Argonne NATIONAL LABORATORY

... for a brighter future

Max-Planck-Institut für Physik (Werner-Heisenberg-Institut)

Development, Deployment and Operations of ATLAS Databases

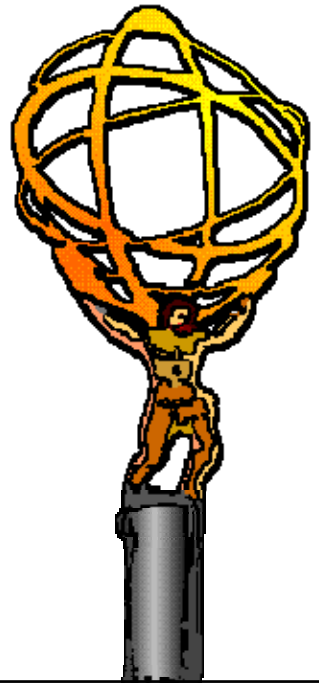
XVI International Conference on Computing in High Energy and Nuclear Physics

Victoria, British Columbia, Canada

September 5, 2007

Hans von der Schmitt (MPI for Physics, Munich)

Alexandre Vaniachine (Argonne)



U.S. Department of Energy

UChicago Argonne LLC

Office of Science U.S. DEPARTMENT OF ENERGY

A U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

Outline and References

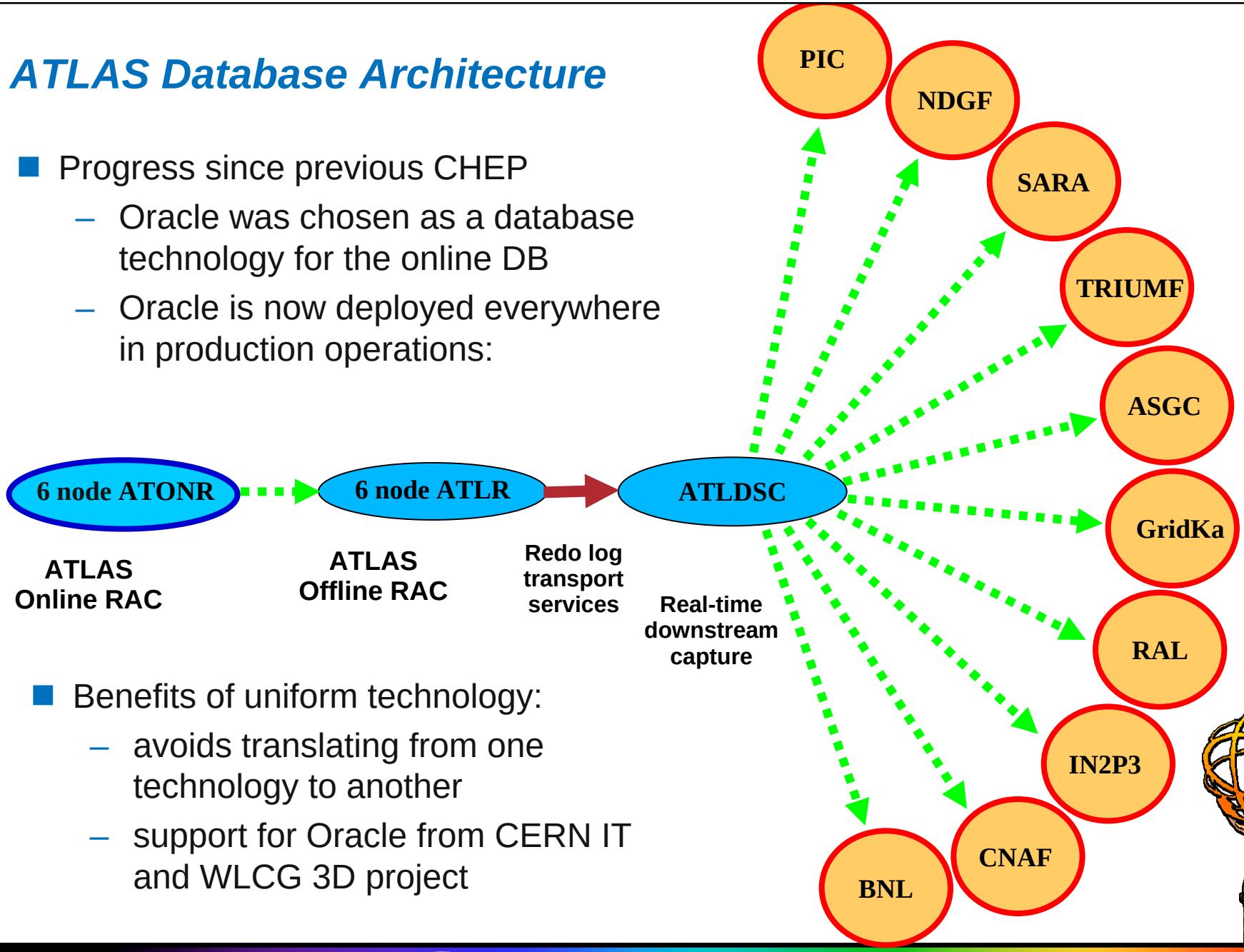
- ATLAS Database Architecture and Applications:
  - PVSS DB and Conditions DB Challenges
- Distributed Database Operations in Production
- ATLAS Scalability Tests: purpose, first results and next steps
- Work-in-progress snapshots:
  - Ramping Up Database Capacities for ATLAS
  - Replication of Conditions DB Payload Data to Tier-1s
  - Muon MDT Calibration Operations
  - Database Operations Monitoring
- Conclusions and Credits

We wish to thank all our ATLAS collaborators who contributed to and facilitated ATLAS database development, deployment and operations

- Their activities are presented in many talks at this Conference:
  Mario Lassnig et al. [id 64]; Alexei Klimentov et al. [id 84]; Caitriana Nicholson et al. [id 85]; Monica Verducci [id 90]; Kathy Pommes et al. [id 100]; Florbela Viegas et al. [id 122]; Helen McGlone et al. [id 161]; Ricardo Rocha et al. [id 255]; Antonio Amorim et al. [id 333]; Solveig Albrand et al. [id 430]; Pavel Nevski et al. [id 450]; Manuela Cirilli et al. [id 462]

ATLAS Database Architecture

- Progress since previous CHEP
  - Oracle was chosen as a database technology for the online DB
  - Oracle is now deployed everywhere in production operations:



- Benefits of uniform technology:
  - avoids translating from one technology to another
  - support for Oracle from CERN IT and WLCG 3D project

Distributed Database Applications

A subset of ATLAS database applications must be distributed world-wide (for scalability) since they are accessed by many computers on the Grid:

