



# OPPO百万级高并发mongodb集群性能数十倍提升优化实践

OPPO互联网文档数据库研发负责人  
杨亚洲



# 背景

1. 峰值TPS超过100万/秒
2. 主要为写流量(insert+update)
3. 读流量几百上千
4. 主从读写分离
5. 分片模式
6. 集群总文档数百亿条
7. 文档二天后散列任意时间点过期



|            |       |    |       |     |      |      |   |       |       |   |       |       |     |   |      |       |       |        |                |                |                  |     |
|------------|-------|----|-------|-----|------|------|---|-------|-------|---|-------|-------|-----|---|------|-------|-------|--------|----------------|----------------|------------------|-----|
| 14.10000   | 2499  | *0 | 66591 | *0  | 280  | 622  | 0 | 5.9%  | 75.1% | 0 | 26.36 | 38.2  | n/a | 0 | 1127 | 30.7m | 70.7m | 71.24  | shard_4093D617 | PRI            |                  |     |
| 1.0000     | 2397  | *0 | 69590 | *0  | 863  | 901  | 0 | 3.5%  | 75.0% | 0 | 78.06 | 56.86 | n/a | 0 | 10   | 1111  | 41.2m | 68.7m  | 3104           | shard_4093D617 | PRI              |     |
| 1.0001     | 2654  | *0 | 74101 | *0  | 347  | 104  | 0 | 19.3% | 73.4% | 0 | 63.06 | 43.66 | n/a | 0 | 12   | 212   | 42.9m | 85.5m  | 3461           | shard_A942EA94 | PRI              |     |
| 1.0002     | 2442  | *0 | 64236 | *0  | 1044 | 928  | 0 | 6.3%  | 75.0% | 0 | 118.6 | 46.06 | n/a | 0 | 1    | 114   | 38.3m | 82.5m  | 3475           | shard_C81C782E | PRI              |     |
| 1.0003     | 2266  | *0 | 72669 | *0  | 28   | 802  | 0 | 5.0%  | 75.0% | 0 | 62.46 | 45.76 | n/a | 0 | 1    | 1127  | 42.0m | 53.7m  | 3524           | shard_66258C06 | PRI              |     |
| 1.0004     | 2660  | *0 | 70757 | *0  | 644  | 852  | 0 | 4.2%  | 75.0% | 0 | 117.6 | 46.36 | n/a | 0 | 1    | 110   | 41.8m | 69.9m  | 4160           | shard_4FC5EC6E | PRI              |     |
| 1.0005     | 2558  | *0 | 67990 | *0  | 938  | 1108 | 0 | 3.8%  | 75.0% | 0 | 59.26 | 45.06 | n/a | 0 | 1    | 110   | 40.5m | 67.5m  | 3487           | shard_F0A9938E | PRI              |     |
| 1.0006     | 2639  | *0 | 74750 | *0  | 770  | 679  | 0 | 7.2%  | 62.2% | 0 | 61.86 | 43.96 | n/a | 0 | 1    | 115   | 43.9m | 73.5m  | 3667           | shard_A899D725 | PRI              |     |
| 1.0007     | 2675  | *0 | 70096 | *0  | 358  | 653  | 0 | 12.5% | 67.1% | 0 | 61.86 | 43.96 | n/a | 0 | 1    | 113   | 41.1m | 48.3m  | 3798           | shard_143EC928 | PRI              |     |
| 1.0008     | 2329  | *0 | 59436 | *0  | 1059 | 802  | 0 | 4.2%  | 75.0% | 0 | 110.6 | 45.26 | n/a | 0 | 1    | 113   | 35.5m | 59.0m  | 3466           | shard_A71FA15C | PRI              |     |
| 1.0009     | 2322  | *0 | 71195 | *0  | 562  | 665  | 0 | 3.9%  | 75.0% | 0 | 56.56 | 36.26 | n/a | 0 | 1    | 114   | 41.6m | 49.2m  | 3437           | shard_110AFE67 | PRI              |     |
| 1.0010     | 2438  | *0 | 84176 | *0  | 484  | 972  | 0 | 5.6%  | 80.1% | 0 | 58.86 | 48.96 | n/a | 0 | 1    | 114   | 41.6m | 57.2m  | 4752           | shard_A0FB523  | PRI              |     |
| 1.0011     | 2039  | *0 | 67802 | *0  | 856  | 852  | 0 | 3.6%  | 75.0% | 0 | 78.06 | 56.86 | n/a | 0 | 1    | 113   | 40.1m | 66.7m  | 3105           | shard_4093D617 | PRI              |     |
| 1.0012     | 1970  | *0 | 62357 | *0  | 360  | 364  | 0 | 19.1% | 73.4% | 0 | 63.06 | 43.66 | n/a | 0 | 1    | 111   | 36.4m | 76.9m  | 3461           | shard_A942EA94 | PRI              |     |
| 1.0013     | 2028  | *0 | 62081 | *0  | 751  | 630  | 0 | 6.1%  | 75.0% | 0 | 118.6 | 46.06 | n/a | 0 | 1    | 112   | 36.6m | 79.0m  | 3475           | shard_C81C782E | PRI              |     |
| 1.0014     | 1942  | *0 | 65749 | *0  | 80   | 412  | 0 | 5.1%  | 75.0% | 0 | 62.46 | 45.76 | n/a | 0 | 1    | 112   | 37.9m | 46.1m  | 3601           | shard_66258C06 | PRI              |     |
| 1.0015     | 1969  | *0 | 59788 | *0  | 705  | 864  | 0 | 4.5%  | 75.2% | 0 | 117.6 | 46.36 | n/a | 0 | 1    | 115   | 35.4m | 59.0m  | 4043           | shard_4FC5EC6E | PRI              |     |
| 1.0016     | 2002  | *0 | 57297 | *0  | 968  | 1174 | 0 | 3.8%  | 75.0% | 0 | 59.26 | 45.06 | n/a | 0 | 1    | 110   | 34.4m | 56.9m  | 3487           | shard_F0A9938E | PRI              |     |
| 1.0017     | 2118  | *0 | 61419 | *0  | 821  | 1115 | 0 | 6.3%  | 62.3% | 0 | 61.86 | 43.96 | n/a | 0 | 1    | 116   | 36.4m | 60.7m  | 3668           | shard_A899D725 | PRI              |     |
| 1.0018     | 2251  | *0 | 67969 | *0  | 355  | 689  | 0 | 12.2% | 67.3% | 0 | 61.86 | 43.96 | n/a | 0 | 1    | 110   | 39.9m | 47.0m  | 3798           | shard_143EC928 | PRI              |     |
| 1.0019     | 1654  | *0 | 66162 | *0  | 819  | 1083 | 0 | 4.3%  | 75.0% | 0 | 110.6 | 45.26 | n/a | 0 | 1    | 119   | 39.2m | 64.9m  | 3478           | shard_A71FA15C | PRI              |     |
| 1.0020     | 2011  | *0 | 69408 | *0  | 576  | 676  | 0 | 3.8%  | 75.0% | 0 | 56.56 | 36.26 | n/a | 0 | 1    | 127   | 40.6m | 48.0m  | 3435           | shard_110AFE67 | PRI              |     |
| 1.0021     | 2012  | *0 | 84269 | *0  | 359  | 321  | 0 | 5.5%  | 80.1% | 0 | 58.86 | 48.96 | n/a | 0 | 1    | 161   | 47.2m | 56.7m  | 4758           | shard_A0FB523  | PRI              |     |
| 1.0022     | 1806  | 9  | 76857 | *0  | 849  | 886  | 0 | 3.6%  | 75.0% | 0 | 78.06 | 56.86 | n/a | 0 | 1    | 121   | 45.2m | 75.4m  | 3107           | shard_4093D617 | PRI              |     |
| 1.0023     | 1852  | *0 | 79962 | *0  | 362  | 886  | 0 | 19.2% | 73.6% | 0 | 63.06 | 43.66 | n/a | 0 | 1    | 119   | 46.7m | 89.4m  | 3462           | shard_A942EA94 | PRI              |     |
| 1.0024     | 1735  | *0 | 71078 | *0  | 1071 | 92   | 0 | 5.7%  | 75.0% | 0 | 118.6 | 46.06 | n/a | 0 | 1    | 111   | 42.0m | 90.1m  | 3473           | shard_C81C782E | PRI              |     |
| 1.0025     | 1727  | *0 | 80514 | *0  | 356  | 353  | 0 | 4.8%  | 75.0% | 0 | 62.46 | 45.76 | n/a | 0 | 1    | 111   | 46.5m | 55.0m  | 3612           | shard_66258C06 | PRI              |     |
| 1.0026     | 1838  | *0 | 75644 | *0  | 688  | 839  | 0 | 4.2%  | 75.0% | 0 | 117.6 | 46.36 | n/a | 0 | 1    | 113   | 44.4m | 73.9m  | 3885           | shard_4FC5EC6E | PRI              |     |
| 1.0027     | 1816  | *0 | 71726 | *0  | 1031 | 1220 | 0 | 3.7%  | 75.0% | 0 | 59.26 | 45.16 | n/a | 0 | 1    | 124   | 42.7m | 70.6m  | 3487           | shard_F0A9938E | PRI              |     |
| 1.0028     | 1921  | *0 | 79747 | *0  | 1021 | 1377 | 0 | 5.6%  | 62.4% | 0 | 61.86 | 43.96 | n/a | 0 | 1    | 110   | 47.3m | 78.4m  | 3683           | shard_A899D725 | PRI              |     |
| 1.0029     | 1858  | *0 | 77566 | *0  | 326  | 520  | 0 | 12.4% | 67.5% | 0 | 61.86 | 43.96 | n/a | 0 | 1    | 116   | 45.2m | 53.2m  | 3798           | shard_143EC928 | PRI              |     |
| 1.0030     | 1566  | *0 | 72560 | *0  | 907  | 1119 | 0 | 4.2%  | 75.0% | 1 | 110.6 | 45.26 | n/a | 0 | 1    | 113   | 42.9m | 71.3m  | 3478           | shard_A71FA15C | PRI              |     |
| 1.0031     | 1639  | *0 | 79236 | *0  | 501  | 131  | 0 | 3.9%  | 75.0% | 0 | 56.56 | 36.26 | n/a | 0 | 1    | 114   | 45.9m | 54.2m  | 3431           | shard_110AFE67 | PRI              |     |
| 1.0032     | 1694  | *0 | 88653 | *0  | 358  | 355  | 0 | 5.6%  | 80.1% | 0 | 58.86 | 48.96 | n/a | 0 | 1    | 135   | 50.0m | 59.8m  | 4752           | shard_A0FB523  | PRI              |     |
| 1.0033     | 1912  | *0 | 71707 | *0  | 1164 | 1524 | 0 | 3.5%  | 75.1% | 0 | 78.06 | 56.86 | n/a | 0 | 1    | 116   | 43.0m | 70.8m  | 3062           | shard_4093D617 | PRI              |     |
| 1.0034     | 1850  | *0 | 71947 | *0  | 373  | 335  | 0 | 19.2% | 73.8% | 0 | 63.06 | 43.66 | n/a | 0 | 1    | 119   | 42.1m | 101.1m | 3472           | shard_A942EA94 | PRI              |     |
| 1.0035     | 1839  | *0 | 66046 | *0  | 1065 | 534  | 0 | 5.3%  | 75.0% | 0 | 118.6 | 46.06 | n/a | 0 | 1    | 119   | 39.2m | 84.8m  | 3438           | shard_C81C782E | PRI              |     |
| 1.0036     | 1764  | *0 | 72823 | *0  | 372  | 391  | 0 | 4.5%  | 75.0% | 0 | 62.46 | 45.76 | n/a | 0 | 1    | 1128  | 42.2m | 49.9m  | 3620           | shard_66258C06 | PRI              |     |
| 1.0037     | 1844  | *0 | 68391 | *0  | 269  | 483  | 0 | 4.5%  | 75.0% | 0 | 117.6 | 46.46 | n/a | 0 | 1    | 112   | 39.7m | 64.8m  | 3712           | shard_4FC5EC6E | PRI              |     |
| 1.0038     | 1845  | *0 | 65310 | *0  | 1115 | 1397 | 0 | 3.7%  | 75.1% | 0 | 59.16 | 45.16 | n/a | 0 | 1    | 116   | 39.2m | 65.0m  | 3487           | shard_F0A9938E | PRI              |     |
| 1.0039     | 1878  | *0 | 72202 | *0  | 1140 | 1473 | 0 | 5.4%  | 62.5% | 0 | 61.86 | 43.96 | n/a | 0 | 1    | 111   | 43.2m | 71.3m  | 3688           | shard_A899D725 | PRI              |     |
| 1.0040     | 2044  | *0 | 70414 | 24  | 335  | 520  | 0 | 12.2% | 67.8% | 0 | 61.86 | 43.96 | n/a | 0 | 1    | 112   | 43.2m | 53.2m  | 3798           | shard_143EC928 | PRI              |     |
| 1.0041     | 1540  | *0 | 61276 | *0  | 867  | 1117 | 0 | 4.1%  | 75.0% | 0 | 110.6 | 45.26 | n/a | 0 | 1    | 118   | 36.5m | 60.4m  | 3488           | shard_A71FA15C | PRI              |     |
| 1.0042     | 1738  | *0 | 73710 | *0  | 495  | 416  | 0 | 3.8%  | 75.0% | 1 | 56.56 | 36.36 | n/a | 0 | 1    | 113   | 42.8m | 50.5m  | 3440           | shard_110AFE67 | PRI              |     |
| 1.0043     | 1831  | *0 | 81740 | *0  | 405  | 761  | 0 | 5.5%  | 80.1% | 0 | 58.86 | 49.06 | n/a | 0 | 1    | 113   | 46.6m | 55.4m  | 4752           | shard_A0FB523  | PRI              |     |
| 1.0044     | 1781  | *0 | 70994 | *0  | 1180 | 1525 | 0 | 3.4%  | 75.1% | 0 | 77.96 | 56.86 | n/a | 0 | 1    | 117   | 42.6m | 70.6m  | 3062           | shard_4093D617 | PRI              |     |
| 1.0045     | 1803  | *0 | 69043 | *0  | 374  | 641  | 0 | 19.1% | 73.8% | 0 | 63.06 | 43.66 | n/a | 0 | 1    | 117   | 40.4m | 98.8m  | 3472           | shard_A942EA94 | PRI              |     |
| 1.0046     | 1744  | *0 | 65738 | *0  | 1081 | 1009 | 0 | 5.0%  | 75.0% | 0 | 118.6 | 46.06 | n/a | 0 | 1    | 115   | 39.2m | 83.8m  | 3391           | shard_C81C782E | PRI              |     |
| 1.0047     | 1723  | *0 | 70072 | *0  | 296  | 162  | 0 | 4.4%  | 75.0% | 0 | 62.46 | 45.76 | n/a | 0 | 1    | 111   | 40.4m | 48.0m  | 3630           | shard_66258C06 | PRI              |     |
| 1.0048     | 1703  | *0 | 65950 | *0  | 571  | 520  | 0 | 4.2%  | 75.0% | 0 | 117.6 | 46.46 | n/a | 0 | 1    | 111   | 38.5m | 66.5m  | 3707           | shard_4FC5EC6E | PRI              |     |
| 1.0049     | 1689  | *0 | 62819 | *0  | 1183 | 1524 | 0 | 3.6%  | 75.0% | 0 | 59.16 | 45.16 | n/a | 0 | 1    | 118   | 37.9m | 62.5m  | 3487           | shard_F0A9938E | PRI              |     |
| 1.0050     | 1770  | *0 | 70549 | *0  | 1132 | 1455 | 0 | 4.9%  | 62.6% | 0 | 61.86 | 43.96 | n/a | 0 | 1    | 112   | 42.2m | 69.9m  | 3688           | shard_A899D725 | PRI              |     |
| 1.0051     | 1714  | *0 | 68484 | *0  | 343  | 598  | 0 | 12.3% | 68.0% | 0 | 61.86 | 43.96 | n/a | 0 | 1    | 138   | 41.7m | 49.1m  | 3822           | shard_143EC928 | PRI              |     |
| 1.0052     | 1226  | *0 | 48998 | *0  | 856  | 1089 | 0 | 4.1%  | 75.0% | 0 | 110.6 | 45.26 | n/a | 0 | 1    | 111   | 29.4m | 48.7m  | 3478           | shard_A71FA15C | PRI              |     |
| 1.0053     | 1793  | 5  | 73368 | 136 | 515  | 641  | 0 | 3.7%  | 75.0% | 0 | 56.56 | 36.36 | n/a | 0 | 1    | 118   | 42.7m | 50.6m  | 3462           | shard_110AFE67 | PRI              |     |
| 1.0054     | 1693  | *0 | 81240 | *0  | 451  | 882  | 0 | 5.5%  | 80.0% | 0 | 58.86 | 49.06 | n/a | 0 | 1    | 1175  | 46.5m | 55.2m  | 4751           | shard_A0FB523  | PRI              |     |
| 1.150.0000 | 22940 | *0 | 56997 | *0  | 129  | 2508 | 0 | 8.6%  | 75.7% | 0 | 76.96 | 63.26 | n/a | 0 | 1    | 110   | 31.0  | 21.9m  | 50.6m          | 2028           | shard_4093D617   | PRI |
| 1.162.1000 | 16525 | *0 | 64880 | *0  | 774  | 1158 | 0 | 8.2%  | 75.9% | 0 | 60.86 | 45.56 | n/a | 0 | 1    | 110   | 11.6  | 19.8m  | 46.3m          | 1237           | shard_A942EA94   | PRI |
| 1.162.1001 | 16244 | *0 | 63819 | *0  | 562  | 3050 | 0 | 9.0%  | 75.7% | 0 | 118.6 | 45.06 | n/a | 0 | 1    | 110   | 21.68 | 19.8m  | 61.4m          | 1925           | shard_C81C782E   | PRI |
| 1.162.1002 | 16528 | *0 | 61351 | *0  | 534  | 806  | 0 | 5.1%  | 75.0% | 0 | 62.46 | 45.36 | n/a | 0 | 1    | 119   | 30.0m | 30.0m  | 1154           | shard_66258C06 | PRI              |     |
| 1.162.1003 | 15082 | *0 | 62414 | *0  | 557  | 1894 | 0 | 5.9%  | 75.7% | 0 | 117.6 | 45.86 | n/a | 0 | 1    | 116   | 21.32 | 18.8m  | 44.1m          | 1561           | shard_4FC5EC6E   | PRI |
| 1.162.1004 | 15721 | *0 | 62515 | *0  | 1149 | 1654 | 0 | 8.7%  | 75.0% | 0 | 60.36 | 46.46 | n/a | 0 | 1    | 110   | 11.5  | 19.9m  | 44.8m          | 1174           | shard_F0A9938E   | PRI |
| 1.162.1005 | 23626 | *0 | 58292 | *0  | 1119 | 2359 | 0 | 7.5%  | 75.8% | 0 | 59.66 | 46.36 | n/a | 0 | 1    | 110   | 23.6m | 50.9m  | 1433           | shard_A899D725 | PRI              |     |
| 1.164.1000 | 21763 | *0 | 64366 | *0  | 1112 | 2416 | 0 | 6.3%  | 75.5% | 0 | 59.46 | 45.56 | n/a | 0 | 1    | 112   | 21.44 | 23.5m  | 52.0m          | 1377           | shard_143EC928</ |     |

