# 问题背景:

客户现网出现pg状态处于active+ remapped， 且无法恢复

当时做过的操作有给其中一个rack下的两个主机降权重，从而进行缩容操作，从ceph osd tree信息看也修改了这些osd的reweight 权重

# **问题现象:**

现网集群针对一个rack下的主机host做过osd的权重归0 即缩容，但是后面对应pg的状态一直卡在active + remapped， 无法恢复，截图如下，但是下面这个截图是操作之后的截图，应该是有16个pg处于active + remapped

[root@rg15-ceph001 ~]# ceph -s

cluster da895fb8-36f1-311d-877d-3d41768303ee

health HEALTH\_WARN

38 pgs backfill\_wait

4 pgs backfilling

55 pgs stuck unclean

recovery 22310/414242202 objects misplaced (0.005%)

noscrub,nodeep-scrub flag(s) set

mon.rg15-ceph001 low disk space; store is getting too big! 37180 MB >= 15360 MB

mon.rg15-ceph002 store is getting too big! 37694 MB >= 15360 MB

mon.rg15-ceph003 store is getting too big! 43563 MB >= 15360 MB

monmap e19: 3 mons at {rg15-ceph001=172.30.2.5:6789/0,rg15-ceph002=172.30.2.6:6789/0,rg15-ceph003=172.30.2.7:6789/0}

election epoch 84894, quorum 0,1,2 rg15-ceph001,rg15-ceph002,rg15-ceph003

osdmap e985004: 678 osds: 678 up, 678 in; 57 remapped pgs

flags noscrub,nodeep-scrub,sortbitwise,require\_jewel\_osds

pgmap v112117641: 37184 pgs, 7 pools, 524 TB data, 131 Mobjects

1577 TB used, 2636 TB / 4214 TB avail

22310/414242202 objects misplaced (0.005%)

37127 active+clean

38 active+remapped+wait\_backfill

15 active+remapped

4 active+remapped+backfilling

recovery io 58202 kB/s, 14 objects/s

client io 1627 MB/s rd, 533 MB/s wr, 19864 op/s rd, 30334 op/s wr

# **问题简单分析：**

简单查看了pg的几种状态，active+ remapped， 说明处于这个状态的pg是能够正常承接客户业务，而且也是三副本，倒不是两副本冗余，因为acting\_set集合是三个osd，单独查看一个pg的query信息下:

[root@rg15-ceph001 ~]# ceph pg 3.1d85 query

{

"state": "active+remapped",

"snap\_trimq": "[]",

"epoch": 985064,

"up": [

96,

86

],

"acting": [

96,

86,

183

],

"actingbackfill": [

"86",

"96",

"183"

],

"info": {

"pgid": "3.1d85",

"last\_update": "965083'25950512",

"last\_complete": "965083'25950512",

"log\_tail": "896756'25947438",

"last\_user\_version": 25950512,

"last\_backfill": "MAX",

"last\_backfill\_bitwise": 1,

"purged\_snaps": "[1~111,113~1,115~271,388~19,3a2~4b,3ee~214]",

"history": {

"epoch\_created": 332,

"last\_epoch\_started": 976195,

"last\_epoch\_clean": 976195,

"last\_epoch\_split": 0,

"last\_epoch\_marked\_full": 0,

"same\_up\_since": 976194,

"same\_interval\_since": 976194,

"same\_primary\_since": 812515,

"last\_scrub": "121885'11927328",

"last\_scrub\_stamp": "2022-05-18 05:39:19.257154",

"last\_deep\_scrub": "14653'445629",

"last\_deep\_scrub\_stamp": "2021-06-03 07:20:46.715152",

"last\_clean\_scrub\_stamp": "2022-05-18 05:39:19.257154"

