RADOS - Bug #20959

cephfs application metdata not set by ceph.py

08/09/2017 03:06 PM - Sage Weil

Status:	Resolved	**Start date:**	08/09/2017
Priority:	Immediate	**Due date:**	
Assignee:	Greg Farnum	**% Done:**	0%
Category:	Correctness/Safety	**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:		**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Component(RADOS):**	Monitor
Severity:	3 - minor	**Pull request ID:**	

Description

"2017-08-09 06:52:11.115593 mon.a mon.0 172.21.15.12:6789/0 154 : cluster [WRN] Health check failed: application not enabled on 1 pool(s) (POOL_APP_NOT_ENABLED)" in cluster log

/a/sage-2017-08-09_05:29:48-rados-luminous-distro-basic-smithi/1500948

The log shows

```
2017-08-09T06:51:53.348 INFO:teuthology.orchestra.run.smithi012:Running: 'sudo adjust-ulimits ceph
-coverage /home/ubuntu/cephtest/archive/coverage timeout 120 ceph --cluster ceph fs new cephfs cep
hfs_metadata cephfs_data'
...
2017-08-09T06:51:53.845 INFO:teuthology.orchestra.run.smithi012.stderr:2017-08-09 06:51:53.837735
7f4347fff700  1 -- 172.21.15.12:0/147144223 <== mon.0 172.21.15.12:6789/0 9 ==== mon_command_ack([
{"prefix": "fs new", "data": "cephfs_data", "fs_name": "cephfs", "metadata": "cephfs_metadata"}]=0
 new fs with metadata
pool 2 and data pool 3 v2) v1 ==== 172+0+0 (3358715605 0 0) 0x7f4348002010 con 0x7f4358192360
```

but 10s of seconds later,

```
"application_metadata":{}}
...
ions":{},"application_metadata":{}}],"o
```

for the metadata pools. rbd one is fine,

```
,"application_metadata":{"rbd":{}}},
```

that's at

```
2017-08-09T06:52:14.246
```

Related issues:

Related to Ceph - Bug #20891: mon: mysterious "application not enabled on <N>...	**Resolved**	**08/03/2017**

History

#1 - 08/09/2017 03:06 PM - Sage Weil

- Description updated

#2 - 08/09/2017 03:14 PM - Greg Farnum

- Related to Bug #20891: mon: mysterious "application not enabled on <N> pool(s)" added

#3 - 08/09/2017 03:14 PM - Greg Farnum

- Assignee set to Greg Farnum

#4 - 08/09/2017 03:23 PM - Greg Farnum

Sage was right, the MDSMonitor unconditionally calls do_application_enable() and that unconditionally sets application metadata on the pending_inc's pool.

So if this is done prior to setting the luminous flags, that might be able to prompt a divergence in what's encoded where?

#5 - 08/09/2017 03:29 PM - Greg Farnum

We're encoding with the quorum features, though, so I don't think that could actually cause a problem, Maybe though.

#6 - 08/09/2017 04:17 PM - Sage Weil

- Assignee deleted (Greg Farnum)

#7 - 08/09/2017 04:19 PM - Sage Weil

The bug I hit before was doing the right checks on encoding, **but** the pending_inc was applied to the in-memory mon copy. So the in-memory osdmap had the fields populated but didn't encode them. then later when the flag **was** set we would get a crc mismatch because suddenly mon encoded fields that were populated with values but the osds didn't get that when they applied their incrementals.

It sounds like this is a somewhat different problem, but I suspect the root cause is the same: we shouldn't populate pending_inc app metadata unless require_osd_release>=luminous is set.

#8 - 08/09/2017 07:53 PM - Greg Farnum

Hmm, this still doesn't make sense. The cluster started out as luminous and so the maps would always have the luminous flags set. (And indeed, on the first dump it's got "require_osd_release":"luminous".)

#9 - 08/09/2017 07:56 PM - Greg Farnum

Okay, unlike the previous log I looked at, the "fs new" command is clearly **not** triggering a new osd map commit. We run do_application_enable against osdmap epoch 19:

```
2017-08-09 06:51:53.748483 7f6980098700  7 mon.a@0(leader).mds e1 prepare_update mon_command({"prefix": "fs ne
w", "data": "cephfs_data", "fs_name": "cephfs", "metadata": "cephfs_metadata"} v 0) v1
2017-08-09 06:51:53.748505 7f6980098700 20 mon.a@0(leader).osd e19 do_application_enable: pool_id=3, app_name=
cephfs
2017-08-09 06:51:53.748513 7f6980098700 20 mon.a@0(leader).osd e19 do_application_enable: pool_id=2, app_name=
cephfs
```

But at the end of the run it still hasn't updated; the final osd monitor log output is

```
2017-08-09 07:13:54.959684 7f6980098700 10 mon.a@0(leader).osd e19 should_propose
```

#10 - 08/09/2017 07:57 PM - Nathan Cutler

As I reported in #20891 I am seeing this on fresh luminous clusters.

#11 - 08/09/2017 09:15 PM - Greg Farnum

- Project changed from Ceph to RADOS

- Category set to Correctness/Safety

- Status changed from Verified to In Progress

- Assignee set to Greg Farnum

- Component(RADOS) Monitor added

So far I've identified three problems in the source:
1) we don't check that we're in luminous mode before the MDS sets pool application metadata, and that may have unpredictable results on which daemons see what state when we're upgrading.
2) the mdsmonitor doesn't check that the osdmonitor is_writeable() before making changes, so it may lose the pool updates
3) the mdsmonitor doesn't trigger an osdmonitor commit, so it may never actually persist the changes if the osdmap doesn't change further

None of these really satisfy me about how a brand-new luminous cluster would be missing the application_metadata in the json dump output while not warning about missing applications, but they're definitely all issues and maybe in conjunction they're having a wider and less visible impact than I think they should.

#12 - 08/11/2017 02:29 AM - Sage Weil

- Status changed from In Progress to Resolved

#13 - 08/11/2017 10:12 PM - Greg Farnum

https://github.com/ceph/ceph/pull/16954