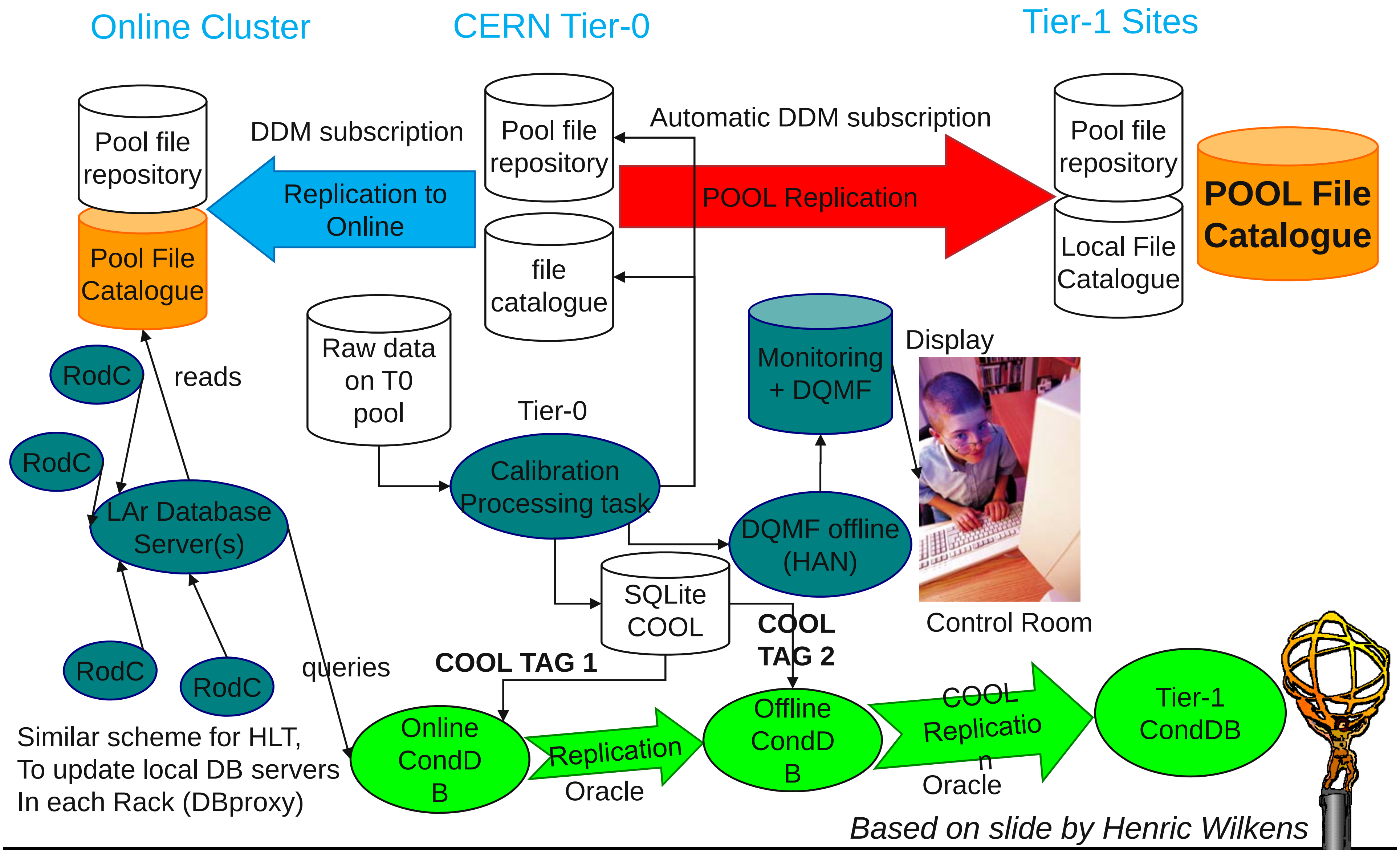
Conditions DB Data Distribution

- For Conditions DB data distribution beyond Tier-0 ATLAS adopted
  - 3D Oracle Streams technology for replication to Tier-1 sites
    - *for replication of COOL Conditions DB IOVs and 'in-lined' payload*
  - ATLAS ADC DDM for file-based database data replication
    - *for replication of POOL files with Conditions DB payload*
- Critical operational infrastructure for these technologies is delivered by:
  - WLCG 3D service
    - *provided by Distributed Deployment of Databases (3D) Project*
  - ATLAS Distributed Computing (ADC) organization
    - *Distributed Data Management (DDM) operations*

