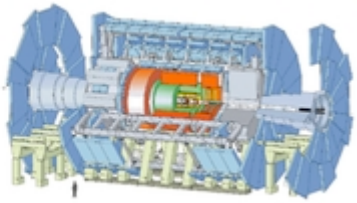
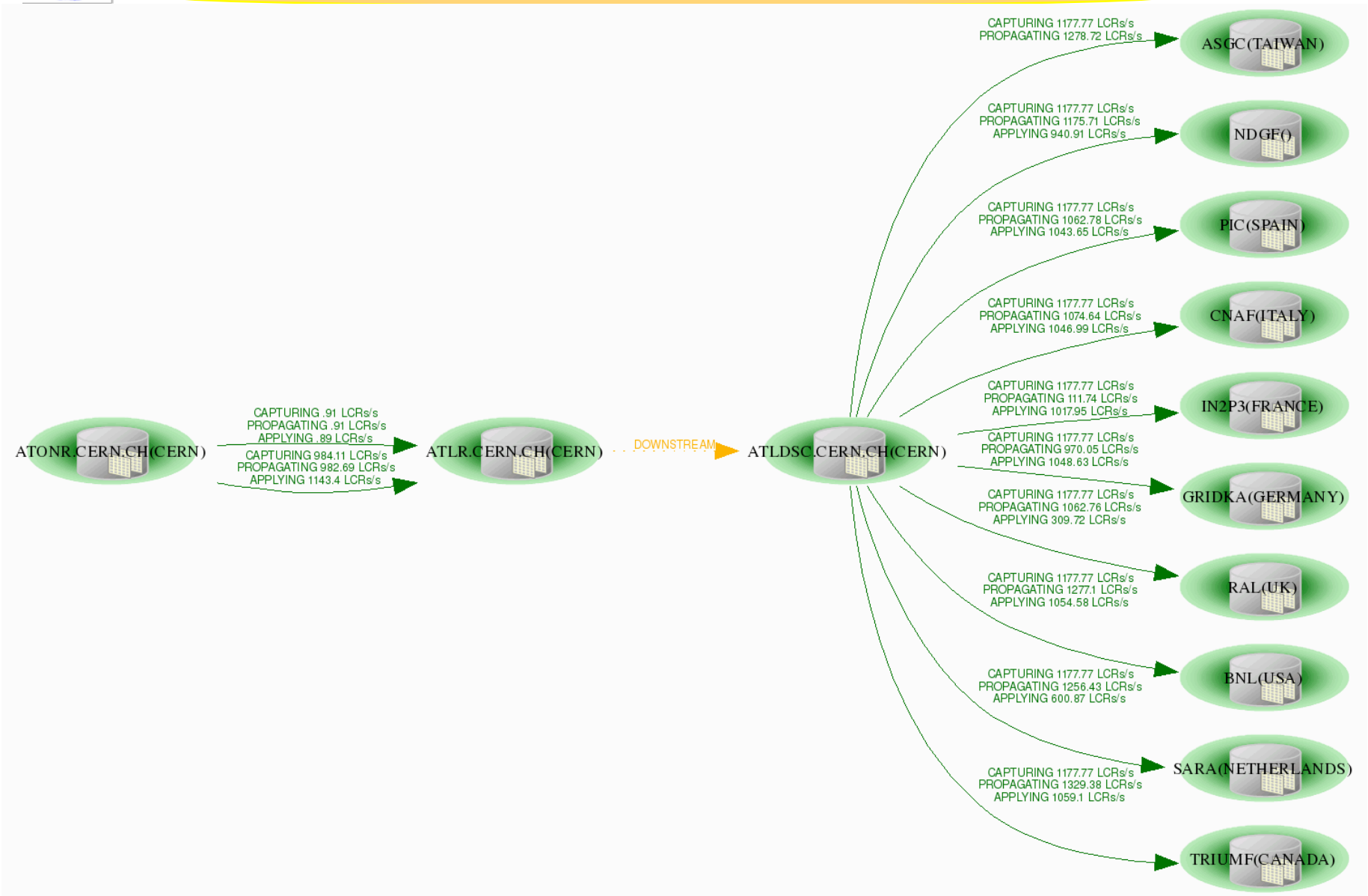
the ATLAS Experiment



Memory and process limit review on the Tier1s Cond. databases for ATLAS

Gancho Dimitrov (DESY)

The ATLAS Oracle replication data flow



Outline

- What are the ATLAS conditions databases @ the T1s used for?
- The current hw resources
- How much memory a single COOL reader process needs on the server?
- The typical ATLAS queries on the conditions data
- The memory used for sorting
- The current settings of sessions, processes, SGA and PGA @ T1s
- Setting the max number of sessions for each of the READER accounts
- Setting the limits using Oracle profiles

What are the ATLAS conditions databases @ the T1s used for?

