# 问题定位分析

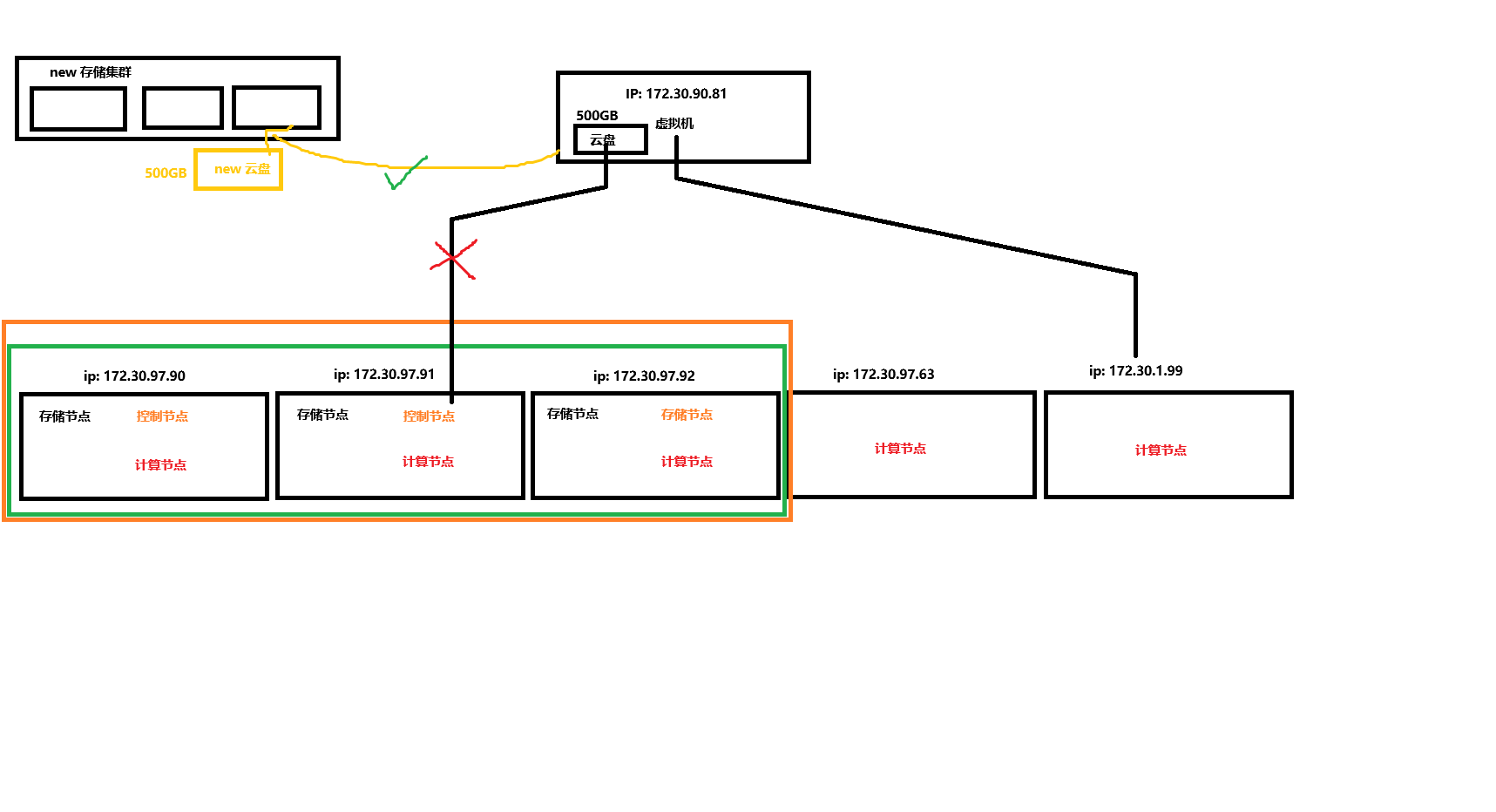
## 一.问题背景

讯飞云存储要进行中心化集群建设，因为现网有很多存量集群建设较早，大部分安装部署的是ceph10.2.10版本，且很多集群部署比较割裂，有些集群容量使用较少，相对应的占用服务器资源较多，增加运维成本的开销，所以才要进行中心化集群建设。

即将现网的很多云主机，镜像和云盘通过热迁移方式迁移到新建集群上去，但是现在有一个问题就是比如一块云盘或虚机镜像创建大小是50GB，实际占用大小是1GB，但是

通过热迁移方式迁移到新建集群后会变成实际大小，即50GB，这会使集群容量膨胀，所以就要解决该问题。

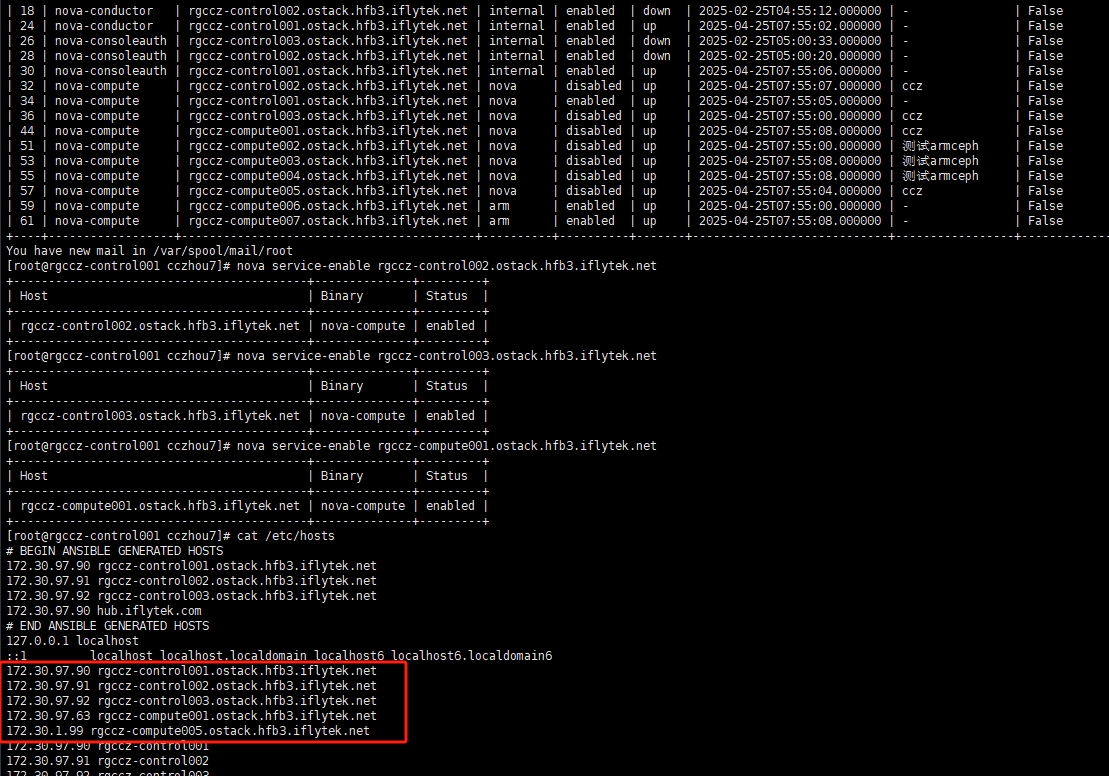
## **二.验证环境**



老存储节点如下图有三个:

