# Ceph filestore\_op\_threads 测试报告

## 基本情况

针对某项目的 Ceph 集群优化分析中给出了初步的配置参数调整方案，方案中大部份配置参数的调整对于 peering 这一过程的影响都不明显，只有 filestore\_op\_threads 这一配置参数的改动影响较大，具体测试数据和分析参见“[Ceph peering 相关参数测试](https://docs.qq.com/doc/DYndyYlJNSE9UckR1)”。

不过其他同事也提出 Ceph 集群本身的性能不在于线程并发，在 peering 相关参数的测试中，虽然增加 filestore\_op\_threads 看起来缩短了 peering 这一过程，但是同时也会增加客户端 I/O 的延时，降低客户端的性能。

针对这一情况，继续对 filestore\_op\_threads 的调整可能造成的影响进行测试验证。

## 测试方案

### 2.1 测试环境

节点：x.x.x.x、x.x.x.x、x.x.x.x

Ceph 集群：

* OSD 数量， 15
* 每个 OSD 的 PG 数量，在 1658 ~ 1810 之间。

测试 RBD ：

* volulmes/fio\_background ，100G ，用于背景 I/O
* volumes/fio\_test ，100G ，用于性能测试

### 2.2 测试方法

使用 fio 进行测试，blocksize 选择 4k 和 4M 两种，分别进行 randread/randwrite/read/write 性能测试。

针对于不同的配置参数，分别进行 3 次测试：

1. 直接测试 I/O 性能指标，作为当前配置下的基准数据，每项测试持续时间 60s ，测试之间间隔 60s；
2. 增加专门的背景 I/O （背景 I/O 选择一个 500G 的 rbd ，持续执行 blocksize 为 4k 的 randwrite ），重新测试 I/O 性能；
3. 保持背景 I/O 继续运行，停止其中一个 ceph-osd ，等待 5 分钟，重新启动 ceph-osd ，记录整个集群完全恢复（所有 PG 恢复到 active+clean 状态）需要的时间。

### 2.3 测试命令

* 性能测试：

# fio --direct=1 --numjobs=4 -rw=randread --bs=4k --iodepth=4 --name=4\_randread\_4k\_4 --ioengine=rbd --clientname=cinder --pool=volumes --rbdname=fio\_test --time\_based --runtime=60 --group\_reporting

# sleep 60

# fio --direct=1 --numjobs=4 -rw=randwrite --bs=4k --iodepth=4 --name=4\_randwrite\_4k\_4 --ioengine=rbd --clientname=cinder --pool=volumes --rbdname=fio\_test --time\_based --runtime=60 --group\_reporting

# sleep 60

# fio --direct=1 --numjobs=4 -rw=read --bs=4k --iodepth=4 --name=4\_write\_4k\_4 --ioengine=rbd --clientname=cinder --pool=volumes --rbdname=fio\_test --time\_based --runtime=60 --group\_reporting

# sleep 60

# fio --direct=1 --numjobs=4 -rw=write --bs=4k --iodepth=4 --name=4\_read\_4k\_4 --ioengine=rbd --clientname=cinder --pool=volumes --rbdname=fio\_test --time\_based --runtime=60 --group\_reporting

# sleep 60

* 背景 I/O

# fio --direct=1 --numjobs=4 -rw=randwrite --bs=4k --iodepth=4 --name=4\_randwrite\_4k\_4 --ioengine=rbd --clientname=cinder --pool=volumes --rbdname=fio\_background --time\_based --runtime=1d --group\_reporting

## 测试数据

### 3.1 数据说明

测试并不是在短时间完成的，其中下列数据中前三项，即默以配置、filestore\_op\_threads = 10 、filestore\_op\_threads = 6 完成时间较为按近，后两项，即filestore\_op\_threads = 4 以及 filestore\_op\_threads 使用默认同时保留其它优化配置执行时间在几天之后，测试结果的一些明显差异（如 4k I/O 中 randwrite/read/write IOPS 大幅增加）无法直接采信，因为不能排除整个 Ceph 集群有其它变动。最后两项为时间较为接近的对照测试。

