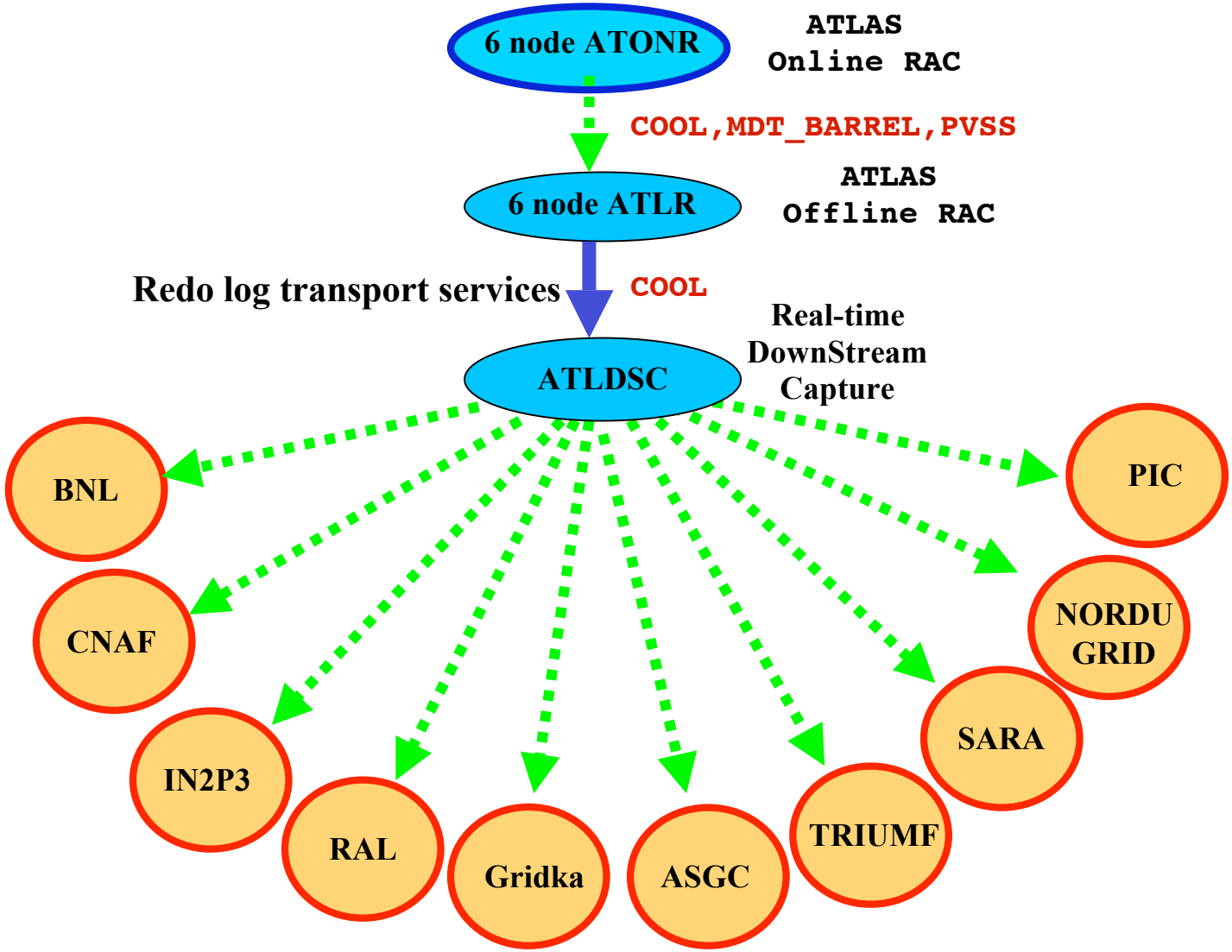
Experiment Scalability Tests and Deployment Schedule

