Castor Databases at CERN

Nilo Segura Chinchilla

Oracle Support - IT/DES

CERN - Geneva

Agenda

- Configuration
- Monitoring
- Performance/Tuning
- Issues
- Future

Configuration

- Hardware
  - Three two-nodes RAC (Stager,SRM,DLF) per experiment plus another one for non-LHC
    - RACs also for special tasks (Name Server,Repack,T3)
    - Single database instances for Dev-Test/Stress-Test/Preproduction
  - Four cores with 8Gb memory
  - Netapp boxes for the storage
    - Access to the database filesystems via NFS
    - RAID 6 DP
    - Automatically resizable filesystems
  - Redundant configuration in all components

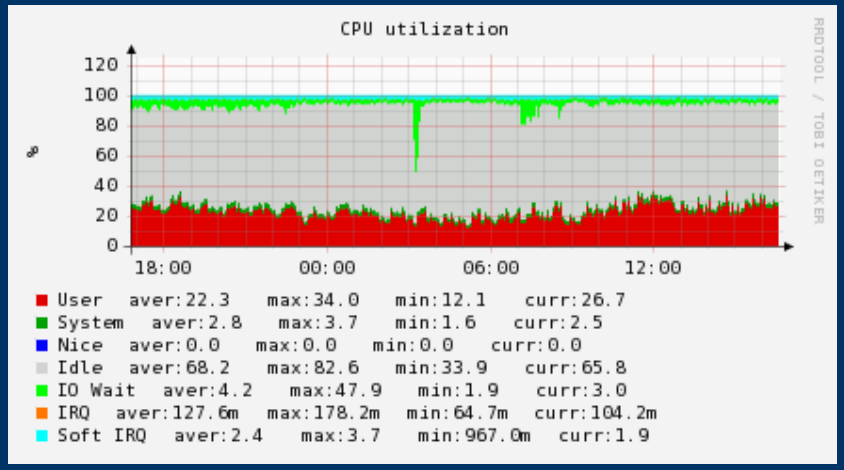
Configuration

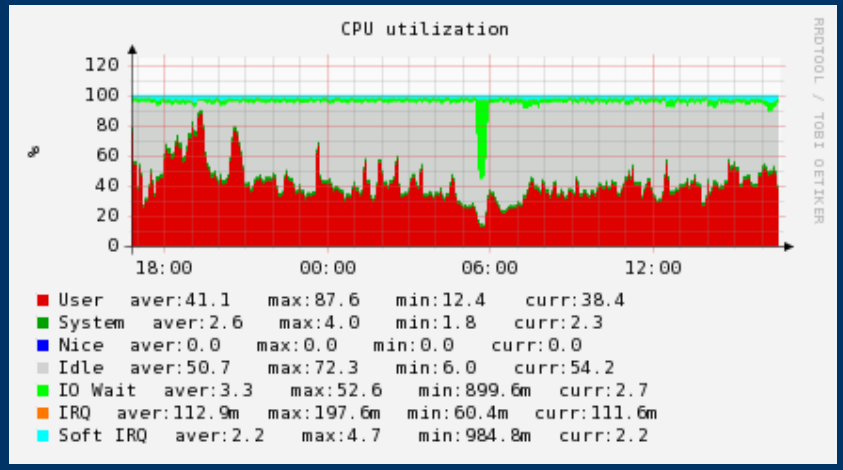
- Software
  - Red Hat 4.0 Kernel 2.6.9-78.0.5.ELsmp
  - RDBMS 10.2.0.4 + CPU Oct 2008 + List of relevant patches
    - Name Server still 10.2.0.3!!!
    - CPU Jan 2009 ready for deployment
  - EM Agent 10.2.0.4 + List of relevant patches
  - RPMs for CRS/RDBMS/Agent
- More info can be found here :
  - https://twiki.cern.ch/twiki/bin/view/DESgroup/CastorDatabases