- The conditions database store the detector environment conditions (e.g. temp, low, high voltages … etc)
- The data is used from the jobs to analyze at what particular conditions certain events have happened

The current hw resources @ T1s

Current Hardware Resources of the ATLAS Cond. databases at the Tier-1

Site	CPU cores	RAM GB	Instances	Total CPU cores
ASGC3D.GRID.SINICA.EDU.TW	4	8	2	8
ATLAS.DB1TIER1.NDGF.ORG	8	16	3	24
ATLASPIC.PIC.ES	4	8	3	12
CONDATLA.CR.CNAF.INFN.IT	4	8	3	12
DBATL.IN2P3.FR	8	16	4	32
LCGDB1.GRIDKA.DE	4	4	2	8
OGMA.GRIDPP.RL.AC.UK	8	16	3	24
USATCOND.BNL.GOV	8	16	2	16
SARADB.GRID.SARA.NL	4	16	7	28
TRAC.TRIUMF.CA	4	10	2	8

Note:

1) The four IN2P3 nodes are shared between the ATLAS cond. DB and the ATLAS AMI DB

2) The seven SARA nodes has one Oracle DB, but it hosts the database schemes of ATLAS condition, LHCb condtion, ATLAS LFC and FTS and LHCb LFC

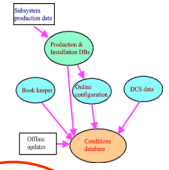
How much memory a single COOL reader process needs on the server?

- When a session is created, often Oracle allocates ~ 3 MB from the PGA (but only 0.5 MB when I reconnect shortly after the previous session is closed, other times (on first login) is 6 MB !?)
- For the COOL queries, Oracle does NOT apply hash or merge joins, but uses only nested loops ('guaranteed' by the hints in the queries). In that sense the PGA memory requirements are minimal.
- But the COOL API requires always the user to specify a sort order (either by CHANNEL ID first or by IOV_SINCE first). For this sorting memory is necessary … (how much?)
- The 'usual' interval which the ATLAS Athena jobs use by default for timestamp folders is 600 seconds (10 minutes)

Typical ‘heavy’ query in the COOL DB

```
SELECT /*+ QB_NAME(MAIN) INDEX(@MAIN "COOL_I3"@MAIN
("CHANNEL_ID" "IOV_SINCE" "IOV_UNTIL")) LEADING(@MAIN
"COOL_C2"@MAIN "COOL_I3"@MAIN) USE_NL(@MAIN
"COOL_I3"@MAIN) INDEX(@MAX1 COOL_I1@MAX1 ("CHANNEL_ID"
"IOV_SINCE" "IOV_UNTIL")) */ "COOL_I3"."OBJECT_ID",
"COOL_I3"."CHANNEL_ID", "COOL_I3"."IOV_SINCE",
"COOL_I3"."IOV_UNTIL", "COOL_I3"."USER_TAG_ID",
"COOL_I3"."SYS_INSTIME", "COOL_I3"."LASTMOD_DATE",
"COOL_I3"."ORIGINAL_ID", "COOL_I3"."NEW_HEAD_ID",
"COOL_I3"."HVCHVOLT_RECV", "COOL_I3"."HVCHCURR_RECV"
FROM ATLAS_COOLOFL_DCS."COMP200_F0023_CHANNELS"
"COOL_C2", ATLAS_COOLOFL_DCS."COMP200_F0023_IOVS" "COOL_I3"
WHERE "COOL_I3"."CHANNEL_ID"="COOL_C2"."CHANNEL_ID" AND
"COOL_I3"."IOV_SINCE">=COALESCE(( SELECT /*+ QB_NAME(MAX1) */
MAX(COOL_I1."IOV_SINCE") FROM
ATLAS_COOLOFL_DCS.COMP200_F0023_IOVS COOL_I1
WHERE COOL_I1."CHANNEL_ID"="COOL_C2"."CHANNEL_ID" AND
COOL_I1."IOV_SINCE"<=:"since1" ),:"sinc3s") AND
"COOL_I3"."IOV_SINCE"<=:"until3" AND "COOL_I3"."IOV_UNTIL">:"sinc3u"
ORDER BY "COOL_I3"."CHANNEL_ID" ASC, "COOL_I3"."IOV_SINCE" ASC
```