The distribution Conditions DB data is not enough:

- To allow data reprocessing at Tier-1 sites we need a model for Conditions DB operations to support access to COOL+POOL conditions/calibrations data by the reconstruction jobs
  - To achieve that in addition to ATLAS teams operating in a coordinated way the Tier-1 sites have to be engaged

Team Work in Conditions & Calibrations Distribution



Online Cluster

CERN Tier-0

Tier-1 Sites

Pool file repository

Pool File Catalogue

DDM subscription

Replication to Online

Pool file repository

file catalogue

Automatic DDM subscription

POOL Replication

Pool file repository

Local File Catalogue

POOL File Catalogue

RodC

reads

RodC

LAr Database Server(s)

RodC

RodC

Raw data on T0 pool

Tier-0

Calibration Processing task

Monitoring + DQMF

Display

DQMF offline (HAN)

SQLite COOL

Control Room

queries

COOL TAG 1

COOL TAG 2

Similar scheme for HLT,
To update local DB servers
In each Rack (DBproxy)

Online CondD B

Replication

Oracle

Offline CondD B

COOL Replicatio n

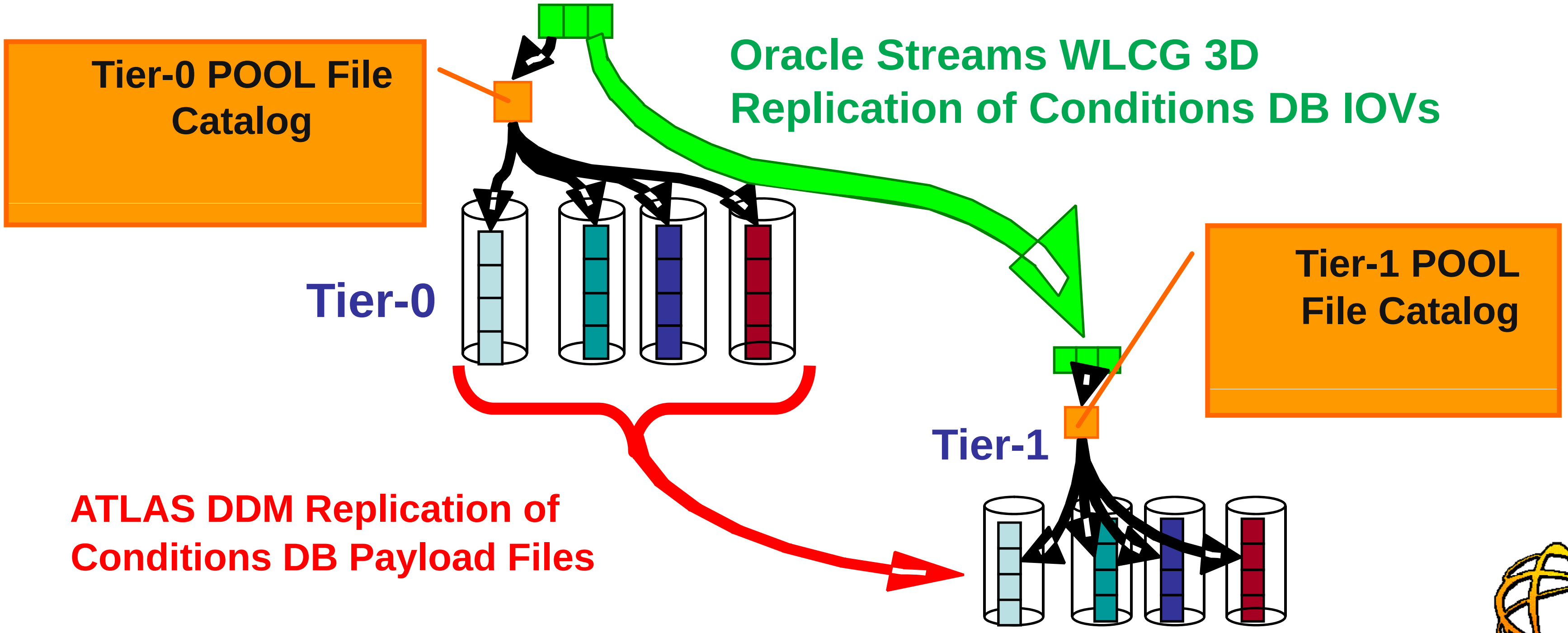
Oracle

Tier-1 CondDB

Based on slide by Henric Wilkens

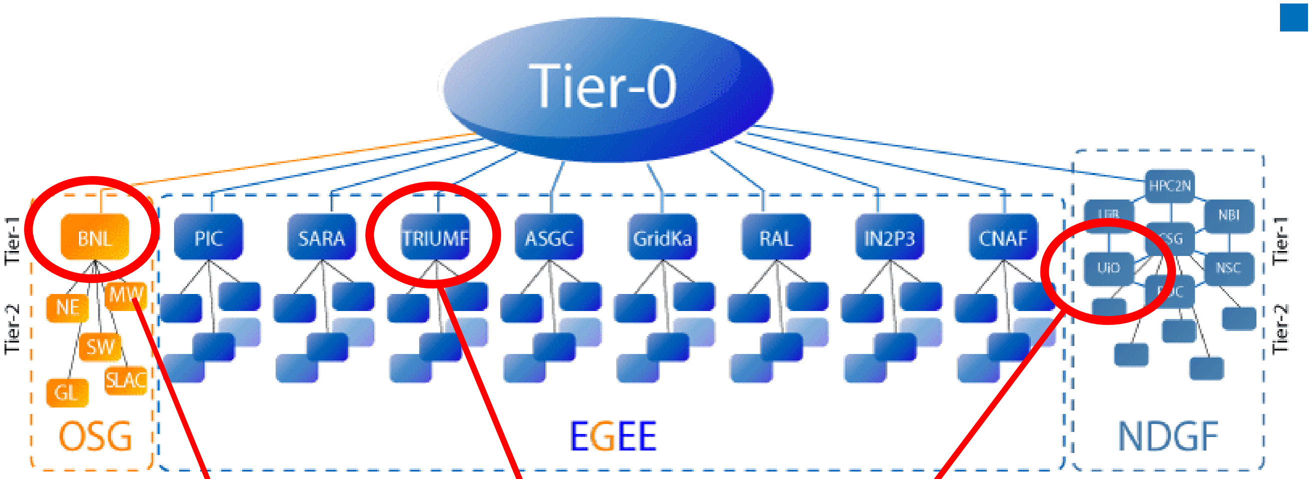
Access to Calibrations Data

- To access calibrations data in POOL files at the Tier-1 site the Athena job must navigate from the POOL token
  - related in the COOL DB schema to specific IOV and Tag values

