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

CERN

ATLAS Scalability Tests of Tier-1 Database Replicas

WLCG Collaboration Workshop (Tier0/Tier1/Tier2)

Victoria, British Columbia, Canada

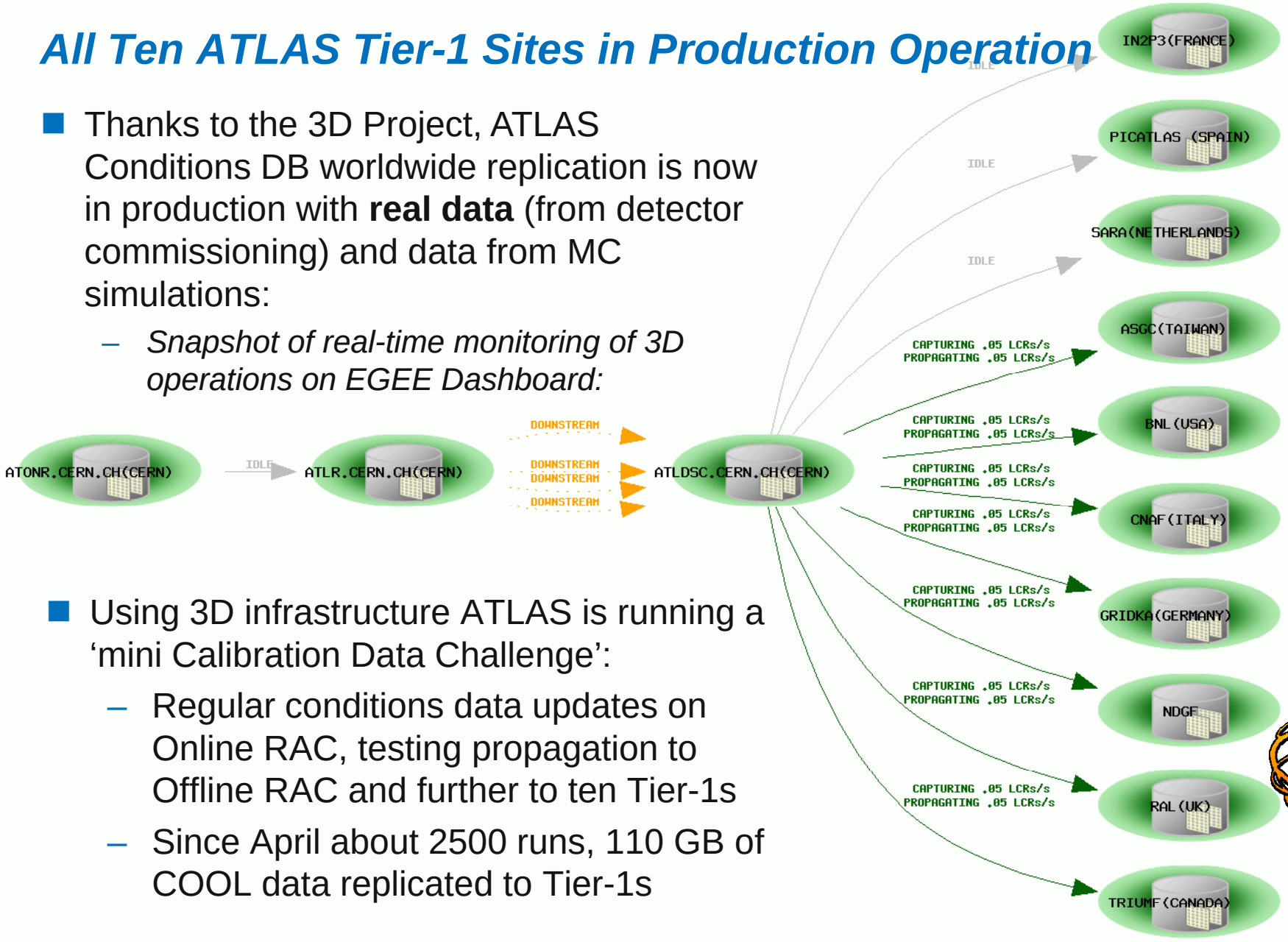
September 1-2, 2007

Richard Hawkings (CERN)

Alexandre Vaniachine (Argonne)

All Ten ATLAS Tier-1 Sites in Production Operation

- Thanks to the 3D Project, 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 about 2500 runs, 110 GB of COOL data replicated to Tier-1s

DB Replicas at Tier-1s: Critical for Reprocessing

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

Schema	#folders	#chan	Chan 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**

Why we are replicating all these data?

