新华三技术有限公司 New H3C Technologies Co.,Ltd.	产品名称 Product Name	
	H3C UniStor X10000	
	产品版本 Product Version	共 20 页 20Pages in all
	NAS 2.0	

H3C UniStor X10000 系列存储（NAS 2.0）用户关机指导书

数字化解决方案领导者

New H3C Technologies Co., Ltd.

新华三技术有限公司

Revision Record 修订记录

Date 日期	Revision Version 修订版本	Change Description 修改描述	Author 作者
2019-12-15	V1.0	第一次发布	
2020-01-15	V1.1	修改3.3章节维护模式	

安全声明

IMPORTANT! See Compliance and Safety information for the product before connecting to the supply. To obtain Compliance and Safety information, go to

http://www.h3c.com

重要！在产品上电启动之前，请阅读本产品的安全与兼容性信息。您可以通过以下步骤获取本产品的安全与兼容性信息：

（1） 请访问网址：http://www.h3c.com/cn/Technical_Documents;

（2） 选择产品类型以及产品型号；

（3） 您可以从安全与兼容性手册中获取安全与兼容性信息。

环境保护

本产品符合关于环境保护方面的设计要求，产品的存放、使用和弃置应遵照相关国家法律、法规要求进行。

技术支持

- 用户支持邮箱：service@h3c.com
- 技术支持热线电话：400-810-0504（手机、固话均可拨打）
- 网址：http://www.h3c.com

目录

1 概述

1.1 文档使用范围

本文档（指南）主要讲解 H3C Unistor X10000 （NAS 2.0）正确的关机操作步骤，NAS 2.0 包括的版本为：E12XX 版本和 R12XX 的所有版本。

本文档仅适用于文件存储场景，如果关机场景涉及到块存储和对象存储，请参考对应指导书。

说明

版本号查看方法：

1. 登录handy界面，点击右上角“i”图标查看；
2. 如果handy节点因硬件故障、网络等原因暂时无法登录，可登录存储集群任意节点的系统后台，使用命令：**cat /etc/onestor_external_version** 命令查看对应版本号。

1.2 使用注意事项

本文档主要讲解 H3C UniStor X10000 系列存储（NAS 2.0）正确的关机步骤。操作前，请仔细阅读 H3C 官方文档避免出现技术风险。

本文档不定时更新，使用前请访问 H3C 官网下载最新版本或者联系 H3C 400 技术支持工程师获取当前最新版本。

执行本文档相关操作前，请仔细阅读本文档内容和相关其他文档，包括但不限于《H3C UniStor X10000 G3 系列存储用户指南》、《H3C UniStor X10000 G3 系列存储部件安装&更换视频》、《H3C Unistor X10000 G3 系列存储配置指导》等。

为确保数据安全、业务稳定，H3C 建议您在相关操作变更前备份重要数据及相关配置信息，选择业务量小的场景或者停业务等维护时间窗口期进行变更操作。如有疑问，请及时联系 H3C 400 工程师获取相应的技术支持。

如果您继续阅读并按照本文档以下步骤进行相关操作，说明您对本文档的文档适用范围及使用注意事项章节已有充分的理解，并不存在任何歧义；同时也表明您已经完全知悉并接受任何可能存在的潜在风险。