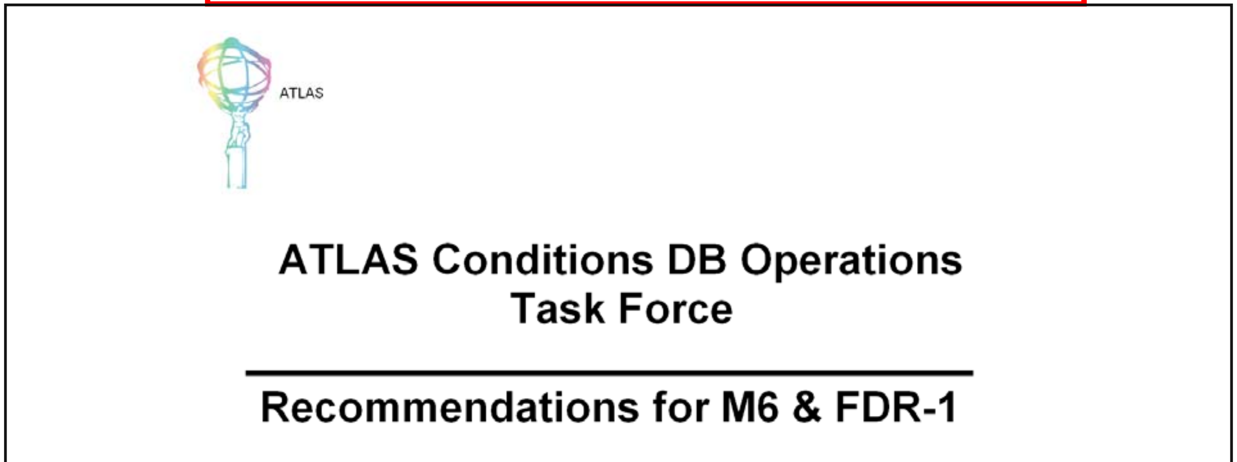

Tier-0 POOL File Catalog
Oracle Streams WLCG 3D Replication of Conditions DB IOVs
Tier-0
Tier-1 POOL File Catalog
Tier-1
ATLAS DDM Replication of Conditions DB Payload Files

- The GUID->PFN lookup needed for that is provided by the POOL File Catalog providing PFN, which is directly accessible from the worker node
  - At Tier-0 the POOL File Catalog is maintained by a cron job

Recommendations for Reprocessing



- M5 reprocessing was exercised at sites from all three ATLAS grids



ATLAS

ATLAS Conditions DB Operations Task Force

Recommendations for M6 & FDR-1

http://doc.cern.ch//archive/electronic/cern/others/atlnot/Communication/soft/com-soft-2008-003.pdf

- To develop a model for distributed Conditions DB operations a dedicated Task Force was created
  - a joint effort of Database and Distributed Computing Operations:
- Based upon the analysis of M5 reprocessing results, the Task Force provided recommendations for the M6, and FDR-1 reprocessing

Roadmap to Reprocessing of the LHC Data

LHC data reprocessing is expected to start during this fall

- The M5 reprocessing exercise was the first ATLAS experience with massive reconstruction of real data at the Tier-1 sites that exercised ATLAS distributed database infrastructure
  - It was extremely useful, with many lessons learnt from that
- Based on this experience, the Conditions DB Operations Task Force developed recommendations for FDR-1 and M6 data reprocessing
- Using these recommendations ATLAS Tier-1 sites are now preparing for the bulk reprocessing of the Mx and FDR-1 data
  - The CCRC08-2 provides an opportunity to test Tier-1 database infrastructure under conditions that are similar to real LHC data
- On a longer timescale the Conditions DB Operations Task Force is looking forward to develop recommendations for LHC data-taking based upon experience gathered during FDR-1 and M6 data reprocessing (and can be validated in FDR-2 reprocessing)
  - Tier-1 DBAs efforts during CCRC08-2 reprocessing are critical to that

Proposed Targets for 3D Milestones in CCRC08-2

- Last LCG Management Board requested to collect any remaining ATLAS milestones for the 3D setup
- Reprocessing during CCRC08-2 in May provides an opportunity to validate that the 3D deployment is ready for LHC data taking

Tier-1 site	**Sessions/Node**
TRIUMF	**40**
FZK	**80**
IN2P3	**60**
CNAF	**40**
SARA	**140**
NDGF	**100**
ASGC	**80**
RAL	**90**
BNL	**220**
PIC	**40**

- Robust Oracle RAC operation at the Session/Node target demonstrates that 3D setup at the site is ready for ATLAS data taking
  - Achieving 50% of the target shows that site is 50% ready
- In addition, during ATLAS reprocessing week DBAs at the Tier-1 sites should collect Oracle performance metrics: CPU, I/O and cache hit ratios
  - Details will be presented to DBAs at the Database Track

Oracle Performance Metrics

We should 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.

