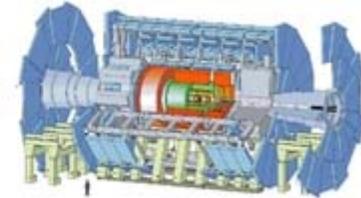




CHEP 2007, Victoria, B.C.



the **ATLAS Experiment**



Conditions Databases and Event Selection

Florbela Viegas

Richard Hawkings

Gancho Dimitrov



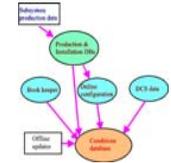
Databases Overview-Types of Data



- Data Storage in ATLAS:
 - File based Data : event data and large volume conditions data
 - Database based Data:
 - “where concurrent writes and transactional consistency are required;
 - where data handling is inherently distributed, typically with centralized writers and distributed readers;
 - where indexing and rapid querying across moderate data volumes is required; and
 - where structured archival storage and query-based retrieval is required. “
- (ATLAS Computing TDR)



Databases Overview-Types of Data



- So, by this criteria, we have resident in the database:
 - **LCG COOL conditions database**, for DCS (Detector control system) and non-DCS data used in offline reconstruction/analysis.
 - PVSS Oracle Archive, i.e. the archive for the DCS « slow controls » data.
 - Detector configuration and connectivity data, specific subdetector data
 - **Event Metadata –aka the TAG database.**
 - File catalog (DQ2) – The DDM (Distributed Data Management) central catalog for file distribution across the grid.
 - Bookkeeping Production System (T0) – Job scheduling and tracking at CERN Tier-0.



Databases Overview - Size



- ATLAS Database Storage Requirements: ~16 TB in 2009 at CERN *

Application with indexes	Estimate (cap/year)
Conditions Data	0.5 TB - 2008 1 TB - 2009
TAGS Event Metadata	1 TB – 2007 1.6 TB – 2008 6 TB – 2009
PVSS Data	2-3 TB x 2 - 2008 and 2009

- Tier -1 will host Conditions Data, and some are going to upload TAGS data as well, but not all.
- * (plus 1 TB for file catalog, monitoring and bookkeeping system and other data)



Databases Overview - Technology



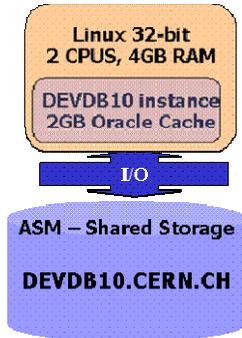
- Which database technology ?
 - Preferred Technology: SQL-based relational databases
- ATLAS values the flexibility in using different RDBMS engines:
 - Oracle - for a large-scale production-level DB Service, taking advantage of the experience of the CERN-IT Physics Database Service
 - MySQL – installations outside of institutional IT departments.
 - SQLite – SQL relational DB access combined with local file-based storage, specially suited for replicas of small subsets of data.
- We maintain database transparency through the client software:
 - Development of the Relational Access Layer (CORAL) within the LCG POOL project to provide performance-optimized, vendor-neutral, C++-based DB access.
 - ATLAS utilizes CORAL both directly and through layered services for DB storage via C++ with either Oracle, MySQL or SQLite back-end engines.



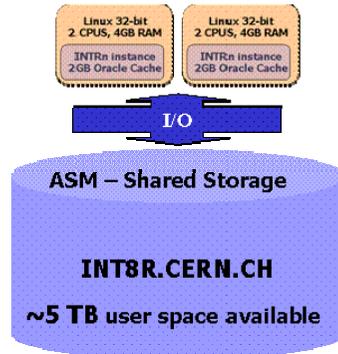
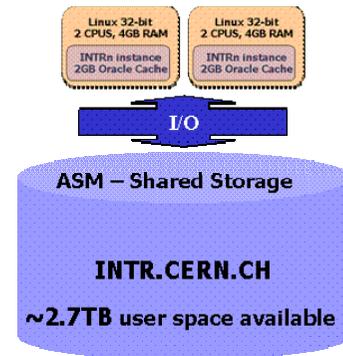
Databases Overview-Technology