### 3.2 4k I/O

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | randread | randwrite | read | write |
| 默认配置 | iops: 11126bw: 43.46MB/s | iops: 260bw: 1.01MB/s | iops: 11849bw: 46.29MB/s | iops: 288bw: 1.12MB/s |
| filestore\_op\_threads = 10 | iops: 10561bw: 41.25 MB/s | iops: 256bw: 1.00 MB/s | iops: 10970bw: 42.85 MB/s | iops: 279bw: 1.09 MB/s |
| filestore\_op\_threads = 6 | iops: 10938bw: 42.73 MB/s | iops: 258bw: 1.01 MB/s | iops: 10965bw: 42.83 MB/s | iops: 276bw: 1.08 MB/s |
| filestore\_op\_threads = 4 | iops: 4484bw: 17.51 MB/s | iops: 390bw: 1.52 MB/s | iops: 10342bw: 40.40 MB/s | iops: 468bw: 1.83 MB/s |
| filestore\_op\_threads 恢复默认，保留其它优化配置调整 | iops: 5804bw: 22.67 MB/s | iops: 398bw: 1.55 MB/s | iops: 13175bw: 51.47 MB/s | iops: 486bw: 1.90 MB/s  |
| filestore\_op\_threads 恢复默认，保留其它优化配置调整（对照测试） | iops: 13672bw: 53.40 MB/s | iops: 407bw: 1.59 MB/s | iops: 18605bw: 72.68 MB/s | iops: 476bw: 1.86 MB/s |
| 默认配置（对照测试） | iops: 11013bw: 43.02 MB/s | iops: 409bw: 1.60 MB/s | iops: 19782bw: 77.27 MB/s | iops: 498bw: 1.95 MB/s |
| filestore\_op\_threads = 6（对照测试） | iops: 14976bw: 58.50 MB/s | iops: 402bw: 1.57 MB/s | iops: 19097bw: 74.60 MB/s | iops: 481bw: 1.88 MB/s |
| 默认配置（对照测试2 ，numjobs=1 iodepth=128） | iops: 7971bw: 31.14 MB/s | iops: 2435bw: 9.51 MB/s | iops: 9592bw: 37.47 MB/s | iops: 3425bw: 13.38 MB/s |
| filestore\_op\_threads = 6（对照测试2 ，numjobs=1 iodepth=128） | iops: 18357bw: 71.71 MB/s | iops: 2392bw: 9.34 MB/s | iops: 7430bw: 29.02 MB/s | iops: 2332bw: 9.11 MB/s |