# 业务层面优化

优化前:

```
db.collection.insert( {"expireAt": new Date('July 22, 2019 01:11:00'), "xxx": "xxxx!"} ) //白天过期
db.collection.insert( { "expireAt": new Date('July 22, 2019 11:11:00'), "xxx": "xxxxx!"} ) //晚上过期
```

过期索引:

```
Db.collection.createIndex( { "expireAt": 1 }, { expireAfterSeconds: 0 } )
```

优化后:

```
db.collection.insert( {"expireAt": new Date('July 22, 2019 01:11:00'), "xxx": "xxxx!"} ) //之前白天过
期的数据, 现在业务优化为晚上过期
db.collection.insert( { "expireAt": new Date('July 22, 2019 02:11:00'), "xxx": "xxxxx!"} ) //晚上过期
```

# mongodb配置优化(网络IO复用，网络IO和磁盘IO做分离)

默认网络线程模型架构：

默认网络模型架构是每收到一个客户端链接，mongodb会创建一个线程处理该链接fd的所有读写请求及磁盘IO操作。

serviceExecutor: adaptive线程模型架构：

adaptive网络配置线程模型，不再是一个链接创建一个线程，新模式实现了网络IO复用，同时确保了网络IO和磁盘IO做分离。

**说明：**建议存在高并发的业务场景使用adaptive配置，低流量场景不建议开启



# 同一集群不同分片同一时刻不同配置性能对比

默认网络线程配置对应分片负载:

```
top - 11:02 up 133 days, 1:02, 3 users, load average: 38.46, 26.74, 23.24
Tasks: 1128 total, 1 running, 1127 sleeping, 0 stopped, 0 zombie
Cpu(s): 8.8%us, 1.7%sy, 0.0%ni, 83.7%id, 5.7%wa, 0.0%hi, 0.2%si, 0.0%st
Mem: 189669116k total, 8334924k free, 8334924k buffers
Swap: 0k total, 0k used, 0k free, 42893772k cached

  PID USER      PR  NI  VIRT  RES  SHR S %CPU  %MEM    TIME+  COMMAND
 37654 mongod  20   0 163M 135M 8180 S  50.0   0.0   45120.38 mongod
```

adaptive网络线程配置对应分片负载:

```
top - 11:02 up 133 days, 1:02, 1 user, load average: 3.05, 3.09, 3.56
Tasks: 631 total, 1 running, 628 sleeping, 0 stopped, 2 zombie
Cpu(s): 2.5%us, 0.8%sy, 0.0%ni, 96.7%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%st
Mem: 189669116k total, 14371260k free, 14371260k buffers
Swap: 0k total, 0 free, 0 used, 42893772k cached

  PID USER      PR  NI  VIRT  RES  SHR S %CPU  %MEM    TIME+  COMMAND
 37654 mongod  20   0 163M 135M 8180 S   0.0   0.0   45120.38 mongod
```

