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

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!

Incident documentation/20150925-LabsOutage

< Incident documentation

Contents [hide]

Summary

At around 11:00 UTC, labnet1002 suffered a kernel failure. This caused a complete labs network outage. A reboot of labnet1002 restored service; unfortunately the original issue failed to page so the outage was prolonged by 20 minutes because no one noticed that that the icinga alert was a critical failure.

Full network service was restored by 11:30; a bit of cleanup for other labs services (specifically, puppet on instances and instance deletion/creation) was sorted out a few hours later.

Timeline

- [11:03] Icinga reports PROBLEM - Host labnet1002 is DOWN: PING CRITICAL - Packet loss = 100%
- [11:04] Icinga reports PROBLEM - NFS read/writeable on labs instances on labstore1002 is CRITICAL: Connection timed out
- [11:12] The first labs users comment about problems on IRC
- [11:20] Mark and others notice the outage. Mark calls Andrew. Alexandros starts to investigate.
- [11:23] Alexandros captures some kernel stack traces (https://phabricator.wikimedia.org/P2092) and reboots.
- [11:29] labnet1002 is back up, most services are restored
- [11:30 onwards] Andrew sorts through some Labs cluster fallout from the outage, eventually restoring puppet and proper scheduler behavior.
- [13:40] All normal labs services are restored.

Conclusions

There were two major contributors to this outage: a kernel failure, and a monitoring failure.

Kernel: labnet1002 is running 3.13.0-59-generic. Moritz suspect that known issues in the -59 release are responsible for today's crash.

Paging: Ping tests do not page, so the original failure of labnet1002 would not have paged. Once a host is down, other icinga tests are suppressed, so no other paging was likely from labnet1002. The second alert (NFS read/writeable on labs instances) should have sent a page, but was misconfigured due to a mismatch between the string-literal 'true' in the monitoring base class and the boolean value true passed in.

Labnet1002 is a known single point of failure. Labnet1001 is available for fail-over; the fail-over process is manual, and probably no faster than a reboot. With that in mind, the recovery time between the problem being first noticed and resolution (less than 10 minutes) was about as good as could be hoped.

Actionables

Kernel

Labnet1002 should be upgraded away from the presumed buggy 3.13.0-59 kernel. The upgrade as already been done in place; all that remains is a reboot. Since a reboot or switch to labnet1001 will cause downtime, we're in a holding pattern. It may or may not be worthwhile to schedule an intentional reboot in the near future. labnet1001 is already running a believed-reliable kernel, 3.13.0-62.

Paging

The quoted-boolean mismatch is a serious hazard that has long been present in our puppet code. Quoted booleans should be removed wherever possible, and CI tests put in place to prevent any future appearance of same.

As of today the nova-network service is marked as critical, so it will page in the future. That doesn't help us if the whole box goes down, though, due to the ping test suppressing other host tests. We need a way to configure a particular host as page-worthy.

Category: Incident documentation