Experiments' updates: new DB applications and deployment plans

- **ATLAS: Florbela Viegas, Gancho Dimitrov, Alexander Vaniachine, Solveig Albrand, Pedro Salgado**
- **LHCb: Marco Clemencic**
- **CMS: Lee Lueking**

Presentation Summary

- CCRC’02 exercise
- CCRC Metrics and statistics collection
- Monitoring and ADC shift work integration
- MDT Calibration dataflow and replication
- PVSS activities and future plans
- AMI replication status and plans
- DQ2 criticality and disaster recovery options
- LHCb activities
- CMS activities

CCRC02 activity and plans

- Sasha’s slides

Additionally to this metric (sessions per node) we want to collect statistics for:

1. Resource usage per process:
    - Physical and Logical I/O used
    - CPU used
    - Cache hit ratio for the session

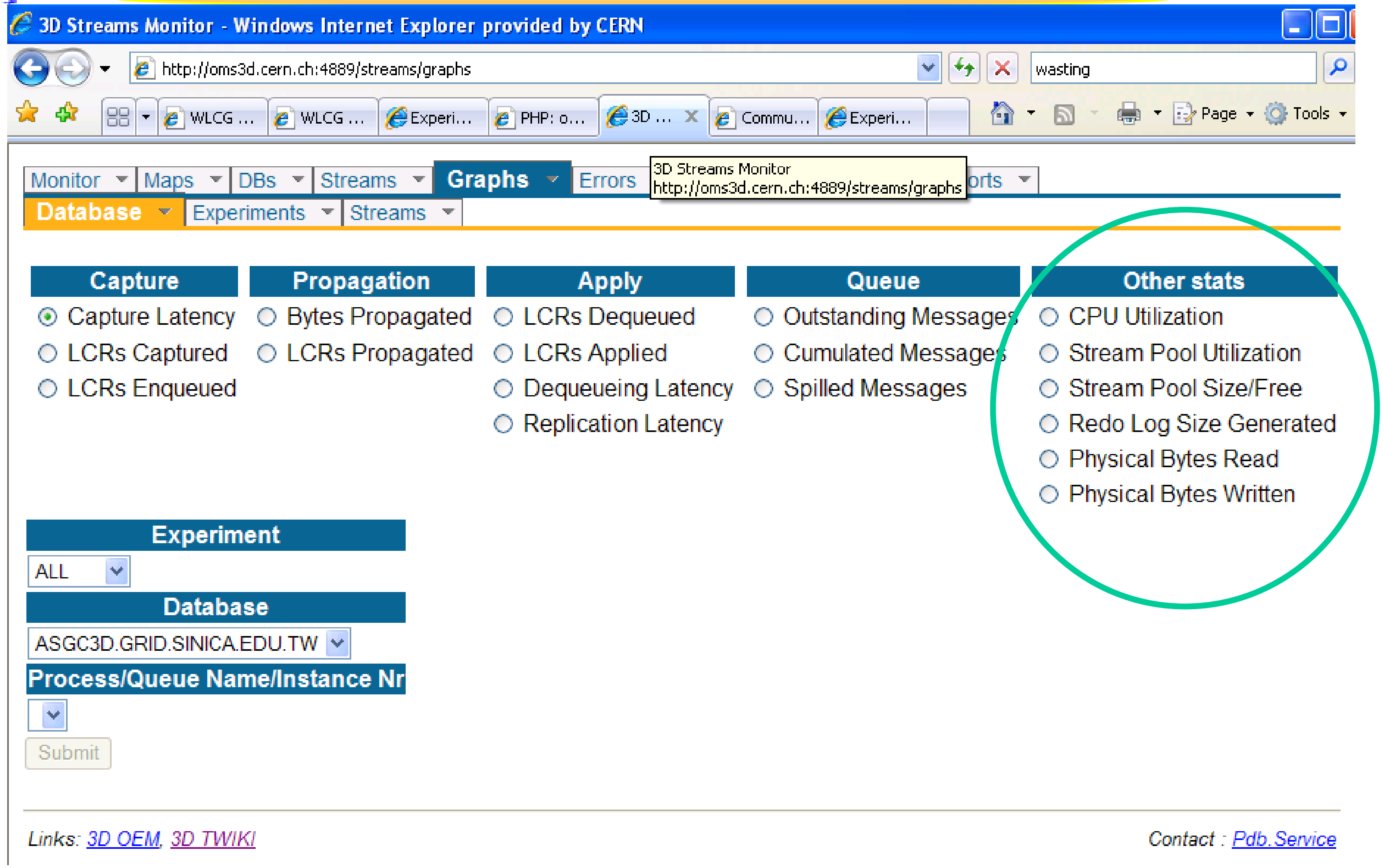
    setting the standard session audit is enough to extract them.

