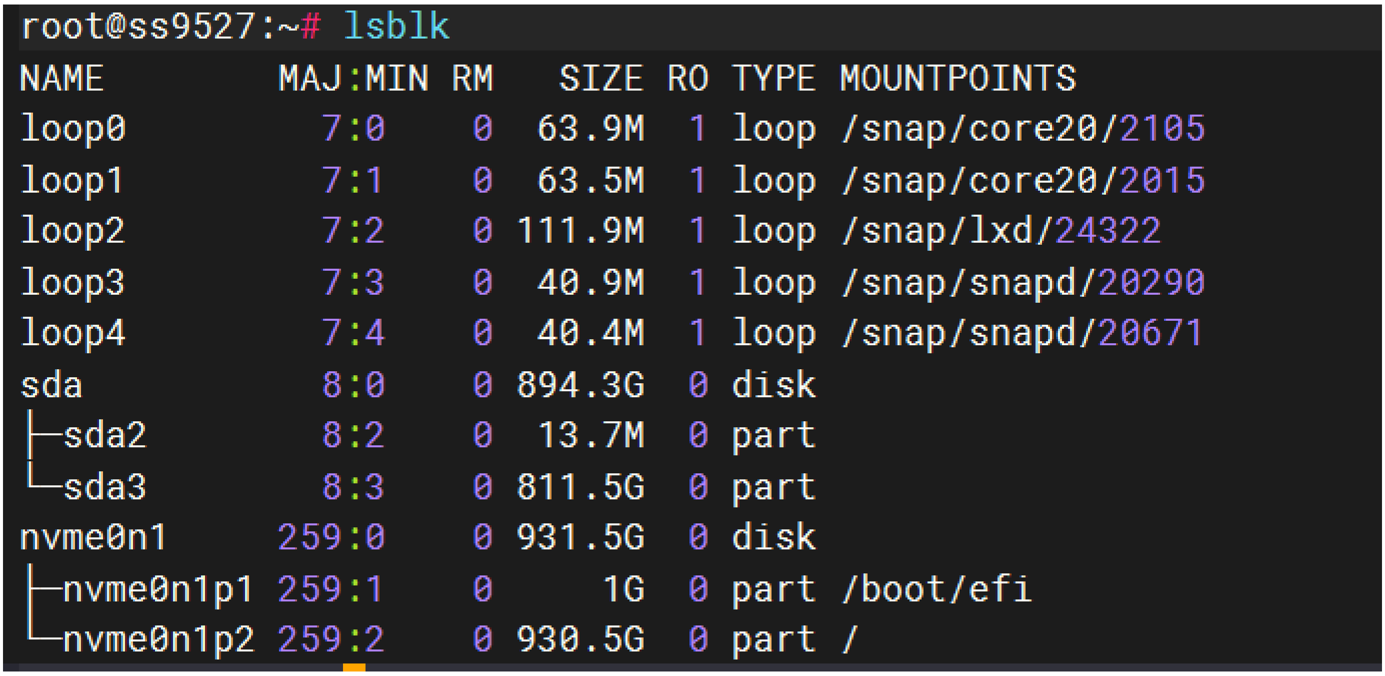
Ceph性能测试

三台机器，分别命名为ss9527, ss01, ss02，先对每个单机进行测试，最后测试ceph集群性能，其中ss9527是主节点，其余为从节点，每台主机两块盘，其中sda作为osd，nvme0n1作为系统盘



```
root@ss9527:~# lsblk
NAME        MAJ:MIN RM   SIZE RO TYPE MOUNTPOINTS
loop0         7:0    0  63.9M  1 loop /snap/core20/2105
loop1         7:1    0  63.5M  1 loop /snap/core20/2015
loop2         7:2    0 111.9M  1 loop /snap/lxd/24322
loop3         7:3    0  40.9M  1 loop /snap/snapd/20290
loop4         7:4    0  40.4M  1 loop /snap/snapd/20671
sda           8:0    0 894.3G  0 disk
├─sda2        8:2    0  13.7M  0 part
└─sda3        8:3    0 811.5G  0 part
nvme0n1     259:0    0 931.5G  0 disk
├─nvme0n1p1 259:1    0     1G  0 part /boot/efi
└─nvme0n1p2 259:2    0 930.5G  0 part /
```

