ORACLE Scalability Tests

Xavier Espinal (PIC/IFAE)

PIC
port d'informació
científica

Outline

- Oracle scalability tests purpose
- Setup for exercising scalability tests
- Scalability tests performed at PIC
- Tentative plan for further scalability tests
- Conclusions

Oracle scalability tests purpose

- Oracle overload conditions observed at 50% of the ATLAs Tier-1s
  - Producing degradation of 3D streaming
  - If single site suffers degradation >4h start affecting streaming to all other Tier-1s
    - Improved to >6.5h (HW upgrades at CERN)
- Need to spot the reasons for overload
  - Efficient queries
  - Reduce opened sessions per job (one in next release)
  - Pilot oracle query: hold the job until the DB is free (see Florbela's talk)
- Controlled scalability tests help in finding the limit of the Tier-1s HW
  - Compute a reasonable benchmarking
    - Need to cope with the reprocessing demands at the site
      - ~1000 slots for a 10% Tier-1
        - Capability to run 1k concurrent jobs without affecting 3D streaming
        - Ensure good flow of jobs during repro campaigns

Needed setup for exercising scalability tests

- Required ATLAS release for reprocessing (now rel. 15.X) installed
- Tune the test job to work with site configuration
  - SW release path
  - SQLite Copy script (local access)
  - Random run number generator
- Useful monitoring:
  - Site ORACLE OEM
  - 3D stream monitor
  - Internal BW: storage-WNs, instances
- Possibility to hammer single instances
  - tnsnames.ora (LOAD_BALANCE = no)
- Possibility to use a reasonable number of CPUs
  - Better 100% of the needed reprocessing metrics
  - Lightweight jobs (single event processing)
- Tune DB connections user limits to allow (NConPerJob*NJobs)

Stress tests at PIC March09 (1/7)

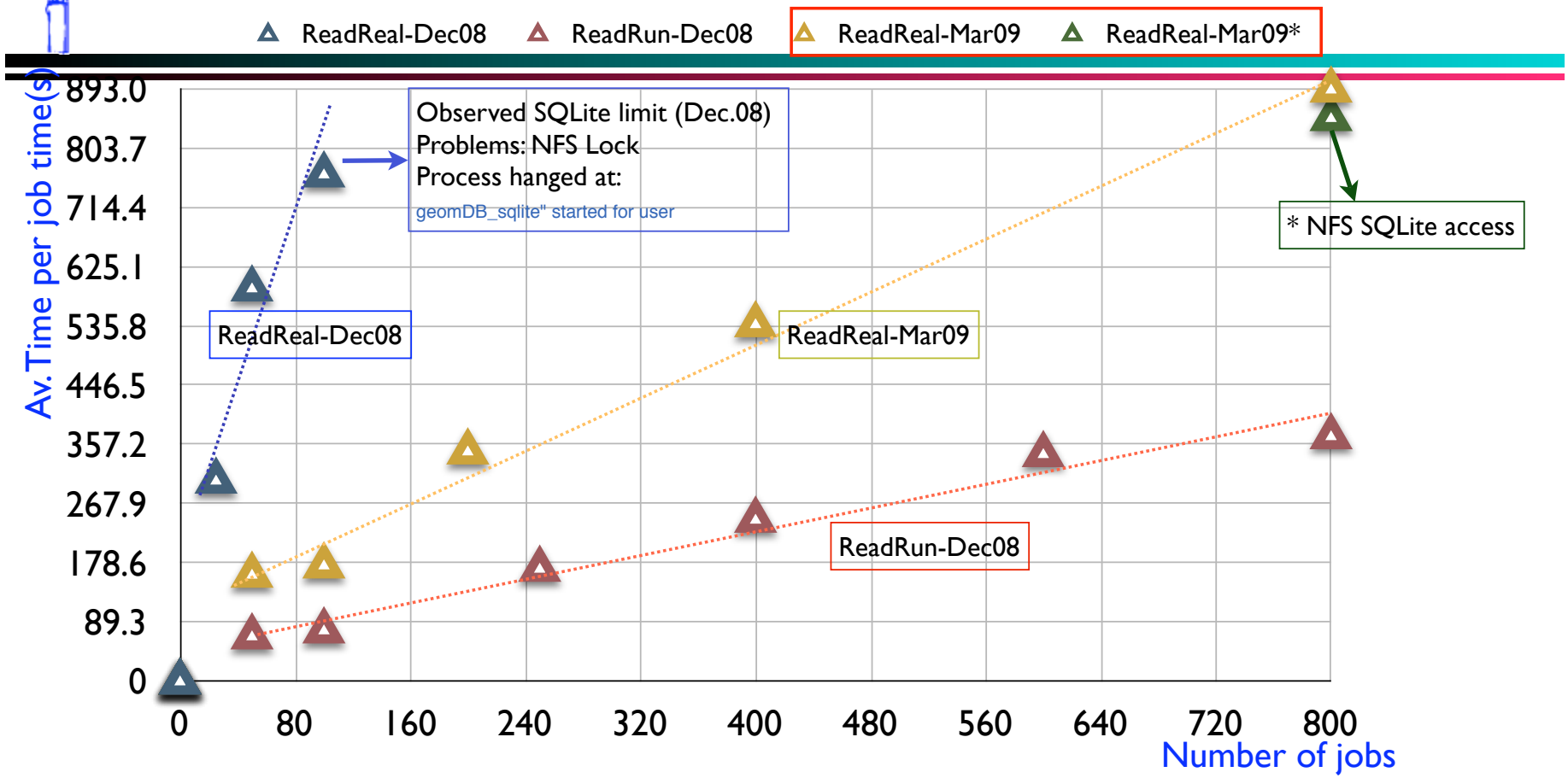
- Target
  - Throttle DB access jobs to fill all CPUs
    - COncurrent jobs sent in bulks: 50, 100, 200, 400 and 800 (farm limit)
    - Ensure concurrency in BS schedulers
- Job characteristics:
  - ATLAS SW used: rel. 14.5.0
  - 6 COOL connections opened per job
  - Single event processing
  - SQLite local acces (NFS SQLite access tested in parallel)
- DB settings:
  - Processes: 2k
  - Sessions allowed: 6k
  - Sessions per user: 5k (800 Athena jobs x6 connections)