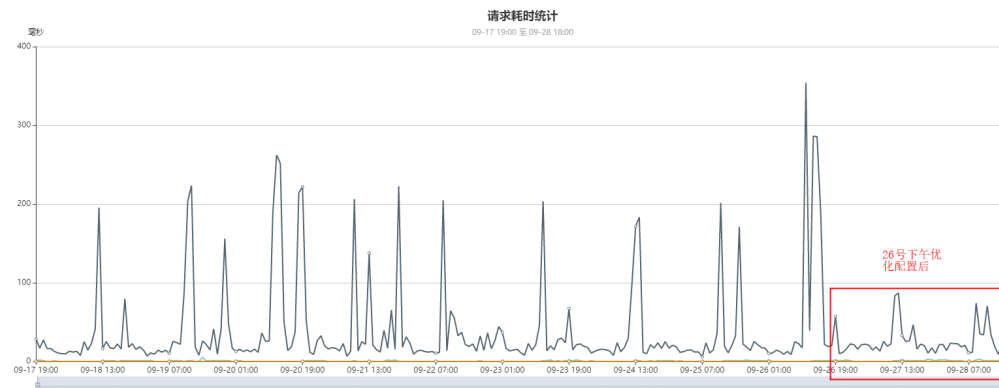
默认配置对应分片慢日志数(19621):

```
~]$ tail /var/log/mongodb/10000/mongodb.log -n 100000 | grep "2019-09-23T20:2" | grep keysExamined | wc -l
19621
```

adaptive配置对应慢日志数(5222):

```
~]$ tail /var/log/mongodb/10000/mongodb.log -n 100000 | grep "2019-09-23T20:2" | grep keysExamined | wc -l
5222
```

整体配置替换前后时延对比:



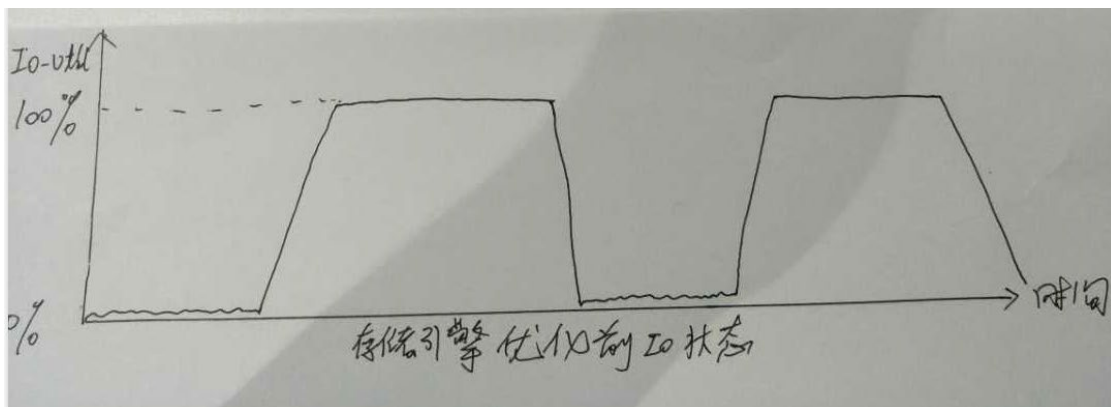
# wiredtiger存储引擎优化



问题背景：磁盘IO一会儿0%，一会儿持续性100%，tps有跌现象

|                                                             |        |        |      |      |       |       |          |          |       |         |         |       |       |
|-------------------------------------------------------------|--------|--------|------|------|-------|-------|----------|----------|-------|---------|---------|-------|-------|
| avg-cpu:  %user   %nice %system %iowait  %steal   %idle     |        |        |      |      |       |       |          |          |       |         |         |       |       |
| 0.03       0.00       0.03       0.00       0.00      99.94 |        |        |      |      |       |       |          |          |       |         |         |       |       |
| Device:                                                     | rrqm/s | wrqm/s | r/s  | w/s  | rkB/s | wkB/s | avgrq-sz | avgqu-sz | await | r_await | w_await | svctm | %util |
| nvme0n1                                                     | 0.00   | 1.48   | 0.00 | 2.63 | 0.00  | 22.88 | 17.40    | 0.00     | 0.02  | 0.03    | 0.02    | 0.01  | 0.00  |
| sda                                                         | 0.00   | 0.68   | 0.02 | 0.81 | 0.16  | 9.61  | 23.55    | 0.00     | 0.41  | 0.18    | 0.42    | 0.06  | 0.00  |
| avg-cpu:  %user   %nice %system %iowait  %steal   %idle     |        |        |      |      |       |       |          |          |       |         |         |       |       |
| 0.05       0.00       0.09       0.00       0.00      99.86 |        |        |      |      |       |       |          |          |       |         |         |       |       |
| Device:                                                     | rrqm/s | wrqm/s | r/s  | w/s  | rkB/s | wkB/s | avgrq-sz | avgqu-sz | await | r_await | w_await | svctm | %util |
| nvme0n1                                                     | 0.00   | 0.00   | 0.00 | 0.00 | 0.00  | 0.00  | 0.00     | 0.00     | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
| sda                                                         | 0.00   | 0.00   | 0.00 | 1.00 | 0.00  | 8.00  | 16.00    | 0.00     | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
| avg-cpu:  %user   %nice %system %iowait  %steal   %idle     |        |        |      |      |       |       |          |          |       |         |         |       |       |
| 0.02       0.00       0.05       0.00       0.00      99.94 |        |        |      |      |       |       |          |          |       |         |         |       |       |
| Device:                                                     | rrqm/s | wrqm/s | r/s  | w/s  | rkB/s | wkB/s | avgrq-sz | avgqu-sz | await | r_await | w_await | svctm | %util |
| nvme0n1                                                     | 0.00   | 0.00   | 0.00 | 0.00 | 0.00  | 0.00  | 0.00     | 0.00     | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
| sda                                                         | 0.00   | 0.00   | 0.00 | 0.00 | 0.00  | 0.00  | 0.00     | 0.00     | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
| avg-cpu:  %user   %nice %system %iowait  %steal   %idle     |        |        |      |      |       |       |          |          |       |         |         |       |       |
| 0.05       0.00       0.06       0.00       0.00      99.89 |        |        |      |      |       |       |          |          |       |         |         |       |       |
| Device:                                                     | rrqm/s | wrqm/s | r/s  | w/s  | rkB/s | wkB/s | avgrq-sz | avgqu-sz | await | r_await | w_await | svctm | %util |
| nvme0n1                                                     | 0.00   | 0.00   | 0.00 | 1.00 | 0.00  | 24.00 | 48.00    | 0.00     | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
| sda                                                         | 0.00   | 16.00  | 0.00 | 2.00 | 0.00  | 72.00 | 72.00    | 0.00     | 0.50  | 0.00    | 0.50    | 0.50  | 0.10  |

|          |        |        |         |           |          |            |          |           |             |         |             |        |        |
|----------|--------|--------|---------|-----------|----------|------------|----------|-----------|-------------|---------|-------------|--------|--------|
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgqu-sz | avgqu-sz  | await       | r_await | w_await     | svctm  | %util  |
| nvme0n1  | 0.00   | 0.00   | 1373.00 | 269512.00 | 25952.00 | 2156096.00 | 8.06     | 572306.87 | 217.63      | 3.90    | 218.72      | 0.00   | 100.40 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00     | 0.00      | 0.00        | 0.00    | 0.00        | 0.00   | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |          |           |             |         |             |        |        |
|          | 0.13   | 0.00   | 0.52    | 1.52      | 0.00     | 97.83      |          |           |             |         |             |        |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgqu-sz | avgqu-sz  | await       | r_await | w_await     | svctm  | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 8.00     | 580257.45 | 0.00        | 0.00    | 0.00        | 0.00   | 100.90 |
| sda      | 0.00   | 0.00   | 0.00    | 10.00     | 0.00     | 80.00      | 8.00     | 0.00      | 0.00        | 0.00    | 0.00        | 0.00   | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |          |           |             |         |             |        |        |
|          | 0.19   | 0.00   | 0.59    | 1.57      | 0.00     | 97.66      |          |           |             |         |             |        |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgqu-sz | avgqu-sz  | await       | r_await | w_await     | svctm  | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 8.00     | 511828.16 | 29973034.00 | 0.00    | 29973034.00 | 167.17 | 100.30 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00     | 0.00      | 0.00        | 0.00    | 0.00        | 0.00   | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |          |           |             |         |             |        |        |
|          | 0.13   | 0.00   | 0.59    | 1.59      | 0.00     | 97.69      |          |           |             |         |             |        |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgqu-sz | avgqu-sz  | await       | r_await | w_await     | svctm  | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00     | 448339.74 | 0.00        | 0.00    | 0.00        | 0.00   | 100.40 |
| sda      | 0.00   | 1.00   | 0.00    | 3.00      | 0.00     | 32.00      | 10.67    | 0.00      | 0.00        | 0.00    | 0.00        | 0.00   | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |          |           |             |         |             |        |        |
|          | 0.18   | 0.00   | 0.70    | 1.63      | 0.00     | 97.49      |          |           |             |         |             |        |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgqu-sz | avgqu-sz  | await       | r_await | w_await     | svctm  | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00     | 381138.11 | 0.00        | 0.00    | 0.00        | 0.00   | 100.50 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00     | 0.00      | 0.00        | 0.00    | 0.00        | 0.00   | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |          |           |             |         |             |        |        |
|          | 0.17   | 0.00   | 0.54    | 2.11      | 0.00     | 97.18      |          |           |             |         |             |        |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgqu-sz | avgqu-sz  | await       | r_await | w_await     | svctm  | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 14.00     | 0.00     | 112.00     | 8.00     | 316501.10 | 24398576.36 | 0.00    | 24398576.36 | 71.64  | 100.30 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00     | 0.00      | 0.00        | 0.00    | 0.00        | 0.00   | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |          |           |             |         |             |        |        |
|          | 0.09   | 0.00   | 0.38    | 2.87      | 0.00     | 96.66      |          |           |             |         |             |        |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgqu-sz | avgqu-sz  | await       | r_await | w_await     | svctm  | %util  |
| nvme0n1  | 0.00   | 0.00   | 3.00    | 20.00     | 24.00    | 160.00     | 8.00     | 253360.31 | 8316580.96  | 4.00    | 21064067.50 | 43.48  | 100.00 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00     | 0.00      | 0.00        | 0.00    | 0.00        | 0.00   | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |          |           |             |         |             |        |        |
|          | 0.25   | 0.00   | 1.43    | 2.02      | 0.00     | 96.30      |          |           |             |         |             |        |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgqu-sz | avgqu-sz  | await       | r_await | w_await     | svctm  | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 127259.00 | 0.00     | 1018072.00 | 0.00     | 3.00      | 266790      | 46      | 2506.98     | 0.01   | 100.10 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00     | 0.00      | 0.00        | 0.00    | 0.00        | 0.00   | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |          |           |             |         |             |        |        |
|          | 0.65   | 0.00   | 0.60    | 2.04      | 0.00     | 96.71      |          |           |             |         |             |        |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgqu-sz | avgqu-sz  | await       | r_await | w_await     | svctm  | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00     | 241667.13 | 0.00        | 0.00    | 0.00        | 0.00   | 100.60 |
| sda      | 0.00   | 2.00   | 0.00    | 2.00      | 0.00     | 32.00      | 16.00    | 0.00      | 0.00        | 0.00    | 0.00        | 0.00   | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |          |           |             |         |             |        |        |
|          | 0.88   | 0.00   | 0.70    | 1.96      | 0.00     | 96.47      |          |           |             |         |             |        |        |



# 存储引擎调优: cachesize调整优化(120G->50G)

cache调小目的:

1. 减少checkpoint刷脏数据时的数据量, 减少磁盘IO跟不上客户端写入速度引起的持续性IO跌0问题。
2. 预留部分内存给内核pageCache, 尽量避免内存不足引起的应用阻塞。



# 存储引擎调优：dirty脏数据淘汰优化

默认存储引擎配置:

| wiredtiger淘汰相关配置       | 默认值 | 工作原理                                                |
|------------------------|-----|-----------------------------------------------------|
| eviction_target        | 80  | 当用掉的内存超过总内存的百分比达到 eviction_target, 后台evict线程开始淘汰    |
| eviction_trigger       | 95  | 当用掉的内存超过总内存的 eviction_trigger, 用户线程也开始淘汰            |
| eviction_dirty_target  | 5   | 当cache中脏数据比例超过 eviction_dirty_target, 后台evict线程开始淘汰 |
| eviction_dirty_trigger | 20  | 当cache中脏数据比例超过 eviction_dirty_trigger, 用户线程也开始淘汰    |
| evict.threads_min      | 4   | 后台evict线程最小数                                        |
| evict.threads_max      | 4   | 后台evict线程最大数                                        |

引擎优化后的配置:

| wiredtiger淘汰相关配置       | 默认值 | 优化目的                                                               |
|------------------------|-----|--------------------------------------------------------------------|
| eviction_target        | 75  | 总体思想是让后台evict尽量早点淘汰脏页page到磁盘，同时调整evict淘汰线程数来加快两次checkpoint之间的脏数据淘汰 |
| eviction_trigger       | 97  |                                                                    |
| eviction_dirty_target  | 3   |                                                                    |
| eviction_dirty_trigger | 25  |                                                                    |
| evict.threads_min      | 8   |                                                                    |
| evict.threads_max      | 12  |                                                                    |



# 存储引擎调优：checkpoint优化调整

默认配置: checkpoint=(wait=60,log\_size=2GB)

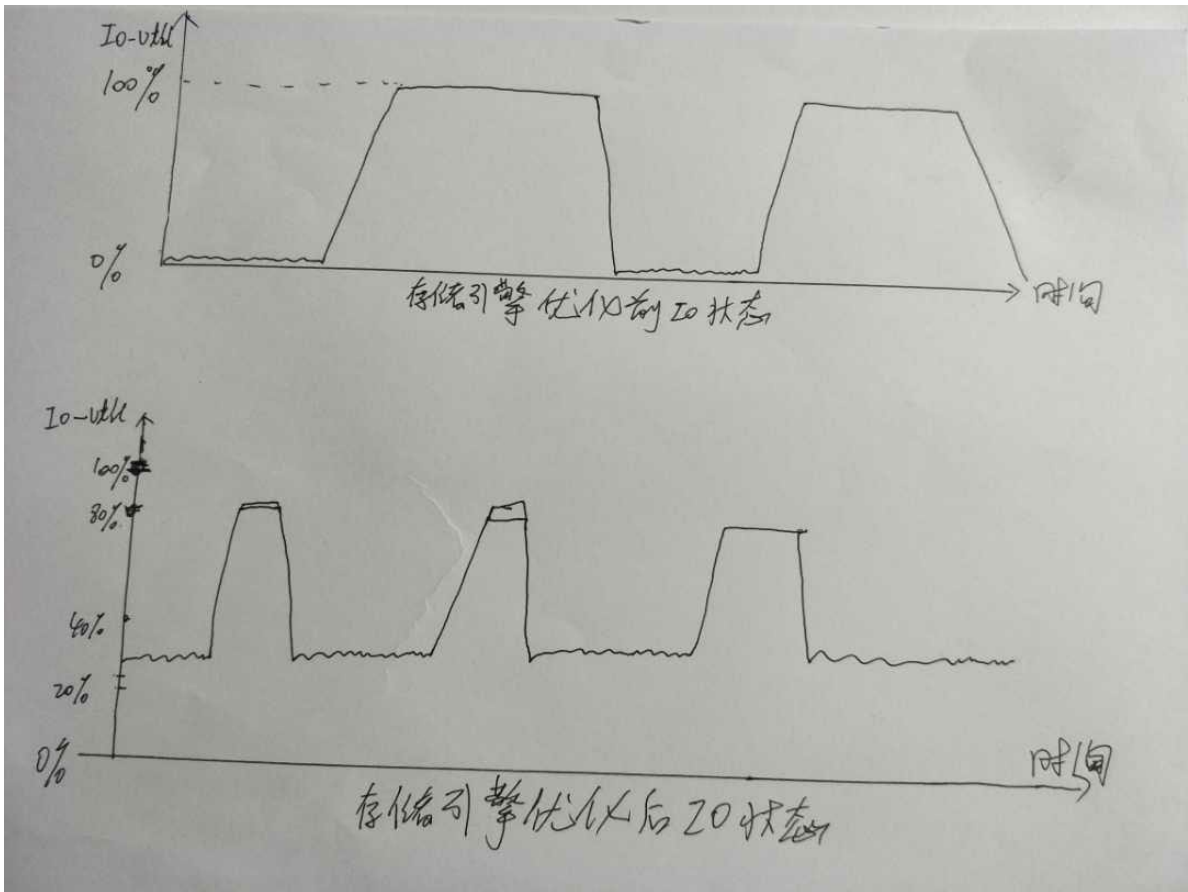
优化后配置: checkpoint=(wait=25,log\_size=1GB)

总体思想:

如果在两次checkpoint的时间间隔内evict淘汰线程淘汰的dirty page越少, 那么积压的脏数据就会越多, 也就是checkpoint的时候脏数据就会越多, 造成checkpoint的时候IO集中大量写磁盘。如果我们把checkpoint的周期缩短, 那么两个checkpoint期间的脏数据相应的也就会减少, 磁盘IO 100%持续的时间也就会缩短。

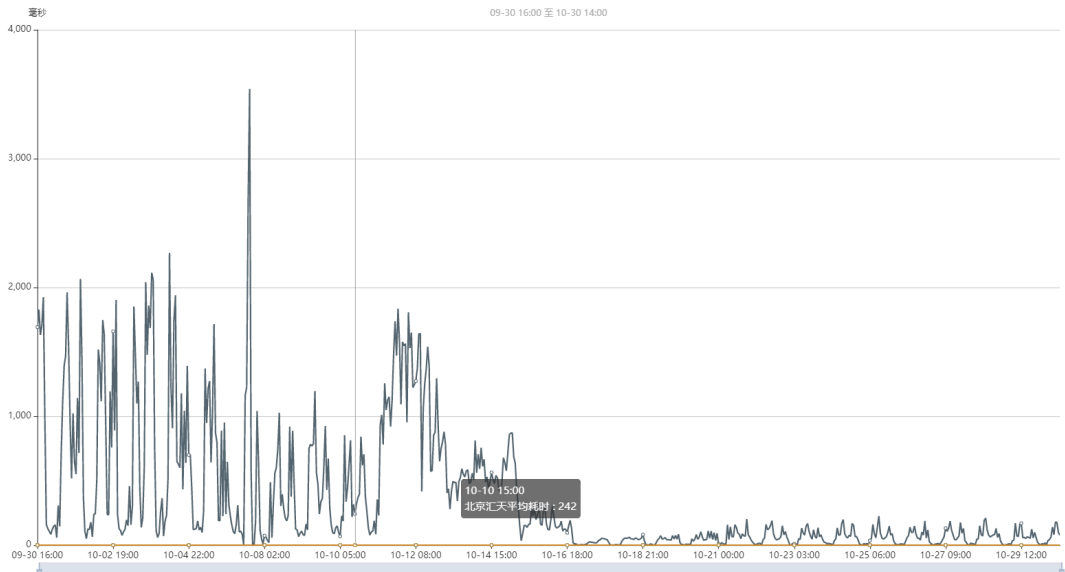
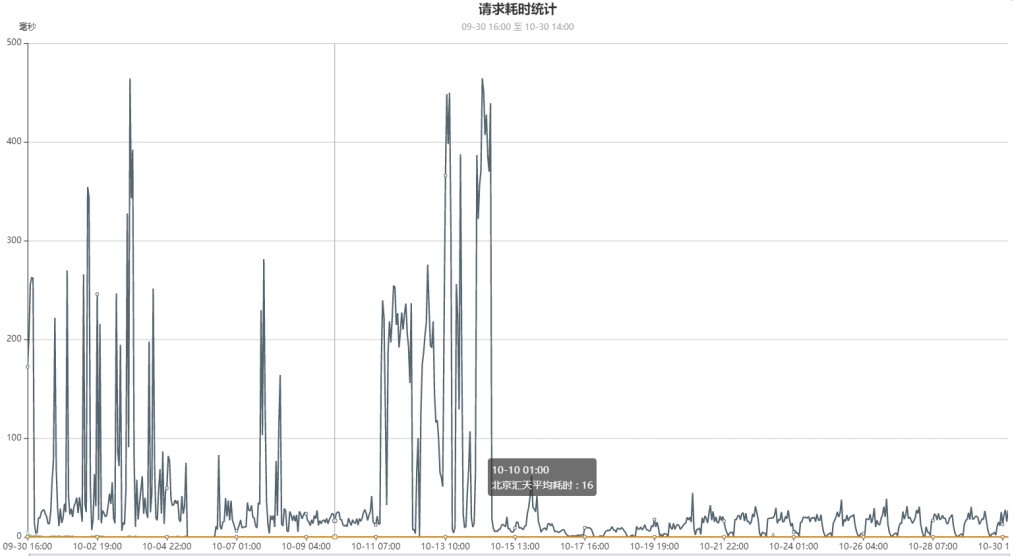
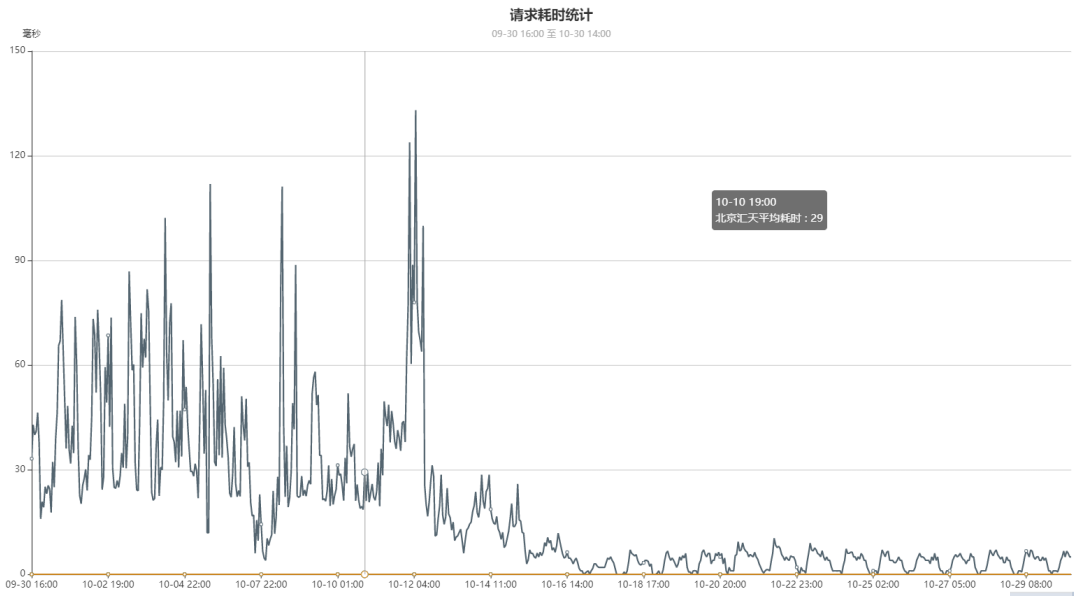
# 存储引擎优化前后IO趋势对比

