存储常见故障排除方法

# 本文目的

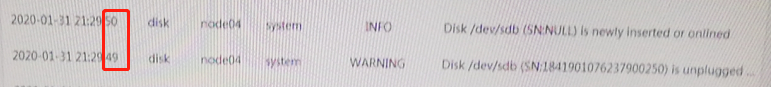
在使用存储过程中，可能会出现一些常见的告警或故障/问题，本文档对这些问题进行整理，并给出可能造成该问题的原因及解决方案，方便运维。

# 节点管理

## 2.1 V3.0-3.5HBA卡系统盘RAID闪断误报

### 问题描述

几乎同一时间出现磁盘告警offline、online(闪断)，如下图：



确认该告警的磁盘是HBA卡系统盘raid，确认raid状态正常。

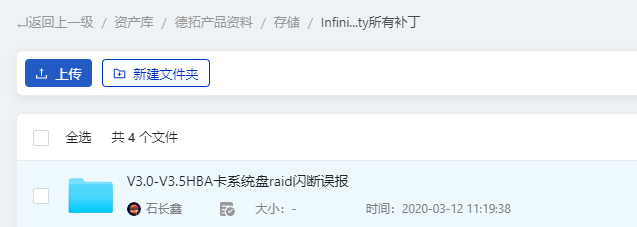
### 可能原因

确认告警出现的时间前后，message、ceph和ctdb日志中未出现业务和服务受到影响，闪断告警未造成其他影响便确认是误报。

### 解决方案

HBA卡系统盘raid闪断误报，通过发布补丁解决。

补丁获取路径：



补丁适配版本：

产品版本：infinity V3.0 至V3.5所有版本适用此补丁升级

注意事项：infinityV3.6版本开始不需要此补丁

升级方案：

备份配置文件：

# cd /opt/infinity/python/apps/misc/infi\_monitor/triggers/

# cp monitor\_disk.py monitor\_disk.py\_bak2020xxxx

补丁升级：

1）该补丁升级不影响业务和界面访问

2）每个节点都需要进行升级操作

3）解压补丁

# tar -zxvf infinityv3.x\_monitor\_disk.tar.gz

4）进入解压目录

# cd infinityv3.x\_monitor\_disk

5）执行安装脚本

# ./upgrade.sh

升级失败处理：

若升级过程出现问题，先联系研发处理，研发确认回退后按照回退方案回退。

回退方案：

1）还原之前备份的配置文件

# cd /opt/infinity/python/apps/misc/infi\_monitor/triggers/

# mv monitor\_disk.py\_bak2020xxxx monitor\_disk.py

2）重启服务

# systemctl restart infi-monitor

## 2.2 V3.0-3.2磁盘异常

注意：本方法适用于V3.0至V3.2版本，可解决60%的磁盘显示问题。

V3.3开始将使用另一种磁盘管理模式，后继版本不要再参考本方法。

### 问题描述

正常使用的磁盘，显示异常，如下图，两个SAS盘，都多显示了一次，一次正常，一次异常。



### 解决方案

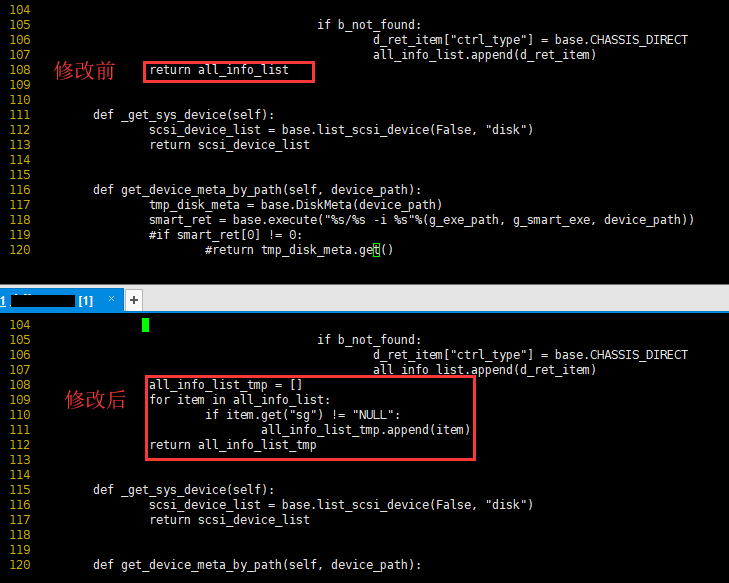
解决方案需每个节点操作。

备份配置文件：

# cd /usr/lib/python2.7/site-packages/infinity/system/disk/

# cp manager.py manager.py\_back

修改/usr/lib/python2.7/site-packages/infinity/system/disk/manager.py配置文件：



修改部分代码为：

all\_info\_list\_tmp = []