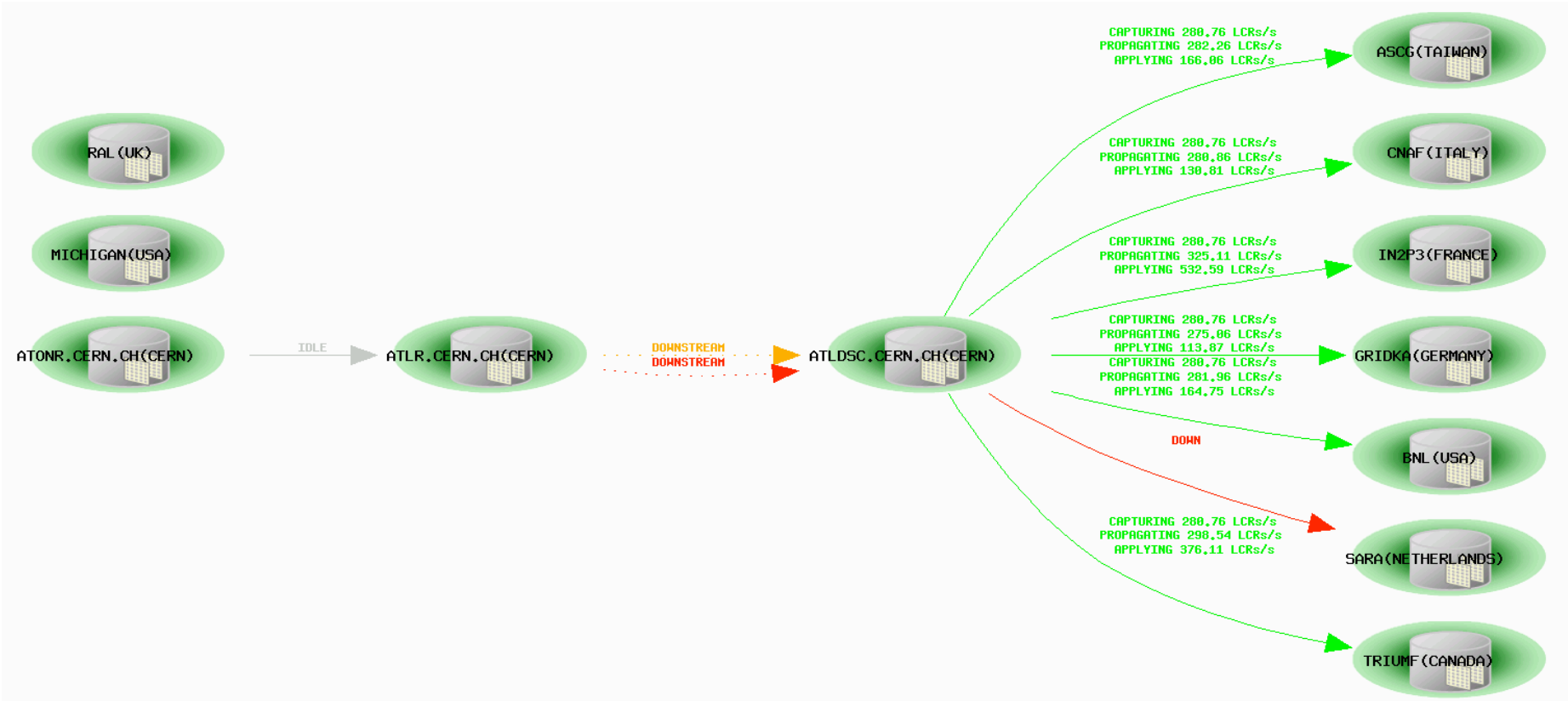
Gancho Dimitrov (ATLAS)

