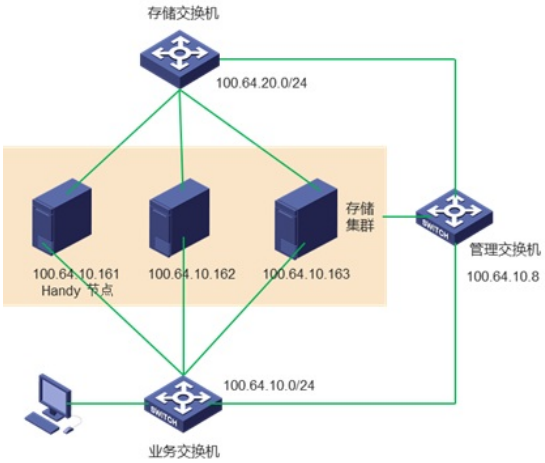
ONEStor3.0中OSD对PG状态的影响

zhiliao_lhl34 2019-10-31 发表

组网及说明

1、实验环境

采用基于CAS 5.0 的虚拟机搭建ONEStor 3.0存储集群



2、网络规划

管理网段： 100.64.10.0/24

存储前端网：100.64.20.0/24

存储后端网：100.64.20.0/24

3、集群部署

VM01 作为管理节点

副本策略选用3副本。

4、查看PG状态

ceph pg stat

5、查看pg组的映射信息

ceph pg dump 【all|summary|sum|delta|pools|osds|pgs|pgs_brief】

6、查看一个PG的map

ceph pg map 1.730

7、其他相关命令：

systemctl stop ceph-osd@3.service

umount /dev/sdc1 /var/lib/ceph/osd/ceph-3

systemctl start ceph-osd@3.service

mount /dev/sdc1 /var/lib/ceph/osd/ceph-3

ceph osd out osd.3

ceph osd in osd.3

ceph osd tree

问题描述

在实际环境中OSD的down状态和out状态会对PG的状态产生影响，本实验探究一下具体的影响。

过程分析

一、 实验步骤及现象

1、down掉一个osd

1）查看健康情况下的pg状态

```
[root@onestor102 ~]# ceph pg stat
1024 pgs: 1024 active+clean; 17495 MB data, 67455 MB used, 1132 GB / 1198 GB avail
[root@onestor102 ~]# ceph osd tree
ID  CLASS WEIGHT  TYPE NAME                          STATUS REWEIGHT PRI-AFF
-10            0 root maintain
 -9       1.17114 root diskpool01
-11       1.17114     rack rack01.diskpool01
-15       0.39038         host onestor101.diskpool01
  2   hdd 0.09760             osd.2                      up  1.00000 1.00000
  5   hdd 0.09760             osd.5                      up  1.00000 1.00000
  8   hdd 0.09760             osd.8                      up  1.00000 1.00000
 11   hdd 0.09760             osd.11                     up  1.00000 1.00000
 -7       0.39038         host onestor102.diskpool01
  0   hdd 0.09760             osd.0                      up  1.00000 1.00000
  3   hdd 0.09760             osd.3                      up  1.00000 1.00000
  6   hdd 0.09760             osd.6                      up  1.00000 1.00000
  9   hdd 0.09760             osd.9                      up  1.00000 1.00000
 -4       0.39038         host onestor103.diskpool01
  1   hdd 0.09760             osd.1                      up  1.00000 1.00000
  4   hdd 0.09760             osd.4                      up  1.00000 1.00000
  7   hdd 0.09760             osd.7                      up  1.00000 1.00000
 10   hdd 0.09760             osd.10                     up  1.00000 1.00000
 -1            0 root default
```