- 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

Purpose of Scalability Tests

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
  - Thus, we do not have to initialize O(100) jobs at once
- Tier-0 uses file-based Conditions DB slice on afs, 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**
- Find out where are the bottlenecks:
  - Hardware (database CPU, storage system (I/O per second))
  - COOL read-back queries (to learn if we have some problems)
- Find out how many clients a Tier-1 database server can support at once
  - Clients using Athena and realistic conditions data workload

ATLAS Conditions DB Tests of 3D Oracle Streams

- Replicated data provide read-back data for scalability tests :

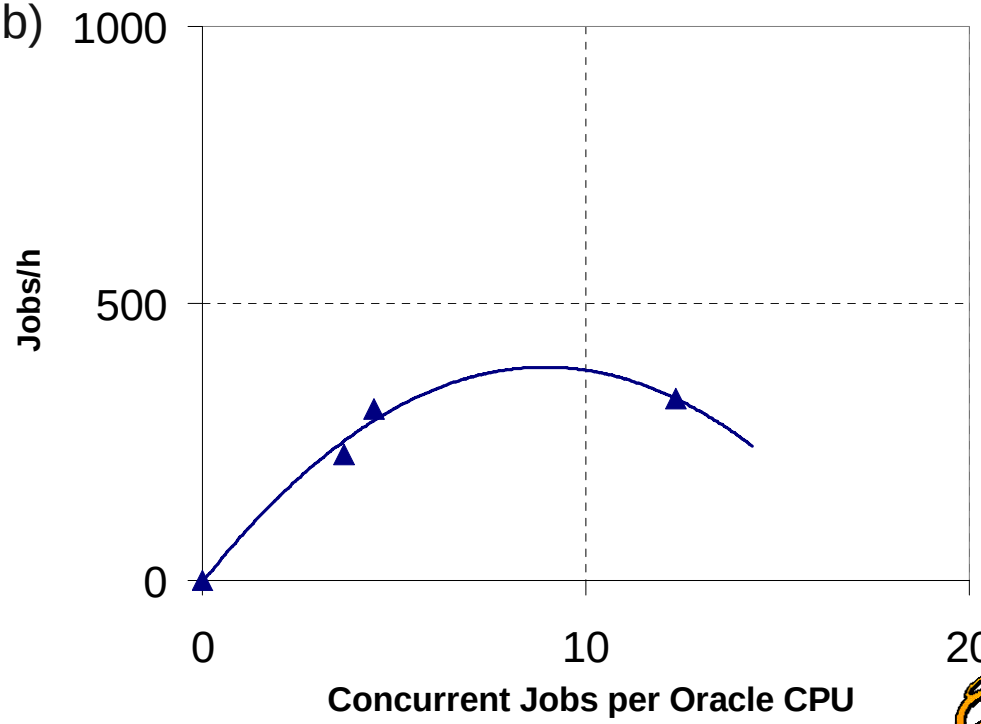
Schema	#folders	#chan	Chan payload	N/run	Total GB
INDET	2	32	160 char	1	**0.21**
CALO	17	32	160 char	1	**1.8**
MDT	1+1	1174	CLOB: 3kB+4.5kB	0.1	**17.0**
GLOBAL	1	50	3 x float	6	**0.25**
TDAQ/DCS	10+5	200+1000	25 x float	12	**80.0**
TRIGGER	1	1000	25 x float	12	**8.0**

'best guess'

- The realistic conditions data workload:
  - a 'best guess' for ATLAS Conditions DB load in reconstruction
    - *dominated by DCS data*
- Three workload combinations were used in the tests:
  - "**no DCS**", "**with DCS**" and "**10xDCS**" (explained on slide 11)

