mgr - Bug #43037

mgr: "mds metadata" to setup new DaemonState races with fsmap

11/26/2019 08:17 PM - Patrick Donnelly

Status:	Resolved	**% Done:**	0%
Priority:	Urgent		
Assignee:	Patrick Donnelly		
Category:			
Target version:	v15.0.0		
Source:	Q/A	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:	nautilus	**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	31899
Severity:	3 - minor	**Crash signature:**	

Description

Symptom: ceph config show mds.b is empty:

```
2019-11-26T13:08:23.402 INFO:tasks.cephfs_test_runner:test_ceph_config_show (tasks.cephfs.test_adm
in.TestConfigCommands) ... FAIL
```

From:
/ceph/teuthology-archive/pdonnell-2019-11-26_04:58:35-fs-wip-pdonnell-testing-20191126.005014-distro-basic-smithi/4543982/teuthology.log

During testing, the daemon is removed from daemon_state after the mons fail the mds:

```
2019-11-26T13:08:05.283+0000 7fb546808700  1 -- 172.21.15.79:0/35406 <== mon.2 v2:172.21.15.79:330
1/0 210 ==== fsmap(e 16) v1 ==== 877+0+0 (crc 0 0 0) 0x557c4b786c00 con 0x557c4b24a400
2019-11-26T13:08:05.283+0000 7fb546808700 10 mgr ms_dispatch2 active fsmap(e 16) v1
2019-11-26T13:08:05.283+0000 7fb546808700 10 mgr ms_dispatch2 fsmap(e 16) v1
2019-11-26T13:08:05.283+0000 7fb546808700 10 mgr notify_all notify_all: notify_all fs_map
2019-11-26T13:08:05.283+0000 7fb546808700 15 mgr notify_all queuing notify to balancer
2019-11-26T13:08:05.283+0000 7fb546808700 15 mgr notify_all queuing notify to crash
2019-11-26T13:08:05.283+0000 7fb546808700 15 mgr notify_all queuing notify to devicehealth
2019-11-26T13:08:05.283+0000 7fb546808700 15 mgr notify_all queuing notify to iostat
2019-11-26T13:08:05.283+0000 7fb546808700 15 mgr notify_all queuing notify to orchestrator_cli
2019-11-26T13:08:05.283+0000 7fb546808700 15 mgr notify_all queuing notify to pg_autoscaler
2019-11-26T13:08:05.283+0000 7fb546808700 15 mgr notify_all queuing notify to progress
2019-11-26T13:08:05.283+0000 7fb546808700 15 mgr notify_all queuing notify to rbd_support
2019-11-26T13:08:05.283+0000 7fb546808700 15 mgr notify_all queuing notify to restful
2019-11-26T13:08:05.283+0000 7fb546808700 15 mgr notify_all queuing notify to status
2019-11-26T13:08:05.283+0000 7fb546808700 15 mgr notify_all queuing notify to telemetry
2019-11-26T13:08:05.283+0000 7fb546808700 15 mgr notify_all queuing notify to volumes
2019-11-26T13:08:05.283+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b2f5970
2019-11-26T13:08:05.283+0000 7fb546808700  4 mgr cull Removing data for a
2019-11-26T13:08:05.283+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b2f5970
2019-11-26T13:08:05.283+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b2f5970
2019-11-26T13:08:05.283+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b2f5970
2019-11-26T13:08:05.283+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b2f5970
2019-11-26T13:08:05.283+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b2f5970
2019-11-26T13:08:05.283+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b2f5970
2019-11-26T13:08:05.283+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b2f5970
2019-11-26T13:08:05.283+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b2f5970
2019-11-26T13:08:05.283+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b2f5970
2019-11-26T13:08:05.283+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b2f5970
2019-11-26T13:08:05.283+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b2f5970
```

```
2019-11-26T13:08:05.283+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b2f5970
2019-11-26T13:08:05.284+0000 7fb546808700  4 mgr cull Removing data for b
2019-11-26T13:08:05.284+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b2f5970
2019-11-26T13:08:05.284+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b2f5970
2019-11-26T13:08:05.284+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b2f5970
2019-11-26T13:08:05.284+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b2f5970
2019-11-26T13:08:05.284+0000 7fb52e881700 20 mgr[restful] Unhandled notification type 'fs_map'
2019-11-26T13:08:05.284+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b2f5970
2019-11-26T13:08:05.284+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b2f5970
2019-11-26T13:08:05.284+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b2f5970
2019-11-26T13:08:05.284+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b2f5970
2019-11-26T13:08:05.284+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b2f5970
2019-11-26T13:08:05.284+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b2f5970
2019-11-26T13:08:05.284+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b2f5970
...
```