The number of block reads ?



Id	Operation	Name	Starts	E-Rows	A-Rows	A-Time	Buffers	Reads
1	SORT ORDER BY		1	1	7093	00:00:10.66	30640	3286
* 2	FILTER		1		7093	00:00:10.62	30640	3286
3	TABLE ACCESS BY INDEX ROWID	COMP200_F0023_IOVS	1	3	7093	00:00:10.59	30640	3286
4	NESTED LOOPS		1	12936	11182	00:00:09.54	24559	3060
5	INDEX FAST FULL SCAN	COMP200_F0023_CHANNELS_PK	1	4088	4088	00:00:00.03	24	7
* 6	INDEX RANGE SCAN	COMP200_F0023_IOVS_CSU_3INDX	4088	1	7093	00:00:09.46	24535	3053
7	SORT AGGREGATE		4088	1	4088	00:00:09.33	12266	3050
	|							
8	FIRST ROW		4088	21703	4088	00:00:09.30	12266	3050
* 9	INDEX RANGE SCAN (MIN/MAX)	COMP200_F0023_IOVS_CSU_3INDX	4088	21703	4088	00:00:09.28	12266	3050
10	SORT AGGREGATE		4088	1	4088	00:00:09.33	12266	3050
11	FIRST ROW		4088	21703	4088	00:00:09.30	12266	3050
* 12	INDEX RANGE SCAN (MIN/MAX)	COMP200_F0023_IOVS_CSU_3INDX	4088	21703	4088	00:00:09.28	12266	3050

Because is hard to justify how many blocks need to be read in the buffer pool when issuing a standard COOL query, calculating an appropriate size of the buffer pool is not possible.

By that reason we try to find the appropriate size of the PGA and the rest of the physical memory can be left for the database SGA, the ASM instance and the operating system

PGA memory used for sorting (test on the ATLAS 'offline' database ATLR)

- Test case 1: the shown query executed for a 10 min time interval

  7093 rows selected (~ 700 KB)

  After the execution : PGA used: 2342973 (~ 2MB),

  PGA allocated: 6017781 (~ 6 MB)

  PGA freeable:1048576 (1 MB)

  PGA max allocated: 6017781 (~ 6 MB)

```
-------------------------------------------------------------------------------------------------------------------------------
| Id  | Operation          | Name  | Starts | E-Rows | A-Rows |   A-Time      | Buffers |  OMem |  1Mem |  O/1/M  |
-------------------------------------------------------------------------------------------------------------------------------
| 1 | SORT ORDER BY |       |      1 |      1 |   7093 |00:00:01.80 |   30633 | 1186K|  567K|    1/0/0|
```

- Test case 2: The same query **without** the ORDER BY clause

  PGA used:493757 (~ 0.5 MB),

  PGA allocated: 905973 (~ 1 MB)

  PGA freeable: 0

  PGA max allocated: 905973 (~ 1 MB)

The difference in the PGA allocated from case 1) and case 2) is 5 MB !!!