1.3 读者对象

本文档（指南）主要适用对象如下：

- 技术支持工程师
- 运维工程师

2 文档使用流程图

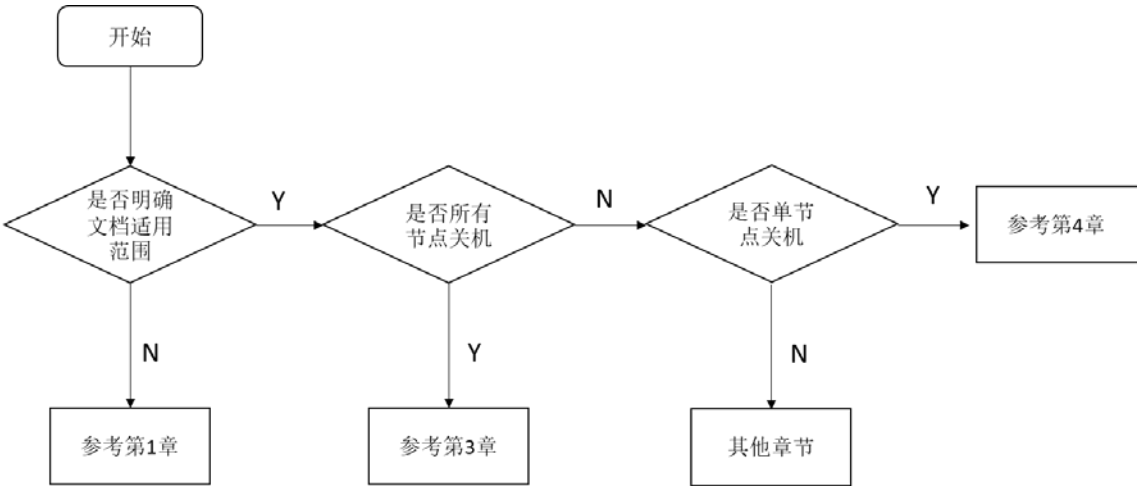


图 1 文档使用流程图

3 集群所有节点关机场景

本章节适用于停止所有业务，集群内所有节点需要关机的场景。

3.1 关机前相关检查

3.1.1 检查集群健康状态

1、登陆 Handy 页面，在“概览”页面，确认集群健康度为 100%，且右上角无告警。若集群健康度不为 100%，或集群有告警，请等待集群自动恢复或排除故障后再操作。若等待一段时间仍然没有恢复进度，则拨打 400 获取帮助。

图 2 确认集群健康度及右上角告警信息

2、在集群中任意节点后台执行 watch ceph -s 持续观察集群健康状态，正常情况下状态为Health_OK。观察一分钟左右，确认健康状态正常。若健康状态不为 Health_OK，请拨打 400 热线确认。

```
[root@node12 ~]# ceph -s
  cluster:
    id:     11943d3b-84cf-47da-af77-6426b20468e2
    health: HEALTH_OK

  services:
    mon: 3 daemons, quorum node12,node13,node14
    mgr: node12(active), standbys: node13, node14
    mds: CAPFS-1/1/1 up  {0=mds0=up:active}, 2 up:standby
    osd: 48 osds: 48 up, 48 in

  data:
    pools:   2 pools, 2304 pgs
    objects: 28 objects, 31724 bytes
    usage:   52427 MB used, 155 TB / 155 TB avail
    pgs:     2304 active+clean

  io:
    client:   3149 B/s rd, 0 B/s wr, 3 op/s rd, 0 op/s wr
```

图 3 后台确认集群健康状态

3.1.2 检查交换机配置

确认存储交换机、业务交换机是否开启 STP，如果开启 STP，检查确认连接服务器的端口已经配置为边缘端口；如果未开启 STP，则可忽略此项检查。具体检查方法请参考具体型号的交换机的命令手册。**注：交换机配置变更请联系 400 确认后再操作。**

3.1.3 检查主机路由信息

在所有主机上执行 route -n 检查主机路由信息。

```
[root@node17 network-scripts]# route -n
Kernel IP routing table
Destination     Gateway         Genmask         Flags Metric Ref    Use Iface
172.16.3.0      0.0.0.0         255.255.255.0   U     0      0        0 ethA01-0
172.16.4.0      0.0.0.0         255.255.255.0   U     0      0        0 bond0
172.16.5.0      0.0.0.0         255.255.255.0   U     0      0        0 bond1
192.168.1.0     172.16.3.254    255.255.255.0   UG    0      0        0 ethA01-0
```

图 4 检查主机路由信息

在所有主机上执行 cat /etc/sysconfig/network-scripts/route-ethxx（ethxx 为该条路由对应的网口，按实际情况修改），查看网络配置文件中是否有相应的路由配置。

```
[root@node17 network-scripts]# cat /etc/sysconfig/network-scripts/route-ethA01-0
192.168.1.0/24 via 172.16.3.254
```

图 5 查看路由配置文件

如果没有，需要将相应的路由配置信息写入/etc/sysconfig/network-scripts/route-ethxx 配置文件。若节点上没有该文件，则需要手动创建。