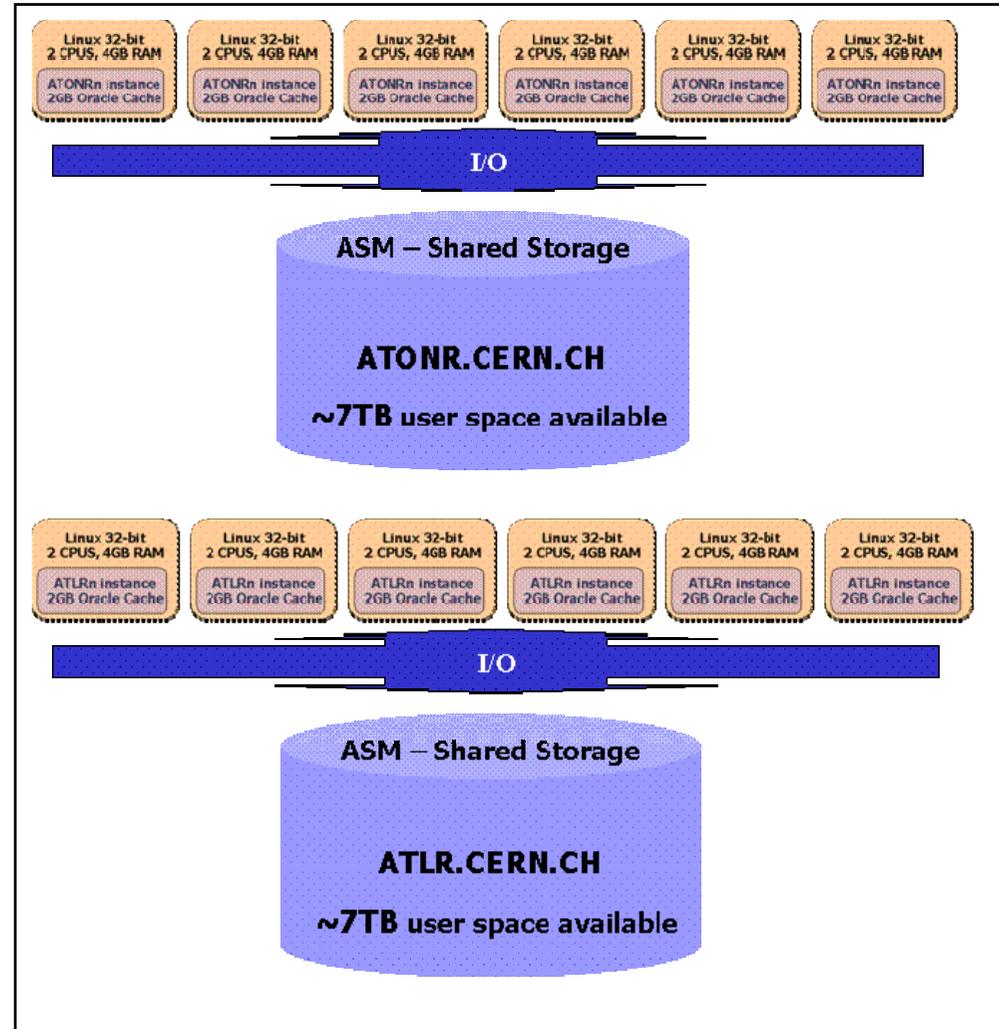
Development



Validation/Testing



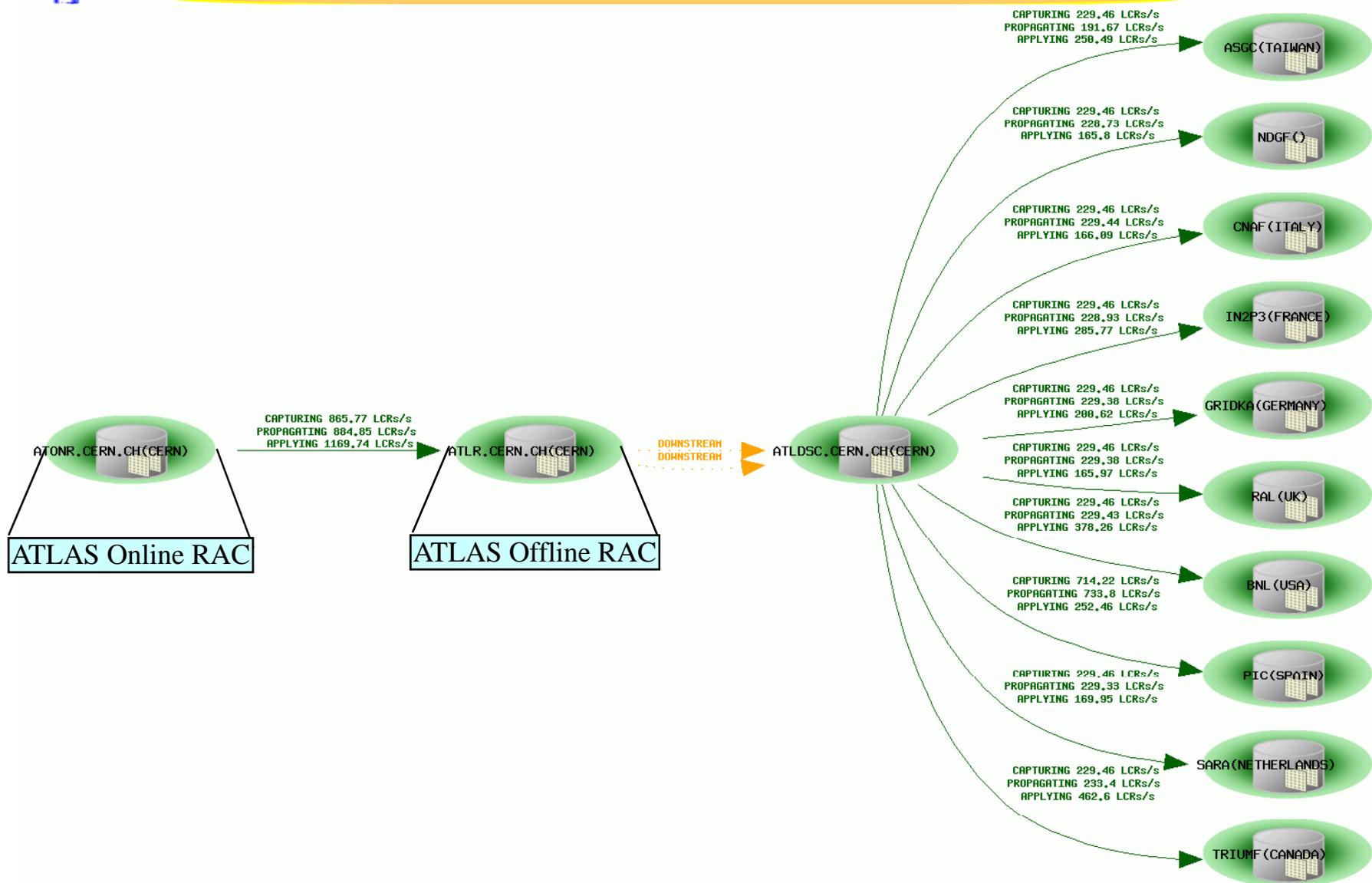
Production



Upgrade 2008Q1
3 quadcore nodes per cluster
16 GB FB-DIMM per node
Expected CPU throughput increase: ~300%
Expected performance increase for non-parallel operations. Ex: streams apply +60%