1. Overall machine/database behavior for I/O and CPU:
    - using the standard OS metrics collected by Oracle this can be easily obtained, might need some saving to tables periodically, as these may be overwritten. So the settings on these statistics gathering must be done by all T1s.

2. Statistics for SQL activity from the sessions:
    - For getting statistics of selects, nature of selects and other more the only way is to set the whole database on trace, and process the resulting files. This adds a bit of CPU overhead to the database, but it might be worth setting for a limited time, as it gives a very good picture of "typical" usage per process.

Some of these statistics were already collected ad-hoc in the last tests at BNL and TRIUMF.

Streams monitor – graphs



Capture	Propagation	Apply	Queue	Other stats
● Capture Latency	○ Bytes Propagated	○ LCRs Dequeued	○ Outstanding Messages	○ CPU Utilization
○ LCRs Captured	○ LCRs Propagated	○ LCRs Applied	○ Cumulated Messages	○ Stream Pool Utilization
○ LCRs Enqueued		○ Dequeueing Latency	○ Spilled Messages	○ Stream Pool Size/Free
		○ Replication Latency		○ Redo Log Size Generated
				○ Physical Bytes Read
				○ Physical Bytes Written

Experiment

ALL

Database

ASGC3D.GRID.SINICA.EDU.TW

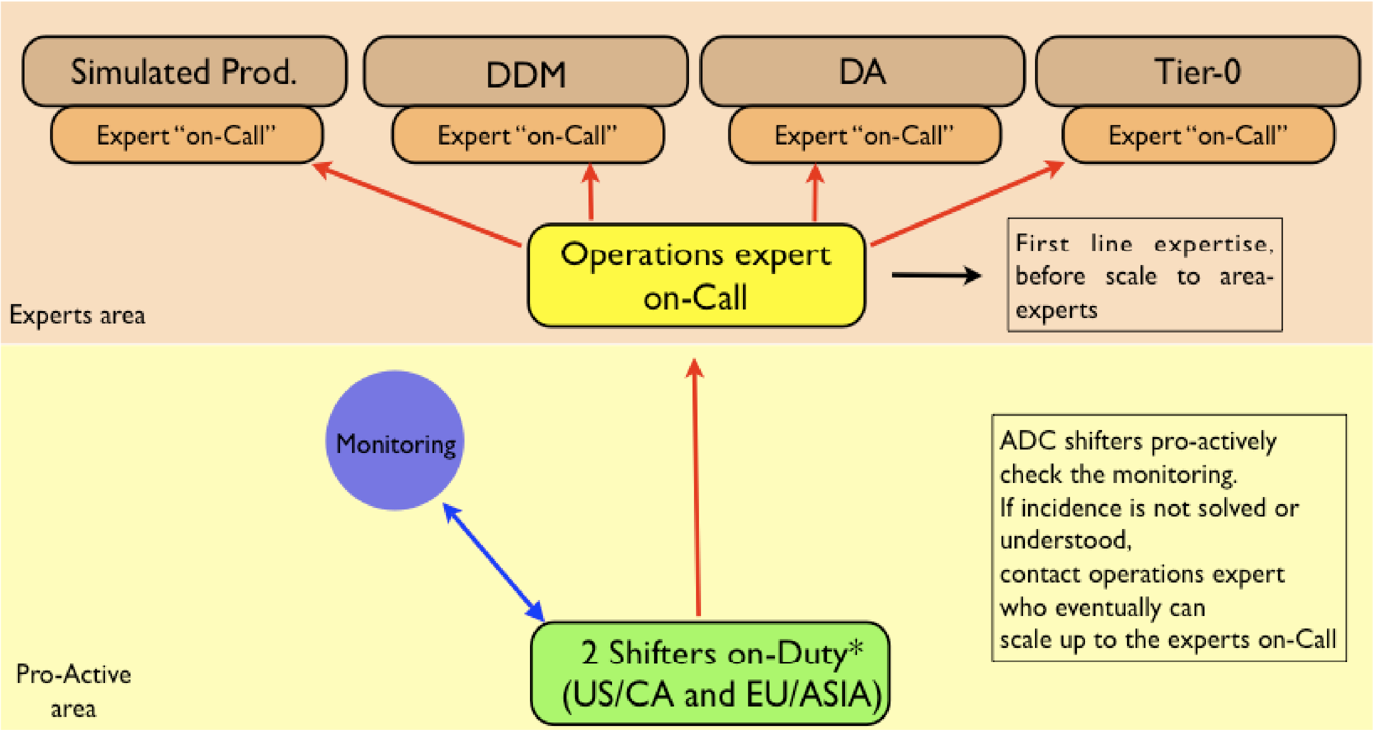
Process/Queue Name/Instance Nr

Submit

Links: 3D OEM, 3D TWIKI

Contact : Pdb.Service

- The ADC shift model:



Simulated Prod.

DDM

DA

Tier-0

Expert "on-Call"

Expert "on-Call"

Expert "on-Call"

Expert "on-Call"

Operations expert on-Call

First line expertise, before scale to area-experts

Experts area

Monitoring

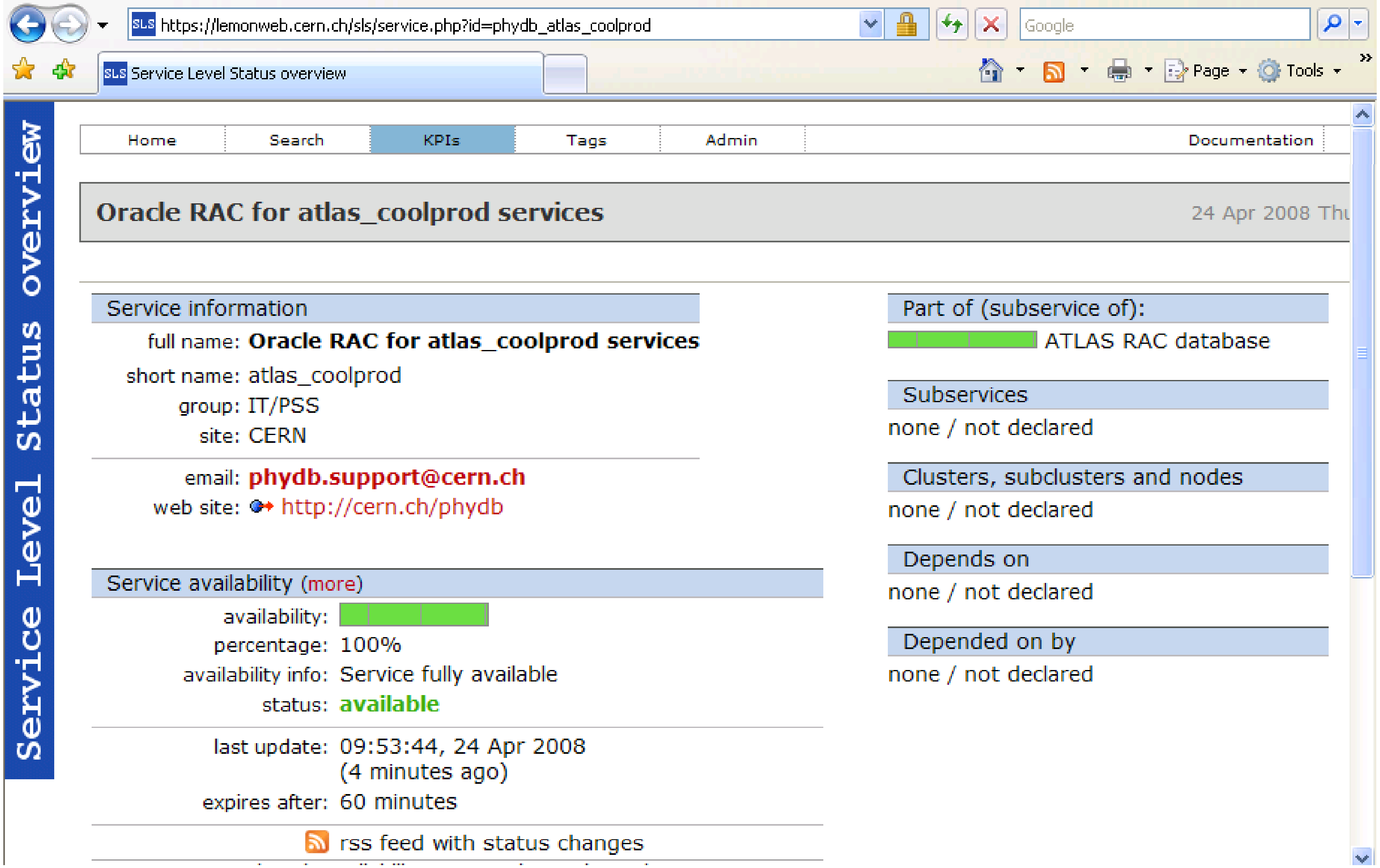
2 Shifters on-Duty*
(US/CA and EU/ASIA)