Stress tests at PIC March09 (2/7)

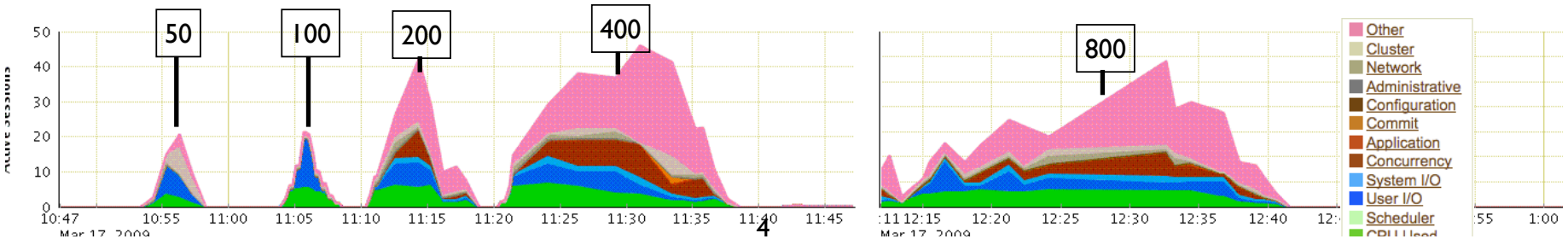
- All jobs finished successfully

#Jobs	Total elapsed time (s)	Mean time per job (s)	Time window
50	8060	161	10:54-10:58
100	17598	175	11:03-11:07
200	69451	347	11:10-11:20
400	215040	538	11:20-11:36
800	698102	893	11:46-12:13
800*	680100	850	12:13-12:40

Stress tests at PIC March09 (3/7)