for item in all\_info\_list:

if item.get("sg") != "NULL":

all\_info\_list\_tmp.append(item)

return all\_info\_list\_tmp

移走旧的磁盘信息：

# cd /etc/dt.d/

# mv disk\_cache.db disk\_cache.db\_back

重启服务，重新生成磁盘信息：

# systemctl restart infi-api

# systemctl restart infi-monitor

# systemctl restart infi-cron

# systemctl restart infi-rpc

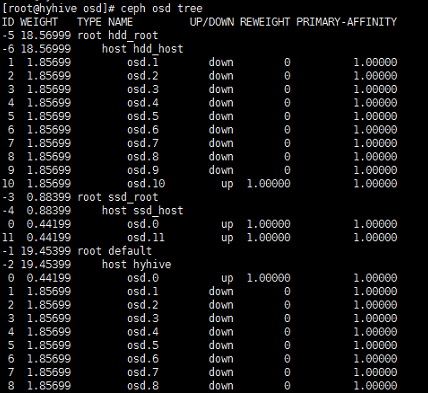
刷新界面，检查磁盘显示是否正常。

# ceph

## 3.1 V2.3版本，OSD无法挂载。

### 问题描述

存储无法访问，ceph状态异常，大量osd down

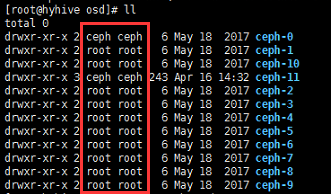


lsblk 查看磁盘，osd down对应磁盘，都未挂载；重启osd，无法自动挂载。

### 问题原因

挂载点对应目录，所属用户、所属用户组异常，异常原因未知，出问题前未有特殊操作。

正常情况所属用户为ceph，所属用户组为ceph。



### 解决方案

1)手动修改挂载点目录所属用户、所属用户组

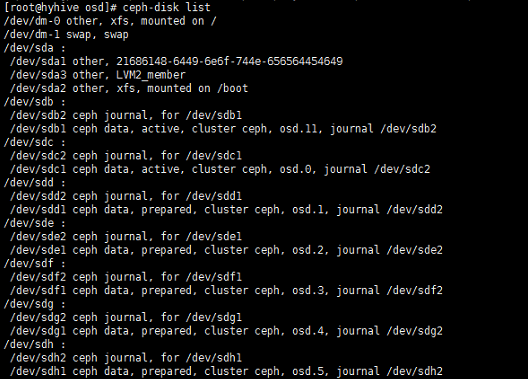
2)重启osd，查看是否自动挂载

3)如无法自动挂载，进行手动挂载

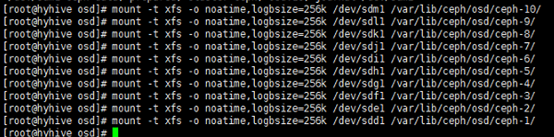
4)确认osd对应磁盘，命令为：

ceph-disk list

如下图：



5)确认挂载参数后，进行手动挂载



6)挂载完成，重启osd服务，服务启动成功，ceph状态恢复。

## 3.2 PG分布不均衡

注意：本节仅适用于infinity3.0至infinity3.6(ceph Luminous)版本，infinity3.7(ceph Nautilus)及以上版本勿用。

### 问题描述

在ceph的实际应用中，我们常常发现存在数据分布不均衡的情况，这个情况在存储池使用率达到70%以上后犹为明显。 在ceph出现了数据分布不均衡的时候，我们可以选择人工干预，通过调整PG分布来达到数据大致均衡。