### 3.3 4k I/O （增加背景 I/O ）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | randread | randwrite | read | write |
| 默认配置 | iops: 10785bw: 42.13 MB/s | iops: 247bw: 0.96 MB/s | iops: 11228bw: 43.86 MB/s | iops: 278bw: 1.08 MB/s |
| filestore\_op\_threads = 10 | iops: 10647bw: 41.59 MB/s | iops: 245bw: 0.96 MB/s | iops: 11058bw: 43.19 MB/s | iops: 265bw: 1.03 MB/ |
| filestore\_op\_threads = 6 | iops: 10569bw: 41.28 MB/s | iops: 248bw: 0.97 MB/s | iops: 10771bw: 42.07 MB/s | iops: 265bw: 1.03 MB/s |
| filestore\_op\_threads = 4 | iops: 7162bw: 27.97 MB/s | iops: 375bw: 1.46 MB/s | iops: 17764bw: 69.39 MB/s | iops: 420bw: 1.64 MB/s |
| filestore\_op\_threads 恢复默认，保留其它优化配置调整 | iops: 10513bw: 41.06 MB/s | iops: 368bw: 1.44 MB/s | iops: 18200bw: 71.09 MB/s | iops: 435bw: 1.70 MB/s |
| filestore\_op\_threads 恢复默认，保留其它优化配置调整（对照测试） | iops: 9457bw: 36.94 MB/s | iops: 373bw: 1.46 MB/s | iops: 15883bw: 62.04 MB/s | iops: 444bw: 1.73 MB/s |
| 默认配置（对照测试） | iops: 10908bw: 42.61 MB/s | iops: 384bw: 1.50 MB/s | iops: 19234bw: 75.13 MB/s | iops: 466bw: 1.82 MB/s |
| filestore\_op\_threads = 6（对照测试） | iops: 14004bw: 54.70 MB/s | iops: 371bw: 1.45 MB/s | iops: 17914bw: 69.98 MB/s | iops: 454bw: 1.77 MB/s |
| 默认配置（对照测试2 ，numjobs=1 iodepth=128） | iops: 1054bw: 4.12 MB/s | iops: 1220bw: 4.77 MB/s | iops: 4627bw: 18.07 MB/s | iops: 1100bw: 4.30 MB/s |
| filestore\_op\_threads = 6（对照测试2 ，numjobs=1 iodepth=128） | iops: 3710bw: 14.49 MB/s | iops: 1134bw: 4.43 MB/s | iops: 5632bw: 22.00 MB/s | iops: 866bw: 3.38 MB/s |

### 3.4 4m I/O

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | randread | randwrite | read | write |
| 默认配置 | iops: 47bw: 188.89 MB/s | iops: 15bw: 59.45 MB/s | iops: 51bw: 202.02 MB/s | iops: 15bw: 60.55 MB/ |
| filestore\_op\_threads = 10 | iops: 45bw: 178.25 MB/s | iops: 15bw: 61.90 MB/s | iops: 45bw: 179.65 MB/s | iops: 15bw: 60.50 MB/s |
| filestore\_op\_threads = 6 | iops: 44bw: 175.29 MB/s | iops: 15bw: 58.81 MB/s | iops: 45bw: 180.01 MB/s | iops: 15bw: 61.74 MB/s |
| filestore\_op\_threads = 4 | iops: 43bw: 172.19 MB/s | iops: 15bw: 59.80 MB/s | iops: 43bw: 171.81 MB/s | iops: 15bw: 60.98 MB/s |
| filestore\_op\_threads 恢复默认，保留其它优化配置调整 | iops: 43bw: 172.68 MB/s | iops: 15bw: 60.69 MB/s | iops: 43bw: 173.38 MB/s | iops: 15bw: 59.99 MB/s |
| filestore\_op\_threads 恢复默认，保留其它优化配置调整（对照测试） | iops: 43bw: 171.16 MB/s | iops: 15bw: 60.50 MB/s | iops: 43bw: 173.87 MB/s | iops: 16bw: 62.85 MB/s |
| 默认配置（对照测试） | iops: 43bw: 172.19 MB/s | iops: 16bw: 62.89 MB/s | iops: 44bw: 174.63 MB/s | iops: 16bw: 63.39 MB/s |
| filestore\_op\_threads = 6（对照测试） | iops: 43bw: 172.24 MB/s | iops: 16bw: 62.18 MB/s | iops: 43bw: 173.92 MB/s | iops: 16bw: 62.58 MB/s |
| 默认配置（对照测试2 ，numjobs=8 iodepth=32） | iops: 41bw: 163.43 MB/s | iops: 21bw: 84.66 MB/s | iops: 41bw: 163.79 MB/s | iops: 21bw: 82.36 MB/s |
| filestore\_op\_threads = 6（对照测试2 ，numjobs=8 iodepth=32） | iops: 41bw: 165.03 MB/s | iops: 21bw: 83.52 MB/s | iops: 40bw: 161.74 MB/s | iops: 22bw: 86.15 MB/s |

