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/20150613-Esams-Text-80

< Incident documentation

Contents [hide]

Summary

The text (primary wiki traffic) service in esams (where most countries in or close to the EU are mapped) failed for all HTTP (but not HTTPS) protocol requests for 19 minutes, from approximately 18:54 to 19:13 UTC on Saturday June 13th, 2015. This was a tertiary casualty of a technical mistake during with the HTTPS-by-default transition of enwiki on Friday morning. We were at risk of this ever since, but it wasn't until Saturday that a catalytic event triggered the fallout.

The original mistake is that the LVS healthcheck for HTTP service on port 80 checks the URL http://en.wikipedia.org/wiki/Main_Page and we should have realized/anticipated this would be a problem and addressed it prior to the enwiki transition. Those checks began failing as soon as that URL began returning a redirect to HTTPS instead of a normal "200 OK" response. This failure affected all backends for the service, but pybal has a depool_threshold parameter of 0.5, which means after the first half have been depooled, it will refuse to depool any further backends regardless of healthcheck state. This condition of pybal having half of its backends for a service depooled, and all failing healthchecks, didn't trigger any icinga alerts. Service continued to be healthy from an external perspective under these conditions.

Later, on Saturday, for unrelated reasons, I depooled a text server in esams using the normal mechanisms (editing the config-master pybal server list on palladium). When the server list data changed, pybal fetched a fresh server list. Apparently when pybal applies a server list change like this at runtime, the depool_threshold no longer applies for the first checks on the new list. So when they all failed their first post-change healthcheck, all servers were marked depooled and removed from the runtime IPVS configuration, causing total service outage.

As a temporary workaround, which is still in effect, I changed all enwiki references in pybal production healthchecks to equivalent dewiki references, since dewiki is not yet redirecting. This obviously needs a better fix before dewiki begins redirecting as well...

Timeline

- [18:53] bblack depooled one esams text cache machine via pybal config (normally, fairly routine)
- [18:54] The moment the problem is obvious in graph history data
- [18:58] Icinga alerted IRC and paged several ops, but notably SMS notifications did not go out to several EU employees
- [19:06] After some digging, bblack discovers cause, begins fix attempts
- [19:13] Service restored

Cause Analysis

1. Human error in lead-up to enwiki HTTPS conversion: failure to recognize that this would break HTTP/port-80 healthchecks for the primary text services in LVS.
2. Lack of monitoring for pybal service states: we monitor externally-visible health, and that pybal is running, but we don't have any monitoring for when a pybal has X/Y backends depooled and/or failing checks, which would have been screaming at us since Friday morning in this case, even though the external view was healthy.
3. Failure to notice port-80 traffic imbalances in graphs. We were staring at a lot of graphs in the aftermath

of the enwiki transition, but apparently never at the right ones to notice the obvious port-80 traffic imbalances per-cluster-machine from half the cluster being depooled for just that port.

4. Additionally, random git-gc invocation with long run-time slowed a puppet-merge while trying to quickly but properly deploy a critical fix. This was worked around, but cost time. I'm not sure if this is realistically actionable in any way.

Actionables

- Fix pybal monitoring of port 80 text/mobile services in light of HTTPS redirects
- Implement pybal pool state monitoring and alerting via icinga
- Investigate smsglobal delivery failures from 2015-06-13 weekend
- icinga log rotation wipes out portions of history

Category: Incident documentation

This page was last edited on 14 June 2015, at 10:06.