3.1.4 检查 RAID 卡电池状态

在服务器的硬件管理页面（如 HDM，iLO 等）查看阵列卡电池状态，确认电池状态正常且处于充满电状态。

图 6 检查阵列卡电池状态

3.1.5 检查硬盘缓存

以下操作在存储集群中每台服务器的后台执行。如果检查结果与预期不符，请查阅维护指导书修改或者联系 400 处理。

1. X10516 G1/X10529 G1 机型

- 检查硬盘写缓存是否关闭：执行 arcconf getconfig 1 pd | grep -i “write cache”，所有的输出结果应为 Disabled (write-through)。

```
root@cvknode3:~# arcconf getconfig 1 pd|grep -i "write cache"
         Write Cache                          : Disabled (write-through)
         Write Cache                          : Disabled (write-through)
         Write Cache                          : Disabled (write-through)
         Write Cache                          : Disabled (write-through)
         Write Cache                          : Disabled (write-through)
         Write Cache                          : Disabled (write-through)
         Write Cache                          : Disabled (write-through)
```

图 7 X10516 G1/X10529 G1 检查硬盘缓存

- 检查所有阵列卡读写缓存是否开启并设置为掉电保护模式。执行 arcconf getconfig 1 ld 查询。

```
Logical Device number 5
   Logical Device name                      : LogicalDrv 5
   Block Size of member drives              : 512 Bytes
   RAID level                               : Simple_volume
   Unique Identifier                        : CF9655AD
   Status of Logical Device                 : Optimal
   Size                                     : 953334 MB
   Parity space                             : Not Applicable
   Read-cache setting                       : Enabled
   Read-cache status                        : On
   Write-cache setting                      : On when protected by battery/ZMM
   Write-cache status                       : On
   Partitioned                              : Yes
   Bootable                                 : No
   Failed stripes                           : No
   Power settings                           : Disabled
   ------------------------------------------------------
   Logical Device segment information
   ------------------------------------------------------
   Segment 0                                : Present (953869MB, SATA, HDD, Enclosure:0, Slot:3)           ZBS0A02F
```

图 8 X10516 G1/X10529 G1 检查掉电保护模式

2. X10536 G1/X10326 G1/ X10360 机型

- 检查硬盘写缓存是否关闭：hpssacli ctrl all show config detail | grep -i cache

```
root@node118:~# hpssacli ctrl all show config detail | grep -i cache
   Cache Serial Number: PBKUD0BRH7P2XI
   Wait for Cache Room: Disabled
   Cache Board Present: True
   Cache Status: OK
   Cache Ratio: 10% Read / 90% Write
   Drive Write Cache: Disabled
   Total Cache Size: 2.0 GB
   Total Cache Memory Available: 1.8 GB
   No-Battery Write Cache: Disabled
   Cache Backup Power Source: Capacitors
   Cache Module Temperature (C): 37
         LD Acceleration Method: Controller Cache
         LD Acceleration Method: Controller Cache
         LD Acceleration Method: Controller Cache
         LD Acceleration Method: Controller Cache
```

图 9 X10536 G1/X10326 G1/X10360 检查硬盘缓存

未做过特殊调整的情况下，Cache Ratio 应为 10%读，90%写；Drive Write Cache 应为 Disabled；No-Battery Write Cache 应为 Disabled。

- 检查各阵列的缓存模式设置是否正确：hpssacli ctrl slot=n ld all show detail（其中 n 为阵列卡槽位号，请按照实际情况修改）

