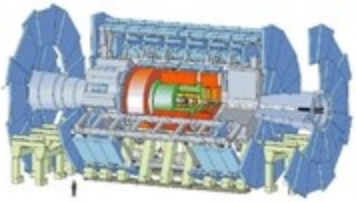
the ATLAS Experiment



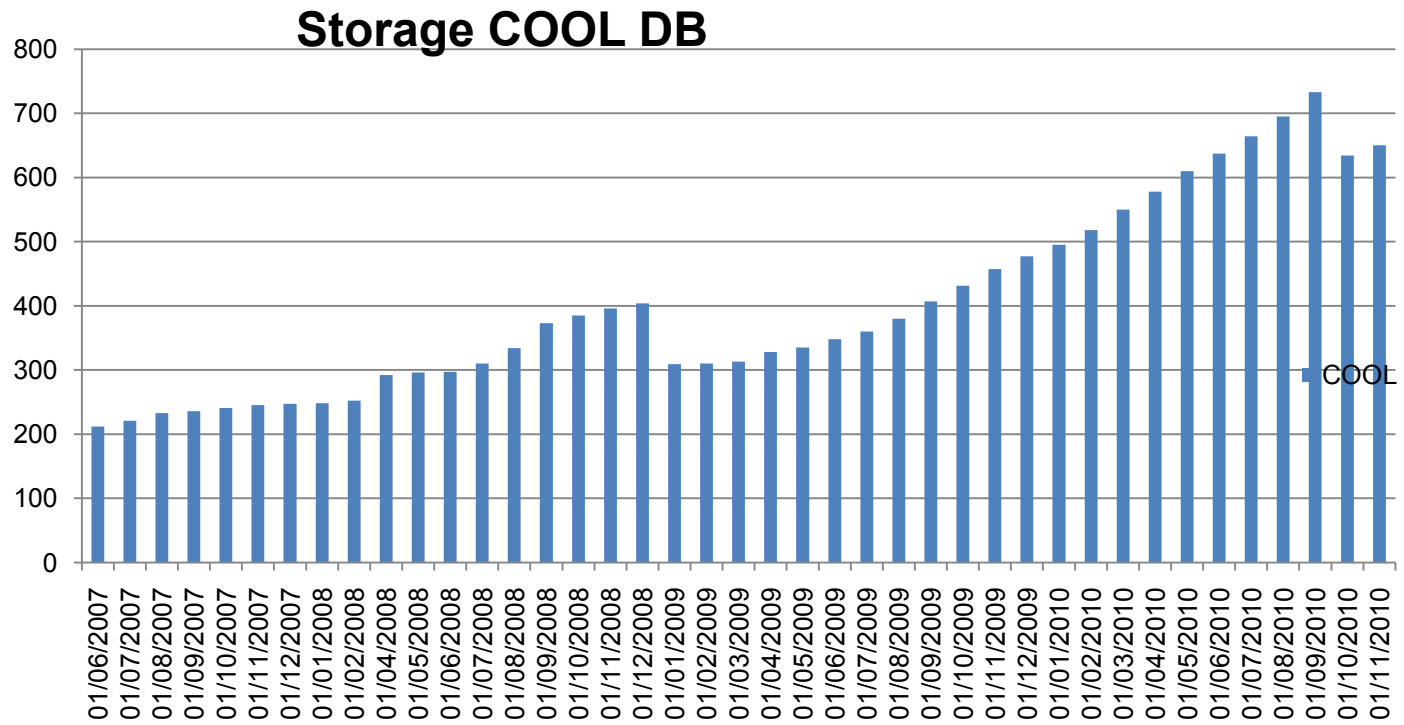
Database developments/optimizations in ATLAS

Gancho Dimitrov (DESY)
Florbela Viegas (CERN)

Outline

- Conditions Database evolution
- The lifecycle of the PVSS data
- New organization of the DQ2 traces data
- Problems with queries that ‘look’ into data of the most recent hours/days
- TAG Databases overview
- Conclusions