How Scalability Test Works

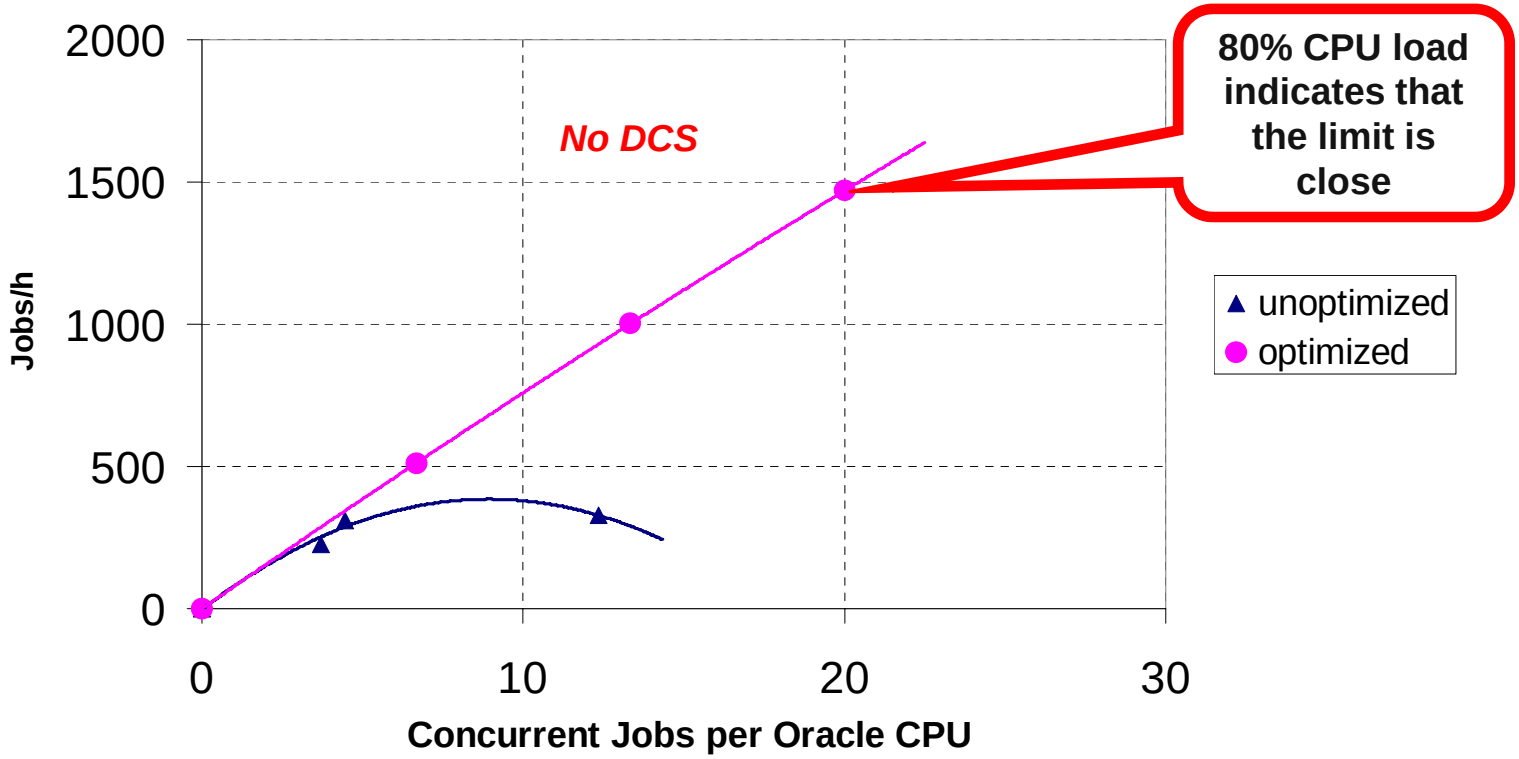
- First ATLAS scalability tests started at the French Tier-1 site at Lyon
- Lyon has a 3-node 64-bit Solaris RAC cluster which is shared with another LHC experiment (LHCb)
- In scalability tests our goal is to overload the database cluster by launching many jobs at parallel
- Initially, the more concurrent 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