  对于 HDD，LD Acceleration Method 应为 Controller Cache。

  对于 SSD，LD Acceleration Method 应为 Disabled 或 Smart IO Path。

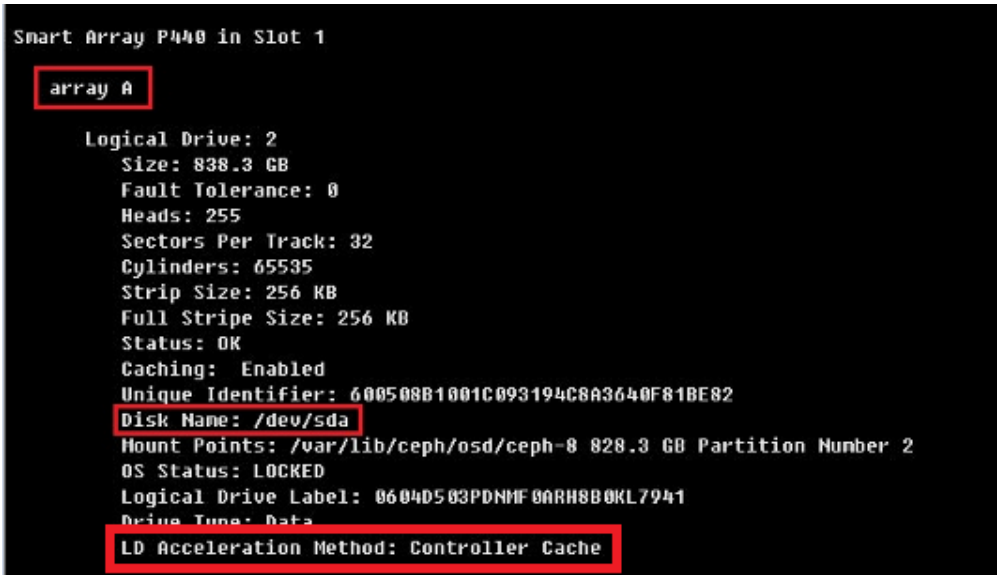


图 10 X10536 G1/X10326 G1/X10360 检查阵列卡缓存模式

3. X10516 G3/X10529 G3 /X10326 G3 /X10536 G3 机型

- 检查硬盘写缓存是否关闭：arcconf getconfig 1 ad|grep " Physical Drive Write Cache Policy Information" -A4（1 为阵列卡槽位号，按实际情况修改），查出来的三项缓存状态都为 Disabled 表示正常

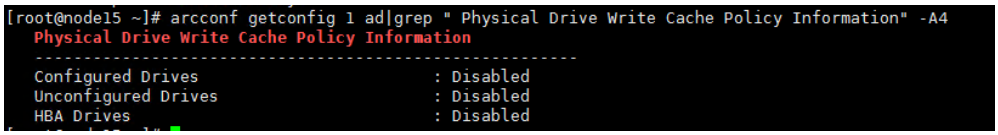


图 11 G3 机型检查硬盘缓存

- 查看逻辑盘的读写缓存的比例，推荐读写比例为 1:9：

  arcconf getconfig 1 ad|grep ‘Cache Properties’ -A6 （1 为阵列卡槽位号，按实际情况修改）

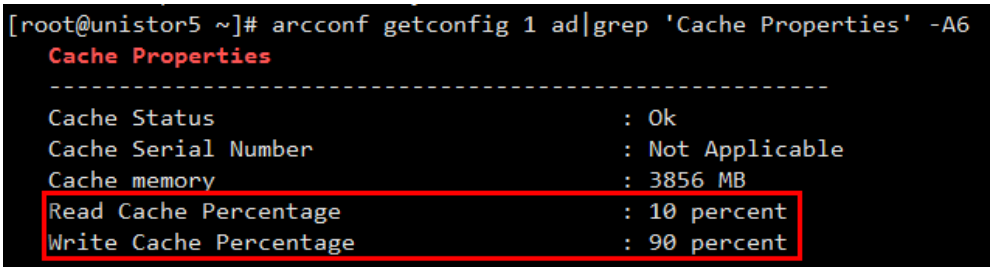


图 12 G3 机型检查逻辑盘读写缓存比例

- 检查各阵列的缓存模式设置是否正确：arcconf getconfig 1 ld （1 为阵列卡槽位号，按实际情况修改），查看 Caching 字段，应为 Enabled 状态。

3.1.6 检查集群硬件状态

登录集群中所有节点的 HDM/iLO，检查是否有硬件报错。确保关机前所有硬件状态正常。

3.1.7 检查 NTP 时钟

在所有节点执行 ntpq -p 检查，所有节点应该指向同一个 ntp server，ntp server 的状态不为 INIT，且 offset 值在 100ms 以内；