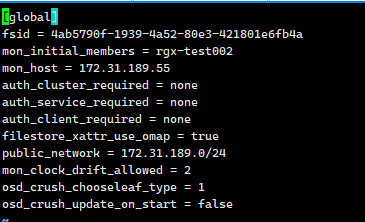
img_v3_02lm_711a6022-c357-42d7-97a4-2bbfa381eenh

控制节点也有三个，控制节点同时还是存储节点，计算节点



新建集群信息如下：

172.31.189.55



****

## **二.问题现象**

此时在其中一台计算节点root@rgccz-compute001 172.30.97.63上进入虚拟化层，执行命令如下：

[root@rgccz-compute001 ~]# docker exec -it -u root nova\_libvirt bash

(nova-libvirt)[root@rgccz-compute001 /]#

(nova-libvirt)[root@rgccz-compute001 /]#

(nova-libvirt)[root@rgccz-compute001 /]#

(nova-libvirt)[root@rgccz-compute001 /]# virsh list

Id Name State

----------------------------------------------------

2 instance-00000016 running

10 instance-000008a3 running

查看其中一台虚机instance-000008a3

(nova-libvirt)[root@rgccz-compute001 /]# ceph osd pool ls

device\_health\_metrics

sata\_pool

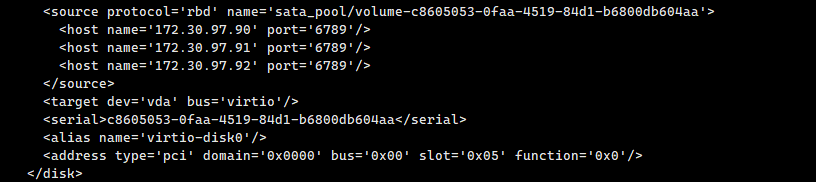
images

vms

执行命令如下:

(nova-libvirt)[root@rgccz-compute001 /]# virsh dumpxml 10





比如对应虚机镜像卷信息如下：

(nova-libvirt)[root@rgccz-compute001 /]# rbd info **sata\_pool/volume-c8605053-0faa-4519-84d1-b6800db604aa**

rbd image 'volume-c8605053-0faa-4519-84d1-b6800db604aa':

size 50 GiB in 12800 objects

order 22 (4 MiB objects)

snapshot\_count: 0

id: 3bd9545e955d8d

block\_name\_prefix: rbd\_data.3bd9545e955d8d

format: 2

features: layering, exclusive-lock, object-map, fast-diff, deep-flatten

op\_features:

flags:

create\_timestamp: Sun Apr 27 11:18:20 2025

access\_timestamp: Sun Apr 27 13:00:08 2025

modify\_timestamp: Sun Apr 27 13:00:08 2025

parent: images/38da0904-629e-4482-a0a8-75f8b9316ac1@snap

overlap: 5 GiB

(nova-libvirt)[root@rgccz-compute001 /]#

(nova-libvirt)[root@rgccz-compute001 /]#

(nova-libvirt)[root@rgccz-compute001 /]#

(nova-libvirt)[root@rgccz-compute001 /]# rbd du sata\_pool/volume-c8605053-0faa-4519-84d1-b6800db604aa

NAME PROVISIONED USED

**volume-c8605053-0faa-4519-84d1-b6800db604aa 50 GiB 900 MiB**

可以看到镜像卷大小是50GB，实际占用磁盘空间大小只有900MB

此时在虚拟化层执行迁移命令，如下所示

virsh qemu-monitor-command instance-000006e6 --pretty '{"execute":"drive-mirror","arguments":{"device":"drive-virtio-disk0","target":"rbd:sata\_pool/volume-c8605053-0faa-4519-84d1-b6800db604aa:auth\_supported=none:mon\_host=172.31.189.55\\:6789","sync":"full","format":"raw"}}'

执行完成后，在172.31.189.55存储节点查看卷占用实际磁盘大小如下所示：

[root@rgx-test002 ~]#

[root@rgx-test002 ~]# rbd info **sata\_pool/volume-c8605053-0faa-4519-84d1-b6800db604aa**

rbd image 'volume-c8605053-0faa-4519-84d1-b6800db604aa':

size 50 GiB in 12800 objects

order 22 (4 MiB objects)

snapshot\_count: 0

id: 7afc6d0ef4b

block\_name\_prefix: rbd\_data.7afc6d0ef4b

format: 2

features: layering, exclusive-lock, object-map, fast-diff, deep-flatten

op\_features:

flags:

create\_timestamp: Sun Apr 27 11:26:03 2025

access\_timestamp: Sun Apr 27 11:26:03 2025

modify\_timestamp: Sun Apr 27 14:09:23 2025

[root@rgx-test002 ~]#

[root@rgx-test002 ~]# rbd du **sata\_pool/volume-c8605053-0faa-4519-84d1-b6800db604aa**

NAME PROVISIONED USED

volume-c8605053-0faa-4519-84d1-b6800db604aa 50 GiB 50 GiB

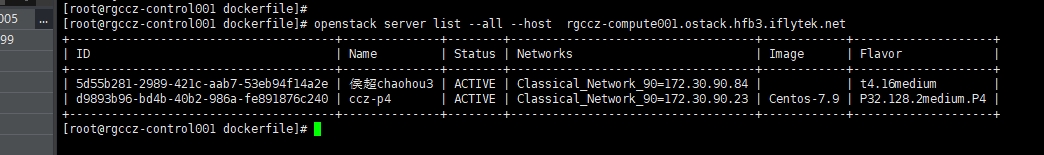
[root@rgx-test002 ~]#

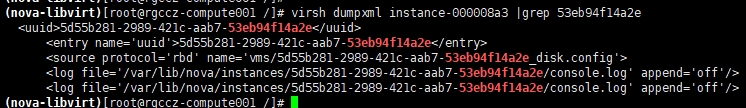
可以看到实际占用大小是50GB

在虚机节点执行virsh shutdown instance-000008a3 关机虚机

此时在新建集群上可以删掉迁移的卷

再控制节点上先找到虚机的uuid





必须要在控制节点上执行如下命令:  
nova start 5d55b281-2989-421c-aab7-53eb94f14a2e开机命令 先在 rgccz-control001 source /root/admin-openrc.sh