- **Geometry DB** – ATLASDD
  - developed by ATLAS
  - first ATLAS database application deployed worldwide
  - distributed world-wide in SQLite files (a part of Database Release)
- **Conditions DB** – COOL
  - developed by LCG with ATLAS contributions
  - distributed world-wide via Oracle streams
  - now in production operations
- **Tag DB** - event-level metadata for physics
  - developed by LCG with ATLAS contributions
  - to be distributed world-wide in files
  - now in central operations for large-scale trials of Event Data streaming models (explicit vs. implicit event streams)
    - *Helen McGlone presents more details later in this session*

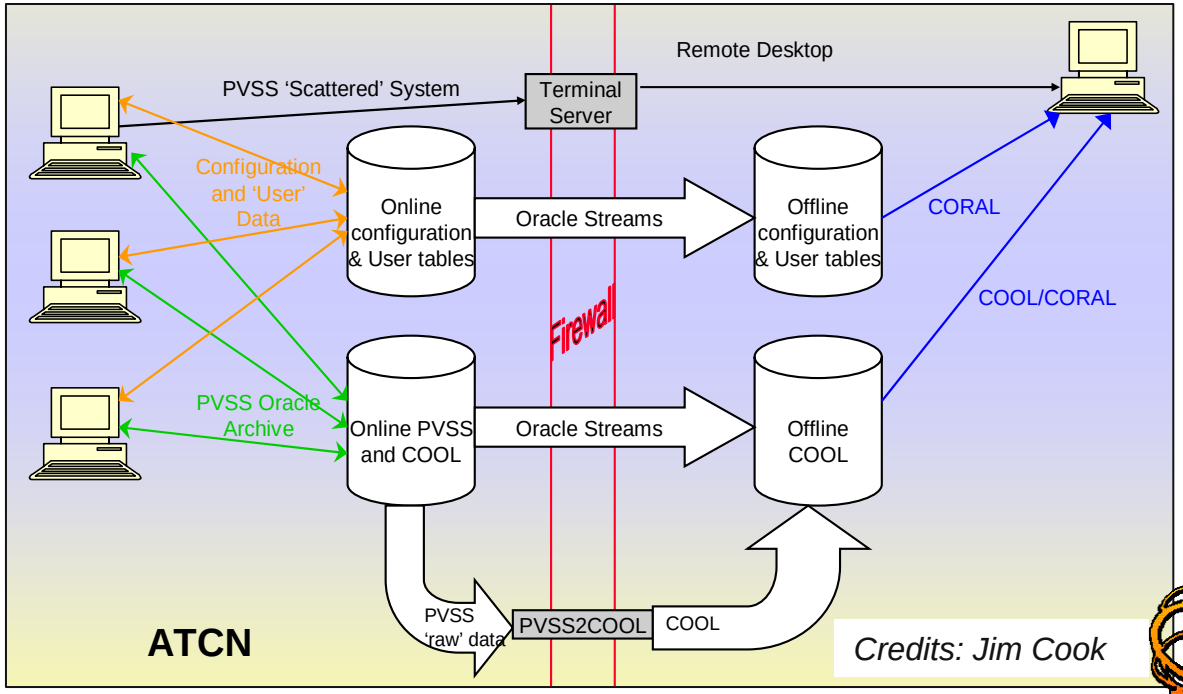
Centralized Database Applications

Those database applications that are accessed by people or by limited number of computers do not have to be distributed world-wide:

- Technical Coordination databases:
  - Detector construction information
- Online databases
  - Slow control data (DCS) stored in the PVSS DB
  - Trigger DB
  - Online databases for subsystems
- Computing Operations databases:
  - Task request database
    - *physics tasks definitions for job submission*
  - Production System DB
    - *job configuration and job completion*
  - Distributed Data Management (DDM) databases:
    - *Dashboard DB from ARDA*
- AMI (ATLAS Metadata Information) database
  - “Where is my dataset?”

Online Databases Replicated for Offline Operations

- However, centralized databases must also be replicated (for other reasons)
- e.g., for Data Quality assessment in ATLAS Data Preparation operations:
  - Online server must not be overloaded by queries used for diagnostics
  - Offline PVSS replica - data are accessible from CERN and outside
- PVSS DB replication via Oracle streams is a challenging and uncharted area:
- Up to 6 GB/day have to be replicated
- PVSS is a commercial product – the PVSS DB schema can't be modified
- In a joint effort with WLCG 3D project the PVSS replication technology was developed by Florbela Viegas and Gancho Dimitrov
  - Current status: two months of successful operations
  - We will share PVSS replication technology with other LHC experiments



Credits: Jim Cook

ATLAS Conditions DB Challenges

- Unprecedented complexity of LHC detectors
  - orders of magnitude more channels than in previous experiments
- Detector operations in a new “Continuous Running” paradigm
  - no start/stop of run upon changes in run conditions
- To address these challenges ATLAS adopted Common LHC solution for Conditions DB – COOL:
  - COOL separates the Interval-of-Validity metadata from the actual Conditions data (“payload”)
  - select “payload” data are stored in files outside of the database server
    - *e.g. calorimeter calibration constants that are not really suited to relational databases because of their size and access patterns*
- COOL technology is successfully deployed in production for ATLAS detector commissioning operations
  - such as major detector commissioning exercise – the M4 Cosmics Run in August
- Current ATLAS Conditions DB operations snapshot:
  - 70 GB of data
  - each day up to 1 GB of COOL data is added and replicated
  - data growth rates are increasing as more subdetector elements are being instrumented during ATLAS detector commissioning
  - average load in current operations: 500 sessions

