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/20161204-labservices1001

< Incident documentation

Summary

At about 02:20 on 2016-12-04, labservices1001 become unresponsive and unreachable. It was rebooted and returned to normal service about 20 minutes later.

Any new VMs created during the outage window (including those used for CI) would have failed to register DNS records -- as it happened, no VMs appear to have been created during the window, so there were few user-facing consequences from this outage. Labservices1001 is also a DNS server, but most DNS traffic fell to the secondary server, labservices1002, which successfully handled it.

The cause of the system crash is yet undetermined. Additionally, there were several monitoring and paging issues revealed by this incident that require investigation.

Timeline

- 02:17 Labservices1001 locks up
- 02:19 Page: 'toolshecker service itself needs to return OK on checker.tools.wmflabs.org os CRITICAL: HTTP CRITICAL: HTTP1.1 502 Bad Gateway'
- 02:21 Given that the toolschecker web service is itself down, a series of other related pages are sent over the next few minutes.
- 02:21 Some CI operations begin to fail due to DNS resolution issues:

```
 02:21:30 git.exc.GitCommandError: 'git remote prune --dry-run origin' returned
with exit code 128
 02:21:30 stderr: 'fatal: unable to access
'https://gerrit.wikimedia.org/r/p/mediawiki/core/'🔗: Could not resolve host:
gerrit.wikimedia.org'
```

- 02:22 Andrew and Alex Monk confer on IRC but both have limited access for troubleshooting (Andrew didn't have his key and Alex didn't have the necessary privileges).
- 02:25 Filippo arrives on IRC (with full access) and begins investigation.
- 02:35 toolschecker starts sending recovery alerts. Presumably it has figured out to switch over to the secondary DNS server, but labservices1001 (aka labs-ns0) is still down.
- 02:40 Filippo reboots labservices1001 from the management console.
- 02:43 Labservices1001 is back up, all services returned to normal.

Conclusions

The actual failure of labservices1001 is unexplained, and may remain so if it does not recur. For the most part, the redundancy in Labs DNS served us well. There are nonetheless a few pressing concerns:

- Why did the labservices1001 failure not page?
- Why was toolschecker so slow to cope with the loss of labs-ns0?
- Why did CI tests not fail over to labs-ns1 gracefully?

Actionables

- Explain the cause of the hardware failure bug T152340
- ✓ **Done** Fix paging for designate services and DNS on labservices1001 bug T152368
- Investigate DNS failover for toolschecker bug T152369
- Add monitoring for a full Labs instance lifecycle bug T123590
- ✓ **Done** Make sure CI boxes know about both DNS servers bug T137460

Category: Incident documentation

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

WIKIMEDIA
a project

Powered By
MediaWiki