ATLAS Online Database, T0 and T1 Architecture - full production in autumn 2007

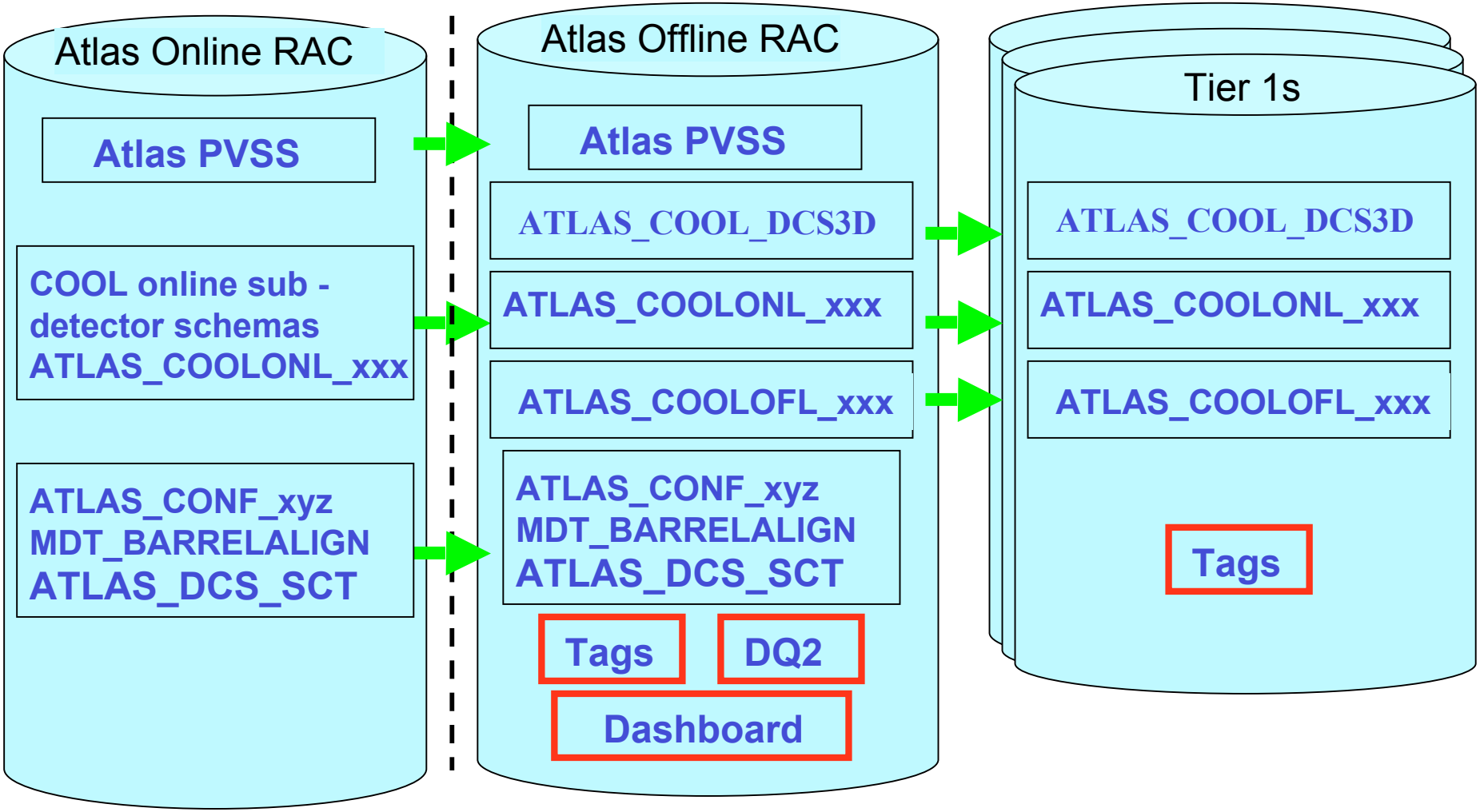


Where are we now?
The ATLAS topology - June, 5th 2007

- Production Phase for conditions data challenges started on April, 1st 2007 with 6 active destinations - GridKa, RAL, IN2P3, CNAF, SARA and ASGC
- BNL and TRIUMF have been included in the beginning of May
- Adding NDGF and PIC is in progress (hopefully on the time of my presentation we will have all 10 ATLAS Tier1s in the Streams flow)



ATLAS schema organization



sub-detectors xxx = INDET, TRT, LAR, TILE , MUON, MUONALIGN , MDT, TDAQ, GLOBAL, SCT, PIXEL, TRIGGER, CALO, RPC, TGC, CSC

Experience so far …

- Since April 13th, about 40Gb of COOL test data have been replicated:

  A cron job runs twice per hour (at xx:25 and xx:55) adding one more run's worth of data to the ATONR database (~20 MB per run)

  Data amount per day ~ 1GB

- Streams behavior

  - Streams flow ‘online’ RAC ATONR ==> ‘offline’ RAC ATLR

  no problems so far

  - Streams flow ‘offline’ RAC ATLR + Downstream capture machine ATLDSC) ==> Tier 1s