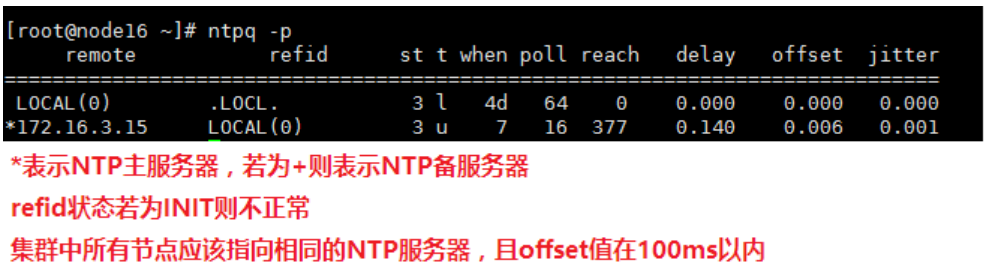


图 13 检查 NTP 时钟

若 NTP 状态不符合预期，请联系 400 确认。

3.2 停止上层业务

1. 如有重要数据，建议备份重要数据。

2. 记录相关客户端挂载目录等信息。

3. 断开与存储相连的所有客户端，避免有客户端在关机期间对存储仍有读写请求，从而出现被动影响业务的情况发生。

断开客户端与 NAS 存储连接后，登录 handy 界面，在 NAS 组里面查看 NFS、CIFS 和 FTP 的连接数，如果都显示为 0，则说明所有连接已断开。

图 14 确认断开客户端与 NAS 存储连接

3.3 开启维护模式

注意

1. Unistor-E1220P13（含）之前版本，在关机操作前**禁止使用维护模**式功能，禁止使用**ceph osd set noup**命令，仅使用**ceph osd set noout**即可；
2. 如果配置了Handy HA，则先关闭非工作Handy节点，再关闭当前工作节点；登录handy管理界面，点击左边导航栏的“管理高可用”查看。

如下两种方法任选其一，推荐选择 handy 界面的方法：

1. Handy 界面操作（注：所有待关机节点都要操作）

登录 Handy 页面，在 Handy 页面点击主机管理→待关机主机→更多→维护模式。

图 15 Handy 页面开启维护模式（a）

在弹出的对话框中，点击“开启”，然后点击“确定”，确认开启维护模式。

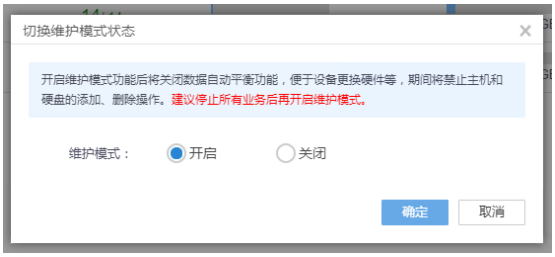


图 16 Handy 页面开启维护模式（b）

2. 后台操作方式（任意节点执行即可）

ssh 登录集群中任意节点的后台，执行 **ceph osd set noout** 和 **ceph osd set noup**。执行完毕后，使用 ceph -s 检查，状态变为 Health_WARN，且提示 noout，noup flags set。

```
root@cvknode3:~# ceph -s
    cluster 74461020-d984-4888-aee4-dbb50d64f487
     health HEALTH_WARN
            noup,noout flag(s) set
     monmap e1: 3 mons at {cvknode1=172.16.4.9:6789/0,cvknode2=172.16.4.10:6789/0,cvknode3=172.16.4.11:6789/0}
            election epoch 38, quorum 0,1,2 cvknode1,cvknode2,cvknode3
     osdmap e118: 12 osds: 12 up, 12 in
            flags noup,noout
      pgmap v5852: 1024 pgs, 1 pools, 4541 MB data, 1433 objects
            12908 MB used, 11033 GB / 11046 GB avail
                1024 active+clean
```

图 17 后台确认已开启维护模式

3.4 正常关机下电

注意

1. 如果配置了Handy HA（管理高可用），则先关闭非工作Handy节点，再关闭当前工作节点。Handy HA配置查看方法：登录handy管理界面，点击左边导航栏的“管理高可用”查看，如果该界面为空则代表未配置Handy HA。如图为配置了Handy HA的图示：

2. 为避免同时关机造成机房电压波动等异常情况，建议逐一正常关闭服务器；
3. 若需要下电操作，请确认操作系统完全关闭（电源开关灯由绿色变为黄色）后，再拔除电源线。

在待关机节点执行如下操作：

1．执行 sync，将内存下刷；

2．执行 hwclock -w，将时钟写入 BIOS；

3．使用 date&hwclock 命令观察输出时间是否一致，如不一致再次执行第 2 步命令并再次检查；

4．执行 shutdown -h now，将服务器正常关机。关机过程中建议关注 HDM 页面电源状态，避免出现关机命令执行失败或关机命令执行卡住的情况。

5．等待当前节点正常关闭后，再重复以上步骤关闭集群内其他节点，建议逐一操作。

3.5 开机操作

3.5.1 上电开机并检查网络

1. 将服务器上电，如果配置了 Handy HA，则先将主节点开机，再将备节点开机；
2. 等待所有服务器开机后，执行 ip addr 检查集群中各节点的管理网、业务网、存储前端网、存储后端网网口是否都为 up 状态；