1 测试准备

1.1 osd磁盘性能

cephadm文件路径问题

如果是cephadm部署，本节的所有osd文件路径都应该为

`/var/lib/ceph/<ceph-id>/osd.x`

对应下文中的

`/var/lib/ceph/osd/ceph-x`

或者进入对应osd容器中使用

osd写性能

```
num=0
# 清除缓存
echo 3 > /proc/sys/vm/drop_caches
```

```
# 使用dd编写一个10G的名为deleteme的文件，该文件填充为0，/dev/zero作为Ceph OSD安装目录的输入文件
dd if=/dev/zero of=/var/lib/ceph/osd/ceph-${num}/deleteme bs=1G count=5
```

> 应该使用oflag=direct绕过内核缓存，但是ubuntu不支持这个选项

osd读性能

```
num=0
# 清除缓存
echo 3 > /proc/sys/vm/drop_caches
# 读取刚刚编写的文件
dd if=/var/lib/ceph/osd/ceph-${num}/deleteme of=/dev/null bs=1G count=5
```

> 应该使用iflag=direct绕过内核缓存，但是ubuntu不支持这个选项

单个机器上全部osd的读写性能

写

```
echo 3 > /proc/sys/vm/drop_caches
for i in `mount | grep osd | awk '{print $3}'`; do (dd if=/dev/zero of=$i/deleteme bs=1G count=1 &) ; done
```

读

```
echo 3 > /proc/sys/vm/drop_caches
for i in `mount | grep osd | awk '{print $3}'`; do (dd if=$i/deleteme of=/dev/null bs=1G count=5 &); done
```

如果是cephadm部署，无法使用这个命令，mount | grep osd | awk '{print $3}'找不到全部的osd文件

测试结果：

	ss01:osd.0	ss02:osd.1	ss9527:osd.2
平均写性能 (GB/s)	2.9	2.5	3.4
平均读性能 (GB/s)	6.8	5.2	7.8

1.2 网络

使用iperf3测试带宽

ss9527:

```
iperf3 -s -f M
```

ss01，ss02:

```
iperf3 -c ss9527 -f M
```

反复运行，稳定在112MBytes/sec

2 Ceph性能测试

2.1 rados bench

测试存储池的读写性能

该工具的语法为：rados bench -p <pool_name> <seconds> <write|seq|rand> -b <block size> -t --no-cleanup

- pool_name：测试所针对的存储池
- seconds：测试所持续的秒数
- <write|seq|rand>：操作模式，write写；seq顺序读；rand随机读
- -b：block size，即块大小，默认为 4M
- -t：读/写并行数，默认为 16
- --no-cleanup 表示测试完成后不删除测试用数据。在做读测试之前，需要使用该参数来运行一遍写测试来产生测试数据，在全部测试结束后可以运行 rados -p <pool_name> cleanup 来清理所有测试数据。

写：

```
rados bench -p volumes 10 write --no-cleanup
hints = 1
Maintaining 16 concurrent writes of 4194304 bytes to objects of size 4194304 for up to 10 seconds or 0 objects
Object prefix: benchmark_data_ss9527_3016569
  sec Cur ops   started  finished  avg MB/s  cur MB/s last lat(s)  avg lat(s)
```

```
    0       0         0         0         0         0           -           0
    1      16        21         5   19.9966        20    0.866188    0.591533
    2      16        39        23    45.992        72      1.5964    0.985411
    3      16        61        45    59.991        88    0.553342    0.936931
    4      16        75        59    58.992        56     1.22173    0.933057
    5      16        91        75   59.9924        64     1.00139    0.973164
    6      16       109        93   61.9927        72     1.00509    0.958637
    7      16       128       112    63.992        76    0.640868    0.935596
    8      16       144       128   63.9922        64    0.506286    0.929051
    9      16       158       142   63.1031        56    0.665602    0.934408
   10      16       177       161   64.3915        76     1.35935    0.947261
Total time run:         10.4884
Total writes made:      177
Write size:             4194304
Object size:            4194304
Bandwidth (MB/sec):     67.5034
Stddev Bandwidth:       18.4222
Max bandwidth (MB/sec): 88
Min bandwidth (MB/sec): 20
Average IOPS:           16
Stddev IOPS:            4.60555
Max IOPS:               22
Min IOPS:               5
Average Latency(s):     0.93303
Stddev Latency(s):      0.310663
Max latency(s):         1.67941
Min latency(s):         0.272525
```