PGA allocated memory
(test on GridKa, IN2P3, BNL)

- Test case 1 @ GridKA:

  After the query execution : PGA used: 1458129 (~ 1.4MB)

  PGA allocated: 3773625 (~ 3.7 MB)

  PGA freeable: 917504

  PGA max allocated: 3773625 (~ 3.7 MB)

- Test case 1 @ IN2P3:

  After the query execution : PGA used: 2343981 (~ 2.3 MB)

  PGA allocated: 4509717 (~ 4.5 MB)

  PGA freeable: 1048576 (~1 MB)

  PGA max allocated: 4509717 (~ 4.5 MB)

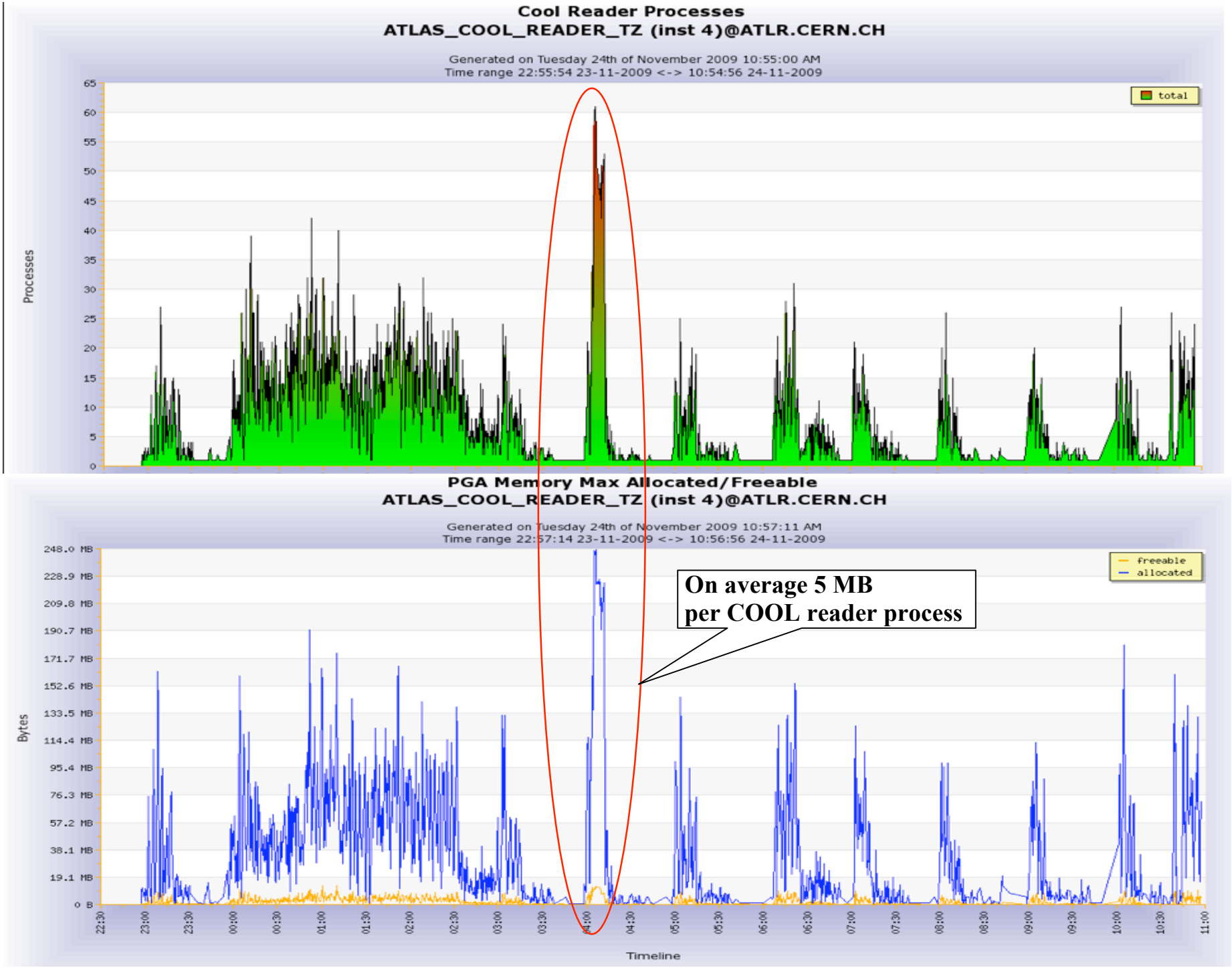
- Test case 1 @ BNL:

  After the query execution : PGA used: 2781573 (~ 2.7 MB)

  PGA allocated: 4706429 (~ 4.7 MB)

  PGA freeable: 0

  PGA max allocated: 4706429 (~ 4.7 MB)