然后执行 nova start 5d55b281-2989-421c-aab7-53eb94f14a2e

也可以 nova reboot --hard 5d55b281-2989-421c-aab7-53eb94f14a2e 这个命令直接重起 watcher 也会 回到老存储集群

不用先关机再开机

此时老集群卷还在

## **问题分析**

1.使用virsh qemu-monitor-command命令进行热迁移，qemu drive-mirror方式默认会按逻辑大小全量迁移数据，忽略实际使用量，所以问题切入点还是在虚拟化层;

2.参考ds提供的两个思路，有一个方法是如果qemu版本过低(低于5.0)，则可以通过离线稀疏迁移和动态切换磁盘方式

(1. 离线稀疏迁移

通过 rbd export 和 rbd import 手动迁移数据：

bash

# 在源集群导出稀疏数据流

rbd export sata\_pool/volume-c8605053-0faa-4519-84d1-b6800db604aa \

--export-format=raw \

--sparse-striping \ # 启用稀疏传输

- \ # 输出到标准输出

| ssh <目标集群节点> \ # 通过管道传输到目标集群

"rbd import - sata\_pool/volume-c8605053-0faa-4519-84d1-b6800db604aa \

--image-format 2 \

--image-feature layering"

(2.动态切换磁盘

创建稀疏目标镜像(同上)。

关闭虚拟机并替换磁盘：

bash

virsh edit instance-000008a3 # 修改磁盘路径为目标镜像virsh start instance-000008a3

virsh start instance-000008a3

但是这种方法就需要在计算节点安装ssh，但是这种就不是热迁移了，会影响现网用户业务，不推荐使用。

3. 一个方法是通过目标卷优化和稀疏迁移模式，确保迁移后目标卷磁盘占用和实际使用大小一致。

(1.目标卷预配置(关键步骤)

在目标 Ceph 集群中手动创建稀疏镜像，禁用全量预分配特性：

bash

# 在目标集群创建稀疏镜像（禁用 object-map 等特性）

rbd create sata\_pool/volume-c8605053-0faa-4519-84d1-b6800db604aa \

--size 50G \

--image-format 2 \

--image-feature layering \ # 仅启用 layering 特性

--stripe-unit 4M \ # 条带单元大小（优化性能）

--stripe-count 8

# 禁用可能导致全量分配的特性

rbd feature disable sata\_pool/volume-c8605053-0faa-4519-84d1-b6800db604aa exclusive-lock object-map fast-diff

(2.使用 bitmap 模式热迁移

通过 QEMU 的 bitmap 同步模式，仅传输实际使用的数据块：

bash

virsh qemu-monitor-command instance-000008a3 --pretty '{

"execute": "drive-mirror",

"arguments": {

"device": "drive-virtio-disk0",

"target": "rbd:sata\_pool/volume-c8605053-0faa-4519-84d1-b6800db604aa:auth\_supported=none:mon\_host=172.31.189.55\\:6789",

"sync": "bitmap", # 使用位图同步（仅传输已使用块）

"mode": "existing", # 复用已存在的目标镜像

"format": "raw",

"bitmap": {

"name": "mirror-bitmap", # 自定义位图名称

"granularity": 65536 # 位图粒度（默认 64KB）

}

}

}'

这种就要求qemu版本大于5.0

实际命令如下:

(nova-libvirt)[root@rgccz-compute005 /]# virsh qemu-monitor-command instance-00000885 --pretty '{"execute":"drive-mirror","arguments":{"job-id": "mirror0","device":"libvirt-1-format","target":"rbd:sata\_pool/volume-05490bca-d9d3-41db-a3e1-71b4cc4ca8aa:auth\_supported=none:mon\_host=172.31.189.55\\:6789","sync":"full","format":"raw"}}'

{

"return": {

},

"id": "libvirt-18304"

}

查询相关信息执行命令: virsh qemu-monitor-command instance-00000885 --pretty '{"execute":"query-block"}'

但是奇怪的是执行如下命令还是报错：  
(nova-libvirt)[root@rgccz-compute005 /]# virsh qemu-monitor-command instance-00000885 --pretty '{"execute":"blockdev-mirror","arguments":{"job-id": "mirror0","device":"libvirt-1-format","target":"rbd:sata\_pool/volume-ec0e79b1-059e-4202-9fae-fdfe381786f1:auth\_supported=none:mon\_host=172.31.189.55\\:6789","sync":"bitmap","bitmap":{"name":"mirror-bitmap","granularity":65536},"filter-node-name":"mirror0","copy-mode":"write-blocking"}}'

{

"id": "libvirt-634",

"error": {

"class": "GenericError",

"desc": "Parameter 'bitmap' is unexpected"

}

}

这个报错问了ds，尽管 QEMU 6.2.0 理论上应支持 bitmap 模式，但实际环境中该参数仍被拒绝。这可能是由于 QEMU 编译选项 或 Libvirt 封装限制 导致的功能裁剪。以下是分步解决方案：

确认QEMU 实际支持的参数

3.1. 检查 QEMU 二进制支持的镜像同步模式

(nova-libvirt)[root@rgccz-compute005 /]# /usr/libexec/qemu-kvm -help | grep -A 10 blockdev-mirror

(nova-libvirt)[root@rgccz-compute005 /]#

若输出中无 bitmap 相关参数，说明该功能被裁剪。输出无，说明是被裁剪的，或者当前qemu版本不支持。

3.2. 查询 QEMU 的块设备同步能力

执行如下命令：  
(nova-libvirt)[root@rgccz-compute005 /]# virsh qemu-monitor-command instance-00000885 --pretty '{"execute": "query-command-line-options","arguments": {"option": "blockdev-mirror"}}'

{

"id": "libvirt-1300",

"error": {

"class": "GenericError",

"desc": "invalid option name: blockdev-mirror"