顺序读：

```
rados bench -p volumes 10 seq
hints = 1
  sec Cur ops   started  finished  avg MB/s  cur MB/s last lat(s)  avg lat(s)
    0       0         0         0         0         0           -           0
    1      16        76        60   239.974       240    0.639121    0.193516
    2      16       116       100   199.978       160    0.712224    0.270871
    3      16       163       147   195.975       188    0.279109     0.29202
Total time run:       3.78572
Total reads made:     177
Read size:            4194304
Object size:          4194304
Bandwidth (MB/sec):   187.019
Average IOPS:         46
Stddev IOPS:          10.1489
Max IOPS:             60
```

```
Min IOPS:             40
Average Latency(s):   0.318287
Max latency(s):       0.85219
Min latency(s):       0.00218221
```

随机读：

```
rados bench -p volumes 10 rand
hints = 1
  sec Cur ops   started  finished  avg MB/s  cur MB/s last lat(s)  avg lat(s)
    0       0         0         0         0         0           -           0
    1      16        79        63   251.916       252    0.483903    0.181538
    2      16       126       110   219.952       188   0.0761377    0.232897
    3      16       179       163   217.294       212    0.250514    0.251589
    4      16       218       202   201.968       156    0.213203    0.282472
    5      16       259       243   194.371       164    0.141211    0.306229
    6      16       302       286   190.636       172    0.142213    0.308882
    7      16       350       334   190.828       192  0.00558097    0.315481
    8      16       404       388    193.97       216   0.0105067    0.314185
    9      16       450       434    192.86       184    0.877584    0.317227
   10      16       502       486   194.371       208    0.356228    0.315628
Total time run:       10.5344
Total reads made:     502
Read size:            4194304
Object size:          4194304
Bandwidth (MB/sec):   190.613
Average IOPS:         47
Stddev IOPS:          7.13676
Max IOPS:             63
Min IOPS:             39
Average Latency(s):   0.326893
Max latency(s):       1.09295
Min latency(s):       0.00127024
```

2.2 rados load-gen （未完善）

用来在Ceph cluster上生成负载和模拟高负载场景，和rados bench相比，rados load-gen 的特点是可以产生混合类型的测试负载

每次操作写入 4 MB 的数据，目标吞吐量为 1 GB/秒，最大吞吐量为1 GB

```
rados -p volumes load-gen \
--read-percent 0 \
--min-object-size 1G \
--max-object-size 1G \
--max-ops 1 \
--read-percent 0 \
--min-op-len 4M \
--max-op-len 4M \
--target-throughput 1G \
--max_backlog 1G
```

--num-objects	初始生成测试用的对象数，默认 200
--min-object-size	测试对象的最小大小，默认 1KB，单位byte
--max-object-size	测试对象的最大大小，默认 5GB，单位byte
--min-op-len	压测IO的最小大小，默认 1KB，单位byte
--max-op-len	压测IO的最大大小，默认 2MB，单位byte
--max-ops	一次提交的最大IO数
--target-throughput	一次提交IO的历史累计吞吐量上限，默认 5MB/s，单位B/s
--max-backlog	一次提交IO的吞吐量上限，默认10MB/s，单位B/s
--read-percent	读写混合中读的比例，默认80，范围[0, 100]
--run-length	运行的时间，默认60s，单位秒

查阅的资料都是这样做的，但我们的集群上却不能使用：

```
run length 0 seconds
preparing 200 objects
aio_write failed
load-gen bootstrap failed
```

半天没找到解决办法，暂时搁置。

2.3 rbd bench-write

测试块设备性能，在pool中创建一个块设备，并挂载到本地系统进行测试

创建

```
# 在volumes中创建一个testdisk，并启用layering
~# rbd create volumes/testdisk --size 10240 --image-feature layering
~# rbd info -p volumes --image testdisk
rbd image 'testdisk':
        size 10 GiB in 2560 objects
        order 22 (4 MiB objects)
        snapshot_count: 0
        id: 234ab0eeae5c6f
        block_name_prefix: rbd_data.234ab0eeae5c6f
        format: 2
        features: layering
        op_features:
        flags:
        create_timestamp: Tue Jan  9 16:43:10 2024
        access_timestamp: Tue Jan  9 16:43:10 2024
        modify_timestamp: Tue Jan  9 16:43:10 2024
~# rbd map volumes/testdisk
/dev/rbd0
~# rbd showmapped
id  pool    namespace  image     snap  device
0   volumes             testdisk  -     /dev/rbd0
```

挂载

```
~# mkfs.ext4 /dev/rbd0
~# mkdir -p /mnt/testdisk
~# mount /dev/rbd0 /mnt/testdisk/
~# df -h /mnt/testdisk/
Filesystem      Size  Used Avail Use% Mounted on
/dev/rbd0        10G  105M  9.9G   2% /mnt/testdisk
```