在现场的环境中当某个osd使用率较高，空间满后，导致ceph集群状态HEALTH\_ERR。

现场环境中：

V3.0、V3.1、V3.2版本，若未打过osd内存限制补丁，在一个osd空间满后，该osd所属存储池将无法写入和无法读取数据。

V3.3至V3.6版本，在一个osd空间满后，该osd所属存储池将无法写入数据，但可以读取数据。

查看ceph状态如下：

[root@node1 ~]# ceph -s

cluster:

id: 3825da34-808e-48e3-865d-15e14ab46d66

health: HEALTH\_ERR

1 full osd(s)

7 pool(s) full

services:

mon: 1 daemons, quorum node1

mgr: node1(active)

mds: infinityfs1-1/1/1 up {0=node1=up:active}

osd: 14 osds: 14 up, 14 in

rgw: 1 daemon active

data:

pools: 9 pools, 680 pgs

objects: 2.23M objects, 12.3TiB

usage: 18.5TiB used, 3.64TiB / 22.1TiB avail

pgs: 680 active+clean

[root@node1 ~]#

ceph osd df 查看同一存储池里面的磁盘利用率相差很大，其中osd.5的利用率达到了99%：

[root@node1 ~]# ceph osd df

ID CLASS WEIGHT REWEIGHT SIZE USE AVAIL %USE VAR PGS

3 ssd 0.13599 1.00000 140GiB 1.00GiB 139GiB 0.72 0.01 128

10 ssd 0.13599 1.00000 140GiB 1.00GiB 139GiB 0.72 0.01 129

0 hdd 1.81898 1.00000 1.82TiB 1.59TiB 236GiB 87.32 1.05 149

1 hdd 1.81898 1.00000 1.82TiB 1.54TiB 286GiB 84.65 1.01 130

2 hdd 1.81898 1.00000 1.82TiB 1.59TiB 237GiB 87.26 1.04 127

4 hdd 1.81898 1.00000 1.82TiB 1.56TiB 267GiB 85.65 1.03 135

5 hdd 1.81898 1.00000 1.82TiB 1.80TiB 18.6GiB 99.00 1.18 149

6 hdd 1.81898 1.00000 1.82TiB 1.59TiB 239GiB 87.19 1.04 151

7 hdd 1.81898 1.00000 1.82TiB 1.54TiB 287GiB 84.59 1.01 129

8 hdd 1.81898 1.00000 1.82TiB 1.32TiB 508GiB 72.73 0.87 121

9 hdd 1.81898 1.00000 1.82TiB 1.44TiB 387GiB 79.24 0.95 135

··· ···

通过界面或 ceph osd tree确认osd5属于哪个存储池。

### 解决方案

1）首先确认环境的ceph版本是 luminous：

root@node1 ~]# ceph version

ceph version 12.2.8 (ae699615bac534ea496ee965ac6192cb7e0e07c0) luminous (stable)

从 luminous v12.2.z开始， osdmap中有一个新的 pg-upmap异常表，允许集群将特定的PG显式映射到特定的osd。

2）设置集群客户端使用L版本（一个节点上执行即可）

ceph osd set-require-min-compat-client luminous

或

ceph osd set-require-min-compat-client luminous --yes-i-really-mean-it

[root@node1 h]# ceph osd set-require-min-compat-client luminous

set require\_min\_compat\_client to luminous

3）获取当前osd map 信息：

ceph osd getmap -o {osdmap\_filename}

[root@node1 ~]# mkdir h

[root@node1 ~]# cd h/

[root@node1 h]# ceph osd getmap -o osd.map

got osdmap epoch 130

4）获取当前集群数据均衡后的优化信息

osdmaptool {osdmap\_filename} --upmap out.txt [--upmap-pool <pool>][--upmap-max n] [--upmap-deviation <max-deviation>]

注：

osdmap\_filename：osdmap的名称

--upmap：优化结果文件

upmap-pool：需要优化均衡的存储池名

upmap-max：指定一次优化的数据条目，默认100 ，可根据环境业务情况调整该值，一次调整的条目越多， 数据迁移会越多，可能对环境业务造成影响。

max-deviation：最大偏差值，默认为0.01(即1％)，该偏差值指的是osd利用率与平均值之间的差值。

[root@node1 h]# osdmaptool osd.map --upmap out.txt --upmap-pool fs\_data

osdmaptool: osdmap file 'osd.map'

writing upmap command output to: out.txt

checking for upmap cleanups

upmap, max-count 100, max deviation 0.01

limiting to pools filepool (1)