| Device:                | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s  | wsec/s     | avgrq-sz | avgqu-sz | await | r_await | w_await | svctm | %util |
|------------------------|--------|--------|---------|-----------|---------|------------|----------|----------|-------|---------|---------|-------|-------|
| sda                    | 0.00   | 0.00   | 0.00    | 0.00      | 0.00    | 0.00       | 0.00     | 0.00     | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
| dfa                    | 0.00   | 0.00   | 184.00  | 3246.00   | 5256.00 | 25968.00   | 9.10     | 0.61     | 0.18  | 0.12    | 0.18    | 0.06  | 21.20 |
| 12/04/2019 11:44:55 AM |        |        |         |           |         |            |          |          |       |         |         |       |       |
| avg-cpu:               | %user  | %nice  | %system | %iowait   | %steal  | %idle      |          |          |       |         |         |       |       |
|                        | 11.36  | 0.00   | 0.81    | 0.02      | 0.00    | 87.81      |          |          |       |         |         |       |       |
| Device:                | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s  | wsec/s     | avgrq-sz | avgqu-sz | await | r_await | w_await | svctm | %util |
| sda                    | 0.00   | 0.00   | 0.00    | 0.00      | 0.00    | 0.00       | 0.00     | 0.00     | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
| dfa                    | 0.00   | 0.00   | 191.00  | 3509.00   | 5280.00 | 28176.00   | 9.04     | 0.28     | 0.08  | 0.23    | 0.07    | 0.04  | 14.40 |
| 12/04/2019 11:44:56 AM |        |        |         |           |         |            |          |          |       |         |         |       |       |
| avg-cpu:               | %user  | %nice  | %system | %iowait   | %steal  | %idle      |          |          |       |         |         |       |       |
|                        | 12.81  | 0.00   | 1.16    | 0.02      | 0.00    | 86.02      |          |          |       |         |         |       |       |
| Device:                | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s  | wsec/s     | avgrq-sz | avgqu-sz | await | r_await | w_await | svctm | %util |
| sda                    | 0.00   | 0.00   | 0.00    | 0.00      | 0.00    | 0.00       | 0.00     | 0.00     | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
| dfa                    | 0.00   | 0.00   | 191.00  | 3970.00   | 5680.00 | 31760.00   | 9.00     | 0.16     | 0.04  | 0.16    | 0.03    | 0.02  | 9.30  |
| 12/04/2019 11:44:57 AM |        |        |         |           |         |            |          |          |       |         |         |       |       |
| avg-cpu:               | %user  | %nice  | %system | %iowait   | %steal  | %idle      |          |          |       |         |         |       |       |
|                        | 12.73  | 0.00   | 1.16    | 0.03      | 0.00    | 86.08      |          |          |       |         |         |       |       |
| Device:                | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s  | wsec/s     | avgrq-sz | avgqu-sz | await | r_await | w_await | svctm | %util |
| sda                    | 0.00   | 0.00   | 0.00    | 0.00      | 0.00    | 0.00       | 0.00     | 0.00     | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
| dfa                    | 0.00   | 0.00   | 192.00  | 3755.00   | 5912.00 | 30040.00   | 9.11     | 0.44     | 0.11  | 0.17    | 0.11    | 0.04  | 17.20 |
| 12/04/2019 11:44:58 AM |        |        |         |           |         |            |          |          |       |         |         |       |       |
| avg-cpu:               | %user  | %nice  | %system | %iowait   | %steal  | %idle      |          |          |       |         |         |       |       |
|                        | 11.98  | 0.00   | 1.11    | 0.02      | 0.00    | 86.89      |          |          |       |         |         |       |       |
| Device:                | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s  | wsec/s     | avgrq-sz | avgqu-sz | await | r_await | w_await | svctm | %util |
| sda                    | 0.00   | 0.00   | 0.00    | 3.00      | 0.00    | 24.00      | 8.00     | 0.00     | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
| dfa                    | 0.00   | 0.00   | 186.00  | 3713.00   | 4704.00 | 29704.00   | 8.82     | 0.48     | 0.12  | 0.15    | 0.12    | 0.05  | 19.00 |
| 12/04/2019 11:44:59 AM |        |        |         |           |         |            |          |          |       |         |         |       |       |
| avg-cpu:               | %user  | %nice  | %system | %iowait   | %steal  | %idle      |          |          |       |         |         |       |       |
|                        | 52.87  | 0.00   | 8.97    | 0.00      | 0.00    | 38.16      |          |          |       |         |         |       |       |
| Device:                | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s  | wsec/s     | avgrq-sz | avgqu-sz | await | r_await | w_await | svctm | %util |
| sda                    | 0.00   | 0.00   | 0.00    | 0.00      | 0.00    | 0.00       | 0.00     | 0.00     | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
| dfa                    | 0.00   | 0.00   | 90.00   | 7455.00   | 2032.00 | 596352.00  | 8.02     | 3.64     | 0.05  | 0.13    | 0.05    | 0.01  | 48.60 |
| 12/04/2019 11:45:00 AM |        |        |         |           |         |            |          |          |       |         |         |       |       |
| avg-cpu:               | %user  | %nice  | %system | %iowait   | %steal  | %idle      |          |          |       |         |         |       |       |
|                        | 80.05  | 0.00   | 14.14   | 0.03      | 0.00    | 5.77       |          |          |       |         |         |       |       |
| Device:                | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s  | wsec/s     | avgrq-sz | avgqu-sz | await | r_await | w_await | svctm | %util |
| sda                    | 0.00   | 0.00   | 0.00    | 0.00      | 0.00    | 0.00       | 0.00     | 0.00     | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
| dfa                    | 0.00   | 0.00   | 9.00    | 162322.00 | 384.00  | 1298624.00 | 8.00     | 0.00     | 9.04  | 0.05    | 0.67    | 0.05  | 90.70 |
| 12/04/2019 11:45:01 AM |        |        |         |           |         |            |          |          |       |         |         |       |       |
| avg-cpu:               | %user  | %nice  | %system | %iowait   | %steal  | %idle      |          |          |       |         |         |       |       |
|                        | 69.88  | 0.00   | 11.66   | 0.03      | 0.00    | 18.43      |          |          |       |         |         |       |       |





# 存储引擎优化前后时延对比



# 服务器IO系统问题

通过和厂商大量测试、固件升级、驱动调试等，最终定位到是操作系统版本问题引起，通过升级该nvme服务器到3.10.0-957.27.2.el7.x86\_64后，问题解决。升级前后IO状况对比如下图所示：