测试，写入5个G的文件

```
# rbd bench-write volumes/testdisk --io-total 5G
# 会警告说bench-write被弃用，使用最新的命令：
~# rbd bench --io-type write volumes/testdisk  --io-total 5G
...测试三次的结果：
elapsed: 162   ops: 1310720   ops/sec: 8063.03   bytes/sec: 31 MiB/s
elapsed: 185   ops: 1310720   ops/sec: 7064.55   bytes/sec: 28 MiB/s
elapsed: 181   ops: 1310720   ops/sec: 7226.28   bytes/sec: 28 MiB/s
```

2.4 fio + rbd ioengine

测RBD块设备的写入性能，使用2.3的/dev/rbd0

下载fio

```
apt install fio -y
```

编辑配置文件

```
~# cat write.fio
[write-4M]
ioengine=rbd
clientname=admin
pool=volumes
rbdname=testdisk
filename=/dev/rbd0
rw=write   # write 表示顺序写，randwrite 表示随机写，read 表示顺序读，randread 表示随机读
bs=4M
size=5g     # 每个fio进程/线程的最大读写
numjobs=1
iodepth=32
direct=1   # 排除OS的IO缓存机制的影响
lockmem=1G   # 锁定所使用的内存大小
runtime=30   #  运行时间
group_reporting   # 多个job合并出报告
```

运行 `fio write.fio` 结果：_

```
write-4M: (g=0): rw=write, bs=(R) 4096KiB-4096KiB, (W) 4096KiB-4096KiB, (T) 4096KiB-4096KiB, ioengine=rbd, iodepth=32
fio-3.28
Starting 1 process
Jobs: 1 (f=1): [W(1)][39.5%][eta 00m:49s]
write-4M: (groupid=0, jobs=1): err= 0: pid=3612300: Tue Jan  9 17:57:20 2024
  write: IOPS=16, BW=67.2MiB/s (70.5MB/s)(2128MiB/31643msec); 0 zone resets
    slat (usec): min=222, max=3880, avg=767.89, stdev=545.41
    clat (msec): min=540, max=3977, avg=1898.67, stdev=595.09
     lat (msec): min=541, max=3978, avg=1899.44, stdev=595.09
    clat percentiles (msec):
     |  1.00th=[  919],  5.00th=[ 1020], 10.00th=[ 1099], 20.00th=[ 1401],
     | 30.00th=[ 1569], 40.00th=[ 1737], 50.00th=[ 1854], 60.00th=[ 1972],
     | 70.00th=[ 2072], 80.00th=[ 2366], 90.00th=[ 2702], 95.00th=[ 3071],
     | 99.00th=[ 3473], 99.50th=[ 3708], 99.90th=[ 3977], 99.95th=[ 3977],
     | 99.99th=[ 3977]
   bw (  KiB/s): min= 8192, max=163840, per=100.00%, avg=74621.67, stdev=36317.96, samples=55
   iops        : min=    2, max=   40, avg=18.22, stdev= 8.87, samples=55
  lat (msec)   : 750=0.38%, 1000=2.82%, 2000=61.28%, >=2000=35.53%
  cpu          : usr=1.00%, sys=0.29%, ctx=173, majf=0, minf=68653
  IO depths    : 1=0.2%, 2=0.4%, 4=0.8%, 8=1.5%, 16=3.0%, 32=94.2%, >=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=99.9%, 8=0.0%, 16=0.0%, 32=0.1%, 64=0.0%, >=64=0.0%
     issued rwts: total=0,532,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=32

Run status group 0 (all jobs):
  WRITE: bw=67.2MiB/s (70.5MB/s), 67.2MiB/s-67.2MiB/s (70.5MB/s-70.5MB/s), io=2128MiB (2231MB), run=31643-31643msec

Disk stats (read/write):
  rbd0: ios=0/0, merge=0/0, ticks=0/0, in_queue=0, util=0.00%
```

2.5 fio + libaio

测试块设备的读写性能，使用2.3的/dev/rbd0

准备工作

```
# 查询块设备是否已经4 KiB对齐。
sudo fdisk -lu
Device     Boot Start      End  Sectors Size Id Type
/dev/vda1  *     2048 83886046 83883999  40G 83 Linux
# 返回的结果中，Start值能被8整除即是4 KiB对齐。否则，请完成4 KiB对齐后再继续性能测试

# 安装libaio和FIO
# 先测试是否有libaio再下载: dpkg -l | grep libaio
sudo apt-get install libaio1 -y
sudo apt-get install libaio-devel -y
sudo apt-get install fio -y

# 切换目录
cd /tmp
```