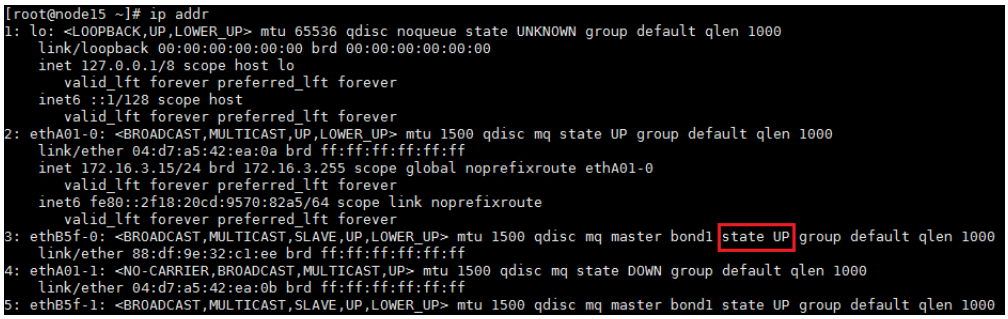


图 18 检查网络均为 up

3. 使用命令 ethtool bond 名检查 bond 聚合口 speed 是否为 20000Mb/s，所有 bond 口都需检查。

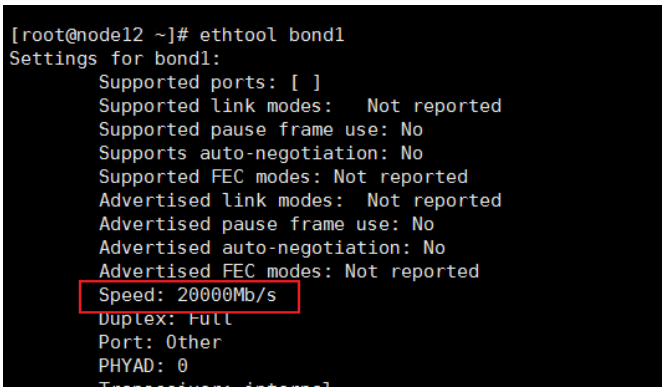


图 19 检查 bond 正常

4. 检查网络连通性：包括集群中所有节点的各网段两两互 ping 能否 ping 通，客户端到集群所有节点的业务网能否 ping 通。

3.5.2 检查 NTP 状态

检查 NTP 状态：在所有节点执行 ntpq -p 检查，所有节点应该指向同一个 ntp server，ntp server 的状态不为 INIT，且 offset 值在 100ms 以内；检查时钟是否有差异，如有差异使用命名 ntpdate -u IP（此处的 IP 是集群 ntp server 的 ip），等待时间同步完成。

```
[root@node16 ~]# ntpq -p
     remote           refid      st t when poll reach   delay   offset  jitter
==============================================================================
 LOCAL(0)        .LOCL.           3 l   4d   64    0    0.000    0.000   0.000
*172.16.3.15     LOCAL(0)         3 u    7   16  377    0.140    0.006   0.001
```

*表示NTP主服务器，若为+则表示NTP备服务器

refid状态若为INIT则不正常

集群中所有节点应该指向相同的NTP服务器，且offset值在100ms以内

图 20 检查 NTP 状态

3.5.3 关闭维护状态

1. 如果之前选择在前台开启的维护模式（注意关闭所有节点的维护模式），需要登录 handy 界面取消维护模式。点击主机管理→选中主机→更多→维护模式；

图 21 关闭维护模式（a）

在弹出的对话框中，点击“关闭”，然后点击“确定”，确认关闭维护模式。

图 22 关闭维护模式（b）

2. 如果之前选择的是在后台开启维护模式需要 ssh 登录集群中任意正常节点的后台，执行 **ceph osd unset noout** 和 **ceph osd unset noup**；
3. 执行 **ceph osd tree**，查看当前节点的 osd 状态是否全部变为 up 状态；
4. 如果在第 2 步检查时发现有 osd 未恢复为 up，在 osd 未 up 的节点，执行命令 **ceph-disk activate-all** 将 osd 拉起。然后再次执行 **ceph osd tree**，检查 osd 是否变为 up。如果仍有 osd 处于 down 的状态，执行 **systemctl restart ceph-osd@id.service**（其中 id 为 down 的 osd 编号），如果仍旧无法拉起 down 的 osd，请设置 **ceph osd set noout**，**并立即联系 400 处理。**