ADC shifters pro-actively check the monitoring.
If incidence is not solved or understood,
contact operations expert who eventually can scale up to the experts on-Call

Pro-Active area

- Operations expert on-call has to determine source of problem and escalation
- There should be a simple way to monitor the « health » of the conditions databases at CERN and T1s
- The diagnosys checklist, and visible indicators of health, should include:
  - Number of processes resource limit: Are the jobs getting errors because they can’t connect?)
  - Load on the database (Are the jobs too slow? Are they accumulating in the database active sessions)
  - Activity on database (Is there I/O contention? Are there locks causing hung jobs?)
  - Is the replication within the expected delay ? (Are jobs not getting new enough data?)
- Should alarms be setup for these events, in a homogeneous fashion for the whole T0/T1 operations? Using nagios, lemon, cron?

SLS for atlas_coolprod service



https://lemonweb.cern.ch/sls/service.php?id=phydb_atlas_coolprod
Google
Service Level Status overview
Page
Tools
Service Level Status overview
Home
Search
KPIs
Tags
Admin
Documentation

Oracle RAC for atlas_coolprod services 24 Apr 2008 Thu

Service information	
full name:	**Oracle RAC for atlas_coolprod services**
short name:	atlas_coolprod
group:	IT/PSS
site:	CERN
email:	**phydb.support@cern.ch**
web site:	http://cern.ch/phydb

Service availability (more)	
availability:	🟩
percentage:	100%
availability info:	Service fully available
status:	**available**
last update:	09:53:44, 24 Apr 2008 (4 minutes ago)
expires after:	60 minutes

rss feed with status changes

Part of (subservice of):