Conditions Data evolution



Storage GB

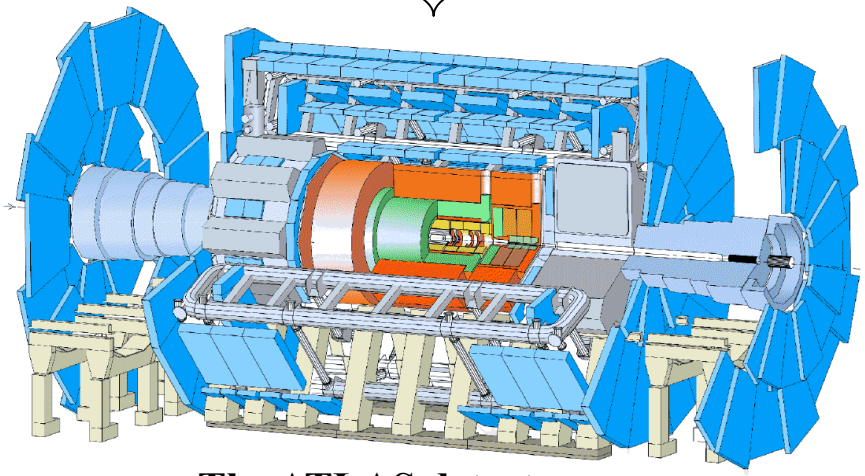
- Well within 1 TB per year evolution, in 2 years we grew from 309 GB to 650 GB, approximately 170 GB/year
- At Tier-1 we have 360 GB replicated at present (OFLP200, CMCP200, COMP200, TMCP200, TDBP200)
- Movement of old cool accounts to Archive DB took place in September

Introduction to the PVSS system and its use in ATLAS

PVSS (Prozessvisualisierungs und Steuerungssystem) is a control and data acquisition system being in use in the LHC experiments since year 2000.

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

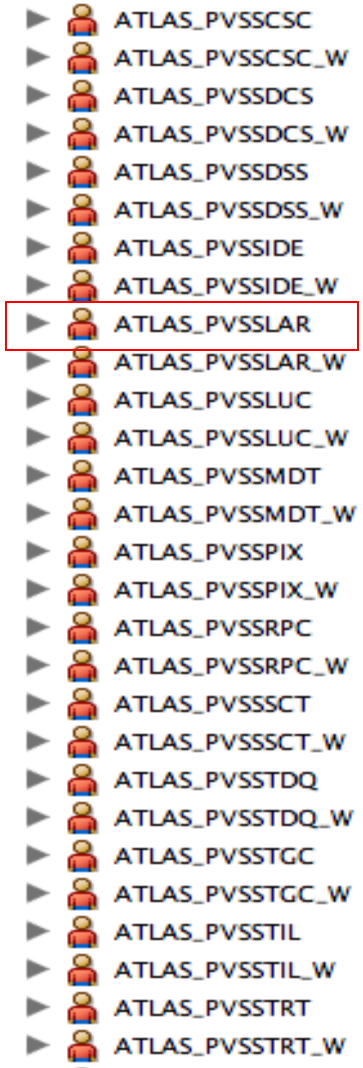
ATONR

ORACLE

The ATLAS 'online' Oracle DB

The ATLAS PVSS DB accounts and table desc.

- A database schema per subdetector (as total 14)



