management
  - Growth of data volume on schema
  - Password expiration notification
- Top Level view for the automatic or manual operation (PopCon and DropBox)
- Error reporting and Logs
- Info for the various stakeholders:
  - Condition DB expert
    - Control of workflows
  - Condition DB developers
    - Control of performances
  - Detector responsible
    - Check status of submitted exportations

Monitoring


CMS Database Monitoring
Last update: 09:47 Europe/Paris - 06.Jun - @366 iTime (auto-refresh in 600 seconds)
[Simple view] [Service detail view]
Database and Streams availability
CMS Offline RAC
CMS Online RAC
CMS Conditions Streams
CMS PVSS Streams
INT2R Validation RAC
INT9R Validation RAC
Streams
Throughput: .2 rows /s
Sync time: 2011-06-06 09:36:57
Throughput: 99.47 rows /s
Sync time: 2011-06-06 09:47:34
Throughput: 549.71 rows /s
Sync time: 2011-06-06 09:47:34
CMSONR.CERN.CH(CERN)
CMSR.CERN.CH(CERN)
Click the image for more information on streams availability


Jun 06 2011 07:50 UTC
perminute
7.2 k
5.4 k
3.6 k
1.8 k
0.0 k
0 2 4 6 8 10 12 14 16 18 20 22 0 2 4 6


CMS Easy Mon
CMS-Databases
06 Jun 2011 07:53 UTC
Healthy:100%
Frontier-Offline
06 Jun 2011 07:53 UTC
Healthy:100%
Frontier-Online
06 Jun 2011 07:45 UTC
Healthy:100%
GlobalTag-for-CondDB
06 Jun 2011 07:48 UTC
Healthy:100%
Luminosity-Monitoring
06 Jun 2011 07:51 UTC
Healthy:100%
Payload-Inspector
06 Jun 2011 07:23 UTC
Healthy:100%
PopCon-Monitoring
06 Jun 2011 07:48 UTC
Healthy:100%
Run-Based-Jobs
06 Jun 2011 07:52 UTC


CMS_COND_31X_RUN_INFO
CMS_COND_31X_PHYSICSTOOLS
CMS_COND_31X_BEAMSPOT
CMS_COND_31X_BTAU
CMS_COND_311X_ECAL_LAS
CMS_COND_STRIP
CMS_COND_310X_ECAL_PED
CMS_COND_42X_HCAL
CMS_COND_31X_HLT
CMS_COND_31X_LIT
CMS_COND_34X_DQM
CMS_COND_42X_RUN_INFO
CMS_COND_31X_DT
CMS_COND_31X_ALIGNMENT
CMS_COND_31X_ECAL
CMS_COND_311X_ECAL_LASP
CMS_COND_31X_HCAL
CMS_COND_31X_PIXEL
CMS_COND_31X_STRIP
SQLITE_FILE
CMS_COND_34X_ECAL
1000
800
600
400
200
0
reset
Sun 08, May
Mon 09, May
Tue 10, May
Wed 11, May
Thu 12, May
Fri 13, May
Sat 14, May
Sun 15, May
Mon 16, May
Tue 17, May
Wed 18, May
Thu 19, May
Fri 20, May
Sat 21, May
Sun 22, May
Mon 23, May
Tue 24, May
Wed 25, May
Thu 26, May
Fri 27, May
Sat 28, May
Sun 29, May
Mon 30, May
Tue 31, May
Wed 01, Jun
Thu 02, Jun
Fri 03, Jun
Sat 04, Jun
Sun 05, Jun
Mon 06, Jun

Operation

- Changes in the Detector's Data Models are reflected in the DB schema
  - New classes describing condition data might require new account on the storage database
- Fixing mistakes or invalid operations
  - On single Tags
  - On the export instructions for the DropBox (further consistency check added)
  - On the GT content
- Dedicated exportation for specific requests
  - Account migration, data set cleanup

Upgrades

- Condition software backend moved from the LCG-AA persistency framework to a CMS built-in software (ORA)
  - Table layout and general schema simplified and rationalized
  - Software stack limited to the effective code for the CMS use cases
  - Existing data need to be moved in new schemas
  - Completed at end of 2011
    - We still rely on ROOT/Reflex for the Class Introspection!

- Upgrade to Oracle 11g
  - Xmas break or spring 2012
  - Could lead to a general review of the architecture (Streams)

Outlook

Data volumes

- Currently around 60 GB
- Increasing of 1.5 GB/month

Activity

- 200 payload types regularly updated
- Around 50 Global Tags produces every month
- Large fraction of the transactional access has long period (quasi read-only!)

we rely on/need…

A Service/Technology providing:

- Storage on scale of 100s GB
- Backup system
- High availability and/or fast failure recovering
- Fast access based on indexes (no complex queries)
- Export or replication facility
  - Imposed by the architecture or the technology?
- Monitoring

A distributed cache for read-only access

- For arbitrary queries
- Supporting large number of clients
- Fast and scalable

what else?

A considerable fraction of our effort is dedicated to the backend software binding OO with Relational data

- Can Oracle provide more helpful features on that?

The security (authentication and authorization) is somewhat unsatisfactory

- Support of certificates used in for the Grid?

The streaming is at some extend one of the most vulnerable part of our architecture

- What are the alternatives (11g)?

Summary

- The CMS Offline Condition DB plays a key role in the CMS database infrastructure.

- Focus of its design is the control of a potentially large set of access patterns into a single software supporting predefined use-cases.
- The successful operation of the system relies on a set of key features that are provided by the IT DB service within the Oracle technology.
- No major change are expected in the system in the near future