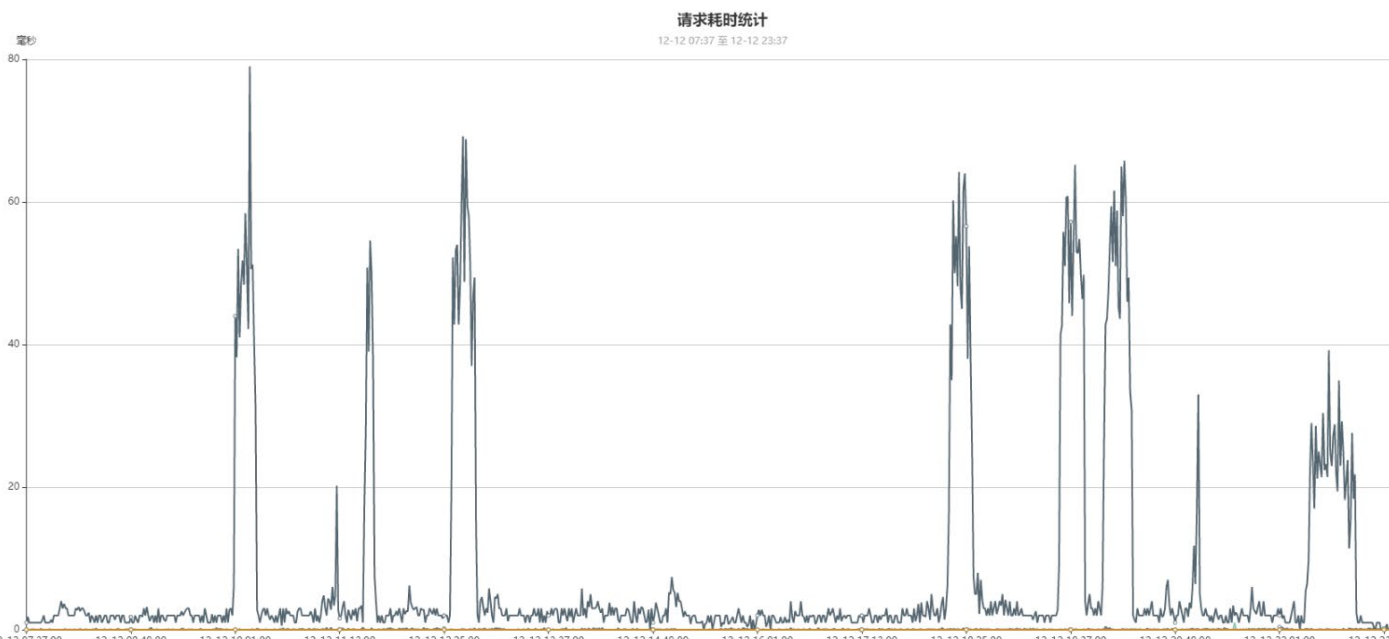
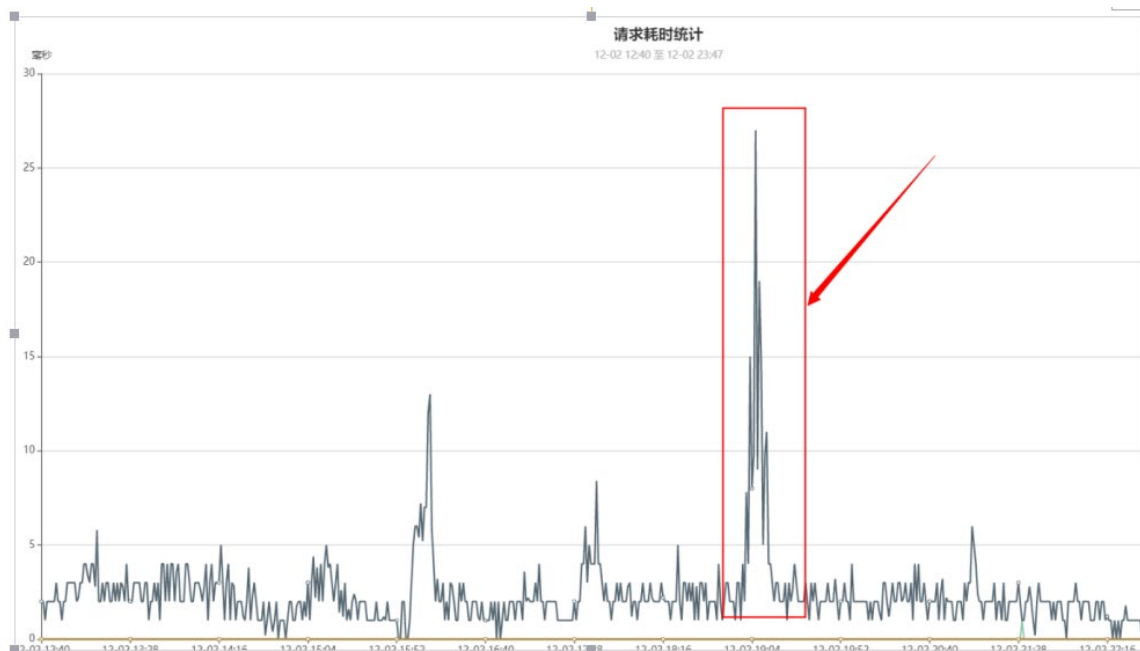
|          |        |        |         |           |          |            |         |            |             |         |             |       |        |
|----------|--------|--------|---------|-----------|----------|------------|---------|------------|-------------|---------|-------------|-------|--------|
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgq-sz | avgq-sz    | await       | r_await | w_await     | svctm | %util  |
| nvme0n1  | 0.00   | 0.00   | 1373.00 | 269512.00 | 25952.00 | 2156096.00 | 8.06    | 572306.87  | 217.63      | 3.90    | 218.72      | 0.00  | 100.40 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 0.00       | 0.00        | 0.00    | 0.00        | 0.00  | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |         |            |             |         |             |       |        |
|          | 0.13   | 0.00   | 0.52    | 1.52      | 0.00     | 97.83      |         |            |             |         |             |       |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgq-sz | avgq-sz    | await       | r_await | w_await     | svctm | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 580257.45  | 0.00        | 0.00    | 0.00        | 0.00  | 100.90 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 0.00       | 0.00        | 0.00    | 0.00        | 0.00  | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |         |            |             |         |             |       |        |
|          | 0.19   | 0.00   | 0.59    | 1.57      | 0.00     | 97.66      |         |            |             |         |             |       |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgq-sz | avgq-sz    | await       | r_await | w_await     | svctm | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 511828.16  | 29973034.00 | 0.00    | 0.00        | 0.00  | 100.30 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 0.00       | 0.00        | 0.00    | 0.00        | 0.00  | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |         |            |             |         |             |       |        |
|          | 0.13   | 0.00   | 0.59    | 1.59      | 0.00     | 97.69      |         |            |             |         |             |       |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgq-sz | avgq-sz    | await       | r_await | w_await     | svctm | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 448339.74  | 0.00        | 0.00    | 0.00        | 0.00  | 100.40 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 32.00      | 10.67       | 0.00    | 0.00        | 0.00  | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |         |            |             |         |             |       |        |
|          | 0.18   | 0.00   | 0.70    | 1.63      | 0.00     | 97.49      |         |            |             |         |             |       |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgq-sz | avgq-sz    | await       | r_await | w_await     | svctm | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 381138.11  | 0.00        | 0.00    | 0.00        | 0.00  | 100.50 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 0.00       | 0.00        | 0.00    | 0.00        | 0.00  | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |         |            |             |         |             |       |        |
|          | 0.17   | 0.00   | 0.54    | 2.11      | 0.00     | 97.18      |         |            |             |         |             |       |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgq-sz | avgq-sz    | await       | r_await | w_await     | svctm | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 14.00     | 0.00     | 0.00       | 0.00    | 316501.10  | 4398576.36  | 0.00    | 24398576.36 | 71.64 | 100.30 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 0.00       | 0.00        | 0.00    | 0.00        | 0.00  | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |         |            |             |         |             |       |        |
|          | 0.09   | 0.00   | 0.38    | 2.87      | 0.00     | 96.66      |         |            |             |         |             |       |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgq-sz | avgq-sz    | await       | r_await | w_await     | svctm | %util  |
| nvme0n1  | 0.00   | 0.00   | 3.00    | 20.00     | 24.00    | 0.00       | 0.00    | 253360.31  | 8316580.96  | 4.00    | 21064067.50 | 43.48 | 100.00 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 0.00       | 0.00        | 0.00    | 0.00        | 0.00  | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |         |            |             |         |             |       |        |
|          | 0.25   | 0.00   | 1.43    | 2.02      | 0.00     | 96.30      |         |            |             |         |             |       |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgq-sz | avgq-sz    | await       | r_await | w_await     | svctm | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 127259.00 | 0.00     | 0.00       | 0.00    | 1018072.00 | 8.00        | 0.00    | 2506.98     | 0.01  | 100.10 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 0.00       | 0.00        | 0.00    | 0.00        | 0.00  | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |         |            |             |         |             |       |        |
|          | 0.65   | 0.00   | 0.60    | 2.04      | 0.00     | 96.71      |         |            |             |         |             |       |        |
| Device:  | rrqm/s | wrqm/s | r/s     | w/s       | rsec/s   | wsec/s     | avgq-sz | avgq-sz    | await       | r_await | w_await     | svctm | %util  |
| nvme0n1  | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 0.00       | 0.00        | 0.00    | 0.00        | 0.00  | 100.60 |
| sda      | 0.00   | 0.00   | 0.00    | 0.00      | 0.00     | 0.00       | 0.00    | 32.00      | 16.00       | 0.00    | 0.00        | 0.00  | 0.00   |
| avg-cpu: | %user  | %nice  | %system | %iowait   | %steal   | %idle      |         |            |             |         |             |       |        |
|          | 0.88   | 0.00   | 0.70    | 1.96      | 0.00     | 96.47      |         |            |             |         |             |       |        |

|             |        |        |         |           |        |            |         |         |       |         |         |       |       |
|-------------|--------|--------|---------|-----------|--------|------------|---------|---------|-------|---------|---------|-------|-------|
| 12:59:16 PM | %user  | %nice  | %system | %iowait   | %steal | %idle      |         |         |       |         |         |       |       |
|             | 13.19  | 0.00   | 4.66    | 0.02      | 0.00   | 82.13      |         |         |       |         |         |       |       |
|             | rrqm/s | wrqm/s | r/s     | w/s       | rkB/s  | wkB/s      | avgq-sz | avgq-sz | await | r_await | w_await | svctm | %util |
|             | 0.00   | 0.00   | 0.00    | 10794.00  | 0.00   | 0.00       | 0.00    | 0.00    | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
|             | 0.00   | 0.00   | 1.00    | 12.00     | 0.00   | 96292.00   | 17.84   | 2.19    | 0.20  | 1.00    | 0.20    | 0.00  | 5.10  |
| 12:59:17 PM | %user  | %nice  | %system | %iowait   | %steal | %idle      |         |         |       |         |         |       |       |
|             | 2.43   | 0.00   | 3.74    | 0.02      | 0.00   | 93.82      |         |         |       |         |         |       |       |
|             | rrqm/s | wrqm/s | r/s     | w/s       | rkB/s  | wkB/s      | avgq-sz | avgq-sz | await | r_await | w_await | svctm | %util |
|             | 0.00   | 3.00   | 0.00    | 4.00      | 0.00   | 28.00      | 14.00   | 0.00    | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
|             | 0.00   | 0.00   | 0.00    | 57362.00  | 0.00   | 1792740.00 | 62.51   | 24.09   | 0.42  | 0.00    | 0.42    | 0.00  | 75.60 |
| 12:59:18 PM | %user  | %nice  | %system | %iowait   | %steal | %idle      |         |         |       |         |         |       |       |
|             | 5.54   | 0.00   | 4.44    | 0.02      | 0.00   | 90.00      |         |         |       |         |         |       |       |
|             | rrqm/s | wrqm/s | r/s     | w/s       | rkB/s  | wkB/s      | avgq-sz | avgq-sz | await | r_await | w_await | svctm | %util |
|             | 0.00   | 0.00   | 0.00    | 2.00      | 0.00   | 8.00       | 8.00    | 0.00    | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
|             | 0.00   | 0.00   | 5.00    | 104429.00 | 0.00   | 2141928.00 | 41.02   | 31.34   | 0.30  | 0.80    | 0.30    | 0.00  | 94.40 |
| 12:59:19 PM | %user  | %nice  | %system | %iowait   | %steal | %idle      |         |         |       |         |         |       |       |
|             | 21.54  | 0.00   | 8.35    | 0.05      | 0.00   | 70.07      |         |         |       |         |         |       |       |
|             | rrqm/s | wrqm/s | r/s     | w/s       | rkB/s  | wkB/s      | avgq-sz | avgq-sz | await | r_await | w_await | svctm | %util |
|             | 0.00   | 0.00   | 0.00    | 0.00      | 0.00   | 0.00       | 0.00    | 0.00    | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
|             | 0.00   | 0.00   | 41.00   | 59080.00  | 0.00   | 858904.00  | 29.09   | 32.12   | 0.54  | 0.83    | 0.54    | 0.00  | 43.80 |
| 12:59:20 PM | %user  | %nice  | %system | %iowait   | %steal | %idle      |         |         |       |         |         |       |       |
|             | 3.06   | 0.00   | 1.70    | 0.00      | 0.00   | 95.24      |         |         |       |         |         |       |       |
|             | rrqm/s | wrqm/s | r/s     | w/s       | rkB/s  | wkB/s      | avgq-sz | avgq-sz | await | r_await | w_await | svctm | %util |
|             | 0.00   | 0.00   | 0.00    | 0.00      | 0.00   | 0.00       | 0.00    | 0.00    | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
|             | 0.00   | 0.00   | 4.00    | 671.00    | 0.00   | 4160.00    | 12.48   | 0.00    | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
| 12:59:21 PM | %user  | %nice  | %system | %iowait   | %steal | %idle      |         |         |       |         |         |       |       |
|             | 5.48   | 0.00   | 3.38    | 0.00      | 0.00   | 91.14      |         |         |       |         |         |       |       |
|             | rrqm/s | wrqm/s | r/s     | w/s       | rkB/s  | wkB/s      | avgq-sz | avgq-sz | await | r_await | w_await | svctm | %util |
|             | 0.00   | 0.00   | 0.00    | 0.00      | 0.00   | 0.00       | 0.00    | 0.00    | 0.00  | 0.00    | 0.00    | 0.00  | 0.00  |
|             | 0.00   | 0.00   | 0.00    | 889.00    | 0.00   | 9336.00    | 21.00   | 0.03    | 0.03  | 0.00    | 0.03    | 0.00  | 0.30  |
| 12:59:22 PM |        |        |         |           |        |            |         |         |       |         |         |       |       |