Databases Overview-Replication

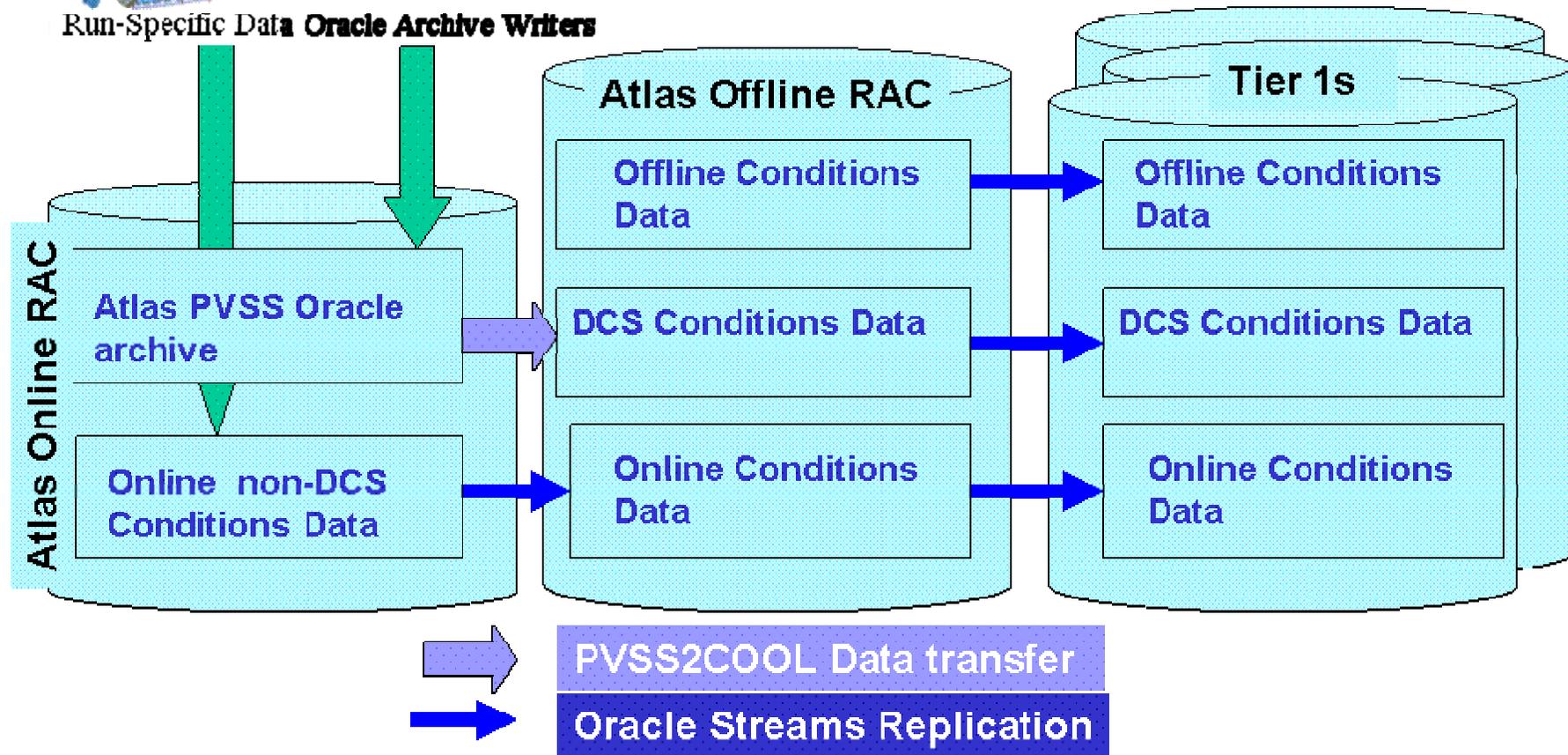




Conditions Data



Run-Specific Data **Oracle Archive Writers**





Conditions Data



- Input Sources at the Online Point:
 - Online DAQ information, Run control, central trigger, High level trigger, Monitoring Services
 - Subdetector specific information, configuration, monitoring and calibration data
- Input Sources at the T0 level
 - Replicated data streamed from the Online Database
 - PVSS2COOL process will update regularly the DCS Conditions data at T0, rate of 1MB/minute, making up most of the COOL data at the offline level (1.4 Gb out of 1.7Gb per day as per best estimates at this time)
- Replicated data from Online Database is read-only down-stream.



Conditions Data



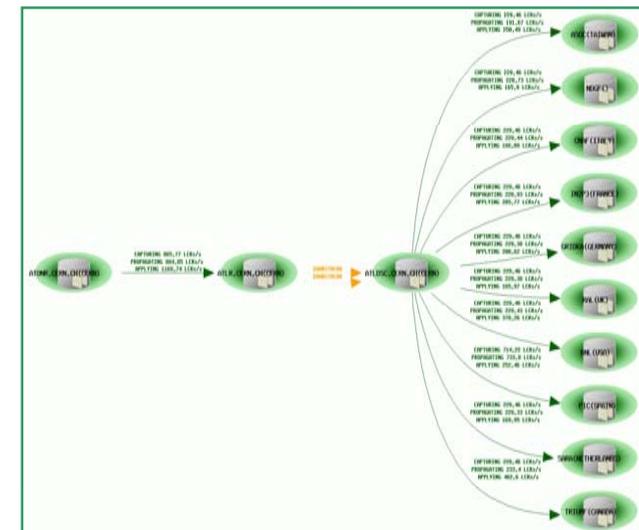
- Some Numbers:
- ATLAS daily reconstruction and/or analysis job rates will be in the range from 100k to 1M jobs/day
- For each of ten Tier-1 centers that corresponds to the Conditions DB access rates of 400- 4000 jobs/hour
- Each reconstruction job will read 10-100MB of data
- Atlas requests to Tier-1s is a 3-node RAC cluster dedicated to the experiment.
- Expected rate of data flow to Tier-1s is between 1-2 GB/day



Oracle Streams Replication Tests



- Production phase started on April, 1st 2007, with 6 destinations
- We have now the **10** destinations actively receiving data
- Since April, 13th more than **60GB** of COOL test data have been replicated
- A cron job runs twice per hour adding one run's worth of data, roughly 20 MB per run, which amounts to 1GB/day volume. Tests have been successful in increasing the volume to **2GB/day** over several days. Problems arise mainly with memory and CPU issues on the replicating machine.
- When a Tier-1 has a failure, procedures are in place to isolate the site and make it « catch up » with the others.
- These procedures were used several times during these tests and were successful
- There was a formal recovery exercise on June, 13th, on the **3D DBA Workshop**, which involved most of the Tier-1s DBAs.