Slide by G. Dimitrov, F. Viegas

Argonne
NATIONAL
LABORATORY

... for a brighter future

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

Backup Slides

Things to Watch

- Using current estimates, Oracle max sessions limit does not seems to became a bottleneck for the Oracle RACs currently deployed at Tier-1
  - this conclusion has to be verified in M6 and FDR-1 reprocessing

Also needs verification:

- Tier-2 CPU capacities that have to access Tier-1 Oracle site
- Variations in job initialization time at each Tier-1 depending on
  - Oracle RAC CPU speed
  - Time it takes to read Conditions DB files from the Tier-1 hardware
- CPU count estimates at each Tier-1 site

as well as the main conclusion:

- If reconstruction jobs spends 1.2% of it’s time during Conditions DB initialization, the number of jobs accessing Oracle RAC concurrently will be on average below 30
  - which is within our previous scalability limit studies
- Based on realistic workloads used in Conditions DB scalability tests
  - See next slide

Realistic Workload for Conditions DB Tests

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

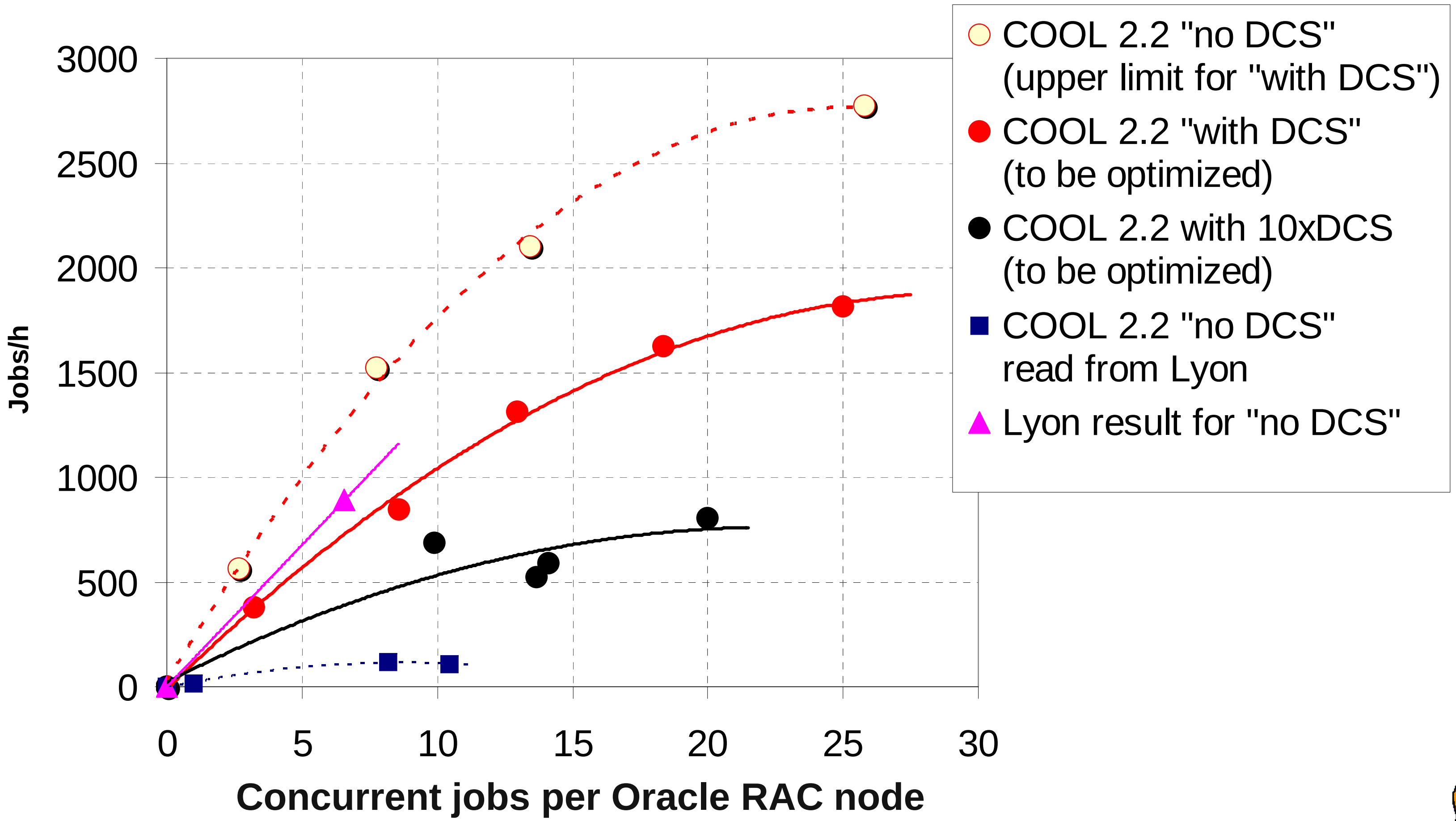
‘best guess’