此时OSD均正常

```
[root@onestor102 ~]# ceph pg dump pgs_brief |more
dumped pgs_brief
PG_STAT STATE        UP        UP_PRIMARY ACTING    ACTING_PRIMARY
2.194   active+clean   [3,8,4]          3   [3,8,4]              3
1.197   active+clean  [5,10,6]          5  [5,10,6]              5
2.195   active+clean  [6,11,1]          6  [6,11,1]              6
1.196   active+clean   [1,9,8]          1   [1,9,8]              1
2.196   active+clean   [5,0,7]          5   [5,0,7]              5
1.195   active+clean   [9,2,4]          9   [9,2,4]              9
2.197   active+clean   [5,9,4]          5   [5,9,4]              5
1.194   active+clean   [7,0,5]          7   [7,0,5]              7
2.190   active+clean   [2,3,1]          2   [2,3,1]              2
1.193   active+clean   [9,7,8]          9   [9,7,8]              9
2.191   active+clean   [5,3,7]          5   [5,3,7]              5
1.192   active+clean   [2,7,9]          2   [2,7,9]              2
2.192   active+clean   [6,7,8]          6   [6,7,8]              6
1.191   active+clean   [8,1,6]          8   [8,1,6]              8
2.193   active+clean  [10,8,3]         10  [10,8,3]             10
1.190   active+clean  [1,9,11]          1  [1,9,11]              1
```

此时PG均正常

2）down掉一个osd进程后查看down掉一个osd后的pg状

```
[root@onestor102 ~]# ceph pg stat
1024 pgs: 258 active+undersized+degraded, 766 active+clean; 18026 MB data, 69055 MB used, 1131 GB / 11
98 GB avail; 102 B/s rd, 11812 KB/s wr, 0 op/s rd, 103 op/s wr; 1181/13548 objects degraded (8.717%)
[root@onestor102 ~]# ceph health detail
HEALTH_WARN 1 osds down; Degraded data redundancy: 1181/13548 objects degraded (8.717%), 233 pgs uncle
an, 258 pgs degraded
OSD_DOWN 1 osds down
    osd.3 (root=diskpool01,rack=rack01.diskpool01,host=onestor102.diskpool01) is down
PG_DEGRADED Degraded data redundancy: 1181/13548 objects degraded (8.717%), 233 pgs unclean, 258 pgs d
egraded
    pg 1.137 is stuck unclean for 303.389532, current state active+undersized+degraded, last acting [4
,5]
    pg 1.139 is active+undersized+degraded, acting [10,11]
    pg 1.140 is active+undersized+degraded, acting [2,1]
    pg 1.148 is active+undersized+degraded, acting [2,7]
    pg 1.14b is active+undersized+degraded, acting [8,7]
    pg 1.14c is active+undersized+degraded, acting [5,1]
    pg 1.15e is active+undersized+degraded, acting [2,7]
```