# 服务器IO问题解决后时延

升级服务器系统版本后，服务器IO能力有了很大的提升，于是替换集群所有分片的主节点为升级后的高IO服务器(由于升级后的服务器没有大规模使用过，为了谨慎保险起见，从节点还是未升级的低IO服务器)，时延进一步降低到平均2-4ms，但是有80ms左右的毛刺，业务侧核心的两个接口时延如下图所示：



oppo

mongodb-3.6默认启用了 **MajorityReadConcern**。启用该功能后，mongodb必须在内存中借助snapshot 及主从通信来维护更多的版本信息，以达到主从一致性的目的。由于从节点是低IO服务器，很容易造成阻塞，这样拉取oplog的速度就会跟不上进度，造成主节点消耗大量的内存来维护快照信息，这样就会导致大量的内存及cpu消耗，最终导致脏数据瞬间剧增，很快达到eviction\_dirty\_trigger阈值，业务也因此抖动。由于业务不需要readConcern("Majority")功能，因此禁掉该配置。

|      |   |     |   |   |   |   |   |     |       |       |   |       |       |     |   |      |      |       |       |      |                |     |
|------|---|-----|---|---|---|---|---|-----|-------|-------|---|-------|-------|-----|---|------|------|-------|-------|------|----------------|-----|
| 37.1 | 1 | 000 | 0 | 0 | 5 | 0 | 0 | 5   | 24.2% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 10   | 1    | 3.45K | 72.3K | 2047 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 6   | 23.8% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 10   | 1    | 1.03K | 70.2K | 2047 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 6   | 23.5% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 10   | 1    | 1.03K | 70.2K | 2047 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 5   | 23.2% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 10   | 1    | 1.03K | 70.0K | 2047 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 5   | 22.9% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 10   | 1    | 1.03K | 70.0K | 2047 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 143 | 22.6% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 10   | 1    | 9.46K | 70.2K | 1823 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 13  | 22.3% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 10   | 1    | 1.66K | 70.1K | 1823 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 23  | 22.2% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 10   | 1    | 2.85K | 90.8K | 1823 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 11  | 21.7% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 10   | 1    | 1.57K | 69.1K | 1823 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 6   | 21.5% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 10   | 1    | 1.09K | 74.3K | 1823 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 5   | 21.3% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 10   | 1    | 1.03K | 69.8K | 1823 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 5   | 20.9% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 21   | 2    | 2.5K  | 70.2K | 1823 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 8   | 48.5% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 21   | 1    | 1.59K | 71.9K | 1823 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 5   | 58.1% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 10   | 2123 | 26.4K | 69.9K | 1823 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 60  | 57.5% | 95.0% | 0 | 117.6 | 55.9% | n/a | 0 | 111  | 2128 | 80.3K | 39.2K | 1844 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 561 | 56.6% | 95.0% | 0 | 118.6 | 56.0% | n/a | 0 | 607K | 2    | 1.66K | 70.2K | 1844 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 10  | 55.7% | 94.0% | 0 | 118.6 | 56.0% | n/a | 0 | 691  | 1128 | 1.31K | 80.0K | 1944 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 11  | 55.0% | 93.9% | 0 | 118.6 | 56.0% | n/a | 0 | 691  | 1128 | 2.56K | 78.7K | 1944 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 8   | 54.7% | 93.8% | 0 | 118.6 | 56.0% | n/a | 0 | 691  | 1128 | 2.56K | 71.1K | 1944 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 8   | 54.2% | 93.3% | 0 | 118.6 | 56.0% | n/a | 0 | 691  | 1128 | 2.56K | 71.1K | 1944 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 10  | 54.2% | 93.3% | 0 | 118.6 | 56.0% | n/a | 0 | 691  | 1128 | 2.67K | 73.0K | 1944 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 | 0 | 0 | 26  | 54.2% | 93.3% | 0 | 118.6 | 56.0% | n/a | 0 | 691  | 2128 | 7.85K | 88.5K | 1944 | shard_4FC5CE6E | PR1 |
| 37.1 | 1 | 000 | 0 | 0 | 0 |   |   |     |       |       |   |       |       |     |   |      |      |       |       |      |                |     |

```

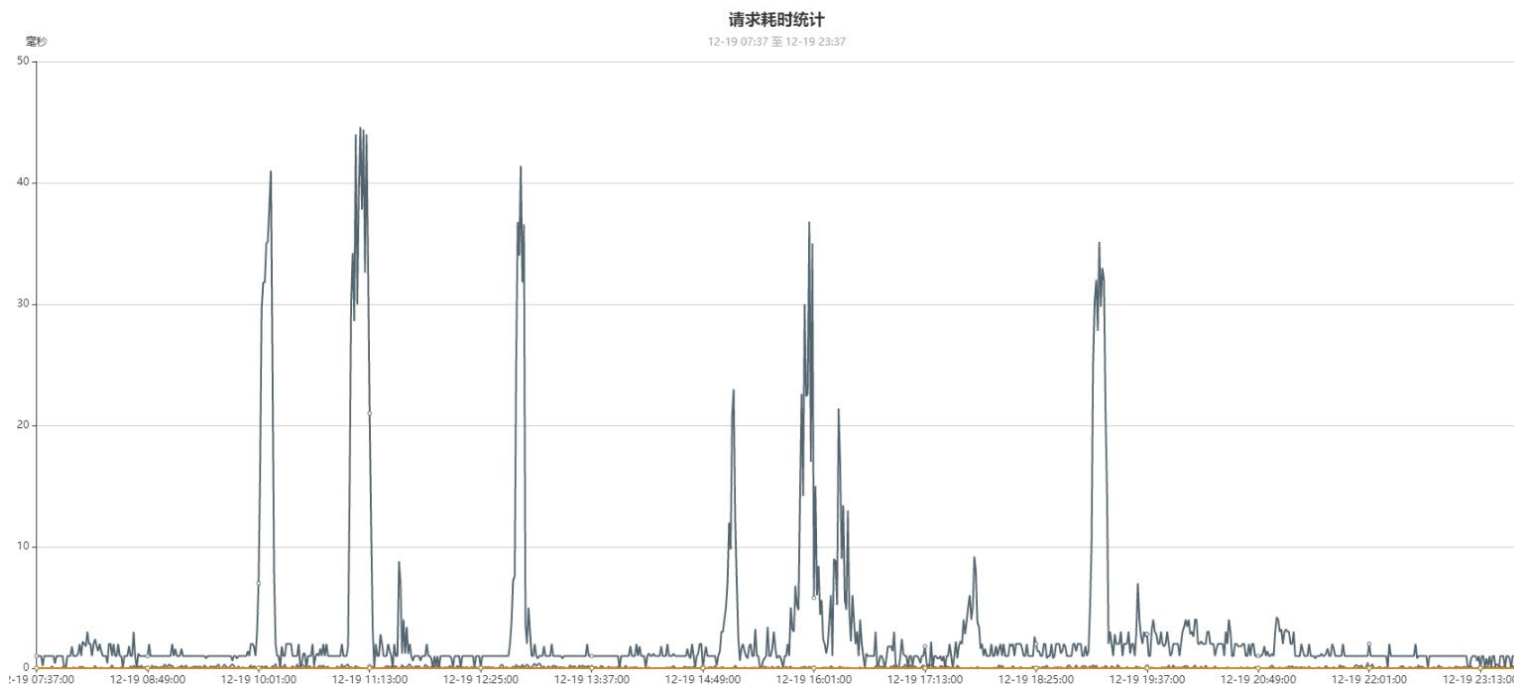
syncedTo: Tue Dec 17 2019 00:19:40 GMT+0800 (CST)
1100 secs (0.31 hrs) behind the primary
source: 10.10.10.100
syncedTo: Tue Dec 17 2019 00:25:44 GMT+0800 (CST)
736 secs (0.2 hrs) behind the primary
shard_F0A9938E:PRIMARY> db.printSlaveReplicationInfo()
source: 10.10.10.100
syncedTo: Tue Dec 17 2019 00:19:41 GMT+0800 (CST)
1100 secs (0.31 hrs) behind the primary
source: 10.10.10.100
syncedTo: Tue Dec 17 2019 00:25:44 GMT+0800 (CST)
737 secs (0.2 hrs) behind the primary
shard_F0A9938E:PRIMARY> db.printSlaveReplicationInfo()
source: 10.10.10.100
syncedTo: Tue Dec 17 2019 00:19:45 GMT+0800 (CST)
1098 secs (0.31 hrs) behind the primary
source: 10.10.10.100
syncedTo: Tue Dec 17 2019 00:25:46 GMT+0800 (CST)
737 secs (0.2 hrs) behind the primary
shard_F0A9938E:PRIMARY> db.printSlaveReplicationInfo()
source: 10.10.10.100
syncedTo: Tue Dec 17 2019 00:19:45 GMT+0800 (CST)
1098 secs (0.31 hrs) behind the primary
source: 10.10.10.100
syncedTo: Tue Dec 17 2019 00:25:47 GMT+0800 (CST)
736 secs (0.2 hrs) behind the primary
shard_F0A9938E:PRIMARY> db.printSlaveReplicationInfo()
source: 10.10.10.100
syncedTo: Tue Dec 17 2019 00:19:45 GMT+0800 (CST)
1099 secs (0.31 hrs) behind the primary
source: 10.10.10.100
syncedTo: Tue Dec 17 2019 00:25:48 GMT+0800 (CST)
736 secs (0.2 hrs) behind the primary
shard_F0A9938E:PRIMARY> db.printSlaveReplicationInfo()
source: 10.10.10.100
syncedTo: Tue Dec 17 2019 00:19:45 GMT+0800 (CST)
1101 secs (0.31 hrs) behind the primary
source: 10.10.10.100
syncedTo: Tue Dec 17 2019 00:25:59 GMT+0800 (CST)
727 secs (0.2 hrs) behind the primary
shard_F0A9938E:PRIMARY>

```



# 升级所有分片从节点到高IO服务器

禁用enableMajorityReadConcern功能及升级所有从节点到高IO服务器后，时延从80ms峰值毛刺减少到了40ms，如下图所示：



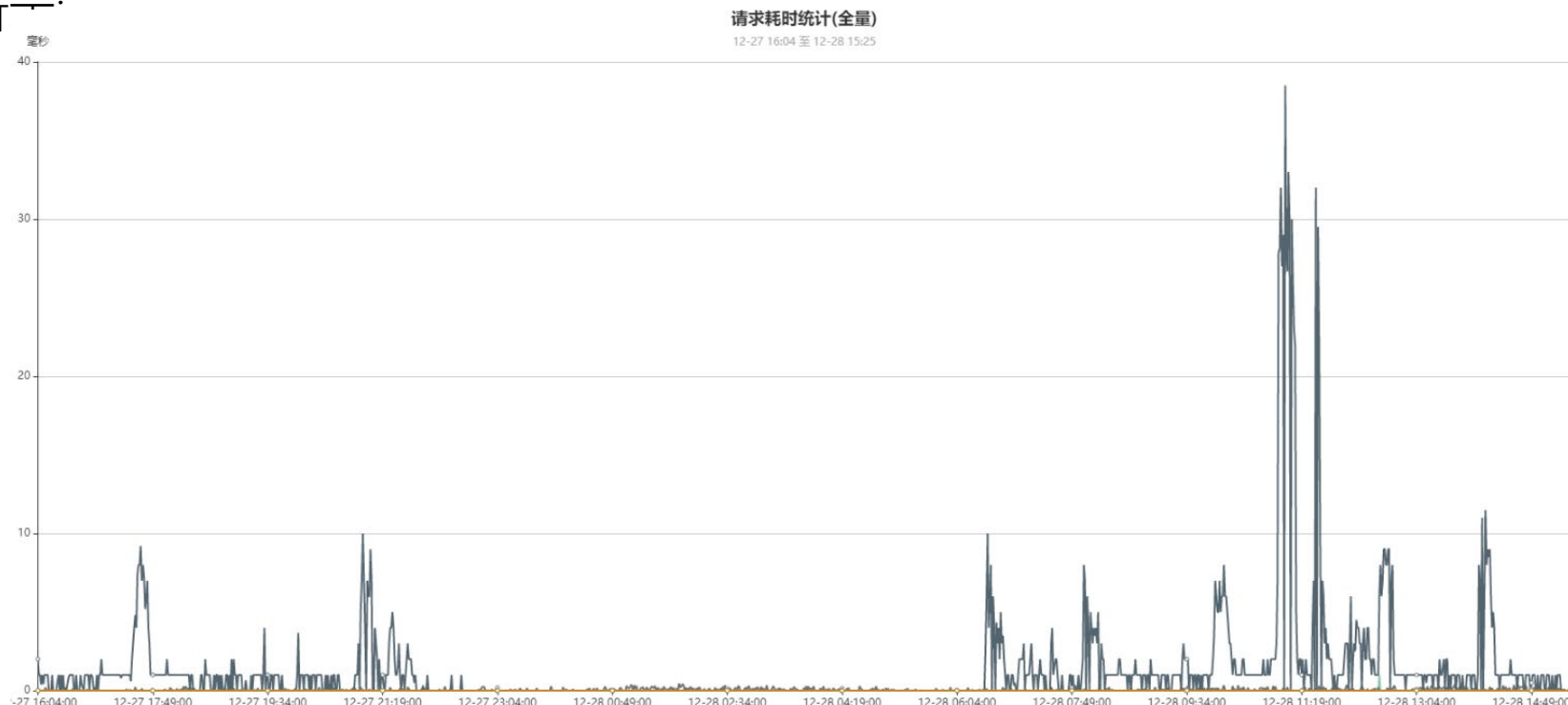
# 存储引擎继续优化



为了进一步减缓时延尖刺，我们继续在之前基础上对存储引擎调优，调整后配置如下：

eviction\_dirty\_trigger: 30% evict.threads\_min: 12 evict.threads\_max:18 checkpoint=(wait=20,log\_size=1GB)