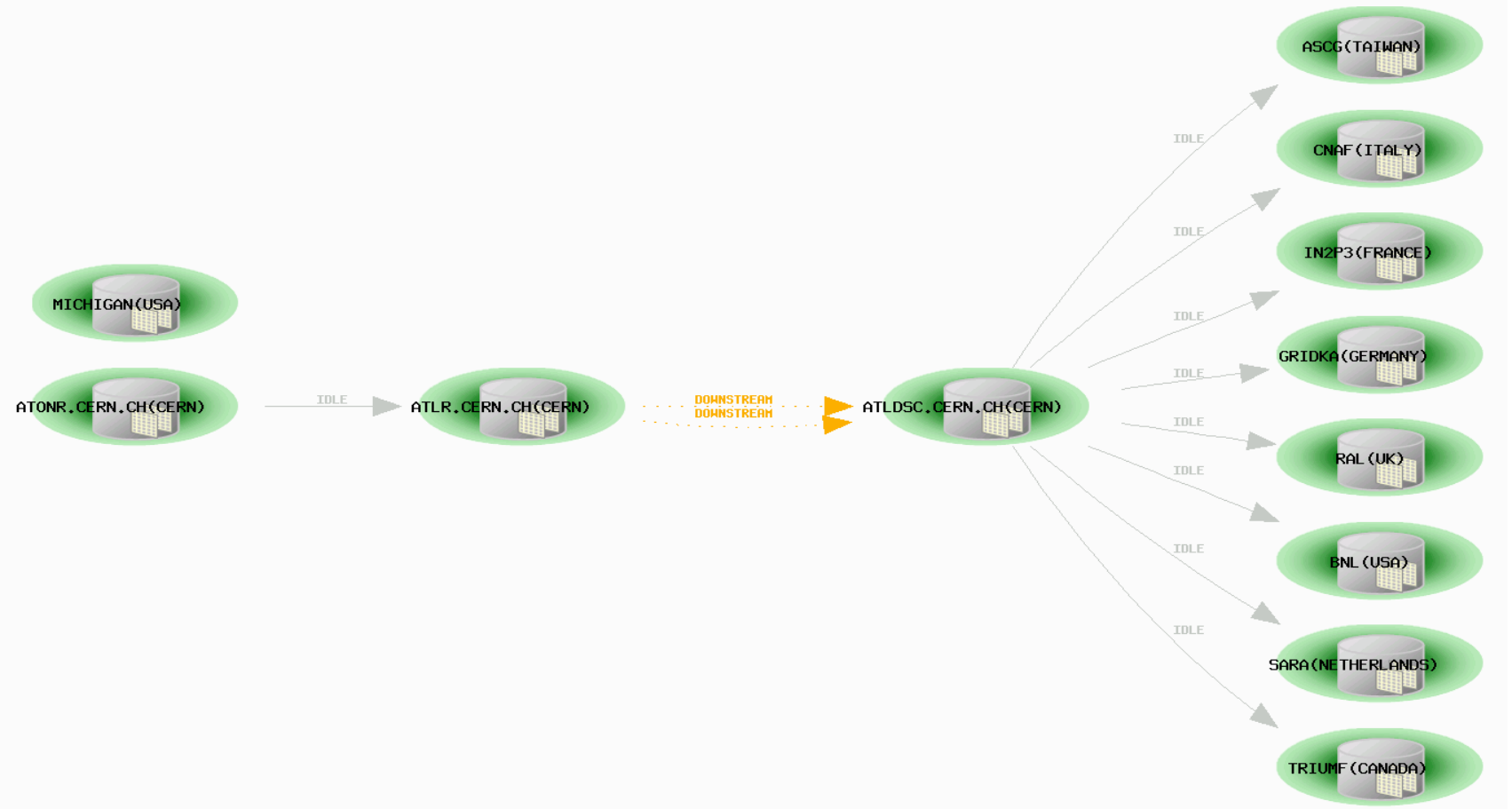
  Works fine, but still some improvements to be done (sometimes CAPTURE gets aborted because of memory problems - currently under investigation by Oracle)

  The procedure of ‘isolating’ a problematic Tier1 has been applied several times and proved that works fine (thanks to Eva)

ATLAS Streams topology - 10th June, 2007

Streams in process of synchronization -

downstream capture machine having 2 queues and 2 CAPTUREs : basic one for the well behaving destinations and temporary one for Tier1s having problems or performing intervention on the DB