},

"stats": {

"version": "965083'25950512",

"reported\_seq": "22511376",

"reported\_epoch": "985060",

"state": "active+remapped",

"last\_fresh": "2025-03-20 10:58:47.993198",

"last\_change": "2025-03-17 12:03:40.963360",

"last\_active": "2025-03-20 10:58:47.993198",

"last\_peered": "2025-03-20 10:58:47.993198",

"last\_clean": "2025-03-17 11:18:28.114977",

"last\_became\_active": "2025-03-17 12:03:40.963360",

"last\_became\_peered": "2025-03-17 12:03:40.963360",

"last\_unstale": "2025-03-20 10:58:47.993198",

"last\_undegraded": "2025-03-20 10:58:47.993198",

"last\_fullsized": "2025-03-20 10:58:47.993198",

"mapping\_epoch": 976112,

"log\_start": "896756'25947438",

"ondisk\_log\_start": "896756'25947438",

"created": 332,

"last\_epoch\_clean": 976195,

"parent": "0.0",

"parent\_split\_bits": 0,

"last\_scrub": "121885'11927328",

"last\_scrub\_stamp": "2022-05-18 05:39:19.257154",

"last\_deep\_scrub": "14653'445629",

"last\_deep\_scrub\_stamp": "2021-06-03 07:20:46.715152",

"last\_clean\_scrub\_stamp": "2022-05-18 05:39:19.257154",

"log\_size": 3074,

"ondisk\_log\_size": 3074,

"stats\_invalid": false,

"dirty\_stats\_invalid": false,

"omap\_stats\_invalid": false,

"hitset\_stats\_invalid": false,

"hitset\_bytes\_stats\_invalid": false,

"pin\_stats\_invalid": false,

"stat\_sum": {

"num\_bytes": 1223372800,

"num\_objects": 302,

"num\_object\_clones": 13,

"num\_object\_copies": 906,

"num\_objects\_missing\_on\_primary": 0,

"num\_objects\_missing": 0,

"num\_objects\_degraded": 0,

"num\_objects\_misplaced": 302,

"num\_objects\_unfound": 0,

"num\_objects\_dirty": 302,

"num\_whiteouts": 0,

"num\_read": 12777793,

"num\_read\_kb": 964341475,

"num\_write": 25957257,

"num\_write\_kb": 950386041,

"num\_scrub\_errors": 0,

"num\_shallow\_scrub\_errors": 0,

"num\_deep\_scrub\_errors": 0,

"num\_objects\_recovered": 94007,

"num\_bytes\_recovered": 388962918744,

"num\_keys\_recovered": 173,

"num\_objects\_omap": 0,

"num\_objects\_hit\_set\_archive": 0,

"num\_bytes\_hit\_set\_archive": 0,

"num\_flush": 0,

"num\_flush\_kb": 0,

"num\_evict": 0,

"num\_evict\_kb": 0,

"num\_promote": 0,

"num\_flush\_mode\_high": 0,

"num\_flush\_mode\_low": 0,

"num\_evict\_mode\_some": 0,

"num\_evict\_mode\_full": 0,

"num\_objects\_pinned": 0

},

"up": [

96,

86

],

"acting": [

96,

86,

183

],

"blocked\_by": [],

"up\_primary": 96,

"acting\_primary": 96

},

"empty": 0,

"dne": 0,

"incomplete": 0,

"last\_epoch\_started": 976195,

"hit\_set\_history": {

"current\_last\_update": "0'0",

"history": []

}

},

"peer\_info": [

{

"peer": "86",

"pgid": "3.1d85",

"last\_update": "965083'25950512",

"last\_complete": "965083'25950512",

"log\_tail": "896756'25947438",

"last\_user\_version": 25950512,

"last\_backfill": "MAX",

"last\_backfill\_bitwise": 1,

"purged\_snaps": "[1~111,113~1,115~271,388~19,3a2~4b,3ee~214]",

"history": {

"epoch\_created": 332,

"last\_epoch\_started": 976195,

"last\_epoch\_clean": 976195,

"last\_epoch\_split": 0,

"last\_epoch\_marked\_full": 0,

"same\_up\_since": 976194,

"same\_interval\_since": 976194,

"same\_primary\_since": 812515,

"last\_scrub": "121885'11927328",

"last\_scrub\_stamp": "2022-05-18 05:39:19.257154",

"last\_deep\_scrub": "14653'445629",

"last\_deep\_scrub\_stamp": "2021-06-03 07:20:46.715152",

"last\_clean\_scrub\_stamp": "2022-05-18 05:39:19.257154"