开始测试

```
# 随机写IOPS
```

```
fio -direct=1 -iodepth=128 -rw=randwrite -ioengine=libaio -bs=4k -size=1G -numjobs=1 -runtime=1000 -group_reporting -filename=/dev/rbd0 -name=Rand_Write_Testing --allow_mounted_write=1

# 随机读IOPS
fio -direct=1 -iodepth=128 -rw=randread -ioengine=libaio -bs=4k -size=1G -numjobs=1 -runtime=1000 -group_reporting -filename=/dev/rbd0 -name=Rand_Read_Testing --allow_mounted_write=1

# 顺序写吞吐量
fio -direct=1 -iodepth=64 -rw=write -ioengine=libaio -bs=1024k -size=1G -numjobs=1 -runtime=1000 -group_reporting -filename=/dev/rbd0 -name=Write_PPS_Testing --allow_mounted_write=1

# 顺序读吞吐量
fio -direct=1 -iodepth=64 -rw=read -ioengine=libaio -bs=1024k -size=1G -numjobs=1 -runtime=1000 -group_reporting -filename=/dev/rbd0 -name=Read_PPS_Testing --allow_mounted_write=1

# 随机写时延
fio -direct=1 -iodepth=1 -rw=randwrite -ioengine=libaio -bs=4k -size=1G -numjobs=1 -runtime=1000 -group_reporting -filename=/dev/rbd0 -name=Rand_Write_Latency_Testing --allow_mounted_write=1

# 随机读时延
fio -direct=1 -iodepth=1 -rw=randread -ioengine=libaio -bs=4k -size=1G -numjobs=1 -runtime=1000 -group_reporting -filename=/dev/rbd0 -name=Rand_Read_Latency_Testing --allow_mounted_write=1

# 顺序写时延
fio -direct=1 -iodepth=1 -rw=write -ioengine=libaio -bs=4k -numjobs=1 -time_based=1 -runtime=1000 -group_reporting -filename=/dev/rbd0 -name=Write_Latency_Testing --allow_mounted_write=1

# 顺序读时延
fio -direct=1 -iodepth=1 -rw=read -ioengine=libaio -bs=4k -numjobs=1 -time_based=1 -runtime=1000 -group_reporting -filename=/dev/rbd0 -name=Read_Latency_Testing --allow_mounted_write=1
```

	随机写IOPS	随机读IOPS	顺序写吞吐量	顺序读吞吐量	随机写时延	随机读时延
吞吐量bw	51.7MiB/s	148MiB/s	69.2MiB/s	160MiB/s	1.744MiB/s	10.5MiB/s
每秒操作数IOPS	13.2k	37.9k	69	160	436	2675

延迟 lat	9.68ms	3.37622ms	921.71ms	398.44ms	2.29042ms	0.37270ms

随机写IOPS

```
Rand_Write_Testing: (g=0): rw=randwrite, bs=(R) 4096B-4096B, (W) 4096B-4096B, (T) 4096B-4096B, ioengine=libaio, iodepth=128
fio-3.28
Starting 1 process
Jobs: 1 (f=1): [w(1)][100.0%][w=54.0MiB/s][w=13.8k IOPS][eta 00m:00s]
Rand_Write_Testing: (groupid=0, jobs=1): err= 0: pid=3936565: Thu Jan 11 11:51:52 2024
  write: IOPS=13.2k, BW=51.7MiB/s (54.2MB/s)(1024MiB/19825msec); 0 zone resets
    slat (nsec): min=457, max=8567.6k, avg=2146.02, stdev=24154.06
    clat (usec): min=112, max=159334, avg=9675.73, stdev=5624.93
     lat (msec): min=2, max=159, avg= 9.68, stdev= 5.62
    clat percentiles (msec):
     |  1.00th=[    6],  5.00th=[    7], 10.00th=[    8], 20.00th=[    8],
     | 30.00th=[    9], 40.00th=[    9], 50.00th=[   10], 60.00th=[   10],
     | 70.00th=[   11], 80.00th=[   11], 90.00th=[   12], 95.00th=[   13],
     | 99.00th=[   21], 99.50th=[   33], 99.90th=[  110], 99.95th=[  138],
     | 99.99th=[  157]
   bw (  KiB/s): min=30240, max=56512, per=100.00%, avg=52901.95, stdev=5872.82, samples=39
   iops        : min= 7560, max=14128, avg=13225.49, stdev=1468.21, samples=39
  lat (usec)   : 250=0.01%
  lat (msec)   : 4=0.01%, 10=67.78%, 20=31.19%, 50=0.69%, 100=0.22%
  lat (msec)   : 250=0.11%
  cpu          : usr=1.71%, sys=3.40%, ctx=125382, majf=0, minf=13
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=100.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.1%
     issued rwts: total=0,262144,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=128

Run status group 0 (all jobs):
  WRITE: bw=51.7MiB/s (54.2MB/s), 51.7MiB/s-51.7MiB/s (54.2MB/s-54.2MB/s), io=1024MiB (1074MB), run=19825-19825msec

Disk stats (read/write):
  rbd0: ios=0/259688, merge=0/0, ticks=0/2510335, in_queue=2510335, util=99.55%
```