}

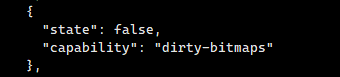
}

检查返回结果中是否包含 bitmap 选项，该命令直接报错，查询ds，从报错信息来看，query-command-line-options 命令不支持直接查询 blockdev-mirror 参数选项，这说明我们需要换一种方式来验证 QEMU 对 bitmap 模式的支持。以下是分步解决方案：

3.3.检查 QEMU 的块设备同步能力，改为改用更通用的命令查询 QEMU 支持的迁移模式：

(nova-libvirt)[root@rgccz-compute005 /]# virsh qemu-monitor-command instance-00000885 --pretty '{"execute": "query-migrate-capabilities"}'





状态是false，是否是不支持呢 ----待后继续查询

3.4.直接测试 bitmap 功能

尝试创建一个脏位图（Dirty Bitmap），验证基础功能是否可用：

(nova-libvirt)[root@rgccz-compute005 /]# virsh qemu-monitor-command instance-00000885 --pretty '{"execute": "block-dirty-bitmap-add","arguments": {"node": "libvirt-1-format","name": "test-bitmap","granularity": 65536}}'

{

"return": {

},

"id": "libvirt-1458"

}

如果成功，说明 QEMU 支持位图功能，但可能 blockdev-mirror 的语法需要调整。

3.5.调整 blockdev-mirror 命令语法，移除显式 bitmap 参数

某些 QEMU 版本可能通过 sync: "bitmap" 隐式启用位图同步，无需单独声明：

(nova-libvirt)[root@rgccz-compute005 /]# virsh qemu-monitor-command instance-00000885 --pretty '{"execute": "blockdev-mirror","arguments": {"job-id": "mirror0","device": "libvirt-1-format","target": "rbd:sata\_pool/volume-ec0e79b1-059e-4202-9fae-fdfe381786f1:auth\_supported=none:mon\_host=172.31.189.55\\:6789","sync": "bitmap","filter-node-name": "mirror0","copy-mode": "write-blocking"}}'

{

"id": "libvirt-1600",

"error": {

"class": "GenericError",

"desc": "Cannot find device='rbd:sata\_pool/volume-ec0e79b1-059e-4202-9fae-fdfe381786f1:auth\_supported=none:mon\_host=172.31.189.55\\:6789' nor node-name='rbd:sata\_pool/volume-ec0e79b1-059e-4202-9fae-fdfe381786f1:auth\_supported=none:mon\_host=172.31.189.55\\:6789'"

}

}

报错了

3.6.使用 drive-mirror 的兼容模式，部分 QEMU 版本仍支持旧版命令的位图功能

(nova-libvirt)[root@rgccz-compute005 /]# virsh qemu-monitor-command instance-00000885 --pretty '{"execute": "drive-mirror","arguments": {"job-id": "mirror0","device": "libvirt-1-format","target": "rbd:sata\_pool/volume-ec0e79b1-059e-4202-9fae-fdfe381786f1:auth\_supported=none:mon\_host=172.31.189.55\\:6789","sync": "bitmap","mode": "existing","format":"raw"}}'

{

"id": "libvirt-1616",

"error": {

"class": "GenericError",

"desc": "Sync mode 'bitmap' not supported"

}

}

报错了

结论: 当前qemu版本对bitmap做了编译裁剪，应该是不行的。

但是通过如下方式进行强行迁移，命令如下：

(nova-libvirt)[root@rgccz-compute005 /]# virsh qemu-monitor-command instance-00000885 --pretty '{"execute":"drive-mirror","arguments":{"job-id": "mirror0","device":"libvirt-2-format","target":"rbd:sata\_pool/volume-05490bca-d9d3-41db-a3e1-71b4cc4ca8aa:auth\_supported=none:mon\_host=172.31.189.55\\:6789","sync":"bitmap","format":"raw"}}'

{

"id": "libvirt-880",

"error": {

"class": "GenericError",

"desc": "Sync mode 'bitmap' not supported"

}

}

virsh qemu-monitor-command instance-00000885 --pretty '{"execute":"drive-mirror","arguments":{"job-id": "mirror0","device":"libvirt-1-format","target":"rbd:sata\_pool/volume-ec0e79b1-059e-4202-9fae-fdfe381786f1:auth\_supported=none:mon\_host=172.31.189.55\\:6789","sync":"bitmap","format":"raw"}}'

在新集群可以看到有迁移过来

[root@rgx-test002 ~]# rbd du sata\_pool/volume-05490bca-d9d3-41db-a3e1-71b4cc4ca8aa

NAME PROVISIONED USED

volume-05490bca-d9d3-41db-a3e1-71b4cc4ca8aa 500 GiB 0 B

[root@rgx-test002 ~]#

[root@rgx-test002 ~]# rbd status sata\_pool/volume-05490bca-d9d3-41db-a3e1-71b4cc4ca8aa

Watchers: none

但是无挂载，即IO不会进行双写，此方法无效。

4.qemu sync同步模式有如下几种

|  |  |  |
| --- | --- | --- |
| **sync 模式** | **作用** | **适用场景** |
| full | 完整备份，会将指定磁盘及其链上的所有母盘（backing file），合并输出到本地 | 目标卷需与源卷完全一致（包括未使用空间） |
| top | 仅备份当前磁盘，而不包含backing file。此项仅对本地盘有效，对于ceph rbd，top和full的效果一样都是完整备份 | 源卷有快照链时减少数据量 |
| none | new I/O，即从执行备份命令起，一旦有IO变化，就会实时导出有变化的数据，直到虚拟机关机或者手工停止job才会停止。仅支持qcow2格式的块设备，不支持raw格式（注意，这里指的是qcow2格式，而并非要求qcow2本地盘） | 手动控制同步过程 |
| incremental | 指定dirty bitmap进行备份，目前qemu的增量备份就是指基于dirty bitmap的备份 | 已有基础镜像的增量更新 |
| bitmap | 基于脏位图（Dirty Bitmap）同步已修改的块（需提前创建位图） | 稀疏传输，仅同步实际数据（需 QEMU 支持） |

尝试如下几种同步模式

top模式: 容量存在膨胀，和full相一致

none模式: 直接报错如下所示

(nova-libvirt)[root@rgccz-compute005 /]# virsh qemu-monitor-command instance-00000885 --pretty '{"execute":"drive-mirror","arguments":{"job-id": "mirror0","device":"libvirt-1-format","target":"rbd:sata\_pool/volume-ec0e79b1-059e-4202-9fae-fdfe381786f1:auth\_supported=none:mon\_host=172.31.189.55\\:6789","sync":"none","format":"raw"}}'

{

"id": "libvirt-414",

"error": {

"class": "GenericError",

"desc": "Backing file not supported for file format 'raw'" -----------------------> 不支持raw格式???

