# 【现网问题】【20250418】ceph10.2.10版本ceph-osd进程占用内存达9G问题定位

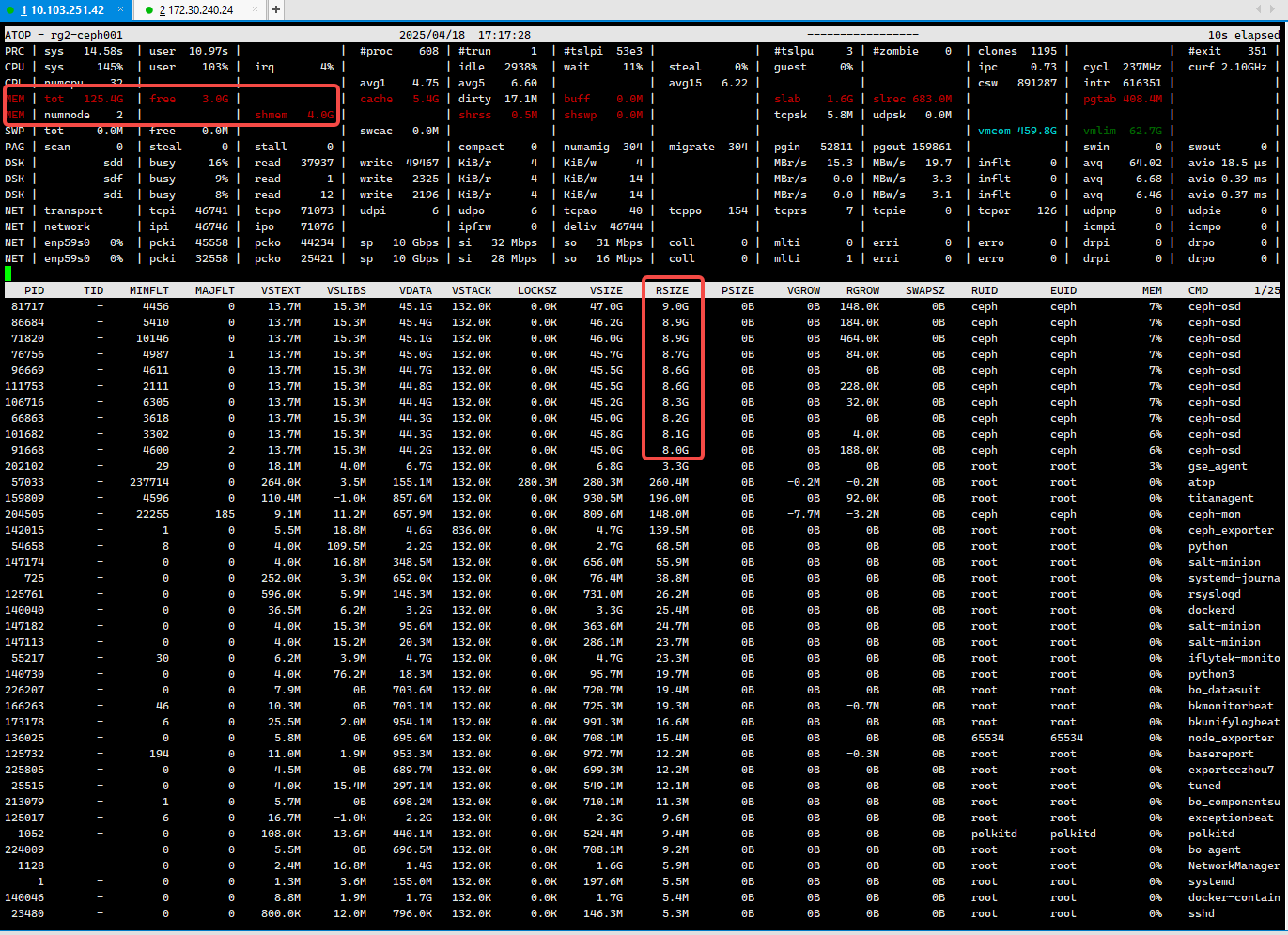
# 一、问题背景

存储运维操作6节点集群进行缩容两个节点操作，发现其他4个节点内存占用过高，几乎内存跑满。

问题现象

节点IP：[172.16.138.101](http://172.16.138.101" \t "_blank)-104

查看其中一个节点内存情况如下所示：