随机读IOPS

```
Rand_Read_Testing: (g=0): rw=randread, bs=(R) 4096B-4096B, (W) 4096B-4096B, (T) 4096B-4096B, ioengine=libaio, iodepth=128
fio-3.28
Starting 1 process
Jobs: 1 (f=1): [r(1)][100.0%][r=149MiB/s][r=38.0k IOPS][eta 00m:00s]
Rand_Read_Testing: (groupid=0, jobs=1): err= 0: pid=3944518: Thu Jan 11 11:53:41 2024
  read: IOPS=37.9k, BW=148MiB/s (155MB/s)(1024MiB/6918msec)
    slat (nsec): min=453, max=198633, avg=2246.93, stdev=3840.81
    clat (usec): min=60, max=13000, avg=3373.92, stdev=3680.78
     lat (usec): min=61, max=13001, avg=3376.22, stdev=3680.74
    clat percentiles (usec):
     |  1.00th=[  129],  5.00th=[  147], 10.00th=[  176], 20.00th=[  249],
     | 30.00th=[  400], 40.00th=[  586], 50.00th=[  816], 60.00th=[ 1483],
     | 70.00th=[ 7701], 80.00th=[ 8094], 90.00th=[ 8455], 95.00th=[ 8717],
     | 99.00th=[ 9110], 99.50th=[ 9241], 99.90th=[ 9503], 99.95th=[ 9634],
     | 99.99th=[10421]
   bw (  KiB/s): min=150400, max=152816, per=100.00%, avg=151569.85, stdev=761.09, samples=13
   iops        : min=37600, max=38204, avg=37892.46, stdev=190.27, samples=13
  lat (usec)   : 100=0.01%, 250=20.13%, 500=15.39%, 750=12.07%, 1000=7.35%
  lat (msec)   : 2=6.27%, 4=1.18%, 10=37.59%, 20=0.02%
  cpu          : usr=5.33%, sys=9.02%, ctx=212188, majf=0, minf=139
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=100.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.1%
     issued rwts: total=262144,0,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=128

Run status group 0 (all jobs):
   READ: bw=148MiB/s (155MB/s), 148MiB/s-148MiB/s (155MB/s-155MB/s), io=1024MiB (1074MB), run=6918-6918msec

Disk stats (read/write):
  rbd0: ios=261317/0, merge=0/0, ticks=879433/0, in_queue=879433, util=98.63%
```

