Ceph - Bug #22944

Infiniband send_msg send returned error 32: (32) Broken pipe

02/07/2018 10:59 AM - Radosław Piliszek

Status:	New	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:	Haomai Wang		
Category:			
Target version:			
Source:	Community (user)	**Affected Versions:**	v12.2.2
Tags:	rdma, infiniband	**ceph-qa-suite:**	
Backport:		**Pull request ID:**	
Regression:	No	**Crash signature (v1):**	
Severity:	2 - major	**Crash signature (v2):**	
Reviewed:			

Description

Using CentOS 7.4, Mellanox OFED 4.2 on Connect-X 3 in Infiniband mode and Ceph 12.2.2 compiled with RDMA.

I've set:

```
ms type = async+rdma
```

I've fixed systemd units to allow RDMA device usage.

But I cannot get RDMA to work. Connections time out.

I sometimes get:

```
Infiniband send_msg send returned error 32: (32) Broken pipe
```

in monitor service journal.

RDMA by itself works just fine (verified with rping).

Ceph by itself also works just fine (verified without RDMA).

This happens even on one-node cluster when trying to access it from the very same node.

Please let me know how I can get you more info to debug it.

History

#1 - 02/12/2018 07:13 PM - Greg Farnum

The RDMA support in AsyncMessenger is experimental and I think the guys building it are planning to rip it apart. I would just use normal ethernet.

#2 - 02/12/2018 07:13 PM - Greg Farnum

- Assignee set to Haomai Wang

Haomai, can you follow up if this interests you, or else close it? :)

#3 - 02/13/2018 02:10 AM - Haomai Wang

hmm, I don't think the provided log is the reason. maybe you can set debug_ms=20/20 to output more?

#4 - 02/13/2018 08:21 AM - Radosław Piliszek

- File ceph-client.smp-016.log added

- File ceph-mon.smp-016.log added

- File ibdump.smp-016.pcap added

Hi Greg, Hi Haomai,

I increased debug to 20. Please find mon and client logs attached.

I also ran ibdump to discover that packets are malformed. There is no LID 4 in the fabric (used as DLID), SGID is invalid and DGID looks like repeated random sequence (also invalid). DQPN is invalid as well (it does not agree with logs). Also SL/TC look random to me but this they are valid and irrelevant when considering same node. I attach the capture. Recent Wireshark opens it just fine.

DGID seems to change from run to run. DLID, SGID, TC, SL do not change ever. DQPN and PSN seem to change between retries but DQPN incorrectly and PSN effectively unnecessarily (because QP is changed anyway).

Correct DLID would be equal to SLID (79).

As for the logs. Client gets error CQE and retries again and again. This is understandable from capture entries - packets are simply unroutable. This looks like addresses are populated incorrectly.

I saw articles mentioning people running Ceph on RDMA so I thought it was in a working state (maybe with some bugs to crush) but this looks like either the build I use (12.2.2) is broken or my environment is unsupported (for some unknown reason because it is rather standard stuff).

#5 - 02/13/2018 01:54 PM - Haomai Wang

actually I always tested with connect3-x with ib mode.

from log, it seemed client have a good handshake with mon. but when client post write request, monitor doesn't receive any message. so client wait timeout then close connection.

what's your ceph.conf?

#6 - 02/13/2018 04:39 PM - Radosław Piliszek

At the moment the ceph.conf is:

```
[global]
mon host = 192.168.4.96
fsid = 6de5466c-d11f-4e59-857b-b11cb0bc4d9b
public network = 192.168.4.0/24
cephx require signatures = true
ms type = async+rdma
ms_async_rdma_device_name = mlx4_0
```

```
ms_async_rdma_polling_us = 0
debug ms = 20/20

[mon]
mon allow pool delete = true
osd pool default size = 2
osd pool default min size = 2
```

Also:

```
[root@smp-016 ~]# ibdev2netdev
mlx4_0 port 1 ==> ib0 (Up)
[root@smp-016 ~]# ip -4 a s ib0
4: ib0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 2044 qdisc mq state UP qlen 1024
    inet 192.168.4.96/24 brd 192.168.4.255 scope global ib0
       valid_lft forever preferred_lft forever
```

#7 - 04/16/2018 01:56 PM - Jay Munsterman

Just adding to the conversation: We appear to be experiencing the same thing here with the same configuration...

Files

ceph-client.smp-016.log	46.4 KB	02/13/2018	Radosław Piliszek
ceph-mon.smp-016.log	54 KB	02/13/2018	Radosław Piliszek
ibdump.smp-016.pcap	3.46 KB	02/13/2018	Radosław Piliszek