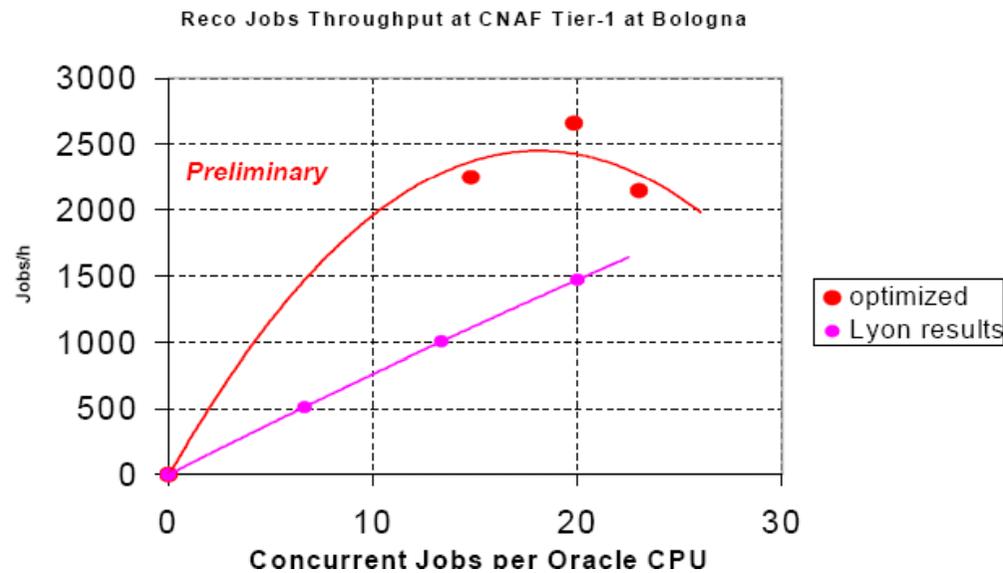




Conditions Data



- Client Stress Tests at IN2P3 and CNAF:



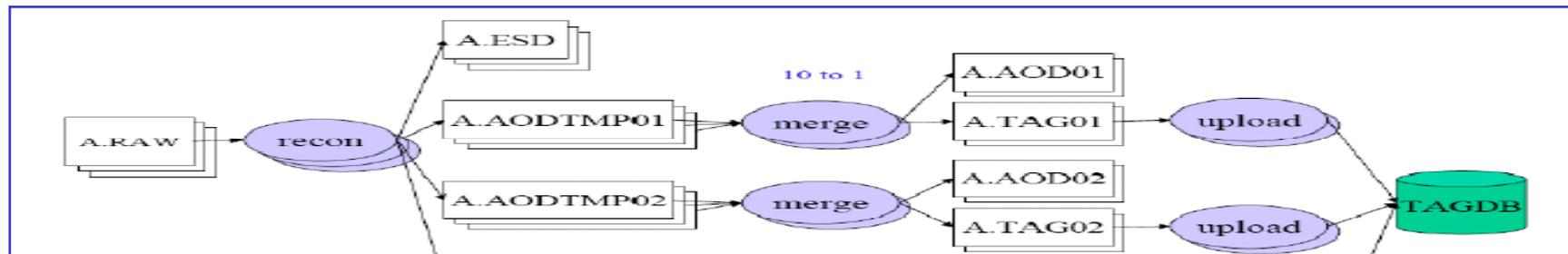
- Results of tests within ATLAS requirements (400-4000 jobs/hour) , further improvements expected with new COOL release.



TAGS - Event Metadata



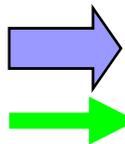
- Tag Data (TAG): When an event is written, metadata about the event (a “tag”) is exported along with references to the event data. These TAG files hold about 1 kB per event.
- Relational TAGS enables the searching of interesting events at a Terabyte scale, in a highly efficient way, instead of opening Petabytes of files.



- Sample queries issued for analysis:
 - « Give me all the events with at least 2 electrons and missing Et greater than 10GeV »
 - « Give me the list of AOD files that satisfy a given query » (e.g. TNT GANGA plug-in typical query for creation of ROOT file. This is used as input to grid jobs in GANGA)