顺序写吞吐量

```
Write_PPS_Testing: (g=0): rw=write, bs=(R) 1024KiB-1024KiB, (W) 1024KiB-1024KiB, (T) 1024KiB-1024KiB, ioengine=libaio, iodepth=64
fio-3.28
Starting 1 process
Jobs: 1 (f=1): [W(1)][100.0%][w=74.1MiB/s][w=74 IOPS][eta 00m:00s]
Write_PPS_Testing: (groupid=0, jobs=1): err= 0: pid=3947247: Thu Jan 11 11:54:45 2024
```

```
  write: IOPS=69, BW=69.2MiB/s (72.6MB/s)(1024MiB/14799msec); 0 zone resets
    slat (usec): min=27, max=34187, avg=112.50, stdev=1066.42
    clat (msec): min=8, max=2202, avg=921.59, stdev=532.14
     lat (msec): min=42, max=2202, avg=921.71, stdev=532.08
    clat percentiles (msec):
     |  1.00th=[   79],  5.00th=[  155], 10.00th=[  241], 20.00th=[  334],
     | 30.00th=[  558], 40.00th=[  760], 50.00th=[  911], 60.00th=[ 1070],
     | 70.00th=[ 1250], 80.00th=[ 1401], 90.00th=[ 1636], 95.00th=[ 1821],
     | 99.00th=[ 2106], 99.50th=[ 2140], 99.90th=[ 2165], 99.95th=[ 2198],
     | 99.99th=[ 2198]
   bw (  KiB/s): min=43008, max=88064, per=99.20%, avg=70290.29, stdev=8778.81, samples=28
   iops        : min=   42, max=   86, avg=68.64, stdev= 8.57, samples=28
  lat (msec)   : 10=0.10%, 50=0.10%, 100=2.25%, 250=8.50%, 500=16.99%
  lat (msec)   : 750=11.52%, 1000=17.29%, 2000=41.11%, >=2000=2.15%
  cpu          : usr=0.40%, sys=0.25%, ctx=1075, majf=0, minf=13
  IO depths    : 1=0.1%, 2=0.2%, 4=0.4%, 8=0.8%, 16=1.6%, 32=3.1%, >=64=93.8%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=99.9%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.1%, >=64=0.0%
     issued rwts: total=0,1024,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=64

Run status group 0 (all jobs):
  WRITE: bw=69.2MiB/s (72.6MB/s), 69.2MiB/s-69.2MiB/s (72.6MB/s-72.6MB/s), io=1024MiB (1074MB), run=14799-14799msec

Disk stats (read/write):
  rbd0: ios=0/1015, merge=0/253, ticks=0/908185, in_queue=908185, util=98.13%
```

顺序读吞吐量

```
Read_PPS_Testing: (g=0): rw=read, bs=(R) 1024KiB-1024KiB, (W) 1024KiB-1024KiB, (T) 1024KiB-1024KiB, ioengine=libaio, iodepth=64
fio-3.28
Starting 1 process
Jobs: 1 (f=1): [R(1)][100.0%][r=138MiB/s][r=138 IOPS][eta 00m:00s]
Read_PPS_Testing: (groupid=0, jobs=1): err= 0: pid=3953042: Thu Jan 11 11:56:17 2024
  read: IOPS=160, BW=160MiB/s (168MB/s)(1024MiB/6398msec)
    slat (usec): min=10, max=552, avg=73.21, stdev=82.57
    clat (msec): min=3, max=1147, avg=398.37, stdev=384.26
     lat (msec): min=3, max=1147, avg=398.44, stdev=384.25
    clat percentiles (msec):
     |  1.00th=[    4],  5.00th=[    5], 10.00th=[    6], 20.00th=[    8],
     | 30.00th=[   23], 40.00th=[   89], 50.00th=[  268], 60.00th=[  535],
     | 70.00th=[  709], 80.00th=[  869], 90.00th=[  936], 95.00th=[  986],
     | 99.00th=[ 1083], 99.50th=[ 1099], 99.90th=[ 1133], 99.95th=[ 1150],
     | 99.99th=[ 1150]
   bw (  KiB/s): min=110592, max=227328, per=100.00%, avg=164010.67, stdev=36305.97, samples=12
   iops        : min=  108, max=  222, avg=160.17, stdev=35.46, samples=12
  lat (msec)   : 4=4.79%, 10=19.82%, 20=3.91%, 50=5.86%, 100=7.32%
  lat (msec)   : 250=7.23%, 500=8.50%, 750=13.87%, 1000=24.90%, 2000=3.81%
  cpu          : usr=0.05%, sys=1.23%, ctx=1256, majf=0, minf=16397
  IO depths    : 1=0.1%, 2=0.2%, 4=0.4%, 8=0.8%, 16=1.6%, 32=3.1%, >=64=93.8%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=99.9%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.1%, >=64=0.0%
     issued rwts: total=1024,0,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=64

Run status group 0 (all jobs):
   READ: bw=160MiB/s (168MB/s), 160MiB/s-160MiB/s (168MB/s-168MB/s), io=1024MiB (1074MB), run=6398-6398msec

Disk stats (read/write):
  rbd0: ios=995/0, merge=0/0, ticks=374371/0, in_queue=374371, util=98.54%
```