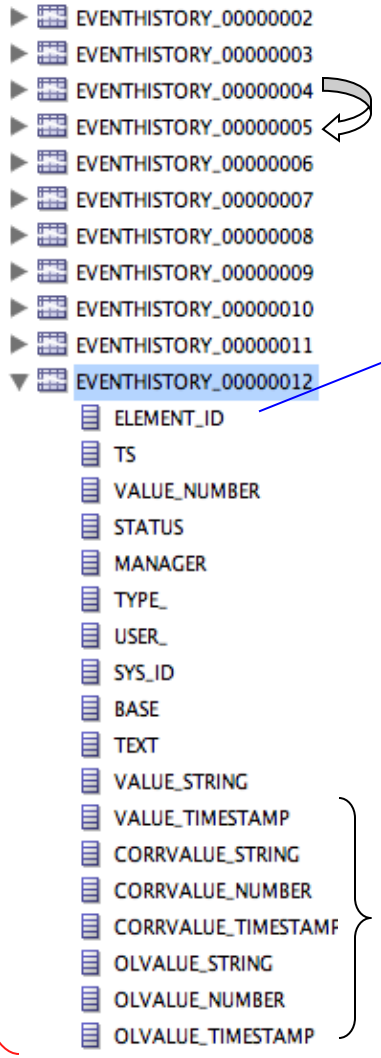


Table is 'switched' when it reaches a certain size and a view is updated to keep them together for the application to access the data (the EVENTHISTORY view)

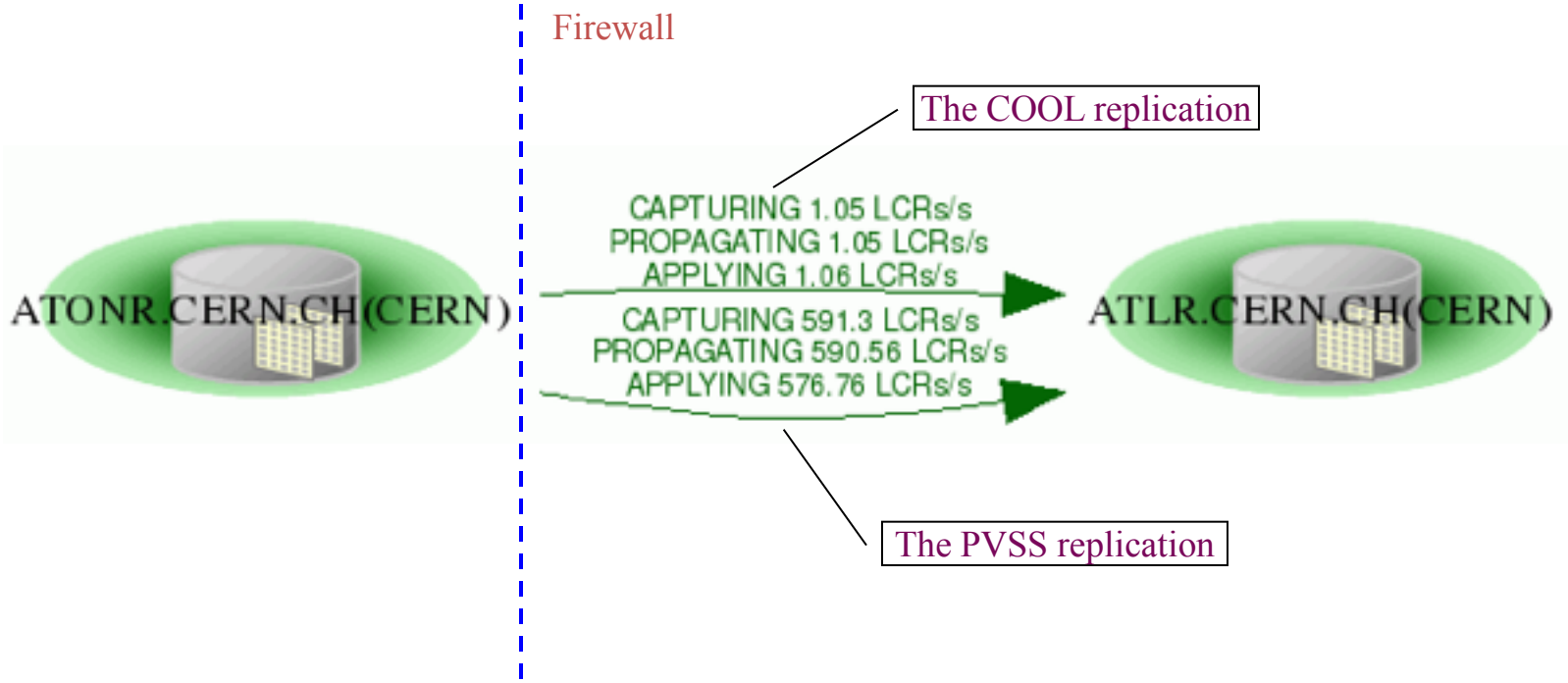
Data point elements, in the LAR case are about 4500

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

The row length is in the range 55-60 bytes

The need of having PVSS data replication from ATONR to ATLR ('online' => 'offline')

- In order to have the PVSS data accessible for the sub-detector expert analysis from the CERN public network and even from outside CERN a need for its replication showed up.