}

}

继续修改报错如下:

(nova-libvirt)[root@rgccz-compute005 /]# virsh qemu-monitor-command instance-00000885 --pretty '{"execute":"drive-mirror","arguments":{"job-id": "mirror0","device":"libvirt-1-format","target":"rbd:sata\_pool/volume-ec0e79b1-059e-4202-9fae-fdfe381786f1:auth\_supported=none:mon\_host=172.31.189.55\\:6789","sync":"none","format":"raw","mode":"existing"}}'

{

"id": "libvirt-425",

"error": {

"class": "GenericError",

"desc": "error reading header from volume-ec0e79b1-059e-4202-9fae-fdfe381786f1: No such file or directory"

}

}

incremental模式:

(nova-libvirt)[root@rgccz-compute005 /]# virsh qemu-monitor-command instance-00000885 --pretty '{"execute":"drive-mirror","arguments":{"job-id": "mirror0","device":"libvirt-1-format","target":"rbd:sata\_pool/volume-ec0e79b1-059e-4202-9fae-fdfe381786f1:auth\_supported=none:mon\_host=172.31.189.55\\:6789","sync":"incremental","format":"raw"}}'

{

"id": "libvirt-419",

"error": {

"class": "GenericError",

"desc": "Sync mode 'incremental' not supported" ---------------------------------> 不支持???

}

}

5.尝试去掉镜像卷全量分配属性，进行热迁移验证效果

执行如下:

(nova-libvirt)[root@rgccz-compute005 /]# rbd feature disable sata\_pool/volume-ec0e79b1-059e-4202-9fae-fdfe381786f1 deep-flatten

(nova-libvirt)[root@rgccz-compute005 /]#

(nova-libvirt)[root@rgccz-compute005 /]#

(nova-libvirt)[root@rgccz-compute005 /]# rbd info sata\_pool/volume-ec0e79b1-059e-4202-9fae-fdfe381786f1

rbd image 'volume-ec0e79b1-059e-4202-9fae-fdfe381786f1':

size 20 GiB in 5120 objects

order 22 (4 MiB objects)

snapshot\_count: 0

id: 235f8b6940657

block\_name\_prefix: rbd\_data.235f8b6940657

format: 2

features: layering

op\_features:

flags:

create\_timestamp: Mon Apr 28 10:37:32 2025

access\_timestamp: Mon Apr 28 16:16:17 2025

modify\_timestamp: Mon Apr 28 10:37:32 2025

结论: 卷仍然存在膨胀问题

今日进展:

1.研究了下virsh qemu sync的几种同步模式(full/ top/ none / bitmap / incremental) full/top 都是全量拷贝 none 不支持ceph rbd raw卷格式 bitmap/incremental 当前qemu虚拟化层不支持，分析是qemu编译将两种功能裁剪掉了，卷可以迁移过来，只是未生效

2.尝试去掉ceph rbd镜像卷全量分配属性，进行热迁移验证效果 验证： 无效果

明日计划: 1.要看一下kvm-qemu官网热迁移相关支持说明，是命令配置问题还是版本问题

6.

该问题原因在于Qemu虚机对rbd卷进行热迁移时，非写入数据读取范围后会重新填零操作，所以50GB卷从老集群迁移到新集群后占用磁盘空间会算上填零区域，所以占用磁盘大小就是50GB

分析qemu官网([Invocation — QEMU documentation](https://www.qemu.org/docs/master/system/invocation.html) 对应中文翻译: [[QEMU]块设备选项 - 哔哩哔哩](https://www.bilibili.com/opus/685381105765842960))，其中有一个参数detect-zeroes 可以对该情况进行优化，对于qemu版本高于V2.6.0版本(包含)的才支持该功能。

继续查看虚拟化层版本信息如下所示:

(nova-libvirt)[root@rgccz-compute001 /]# virsh version

Compiled against library: libvirt 8.0.0 //表示当前virsh工具在编译时链接的 Libvirt 库版本为8.0.0

Using library: libvirt 8.0.0 //当前系统运行时实际加载的 Libvirt 库版本为 8.0.0，与编译版本一致

Using API: QEMU 8.0.0 //Libvirt 使用的 QEMU 管理接口（API）版本为 8.0.0。此版本号是 Libvirt 内部定义的 QEMU 接口兼容性标识，并非 QEMU 的实际版本

Running hypervisor: QEMU 6.2.0 //当前宿主机上运行的 QEMU 程序版本为 6.2.0

Qemu版本为V6.2.0，所以是支持的，目前所有虚机未添加该参数，需要在虚机侧添加适配该参数。

[root@rgccz-control001 ~]# openstack server list --all --host rgccz-compute005.ostack.hfb3.iflytek.net

+--------------------------------------+--------------+--------+-----------------------------------+---------------------+-------------+

| ID | Name | Status | Networks | Image | Flavor |

+--------------------------------------+--------------+--------+-----------------------------------+---------------------+-------------+

| 8459a89f-6fcb-43ae-a4ad-c5e430f0f243 | vm-cczhou7-1 | ACTIVE | Classical\_Network\_90=172.30.90.80 | | t4.16medium |

| 9d554b47-a805-4fcd-ac1f-4b73d01fdd94 | q35 | ACTIVE | Classical\_Network\_90=172.30.90.83 | openEuler-22.03.LTS | t4.16medium |

| 629b29b9-b0fb-4766-b27d-f4a686d5cbda | 侯超chaohou3 | ACTIVE | Classical\_Network\_90=172.30.90.81 | Centos-7.9-old | t8.16medium |

+--------------------------------------+--------------+--------+-----------------------------------+---------------------+-------------+

[root@rgccz-control001 ~]#

尝试对计算节点5上的虚机进行添加:

6.1使用virsh edit instance-00000885 方式修改添加，并重启虚机，发现未生效，经了解，openstack重启虚机会重新生成虚机xml文件并覆盖，该方法无效。

6.2通过openstack image set修改镜像属性，操作如下:

[2025-04-29 16:06:43][root][10.43.11.153] openstack image set --property detect-zeroes=unmap --property discard=unmap Centos-7.9-old

修改前:



修改后:



通过执行命令: openstack image show Centos-7.9-old 命令来看属性: | properties | cluster='cephA', detect-zeroes='unmap', direct\_url='rbd://ebe610f1-e769-447e-8447-100e9d949e6b/images/38da0904-629e-4482-a0a8-75f8b9316ac1/snap', discard='unmap', hw\_qemu\_guest\_agent='yes', hw\_vif\_multiqueue\_enabled='true', locations='[{u'url': u'rbd://ebe610f1-e769-447e-8447-100e9d949e6b/images/38da0904-629e-4482-a0a8-75f8b9316ac1/snap', u'metadata': {}}]' |

可以看到该虚机属性里已添加 detect-zeroes配置

重启虚机后，继续查看虚机挂载磁盘属性，如下所示：



<disk type='network' device='disk'>

<driver name='qemu' type='raw' cache='none' discard='unmap'/>

<auth username='cinder'>

<secret type='ceph' uuid='2786c4a2-156f-49b3-8c08-d7e73fba6996'/>

</auth>

<source protocol='rbd' name='sata\_pool/volume-ec0e79b1-059e-4202-9fae-fdfe381786f1' index='1'>

<host name='172.30.97.90' port='6789'/>

<host name='172.30.97.91' port='6789'/>

<host name='172.30.97.92' port='6789'/>

</source>

<target dev='vdc' bus='virtio'/>

<serial>ec0e79b1-059e-4202-9fae-fdfe381786f1</serial>

<alias name='virtio-disk2'/>

<address type='pci' domain='0x0000' bus='0x00' slot='0x07' function='0x0'/>

</disk>

挂载磁盘上并无该属性信息

今日进展:

6.该问题原因在于Qemu虚机对rbd卷进行热迁移时，非写入数据读取范围后会重新填零操作，所以50GB卷从老集群迁移到新集群后占用磁盘空间会算上填零区域，所以占用磁盘大小就是50GB

分析qemu官网([Invocation — QEMU documentation](https://www.qemu.org/docs/master/system/invocation.html" \t "_blank) 对应中文翻译: [[QEMU]块设备选项 - 哔哩哔哩](https://www.bilibili.com/opus/685381105765842960" \t "_blank))，其中有一个参数detect-zeroes 可以对该情况进行优化，对于qemu版本高于V2.6.0版本(包含)的才支持该功能。

继续查看虚拟化层版本信息如下所示:

(nova-libvirt)[root@rgccz-compute001 /]# virsh version

Compiled against library: libvirt 8.0.0 //表示当前virsh工具在编译时链接的 Libvirt 库版本为8.0.0

Using library: libvirt 8.0.0 //当前系统运行时实际加载的 Libvirt 库版本为 8.0.0，与编译版本一致

Using API: QEMU 8.0.0 //Libvirt 使用的 QEMU 管理接口（API）版本为 8.0.0。此版本号是 Libvirt 内部定义的 QEMU 接口兼容性标识，并非 QEMU 的实际版本 Running hypervisor: QEMU 6.2.0 //当前宿主机上运行的 QEMU 程序版本为 6.2.0 Qemu版本为V6.2.0，所以是支持的，目前所有虚机未添加该参数，需要在虚机侧添加适配该参数。

尝试对计算节点5上的虚机进行添加，目前尝试了2种方式

6.1使用virsh edit instance-00000885 方式修改添加，并重启虚机，发现未生效，经了解，openstack重启虚机会重新生成虚机xml文件并覆盖，该方法无效。

6.2通过openstack image set修改镜像属性，修改成功，但是虚机挂载磁盘属性未带上参数detect-zeroes，尝试热迁，卷容量未减少，修改新老ceph集群支持对应discard 参数，重启后老集群的一个osd down，计划先修复集群osd，再继续验证。

修复ceph osd进程正常拉起后，验证卷热迁移，暂无明显效果

基于该虚拟机重新挂载10GB卷进行热迁移，暂无明显效果

通过执行virsh qemu-monitor-command instance-00000885 --pretty '{"execute":"query-block"}' 命令查看，输出如下：



{

"io-status": "ok",

"device": "",

"locked": false,

"removable": false,

"inserted": {

"iops\_rd": 0,

"detect\_zeroes": "off",

"image": {

"virtual-size": 10737418240,

"filename": "json:{\"driver\": \"raw\", \"file\": {\"pool\": \"sata\_pool\", \"image\": \"volume-c22acb18-663c-4621-b744-5f9fbf51c991\", \"server.0.host\": \"172.30.97.90\", \"server.1.host\": \"172.30.97.91\", \"server.2.host\": \"172.30.97.92\", \"driver\": \"rbd\", \"server.0.port\": \"6789\", \"server.1.port\": \"6789\", \"server.2.port\": \"6789\", \"user\": \"cinder\"}}",

"cluster-size": 4194304,

"format": "raw",

"dirty-flag": false

},

"iops\_wr": 0,

"ro": false,

"node-name": "libvirt-5-format",

"backing\_file\_depth": 0,

"drv": "raw",

"iops": 0,

"bps\_wr": 0,

"write\_threshold": 0,

"encrypted": false,

"bps": 0,

"bps\_rd": 0,

"cache": {

"no-flush": false,

"direct": true,

"writeback": true

},

"file": "json:{\"driver\": \"raw\", \"file\": {\"pool\": \"sata\_pool\", \"image\": \"volume-c22acb18-663c-4621-b744-5f9fbf51c991\", \"server.0.host\": \"172.30.97.90\", \"server.1.host\": \"172.30.97.91\", \"server.2.host\": \"172.30.97.92\", \"driver\": \"rbd\", \"server.0.port\": \"6789\", \"server.1.port\": \"6789\", \"server.2.port\": \"6789\", \"user\": \"cinder\"}}"

},

"qdev": "/machine/peripheral/virtio-disk3/virtio-backend",

"type": "unknown"

}

可以看到 "detect\_zeroes": "off",未生效，需要查看是否有命令可以修改该参数

6.3尝试执行命令[root@rgccz-control001 ~]# openstack flavor set --property hw:detect\_zeroes=unmap --property hw:discard=unmap t8.16medium

执行成功

其中查叙虚机命令如下:

[root@rgccz-control001 ~]# openstack server list --all --host rgccz-compute005.ostack.hfb3.iflytek.net