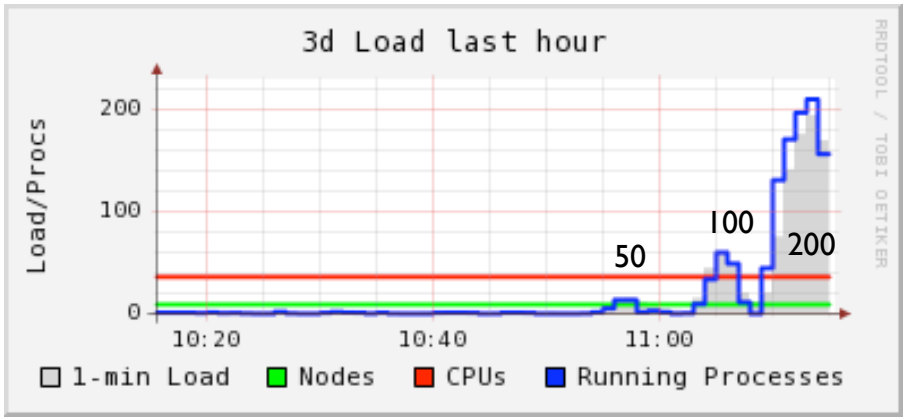


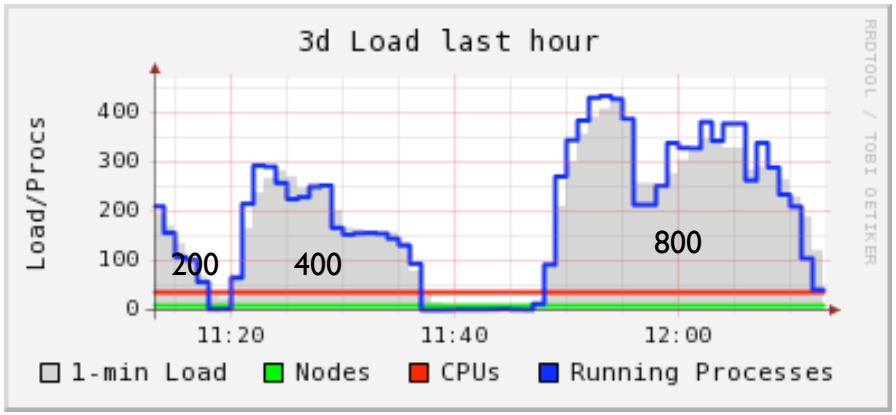
ATL3D all instances



Stress tests at PIC March09 (4/7)

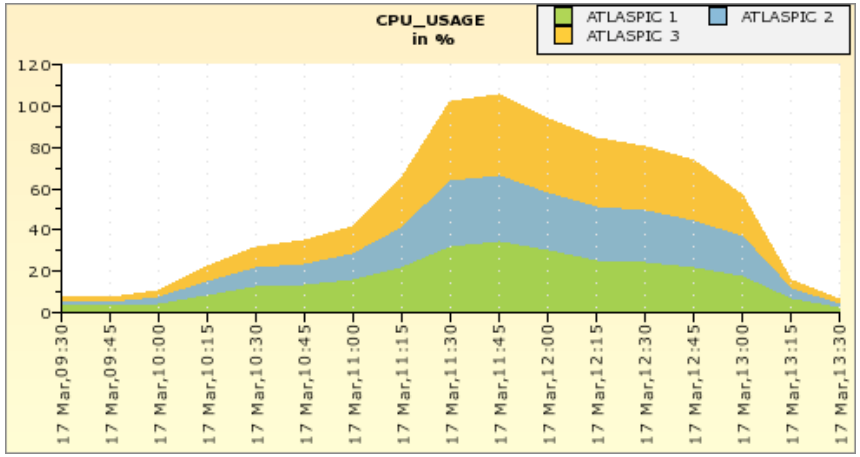
- Rac loads:

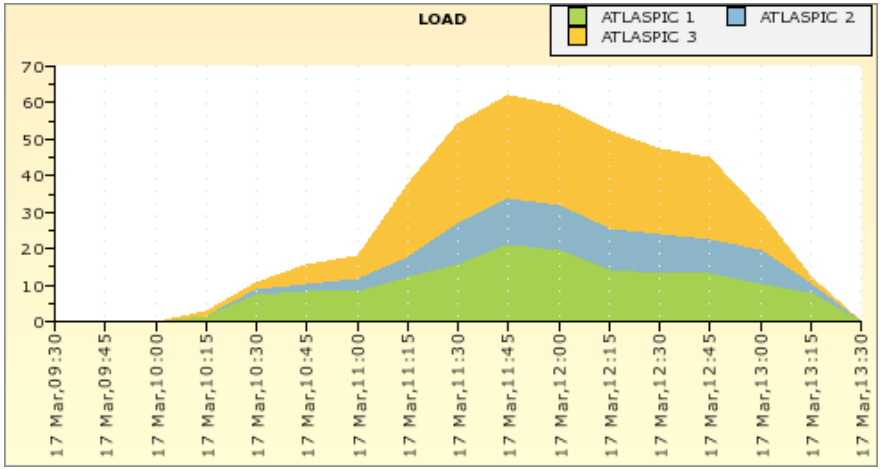


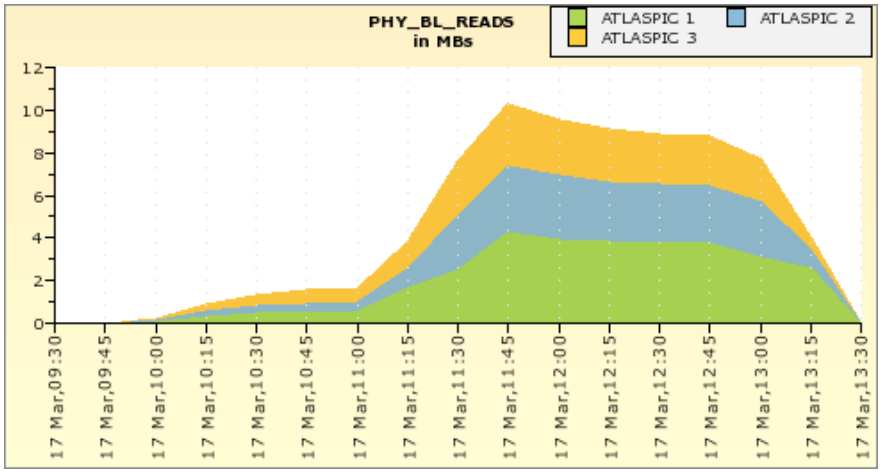


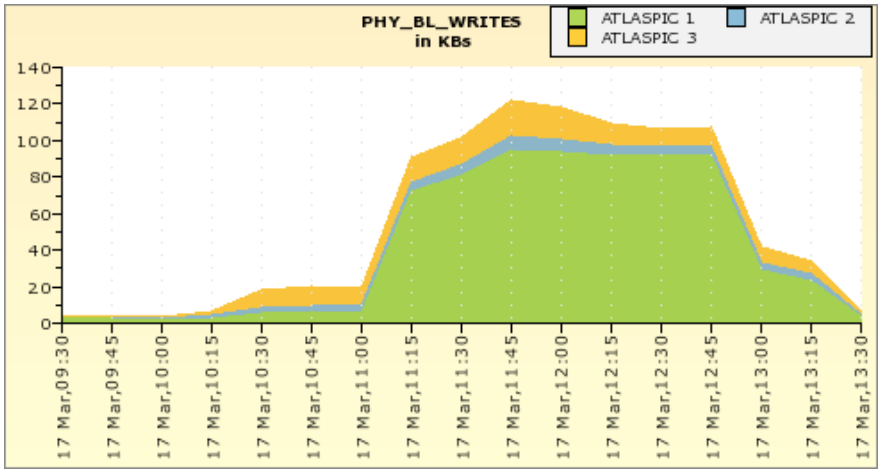
Stress tests at PIC March09 (5/7)

- Loads seen from Stream Monitoring:



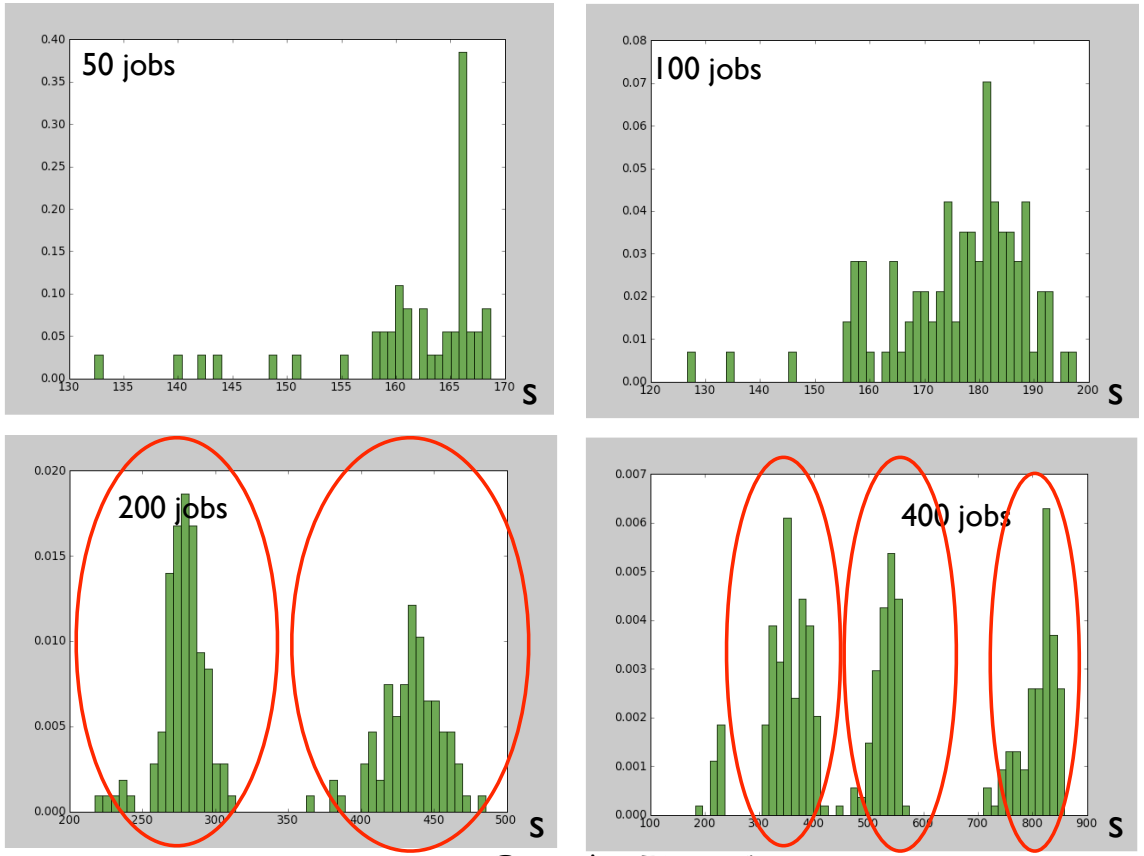






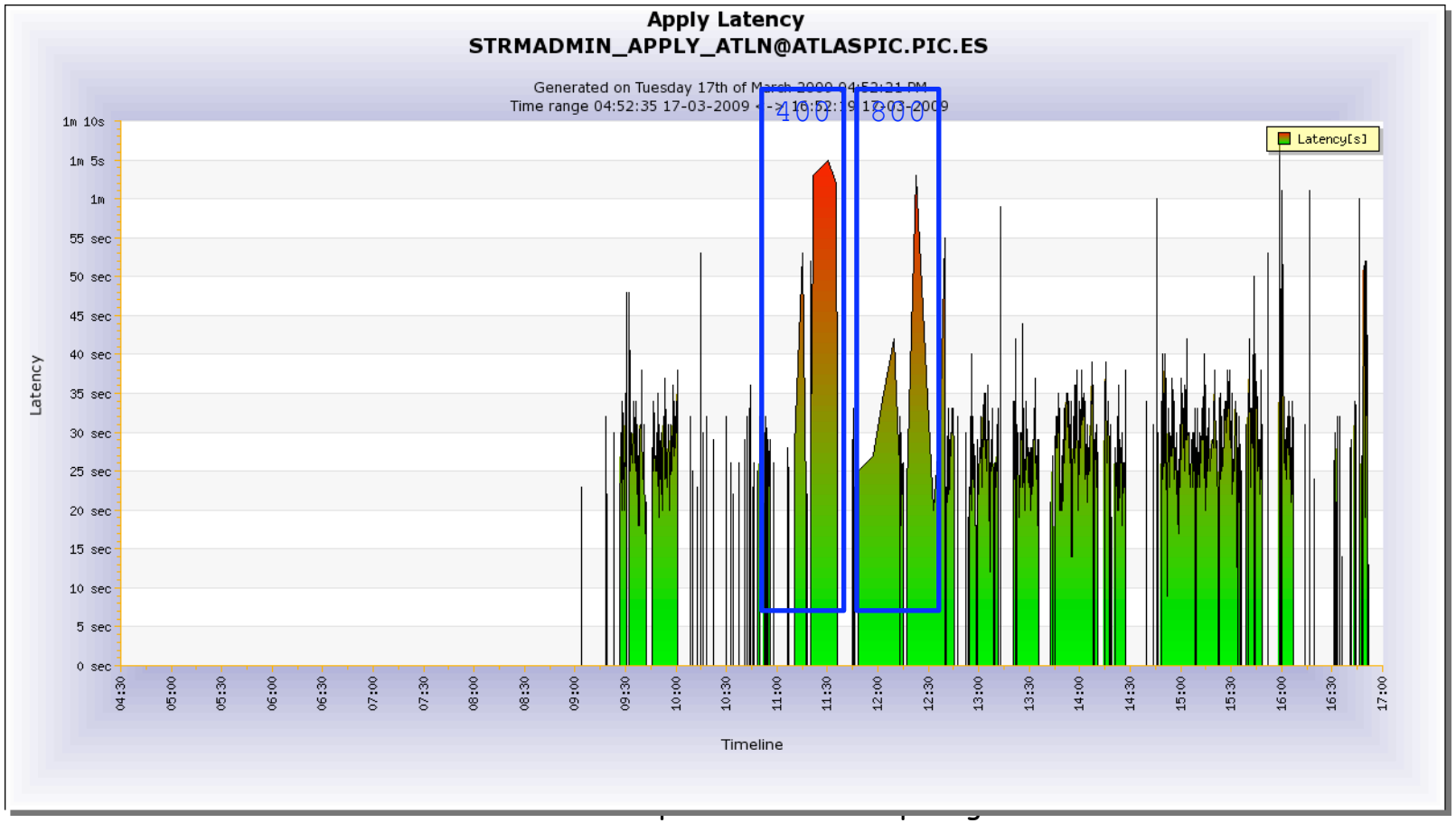
Stress tests at PIC March09 (6/7)

- Structures in job processing time appear when CPU is overloaded
  - Need further investigation by DBA experts (PIC: Luis/Elena and ATLAS:Gancho/Florbela)
  - Curious to see if this happen at other sites



Stress tests at PIC March09 (7/7)

- Tests impact on 3D streaming
  - Latency grew to ~60 sec (x2 normal behavior)
    - No major impacts, quick recovering



Further tests: Tier-1 benchmarking (1/2)

- We have the intention to perform further tests at all Tier-1s
  - Using newer release 15.X
    - Modified access pattern (single COOL connection per job)
    - Test Oracle pilot queries
  - Try to mimic needed reprocessing metrics
    - 1k CPUs for a 10% Tier-1
  - Take profit of Scheduled Downtimes or low MC activity
    - Ramping up to 100% of the repro requirement
  - Follow same framework as the tests performed at PIC
    - Single event processing
    - Local SQLite copy

Further tests: Tier-1 benchmarking (2/2)

- Would be good to coordinate the exercise within ATLAS ADC operations group
  - Coordinate different dates for each Tier-1:
    - Need follow up from ATLAS CERN DBAs:
      - Site ORACLE tuning
      - 3D Stream monitor:
        - Stream replication delays
        - Integral induced loads (all instances view)
    - Need involvement from site DB expert
      - ORACLE OEM monitor
        - Internal performance
    - Need involvement from site ATLAS contact
      - Prepare jobs, submit locally and force concurrency
      - Job monitoring
  - Need aprox. 2h time window to use req. number of CPUs

Conclusions

- ATLAS April repro campaign was a success:
  - Majority of jobs finished at all clouds (except TW)
  - Forcing pre-stage from tapes
    - Write and read buffers cleaned in advance (except FR cloud)
      - ➡ Worst case: this will not happen during real repro
  - But no 3D, local DB access (SQLite)
- CondDB need to be stressed as well
  - New release will give a good feeling about the loads for real data repro
  - Expected loads not fully tested
    - Need for full Tier-1 condDB stress tests
  - Spot and target possible showstoppers
    - HW, SW, access patterns
  - Still some months until start of collisions
    - But less time before STEP09
      - ➡ Test multi-VO ORACLE access before collisions?