From:
/ceph/teuthology-archive/pdonnell-2019-11-26_04:58:35-fs-wip-pdonnell-testing-20191126.005014-distro-basic-smithi/4543982/remote/smithi079/log/ceph-mgr.x.log.gz

MDS becomes standby in e21 and opens a connection with the mgr:

```
2019-11-26T13:08:17.685+0000 7f30c5187700  1 mds.b Updating MDS map to version 21 from mon.1
2019-11-26T13:08:17.685+0000 7f30c5187700 10 mds.b      my compat compat={},rocompat={},incompat={1=base v0.20,2=client writeable ranges,3=default file layouts on dirs,4=dir inode in separate object,5=mds uses versioned encoding,6=dirfrag is stored in omap,7=mds uses inline data,8=no anchor table,9=file layout v2,10=snaprealm v2}
2019-11-26T13:08:17.685+0000 7f30c5187700 10 mds.b  mdsmap compat compat={},rocompat={},incompat={1=base v0.20,2=client writeable ranges,3=default file layouts on dirs,4=dir inode in separate object,5=mds uses versioned encoding,6=dirfrag is stored in omap,8=no anchor table,9=file layout v2,10=snaprealm v2}
2019-11-26T13:08:17.685+0000 7f30c5187700 10 mds.b my gid is 5284
2019-11-26T13:08:17.685+0000 7f30c5187700 10 mds.b map says I am mds.-1.0 state up:standby
2019-11-26T13:08:17.685+0000 7f30c5187700 10 mds.b msgr says I am [v2:172.21.15.79:6834/3636622064,v1:172.21.15.79:6835/3636622064]
2019-11-26T13:08:17.685+0000 7f30c5187700  1 -- [v2:172.21.15.79:6834/3636622064,v1:172.21.15.79:6835/3636622064] --> [v2:172.21.15.79:3300/0,v1:172.21.15.79:6789/0] -- mon_subscribe({mgrmap=0+}) v3 -- 0x55f4168b6200 con 0x55f415c49800
2019-11-26T13:08:17.685+0000 7f30c5187700 10 mds.b handle_mds_map: handling map in rankless mode
2019-11-26T13:08:17.685+0000 7f30c5187700  1 mds.b Monitors have assigned me to become a standby.
2019-11-26T13:08:17.685+0000 7f30c5187700  5 mds.beacon.b set_want_state: up:boot -> up:standby
2019-11-26T13:08:17.686+0000 7f30c818d700  1 -- [v2:172.21.15.79:6834/3636622064,v1:172.21.15.79:6835/3636622064] <== mon.1 v2:172.21.15.79:3300/0 7 ==== mdsbeacon(5284/b up:boot seq 1 v21) v7 ==== 130+0+0 (crc 0 0 0) 0x55f4168a4600 con 0x55f415c49800
2019-11-26T13:08:17.686+0000 7f30c818d700  5 mds.beacon.b received beacon reply up:boot seq 1 rtt 0.191998
2019-11-26T13:08:17.686+0000 7f30c5187700  1 -- [v2:172.21.15.79:6834/3636622064,v1:172.21.15.79:6835/3636622064] <== mon.1 v2:172.21.15.79:3300/0 8 ==== mgrmap(e 3) v1 ==== 33895+0+0 (crc 0 0 0) 0x55f4168a4600 con 0x55f415c49800
2019-11-26T13:08:17.686+0000 7f30c5187700  1 --2- [v2:172.21.15.79:6834/3636622064,v1:172.21.15.79:6835/3636622064] >> [v2:172.21.15.79:6800/35406,v1:172.21.15.79:6801/35406] conn(0x55f416906000 0x55f415c1b080 unknown :-1 s=NONE pgs=0 cs=0 l=1 rx=0 tx=0).connect
2019-11-26T13:08:17.687+0000 7f30c798c700  1 --2- [v2:172.21.15.79:6834/3636622064,v1:172.21.15.79:6835/3636622064] >> [v2:172.21.15.79:6800/35406,v1:172.21.15.79:6801/35406] conn(0x55f416906000 0x55f415c1b080 unknown :-1 s=BANNER_CONNECTING pgs=0 cs=0 l=1 rx=0 tx=0)._handle_peer_banner_payload supported=0 required=0
2019-11-26T13:08:17.687+0000 7f30c5187700  1 -- [v2:172.21.15.79:6834/3636622064,v1:172.21.15.79:6835/3636622064] --> [v2:172.21.15.79:6800/35406,v1:172.21.15.79:6801/35406] -- mgropen(unknown.b) v3 -- 0x55f41690c000 con 0x55f416906000
```