经过此轮的存储引擎调优后，该业务的核心接口时延进一步好转，时延尖刺相比之前有了进一步的改善，时延最大尖刺时间从前面的40ms降低到了30ms，同时尖刺出现的频率明显降低了，如下图所示：



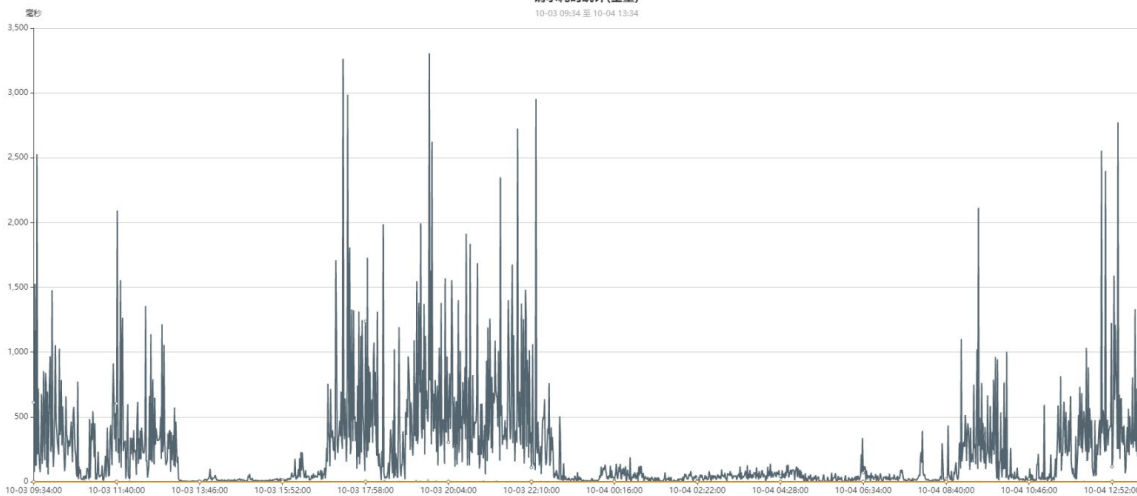


# 总结

通过软件层面(mongodb服务配置、业务优化、存储引擎优化)及硬件优化(升级操作系统)后, 该大流量集群的核心接口时延从最初的平均成百上千ms降低到了现在的平均1-2ms, 性能提升比较可观, 整体时延性能提升数十倍。优化前主要接口时延如下:

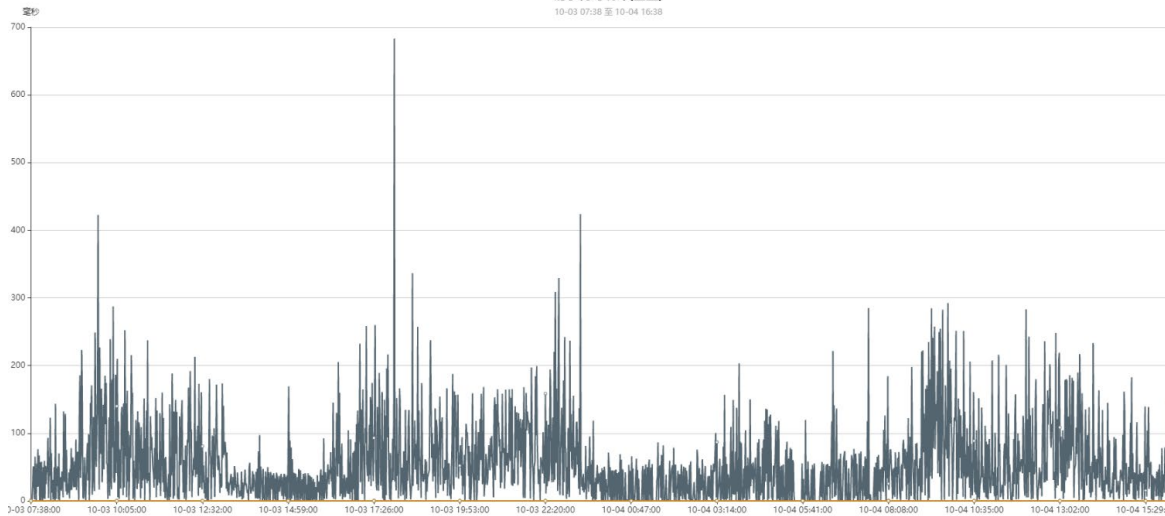
请求耗时统计(全量)

10-03 09:34 至 10-04 13:34



请求耗时统计(全量)

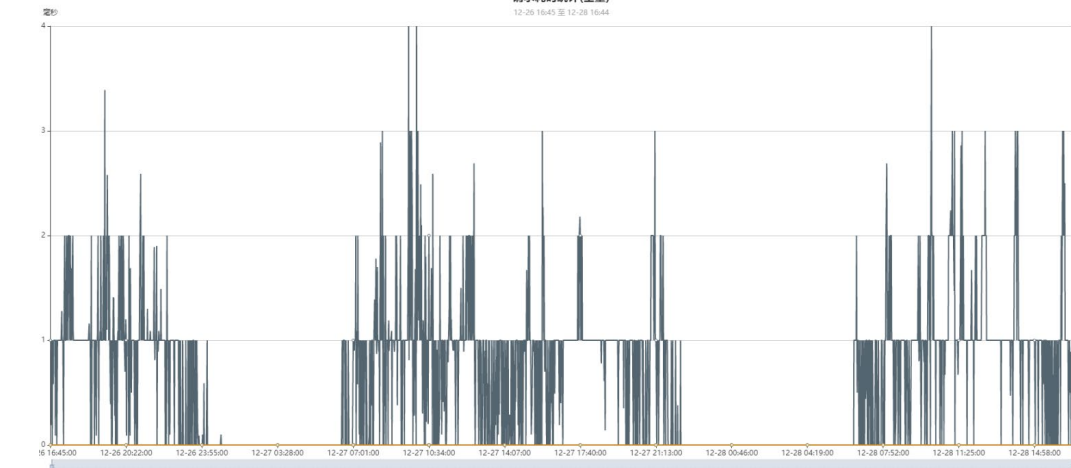
10-03 07:30 至 10-04 16:08



优化后主要接口时延如下:

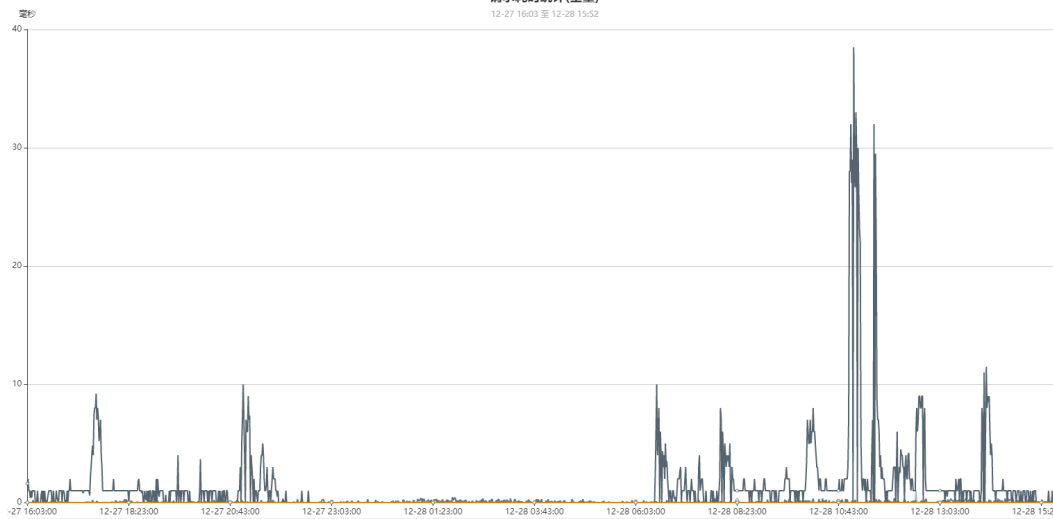
请求耗时统计(全量)

12-26 16:45 至 12-28 16:44



请求耗时统计(全量)

12-27 16:03 至 12-28 15:52



# OPPO互联网服务简介



可能外界对OPPO的认识仅仅停留在手机业务，这其实是一种误解。OPPO的互联网业务流量绝不输于部分一线互联网公司，当前OPPO互联网服务部分主流中间件、数据库等单集群流量峰值已经突破了百万级tps，其中OPPO文档数据库单集群最大峰值tps突破150万/s，总集群峰值tps突破400万/s，预计2020年左右文档数据库总tps突破1000万/s大关，欢迎加入OPPO互联网云存储团队参与挑战公司百万级高并发文档数据库研发。

## OPPO互联网文档数据库团队后续分享

近期继续分享如下主题，敬请关注：

1. 《百万级高并发mongodb集群性能优化采坑记》
2. 《线上典型集群抖动、不可用等问题汇总分析》
3. 《Mongodb文档数据库业务使用最佳案例分享》
4. 《Mongodb-3.6网络IO线程模型设计及代码实现》
5. 其他

## 作者简介

杨亚洲，前滴滴出行技术专家，现任OPPO文档数据库研发负责人，一直专注于分布式缓存、高性能服务器、分布式存储、数据库、中间件等相关研发，后续会持续分析Mongodb内核、wiredtiger存储引擎、rocksdb存储引擎源码及实现，相关分享会发布到我的git账号和oppo互联网技术微信公众号。

Github账号地址: <https://github.com/y123456yz>

邮箱: yangyazhou@oppo.com



## 建议:

文章中的一些优化方法并不是一定适用于所有mongodb场景，请根据实际业务场景和硬件物理资源进行优化，而不要按部就班。

# 欢迎关注MongoDB中文社区



MongoDB中文社区网址：  
<http://www.mongoring.com/>

MongoDB中文社区技术交流群：  
1.请关注微信服务号  
2.点击“社区互联”  
3.选择“联系我们”

MongoDB中文社区微信服务号+订阅号





mongoing  
中文社区

**mongoing.com**