查看需要进行调整的PG，和新的osd映射关系：

[root@node1 h]# cat out.txt

ceph osd pg-upmap-items 1.1 5 8

ceph osd pg-upmap-items 1.2 0 8 5 11

ceph osd pg-upmap-items 1.8 5 8

ceph osd pg-upmap-items 1.10 5 8 6 11

ceph osd pg-upmap-items 1.15 5 11

ceph osd pg-upmap-items 1.18 5 8 2 9

ceph osd pg-upmap-items 1.1f 5 8

ceph osd pg-upmap-items 1.20 5 11

ceph osd pg-upmap-items 1.21 5 9

ceph osd pg-upmap-items 1.22 5 8

··· ···

每一条记录代表了1个PG映射 OSD ID的调整，调整映射就会使得数据进行迁移。

第一条信息中： ceph osd pg-upmap-items 1.1 5 8，PG1.1从osd5迁移到osd8。

5）均衡

执行均衡 source out.txt ：

[root@node1 h]# source out.txt

set 1.1 pg\_upmap\_items mapping to [5->8]

set 1.2 pg\_upmap\_items mapping to [0->8,5->11]

set 1.8 pg\_upmap\_items mapping to [5->8]

set 1.10 pg\_upmap\_items mapping to [5->8,6->11]

set 1.15 pg\_upmap\_items mapping to [5->11]

set 1.18 pg\_upmap\_items mapping to [5->8,2->9]

set 1.1f pg\_upmap\_items mapping to [5->8]

set 1.20 pg\_upmap\_items mapping to [5->11]

set 1.21 pg\_upmap\_items mapping to [5->9]

set 1.22 pg\_upmap\_items mapping to [5->8]

··· ···

运行数据分布调整操作，这一步将会使集群开始进行PG重新映射，同时集群数据开始迁移均衡 。

[root@node1 h]# ceph -s

cluster:

id: 3825da34-808e-48e3-865d-15e14ab46d66

health: HEALTH\_ERR

1 full osd(s)

7 pool(s) full

113103/6686299 objects misplaced (1.692%)

等到数据迁移完成后，再来观察磁盘的使用率，发现较调整前更均衡。

[root@node1 ~]# ceph osd df

ID CLASS WEIGHT REWEIGHT SIZE USE AVAIL %USE VAR PGS

3 ssd 0.13599 1.00000 140GiB 1.01GiB 139GiB 0.72 0.01 128

10 ssd 0.13599 1.00000 140GiB 1.01GiB 139GiB 0.72 0.01 129

0 hdd 1.81898 1.00000 1.82TiB 1.56TiB 261GiB 86.00 1.03 148

1 hdd 1.81898 1.00000 1.82TiB 1.54TiB 286GiB 84.65 1.01 130

2 hdd 1.81898 1.00000 1.82TiB 1.57TiB 259GiB 86.07 1.03 126

4 hdd 1.81898 1.00000 1.82TiB 1.56TiB 267GiB 85.65 1.02 135

5 hdd 1.81898 1.00000 1.82TiB 1.56TiB 262GiB 85.91 1.03 139

6 hdd 1.81898 1.00000 1.82TiB 1.56TiB 263GiB 85.87 1.03 150

7 hdd 1.81898 1.00000 1.82TiB 1.54TiB 289GiB 84.49 1.01 129

8 hdd 1.81898 1.00000 1.82TiB 1.49TiB 333GiB 82.13 0.98 128

9 hdd 1.81898 1.00000 1.82TiB 1.49TiB 336GiB 81.97 0.98 137

6）数据迁移完成后，ceph恢复HEALTH\_OK状态，若是V3.0、V3.1、V3.2版本，且未打osd内存限制补丁，需在此时打上补丁。

7）如果数据迁移完成后，发现osd空间使用率还是不够均衡并且相差较大，可以备份之前的osd.map和out.txt，然后按照之前的方式重新获取osd.map和out.txt，后再次进行PG分布调整。