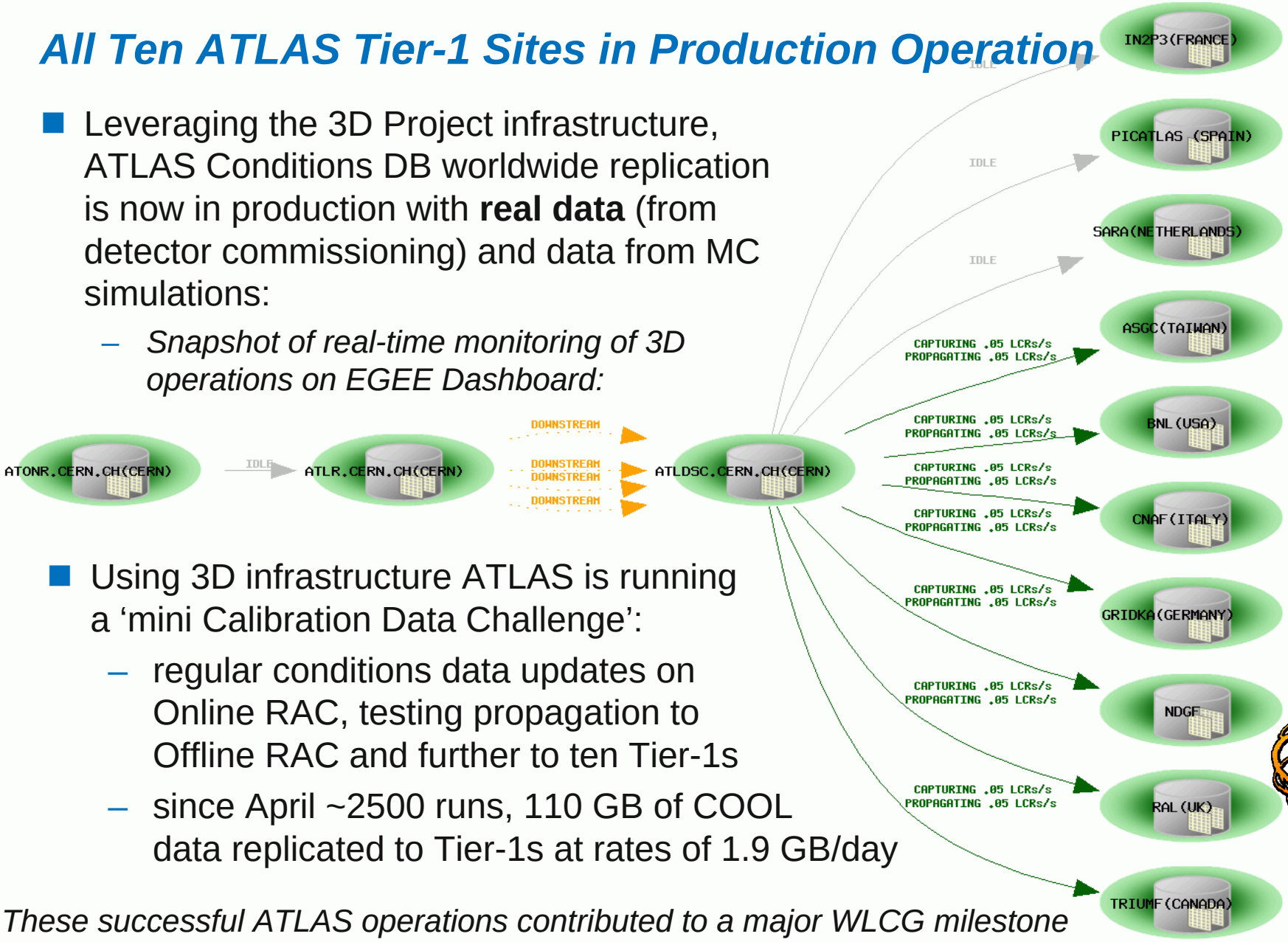
Distributed Database Operations

Access to Conditions DB data is critical for event data reconstruction

- To achieve scalability in Tier-0 operations slices of the corresponding conditions/calibrations data will be delivered to Tier-0 farm via files on afs
- Beyond Tier-0 we need a different technology for data distribution
  - 3D Oracle streams to Tier-1 sites
    - *for replication of fast-varying data*
      - such as Conditions DB data
  - ATLAS DDM DQ2 for file-based database data replication
    - *for replication of slow-varying data*
      - such as 'static' Database Releases (Geometry DB, etc.)
    - *for replication of large files with Conditions DB payload*
- Critical operational infrastructure for these technologies is delivered by:
  - ATLAS Distributed Data Management (DDM)
    - Alexei Klimentov et al. [id 84]
  - WLCG Project on Distributed Deployment of Databases (3D)
    - Dirk Duellmann [id 171]

All Ten ATLAS Tier-1 Sites in Production Operation

- Leveraging the 3D Project infrastructure, ATLAS Conditions DB worldwide replication is now in production with **real data** (from detector commissioning) and data from MC simulations:
  - *Snapshot of real-time monitoring of 3D operations on EGEE Dashboard:*



- Using 3D infrastructure ATLAS is running a ‘mini Calibration Data Challenge’:
  - regular conditions data updates on Online RAC, testing propagation to Offline RAC and further to ten Tier-1s
  - since April ~2500 runs, 110 GB of COOL data replicated to Tier-1s at rates of 1.9 GB/day

These successful ATLAS operations contributed to a major WLCG milestone

Major WLCG Milestones Accomplished

11.06.2007		WLCG High Level Milestones - 2007												
			Done (green)				Late < 1 month (orange)				Late > 1 month (red)			
ID	Date	Milestone	ASGC	CC IN2P3	CERN	FZK GridKa	INFN CNAF	NDGF	PIC	RAL	SARA NIKHEF	TRIUMF	BNL	FNAL
3D Services														
WLCG-07-09	Mar 2007	**3D Oracle Service in Production** Oracle Service in production, and certified by the Experiments						Jun 2007						squid frontier
WLCG-07-10	May 2007	**3D Conditions DB in Production** Conditions DB in operations for ATLAS, CMS, and LHCb. Tested by the Experiments.												squid frontier

Steady progress towards a major achievement:

- 3D production operations for Conditions DB started on April 1 with six active sites: GridKa, RAL, IN2P3, CNAF, SARA and ASGC
- BNL and TRIUMF were included in the beginning of May
- NDGF site joined in the first week of June
- PIC (Barcelona) was connected in 3D production in the third week of June

Production operations were tested by ATLAS experiment:

- Replication is used for reading back conditions data at the Tier-1s
  - 3D production operations with real data started in July

Together we have accomplished 'something big' (see next slide)

Fruitful Collaboration with WLCG 3D Project

PSS

Conclusions

CERN IT Department

- We have one of the largest distributed database systems world-wide up and running!
  - Many operational problems will still have to be resolved but the basic infrastructure, operational procedures and monitoring are in place.
- The contribution of the ATLAS database team has always been essential and continues to push the limits of the 3D project.
- The conclusions from the ATLAS scalability tests will be the next major milestone to validate the 3D setups at CERN and Tier 1

Dirk.Duellmann@cern.ch

Credits: Dirk Duellmann

DB Replicas at Tier-1s: Critical for Reprocessing

What for we are building such large distributed database system?