3.5.4 检查集群状态

登录 Handy 界面，持续观察集群健康度，直到集群健康度恢复 100%且所有告警消除。

3.5.5 恢复业务

持续观察存储没有异常告警后，按照关机前记录的业务连接情况恢复业务。为避免所有业务同时上线对集群造成瞬时冲击，建议有计划的进行业务恢复操作。

4 集群内单节点关机场景

本章节适用于在线情况下的单节点关机操作。带业务情况下，一次最多只能关机一个节点。若有多个节点需要关机，则需要停业务操作。

在执行在线单节点关机的过程中，不要有新业务连接到存储集群。

4.1 关机前相关检查

4.1.1 检查集群健康状态

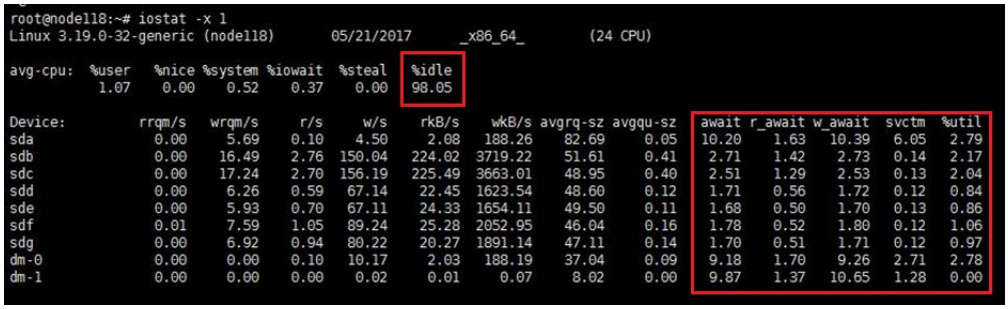
请参考章节 3.1.1。

4.1.2 检查集群业务压力

1. 检查 iostat 状态

ssh 到集群中**所有主机**的后台命令行，执行 iostat -x 1 持续观察所有节点的 CPU 使用率和磁盘压力，该命令会每 1s 刷新输出 iostat，建议每台主机观察 2min 左右。空闲的 CPU %idle 应该在 50 以

上；%util（磁盘 IO 繁忙度）需在 40%以下；svctm（平均每次 IO 请求的处理时间）需在 20 以下(单位为 ms)；await（平均 IO 等待时间）和 r_await（平均读操作等待时间） 以及 w_await（平均写操作等待时间）需在 20 以下（单位为 ms）。如果偶有超过上限的情况，属于正常现象，但如果持续保持在上限以上，则需要等待业务压力变小或暂停部分业务，直到集群业务压力满足条件。



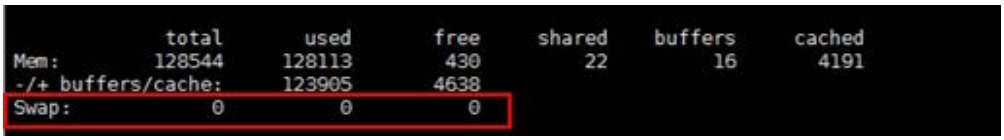
```
root@node118:~# iostat -x 1
Linux 3.19.0-32-generic (node118)       05/21/2017      _x86_64_        (24 CPU)

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           1.07    0.00    0.52    0.37    0.00   98.05

Device:         rrqm/s   wrqm/s     r/s     w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await r_await w_await  svctm  %util
sda               0.00     5.69    0.10    4.50     2.08   188.26    82.69     0.05   10.20    1.63   10.39   6.05   2.79
sdb               0.00    16.49    2.76  150.04   224.02  3719.22    51.61     0.41    2.71    1.42    2.73   0.14   2.17
sdc               0.00    17.24    2.70  156.19   225.49  3663.01    48.95     0.40    2.51    1.29    2.53   0.13   2.04
sdd               0.00     6.26    0.59   67.14    22.45  1623.54    48.60     0.12    1.71    0.56    1.72   0.12   0.84
sde               0.00     5.93    0.70   67.11    24.33  1654.11    49.50     0.11    1.68    0.50    1.70   0.13   0.86
sdf               0.01     7.59    1.05   89.24    25.28  2052.95    46.04     0.16    1.78    0.52    1.80   0.12   1.06
sdg               0.00     6.92    0.94   80.22    20.27  1891.14    47.11     0.14    1.70    0.51    1.71   0.12   0.97
dm-0              0.00     0.00    0.10   10.17     2.03   188.19    37.04     0.09    9.18    1.70    9.26   2.71   2.78
dm-1              0.00     0.00    0.00    0.02     0.01     0.07     8.02     0.00    9.87    1.37   10.65   1.28   0.00
```

