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/20180110-swift

< Incident documentation

Summary

Swift suffered a brief unavailability period in eqiad during roll restarts for kernel upgrades.

Timeline

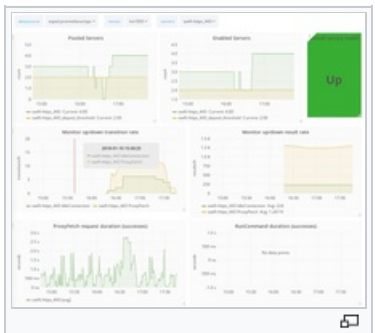
- 20180109 - Filippo is upgrading kernel on swift fleet, swift eqiad frontends is roll-restarted without incident, though the kernel package wasn't upgraded on those machines, requiring another roll-restart. During this operation ms-fe1008 is inadvertently not repooled as it should be. 3/4 machines are serving traffic.
- 20180110T16:11 - Roll restarts for ms-fe1* resumes, ms-fe1005 is depooled. 2/4 machines are serving traffic.
- 20180110T16:18 - ms-fe1005 repooled. 3/4 machines are serving traffic.
- 20180110T16:21 - ms-fe1006 depooled. 2/4 machines are serving traffic.
- 20180110T16:22 - ms-fe1007 cannot cope with the load. It is marked as down by PyBal and depooled. 1/4 machines are serving traffic.
- 20180110T16:29 - thumbor.svc.eqiad pages
- 20180110T16:29 - thumbor.svc.eqiad recovers
- 20180110T16:34 - ms-fe1005 also goes down and is depooled by PyBal. 0/4 machines are serving traffic.
- 20180110T16:36 - ms-fe.svc.eqiad pages
- 20180110T16:38 - ms-fe1008 repooled
- 20180110T16:39 - ms-fe1006 repooled
- 20180110T16:39 - ms-fe.svc.eqiad recovers

Conclusions

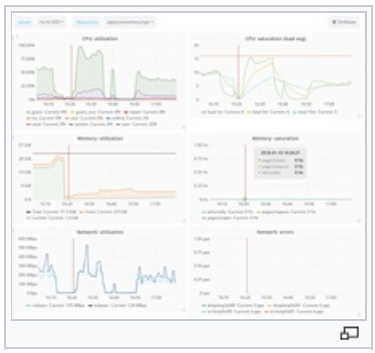
Swift frontends safety margin is two machines out of four, though this margin was violated due to a combination of factors: namely one less machine in the pool than assumed and too fast traffic swings between the remaining machines in service. Further, PyBal does not take into account administratively depooled servers T184715 when checking whether a host can be depooled or not. The traffic swings ended up overloading ms-fe1007, the only machine fully in service at the time, and subsequently drop of traffic from all frontends in service at the time (ms-fe1005, ms-fe1007).

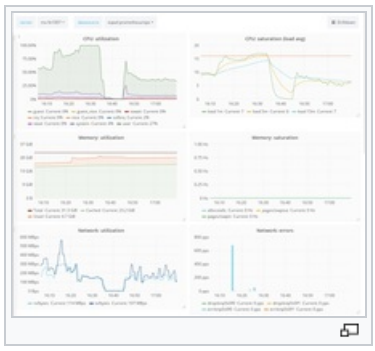
Actionables

Mostly operator error and failure to verify preconditions before starting procedures on high traffic/critical services.



- Verify preconditions before start of operations (e.g. service pool is healthy)
- Patch PyBal to properly enforce depool-threshold phab:T184715





Category: Incident documentation