Sliding window for the PVSS Archive on the ATONR

- An idea of keeping only the data of the most recent 12 months on the ATONR (sliding window) popped up naturally.

  The reasons are:

  - the operators in the ATLAS control room do NOT need to look furhter than 12 months in the past.

  - the complete archive is already on the ATLAS ‘offline’

  - the ‘online’ DB is vital for the datataking and is wise to be kept smaller in case of a need of recovery operation.

  Currently the PVSS data (all tables and index segments) of the last 12 months occupies ~ 2.5 TB

Sliding window for the PVSS Archive on the ATONR (2)

1) That approach implies a move from the current « tablespace size threshold » to a « time interval » one – promising results from the tests

2) As each PVSS table resides into its own tablespace, for ATLAS that would mean ~ 100 tablespaces / year. Producing so many tablespaces(files) on the 'offline' side is not acceptable from administration POV. To address the last, a special code was introduced in the Streams Apply handler which combines the PVSS tables of each sub-detector and an year in a common tablespace.

3) An important is to prevent table dropping on the source DB from being propagated on the destination DB.

A double protection is foreseen – a tagged session on the source DB and special code in the APPLY handler on the destination DB that discards any dropping table messages.

The tests so far are very positive. The move towards of putting the changes on production is to be agreed … Naturally this would be when there is a LHC technical stop

Distributed Data Management System (DDM) – a move to a new organization of the traces data

- Each operation on ATLAS dataset level on the grid get registered on the DDM database (hosted on the ATLAS ‘offline’ database)
- So far the data was kept in a range partitioned table (an Oracle partition per month). Each partition having more than 100 mln rows and is expected to be more and more in the future.
- The table has an index on a column of timestamp type. This index often becomes a hot spot as contention is caused on high concurrent inserts.