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**

- 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**”

Based on slide by R. Hawkings

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 were not optimized

Requirements for Oracle Capacities at Each Tier-1

- Current ATLAS production operations finishes up to 60,000 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
- From ATLAS COOL scalability tests we got initial confirmation that Tier-1 capacities request to WLCG (for 3-node clusters) is close to what will be needed for reprocessing in the first year of ATLAS operations
  - During scalability test a single Linux RAC node served 1,500 jobs/hour
- The next iteration on the ATLAS requirements for Oracle capacities at Tier-1 should be aligned with ATLAS computing resources at Tier-1, such as raw CPU count
    - *at ATLAS Tier-1s the computing resources pledges for 2008 vary from 5% to 25%*
  - These variations have to be matched in Oracle capacities
- The M5 reprocessing tests at Tier-1 sites identified another potential bottleneck
  - Oracle sessions count

Assumptions for Reconstruction Job

- 15 COOL schemas has to be accessed (out of 17+1)
- Conditions DB data initialization times measured by Richard for the test job at CNAF were 100-200 s (depending on workload) with 40-50 concurrent jobs
  - Use a conservative number of 300 s to account for time needed to access Conditions DB files
- Reconstruction time of one event is 10 s
  - Taken from FDR-1 at Tier-0
- Number of events in one reconstruction job is 2,500
  - Corresponds to five SFO files merged into one
  - For five streams the average is of 40 Hz/stream, then 2,500 events on average correspond to a time slice of about 1 min
    - *We assume that all Conditions DB data read at initialization will be valid for that period of less than 1 min*
- These assumptions result that a seven-hour reconstruction job will spend 1.2% of the time in Conditions DB data initialization phase

Calculations of the CCRC08-2 3D Targets

- Core count is approximated by 2008 CPU pledge in kSI2K divided by 2
  - a conservative estimate to account for old machines at Tier-1s
    - *Roger's updated number for Tier-0 is 2.25*
- Number of reconstruction jobs accessing Oracle RAC concurrently is calculated assuming time spent in Conditions DB initialization of 1.2%
  - 5 min out of 7 hours

	ATLAS kSI2K	Estimated CPU cores	Concurrent Jobs	Sessions /RAC	RAC nodes	Sessions/Node	**Target**
TRIUMF	905	453	5	81	2	41	**40**
FZK	1812	906	11	163	2	82	**80**
IN2P3	2066	1033	12	186	3	62	**60**
CNAF	960	480	6	86	2	43	**40**
SARA	3048	1524	18	274	2	137	**140**
NDGF	1070	535	6	96	1	96	**100**
ASGC	1700	850	10	153	2	77	**80**
RAL	1925	963	12	173	2	87	**90**
BNL	4844	2422	29	436	2	218	**220**
PIC	865	433	5	78	2	39	**40**

A Reminder: ATLAS Conditions DB Technology

- For Conditions DB technology ATLAS adopted Common LHC solution:
  - COOL (Conditions database Of Objects for LHC)
- In COOL database architecture the conditions data are usually characterized by the Interval-of-Validity (IOV) metadata and a data payload, with an optional version tag
  - Separation of the IOV and payload allows storage of the select payload data in POOL files outside of the database server
    - *An example of such data is calorimeter calibration constants that are not really suited to relational databases because of their size and access patterns*
  - Proper separation reduces requirements for Oracle capacities