First Scalability Tests at Lyon CC IN2P3 Tier-1

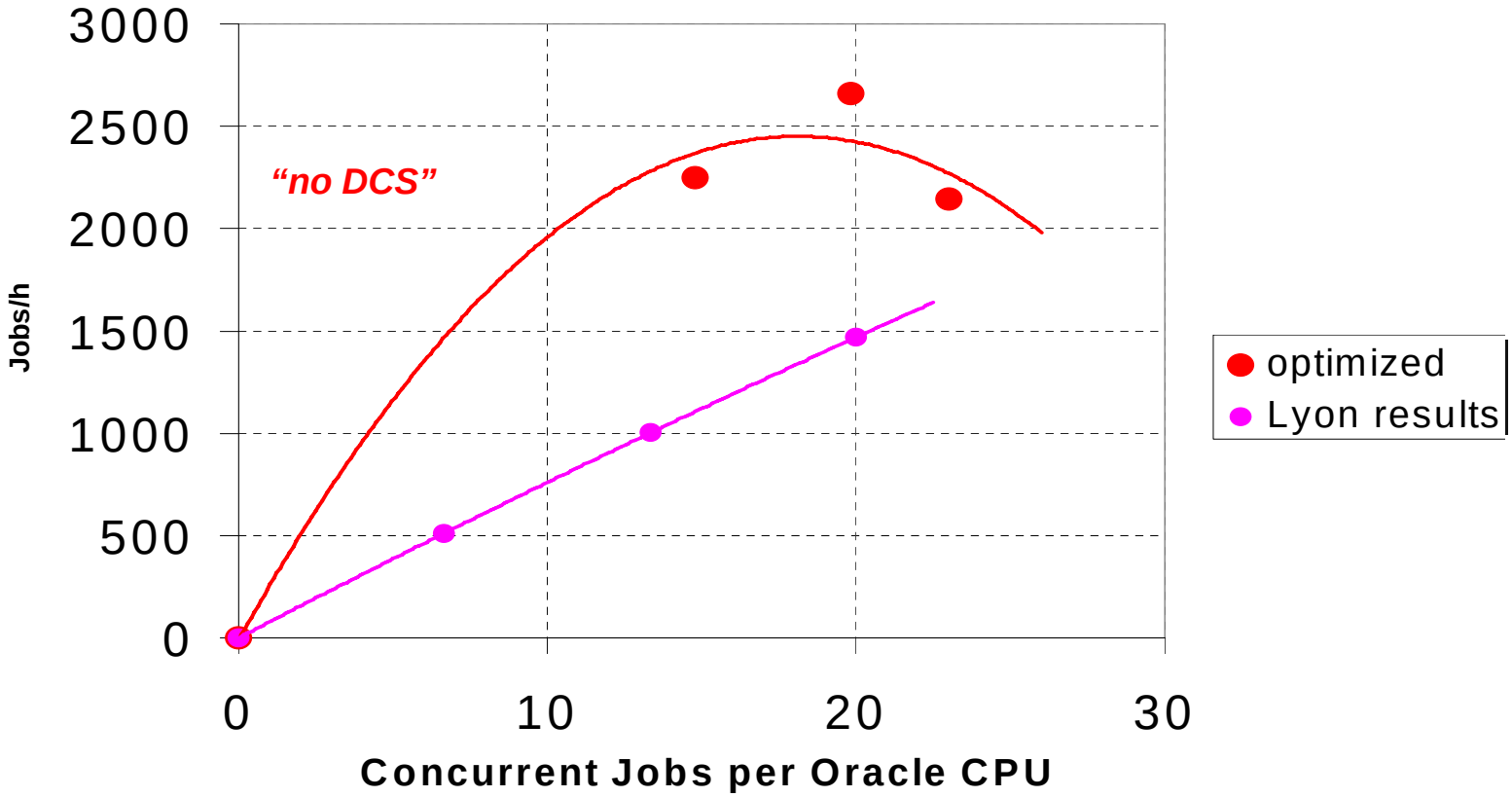
- Even with that old COOL 2.1 version we got useful “no DCS” results by engaging “manual” optimization via actions on the server side