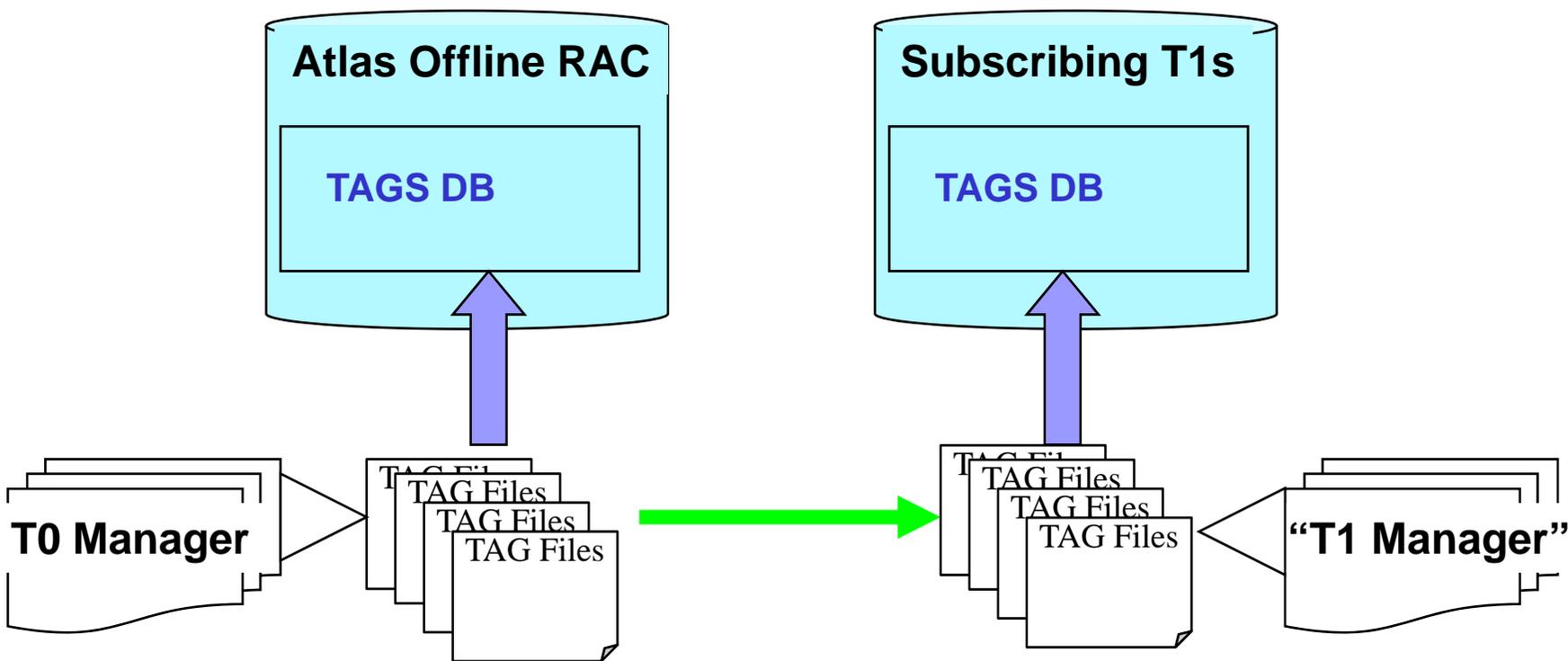


TAGS - Event Metadata



POOL Collection utilities

DDM File Replication





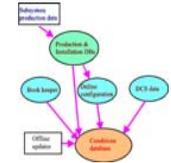
TAGS - Event Metadata



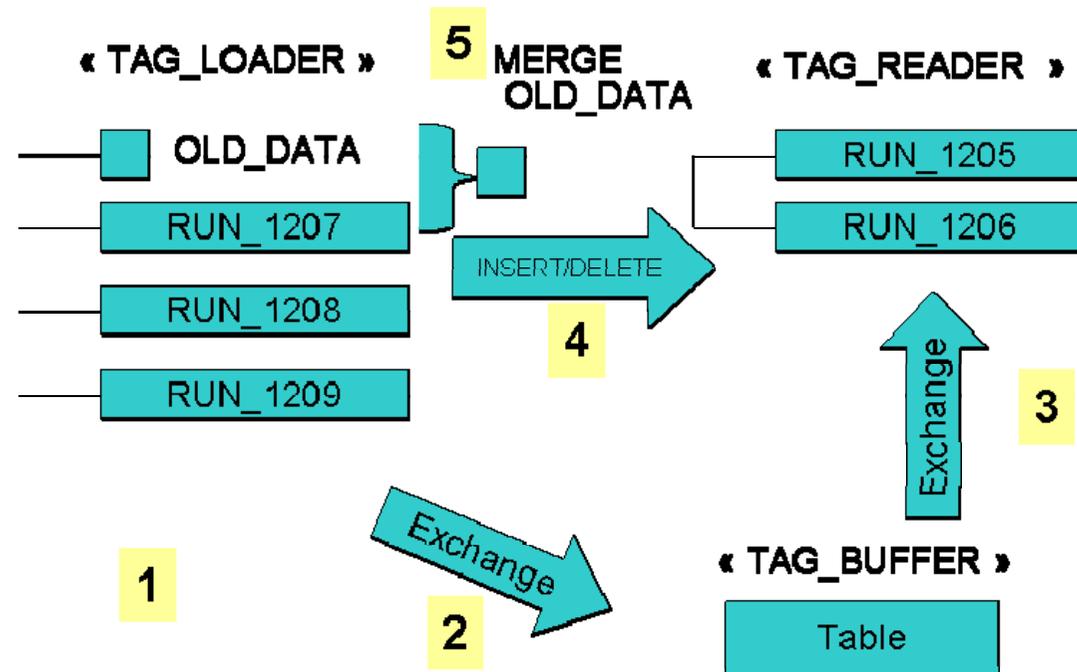
- Event metadata is read from TAG files, and stored in the database as POOL Collections. These are supported in ROOT format as well.
- TOM at T0 – responsible for the automatic load of the TAG files, as they become available.
- T1s will receive the files, but most don't have database storage capacity for the loading them in the database. Select subscribers are volunteering.
- Challenges of Event Metadata in the data base:
 - Huge amount of data per year ~6TB+
 - Unpredictability of querying criteria suggests full indexing BUT the 200Hz loading rate requirement does not allow this.
 - Different use cases:
 - Web query
 - TNT Navigation Tool.
 - Pool Collection utilities.
 - Restrictions of queries to available resources e.g. limit the number of active queries in a system while keeping fair-share of usage.



TAGS - Event Metadata



- Schema Architecture tested at last year T0 TAGS loading test
- Partitioned architecture tested.
- Concept of LOADER table and READER table, one with no indexes and the latter with indexing when partition is exchanged from LOADER to READER
- Test successful for achieving the 200Hz loading rate. A lot of open questions and followups on subsequent tests





TAGS - Event Metadata



- To test if the technology was capable of handling such a complex problem, a 10^9 row database(1 TB+ storage) with known distributions was built.

- Details and Conclusions of these tests are presented in this conference:

<http://indico.cern.ch/contributionDisplay.py?contribId=161&sessionId=31&confId=3580>

- Example of an optimized query made in this database:

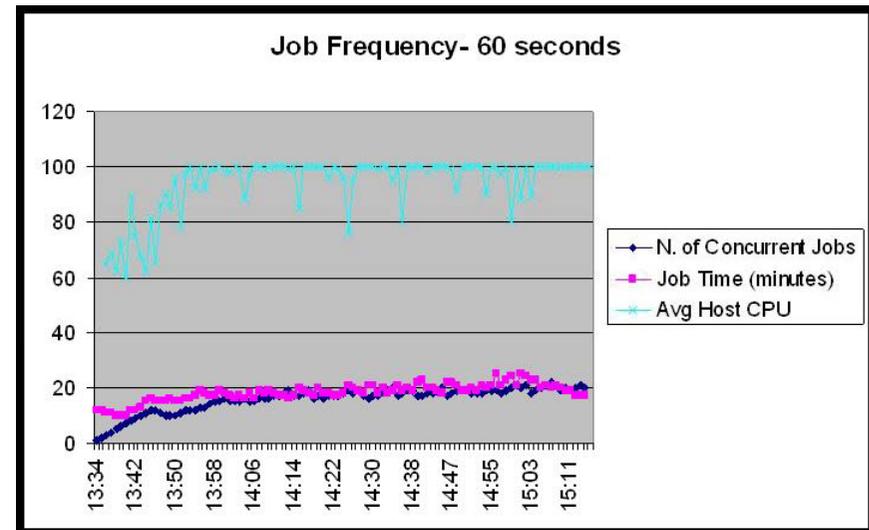
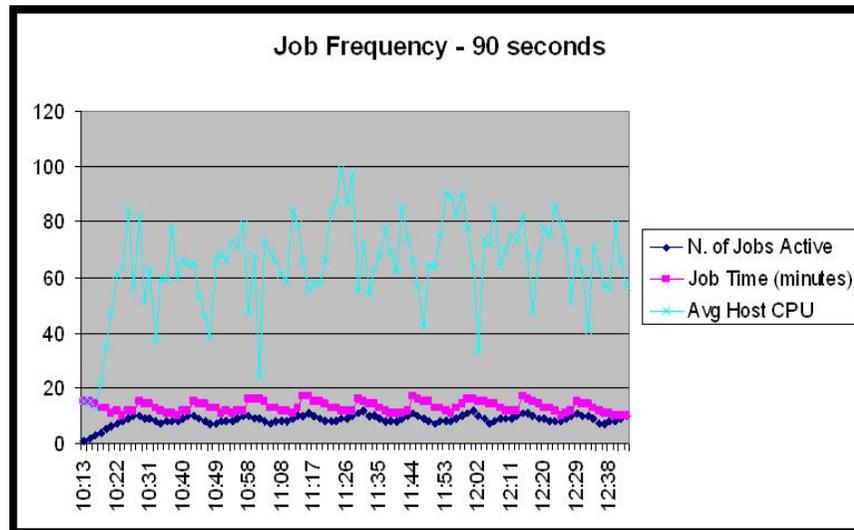
```
select /*+ index_join(e ICMg1_3_nor1num01 ICMg1_3_nor10num01
ICMg1_3_nor100num01 ICMg1_3_ID) */
id from event_g1_3_o100m partition (runnr_1441_1480) e
where id > 1 and nor1num01 > 498000 and nor10num01 > 485000 and nor100num01
> 300000
INTERSECT
select /*+ index(e2 ICMg1_3_ID)*/ id from event_g1_3_o100m partition
(runnr_1441_1480) e2 where id is not null and rowid in (
select /*+ index_combine(e) */
rowid from event_g1_3_o100m partition (runnr_1441_1480) e
where enumuni1000num01 > 25 and enumuni100num01 > 3 and
uni10Knum01 < 9900);
```



TAGS - Event Metadata



- Stress Tests were also made, to determine CPU and I/O capacity needed.



- Each job divides its time between CPU and I/O, having some cluster activity as well when saturation is not near.
- I/O rate topped at 35MB/sec on 60 sec intervals, 25MB/second on 90 sec interval.
- 1 Job every 90 seconds generates ~9000 queries per day



Challenges and Looking ahead



- Individual components are being tested
- Most limits known. (Oracle Streams throughput, indexing overhead, partitioning strategies, maximum I/O rates and others)
- Conditions Data: individual performance issues and Streaming caveats
- TAGS: testing of architecture and streaming to T1/T2
- Long-Running real data tests: Cosmic Run started August 23rd → putting it all together, from online to T1s
- Ongoing improvement: Backup and Recovery, Disaster Recovery, availability of online, shift work, addressing user error failure.

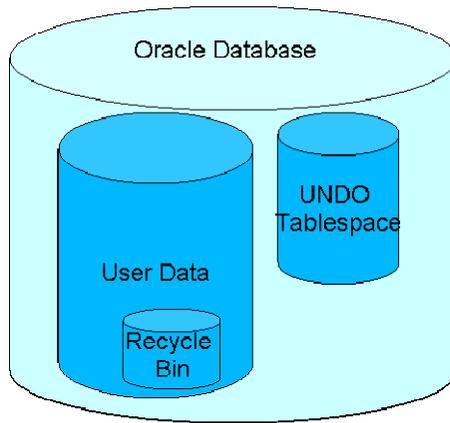




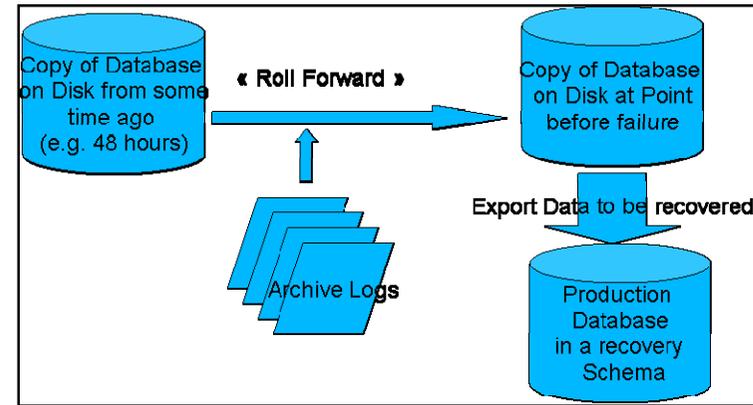
Data Protection



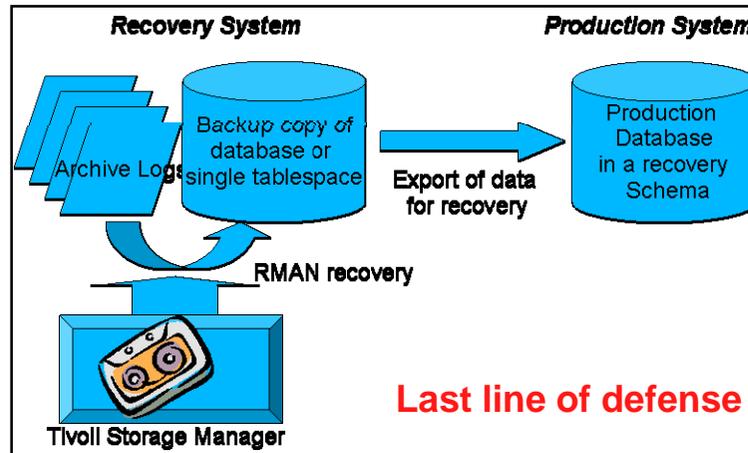
- Backup and Recovery policy agreed with IT PSS and already tested.