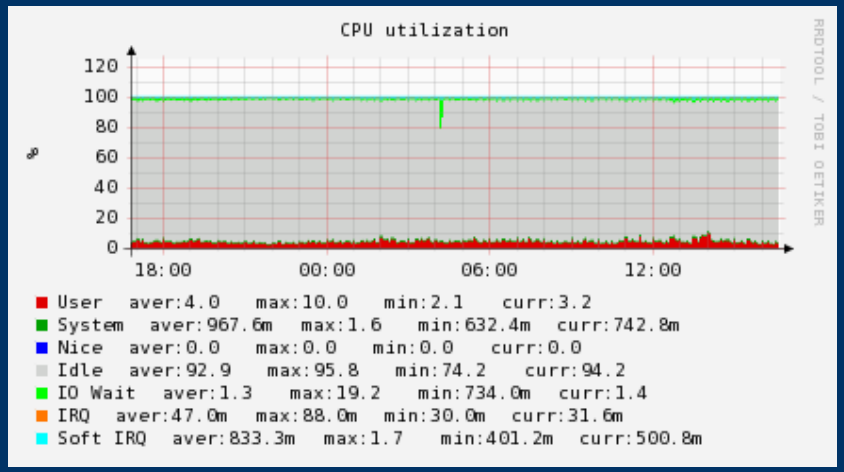
Monitoring

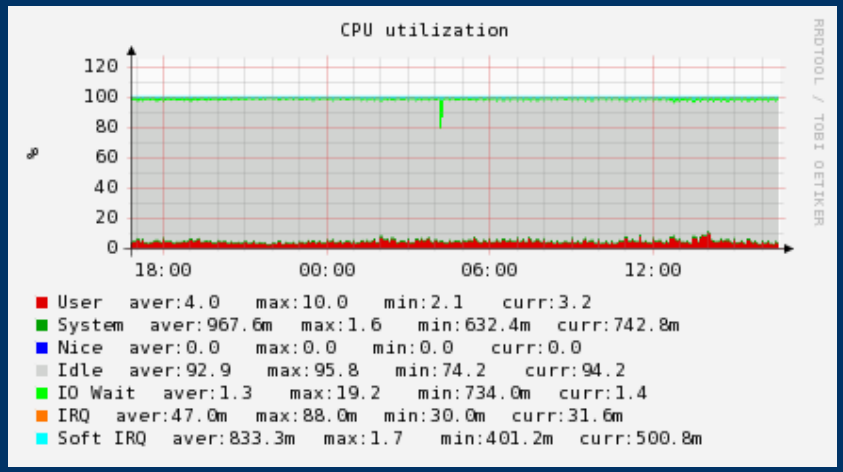
- Lemon gives a good idea of the system status
- AWR snapshots : every 20m (default 1h)
- Oracle Enterprise Manager
  - Top activity
  - Wait times
  - Execution plan & statistics
- And a few sqlplus based scripts..
- SMS messages