From:

/ceph/teuthology-archive/pdonnell-2019-11-26_04:58:35-fs-wip-pdonnell-testing-20191126.005014-distro-basic-smithi/4543982/remote/smithi079/log/ceph-mds.b.log.gz

mgr receives e21 and fires off a request for the mds metadata:

```
2019-11-26T13:08:17.685+0000 7fb546808700 10 mgr ms_dispatch2 active fsmap(e 21) v1
2019-11-26T13:08:17.685+0000 7fb546808700 10 mgr ms_dispatch2 fsmap(e 21) v1
2019-11-26T13:08:17.685+0000 7fb546808700 10 mgr notify_all notify_all: notify_all fs_map
2019-11-26T13:08:17.685+0000 7fb546808700 15 mgr notify_all queuing notify to balancer
2019-11-26T13:08:17.685+0000 7fb546808700 15 mgr notify_all queuing notify to crash
2019-11-26T13:08:17.685+0000 7fb546808700 15 mgr notify_all queuing notify to devicehealth
2019-11-26T13:08:17.685+0000 7fb546808700 15 mgr notify_all queuing notify to iostat
2019-11-26T13:08:17.685+0000 7fb546808700 15 mgr notify_all queuing notify to orchestrator_cli
2019-11-26T13:08:17.685+0000 7fb546808700 15 mgr notify_all queuing notify to pg_autoscaler
2019-11-26T13:08:17.685+0000 7fb546808700 15 mgr notify_all queuing notify to progress
2019-11-26T13:08:17.685+0000 7fb546808700 15 mgr notify_all queuing notify to rbd_support
2019-11-26T13:08:17.685+0000 7fb546808700 15 mgr notify_all queuing notify to restful
2019-11-26T13:08:17.685+0000 7fb546808700 15 mgr notify_all queuing notify to status
2019-11-26T13:08:17.685+0000 7fb546808700 15 mgr notify_all queuing notify to telemetry
2019-11-26T13:08:17.685+0000 7fb546808700 15 mgr notify_all queuing notify to volumes
2019-11-26T13:08:17.685+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b82e370
2019-11-26T13:08:17.685+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b82e370
2019-11-26T13:08:17.685+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b82e370
2019-11-26T13:08:17.685+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b82e370
2019-11-26T13:08:17.685+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b82e370
2019-11-26T13:08:17.685+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b82e370
2019-11-26T13:08:17.685+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b82e370
2019-11-26T13:08:17.685+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b82e370
2019-11-26T13:08:17.685+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b82e370
2019-11-26T13:08:17.685+0000 7fb52e881700 20 mgr ~Gil Destroying new thread state 0x557c4b82e370
2019-11-26T13:08:17.685+0000 7fb52e881700 20 mgr Gil Switched to new thread state 0x557c4b82e370
2019-11-26T13:08:17.685+0000 7fb546808700  1 -- 172.21.15.79:0/35406 --> [v2:172.21.15.79:3301/0,v
1:172.21.15.79:6790/0] -- mon_command({"prefix": "mds metadata", "who": "b"} v 0) v1 -- 0x557c4b4c
a000 con 0x557c4b24a400
```