First line of defense



Second line of defense



Last line of defense



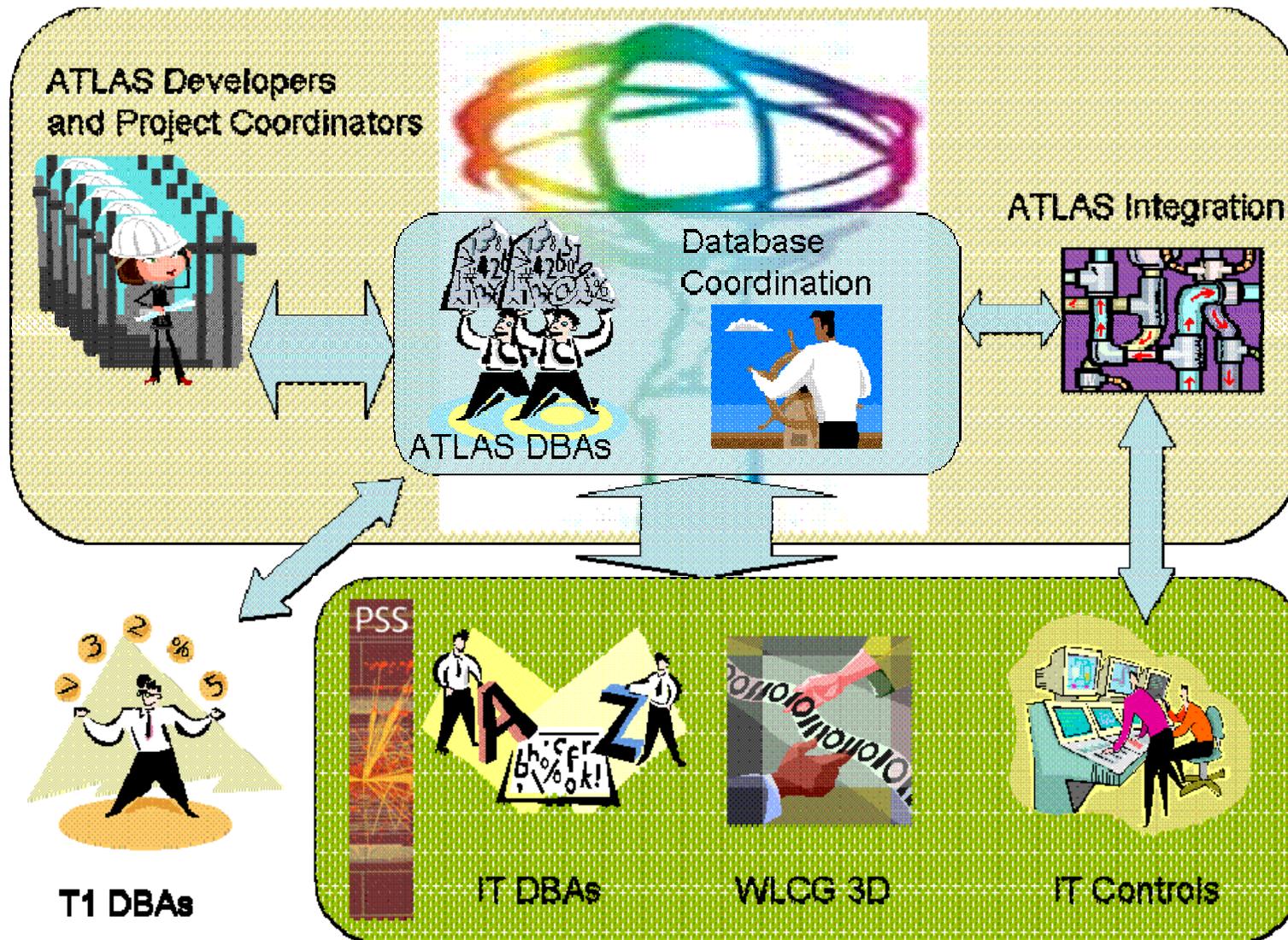
Data Protection-Monitoring



- The monitoring of the database is a highly critical task for data protection and assurance of service.
- The databases are monitored by:
 - Standard Oracle tools, by ATLAS DBAs and IT, with sets of alerts and management activities
 - IT developed it's own monitoring tools for the Oracle Databases, reports weekly on activity to ATLAS
 - ATLAS DBAs developed additional tools for monitoring at the Subdetector and Service level. This empowers users to look at their own applications in a DBA fashion: Session history, active sessions, SQL performance, database resource usage, etc.
 - Many tasks are automated e.g. grants and synonyms between application owners, readers and writers.



Human Resources – Effort and Interaction





Human Resources – Effort and Interaction



- Open door policy with ATLAS Developers, problems are discussed and solutions found before moving to production
- Great effort in reusing code and sharing information between developers and experiments, such as:
 - COOL and POOL development
 - PVSS project
 - WLCG 3D Team
- Great collaboration between IT groups and ATLAS Database group
 - Interaction well defined on administration tasks
 - Agreement of policies, like security and backup
 - Knowledge-sharing and problem-solving together
 - Human effort optimized for both groups – IT focuses on SLA tasks and ATLAS DBAs on the application management.