随机写时延

```
Rand_Write_Latency_Testing: (g=0): rw=randwrite, bs=(R) 4096B-4096B, (W) 4096B-4096B, (T) 4096B-4096B, ioengine=libaio, iodepth=1
fio-3.28
Starting 1 process
Jobs: 1 (f=1): [w(1)][100.0%][w=1505KiB/s][w=376 IOPS][eta 00m:00s]
Rand_Write_Latency_Testing: (groupid=0, jobs=1): err= 0: pid=3956461: Thu Jan 11 12:07:15 2024
  write: IOPS=436, BW=1744KiB/s (1786kB/s)(1024MiB/601188msec); 0 zone resets
    slat (nsec): min=795, max=2047.3k, avg=8616.42, stdev=9983.26
    clat (nsec): min=1185, max=623590k, avg=2281590.00, stdev=5502401.52
     lat (usec): min=1322, max=623593, avg=2290.42, stdev=5502.51
    clat percentiles (usec):
     |  1.00th=[  1582],  5.00th=[  1680], 10.00th=[  1729], 20.00th=[  1811],
     | 30.00th=[  1860], 40.00th=[  1909], 50.00th=[  1942], 60.00th=[  1991],
     | 70.00th=[  2040], 80.00th=[  2114], 90.00th=[  2278], 95.00th=[  2737],
     | 99.00th=[ 11207], 99.50th=[ 16581], 99.90th=[ 22676], 99.95th=[ 36439],
     | 99.99th=[312476]
   bw (  KiB/s): min=  448, max= 2248, per=100.00%, avg=1747.94, stdev=293.72, samples=1200
   iops        : min=  112, max=  562, avg=436.89, stdev=73.45, samples=1200
  lat (usec)   : 2=0.01%, 4=0.01%
  lat (msec)   : 2=61.32%, 4=36.37%, 10=1.18%, 20=0.94%, 50=0.14%
  lat (msec)   : 100=0.02%, 250=0.01%, 500=0.02%, 750=0.01%
  cpu          : usr=0.28%, sys=0.56%, ctx=269359, majf=0, minf=15
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     issued rwts: total=0,262144,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
  WRITE: bw=1744KiB/s (1786kB/s), 1744KiB/s-1744KiB/s (1786kB/s-1786kB/s), io=1024MiB (1074MB), run=601188-601188msec

Disk stats (read/write):
  rbd0: ios=45/262124, merge=0/0, ticks=20/594599, in_queue=594619, util=100.00%
```

随机读时延

```
Rand_Read_Latency_Testing: (groupid=0, jobs=1): err= 0: pid=4001754: Thu Jan 11 12:12:28 2024
  read: IOPS=2675, BW=10.5MiB/s (11.0MB/s)(1024MiB/97976msec)
    slat (nsec): min=688, max=156396, avg=3788.07, stdev=3498.03
    clat (usec): min=56, max=42445, avg=368.83, stdev=311.69
     lat (usec): min=57, max=42448, avg=372.70, stdev=311.84
    clat percentiles (usec):
     |  1.00th=[  198],  5.00th=[  223], 10.00th=[  247], 20.00th=[  281],
     | 30.00th=[  314], 40.00th=[  334], 50.00th=[  351], 60.00th=[  371],
     | 70.00th=[  396], 80.00th=[  437], 90.00th=[  494], 95.00th=[  545],
     | 99.00th=[  660], 99.50th=[  758], 99.90th=[ 1156], 99.95th=[ 2212],
     | 99.99th=[15401]
   bw (  KiB/s): min= 8368, max=12376, per=100.00%, avg=10705.60, stdev=665.60, samples=195
   iops        : min= 2092, max= 3094, avg=2676.40, stdev=166.40, samples=195
  lat (usec)   : 100=0.01%, 250=11.27%, 500=79.46%, 750=8.74%, 1000=0.35%
  lat (msec)   : 2=0.12%, 4=0.02%, 10=0.02%, 20=0.01%, 50=0.01%
  cpu          : usr=0.67%, sys=1.73%, ctx=267155, majf=0, minf=13
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     issued rwts: total=262144,0,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
   READ: bw=10.5MiB/s (11.0MB/s), 10.5MiB/s-10.5MiB/s (11.0MB/s-11.0MB/s), io=1024MiB (1074MB), run=97976-97976msec

Disk stats (read/write):
  rbd0: ios=261938/0, merge=0/0, ticks=95703/0, in_queue=95703, util=99.96%
```

相关文档

管理指南 Red Hat Ceph Storage 4 | Red Hat Customer Portal

理解 OpenStack + Ceph (8): 基本的 Ceph 性能测试工具和方法 - SammyLiu - 博客园

如何在Linux实例中使用FIO工具测试块存储性能_云服务器 ECS(ECS)-阿里云帮助中心