Access to Conditions DB data is critical for event data reconstruction

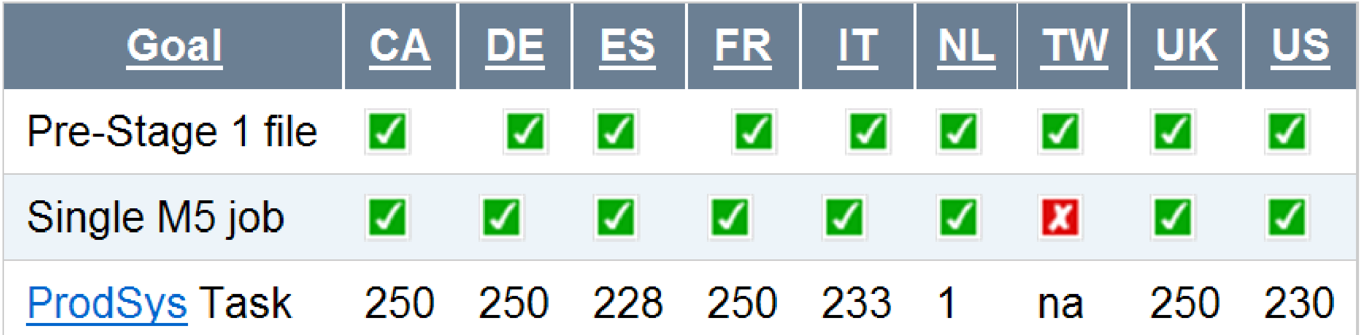
- To achieve scalability in Tier-0 operations slices of the corresponding conditions/calibrations data will be delivered to Tier-0 farm via files on afs
  - Beyond Tier-0 we need a different technology for data distribution and a new model of Conditions DB operations that must take into account grid computing realities and ATLAS Computing Model that assigned massive reprocessing of the LHC data to the Tier-1 sites

Choice of POOL File Catalog Technology for M6 & FDR-1

- One of the principal open question addressed by the Task Force was:
  - What is the model for reconstruction jobs to lookup physical file replicas at sites?
    - *are they expected to be fed local XML-file catalogues, or can they make lookups ‘on demand’ in LFC file catalogues?*
- Based upon findings from the initial tests of these technologies the Task Force recommended to adopt a site-specific model at Tier-1s
  - The recommended model has been tested successfully in a large-scale M5 reconstruction at both Tier-0 and Tier-1
- Detailed description of these site-specific PFC settings is documented in

http://twiki.cern.ch/twiki/bin/view/Atlas/CDBOTaskForce?topic=T1Reprocessing

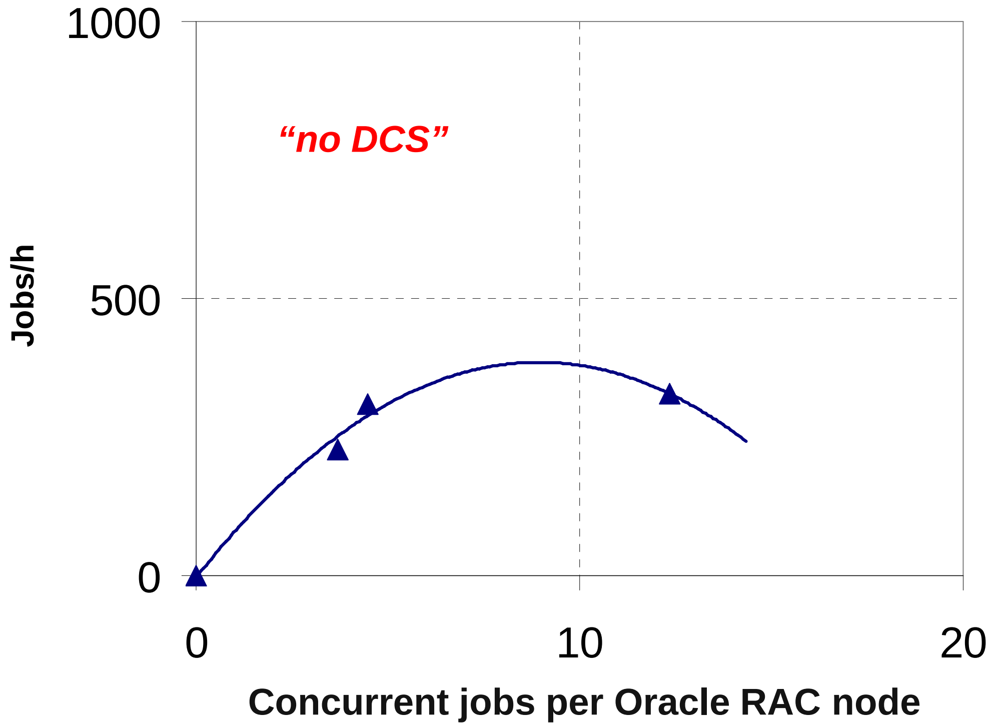
- After the Task Force recommendations were discussed at the previous ATLAS Tier-0/1/2/3 Jamboree all ten ATLAS Tier-1 sites implemented the recommendations (by April 2), with nine sites finished M5 reprocessing:



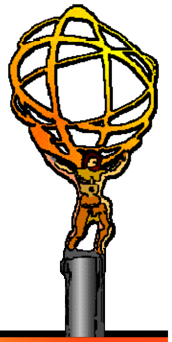
Goal	CA	DE	ES	FR	IT	NL	TW	UK	US
Pre-Stage 1 file	✓	✓	✓	✓	✓	✓	✓	✓	✓
Single M5 job	✓	✓	✓	✓	✓	✓	✗	✓	✓
ProdSys Task	250	250	228	250	233	1	na	250	230

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



Databases are Critical to Keeping Track of Jobs and Files

- To achieve robust operations on the grids ATLAS splits data processing tasks of petabytes of event data into smaller units - jobs and files
  - An overstretched analogy is splitting of Internet communications into small TCP/IP packets (to make re-sending of failed packets efficient)
    - *In a similar way, the failed grid jobs and files transfers are re-tried*

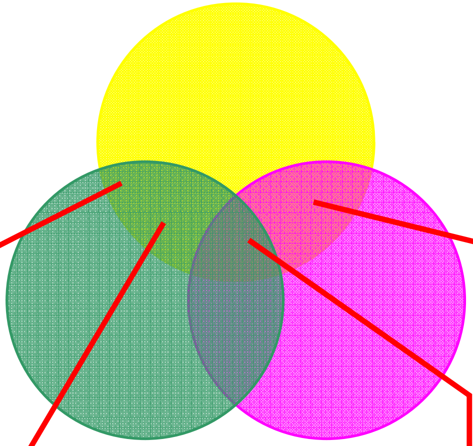
- Job – unit for data processing workflow management in Grid Computing
  - Managed by ATLAS Production System:

- ATLAS job configuration and job completion are stored in prodDB

- ATLAS jobs are grouped in Tasks:

- ATLAS Tasks DB

Databases



Jobs

- Metadata Information DB

- File - unit for data management in Grid Computing
  - Managed by ATLAS Distributed Data Management (DDM) System

Files

- ATLAS files are grouped in Datasets:

- Central File Catalogs

Eliminating Critical Services Downtime

- Among all ATLAS Computing Services the Database Services got the highest rating in terms of criticality:
  - http://twiki.cern.ch/twiki/bin/view/Atlas/CriticalDatabases
- When Critical Database Services are down many operations can be stopped
- To avoid that we are now investigating the possibility to set up a fall-back off-site server for the CERN Oracle system
- Oracle technology we are evaluating for that called transportable tablespaces:
  - Transportable tablespaces enables you to unplug a set of tablespaces from a database, move or copy them to another location, and then plug them into another database
  - Transportable tablespace from backup with RMAN enables you to move or copy a tablespace set while the tablespaces remain online (writeable)
- Select Centralized Databases will become distributed:
  - Replicated to SARA Tier-1 center in Amsterdam, The Netherlands