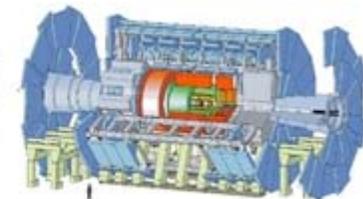
« Will it all work? »



- Yes, because:
- Technology Knowledge and Testing – by now we have a long experience with the technology, very few surprises.
- Procedures and Protocols in place – Monitoring and response procedures are ready and being deployed
- Hardware upgrade will increase capacity many-fold – upgrade of hardware, evolution of software also in play.
- Synergy between all human resources involved – Positive attitude of collaboration between all players, to a common goal in the form of: knowledge sharing, harmonization of practices, roles well defined.



the **ATLAS Experiment**





Additional Slides





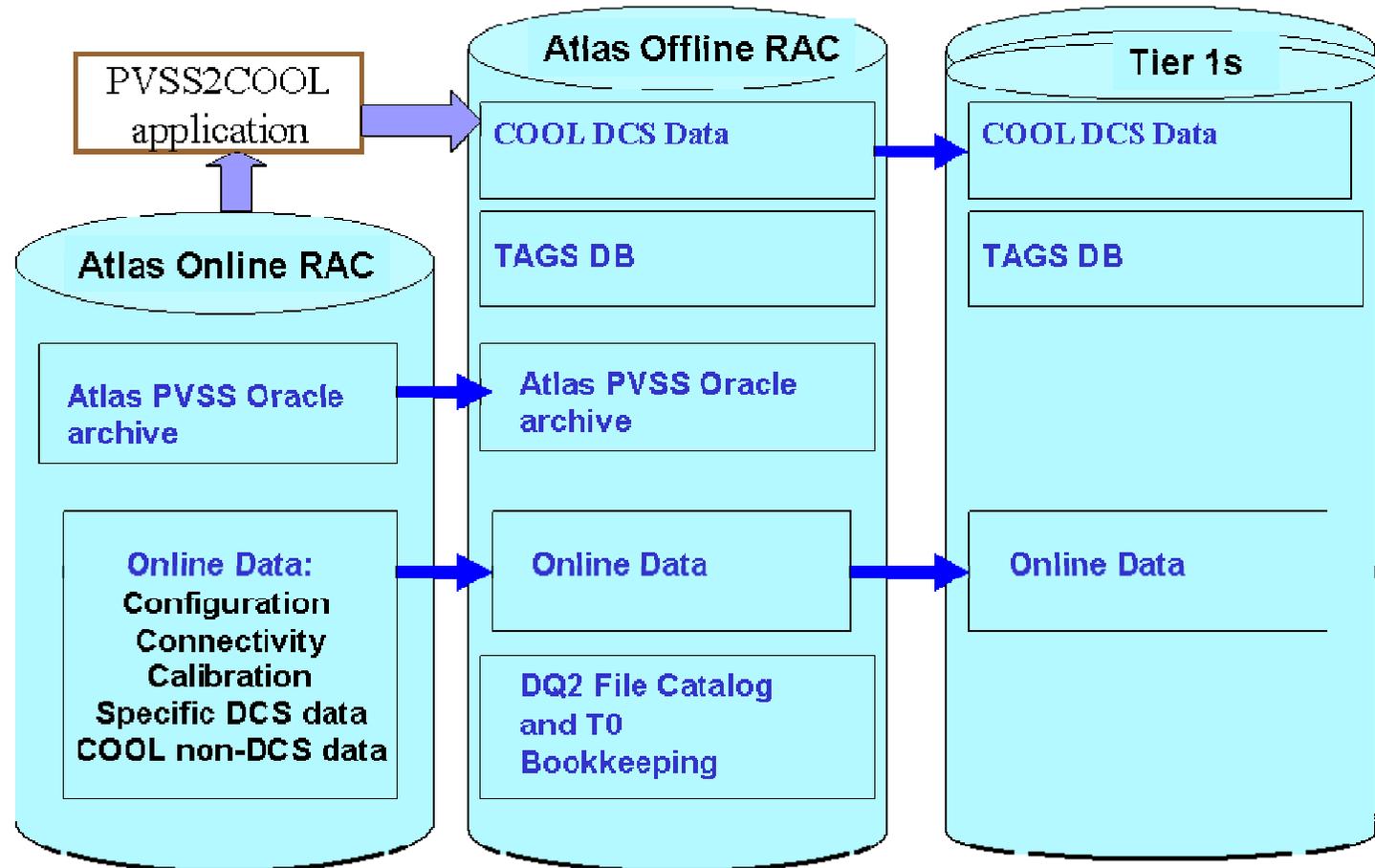
For more information



- ATLAS Computing TDR Database Section – <http://atlas-proj-computing-tdr.web.cern.ch/atlas-proj-computing-tdr/Html/Computing-TDR-35.htm>
- ATLAS Database Project Twiki: <https://twiki.cern.ch/twiki/bin/view/Atlas/DataBases>
- IT Physics Database Services Twiki: <https://twiki.cern.ch/twiki/bin/view/PSSGroup/PhysicsDatabasesSection>
- _Related sessions in this conference:
- Database Operations: Sasha Vaniachine in [Development, Deployment and Operations of ATLAS Databases](#)
- Event Metadata: Helen McGlone in [Building a Scalable Event-Level Metadata System for ATLAS](#)
- Conditions Data and Software: Andrea Valassi in [COOL Software Development and Service Deployment Status](#)
- IT Database Services: Maria Girone in [CERN Database Services for the LHC Computing Grid](#)



Databases Overview-Data Flow





TAGS and COOL - Challenges and lessons learned



- Cutting Edge technology – pushing the limits of Oracle RDBMS :
 - Streaming TAGS not achieved due to limits on the Oracle Streams software, workarounds possible, but benefit-risk ratio not good.
 - Streaming PVSS from online to offline, challenge in throughput and workaround application code. A success due to the effort of subdetectors to minimize amount of data.
 - Partitioning and indexing strategies « thinking out of the box » for TAGS and COOL . Huge challenges in performance, storage space and application transparency
 - Resource management for throughput (TAGS), still being studied how to limit the amount of resources between applications, keeping a fair share of use for analysts.
- Challenge: taking advantage of technology but still maintaining technology independence (important for T2 and « laptop » analysis)
- We have one of the largest distributed database systems world-wide up and running – CERN IT PSS (Physics Support Services) statement from Oracle