### 3.5 4m I/O （增加背景 I/O ）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | randread | randwrite | read | write |
| 默认配置 | iops: 45bw: 179.45 MB/s | iops: 14bw: 57.77 MB/s | iops: 45bw: 180.75 MB/s | iops: 15bw: 61.25 MB/s |
| filestore\_op\_threads = 10 | iops: 44bw: 176.71 MB/s | iops: 15bw: 58.75 MB/s | iops: 45bw: 178.76 MB/s | iops: 15bw: 58.82 MB/s |
| filestore\_op\_threads = 6 | iops: 44bw: 175.45 MB/s | iops: 14bw: 57.32 MB/s | iops: 45bw: 178.09 MB/s | iops: 15bw: 60.13 MB/s |
| filestore\_op\_threads = 4 | iops: 43bw: 172.25 MB/s | iops: 15bw: 59.77 MB/s | iops: 43bw: 172.57 MB/s | iops: 15bw: 58.15 MB/s |
| filestore\_op\_threads 恢复默认，保留其它优化配置调整 | iops: 43bw: 171.16 MB/s | iops: 15bw: 59.32 MB/s | iops: 43bw: 173.03 MB/s | iops: 15bw: 59.78 MB/s |
| filestore\_op\_threads 恢复默认，保留其它优化配置调整（对照测试） | iops: 43bw: 171.56 MB/s | iops: 15bw: 60.75 MB/s | iops: 43bw: 173.62 MB/s | iops: 16bw: 62.55 MB/s |
| 默认配置（对照测试） | iops: 43bw: 171.21 MB/s | iops: 15bw: 61.55 MB/s | iops: 43bw: 173.31 MB/s | iops: 15bw: 60.67 MB/s |
| filestore\_op\_threads = 6（对照测试） | iops: 42bw: 169.74 MB/s | iops: 14bw: 57.56 MB/s | iops: 43bw: 171.55 MB/s | iops: 16bw: 62.03 MB/s |
| 默认配置（对照测试2 ，numjobs=8 iodepth=32） | iops: 39bw: 157.61 MB/s | iops: 13bw: 50.84 MB/s | iops: 39bw: 156.49 MB/s | iops: 13bw: 53.03 MB/s |
| filestore\_op\_threads = 6（对照测试2 ，numjobs=8 iodepth=32） | iops: 39bw: 156.39 MB/s | iops: 11bw: 42.42 MB/s | iops: 39bw: 156.35 MB/s | iops: 13bw: 52.75 MB/s |

### 3.6 模拟 OSD 故障集群恢复时间

|  |  |
| --- | --- |
|  | 恢复时间 |
| 默认配置 | 10m25s |
| filestore\_op\_threads = 10 | 10m54s |
| filestore\_op\_threads = 6 | 10m36s |
| filestore\_op\_threads = 4 | 14m38s |
| filestore\_op\_threads 恢复默认，保留其它优化配置调整 | 28m |
| filestore\_op\_threads 恢复默认，保留其它优化配置调整（对照测试） | 25m55s |
| 默认配置（对照测试） | 27m52s |
| filestore\_op\_threads = 6（对照测试） | 25m34s |
| 默认配置（对照测试2） | 16m04s |
| filestore\_op\_threads = 6（对照测试2） | 16m01s |

## 结论

从上面测试数据可以看出，filestore\_op\_threads 的确可能增加客户端 I/O 的延时从而造成性能下降，不过下降幅度不是特别大，可以认为 filestore\_op\_threads 增加对于性能的影响较小，而且可能产生负面作用。

在 Ceph peering 相关参数测试中 filestore\_op\_threads 调整产生的效果非常明显，主要的原因应该是其它 ramdisk OSD 可以并发处理大量 I/O 操作。对于普通的机械硬盘，单纯增加 filestore\_op\_threads 的影响并不大。

综合多次测试的结果，仍然可以考虑在配置参数优化方案中小幅度增大 filestore\_op\_threads 的取值，如按对照测试组中的取值，线程数设置为 6 。