Atlas – CMS – LHCb - Alice

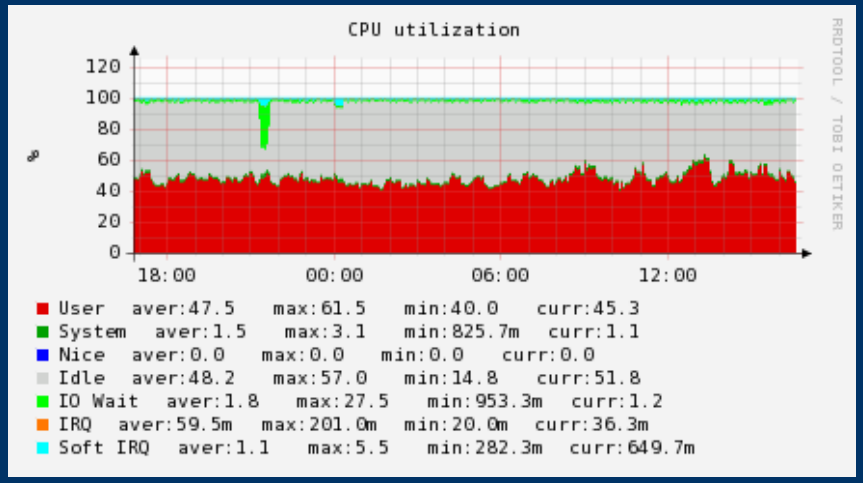


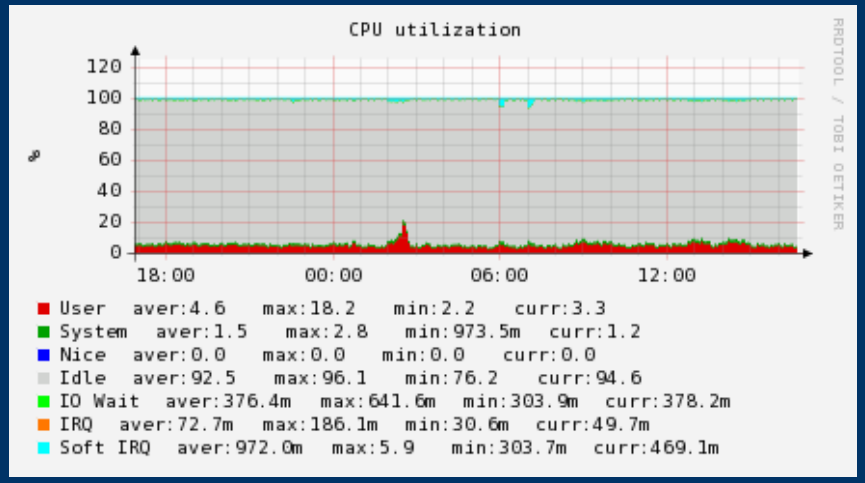




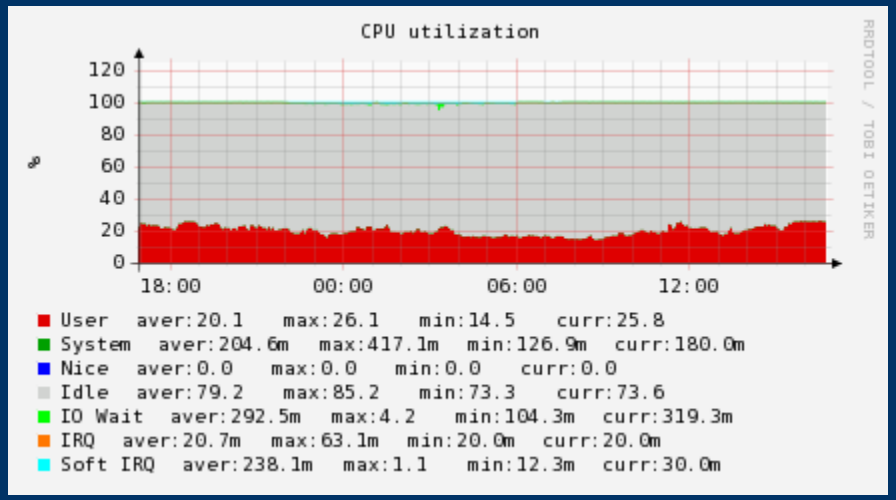


Public - NameServer





Atlas DLF


CPU utilization
%
120
100
80
60
40
20
0
18:00
00:00
06:00
12:00
RRDTOOL / TOBI OETIKER
User aver:20.1 max:26.1 min:14.5 curr:25.8
System aver:204.6m max:417.1m min:126.9m curr:180.0m
Nice aver:0.0 max:0.0 min:0.0 curr:0.0
Idle aver:79.2 max:85.2 min:73.3 curr:73.6
IO Wait aver:292.5m max:4.2 min:104.3m curr:319.3m
IRQ aver:20.7m max:63.1m min:20.0m curr:20.0m
Soft IRQ aver:238.1m max:1.1 min:12.3m curr:30.0m

Performance/Tuning

- Smooth running since Dec2008
- Replaced default db statistics job
  - Estimate percent 80, “for all columns size 1”
- Applied only to Stager and SRM
  - DLF still uses the default db stats job
- Still find (occasionally) potential areas where performance could be improved
  - New single/multi column(s) indexes
  - Modifications in the pl/sql – sql code

Performance

- Recent incident (6/Feb) on SRM/Atlas caused by multiple Full Table Scans on STAGEREQUEST
  - optimizer believed that table was small therefore FTS over Index access
  - wait events: latch buffer cache chains, read by other session
- Fixed by setting manually NUMROWS and NUMBLOCKS
  - dbms_stats.set_table_stats('SRM_ATLAS2','STAGEREQUEST',numblocks=>10000000,numrowss=>10000000,no_invalidate=>false);
  - disabled the new stats jobs and locked the stats for the db account until the problem is understood

Performance/Tuning

- In SRM, we can see temporary row lock contention
  - Several sessions involved
  - It happens when a non-db operation takes longer than expected (while holding a table row lock).
- Identified a new index on (SRM) SUBREQUEST (CASTORFILENAME)
  - In CVS, not deployed

Performance/Tuning

- Different tuning possibilities for SRM and Stager
  - Code rewrite (plsql) - Stager
  - Optimizer hints - Stager
  - Tweak optimizer values for certain objects - Stager/SRM
  - SQL Profiles (dbms_sqltune or via OEM) – Stager/SRM
  - Tune optimizer job parameters
    - “A la carte” settings per table/index

Issues

- Shrink table will occasionally generate INVALID ROWID error
  - Documented behaviour - need to re-issue the stmt and works
    - But not practical!!!!
  - After discussions, no longer SHRINK operations online (not critical anymore after 2.1.8)
    - Run during scheduled maintenance period (?)
- Heavy usage of 2.1.8 in Repack uncovered deadlocks in the code
  - Hot fixes (changes in pl/sql code) deploye

Issues

- Single Name Server for ALL stagers is a bad idea
  - “Logical” single point of failure (crashes, maintenance/upgrades)
- Cleanup job not running since December (CMS,Public)
  - Tables size increased significantly but performance was not affected
- How to protect the system against bad user usage ?

Issues

- Current monitoring jobs (castor_read) running on the Stager are very inefficient
  - Several jobs reading several times large tables (diskcopy,castorfile...)
  - Should there be any problem, the dba disables the jobs.

Future 2009

- Hardware
  - Some servers running around 50-70% CPU user time
    - Increase the number of CPU/Cores ?
  - Group all the DLFs in just one single multi-node RAC ?
    - Reduce by ½ the number nodes : easier to maintain, less power.
  - Storage wise we are good
- Software
  - Clients : have to use 10.2.0.4 Instant Client to match server version
    - But needs to be re-linked due to “naming” problems with Oracle libs and gcc 4.x
  - Server: Migration to Red Hat 5.x, Oracle 11gR1 or (10.2.0.5)
    - No plans yet, but 11.1.0.7 test database available for Castor
    - 10.2.0.5 (Terminal Release) available by Summer 2009 (?)

Future 2009

- Configuration
  - DB side connection pooling for SRM ?
- Performance
  - Get rid of the optimiser hints in the code
    - SQL Profiles, SQL Plan Management (as of 11gR1)
  - Catalog all the “right” execution plans ?
    - Create SQL profiles via OMS and distribute as part of the Castor software ?

Q & A