Cool Reader Processes
ATLAS_COOL_READER_TZ (inst 4)@ATLR.CERN.CH
Generated on Tuesday 24th of November 2009 10:55:00 AM
Time range 22:55:54 23-11-2009 <-> 10:54:56 24-11-2009
total
Processes
65
60
55
50
45
40
35
30
25
20
15
10
5
0
PGA Memory Max Allocated/Freeable
ATLAS_COOL_READER_TZ (inst 4)@ATLR.CERN.CH
Generated on Tuesday 24th of November 2009 10:57:11 AM
Time range 22:57:14 23-11-2009 <-> 10:56:56 24-11-2009
freeable
allocated
On average 5 MB
per COOL reader process
Bytes
248.0 MB
228.9 MB
209.8 MB
190.7 MB
171.7 MB
152.6 MB
133.5 MB
114.4 MB
95.4 MB
76.3 MB
57.2 MB
38.1 MB
19.1 MB
0 B
22:30
23:00
23:30
00:00
00:30
01:00
01:30
02:00
02:30
03:00
03:30
04:00
04:30
05:00
05:30
06:00
06:30
07:00
07:30
08:00
08:30
09:00
09:30
10:00
10:30
11:00
Timeline

Statistics from the Tier1s

Statistics from Florbela's Tier1s monitoring on the average used memory from the ATLAS COOL reader processes on the Tier1s condition databases

(for November 2009)

DB_LINK	AVG_MEM_PROC
ORCL.BNL.GOV	3
ATLASPIC.PIC.ES	3
DBATL.IN2P3.FR	2
SARADB.GRID.SARA.NL	6
ATLAS.DB1TIER1.NDGF.ORG	3
OGMA.GRIDPP.RL.AC.UK	4
ASGC3D.GRID.SINICA.EDU.TW	5
TRAC.TRIUMF.CA	3
LCGDB1.GRIDKA.DE	2
CONDATLA.CR.CNAF.INFN.IT	3

The Oracle memory limits on sorting/hashing

- The total work area cannot exceed 200 MB, because of the default _pga_max_size setting

- The RAM sort cannot use more than 5% of the ‘pga_aggregate_target’ or ‘_pga_max_size’ whichever is smaller.

  For BNL, with pga_aggregate_target=4 GB, the limit is 200 MB

  For GridKa, with pga_aggregate_target= 400 MB, the limit is 20 MB

=>

- Assuming that on average a COOL reader process **will not use more than 10 MB** from the PGA and knowing that since ATLAS sw releases 15.0 the jobs connect, get the data and disconnect immediately, one can calculate the max user processes on the system avoiding swapping to the disk or ORA-04030: out of process memory when trying to allocate nn bytes

The current processes/sessions and SGA/PGA configuration on the ATLAS Oracle Tier1s

Site	processes	sessions	sga_max_size	pga_aggregate_target
ASGC3D.GRID.SINICA.EDU.TW	600	665	2GB	800MB
ATLAS.DB1TIER1.NDGF.ORG	800	1200	3GB	400MB
ATLASPIC.PIC.ES	2000	6000	5GB	2GB
CONDATLA.CR.CNAF.INFN.IT	500	555	2.6GB	1GB
DBATL.IN2P3.FR	1500	5000	8GB	1.6GB
LCGDB1.GRIDKA.DE	500	555	2GB	400MB
OGMA.GRIDPP.RL.AC.UK	800	885	4GB	1GB
ORCL.BNL.GOV	4000	4405	8GB	4GB
SARADB.GRID.SARA.NL	1000	1105	1.6GB	1.5GB
TRAC.TRIUMF.CA	800	1500	6GB	1GB

Note: RAL, NDGF and CNAF have a high difference between the RAM and the PGA+SGA.
For the sites where the sessions are times more the processes, a correction is needed

How to protect the DB from spawning all server processes, because of bursts of client jobs ?

- By Setting limits for each open production account via Oracle profiles.
- The sum of SESSIONS_PER_USER of ALL defined profiles has to be less than the DB level setting of 'sessions'. A hundred sessions (processes) have to be reserved for other activities, e.g. the Streams propagation, Streams apply, 3D and ATLAS monitoring, backup processes, power users ... etc
- The IDLE_TIME to be set to not more than 30 min

List of the open accounts @ the ATLAS Tier1 Oracle cond. databases

1. ATLAS_COOL_READER_U The ATLAS_COOL_READER_U is used from ATLAS sw releases as of version 15.0.0. The important here is that each job opens ONLY one DB session at a time.

