fs - Bug #23538

mds: fix occasional dir rstat inconsistency between multi-MDSes

04/02/2018 06:57 AM - Zhi Zhang

Status:	Resolved	**Start date:**	04/02/2018
Priority:	Normal	**Due date:**	
Assignee:	Zhi Zhang	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:	v13.0.0		
Source:	Community (dev)	**Affected Versions:**	
Tags:		**ceph-qa-suite:**	
Backport:	luminous	**Component(FS):**	Client, MDS
Regression:	No	**Labels (FS):**	
Severity:	3 - minor	**Pull request ID:**	
Reviewed:			

Description

Recently we found dir rstat inconsistency between multi-MDSes on ceph version Luminous.

For example, on client A, we have a dir which is on MDS.0. Then we write some files into this dir and they are also processed on MDS.0. Then we check ceph.dir.rbytes on client A.

```
[ceph@c165 ~]$ sudo getfattr -n ceph.dir.rbytes /mnt/cephfs/test8/test9/
getfattr: Removing leading '/' from absolute path names
# file: mnt/cephfs/test8/test9/
ceph.dir.rbytes="504"
```

On client B, we remount it as a newly-mounted client and then check ceph.dir.rbytes, if it connects to MDS.1 this time.

```
[ceph@c166 ~]$ sudo getfattr -n ceph.dir.rbytes /mnt/cephfs/test8/test9/
getfattr: Removing leading '/' from absolute path names
# file: mnt/cephfs/test8/test9/
ceph.dir.rbytes="315"
```

As above, client B gets old rstat info if it connects to MDS.1. And this rstat info won't get updated for a long time.

From MDS log on MDS.1, we can also see this dir's rstat info is not up-to-date because it is replicated, not auth one.

```
...
2018-03-30 18:25:12.508385 7fa05e7c3700  7 mds.1.locker rdlock_start  on (isnap sync) on [inode 0x
1000621aa64 [...2,head] /test8/test9/ rep@0.1 v36 f(v0 m2018-03-30 18:22:30.501121 8=8+0) n(v0 rc2
018-03-30 18:21:33.550134 b315 6=5+1) (inest mix) (iversion lock) caps={17032=pAsLsXsFs/-@2} | req
uest=1 lock=0 caps=1 waiter=0 0x7fa06f3a4800]
2018-03-30 18:25:12.508401 7fa05e7c3700 10 mds.1.locker  got rdlock on (isnap sync r=1) [inode 0x1
000621aa64 [...2,head] /test8/test9/ rep@0.1 v36 f(v0 m2018-03-30 18:22:30.501121 8=8+0) n(v0 rc20
18-03-30 18:21:33.550134 b315 6=5+1) (isnap sync r=1) (inest mix) (iversion lock) caps={17032=pAsL
sXsFs/-@2} | request=1 lock=1 caps=1 waiter=0 0x7fa06f3a4800]
2018-03-30 18:25:12.508419 7fa05e7c3700 20 Session check_access path /test8/test9
2018-03-30 18:25:12.508421 7fa05e7c3700 10 MDSAuthCap is_capable inode(path /test8/test9 owner 0:0
 mode 040755) by caller 0:0 mask 1 new 0:0 cap: MDSAuthCaps[allow *]
2018-03-30 18:25:12.508434 7fa05e7c3700 20 mds.1.bal hit_dir 0 pop is 8, frag * size 1
2018-03-30 18:25:12.508440 7fa05e7c3700 10 mds.1.server reply to stat on client_request(client.170
32:6 getattr -#0x1000621aa64 2018-03-30 18:25:12.508050 caller_uid=0, caller_gid=0{0,}) v4
```

```
2018-03-30 18:25:12.508448 7fa05e7c3700  7 mds.1.server reply_client_request 0 ((0) Success) clien
t_request(client.17032:6 getattr - #0x1000621aa64 2018-03-30 18:25:12.508050 caller_uid=0, caller_
gid=0{0,}) v4
...
```

Related issues:

Copied to fs - Backport #23835: luminous: mds: fix occasional dir rstat incon...	**Resolved**

History

#1 - 04/02/2018 06:58 AM - Zhi Zhang

I looked through the code and found in MDCache::predirty_journal_parents, the parent's rstat of a file will be updated correctly. But if this parent is auth, it wasn't be marked updated scatterlock on its nestlock. However this file't other ancestor's nestlock could be marked.

So this file's parent won't do scatter/gather process in scatter_tick and hence replcated MDS won't get latest rstat info from auth one.

I made a fix here to mark file's parent nestlock to let it do scatter/gather process. From my testing, clients can get latest rstat info now no matter when annd which MDS they connect to.

#2 - 04/02/2018 07:10 AM - Zhi Zhang

https://github.com/ceph/ceph/pull/21166

#3 - 04/09/2018 01:26 PM - Patrick Donnelly

- Status changed from New to Need Review

- Target version set to v13.0.0

- Backport set to luminous

#4 - 04/24/2018 04:35 AM - Patrick Donnelly

- Category deleted (90)

- Status changed from Need Review to Pending Backport

- Assignee set to Zhi Zhang

- Component(FS) Client added

#5 - 04/24/2018 05:48 AM - Nathan Cutler

- Copied to Backport #23835: luminous: mds: fix occasional dir rstat inconsistency between multi-MDSes added

#6 - 06/10/2018 09:15 AM - Nathan Cutler

- Status changed from Pending Backport to Resolved