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/20190604-blubberoid

< Incident documentation

Contents [hide]

Summary

Blubberoid in codfw paged and was unable to serve the exactly 0 req/s that it receives (that is the service isn't really used). The service should not be paging anyway. However the underlying issue was more important and was related to recovering actions caused by an operator error

Impact

- Nothing, noone.

Detection

Pages for blubberoid arrived, but Alex was aware way before that. The pages were actually the result of alex becoming a bit too confident in his recovering actions.

Timeline

No timeline really

What actually happened

Alex was investigating the deployment of the sessionstore service and some latencies in checks that presented themselves only on kubernetes1004. He realized that the node routing tables of the new nodes were not correct. In fact one of the nodes(kubernetes1006) was advertising pod IPs that belonged to a different node (kubernetes1004) causing a partial backholing of traffic from kubernetes1004 to pods on kubernetes1006. A drain fixed the immediate issue and debugging ensued. Soon it become apparent that calico was assigned a /24 that had split up in 4 /26 networks (the smallest network a calico node seems to be able to address in our version), leaving the 2 new nodes in both DCs without an assigned IP block. A typo during the insertion of a new IPpool block in calico (adding 10.64.64.1/24 instead of 10.64.65.0/24) caused overlapping pools to exist in calico. That did not cause an issue until Alex tried to delete the overlapping IPpool. Seems like calicoctl decided to apply the mask to the IP and instead deleted the original 10.64.64.0/24 supplied network. Nodes very quickly figured that out and abandoned their reservations. However the bird component on them kept on advertising the /32 IP subnets (IPs actually) of each pod, allowing everything to proceed ok. Actions to remove the erroneous IPpool failed at the calico level and the pool was removed using etcd tools. The pools were correctly entered again, but it became evident that the nodes would not reallocate the now orphaned assignments to themselves. The assignments were again manually deleted from etcd and the nodes created finally new assignments, albeit ones that had nothing to do with their old ones. That however did not cause an issue as they kept on advertising the /32s next to their new (and unrelated) /26s. To clear up the state of things, Alex started a rolling restart of all pods in both clusters by draining in sequence all nodes and then uncordoning them. In eqiad he was rather timid about

this, in codfw he actually automated it, causing some increased latencies for kubelet and pod starts. Pods for all services are assigned in multiples of 4 in order to maximize availability so no issue was met as traffic continued flowing normally. However blubberoid is an exception to the above rule and only had 1 instance. Restarting that pod in a different node in tandem with many other pods ended up being slow enough for icinga to catch this and page. Aside from that, no other request was lost.

Conclusions

~~*Blubberoid should not page*~~

We need to upgrade calico ASAP

~~*We need to make sure we have enough pods for all services to survive availability zone (rack row effectively) outages*~~

What went well?

- Kubernetes rules
- BGP rules even more.
- Alex figured out the issue after looking into etcd

What went poorly?

- *Adding a new overlapping IP pool into calico was possible, removing it was not.*

Where did we get lucky?

- We have minimal intra k8s service requests currently.

Links to relevant documentation

- None for this, the actual incident had nothing to do with the alert

Actionables

- Upgrade calico phab:T207804

Category: Incident documentation