},

"stats": {

"version": "965083'25950511",

"reported\_seq": "22510930",

"reported\_epoch": "965083",

"state": "active+clean",

"last\_fresh": "2025-03-12 03:27:59.872142",

"last\_change": "2025-03-12 03:26:34.685945",

"last\_active": "2025-03-12 03:27:59.872142",

"last\_peered": "2025-03-12 03:27:59.872142",

"last\_clean": "2025-03-12 03:27:59.872142",

"last\_became\_active": "2025-03-12 03:26:34.685829",

"last\_became\_peered": "2025-03-12 03:26:34.685829",

"last\_unstale": "2025-03-12 03:27:59.872142",

"last\_undegraded": "2025-03-12 03:27:59.872142",

"last\_fullsized": "2025-03-12 03:27:59.872142",

"mapping\_epoch": 976112,

"log\_start": "896756'25947438",

"ondisk\_log\_start": "896756'25947438",

"created": 332,

"last\_epoch\_clean": 965081,

"parent": "0.0",

"parent\_split\_bits": 0,

"last\_scrub": "121885'11927328",

"last\_scrub\_stamp": "2022-05-18 05:39:19.257154",

"last\_deep\_scrub": "14653'445629",

"last\_deep\_scrub\_stamp": "2021-06-03 07:20:46.715152",

"last\_clean\_scrub\_stamp": "2022-05-18 05:39:19.257154",

"log\_size": 3073,

"ondisk\_log\_size": 3073,

"stats\_invalid": false,

"dirty\_stats\_invalid": false,

"omap\_stats\_invalid": false,

"hitset\_stats\_invalid": false,

"hitset\_bytes\_stats\_invalid": false,

"pin\_stats\_invalid": false,

"stat\_sum": {

"num\_bytes": 1223372800,

"num\_objects": 302,

"num\_object\_clones": 13,

"num\_object\_copies": 909,

"num\_objects\_missing\_on\_primary": 0,

"num\_objects\_missing": 0,

"num\_objects\_degraded": 0,

"num\_objects\_misplaced": 0,

"num\_objects\_unfound": 0,

"num\_objects\_dirty": 302,

"num\_whiteouts": 0,

"num\_read": 12777793,

"num\_read\_kb": 964341475,

"num\_write": 25957257,

"num\_write\_kb": 950386041,

"num\_scrub\_errors": 0,

"num\_shallow\_scrub\_errors": 0,

"num\_deep\_scrub\_errors": 0,

"num\_objects\_recovered": 94007,

"num\_bytes\_recovered": 388962918744,

"num\_keys\_recovered": 173,

"num\_objects\_omap": 0,

"num\_objects\_hit\_set\_archive": 0,

"num\_bytes\_hit\_set\_archive": 0,

"num\_flush": 0,

"num\_flush\_kb": 0,

"num\_evict": 0,

"num\_evict\_kb": 0,

"num\_promote": 0,

"num\_flush\_mode\_high": 0,

"num\_flush\_mode\_low": 0,

"num\_evict\_mode\_some": 0,

"num\_evict\_mode\_full": 0,

"num\_objects\_pinned": 0

},

"up": [

96,

86

],

"acting": [

96,

86,

183

],

"blocked\_by": [],

"up\_primary": 96,

"acting\_primary": 96

},

"empty": 0,

"dne": 0,

"incomplete": 0,

"last\_epoch\_started": 976195,

"hit\_set\_history": {

"current\_last\_update": "0'0",

"history": []

}

},

{

"peer": "183",

"pgid": "3.1d85",

"last\_update": "965083'25950512",

"last\_complete": "965083'25950512",

"log\_tail": "896756'25947438",

"last\_user\_version": 25950512,

"last\_backfill": "MAX",

"last\_backfill\_bitwise": 1,

"purged\_snaps": "[1~111,113~1,115~271,388~19,3a2~4b,3ee~214]",

"history": {

"epoch\_created": 332,

"last\_epoch\_started": 976195,

"last\_epoch\_clean": 976195,

"last\_epoch\_split": 0,

"last\_epoch\_marked\_full": 0,

"same\_up\_since": 976194,

"same\_interval\_since": 976194,

"same\_primary\_since": 812515,

"last\_scrub": "121885'11927328",

"last\_scrub\_stamp": "2022-05-18 05:39:19.257154",

"last\_deep\_scrub": "14653'445629",

"last\_deep\_scrub\_stamp": "2021-06-03 07:20:46.715152",

"last\_clean\_scrub\_stamp": "2022-05-18 05:39:19.257154"