🟩 ATLAS RAC database

Subservices

none / not declared

Clusters, subclusters and nodes

none / not declared

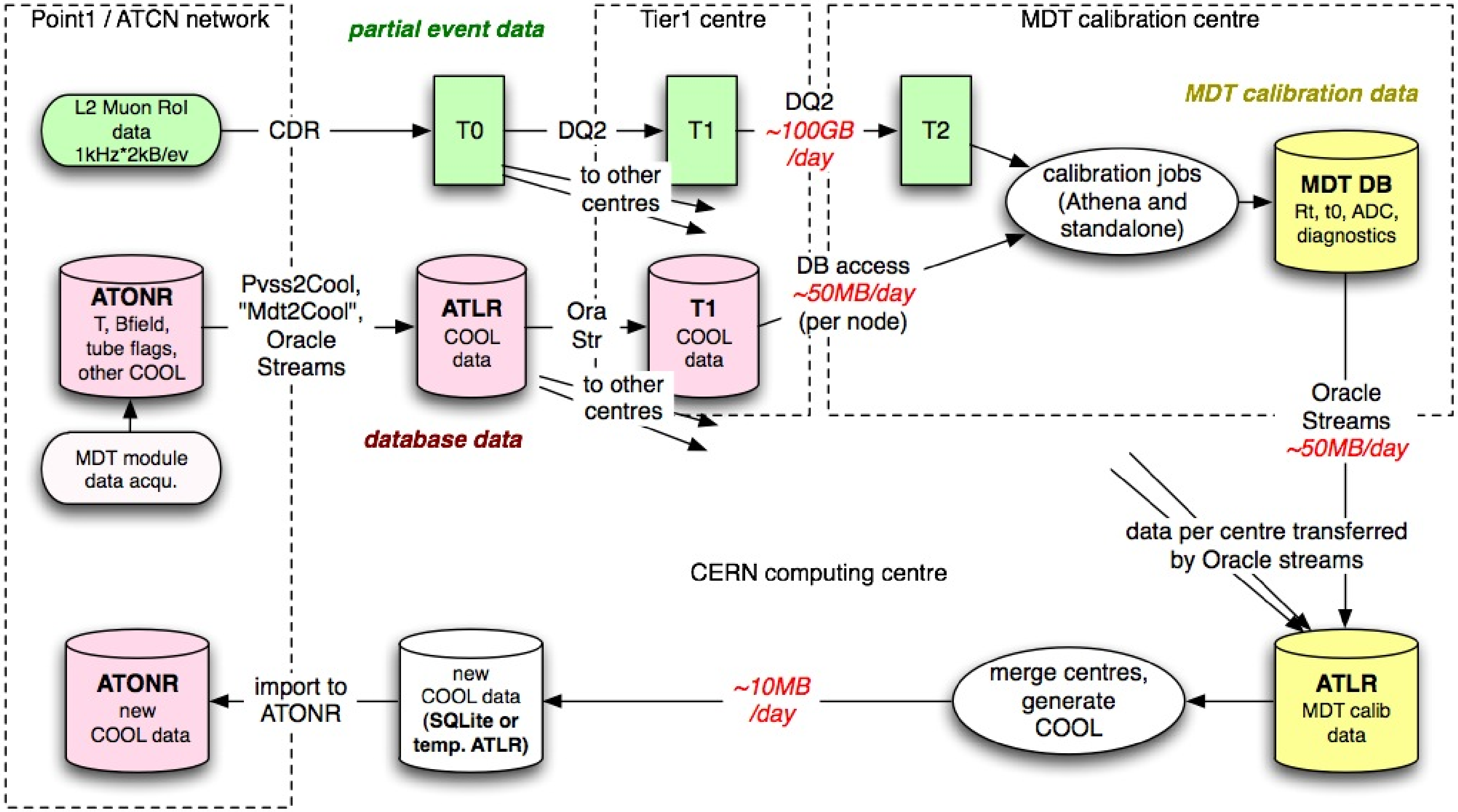
Depends on

none / not declared

Depended on by

none / not declared

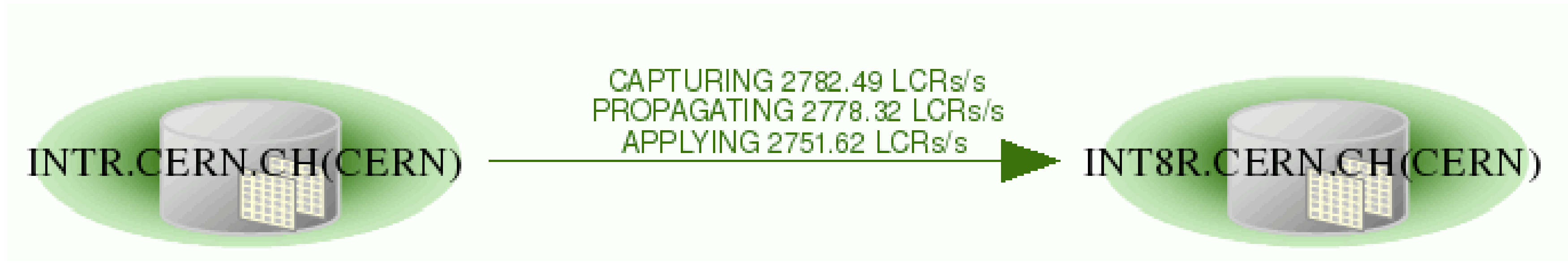
- A new replication model: MDT calibration
- Dataflow:



- Full MDT calibration data replication cycle must be complete in 24 hours.
- For T0-T1 replication, the already setup mechanism is fine, MDT calibration is one more schema to replicate.
- From Rome, Michigan and Munich to CERN, replication is monitored and maintained by site DBAs. A Service Level Agreement should be put in place.
- The jobs at these sites should get their data from the nearest T1 (INFN, BNL and Gridka.
- The other replication components and processes have to be monitored as part of the MDT calibration data replication service, so some tool has to be put in place for the ADC shifters monitoring – Gridmap?

Preparation for the production PVSS replica ATONR => ATLR

- **Test replica from INTR to INT8R has been setup**



- **Achieved improvement with almost a factor of 2 in the PVSS replication throughput since the last tests done in February**

  From 2000-2200 LCR/sec before to 3500-3800 LCRs/sec now (maximum steady rate without accumulation of latency)

