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/20150331-LabsNFS-Filesystem-Switch

< Incident documentation

Contents [hide]

Summary

The planned switch of the filesystem underlying the Labs NFS service unexpectedly caused instances running Ubuntu Precise to require rebooting rather than a simple filesystem remount as expected. As a consequence, a large number of Labs projects (including Tool Labs) were negatively impacted during diagnosis and restart of those instances. In addition, the large number of instances being restarted caused the Labs virtualization infrastructure to overload, slowing recovery down.

Normal operation returned around 22h UTC for Tool Labs, and Labs generally around 00h UTC, with minor issues remaining consistent with a Labs-wide outage that were gradually fixed as they were reported. The rsync did cause another outage a few hours later, however.

Timeline

```
21:00 turned NFS off
21:01 rotated the filesystems between the old (flat) one and the new (thin) one
21:03 turned NFS back on
         At that point, instances were expected to no longer be able to access
files
21:05 unmount and remount the NFS filesystems on the instances of the tools
project (which was first in line)
         At that point, Trusty instances reacted as planned (quickly recovered,
with filesystem available) but Precise instances ended up being unable to
detatch, misoperating on the former filesystem
21:12 Diagnosing the issue took place, with bastion-restricted-01 as the guinea
pig (still running Precise)
21:25 Confirmed that unmounting the previous filesystem was broken (rather than
unmount, the mountpoint was converted to a _(deleted) faux-file that cannot be
operated upon
21:28 Attempted a reboot of bastion-restricted-01 to see if the bootstrap mounts
would work - they did.
21:36 Confirmed that precise instances could be fixed by a reboot by doing so on
tools-master.  Also confirmed.
21:41 Proceed to reboot all of tool's precise instances in an order intended to
speed recovery as much as possible
21:45 Yuvi proceeds to reboot affected instances of deployment-prep
21:56 Confirmation that tool labs is recovering
22:01 Restarted most jobs as gridengine recovers, web services mostly back
online
22:10 Some reports come in of older versions of file being visible; outage seems
to have caused the loss of the most recent sync
22:11 Yuvi proceeds to use salt to remount filesystem on Trusty and Jessie
instances.  Salt succeeds but fails to report success
22:23 Andrew generates a list of Precise instances to reboot, which are then
rebooted at 5s interval
22:30 Tremendous load on virtualization hosts as instances reboot (sometimes
many to a host) cause regular hangs of groups of instances, but manages to push
through
22:53 rsync started to recover the last changes from the old filesystem
23:30 Andrew halts a few especially greedy instances to ease CPU load on
virt1004 and virt1011
23:35 Labs stabilizes
00:09 After some minor point fixes, we declare Labs to be back up
13:42 run rsync one last time to ensure no out of date files
15:20 final filesystem rsync confirms no out of date files
```

Conclusions

The outage to Precise instances seems to have been unavoidable, but if that configuration had been tested we might have been to plan for it and schedule a proper maintenance window for the restarts as opposed to having to recover from an unplanned outage.

Actionables

- Create a method to schedule restart of instances in a way that cannot overload virtualization hosts by staggering them https://phabricator.wikimedia.org/T94613
- Create 'checklists' for all planned maintenance that we should be followed. https://phabricator.wikimedia.org/T94608
- Make certain to include all extant OS flavors when testing a planned change to the infrastructure. While all new installs are Trusty and Jessie, there remains a number of other releases that may react differently in the fleet. (Should be part of checklist)
- Schedule a lot more time for any planned maintenance on NFS, no matter how trivial (should be part of checklist)

Category: Incident documentation