Experience so far … (2)

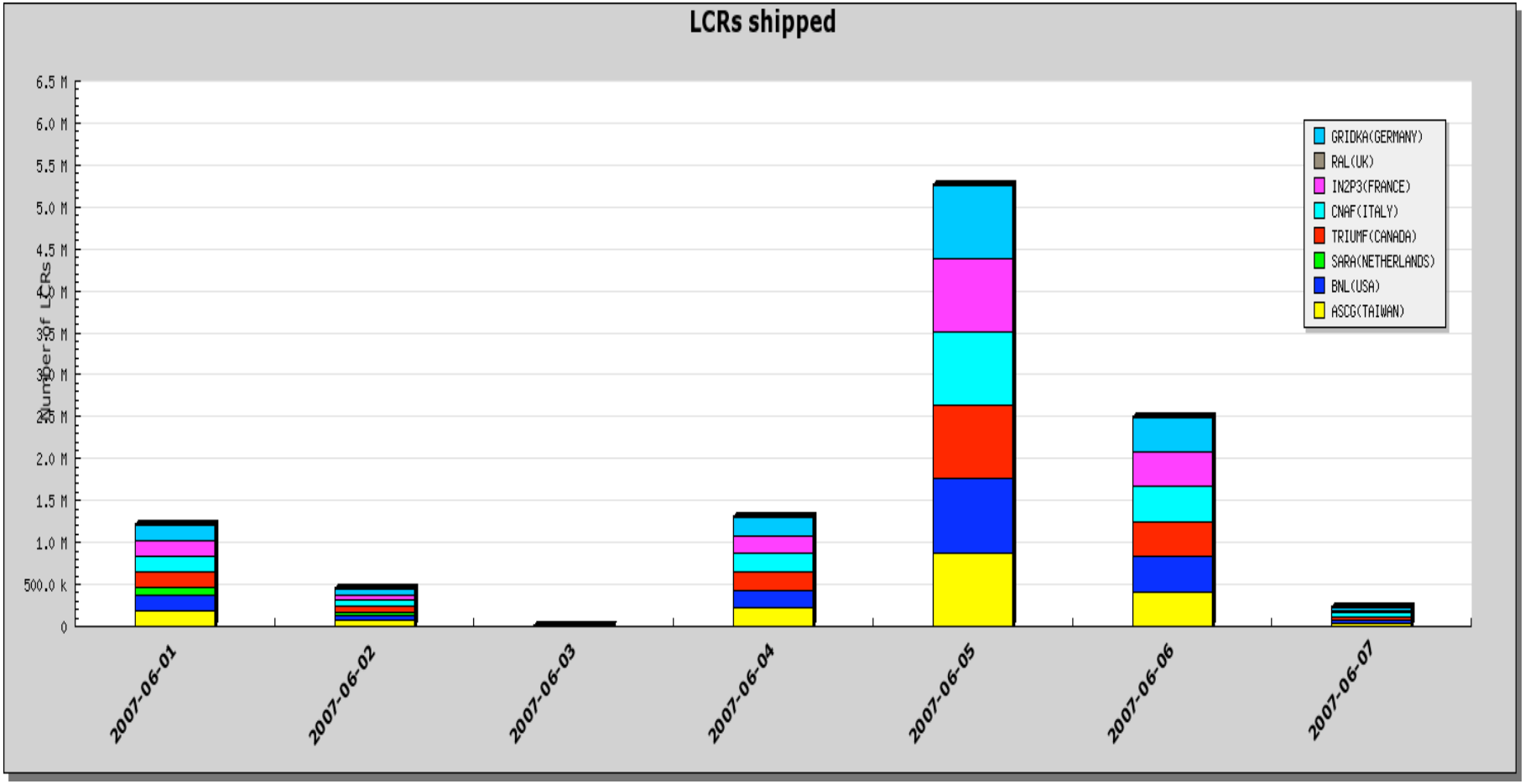
- There is no monitoring on the Streams after the regular working time. Most of the problems happen on Friday evening, Saturday or holidays and noone takes care of that. (network problems, DB blocked because of lack of space for archiving the redo logs and so on). Afterwards, it is quite difficult for the Streams to recover.