- ATLAS Computing Model provides following requirements at Tier-1 with respect to Conditions DB:
  - Running reconstruction re-processing: O(100) jobs in parallel
  - Catering for other 'live' Conditions DB usage at the Tier-1 (Calibration and Analysis), and perhaps for the associated Tier-2/3s
- To provide input to future hardware purchases for Tier-1s
  - *How many servers required, balance between CPU, memory and disk?*

we have to do Oracle scalability tests with the following considerations:

- Although reconstruction jobs last for hours, most conditions data is read at initialization
  - *Unlike the HLT case, we do not have to initialize O(100) jobs at once*
- Tier-0 uses file-based Conditions DB slice, at Tier-1 DB access differs
  - *Because of rate considerations we may have to stage and process all files grouped by physical tapes, rather than datasets*
- Will the database data caching be of value for the Tier-1 access mode?
  - *Our scalability tests should not rely on data caching:*
    - **We should test random data access pattern**

Realistic Workload for Conditions DB Tests

- ATLAS replications workload is using multiple COOL schemas with mixed amount and types of data:

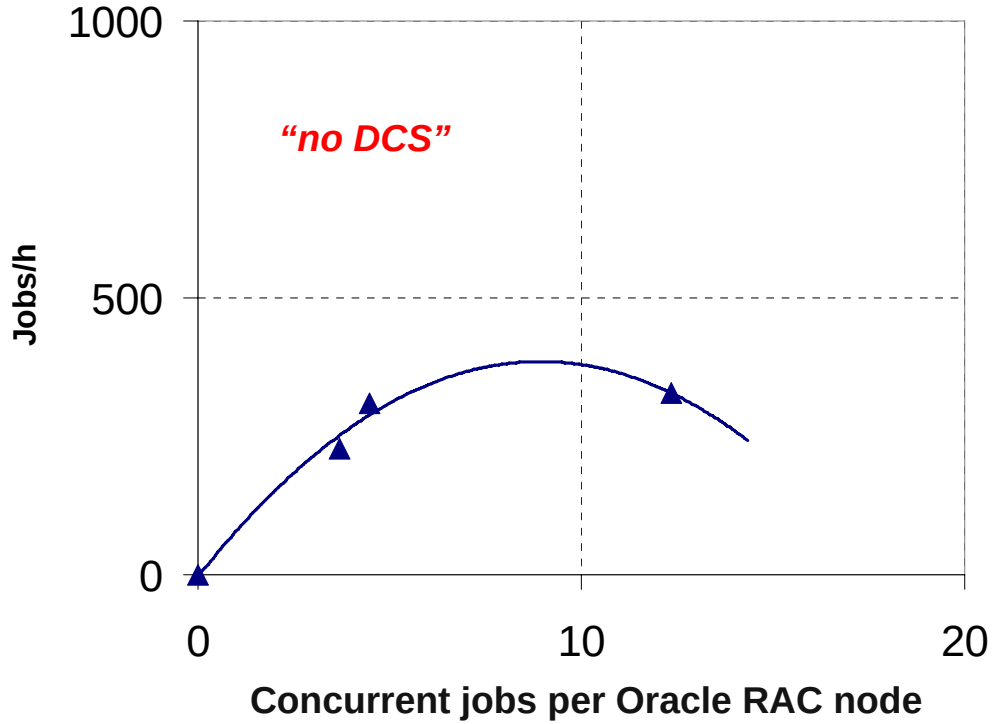
Schema	Folders	Channels	Channel payload	N/Run	**Total GB**
INDET	2	32	160 char	1	**0.21**
CALO	17	32	160 char	1	**1.8**
MDT	1+1	1174	CLOB: 3kB+4.5kB	0.1	**17.0**
GLOBAL	1	50	3 x float	6	**0.25**
TDAQ/DCS	10+5	200+1000	25 x float	12	**80.0**
TRIGGER	1	1000	25 x float	12	**8.0**

'best guess'

- Replicated data provided read-back data for scalability tests
  - The realistic conditions data workload:
    - *a 'best guess' for ATLAS Conditions DB load in reconstruction*
      - dominated by the DCS data
- Three workload combinations were used in the tests:
  - "**no DCS**", "**with DCS**" and "**10xDCS**" (details on slide 26)

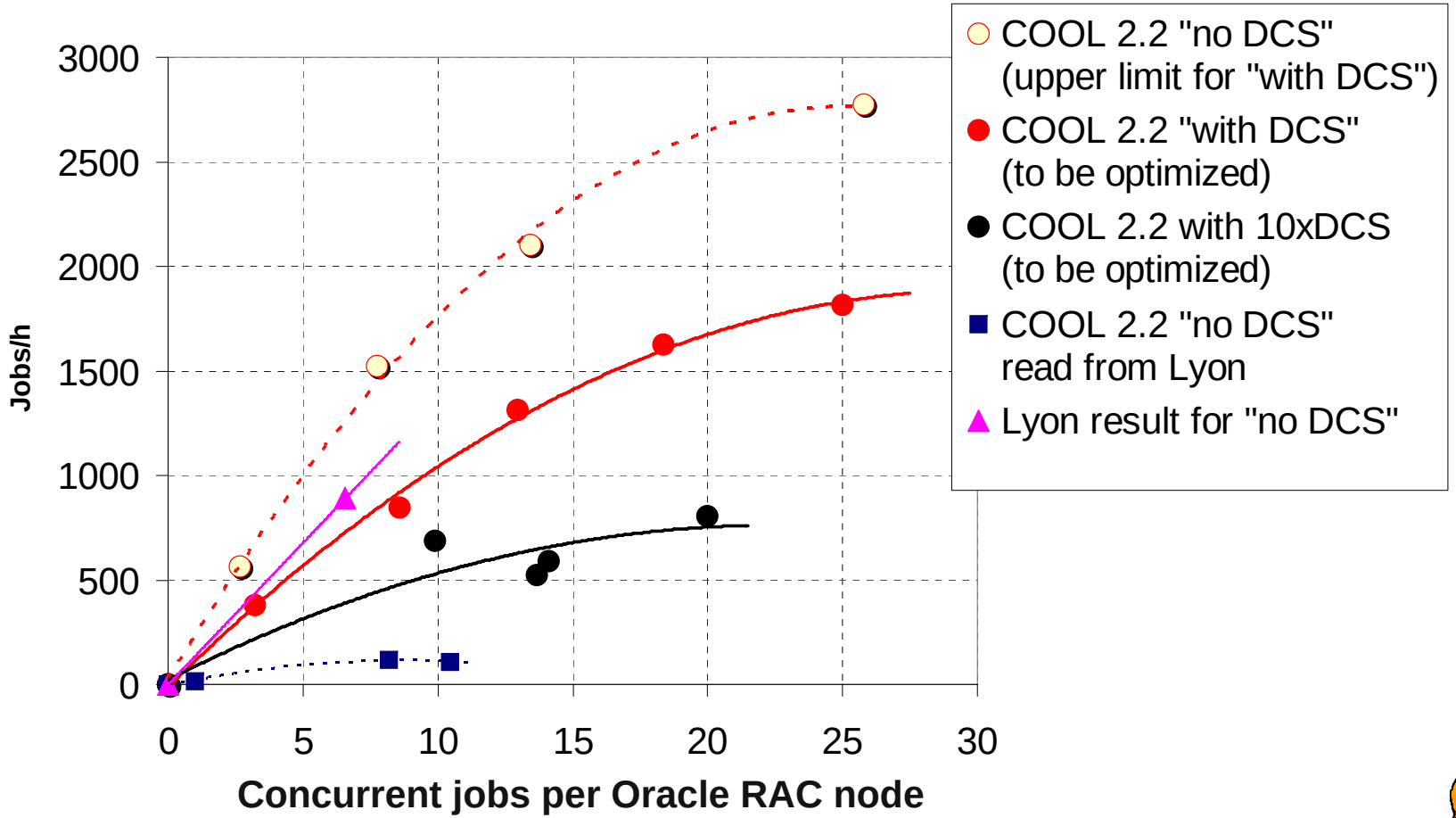
How Scalability Test Works

- First ATLAS scalability tests started at the French Tier-1 site at Lyon. Lyon has a 3-node 64-bit Solaris RAC cluster which is shared with another LHC experiment - LHCb.



