Toolforge webservices are in the final stages of migrating to the toolforge.org domain.
Please help us clean up older documentation referring to tools.wmflabs.org!

Main page
Recent changes
Server admin log (Prod)
Server admin log (RelEng)
Deployments
SRE/Operations Help
Incident status

Cloud VPS & Toolforge
Cloud VPS documentation
Toolforge documentation
Request Cloud VPS project
Server admin log (Cloud VPS)

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Incident documentation/20150527-GridEngine

< Incident documentation

Contents [hide]

Summary

Tool Labs uses GridEngine for distributing jobs across multiple hosts. At about 1830 UTC on 27 May 2015, the GridEngine master died and refused to come back up. This was triggered by a restart due to a package upgrade for T98577. Existing jobs continued fine but no new jobs could be started. Various approaches were tried, and disabling nscd on tools-master made the issue a lot less pressing at around 2000 UTC (one failure in about 1000 attempts). Things remained at this state until about 1120 UTC on 28 May 2015, where it blew up again. Eventually it was traced down to lines in /etc/hosts being too long for gridengine to process, and fixed in 1340 UTC.

Phab ticket: https://phabricator.wikimedia.org/T100554 and subtasks.

Timeline

27th May 2015

18:20 - New gridengine-common package uploaded to carbon for T100073, toollabs instances start updating themselves. 18:30 - Reports of gridengine commands (qstat) failing come in on IRC

18:40 - Yuvi downgrades gridengine-common to previous version, restarts gridengine-master on tools-master. No change.

18:45 - `/data/project/.system/gridengine/default/common/act_qmaster` found to point to localhost, so tools were trying to contact localhost for gridengine-master. This file is on NFS and written to by the gridengine master, so it kept thinking for some reason it was localhost. Forcing that file to point to tools-master also does not work, qstat failing with `error: commlib error: access denied (server host resolves destination host "tools-master.eqiad.wmflabs" as "(HOST_NOT_RESOLVABLE)")`

18:50 - Yuvi tries to switchover to tools-shadow, following documentation at https://wikitech.wikimedia.org/wiki/Nova_Resource:Tools/Admin. Failover fails as well - tools-shadow's gridengine-master does not come up when explicitly started even if tools-master's is killed.

18:50 to 19:50 - Yuvi and Valhallasw try various things, including restarting nscd, restarting tools-master itself, rejigging entries in /etc/hosts (to point to tools-master for 127.0.0.1), stracing to attempt to figure out what's going wrong, read through plenty of other people struggle through GridEngine issues, to mostly no effect. We find that gridengine has its own 'utility' function for gethostbyname, in /usr/lib/gridengine, and that's reporting itself as 127.0.0.1 only.

20:00 - bblack restarts nscd, and this point qstat suddenly starts working again. Suspicion is that it was caching /etc/hosts entries in some form or other, and that somehow affected it working. Mysterious as to why it works this time.

20:10 - Intermittent qstat failures, but mostly working. bblack also finds that dnsmasq, the labs DNS server, returns SERVFAIL for both AAAA and MX records, so that's a possible avenue of exploration - but running dig in a loop fails to produce any issues while qstat is still ocassionally failing. It's still fairly ok - about 1 failure every 1000 qstat calls or so, so everyone calls it a night and goes to bed.

28th May 2015

Yuvi wakes up at indeterminate hour, qstat failure rate still at about 1 in 1000 qstat calls. 11:20 - qstat failures reported again - `error: commlib error: access denied (server host resolves rdata host "tools-bastion-01.eqiad.wmflabs" as "(HOST_NOT_RESOLVABLE)")`

11:25 - petan reboots tools-master again, back to original error of `error: unable to send message to qmaster using port 6444 on host "tools-master": got send error.` Other attempts are made, including turning off nscd (no effect) and entering an entry for tools-bastion-01 on /etc/hosts of tools-master (no effect)

12:55 - bblack figures out that gridengine's gethostbyname reads /etc/hosts and gives up right after, without even hitting DNS, and this might be because of our huge /etc/hosts file. This proves to be correct, as removing the huge lines brings the gridengine-master back online without intermittent failures.

Conclusions

The underlying cause was a huge /etc/hosts file, which is needed until https://phabricator.wikimedia.org/T63897 is fixed. The change that added the huge /etc/hosts file was merged about 8 days before the outage, so wasn't immediately obvious that was the cause. Secondary cause was that a package upgrade wasn't tested well enough but due to communication issues was assumed to have been tested well enough (including a restart of the deamon).

Long term fix is to get rid of GridEngine - it has no active upstream, and Debian doesn't consider is maintained enough to include it in Jessie. Short term fixes listed below.

Actionables

Explicit next steps to prevent this from happening again as much as possible, with Phabricator tasks linked for every step.

- Status: **Unresolved** Move labsdb aliases to DNS https://phabricator.wikimedia.org/T63897
- Status: **Unresolved** Re-test gridengine master / shadow failover https://phabricator.wikimedia.org/T90546

Category: Incident documentation

This page was last edited on 28 May 2015, at 16:46.

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. See Terms of Use for details.

Privacy policy About Wikitech Disclaimers Code of Conduct Developers Statistics Cookie statement Mobile view