注意：再次调整需要等待上次调整的数据迁移完成。

后续处理：

1）让客户清理不需要的数据

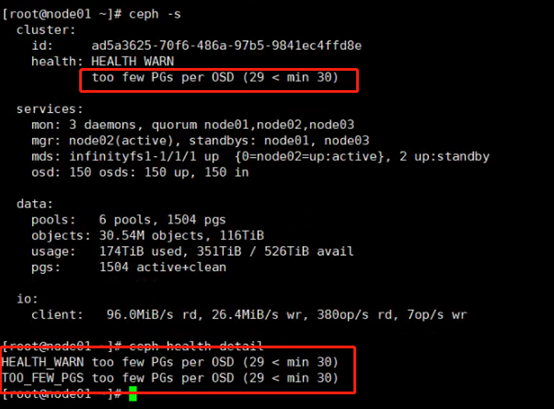
2）建议扩容

## 3.3 osd的pg数过低

注意：本节仅适用于infinityV2至infinityV3.6版本，infinityV3.7及以上版本勿用。

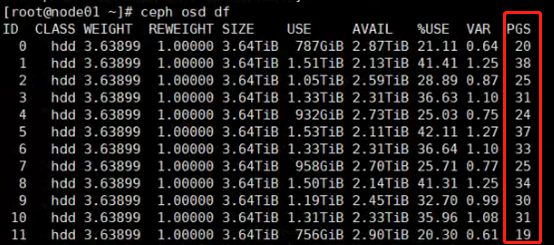
### 问题描述

V3扩容节点扩容完成磁盘组后，ceph出现如下告警：



### 问题原因

出现该告警是因为我们有osd的pg数小于30，如下图：



出现该情况，可能会造成osd性能、数据均衡，等方面影响。

所以在现场，对所有节点扩容存储池完成后，并且数据迁移完成后，我们需要进行调整pg数量。

### 解决方案

#### 1）计算pg\_num和pgp\_num

infinityV2版本pg\_num和pgp\_num计算公式：

pg\_num = osd\_size \* 128 / size，结果取整；

其中 osd\_size表示对应存储池所有osd的个数；

size 表示冗余份数，副本存储池 size 取副本份数，纠删码存储池 size = k+m，例如三幅本存储池 size=3，纠删码6+2存储池 size=6+2=8。

pgp\_num = pg\_num

infinityV3版本pg\_num和pgp\_num计算公式：

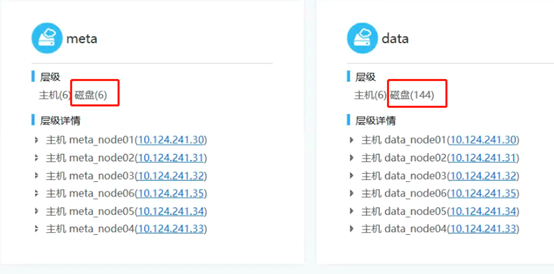
pg\_num = osd\_size \* 64 / size，结果取整；

其中 osd\_size 表示对应存储池所有osd的个数；size表示冗余份数，副本存储池 size 取副本份数，纠删码存储池 size = k+m ，例如三幅本存储池 size=3，纠删码6+2存储池 size=6+2=8。