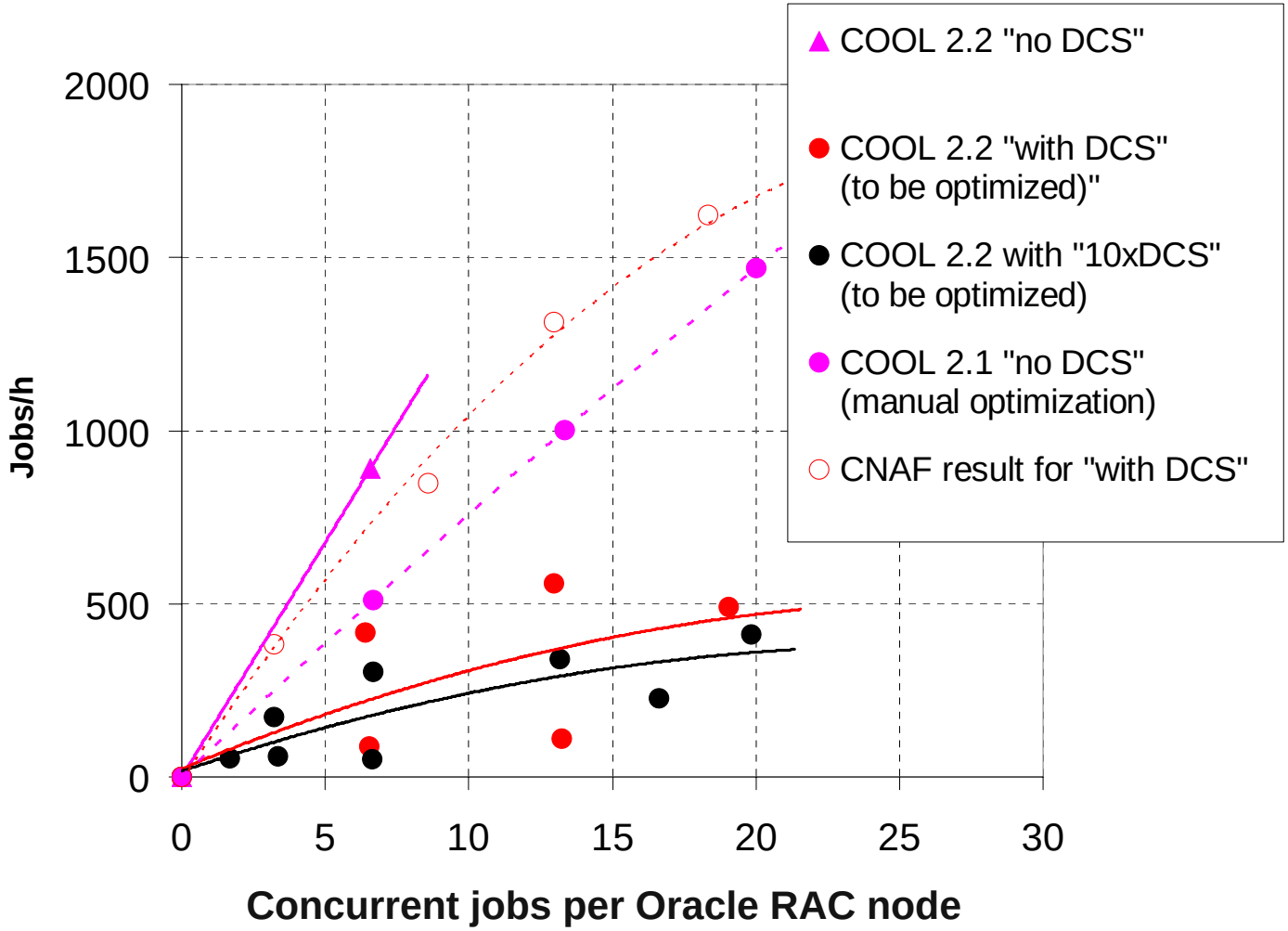
- In scalability tests our goal is to overload the database cluster by launching many jobs at parallel
- Initially, the more concurrent jobs is running (horizontal axis) – the more processing throughput we will get (vertical axis), until the server became overloaded, when it takes more time to retrieve the data, which limits the throughput
- In that particular plot shown the overload was caused by lack of optimization in the COOL 2.1 version that was used in the very first test
  - But it was nice to see that our approach worked

Scalability Tests at Bologna CNAF Tier-1



- CNAF has a 2-node dual-CPU Linux Oracle RAC (dedicated to ATLAS)
- Further COOL 2.2 optimization is expected to provide some increase in performance, since queries for multi-version folders are not optimized yet

Scalability Tests at Lyon CC IN2P3 Tier-1



It is too early to draw many conclusions from the comparison of these two sites, except that it shows the importance of doing tests at several sites

In the Ballpark

- Current ATLAS production operations finishes up to 55,000 jobs/day
- We estimate that during LHC operations ATLAS daily reconstruction and analysis jobs rates will be in the range from 100,000 to 1,000,000 jobs/day
  - For each of ten Tier-1 centers that corresponds to 400 to 4,000 jobs/hour
- For many Tier-1s pledging ~5% capacities (vs. 1/10th of the capacities) that would correspond to the rates of 200 to 2,000 jobs/hour
  - with most of these will be analysis or simulation jobs which do not need so much Oracle Conditions DB access
- Thus, our results from the initial scalability tests are promising
  - We got initial confirmation that ATLAS capacities request to WLCG (3-node clusters at all Tier-1s) is close to what will be needed for reprocessing in the first year of ATLAS operations