+--------------------------------------+--------------+--------+-----------------------------------+----------------+-------------+

| ID | Name | Status | Networks | Image | Flavor |

+--------------------------------------+--------------+--------+-----------------------------------+----------------+-------------+

| 8459a89f-6fcb-43ae-a4ad-c5e430f0f243 | vm-cczhou7-1 | ACTIVE | Classical\_Network\_90=172.30.90.80 | | t4.16medium |

| 629b29b9-b0fb-4766-b27d-f4a686d5cbda | 侯超chaohou3 | ACTIVE | Classical\_Network\_90=172.30.90.81 | Centos-7.9-old | t8.16medium |

+--------------------------------------+--------------+--------+-----------------------------------+----------------+-------------+

通过执行如下命令查询虚机所挂载的卷

[root@rgccz-control001 ~]# openstack server show 629b29b9-b0fb-4766-b27d-f4a686d5cbda -c volumes\_attached

+------------------+-------------------------------------------+

| Field | Value |

+------------------+-------------------------------------------+

| volumes\_attached | id='07e0e805-a3e0-4833-bdf0-70fcecd4c621' |

| | id='05490bca-d9d3-41db-a3e1-71b4cc4ca8aa' |

| | id='ec0e79b1-059e-4202-9fae-fdfe381786f1' |

| | id='c22acb18-663c-4621-b744-5f9fbf51c991' |

+------------------+-------------------------------------------+

执行命令查看挂载卷的类型，这些卷的类型应该都是sata

[root@rgccz-control001 ~]# openstack volume show 07e0e805-a3e0-4833-bdf0-70fcecd4c621 -c type

+-------+-------+

| Field | Value |

+-------+-------+

| type | sata |

+-------+-------+

[root@rgccz-control001 ~]# openstack volume show c22acb18-663c-4621-b744-5f9fbf51c991 -c type

+-------+-------+

| Field | Value |

+-------+-------+

| type | sata |

+-------+-------+

通过如下命令设置对应卷类型的卷修改属性信息

[root@rgccz-control001 ~]# openstack volume type set sata --property detect\_zeroes=unmap --property discard=unmap ----------------> 这里不太对

查询对应的属性信息，该属性信息是虚机侧卷的类型

[root@rgccz-control001 ~]# openstack volume type show sata -c properties

+------------+------------------------------------------------------------------------+

| Field | Value |

+------------+------------------------------------------------------------------------+

| properties | detect\_zeroes='unmap', discard='unmap', volume\_backend\_name='rbd-sata' |

+------------+------------------------------------------------------------------------+

[root@rgccz-control001 ~]# openstack volume type show sata

+--------------------+------------------------------------------------------------------------+

| Field | Value |

+--------------------+------------------------------------------------------------------------+

| access\_project\_ids | None |

| description | None |

| id | f65bed98-1e71-4242-9e22-0f3853f5fb06 |

| is\_public | True |

| name | sata |

| properties | detect\_zeroes='unmap', discard='unmap', volume\_backend\_name='rbd-sata' |

| qos\_specs\_id | None |

+--------------------+------------------------------------------------------------------------+

查询存储侧卷的属性信息，发现此时存储侧卷类型未修改

[root@rgccz-control001 ~]# openstack volume show c22acb18-663c-4621-b744-5f9fbf51c991 -c type -c properties

+------------+--------------------+

| Field | Value |

+------------+--------------------+

| properties | attached\_mode='rw' |

| type | sata |

+------------+--------------------+

通过执行如下命令设置存储侧卷的属性信息

[root@rgccz-control001 ~]# openstack volume set c22acb18-663c-4621-b744-5f9fbf51c991 --property detect\_zeroes=unmap --property discard

通过执行如下命令查看存储卷属性信息，可以确认该卷已经带了detect\_zeroes属性

[root@rgccz-control001 ~]# openstack volume show c22acb18-663c-4621-b744-5f9fbf51c991 -c type -c properties

+------------+------------------------------------------------------------+

| Field | Value |

+------------+------------------------------------------------------------+

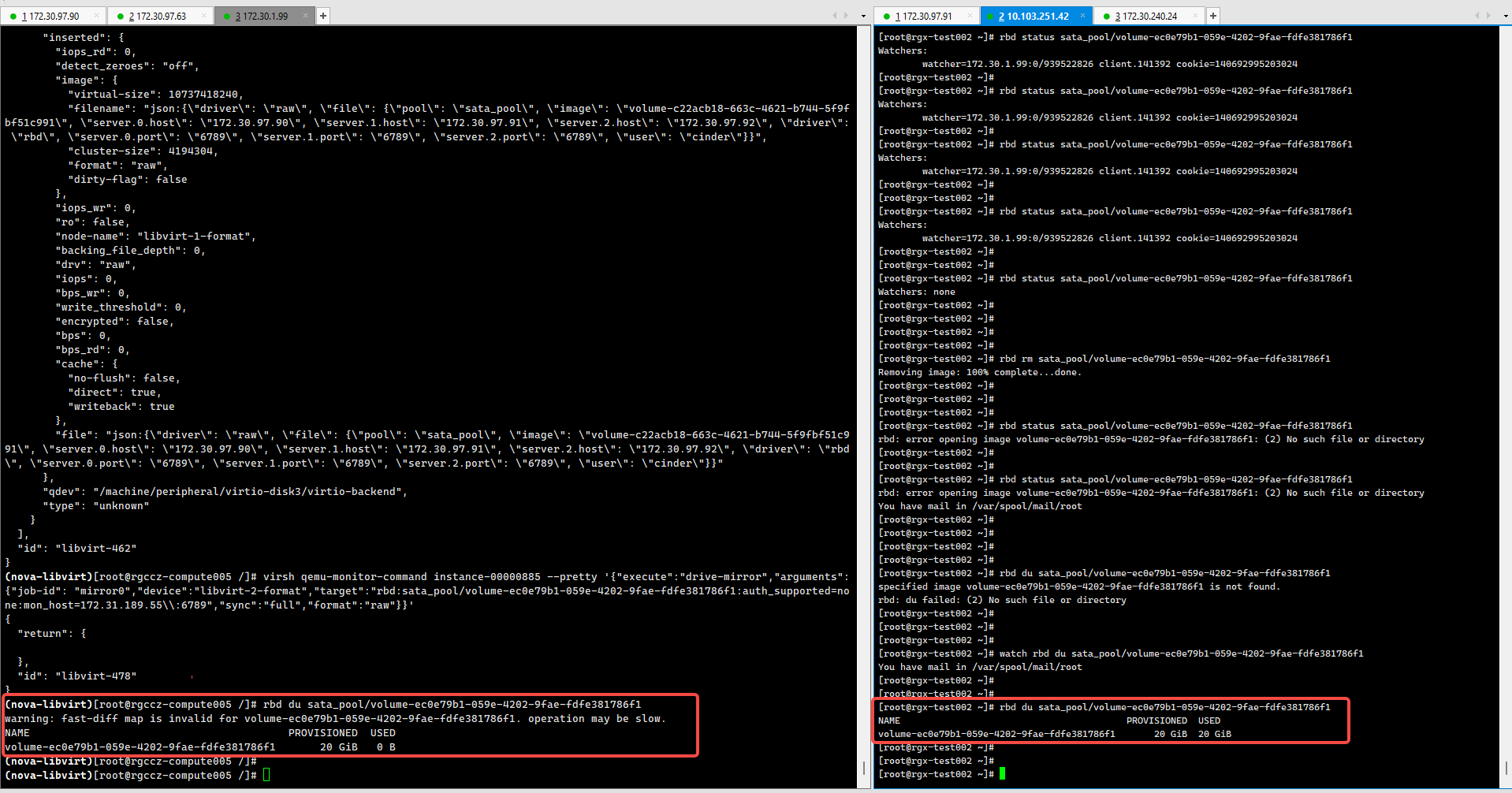
| properties | attached\_mode='rw', detect\_zeroes='unmap', discard='unmap' |