pgp\_num = pg\_num

例如：

上面出现告警的环境(V3环境)，查看相关信息，如下图：





由上面两张图可知(上图为V3环境)：

meta池，osd个数6，数据保护策略3副本，则

pg\_num = 6 \* 64 / 3 = 128

pgp\_num = pg\_num = 128

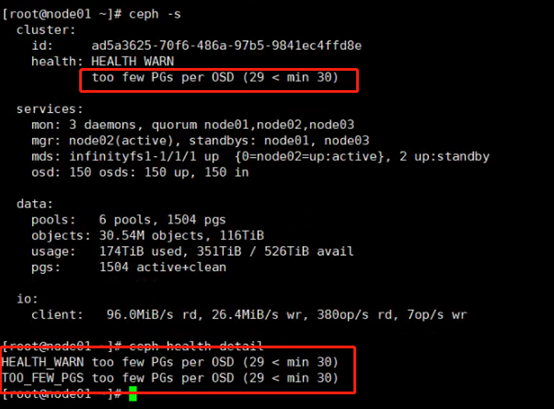
data池，osd个数144，数据保护策略纠删码2+1，则

pg\_num = 144 \* 64 / 3 = 3072

pgp\_num = pg\_num = 3072

#### 2）调整pg\_num和pgp\_num

前置条件：ceph为health\_OK状态，或除了osd上pg个数少于30的告警外，无其他告警，如下图：



判断是否分多次调整：

* 刚才我们计算出pg\_num和pgp\_num需要调整为：

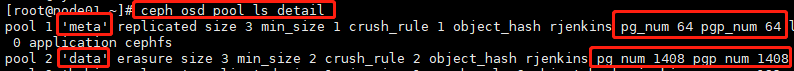
meta池：pgp\_num = pg\_num = 128

data池：pgp\_num = pg\_num = 3072

* 查看为调整前存储池的pg\_num和pgp\_num，命令：

ceph osd pool ls detail

如下图



* 判断元数据池：需要调整的pg\_num和pgp\_num未超过500，可以直接一次性调整。
* 判断数据池：

如果新计算出来的pg\_num超过当前pg\_num的2倍时，分多次调整pg\_num：一次调整pg\_num的数量不超过需要调整的pg\_num总数的30%，分多次调整到计算出来的pg\_num和pgp\_num

（此次目的: 防止扩容过多节点时一次需要调整pg\_num和pgp\_num的数量太多）

例如：上述环境meta池pg\_num和pgp\_num需要调整为128，小于500，可以一次调整到500；

上述环境data池pg\_num和pgp\_num需要调整为3072，当前值1408，超过2倍多一点点，可以一次调整到3072。

假设：存储池需要调整为4096，当前值1024，则第一次调整为：1024+4096 \* 30%=2253；

等待数据迁移完成后，第二次调整为：2253+4096 \* 30%=3482；

等待数据迁移完成后，第三次调整为4096。

* 上述分次调整方法如有无法判断的请联系售后和研发。

#### 3）调整步骤

设置完成后，会有数据迁移，尽量在无业务的时候开始执行。

调整命令：

#ceph osd pool set {pool\_name} pg\_num {pg\_num}

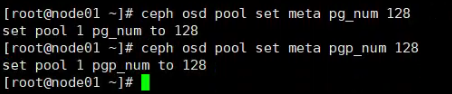
#ceph osd pool set {pool\_name} pgp\_num {pgp\_num}

{pool\_name}：需要调整的具体存储池名称

{pg\_num}：对应存储池计算出来的pg\_num

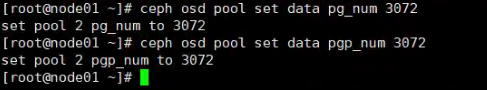
{pgp\_num}：对应存储池计算出来的pg\_num

我们先调整元数据池，因为元数据池恢复速度很快，pg\_num和pgp\_num一起调整，如下图：



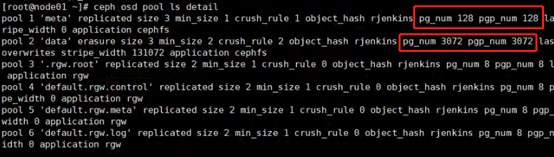
观察ceph数据迁移状态，等待数据迁移完成。

数据迁移完成后，开始调整数据池，pg\_num和pgp\_num一起调整，如下图：



如果存储池是分多次调整的情况，则需要等待一次调整的数据迁移完成后，再进行下一个阶段调整。

所有存储池调整完成后，确认存储池pg\_num和pgp\_num正常，如下图，ceph恢复到HEALTH\_OK状态



数据迁移过程中，可根据业务情况，调整数据恢复速度，使无业务的时候恢复更快一些。

#### 4）注意事项

pg\_num和pgp\_num只能调大，不能调小。

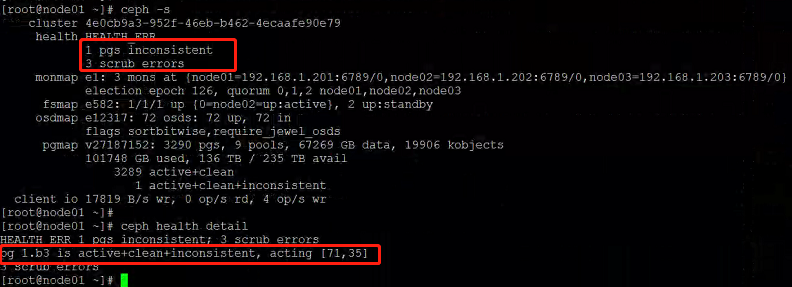
调整pg\_num和pgp\_num后，会有数据迁移，会影响性能，建议在没业务的时候调整或者业务量小的时候调整。