2. ATLAS_CONF_TRIGGER_V2_R Reader account for the trigger configuration database

3. ATLAS_FRONTIER_READER User from the Frontiier servlet to connect to the Oracle DB (instead of the ATLAS_COOL_READER_U)

ATLAS DB TWIKI pages

More information on the profiles (and the script for creating them) can be found on the following page

https://twiki.cern.ch/twiki/bin/view/Atlas/OracleProcessesSessionsUsersProfiles

Further information is on

https://twiki.cern.ch/twiki/bin/view/Atlas/DataBases

https://twiki.cern.ch/twiki/bin/view/Atlas/ResourcesAndDBSettings

Final words …

- Some of the sites seem to have the ‘processes’ setting very high. Either it get changed or set much lower value for the SESSIONS_PER_USER in the COOL reader profile.
- Most of the sites have proper PGA size (certainly, the DBAs need to monitor whether DB swapping is not happening). For some, adjustment is needed if free RAM is available.
- The sites that have setting of sessions times more than the processes (old setting, because of old ATLAS sw was spawning several sessions within an Oracle process) have to re-adjust this using the Oracle formula sessions = processes * 1.1 + 5
- Having large SGA (buffer pool) helps a lot in reducing the disk reads. Some sites have available a lot of RAM, but the SGA set to be small.
- Profiles on the reader accounts to be created for setting limits on the number of sessions and max idle time

Backup slide - tables with most channels (data points)

A list with the ATLAS COOL folders with the highest number of defined channels (data points).

Queries on these tables that perform join with the relevant IOV table are expected to be the heaviest because of the nested loops (thus the high number of block reads)

OWNER	TABLE_NAME	NUM_ROWS
ATLAS_COOLONL_TDAQ	COMP200_F0005_CHANNELS	108392
ATLAS_COOLONL_TDAQ	COMP200_F0003_CHANNELS	13571
ATLAS_COOLONL_TDAQ	COMP200_F0004_CHANNELS	11352
ATLAS_COOLOFL_SCT	COMP200_F0004_CHANNELS	10186
ATLAS_COOLONL_SCT	COMP200_F0054_CHANNELS	10186
ATLAS_COOLONL_TDAQ	COMP200_F0007_CHANNELS	10176
ATLAS_COOLONL_TGC	COMP200_F0008_CHANNELS	5433
ATLAS_COOLONL_MDT	OFLP200_F0006_CHANNELS	4899
ATLAS_COOLONL_MDT	COMP200_F0007_CHANNELS	4899
ATLAS_COOLOFL_DCS	OFLP200_F0033_CHANNELS	4384
ATLAS_COOLOFL_DCS	COMP200_F0014_CHANNELS	4384
ATLAS_COOLONL_SCT	COMP200_F0033_CHANNELS	4108
ATLAS_COOLONL_SCT	COMP200_F0037_CHANNELS	4108
ATLAS_COOLOFL_SCT	OFLP200_F0011_CHANNELS	4108
ATLAS_COOLONL_SCT	OFLP200_F0038_CHANNELS	4108
ATLAS_COOLOFL_SCT	OFLP200_F0008_CHANNELS	4108
ATLAS_COOLONL_SCT	OFLP200_F0035_CHANNELS	4108
ATLAS_COOLOFL_DCS	COMP200_F0048_CHANNELS	4088
ATLAS_COOLOFL_DCS	OFLP200_F0026_CHANNELS	4088
ATLAS_COOLONL_SCT	COMP200_F0071_CHANNELS	4088
ATLAS_COOLOFL_DCS	COMP200_F0023_CHANNELS	4088
ATLAS_COOLOFL_DCS	COMP200_F0046_CHANNELS	4088
ATLAS_COOLOFL_DCS	OFLP200_F0017_CHANNELS	4088
ATLAS_COOLOFL_DCS	OFLP200_F0025_CHANNELS	4088
ATLAS_COOLOFL_DCS	OFLP200_F0020_CHANNELS	4088