},

"stats": {

"version": "965083'25950511",

"reported\_seq": "22510930",

"reported\_epoch": "965083",

"state": "active+clean",

"last\_fresh": "2025-03-12 03:27:59.872142",

"last\_change": "2025-03-12 03:26:34.685945",

"last\_active": "2025-03-12 03:27:59.872142",

"last\_peered": "2025-03-12 03:27:59.872142",

"last\_clean": "2025-03-12 03:27:59.872142",

"last\_became\_active": "2025-03-12 03:26:34.685829",

"last\_became\_peered": "2025-03-12 03:26:34.685829",

"last\_unstale": "2025-03-12 03:27:59.872142",

"last\_undegraded": "2025-03-12 03:27:59.872142",

"last\_fullsized": "2025-03-12 03:27:59.872142",

"mapping\_epoch": 976112,

"log\_start": "896756'25947438",

"ondisk\_log\_start": "896756'25947438",

"created": 332,

"last\_epoch\_clean": 965081,

"parent": "0.0",

"parent\_split\_bits": 0,

"last\_scrub": "121885'11927328",

"last\_scrub\_stamp": "2022-05-18 05:39:19.257154",

"last\_deep\_scrub": "14653'445629",

"last\_deep\_scrub\_stamp": "2021-06-03 07:20:46.715152",

"last\_clean\_scrub\_stamp": "2022-05-18 05:39:19.257154",

"log\_size": 3073,

"ondisk\_log\_size": 3073,

"stats\_invalid": false,

"dirty\_stats\_invalid": false,

"omap\_stats\_invalid": false,

"hitset\_stats\_invalid": false,

"hitset\_bytes\_stats\_invalid": false,

"pin\_stats\_invalid": false,

"stat\_sum": {

"num\_bytes": 1223372800,

"num\_objects": 302,

"num\_object\_clones": 13,

"num\_object\_copies": 909,

"num\_objects\_missing\_on\_primary": 0,

"num\_objects\_missing": 0,

"num\_objects\_degraded": 0,

"num\_objects\_misplaced": 0,

"num\_objects\_unfound": 0,

"num\_objects\_dirty": 302,

"num\_whiteouts": 0,

"num\_read": 12777793,

"num\_read\_kb": 964341475,

"num\_write": 25957257,

"num\_write\_kb": 950386041,

"num\_scrub\_errors": 0,

"num\_shallow\_scrub\_errors": 0,

"num\_deep\_scrub\_errors": 0,

"num\_objects\_recovered": 94007,

"num\_bytes\_recovered": 388962918744,

"num\_keys\_recovered": 173,

"num\_objects\_omap": 0,

"num\_objects\_hit\_set\_archive": 0,

"num\_bytes\_hit\_set\_archive": 0,

"num\_flush": 0,

"num\_flush\_kb": 0,

"num\_evict": 0,

"num\_evict\_kb": 0,

"num\_promote": 0,

"num\_flush\_mode\_high": 0,

"num\_flush\_mode\_low": 0,

"num\_evict\_mode\_some": 0,

"num\_evict\_mode\_full": 0,

"num\_objects\_pinned": 0

},

"up": [

96,

86

],

"acting": [

96,

86,

183

],

"blocked\_by": [],

"up\_primary": 96,

"acting\_primary": 96

},

"empty": 0,

"dne": 0,

"incomplete": 0,

"last\_epoch\_started": 976195,

"hit\_set\_history": {

"current\_last\_update": "0'0",

"history": []

}

}

],

"recovery\_state": [

{

"name": "Started\/Primary\/Active",

"enter\_time": "2025-03-17 12:03:40.931102",

"might\_have\_unfound": [],

"recovery\_progress": {

"backfill\_targets": [],

"waiting\_on\_backfill": [],

"last\_backfill\_started": "MIN",

"backfill\_info": {

"begin": "MIN",

"end": "MIN",

"objects": []

},

"peer\_backfill\_info": [],

"backfills\_in\_flight": [],

"recovering": [],

"pg\_backend": {

"pull\_from\_peer": [],

"pushing": []

}

},

"scrub": {

"scrubber.epoch\_start": "0",

"scrubber.active": 0,

"scrubber.state": "INACTIVE",

"scrubber.start": "MIN",

"scrubber.end": "MIN",

"scrubber.subset\_last\_update": "0'0",

"scrubber.deep": false,

"scrubber.seed": 0,

"scrubber.waiting\_on": 0,

"scrubber.waiting\_on\_whom": []

}

},

{

"name": "Started",

"enter\_time": "2025-03-17 12:03:39.952279"

