Latest Status of the ALICE WMS Usage

WLCG GDB Meeting
CERN 11th Feb 2009

Patricia Méndez Lorenzo. Report based on the discussions with Maarten Litmaath

Reminder of the Christmas Report

- 3 ALICE-dedicated WMS nodes at CERN
  - wms103, wms109 and wms204
  - Submission to CERN and 'failover' for almost all ALICE sites
- wms204 - backlog during the Christmas running
  - gLite3.1, 8 core machine
  - job submission progressively slower
- wms103 and wms109 also a bit unstable
  - gLite3.0 in old hardware
  - **Replaced by two new WMS just after Christmas time: wms215 and wms214**

Follow-up hypotheses

- Several reasons for overload discussed:
  - Destination queue not available or any configuration problem at the destination site
    - Submitted jobs are then kept for 2h (up to 3 retries per job)
    - **We come back to this point**
  - ALICE jdl construction
    - A complicated jdl will slow down the matchmaking process
    - The workload manager will not be able to keep up with all the requests which are being sent by the WMProxy service
    - **We come back also to this point**
  - BDII overloaded
    - Discarded
    - Improvement in the next gLite3.1 version
  - Network problems
    - Discarded
  - Myproxy server overload
    - Discarded
    - **We come back to this point**

WMS behaviour

- As effect of this high load the new submitted requests where in status WAITING or READY forever
  - Suicide mode – new requests still coming and accepted, thus worsening the status….
  - **Request at that time: Could the WMS be configured to avoid new submissions once it gets in such a state?**
- The above hypotheses were considered as ingredients of a possible high load, but not the unique reason
  - Concluded that as soon as the experiment restarted the production we would follow carefully the evolution of the 3 nodes and report further issue to the developers

Current Status (I)

- This week ALICE MC production with many jobs
- High backlogs
  - More than 31000 queued jobs in wms215.cern.ch the 10th of February

ALICE dedicated WMS

WMS	date	RunJ	CurJ	DayJ	load	i.fl	q.fl
wms112	02/10-17:00	3	3	191	0.24	0	0
wms113	02/10-17:00	3	3	84	0.36	0	0
wms114	02/10-17:00	3	5	71	0.25	0	0
wms201	02/10-17:00	30	138	8415	2.01	43	0
wms202	02/10-17:00	3	4	160	0.26	0	0
wms203	02/10-17:00	354	5352	4595	2.52	30	0
wms204	02/10-17:00	229	658	4223	3.22	13612	0
wms205	02/10-17:00	677	1137	6011	8.64	55	0
wms206	02/10-17:00	39	156	8542	2.52	41	0
wms207	02/10-17:00	311	329	113	0.19	0	0
wms208	02/10-17:00	17	42	1059	1.90	169	0
wms209	02/10-17:00	12	51	1064	1.58	115	0
wms210	02/10-17:00	305	326	124	0.27	4	0
wms212	02/10-17:00	3	4	163	0.32	2	0
wms213	02/10-17:00	421	1596	10183	1.37	0	0
wms214	02/10-17:00	35	1062	4303	0.97	8	0
wms215	02/10-17:00	82	153	6521	2.57	31759	0
wms216	02/10-17:00	312	5433	4759	2.21	31	0

ALICE dedicated WMS

Hypotheses (I)

- Overload of the myproxy server
  - The ALICE submission procedure was changed in January to avoid this
    - Proxy delegation request once per hour - 'frugal' usage of myproxy server
- Conclusion - the WMS overload is uncorrelated with the use of myproxy

Hypotheses (II)

- The destination queue is not available
  - If the queue(s) declared in the jdl is not available the request will be in a standby status for 2h and each request will be tried to be matched until 3 times
    - These requests are not tracked
      - glite-wms-job-status would report Waiting for each such request
    - The submissions continue until the input.fl gets (might get) overloaded
    - At CERN (large number of submissions) this is possible
      - Huge number of submissions retried 3 times each…
    - But this 2h can be easily decreased
      - At this moment the value is hardcoded it can be easily decreased

ALICE follow-up (I)

- From the ALICE side
  - Current approach
    - Jobs are submitted when the site has free slots
    - The status of the queue in the BDII is however not queried
    - In the case that the queue is not available, the WMS is in any case used
  - Changes
    - Before submitting a new bunch the status of the queue will be queried
    - If not available, the job(s) will not be submitted
    - A useless usage of the WMS will be avoided

ALICE follow-up (II)

- The jdl is too complicated
  - This reason might be true at CERN, not at the rest of the sites
  - ALICE jdls contain all the available ALICE queues at each site
    - In the case of CERN this is a large number
    - The WMS needs to check all of them for the ranking
  - Using a wildcard expression (*.cern.ch) might help
    - However the problem was not visible with WMS gLite3.0 nor with the old RB
    - This issue is not visible only at CERN also at other sites with a small number of ALICE queues
    - WMS developer Marco Cecchi does not expect a clear gain from simplifying the jdl (all the queues must be checked anyway)

The next WMS gLite3.1

- ALICE has also volunteered to test the new version in a production environment and provide the developers with a feedback
- Some improvements in the new version
  - Allows for dedicated VO view of the BDII
    - Although accessing the whole production BDII, reading the part related to the VO only
    - Faster matchmaking due to fewer entries (linear relation)
  - Matchmaking time window configurable with YAIM
    - Although the current version already allows it

Conclusions

- We cannot conclude at this point the reason of the high backlogs observed by ALICE
- Several improvements from the ALICE side and also from the WMS configuration might help in the future
- The experiment is willing to test the new gLite 3.1 WMS version
- **The CREAM-CE deployment at CERN will ensure a smooth production and it is highly required**