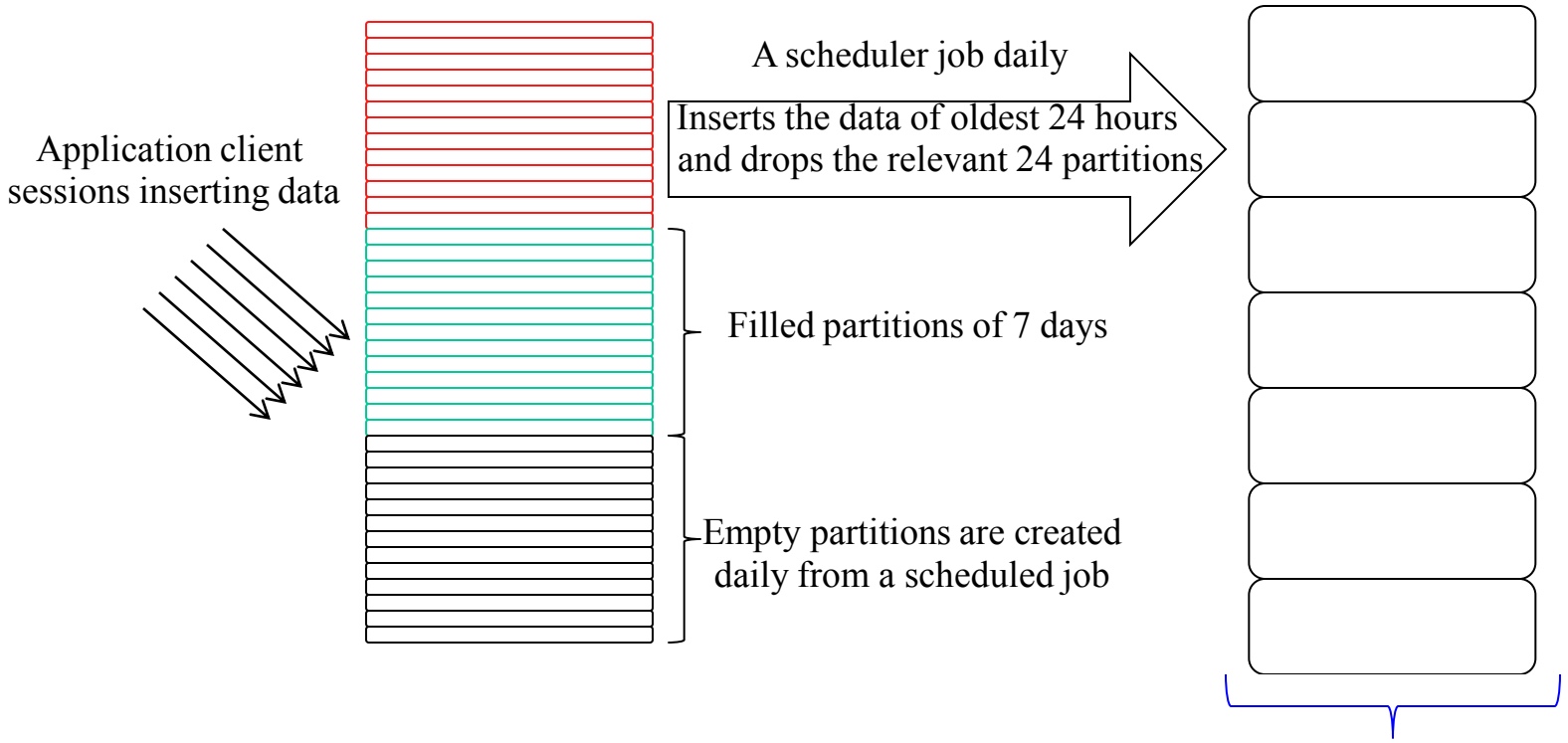
- To address the above, different organization was designed

The idea is NOT to rely on any indexes, but rather have the data ‘chopped’ on pieces appropriate for the queries plus apply data compression as second step.

Schema of the new traces data organization

T_TRACES –
partitioned on a 'time' column.
Each partition covers a time range of an hour

T_TRACESARCH (with compression) –
partitioned on a 'time' column.
Each partition covers a time range of 7 days



1) New partitions are created from a scheduler job weekly
2) The compressed data segments occupy **three times less space** in comparison with the non-compressed T_TRACES ones

A generic problem with the Oracle statistics gathering approach

- For queries that are interested in data of the most recent hours, often get non-optimal execution plan and thus consume a lot of resources.

  e.g. For the 'WHERE modiftime > SYSDATE -1/2' the Optimizer considers that there are only few rows relevant to that condition even if the statistics are very recent (computed from the last night). In reality, for a ½ day in several different schemas we could get tens or hundreds of thousands rows. With the wrong statistics Oracle produces non-optimal execution plans.

  A real case is where more then two indexes exist and Oracle decides for the inappropriate one or when a join of two tables is needed, Oracle chooses NESTED LOOPs within a index range scan is taking place instead of HASH JOIN. That leads to much more buffer reads (respectively IO and CPU)

  To address that problem, the queries need a lot of hints for instructing the Optimizer (e.g. INDEX_RS_ASC, NO_INDEX, CARDINALITY, USE_HASH ...etc), which is not easy to maintain.