- **Special thanks to Luca Canali and Eva Dafonte.**

Changes/setting done for having that improvement

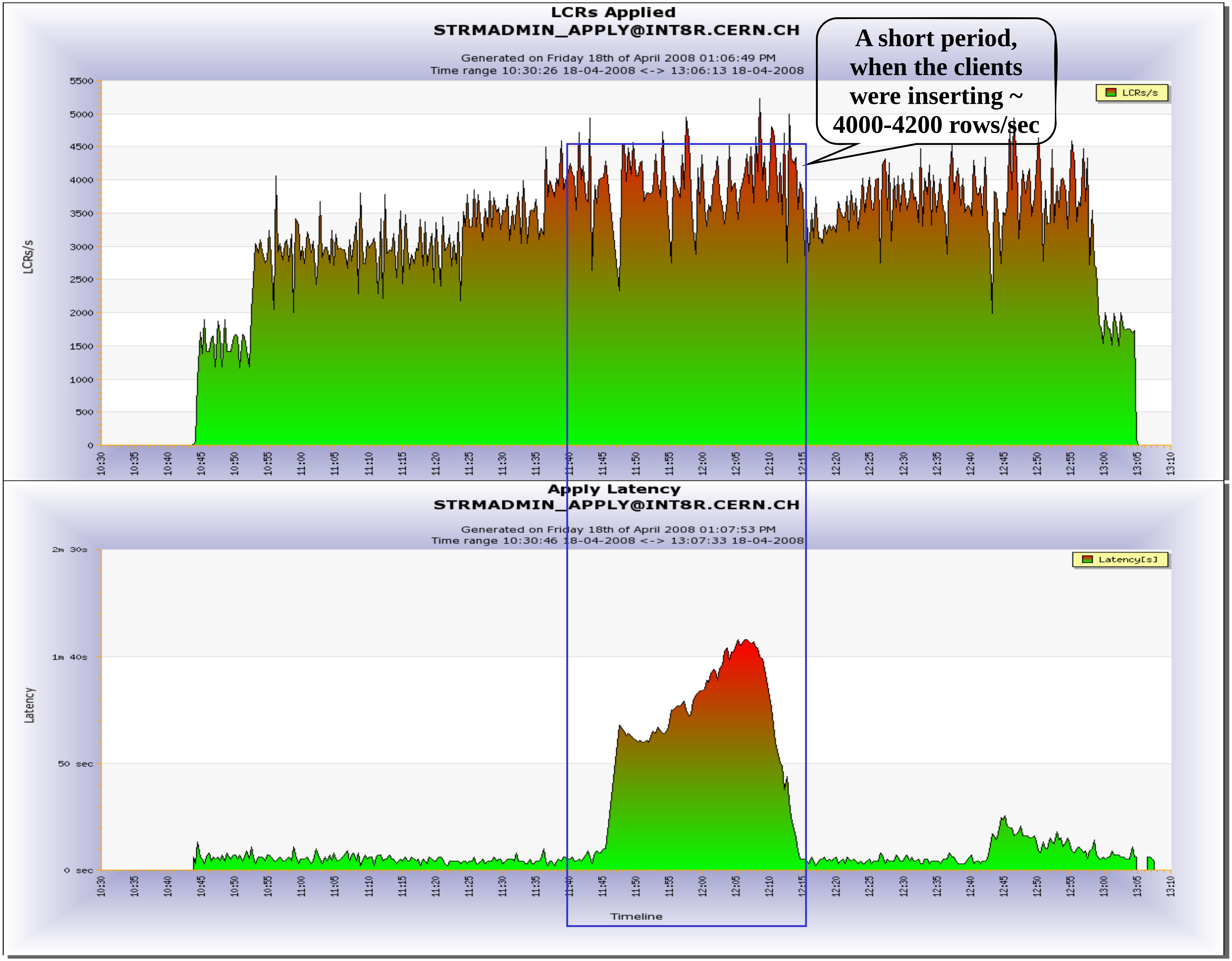
- Instead of the standard heap organized PVSS ‘EVENTHISTORY_xxx’ tables + PK, we create them as Index-Organized ones (IOT), thus reducing the I/Os and the disk space usage.

- Two hidden parameters, found from Luca, determine when a flow control to happen (the CAPTURE default is 15000 unbrowsed mesagges is the queue). Setting it to higher values caused less frequent flow control to happen.