- More details on ATLAS scalability tests were presented at the WLCG Workshop held before this Conference:

  http://indico.cern.ch/contributionDisplay.py?contribId=6&confId=3578

Next Steps

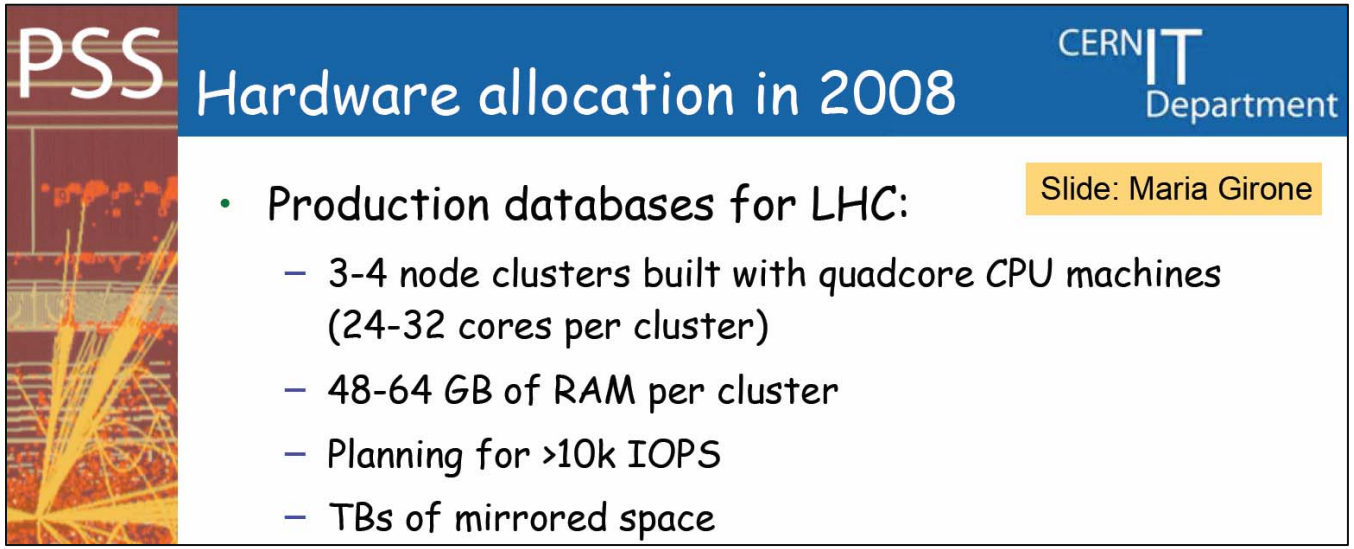
- We plan to continue scalability tests with new COOL releases
  - These tests will result in more precise determination of the actual ATLAS requirements for database capacities in operations
- Having this information we will complete the next iteration on the ATLAS requirements for Oracle capacities at Tier-1 that will be aligned with other ATLAS computing resources at Tier-1, such as raw data storage fraction
  - On average, each Tier-1 will store 1/10th of the raw data
    - *However, actual ATLAS computing resources at Tier-1 currently vary from 4% to 23%*
  - These variations have to be matched in Oracle CPU count
- Also, grouping files on tapes by the datasets (vs. in an arbitrary order) may reduce the Tier-1 requirements for Oracle database capacities during reprocessing
  - However, we don't think we should build into our system a constraint that forces things to be processed linearly

Ramping Up Database Capacities for ATLAS

- Expected growth of ATLAS Oracle data volumes at Tier-0:

Year	Total (TB)	Online	Offline			
		PVSS (TB)	PVSS (TB)	TAG (TB)	COOL (TB)	DDM ARDA Monitoring (TB)
2007	**6.3**	2	2	1	0.3	1
2008	**9.1**	3	3	1.6	0.5	1
2009	**14**	3	3	6	1.0	1

- CERN IT is on track towards delivering the request:

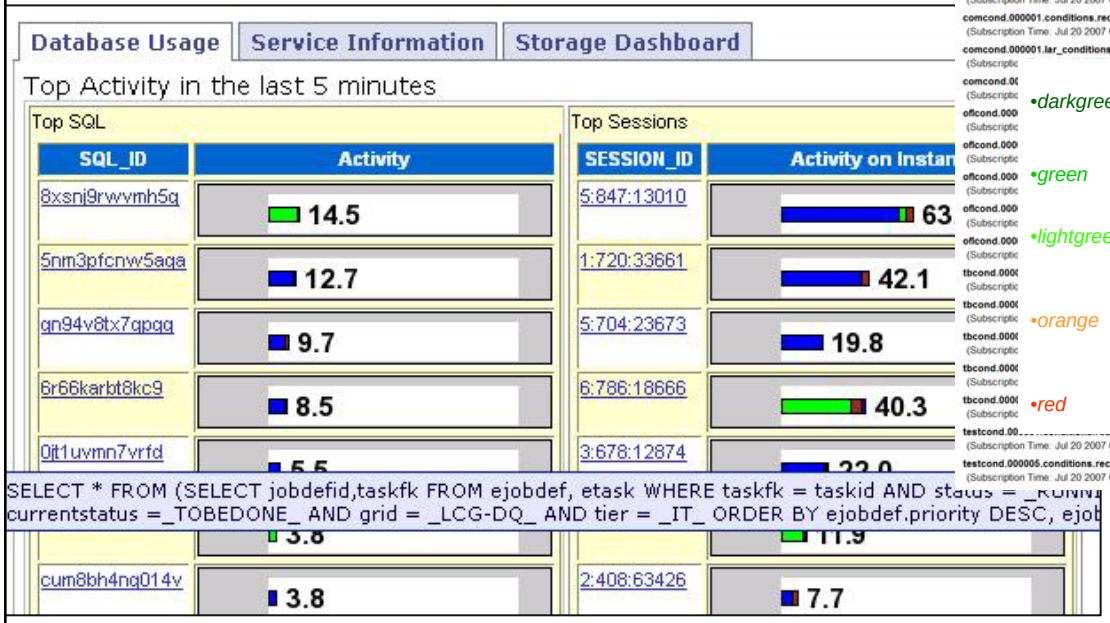


Replication of Conditions DB Payload Data to Tier-1s

- Common LHC software used by ATLAS for data access is technology neutral; currently, two common LHC interfaces are implemented:
  - POOL for access to ROOT files: ATLAS Event Store Database
  - CORAL for access to Relational Databases
    - *Server-based: Oracle, MySQL, and squid/FroNTier*
    - *File-based: SQLite*
- File-based technology is used to achieve scalability for slow-varying data, needed by every data processing job such as the Geometry DB
  - Such data are now packaged in the Database Release (decoupled from the Software Release) and distributed to all sites in an automated way by ATLAS DDM operations
- Also, certain fast-varying Conditions DB payload data are stored in POOL/ROOT files
  - Automatic replication of Conditions DB payload files started by ATLAS DDM operations
- Priority queues for the DDM file transfers has being implemented
  - Since the Conditions DB files should arrive before the event data files