此时PG处于undersized+degraded状态，两副本

4）拉起osd查看pg状态

```
[root@onestor102 ~]# ceph pg dump pgs_brief |more
dumped pgs_brief
PG_STAT STATE                      UP        UP_PRIMARY ACTING    ACTING_PRIMARY
2.194   active+recovering+degraded   [3,8,4]          3   [3,8,4]              3
1.197                active+clean  [5,10,6]          5  [5,10,6]              5
2.195                active+clean  [6,11,1]          6  [6,11,1]              6
1.196                active+clean   [1,9,8]          1   [1,9,8]              1
2.196                active+clean   [5,0,7]          5   [5,0,7]              5
1.195                active+clean   [9,2,4]          9   [9,2,4]              9
2.197                active+clean   [5,9,4]          5   [5,9,4]              5
1.194                active+clean   [7,0,5]          7   [7,0,5]              7
2.190   active+recovering+degraded   [2,3,1]          2   [2,3,1]              2
1.193                active+clean   [9,7,8]          9   [9,7,8]              9
2.191   active+recovering+degraded   [5,3,7]          5   [5,3,7]              5
1.192                active+clean   [2,7,9]          2   [2,7,9]              2
2.192                active+clean   [6,7,8]          6   [6,7,8]              6
1.191                active+clean   [8,1,6]          8   [8,1,6]              8
2.193   active+recovering+degraded  [10,8,3]         10  [10,8,3]             10
1.190                active+clean  [1,9,11]          1  [1,9,11]              1
2.18c                active+clean  [7,6,11]          7  [7,6,11]              7
1.18f                active+clean  [5,0,10]          5  [5,0,10]              5
2.18d                active+clean  [8,6,10]          8  [8,6,10]              8
1.18e                active+clean   [9,7,2]          9   [9,7,2]              9
2.18e                active+clean  [11,4,6]         11  [11,4,6]             11
1.18d                active+clean   [3,2,7]          3   [3,2,7]              3
2.18f                active+clean   [9,7,5]          9   [9,7,5]              9
```

```
[root@onestor102 ceph]# ceph pg stat
1024 pgs: 1024 active+clean; 50116 MB data, 166 GB used, 1032 GB / 1198 GB avail
[root@onestor102 ceph]# ceph health detail
HEALTH_OK
[root@onestor102 ceph]# ceph pg dump pgs_brief |more
dumped pgs_brief
PG_STAT STATE        UP        UP_PRIMARY ACTING    ACTING_PRIMARY
2.194   active+clean   [3,8,4]          3   [3,8,4]              3
1.197   active+clean  [5,10,6]          5  [5,10,6]              5
2.195   active+clean  [6,11,1]          6  [6,11,1]              6
1.196   active+clean   [1,9,8]          1   [1,9,8]              1
2.196   active+clean   [5,0,7]          5   [5,0,7]              5
1.195   active+clean   [9,2,4]          9   [9,2,4]              9
2.197   active+clean   [5,9,4]          5   [5,9,4]              5
1.194   active+clean   [7,0,5]          7   [7,0,5]              7
2.190   active+clean   [2,3,1]          2   [2,3,1]              2
1.193   active+clean   [9,7,8]          9   [9,7,8]              9
2.191   active+clean   [5,3,7]          5   [5,3,7]              5
1.192   active+clean   [2,7,9]          2   [2,7,9]              2
2.192   active+clean   [6,7,8]          6   [6,7,8]              6
1.191   active+clean   [8,1,6]          8   [8,1,6]              8
2.193   active+clean  [10,8,3]         10  [10,8,3]             10
1.190   active+clean  [1,9,11]          1  [1,9,11]              1
2.18c   active+clean  [7,6,11]          7  [7,6,11]              7
```

不正常的PG先处于recovering+degradeed状态，之后数据平衡处于正常状态。

2、out掉一个osd

1）out osd并查看pg状态

```
[root@onestor102 ~]# ceph pg dump pgs_brief |more
dumped pgs_brief
PG_STAT STATE                          UP        UP_PRIMARY ACTING    ACTING_PRIMARY
2.194   active+recovering+degraded   [6,8,4]          6  [6,8,4]              6
1.197                active+clean  [5,10,6]          5 [5,10,6]              5
2.195                active+clean  [6,11,1]          6 [6,11,1]              6
1.196                active+clean   [1,9,8]          1  [1,9,8]              1
2.196                active+clean   [5,0,7]          5  [5,0,7]              5
1.195                active+clean   [9,2,4]          9  [9,2,4]              9
2.197                active+clean   [5,9,4]          5  [5,9,4]              5
1.194                active+clean   [7,0,5]          7  [7,0,5]              7
2.190   active+recovering+degraded   [2,6,1]          2  [2,6,1]              2
1.193                active+clean   [9,7,8]          9  [9,7,8]              9
2.191   active+recovering+degraded   [5,0,7]          5  [5,0,7]              5
1.192                active+clean   [2,7,9]          2  [2,7,9]              2
2.192                active+clean   [6,7,8]          6  [6,7,8]              6
1.191                active+clean   [8,1,6]          8  [8,1,6]              8
2.193   active+recovering+degraded  [10,8,6]         10 [10,8,6]             10
1.190                active+clean   [1,9,11]          1  [1,9,11]             1
2.18c                active+clean   [7,6,11]          7  [7,6,11]             7
1.18f                active+clean   [5,0,10]          5  [5,0,10]             5
2.18d                active+clean   [8,6,10]          8  [8,6,10]             8
1.18e                active+clean   [9,7,2]           9  [9,7,2]              9
2.18e                active+clean  [11,4,6]          11 [11,4,6]             11
1.18d                active+clean   [2,7,6]          2  [2,7,6]              2
2.18f                active+clean   [9,7,5]          9  [9,7,5]              9
1.18c                active+clean   [6,2,7]          6  [6,2,7]              6
2.188   active+recovering+degraded  [11,1,6]         11 [11,1,6]             11
```

