

hswaw - Bugless #52

k0/cluster: Deploy pmtud

09/20/2021 10:48 AM - q3k

Status:	New	
Priority:	High	
Assignee:		
Category:	hsccloud	
Description We are running ECMP for our LoadBalancer Services, which causes us to fuck up whenever we can't serve traffic at more than 1500 byte ethernet frames (ie. because of a <1500 MTU segment on a path to the client). This is a well-known problem with ECMP setups, and there's even a soultion: https://github.com/cloudflare/pmtud		

History

#1 - 09/20/2021 10:50 AM - q3k

Note: deploying pmtud on each machine means that they all need to be in the same broadcast domain. This is currently the case, but might not be if we run of out IP addresses in the L2 segment currently used for k0 machines (and some other machines). Another problem is that we'd also be re-broadcasting that ICMP to machines that aren't even part of k0, just happen to live in that L2 segment.

Alternatively, we could hack pmtud to instead have a target list of machines to broadcast to, possibly over L3.

#2 - 09/20/2021 10:52 AM - q3k

Alternatively alternatively: NIH the fucker and make our own PMTUD-like thing in Rust, as it seems like there isn't much code there (mostly stuff to make C usable, like hash functions and rate limiters): <https://github.com/cloudflare/pmtud/blob/master/src/main.c#L104>

#3 - 09/21/2021 01:51 PM - q3k

Related: it would also be nice to handle ICMP Echo Requests to respond to pings for virtual IPs. Currently traceroutes show repeated bounces between the machine handling the VIP and the ToR switch for that machine (because there's no iptables rule inserted by kube-proxy/metallb(?) to handle ICMP).

#4 - 07/04/2022 01:08 PM - q3k

- Category set to hsccloud