fs - Feature #19135

Multi-Mds: One dead, all dead; what is Robustness??

03/03/2017 02:57 AM - xianglong wang

Status:	Rejected	**Start date:**	03/03/2017
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:		**Affected Versions:**	
Tags:		**Release:**	
Backport:		**Component(FS):**	
Reviewed:		**Needs Doc:**	No
User Impact:			

Description

I have three mds, when I copy file to fuse-type mountpoint for test.
when I shutdown one mds, the client hang, the mds log blew. I want to know wether hava any protocol like raft, paxos to keep cluster more health? or some configure policy?

1. ceph -v
   ceph version 10.2.5 (c461ee19ecbc0c5c330aca20f7392c9a00730367)

---
mds logs:

```
2017-03-03 10:35:28.617671 7f42ecf15700  0 log_channel(cluster) log [WRN] : slow request 241.28685
1 seconds old, received at 2017-03-03 10:31:27.330669: client_request(client.74105:98380 lookup #1
/bin 2017-03-03 10:31:27.331487) currently failed to rdlock, waiting
2017-03-03 10:39:28.670117 7f42ecf15700  0 log_channel(cluster) log [WRN] : 2 slow requests, 2 inc
luded below; oldest blocked for > 481.345120 secs
2017-03-03 10:39:28.670129 7f42ecf15700  0 log_channel(cluster) log [WRN] : slow request 481.34512
0 seconds old, received at 2017-03-03 10:31:27.324945: rejoin:client.74105:98379 currently initiat
ed
2017-03-03 10:39:28.670134 7f42ecf15700  0 log_channel(cluster) log [WRN] : slow request 481.33939
6 seconds old, received at 2017-03-03 10:31:27.330669: client_request(client.74105:98380 lookup #1
/bin 2017-03-03 10:31:27.331487) currently failed to rdlock, waiting
2017-03-03 10:46:28.632377 7f42e8907700  0 -- 192.168.88.136:6801/3074 >> 192.168.88.134:6802/3869
 pipe(0x7f4300ceb400 sd=20 :57180 s=2 pgs=5 cs=1 l=0 c=0x7f43048f4f00).fault, initiating reconnect
2017-03-03 10:46:31.643232 7f42e8705700  0 -- 192.168.88.136:6801/3074 >> 192.168.88.134:6802/3869
 pipe(0x7f4300ceb400 sd=20 :57180 s=1 pgs=5 cs=2 l=0 c=0x7f43048f4f00).fault
2017-03-03 10:47:28.800833 7f42ecf15700  0 log_channel(cluster) log [WRN] : 2 slow requests, 2 inc
luded below; oldest blocked for > 961.475741 secs
2017-03-03 10:47:28.800847 7f42ecf15700  0 log_channel(cluster) log [WRN] : slow request 961.47574
1 seconds old, received at 2017-03-03 10:31:27.324945: rejoin:client.74105:98379 currently initiat
ed
2017-03-03 10:47:28.800854 7f42ecf15700  0 log_channel(cluster) log [WRN] : slow request 961.47001
7 seconds old, received at 2017-03-03 10:31:27.330669: client_request(client.74105:98380 lookup #1
/bin 2017-03-03 10:31:27.331487) currently failed to rdlock, waiting
```

History

#1 - 03/03/2017 03:46 AM - Zheng Yan

- Project changed from Ceph to fs

#2 - 03/03/2017 06:59 AM - Zheng Yan

multiple active mds is mainly for improving performance (balance load to multiple mds). robustness is achieved standby mds

#3 - 03/03/2017 07:54 AM - xianglong wang

Zheng Yan wrote:

> multiple active mds is mainly for improving performance (balance load to multiple mds). robustness is achieved standby mds

Is it possible for both high performance and availability?
have any solution? please...

#4 - 03/03/2017 09:09 AM - John Spray

- Status changed from New to Rejected

Having multiple active MDS daemons does not remove the need for standby daemons. Set max_mds to something less than your actual number of physical MDSs and the remaining daemons will automatically be used as standbys.

#5 - 03/06/2017 04:21 AM - xianglong wang

John Spray wrote:

> Having multiple active MDS daemons does not remove the need for standby daemons. Set max_mds to something less than your actual number of physical MDSs and the remaining daemons will automatically be used as standbys.

I test but the result is not changed! three mds && set max_mds=2 by "ceph mds set_max_mds 2".
is this right

```
[root@admin ceph]# ceph -s
cluster 698db7be-5c6e-4bb3-873f-6ce17c3fad5b
health HEALTH_OK
monmap e1: 3 mons at {ceph1=192.168.88.131:6789/0,ceph2=192.168.88.132:6789/0,ceph3=192.168.88.133:6789/0}
election epoch 92, quorum 0,1,2 ceph1,ceph2,ceph3
fsmap e362: 3/3/2 up {0=md2=up:active,1=md1=up:active,2=md3=up:active}
osdmap e268: 4 osds: 3 up, 3 in
flags sortbitwise,require_jewel_osds
pgmap v2645: 320 pgs, 3 pools, 229 MB data, 1977 objects
607 MB used, 20863 MB / 21470 MB avail
320 active+clean

----------After-Down-A-Mds----------

[root@admin ceph]# ceph -s
cluster 698db7be-5c6e-4bb3-873f-6ce17c3fad5b
health HEALTH_ERR
mds rank 1 has failed
mds cluster is degraded
monmap e1: 3 mons at {ceph1=192.168.88.131:6789/0,ceph2=192.168.88.132:6789/0,ceph3=192.168.88.133:6789/0}
election epoch 92, quorum 0,1,2 ceph1,ceph2,ceph3
fsmap e363: 2/3/2 up {0=md2=up:active,2=md3=up:active}, 1 failed
osdmap e269: 4 osds: 3 up, 3 in
flags sortbitwise,require_jewel_osds
pgmap v2648: 320 pgs, 3 pools, 229 MB data, 1977 objects
607 MB used, 20863 MB / 21470 MB avail
320 active+clean
client io 131 B/s rd, 0 op/s rd, 0 op/s wr
```

#6 - 03/06/2017 01:06 PM - John Spray

After decreasing max_mds, you also need to use "ceph mds deactivate <rank>" to shrink the active cluster, so that one of the daemons will go back to being a standby. Note that this isn't necessarily stable in the jewel code you're using, so if this is just an experimental cluster you could also try using the latest development packages (https://ceph.com/releases/v12-0-0-luminous-dev-released/)

If you have further questions about usage/behaviour please ask on the ceph-users mailing list (feel free to open another tracker ticket if you find a software bug)

#7 - 03/08/2017 08:21 AM - xianglong wang

John Spray wrote:

> After decreasing max_mds, you also need to use "ceph mds deactivate <rank>" to shrink the active cluster, so that one of the daemons will go back to being a standby. Note that this isn't necessarily stable in the jewel code you're using, so if this is just an experimental cluster you could also try using the latest development packages (https://ceph.com/releases/v12-0-0-luminous-dev-released/)
>
> If you have further questions about usage/behaviour please ask on the ceph-users mailing list (feel free to open another tracker ticket if you find a software bug)

OK, I get it. THX!