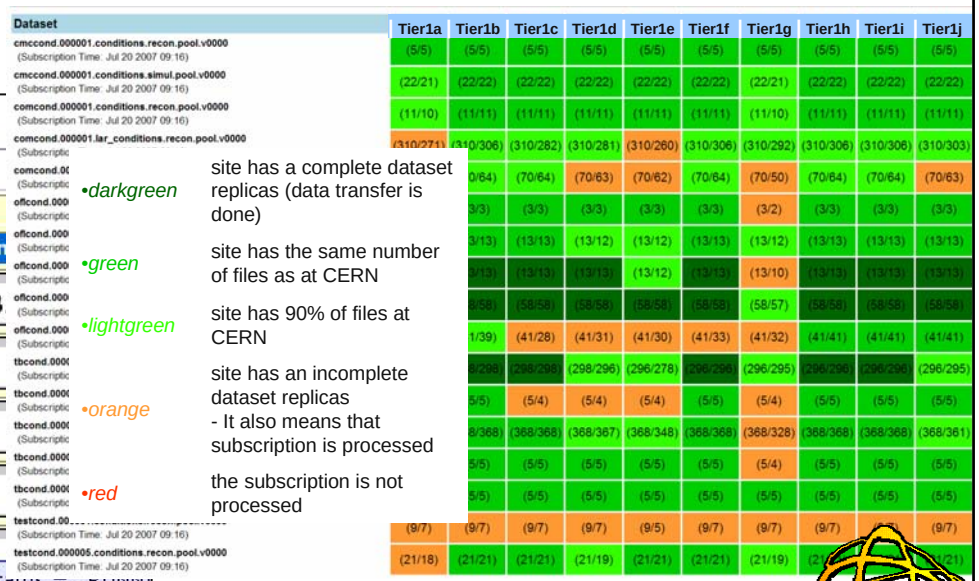
Database Operations Monitoring is in Place

- ATLAS database applications require robust operational infrastructure for data replication between online and offline at Tier-0, and for the distribution of the offline data to Tier-1 and Tier-2 computing centers
- Monitoring is critical to accomplish that:



- ATLAS service monitoring

Credits: Florbela Viegas

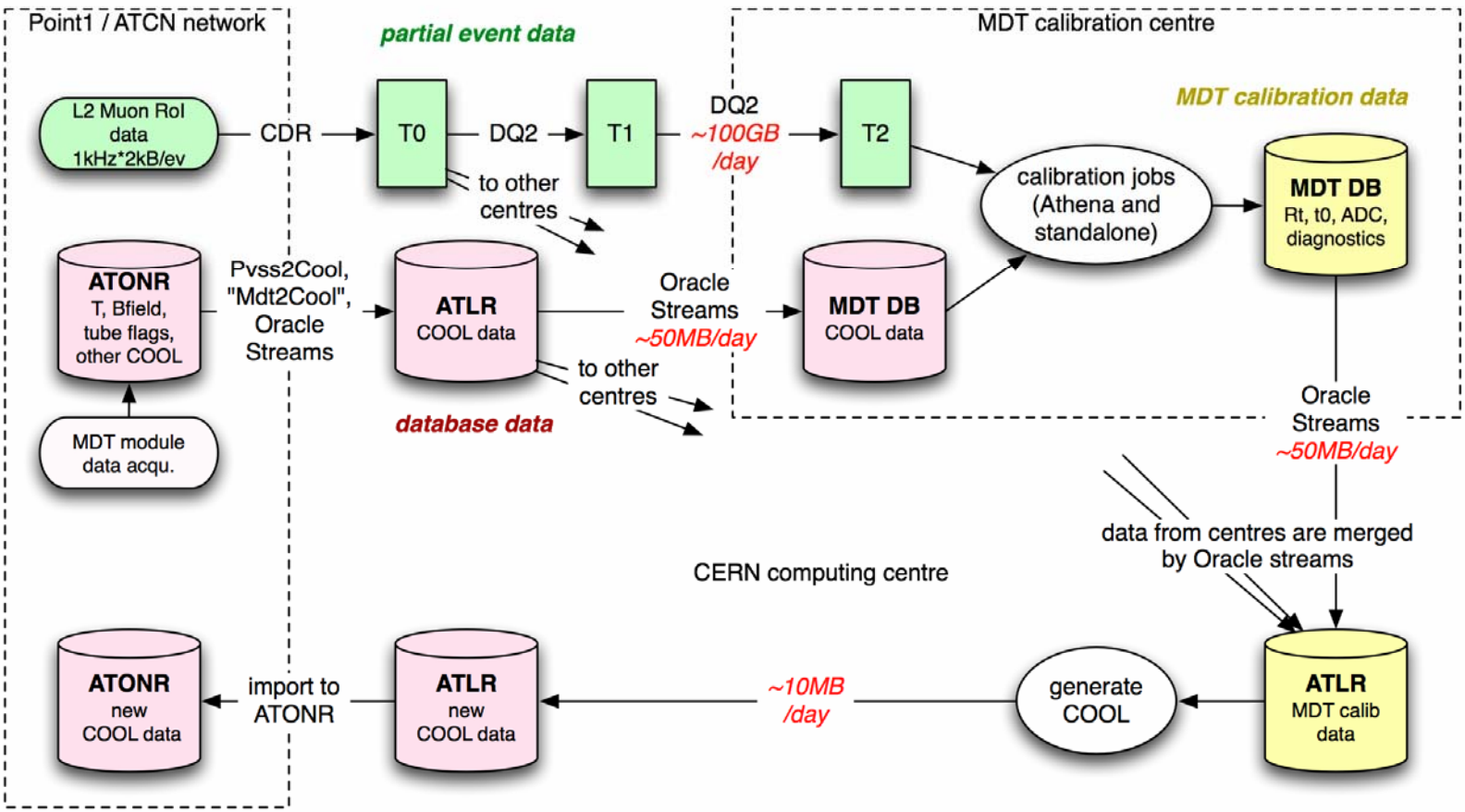


- Conditions DB files replications monitoring

Credits: Alexei Klimentov

Towards MDT Calibration Operations

- ATLAS subdetectors such as Muon System will use the 3D Project database infrastructure for their own calibration databases, e.g. Monitored Drift Tubes (MDT)
- For that Oracle servers are installed at the MDT sites: Michigan, Rome, Munich
  - Replication via Oracle streams tested from Michigan to CERN



Conclusions

- In preparation for ATLAS data taking in ATLAS database activities a coordinated shift from development towards operations has occurred
- In addition to development and commissioning activities in databases, ATLAS is active in the development and deployment of the tools that allow the worldwide distribution and installation of databases and related datasets, as well as the actual operation of this system on ATLAS multi-grid infrastructure
- In collaboration with WLCG 3D Project a major milestone accomplished in database deployment:
  - ATLAS Conditions DB is in production operations with real data on one of the largest distributed database systems world-wide
- A useful framework for Oracle scalability tests has been developed
  - Scalability tests will allow more precise determination of the actual ATLAS requirements for distributed database capacities
- To avoid “unpleasant surprises” we monitor DB volume growth in detector commissioning and test scalability
  - First Oracle scalability tests indicate that WLCG 3D capacities in deployment for ATLAS are in the ballpark of what ATLAS requires