此时PG处于recovering+degradeed状态，三副本，之后数据平衡。

2）将out 掉的osd.3 重新拉回集群，查看此时PG状态

```
[root@onestor102 ceph]# ceph pg dump pgs_brief |more
dumped pgs_brief
PG_STAT STATE                          UP        UP_PRIMARY ACTING    ACTING_PRIMARY
2.194   active+recovering+degraded   [3,8,4]          3  [3,8,4]              3
1.197                active+clean  [5,10,6]          5 [5,10,6]              5
2.195                active+clean  [6,11,1]          6 [6,11,1]              6
1.196                active+clean   [1,9,8]          1  [1,9,8]              1
2.196                active+clean   [5,0,7]          5  [5,0,7]              5
1.195                active+clean   [9,2,4]          9  [9,2,4]              9
2.197                active+clean   [5,9,4]          5  [5,9,4]              5
1.194                active+clean   [7,0,5]          7  [7,0,5]              7
2.190   active+recovering+degraded   [2,3,1]          2  [2,3,1]              2
1.193                active+clean   [9,7,8]          9  [9,7,8]              9
2.191   active+recovering+degraded   [5,3,7]          5  [5,3,7]              5
1.192                active+clean   [2,7,9]          2  [2,7,9]              2
2.192                active+clean   [6,7,8]          6  [6,7,8]              6
1.191                active+clean   [8,1,6]          8  [8,1,6]              8
2.193   active+recovering+degraded  [10,8,3]         10 [10,8,3]             10
1.190                active+clean   [1,9,11]          1  [1,9,11]             1
2.18c                active+clean   [7,6,11]          7  [7,6,11]             7
1.18f                active+clean   [5,0,10]          5  [5,0,10]             5
2.18d                active+clean   [8,6,10]          8  [8,6,10]             8
1.18e                active+clean   [9,7,2]           9  [9,7,2]              9
2.18e                active+clean  [11,4,6]          11 [11,4,6]             11
1.18d                active+clean   [3,2,7]          3  [3,2,7]              3
```

```
PG_STAT STATE                          UP      UP_PRIMARY   ACTING   ACTING_PRIMARY

2.14b   active+remapped+backfilling [4,8,3]      4            [4,8,9]      4
2.9d    active+remapped+backfilling [7,3,2]      7            [7,2,6]      7
2.f6    active+remapped+backfilling [4,3,2]      4            [4,2,9]      4
2.b9    active+remapped+backfilling [3,10,8]     3            [8,10,9]     8
2.112   active+remapped+backfilling [11,3,1]     11           [11,1,9]     11
```

此时PG又处于recovering+degradeed或者remapped+backfilling状态，之后数据平衡，恢复原状态。

解决方法

OSD/PG状态总结

	down	down恢复	out	out恢复
PG状态	active+clean——active+degraded+undersized	active+recovering+degrade——active+clean	active+clean——active+recovering+degrade——active+clean	active+clean——active+recovering+degrade/active+remapped+backfilling——active+clean
副本数	2	3	3	3
PG平衡	不触发	触发	触发	触发
集群健康度	不健康	100%	100%	100%
映射关系	不增新osd	恢复	增加新osd	恢复

附件下载： PG概述及OSD对PG状态的影响.rar