Linux kernel client - Feature #49581

IP dual stack support for cephfs linux kernel client

03/03/2021 02:08 PM - Stefan Kooman

Status:	New	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:	Ilya Dryomov		
Category:	libceph		
Target version:			
Source:	Community (user)	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:			

Description

The linux 5.11 kernel cephfs client can not deal with an IP dual stack cluster. It fails to decode the OSDmap and the MDSmap:

Mar 2 09:01:14 kernel: Key type ceph registered
Mar 2 09:01:14 kernel: libceph: loaded (mon/osd proto 15/24)
Mar 2 09:01:14 kernel: FS-Cache: Netfs 'ceph' registered for caching
Mar 2 09:01:14 kernel: ceph: loaded (mds proto 32)
Mar 2 09:01:14 kernel: libceph: mon4 (1)[mond addr]:6789 session established
Mar 2 09:01:14 kernel: libceph: another match of type 1 in addrvec
Mar 2 09:01:14 kernel: ceph: corrupt mdsmap
Mar 2 09:01:14 kernel: ceph: error decoding mdsmap -22
Mar 2 09:01:14 kernel: libceph: another match of type 1 in addrvec
Mar 2 09:01:14 kernel: libceph: corrupt full osdmap (-22) epoch 98764 off 6357 (0000000027a57a75 of 00000000d3075952-00000000e307797f)
Mar 2 09:02:15 kernel: ceph: No mds server is up or the cluster is laggy

As both an IPv4 and IPv6 address are encoded for the OSDs and MDSs, so there are multiple v1 and v2 addresses to choose from. Example for MDS (ceph fs dump):

[mds.mds1{0:229930080} state up:active seq 144042 addr [v2:[2001:7b8:80:1:0:1:3:1]:6800/2234186180,v1:[2001:7b8:80:1:0:1:3:1]:6801/2234186180,v2:0.0.0.0:6802/2234186180,v1:0.0.0.0:6803/2234186180]]

Note that in this case *no* IPv4 has been explicitly configured, but nevertheless '0.0.0.0' gets configured when the cluster is running with ms_bind_ipv4=true (default).

According to https://docs.ceph.com/en/latest/rados/configuration/network-config-ref/#ipv4-ipv6-dual-stack-mode a future Ceph release will support a IP dual stack cluster. Even with a separate public and cluster network. The kernel client needs to be able to deal with situations like this. I'm not sure if there is a certain IP stack to be preferred, but I guess that would be the easiest way to deal with situations like this (try IPv6 first, IPv4 second). One could argue that it should be possible to set a flag somewhere which IP stack has to be preferred. I don't want to get into details of how this should or could be done, Ceph developers will know that best. Maybe a discussion about details like this have already been held and I just don't know about them. Hack that might already be implemented everywhere else except for the kernel client.

More context in this thread: https://www.spinics.net/lists/ceph-users/msg64928.html

History

#1 - 03/03/2021 02:40 PM - Ilya Dryomov

- Project changed from Ceph to Linux kernel client

- Category set to libceph

- Tags deleted (cephfs kernel)

#2 - 03/03/2021 02:41 PM - Ilya Dryomov

- Assignee set to Ilya Dryomov