图 16 检查 iostat 状态

2. 检查内存使用率

ssh 到集群中**所有主机**的后台命令行，执行 free -m 检查内存使用率，需要满足内存使用率在 60%以下，且 swap 分区为关闭状态（如图，swap 后的数据为 3 个 0 表示 swap 分区已关闭）。若内存使用率超过了 60%，可通过执行 sync;echo 3 > /proc/sys/vm/drop_caches 释放内存 cache，等待约 1 分钟，然后再检查内存使用率。

注：内存使用率为第一行的 used 值与内存总容量的比值。



```
             total       used       free     shared    buffers     cached
Mem:        128544     128113        430         22         16       4191
-/+ buffers/cache:     123905       4638
Swap:            0          0          0
```

图 17 检查内存使用率

4.1.3 检查交换机配置

确认存储交换机、业务交换机是否开启 STP，如果开启 STP，检查确认连接服务器的端口已经配置为**边缘端口**；如果未开启 STP，则可忽略此项检查。具体检查方法请参考具体型号的交换机的命令手册。**注：交换机配置变更请联系 400 确认后再操作。**

4.1.4 检查主机路由信息

请参考章节 3.1.3。

4.1.5 检查 RAID 卡电池状态

请参考章节 3.1.4。

4.1.6 检查硬盘缓存

请参考章节 3.1.5。

4.1.7 检查集群硬件状态

请参考章节 3.1.6。

4.1.8 检查 NTP 时钟

请参考章节 3.1.7。

4.2 开启维护模式

请参考章节 3.3。**注：Handy 页面操作仅对待关机节点开启维护模式；后台操作方法相同**。

4.2.1 备份网络信息

执行 ifconfig -a 命令，并将输出信息保存到本地。

4.2.2 手动停止 osd

1. 在待关机节点执行 systemctl stop ceph-osd.target，停止该节点所有 osd。

2. 等待约 1 分钟，执行 ceph osd tree，确认只有该节点的所有 osd 状态变为 down，其他节点的 osd 状态仍为 up。

3. 执行 ceph -s，确认 pg 状态中不存在 pg peering，pg stale，pg activating 或 pg inactive 中的任一状态。

注意

1. pg peering，pg stale，pg activating属于停止osd后，pg的中间状态，通常在几秒到十几秒之间就会结束，如果等待1分钟左右还未消失，请联系400处理。

4.3 正常关机下电

注意

1. 如果客户端通过实ip连接存储，关机操作会导致连接该关机节点的业务中断，开机后需要手动重新连接；
2. 如果客户端通过cifs连接访问存储，关机操作可能会导致连接该关机节点的业务中断，需要重新触发或手动重新连接；
3. 若需要下电操作，请确认操作系统完全关闭（电源开关灯由绿色变为黄色）后，再拔除电源线。

1．执行 sync，将内存下刷；

2．执行 hwclock -w，将时钟写入 BIOS；

3．使用 date&hwclock 命令观察输出时间是否一致，如不一致再次执行第 2 步命令并再次检查；

4．执行 shutdown -h now，将服务器正常关机。关机过程中建议关注 HDM 页面电源状态，避免出现关机命令执行失败或关机命令执行卡住的情况。

4.4 开机操作

请参考章节 3.5完成开机操作，待 osd 完全拉起后，登录 Handy 界面，持续观察集群健康度，直到集群健康度恢复 100%且所有告警消除，观察业务读写状况。