| type | sata |

+------------+------------------------------------------------------------+

查看的ds信息如下：  


重新挂载卷后执行验证操作，仍然无效果



今日进展:

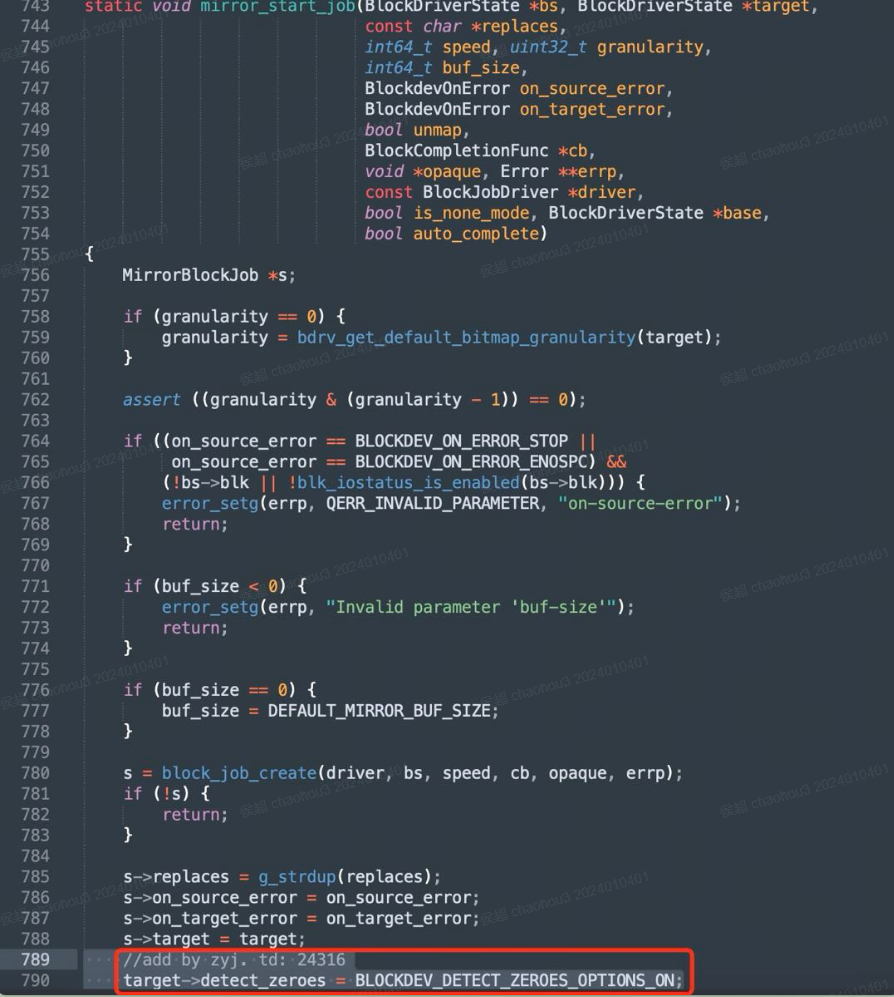
1. 修复ceph老集群osd进程异常
2. 通过openstack image set修改镜像属性成功，验证迁移卷膨胀问题仍存在
3. 通过openstack volume type set sata修改虚机侧挂载卷属性，openstack volume set修改存储侧卷属性均成功，验证迁移卷膨胀问题仍存在 ---- 继续分析该操作未生效原因

05.06

今日进展

1.执行命令openstack volume set c22acb18-663c-4621-b744-5f9fbf51c991 --property detect\_zeroes=unmap --property discard 修改存储侧卷属性，并重新挂载，验证无效果，暂无更进一步分析，通过在qemu6.2.0源码里查询detect\_zeroes，并未在mirror.c 文件查询到，怀疑高版本是否也未必支持ceph rbd卷迁移稀疏优化(qemu低版本明确不支持detect\_zeroes参数传递从而优化卷全量拷贝问题)。

2.通过了解和查寻业界针对这一类问题的做法，目前中国电子云虚拟化部门(qemu版本2.6.0)修改方案如下，在mirror.c mirror\_start\_job函数里添加detect\_zeroes参数，支持稀疏拷贝，从而解决该问题。



下一步计划:

1.了解openstack 架构以及和qemu 关系

2.了解qemu源码架构以及各子模块业务功能

3.块存储卷迁移流程

4.修改qemu源码重新编译替换进行验证

今日进展:

了解openstack、qemu、libvirt层次关系，下载qemu源码，暂无更新进展;

提供一个思路：如果能够修改虚拟机xml配置文件添加detect\_zeroes参数，就可以修改virsh命令直接验证一下



## **参考链接**

[虚拟机通过virsh qemu-monitor-command在线备份 - ishmaelwanglin - 博客园](https://www.cnblogs.com/ishmaelwanglin/p/17053204.html)

[Invocation — QEMU documentation](https://www.qemu.org/docs/master/system/invocation.html)

[[QEMU]块设备选项 - 哔哩哔哩](https://www.bilibili.com/opus/685381105765842960)

[QEMU虚拟机热迁移 | 记录成长的痕迹](http://joyxu.github.io/2025/02/08/QEMU%E8%99%9A%E6%8B%9F%E6%9C%BA%E7%83%AD%E8%BF%81%E7%A7%BB/)