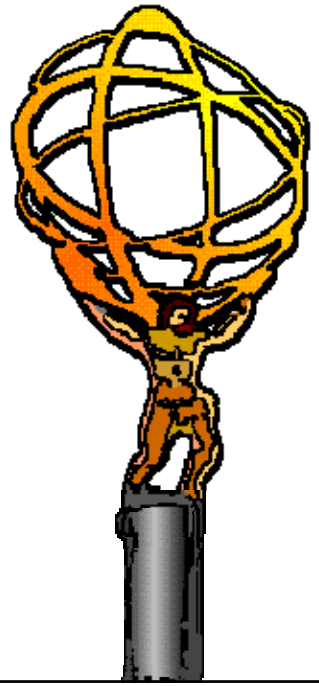
Credits

- Because of large allocation of dedicated resources, scalability tests require careful planning and coordination with Tier-1 sites, which volunteered to participate in these tests
  - Lyon test involved a collaborative effort beyond ATLAS database (R. Hawkings, S. Stonjek, G. Dimitrov and F. Viegas) – many thanks to CC IN2P3 Tier-1 people: G. Rahal, JR Rouet, PE Macchi, and to E. Lancon and RD Schaffer for coordination
  - Bologna test involved a collaborative effort beyond ATLAS database (R. Hawkings, G.Dimitrov and F. Viegas) – many thanks to CNAF Tier-1 people: B. Martelli, A. Italiano, L. dell'Agnello, and to L. Perini and D. Barberis for coordination

Argonne NATIONAL LABORATORY

... for a brighter future

Backup Slides



U.S. Department of Energy

UChicago Argonne LLC

Office of Science
U.S. DEPARTMENT OF ENERGY

A U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

Three Options used for the Test Workloads

The latest COOL 2.2 version enabled more realistic workload testing

- Three different workloads have been used in these tests:

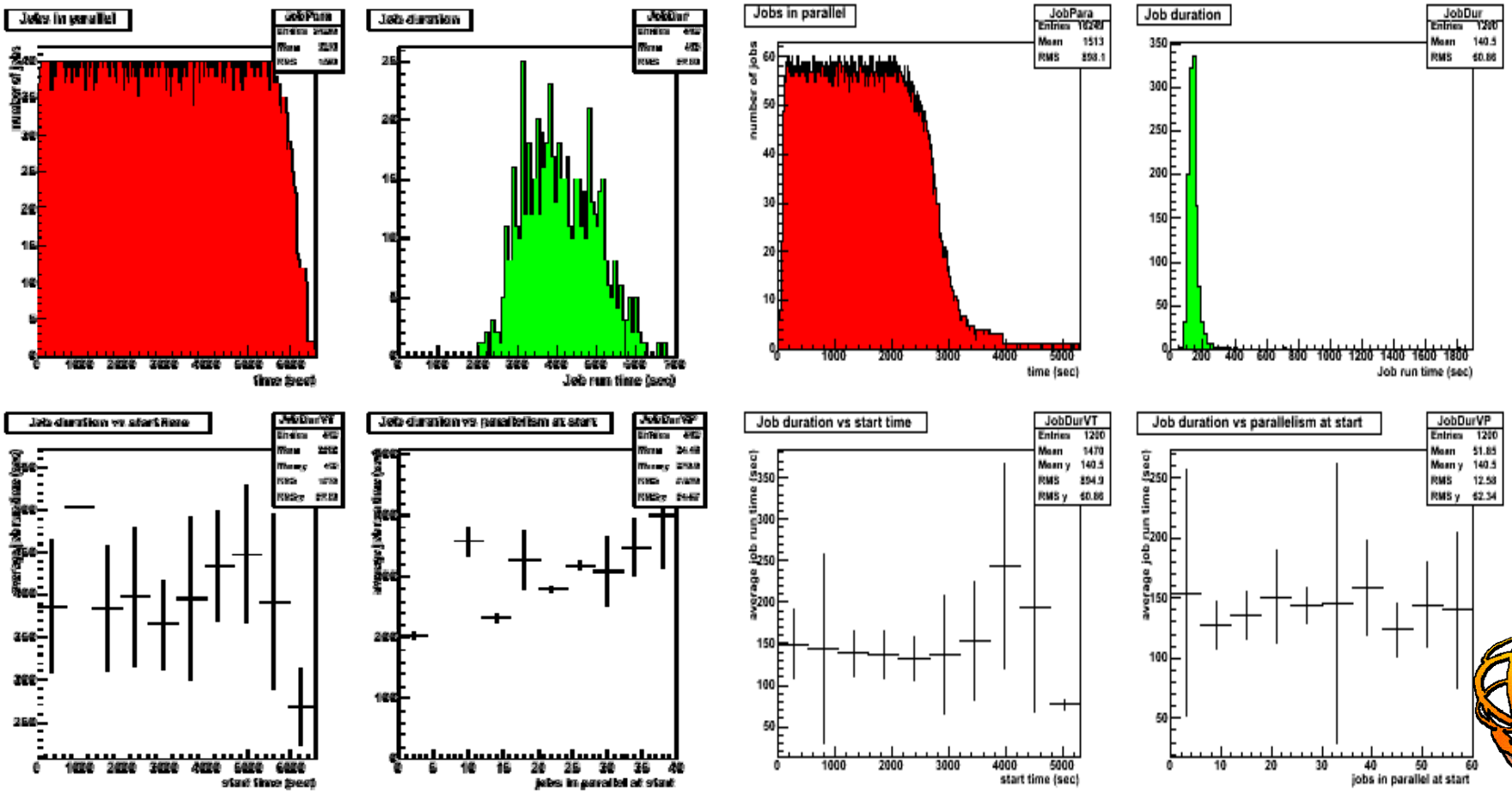
"no DCS" - Data from 19 folders of 32 channels each, POOL reference (string) payload, plus 2 large folders each with 1174 channels, one with 3k string per channel, one with 4.5k string per channel, which gets treated by Oracle as a CLOB, plus 1 folder with 50 channels simulating detector status information. This is meant to represent a reconstruction job running reading calibration but no DCS data. The data is read once per run.

"with DCS" - As above, but an additional 10 folders with 200 channels each containing 25 floats, and 5 folders with 1000 channels of 25 floats, representing some DCS data, again read once per run

"10xDCS" - As above, but processing 10 events spaced in time so that all the DCS data is read again for each event. This represents a situation where the DCS data varies over the course of a run, so each job has to read in 10 separate sets of DCS data.

Scalability tests: detailed monitoring of what is going on

- ATLAS testing framework keeps many things in check and under control:



IN3P3 test

CNAF test

Credits: Richard Hawkings