  _buffered_publisher_flow_control_threshold

  and

  _capture_publisher_flow_control_threshold

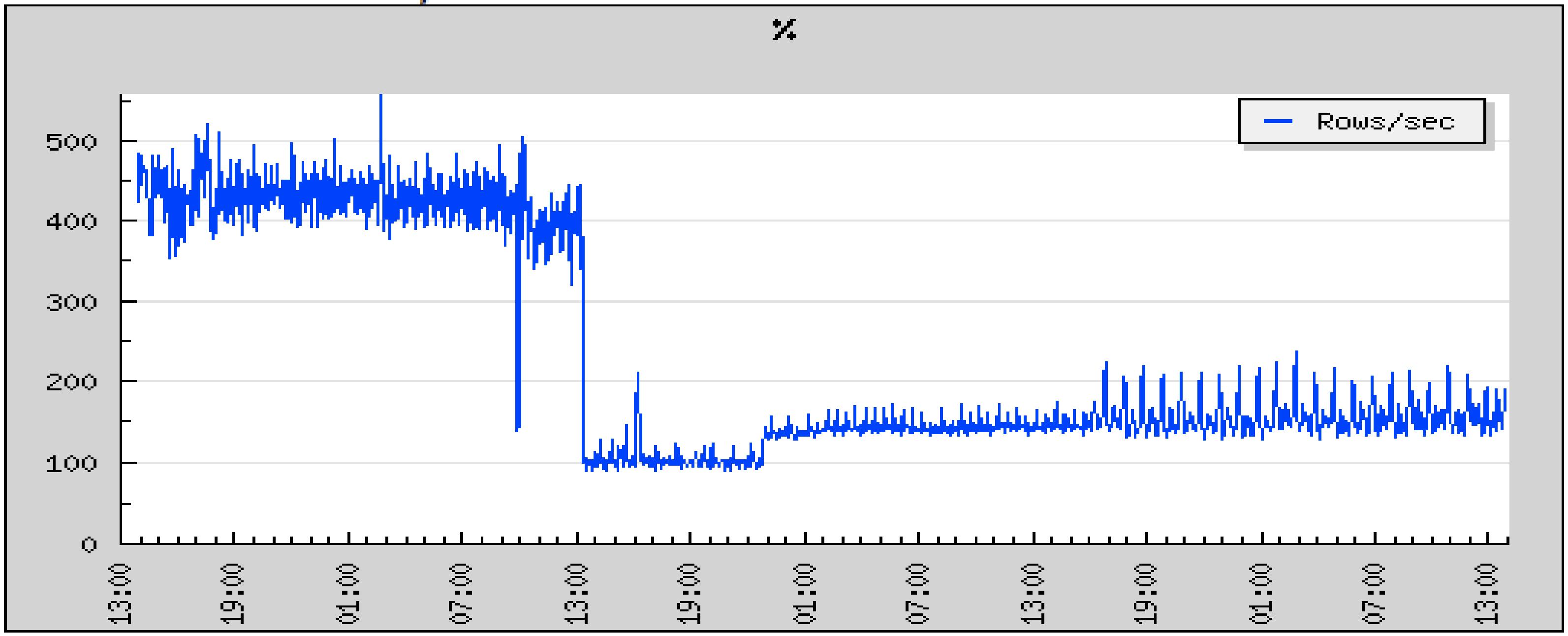


A short period, when the clients were inserting ~ 4000-4200 rows/sec

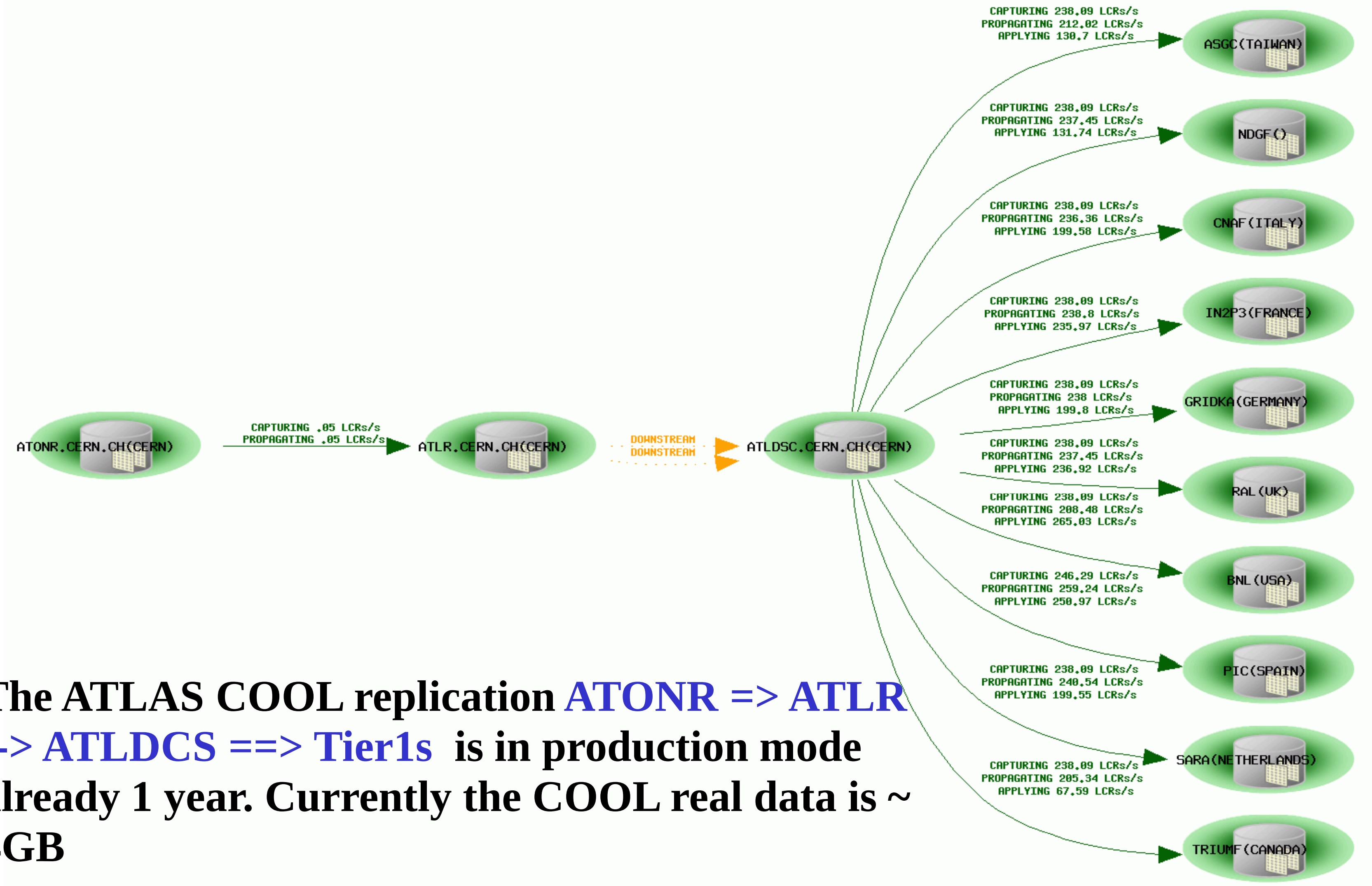
Is the achieved throughput enough?

- The results are promising and we hope that the achieved throughput will be enough for dealing with the ‘normal’ PVSS insertion plus cases of burst of inserts on ‘start of run’ and other events.
- So far the overall INSERT activity hasn’t been more than 500 rows/sec

PVSS Overall Activity



Oracle streams replication of the COOL data



The ATLAS COOL replication ATONR => ATLR --> ATLDCS ==> Tier1s is in production mode already 1 year. Currently the COOL real data is ~ 4GB

The detector conditions data and its replication

- The detector conditions data will be the biggest fraction of data replicated to the Tier1s. This data resides into the ATLAS_COOLOFL_DSC account on the ATLAS ‘offline’ DB ‘ATLR’.

- For the purpose of not flooding the replication, the schema is temporary taken out of the flow.

- After processing the PVSS data gotten in the last few months from all 11 sub-detectors, the above schema will be instantiated on all destination databases using transportable tablespace method (expected to be 10s of GBs)

Summary and plans

- The Oracle replica is active for

  - 16 COOLONL_xxx schemas

  - 16 COOLOFL_xxx schemas

  - CONF schemas (TRIGGER, PIXEL, TGC …)

  - OKS_TDAQ, ATLAS_RUN_NUMBER

  - ATLOG (atlas_logbook), COCA(Collection and Cache), MDA (Monitoring Data Archiving)

- To be added

  - PVSS archive (so far, 11 schemas)

  - PVSS CONF (configuration data)

  - ATLAS_MDT_DCS

- AMI Status – Solveig’s slides

- DQ2 – slides by Pedro

- Darios’s quote:
- “ Data taking may be in trouble after a few hours if no new dataset registration can proceed and therefore no data is moved and no jobs can be submitted (including detector monitoring and calibration), […] consider 4 hours as a kind of absolute maximum.”

LHCb actitives

- slides by Marco Clemencic

CMS activities

- Slides by Lee Lueking