- Test jobs read “no DCS” Conditions DB data workload at random
- The calculated throughput (jobs/hour) is for the RAC tested (3-node SPARC)

First Scalability Tests at Bologna CNAF Tier-1

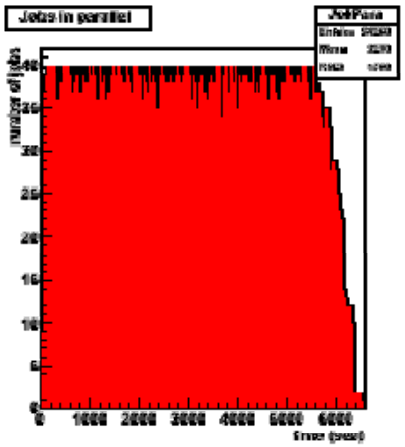
- In a similar way results were obtained at CNAF with the old COOL 2.1

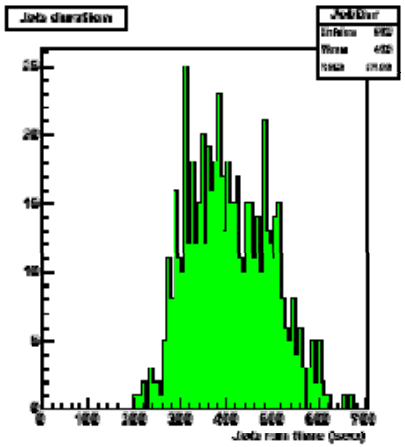


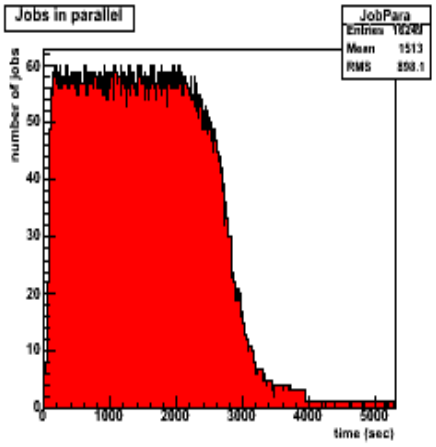
- Test jobs read "no DCS" Conditions DB data workload at random
- The calculated throughput (jobs/hour) is for the RAC tested:
  - 2-node dual-core Linux Oracle RAC (dedicated to ATLAS)

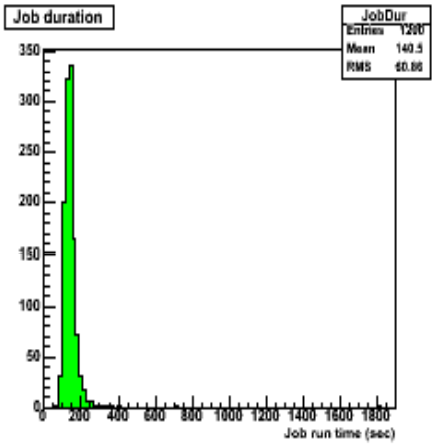
Scalability tests: detailed monitoring of what is going on

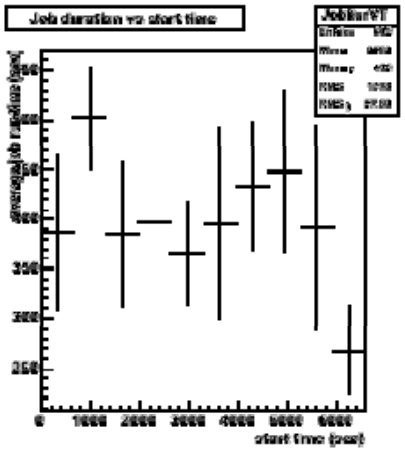
- ATLAS testing framework keeps many things in check and under control:

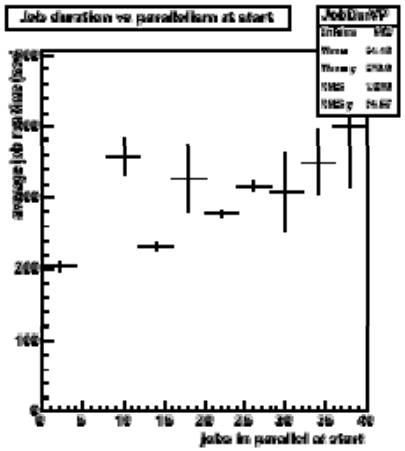




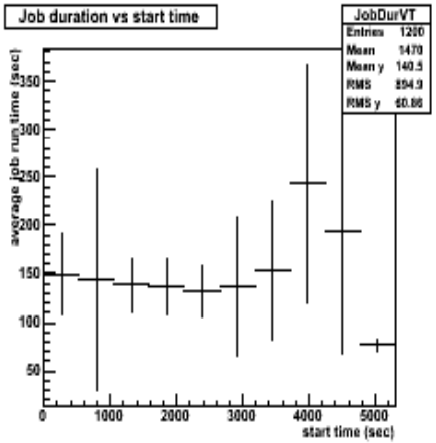


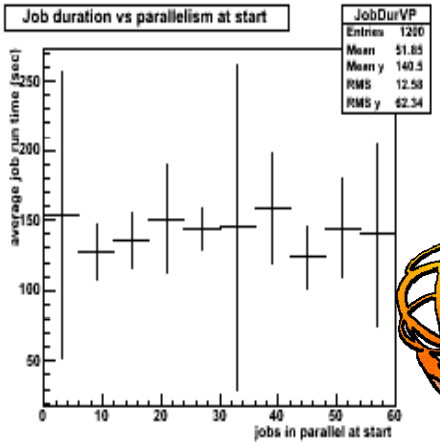






IN3P3 test





CNAF test

Latest Tests with COOL 2.2

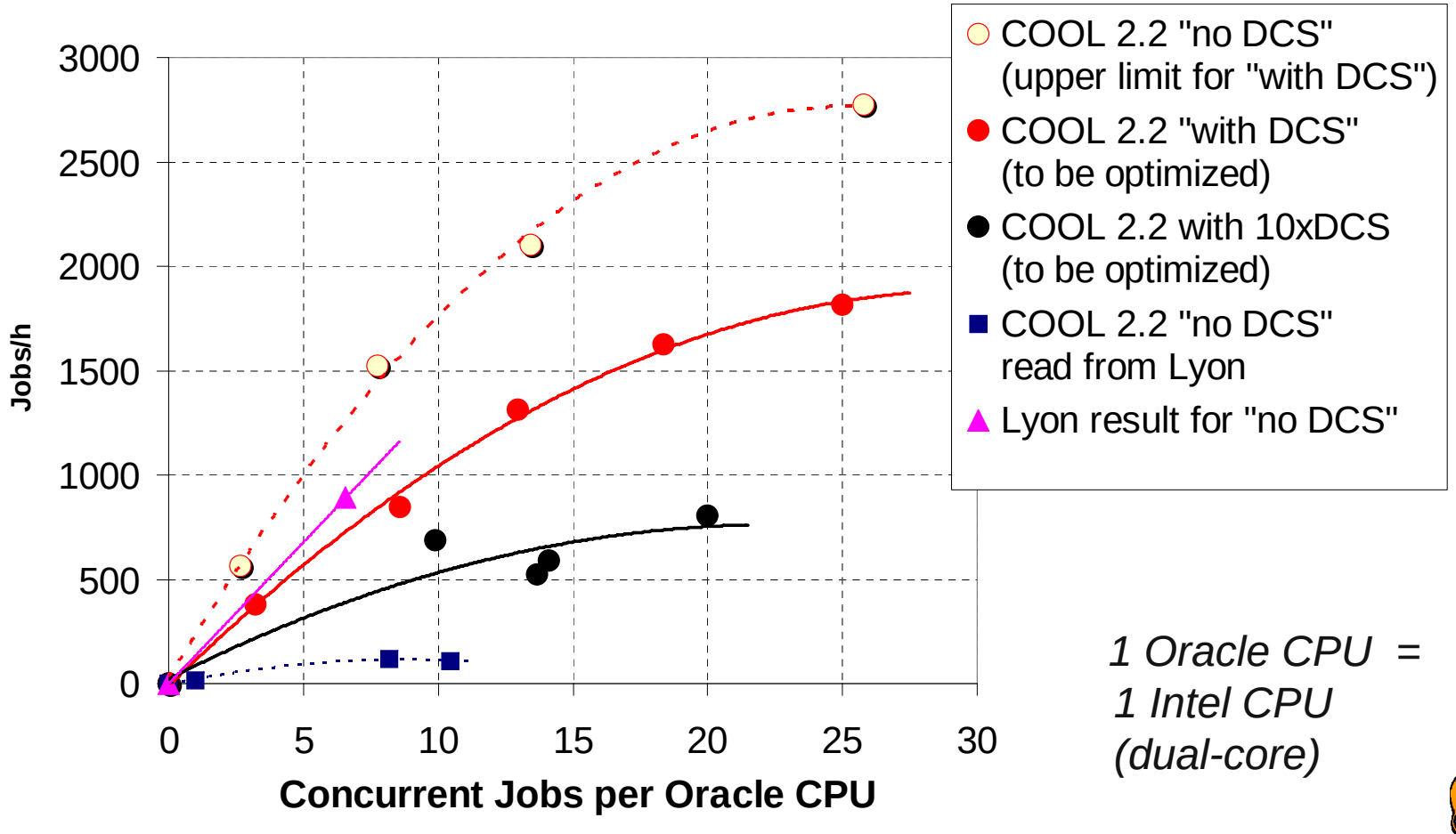
- New COOL 2.2 version enabled more realistic workload testing
  - Now “with DCS”
- Three different workloads have been used in these tests:

“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.

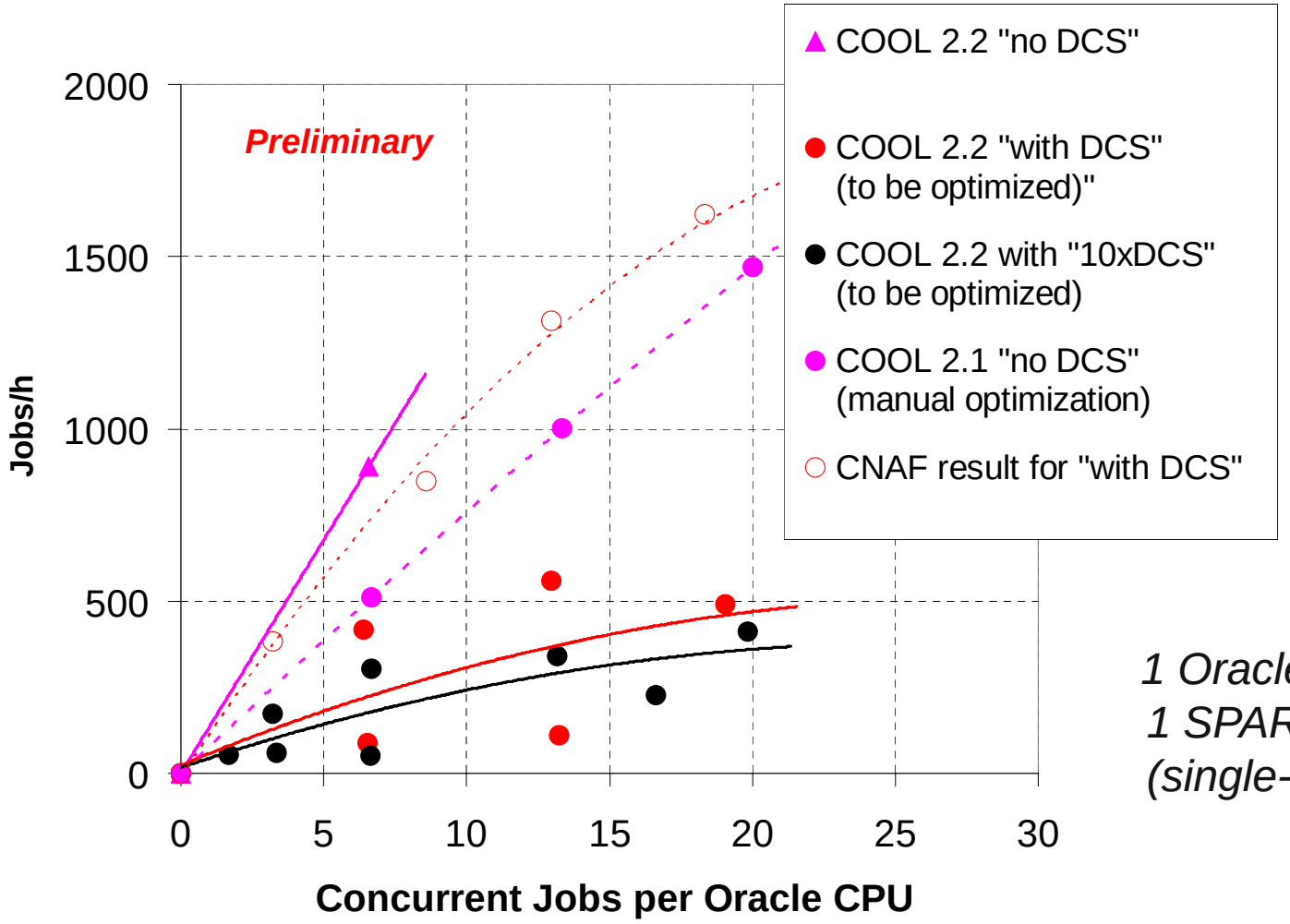
Latest Scalability Tests at Bologna CNAF Tier-1