多次调整 pg\_num和pgp\_num时，等上一次调整恢复完成后，进行下一次调整。

## 3.4 pg\_inconsistent

### 问题描述

我们在现场ceph环境中偶尔会遇到告警，如下图：



pg\_inconsistent 即pg不一致，出现该情况的大多原因：ceph有osd出现慢io时其他节点出现了osd down，或多个osd长时间处于慢io状态，等情况。

### 解决方案

1）ceph osd tree 确认集群中所有osd都是在线状态，如有多个节点或多个osd down，请联系研发处理

2）如果有数据降级，需先等待降级数据恢复完成，再进行pg inconsistent的处理

3）如上图，重启pg 1.b3 的主从osd71和35，后观察pg是否恢复，若未恢复进行下一步

4）修复pg

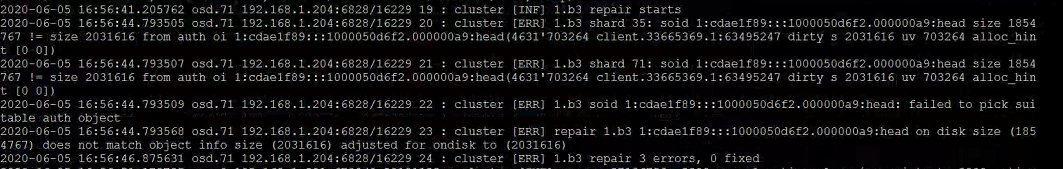
# ceph pg repair <pg\_id>

# ceph pg repair 1.b3

修复时打印日志

tailf /var/log/ceph/ceph.log

查看修复情况，如下图：



5）repair完成后，进行pg校验，如下，查看ceph.log确认校验完成后，确认ceph状态

# ceph pg deep-scrub <pg\_id>

# ceph pg deep-scrub 1.b3

6）若pg\_inconsistent还是未恢复成功，请联系研发处理。

## 3.5 机器异常断电后，有HDD无法拉起

### 可能原因

断电时磁盘写缓存没有完全刷下去。

### 解决方案

## 3.6 ceph告警一个osd full，存储可读不可写

### 可能原因

pg分布不均衡；

集群存储空间不足

### 解决方案

V4.0前的版本需在集群管理中均衡存储pg，若多次均衡后依旧告警osd full，则需检查集群的可用存储空间，若可用空间不足，需清理不需要的数据，或扩容。

V4.0后的版本产品提供了自动均衡，若告警osd full，则检查集群的可用存储空间，空间不足则清理不需要的数据或扩容。均衡pg方法请参照3.2.

## 3.7 ceph有pg恢复和大量慢io

### 可能原因

删除了大量数据，导致磁盘利用率高.

### 解决方案

等待数据清理完成即可。

## 3.8 ceph-mon服务异常，文件系统无法正常使用

### 可能原因

系统盘空间满导致ceph-mon异常。

### 解决方案

清理空间后重启ceph-mon服务。

## 3.9 从界面踢出磁盘，更换新盘后，从界面添加磁盘的窗口没有显示新盘

### 可能原因

新盘有问题；

/var/datatom/osd\_disk.db数据库没有清理踢出的磁盘。

### 解决方案

节点管理中查看新盘SMART是否正常，若正常，则后台检查osd\_disk.db数据库中是否清理踢出的旧盘，若未清理，则手动清理osd\_disk.db数据库。

## 3.10 有元数据盘down/无法up。

### 可能原因

元数据盘故障；

若集群中有文件系统，可能是BlueFS日志膨胀导致。如何检查？

### 解决方案

检查元数据盘是否可用，重新插拔元数据，若元数据盘正常，则...