but MDS connects and sends mgropen before that request returns:

```
2019-11-26T13:08:17.687+0000 7fb54a810700 10 mgr.server ms_handle_authentication ms_handle_authent
ication new session 0x557c4b44d450 con 0x557c4b787400 entity mds.b addr
2019-11-26T13:08:17.687+0000 7fb54a810700 10 mgr.server ms_handle_authentication  session 0x557c4b
44d450 mds.b has caps allow * 'allow *'
2019-11-26T13:08:17.687+0000 7fb54a810700  1 --2- [v2:172.21.15.79:6800/35406,v1:172.21.15.79:6801
/35406] >> [v2:172.21.15.79:6834/3636622064,v1:172.21.15.79:6835/3636622064] conn(0x557c4b787400 0
x557c4b889b80 crc :-1 s=READY pgs=4 cs=0 l=1 rx=0 tx=0).ready entity=mds.? client_cookie=0 server_
cookie=0 in_seq=0 out_seq=0
2019-11-26T13:08:17.687+0000 7fb546808700  1 -- 172.21.15.79:0/35406 <== mon.2 v2:172.21.15.79:330
1/0 240 ==== fsmap(e 22) v1 ==== 1060+0+0 (crc 0 0 0) 0x557c4b6d5800 con 0x557c4b24a400
2019-11-26T13:08:17.687+0000 7fb546808700 10 mgr ms_dispatch2 active fsmap(e 22) v1
2019-11-26T13:08:17.687+0000 7fb546808700 10 mgr ms_dispatch2 fsmap(e 22) v1
...
2019-11-26T13:08:17.687+0000 7fb52e080700  1 -- [v2:172.21.15.79:6800/35406,v1:172.21.15.79:6801/3
5406] <== mds.? v2:172.21.15.79:6834/3636622064 1 ==== mgropen(mds.b) v3 ==== 62129+0+0 (crc 0 0 0
) 0x557c4b81b080 con 0x557c4b787400
2019-11-26T13:08:17.687+0000 7fb52e080700 10 mgr.server handle_open from 0x557c4b787400  mds.b
2019-11-26T13:08:17.687+0000 7fb52e080700  1 -- [v2:172.21.15.79:6800/35406,v1:172.21.15.79:6801/3
5406] --> [v2:172.21.15.79:6834/3636622064,v1:172.21.15.79:6835/3636622064] -- mgrconfigure(period
=5, threshold=5) v3 -- 0x557c4b2d3600 con 0x557c4b787400
2019-11-26T13:08:17.688+0000 7fb546808700  1 -- 172.21.15.79:0/35406 <== mon.2 v2:172.21.15.79:330
1/0 241 ==== mgrdigest v1 ==== 1721+0+0 (crc 0 0 0) 0x557c4b895680 con 0x557c4b24a400
...
2019-11-26T13:08:17.688+0000 7fb546808700  1 -- 172.21.15.79:0/35406 <== mon.2 v2:172.21.15.79:330
1/0 242 ==== mon_command_ack([{"prefix": "mds metadata", "who": "b"}]=0  v22) v1 ==== 72+0+671 (cr
c 0 0 0) 0x557c4b4a1c00 con 0x557c4b24a400
```

The issue appears to be in how we construct daemon_state in DaemonServer::handle_open. We are only creating a new daemon_state if a service daemon:
https://github.com/ceph/ceph/blob/f1e3fec83d50f5536870a9055d390a1daaaae942/src/mgr/DaemonServer.cc#L402-L408

I think the right fix here is to defer mgropen handling until we have the daemon metadata from the mons.

Related issues:

Copied to mgr - Backport #43046: nautilus: mgr: "mds metadata" to setup new D...	**Resolved**

History

#1 - 11/26/2019 08:42 PM - Patrick Donnelly

Recent master run of this job: http://pulpito.ceph.com/pdonnell-2019-11-26_19:25:42-fs-master-distro-basic-smithi/

#2 - 11/26/2019 08:52 PM - Patrick Donnelly

- Status changed from In Progress to Fix Under Review

- Pull request ID set to 31899

#3 - 11/27/2019 02:04 PM - Sage Weil

- Status changed from Fix Under Review to Pending Backport

#4 - 11/27/2019 02:18 PM - Patrick Donnelly

- Copied to Backport #43046: nautilus: mgr: "mds metadata" to setup new DaemonState races with fsmap added

#5 - 02/13/2020 12:09 PM - Nathan Cutler

- Status changed from Pending Backport to Resolved

While running with --resolve-parent, the script "backport-create-issue" noticed that all backports of this issue are in status "Resolved" or "Rejected".