CephFS - Bug #20569

mds: don't mark dirty rstat on non-auth inode

07/11/2017 09:32 AM - Zhi Zhang

Status:	Resolved	**% Done:**	0%
Priority:	Normal		
Assignee:	Patrick Donnelly		
Category:			
Target version:			
Source:	Community (dev)	**ceph-qa-suite:**	
Tags:		**Component(FS):**	
Backport:		**Labels (FS):**	multimds
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature (v1):**	
Reviewed:		**Crash signature (v2):**	
Affected Versions:			

Description

Currently using multi-MDS on Luminous, we found ceph status reported such warning all the time if writing large amount of files (e.g, 1 million) into a directory and migration had happened from MDS.0 to MDS.1

```
cluster 63dd6fd2-78d5-49b0-9b3f-5b69fcbd4b25
     health HEALTH_WARN
            mds1: Too many inodes in cache (602815/100000), 92628 inodes in use by clients, 0 stra
y files
            noscrub,nodeep-scrub flag(s) set
     monmap e3: 1 mons at {c167=xxx:6789/0}
            election epoch 38, quorum 0 c167
      fsmap e10115: 2/2/2 up {0=c166=up:active,1=c167=up:active}, 1 up:standby-replay
```

From MDS perf dump, we can see inodes with caps were less than 100000, but total inodes in cache were too many.

```
[ceph@c167 ~]$ sudo ceph --admin-daemon /var/run/ceph/ceph-mds.c167.asok perf dump | grep inode
        "inode_max": 100000,
        "inodes": 602815,
        "inodes_top": 0,
        "inodes_bottom": 0,
        "inodes_pin_tail": 602815,
        "inodes_pinned": 602815,
        "inodes_expired": 706862,
        "inodes_with_caps": 92628,
        "exported_inodes": 0,
        "imported_inodes": 113536
```

MDS trim job couldn't trim those inodes because they were not expired in LRU. From MDS cache dump, we can see those inodes' "dirtyrstat" was still 1 but "caps" or other flags were already 0.

```
[inode 20001010c3e [2,head] /1499743036981_0/file_895909 auth v130471 s=73728 n(v0 b73728 1=1+0)/n
(v0 1=1+0) (iversion lock) | ptrwaiter=0 request=0 lock=0 caps=0 dirtyrstat=1 dirtyparent=0 dirty=
0 authpin=0 0x7f1cd7e5ca00]
[inode 20001010c3d [2,head] /1499743036981_0/file_895908 auth v137783 s=73728 n(v0 b73728 1=1+0)/n
(v0 1=1+0) (iversion lock) | ptrwaiter=0 request=0 lock=0 caps=0 dirtyrstat=1 dirtyparent=0 dirty=
0 authpin=0 0x7f1cd7e5d000]
```

```
...
```

The reason is that those inodes had been exported to MDS.1 and marked with dirtyrstat in predirty_journal_parents, but predirty_journal_parents was stopped later if this inode's parent was not auth for MDS.1. So its dirtyrstat flag won't been cleared.

I think we don't need to mark dirtyrstat on this inode on non-auth MDS. Auth MDS will do such things on this inode's parent.

History

#1 - 07/11/2017 09:34 AM - Zhi Zhang

https://github.com/ceph/ceph/pull/16253

Within this fix, inodes can be trimmed successfully and no such warning reported on replicated MDS.

#2 - 07/19/2017 02:34 AM - Zhi Zhang

New PR here: https://github.com/ceph/ceph/pull/16337

#3 - 07/20/2017 12:00 AM - Patrick Donnelly

- Status changed from New to 7

#4 - 07/20/2017 12:00 AM - Patrick Donnelly

- Assignee set to Patrick Donnelly

#5 - 07/21/2017 08:31 PM - Patrick Donnelly

- Status changed from 7 to Resolved

#6 - 03/09/2019 12:32 AM - Patrick Donnelly

- Category deleted (90)

- Labels (FS) multimds added