# ctdb

4.1

* 可能原因
* 解决方案

# NAS

## 5.1 NAS状态“初始化/不健康”，ctdb日志中报错

### 问题描述

NAS状态“初始化/不健康”，ctdb日志中报错ERROR:nfs directory "/infinityfs1/infinityfs1" not available

### 可能原因

客户从PC端删除了该目录，导致NFS服务异常。

### 解决方案

检查 /etc/exports的nfs共享配置中是否有/infinityfs1/infinityfs1路径的共享，若有，检查该路径是否存在，若不存在，则删除/infinityfs1/infinityfs1路径的NFS共享，重启NAS服务。

## 5.2 异常断电后，NAS服务启动不正常，winbind服务异常

### 可能原因

未挂载infinityfs1文件系统。

### 解决方案

挂载文件系统后，重启NAS服务。

## 5.3 NFS&CIFS配置在同一目录时，NFS端传入数据，CIFS端无法删除

### 可能原因

NFS写入文件的owner是root，smb写入文件的owner是smb用户，导致nfs写进来的文件smb没法读写

### 解决方案

V4.0及以后版本可设置新文件默认权限为777，V4.0前的版本需后台添加root为samba用户

## 5.4 客户端挂载nfs失败

（其他共享挂载失败除可能原因1外，排除方法同理）

### 可能原因

1）非法端口。

2）nfs目标目录已被删除。

3）挂载命令中的挂载地址输入错误。

### 解决方案

1）在nfs配置文件中添加insecure（insecure表示nfs通过1024以上的端口发送，我们产品使用的nfs端口是2049），若已是insecure则查看可能原因2｡

2）在文件系统中查看nfs共享的目标目录是否存在，若不存在，则重新创建nfs共享并指定存在的目标目录｡若存在，则查看可能原因3。

3）输入正确的挂载命令。

mount -t nfs 公共IP:共享目录的路径 挂载路径

## 5.6 ftp共享丢失

### 问题描述

为test1目录创建ftp共享，选择所有用户，再为test2目录创建ftp共享，选择所有用户，创建完成后，test1目录的ftp共享丢失

### 可能原因

ftp共享原理如此，同一个用户只能拥有一个共享目录，为已有共享的用户再创建共享时，原有共享设置会被覆盖。ftp共享目录是ftp用户登陆的home目录，只能指定一个，如果存在多个，只有第一个有效。

### 解决方案

同一个用户仅指定一个ftp共享目录。

## 5.7 linux系统无法挂载infinityV3的smb共享

前置条件：客户端、权限、共享目录路径均配置正确。

### 可能原因

客户端未安装cifs-utils包。

挂载源路径填写错误，挂载命令中应为共享名称。

### 解决方案

安装cifs-utils包。

挂载命令中IP后应为共享名称：

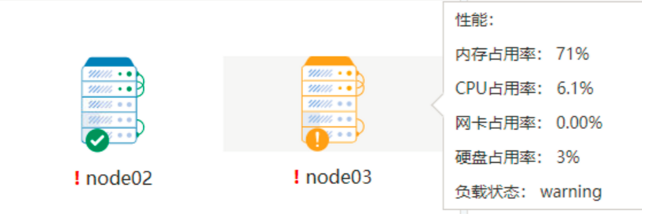
# mount -t cifs -o username=hcm,password=123456,iocharset=utf8,vers=3\.0 //10.0.7.211/CIFS\_share1 /mnt/Windows

# 其他

## 6.1 调整负载状态颜色

### 问题描述

在V3.0-3.6版本的环境中，16盘位、24盘位、36盘位，我们在限制osd内存后，内存使用情况基本都会跑到限制的标准，首页上的负载状态会显示成黄色。针对这种我们平台正常占用内存，但界面显示为黄色的状态，有的客户比较在意时，可以优化负载状态内存使用颜色变化的百分比。



infinity3.7版本开始将调整内存黄色告警阈值为80%。infinity3.6以前版本，如客户比较在意，可进行手动修改。

### 解决方案

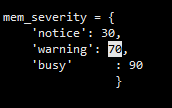
备份配置文件：

cp /opt/infinity/python/apps/cluster/stats/sys\_percent.py /opt/infinity/python/apps/cluster/stats/sys\_percent.py\_bak

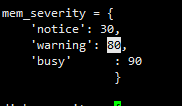
修改配置：

/opt/infinity/python/apps/cluster/stats/sys\_percent.py

默认的节点内存黄色警告值是70：



将节点内存黄色警告值改为80：



修改后重启服务：infi-api，infi-rpc，dt-rpc。