TAG Production Data Details

TAG TYPE (GB)	# events (M)	DB Storage(GB)	Avg space per event (kB)
TAG – First pass 2010	1155	3978	3.8 (1)
TAG _ Repro September	557	940	1.6(2)
Monte Carlo 2009	865	2552	2.9

- 1-Average for latest events (first pass is a mix of ancient TAG data and merged TAG data and comm).
- 2- We are applying vertical partitioning to reprocessed data. Details here: ATLAS S&C 15July (http://indico.cern.ch/getFile.py/access?contribId=38&sessionId=16&resId=0&materialId=slides&confId=76895)
- Monte Carlo is not compressed (>255 columns)

Database organization of TAG data

- Data is organized in schemas per each pass. Eg:
- ATLAS_TAGS_DATA_F_2010 is for first pass 2010 data
- ATLAS_TAGS_DATA_R_SEP2010 is for September reprocessed data. It includes data from 2009 that was reprocessed as well
- ATLAS_TAGS_MC09 is Monte Carlo data generated for 2009.

- Pledge for the whole system is 2 passes of data per year. This has been traditionally estimated as 11.5 TB, but with TAG merging and compression enhancements, can be lowered to 5.3 TB per nominal year.

Database organization of TAG data

- With the advent of a central catalog of data (by PhD student Elisabeth Vinek) the distribution of data became transparent to the users of the system
- So, now data is distributed mainly for:
  - Redundancy, as there is no Oracle backup of this data
  - Closeness to the ELSSI WebSite (CERN, TRIUMF, BNL)
- Model of databases as storage elements
- With this new model, each database committed to TAGS has to provide the minimum amount of data for ONE pass, the lowest chunk of data we process.

- Estimate for 2011, 10 months at current rate:
  - 5.53 TB for first pass until the end of the year
  - 2.21 TB for full year reprocessing (avg reprocessed data was ~50%)
  - Monte Carlo was initially estimated at 8 TB, MC09 is 2.5 TB… quite overshot, expect the same for 2011

Minimum storage pledges per site

- In the TAG DB universe we have different policies:
  - BNL keeps two of the latest reprocessings.
  - TRIUMF keeps first pass data and several reprocessings. Regular cleanup of first pass is done.
  - CERN keeps everything but Monte Carlo.
  - DESY keeps first pass data, Monte Carlo and occasionally reprocessings
  - PIC only keeps latest reprocessing and some first pass data. Regular cleanup of first pass has to be done, as storage is limited.
  - RAL will keep Monte Carlo and one reprocessing (10TB) usable space.
- How do we manage this at the database level? What problems we face?

TAG DBAs Challenges

- First pass data:
  - Continuously uploaded.
  - Deletion should be made when data is reprocessed at least twice
  - NOT all data is reprocessed.
  - Partial Deletion => Fragmentation
- Strategy:
  - Switch default tablespace regularly
  - Move non-reprocessed data to new tablespace, only possible when it is small. Otherwise might as well leave it there, as space cannot be reclaimed.
  - Problem: work intensive
- Reprocessed data – will be frequent for TAGs
  - Nice to be exported in a single chunk.
  - Transportable tablespace technology a very good candidate for this
  - Results have been discouraging: network issues

TAG DBAs Challenges

- Speed of file transfer, single file, 940 GB September Repro:
  - CERN-> TRIUMF 18 GB/H (41 Mbit/s)
  - CERN-> BNL 313 GB/H (712 Mbit/s)
  - TRIUMF->BNL 3 GB/H (7 Mbit/s)
  - CERN-> PIC 3 GB/H
  - CERN-> DESY 3 GB/H
- Discouraging and disconcerting numbers
  - Databases not in OPN, but GPN
  - Can we find a way to schedule to achieve CERN<->BNL speed?
  - In talks with network people from CERN and at sites
- Need your collaboration as DBAs to liase with your network admins.
- Alternatives to TTS:
  - direct upload – weight on Tier-0 and TAG operations
  - Impdp – being tested, very slow, very manual.
  - ftp versus dbms_file_transfer?

Conclusions

- With the current successful year of datataking the data volumes on the ATLAS databases grown progressively.

  The challenge is to keep the DB applications that rely on the Oracle databases well tuned and perform as the user expects.

- To fulfill the above new design and tuning techniques were (or planned to be) put in place (some of them presented into these slides)