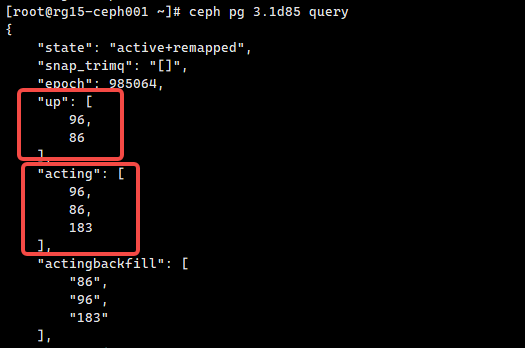
}

],

"agent\_state": {}

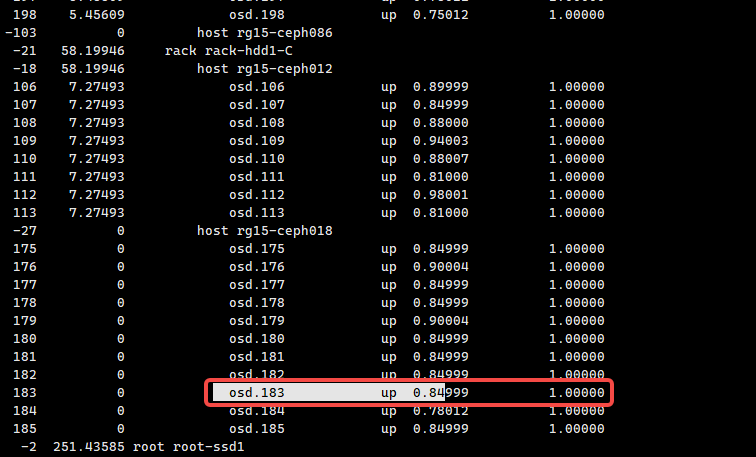
}

其中比较关键的点在于：



此时该pg的acting\_set 是[96, 86, 183] , 但是up\_set是[96, 86] 也就是说183 osd 没有被选出来

我们看一下184 osd所在的主机

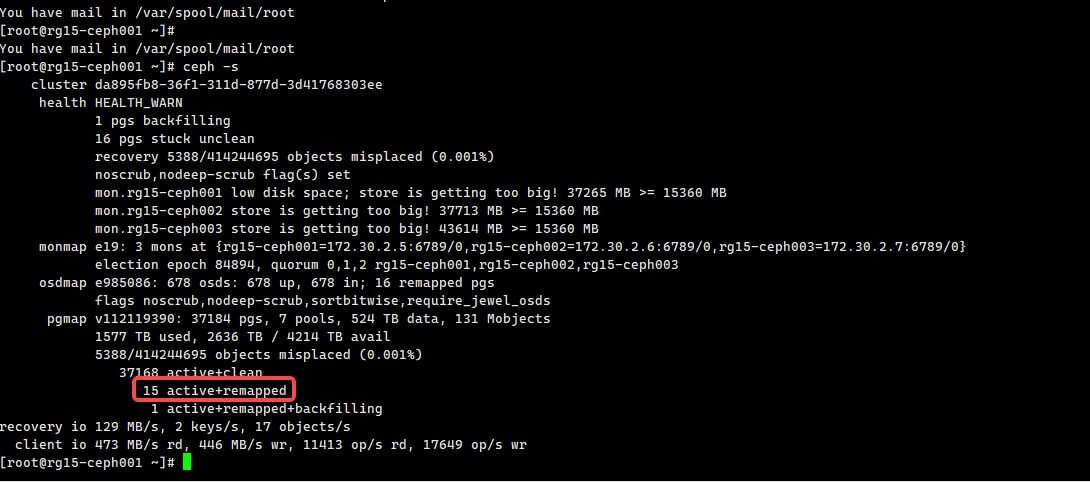


可以看到这个osd所在的主机其权重weight都是0, 但是acting\_set 还包含183， 所以这个osd还是会承担业务，现在问题是up\_set 选不出来，问题点主要是crush算法通过该集群crushmap无法选出该rack下的host；

再看ceph osd tree 对应该rack的主机osd，发现该rack下所有的主机的reweight 都被改小了, < 1.000000， pg有两个weight权重，一个是按照容量weight进行计算，还有一个是reweight微调，这两个权重的区别在于weight 会改变该主机在整个集群的权重，当然osd里的pg会迁移出去，从而造成大量的数据迁移，reweight 只针对该主机内部的数据迁移，主机的权重不会改变，变的是host内部的权重，即pg迁移的范围减少，从而对集群的影响减少

但是改小reweight可能会造成crush算法对osd的选取的影响，过载测试过不了，从而造成当pg处于remap状态一直卡住

重新调整osd reweight，此时对应的pg重新处于选择，就能够选出对应的osd，从而跳过remapped状态，进入recovery流程



这个集群对应的ceph版本是10.2.10， 是一个稳定版本

[root@rg15-ceph001 ~]# ceph -v

ceph version 10.2.10 (5dc1e4c05cb68dbf62ae6fce3f0700e4654fdbbe)

理论上不太可能存在软件bug，所以定位思路不要往软件缺陷方面想，成熟的软件产品多半是维护过程中操作导致了软件无法正常执行，其实ceph crush算法本身也存在一些小问题，包括递归选取，选不出卡住，但是更多的是因操作导致crush条件不满足算法的要求，从而返回-1， 在集群层面看到的问题就是pg状态的卡住