WLCG Service Incident Report

KIT firewall and OPN overload by ALICE jobs

May 6, 2014

Time line

In the WLCG operations meeting of April 3 an operational instability was reported for ALICE jobs at KIT:

- around 14:00 CEST on April 1 the number of running jobs in MonALISA started going down from 5k+ to ~1k for unknown reasons
- since then the numbers fluctuated wildly around that low level, while the batch system typically had ~3k ALICE jobs running at any time
- no changes were made by ALICE and other sites appeared to be working OK

In the next many days ALICE operations experts and ALICE support experts at KIT spent a very large effort to understand what was going on and to implement mitigations while the true solution was not yet known.

As the ALICE VOBOX at KIT got overloaded with the error messages from the many failing jobs, the focus was first on that service:

- the VOBOX kernel network parameters were improved, which did not make much of a difference
- the error reporting code in AliEn was made more efficient, which allowed the VOBOX to cope, but did not yet reveal why the jobs were failing

A reboot of the VOBOX first seemed to cure the problems, but only for a day.

In the WLCG operations meeting of April 7 ATLAS and CMS first reported job and data transfer errors due to "network issues" at KIT.

In the meantime a cap of 1.5k was put on the number of concurrently running ALICE jobs at KIT and the situation stabilized at that low level, allowing significant work to get done at KIT, while experts continued looking into the matter.

A large fraction of those jobs were doing analysis (in trains) and from MonALISA plots and job logs it became clear that a sizable subset could not access the local SE for reading their input data and thereby had to resort to reading the data remotely from other sites, thereby putting an unexpected load on the KIT site firewall and the OPN link to CERN.

This observation led the KIT network experts to discover and fix a rack of WN whose network settings indeed prevented access to the local SE.

Even with that fix in place, large numbers of analysis jobs intermittently read from remote SEs, while the job logs did not contain errors pertaining to the local SE: the jobs appeared to be picking remote SEs right away, despite the local presence of the data and the local SE being in good shape.

To reduce the load on the firewall and the OPN, the jobs cap was further lowered to 1k.

The cause of the remote reading was finally identified on Fri April 11:

- for analysis jobs the location of the WN was not propagated to the central services
- the client code then ended up using not only the local replica, but close replicas as well

A patch was prepared during the weekend and a first analysis tag with the fix became available on Tue April 15 and was verified in a few workflows.

The fix has been and will be included in all subsequent analysis tags (1 per day in the months leading up to Quark Matter 2014), while older tags will no longer be readily available to users after QM.

Though the vast majority of analysis jobs (trains) quickly moved to recent tags with the fix, older tags were still occasionally used by significant numbers of jobs, leading to significant load on the firewall and OPN link. ALICE therefore has needed to keep the KIT jobs cap lower than usual until early May, at 3k or less.

Conclusions

- Though the problem in principle was affecting jobs at all sites, KIT turned out to be the most affected.
- As a result of the thorough investigations, a number of improvements were implemented in AliEn, MonALISA and the network configuration at KIT.
- We thank the experts at KIT for all their efforts and the other experiments for their patience!

Links

MonALISA:

- http://alimonitor.cern.ch/map.jsp

OPN statistics:

- https://netstat.cern.ch/monitoring/network-statistics/ext/?p=LHCOPN

WLCG operations meetings:

- https://twiki.cern.ch/twiki/bin/view/LCG/WLCGOperationsMeetings