- A fast reaction to the problems from the Teir1 DBAs is essential

- There are certain problems on the CAPTURE side as well.
  Often gets aborted because with “ORA-01280: Fatal LogMiner Error” error

- The character set conversion AL32UTF8 => WE8ISO8859P1 at RAL using ‘CSALTER’ tool was successful, but didn’t help for solving the problem with the very slow import of the CLOB data. An import of 40GB file took 8 hours, while the same at NDGF took 3 hours !?!

Streams activity from the last week - nice report from Zbigniew's Streams monitoring page



Further steps …

- Having all 10 Tier1s synchronized on the production Streams flow. We couldn't add PIC to the Streams flow on Friday as the file transfer was very slow and the failed import afterwards because of the many invalid packages in the SYS schema

- Double the inserted amount of data from 1 to 2 GB and monitor the overall performance and stability. If the replication goes smoothly and there is no latency from the Apply side, than no need for introducing Materialized views.

  (the expected amount of data in COOL is ~ 1.7 GB/ day)

- Massive parallel readback tests at IN2P3 (later in CNAF as well)

  Test massive parallel access to the Tier-1 servers by many simultaneous jobs, to see what readback load they can sustain

- Preparations for the data readback tests at BNL with real data from one subsystem (LAr calibrations data from comissioning).

Open questions …

- Shall we have 2 databases on the offline RAC (or only one as it is now)

  - 1st one, for all schemas that will be replicated to the T1s

  COOLONL_xxx, COOLOFL_xxx, COOL_DCS3D

  - 2th one, for all applications that will be not ‘streamed’

  PVSS, Tags, DQ2, Dashboard_dm, Dashboard_job, ATLAS_PS

  The idea is to assign the Downstream Capture to the first DB, offloading it from scanning huge amount of redo logs data from the intensive writing applications as Tags, PVSS, DQ2 …etc

- Replication of the PVSS data from online cluster ‘ATONR’ to the offline ‘ATLR’ via Streams is currently under testing. So far promissing results - 300Mb per hour is not a problem (~ 1500LCRs/sec), CPU at the destination is ~ 40-45 %, no latency

ATLAS resource requirements to the T1s

- No change in the requirements since the last 3D workshop at SARA, Amsterdam (for ref. see the next 2 slides)
- Tags data will definitely not be replicated via the Oracle Streams
- ATLAS does not require Tags databases to be hosted at all Tier-1s

  Instead, Tier-1s are encouraged to volunteer to host Tags database in Oracle. (BNL is already one of them)

ATLAS resource requirements to the T1s

- ATLAS will probably have maximum 200 days data taking per year
- 50K active seconds/day = 58% efficiency for each active day, 10^7 events/day
- 200 Hz during active data taking

Year 2008

40% of a nominal year gives the following estimate

TAGs 2,44 TB , COOL 327 GB TOTAL : 2,77 TB

Year 2009

60% of a nominal year gives the following estimate

TAGs 3.65 TB , COOL 490 GB TOTAL: 4.14 TB

For each additional nominal year

TAGs 6.09 TB , COOL 818 GB TOTAL: 6.9 TB

ATLAS resource requirements to the T1s

For each nominal year

TAGs 6.09 TB , COOL 818 GB TOTAL: 6.9 TB

This acounts are for « real » data: All the indexing, materialized view structures for COOL and single collection for TAGS

T1s should add to this numbers the Oracle overhead:

- Auxiliary tablespaces (SYSTEM,SYSAUX,UNDO)
- Space for Backup and Recovery Policy agreed with CERN IT for consistency :
  - Archive log management
  - Flashback recovery area
  - Backup on disk
- Space for the mirroring in the ASM storage system

LHCb plans - update from Marco Clemencic

- Tests

  - currently testing the software infrastructure to access the conditions database from the GRID

  - scalability tests foreseen in a couple of months

  - still to validate the configuration of the downstream capture for CondDB

- Deployment

  1. Conditions Database
     - all LHCb Tier-1s connected via Streams
     - by the end of the summer it should be used on a production basis(accessed by reconstruction and analysis jobs)
  2. LFC
     - preparing a plan with Tier-1 service responsible and DBAs for the deployment of LFC replicas(the deployment should be completed after the summer)

LHCb Streams topology