1 Oracle CPU = 1 Intel CPU (dual-core)

- The top query for "**no DCS**"/"**with DCS**" cases is the COOL 'listChannels' call for Multi-Version folders - this query is not optimized in COOL 2.2
  - optimization is expected to result in further increase in performance

Latest Scalability Tests at Lyon CC IN2P3 Tier-1



1 Oracle CPU = 1 SPARC CPU (single-core)

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

Importance of Tests at Different Sites

- Among the reasons for performance differences between the sites is that we have seen the Oracle optimizer is choosing different query plans at Lyon and CNAF
  - For some yet unknown reason, at Lyon, optimizer is choosing not to use an index, and thus getting worse performance
    - *We don't yet know if this is because of the differences in hardware, different server parameters, different tuning or what*
- Also, at Lyon different throughputs were observed at different times
  - Thus, the performance is not yet fully understood
    - *Is it because of the shared server?*
- The calculated throughput (jobs/hor) is for the RAC tested (not per CPU)
  - Sites had very different hardware configurations:
    - *at CNAF, Bologna: 4 Intel cores = 2 "Oracle CPUs"*
    - *at CC IN2P3, Lyon: 3 SPARC cores = 3 "Oracle CPUs"*
  - There were more actual CPUs at CNAF than at Lyon, which accounts for some difference in RAC performances observed

In the Ballpark

- We estimate that ATLAS daily reconstruction and/or analysis jobs rates will be in the range from 100,000 to 1,000,000 jobs/day
  - Current ATLAS production finishes up to 55,000 jobs/day
- For each of ten Tier-1 centers that corresponds to the rates of 400 to 4,000 jobs/hour
- For many Tier-1s pledging ~5% capacities (vs. 1/10th of the capacities) that corresponds 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

Conclusions, Plans and Credits

- A useful framework for Oracle scalability tests has been developed
- Initial Oracle scalability tests indicate that WLCG 3D capacities in deployment for ATLAS are in the ballpark of what ATLAS requested
- We plan to continue scalability tests with new COOL releases
  - Initial results indicate better COOL performance
- Future scalability tests will allow more precise determination of the actual ATLAS requirements for distributed database capacities
- 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 DB (**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 DB (**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*