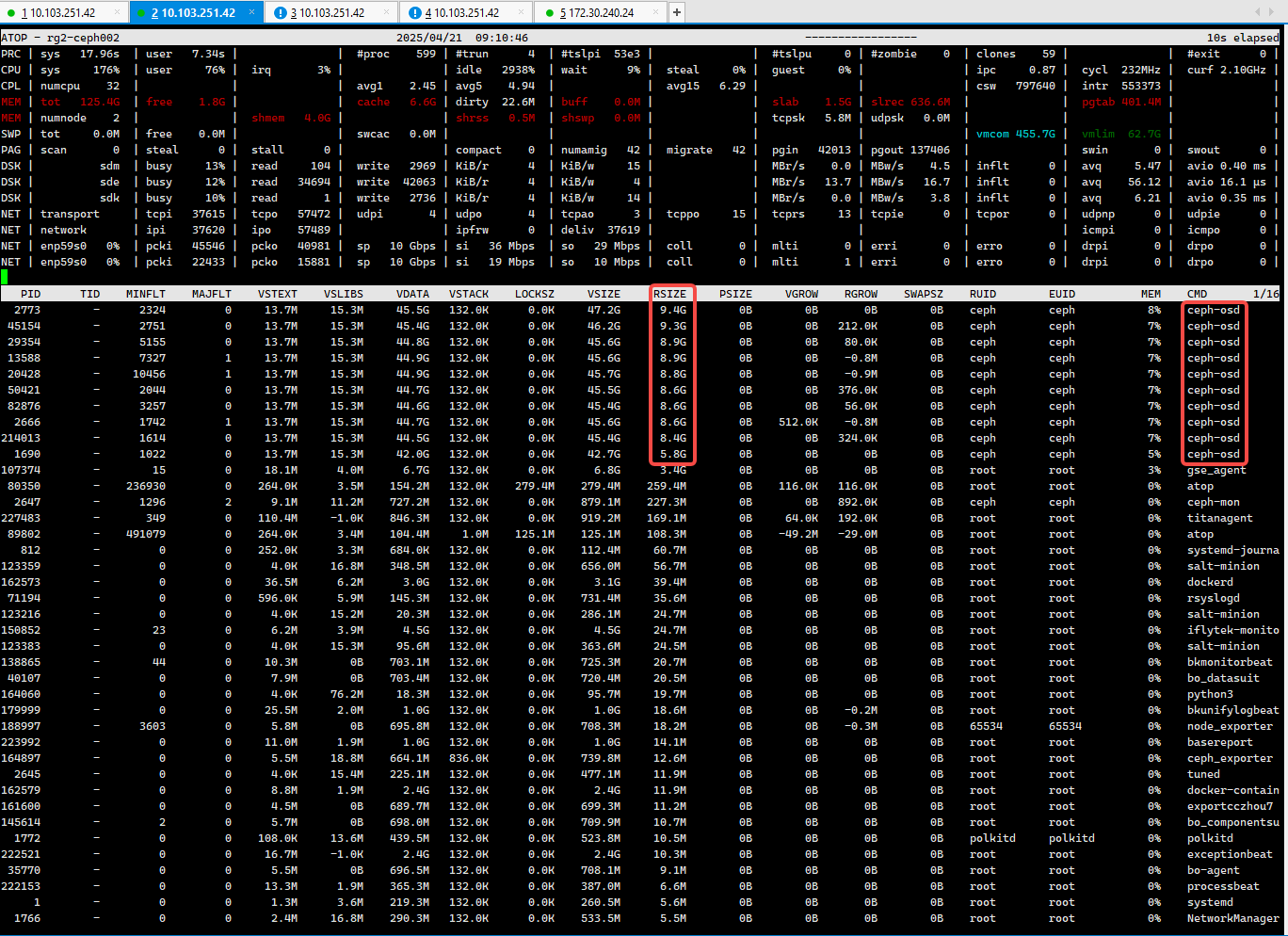
# **二、问题定位**

### 通过atop命令可以确认是ceph -osd进程占用内存过高，这个IP节点一共是10个OSD，每个OSD进程过去10s物理内存大小是9G，10个OSD就是90GB

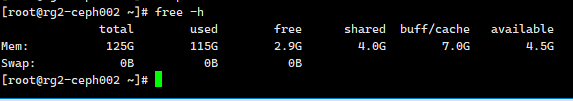
### **接下去就是为什么osd进程的内存占用高？是哪些占用**

### **查看节点内存以及iostat磁盘压力**

节点172.16.138.101，其他3个节点也一样

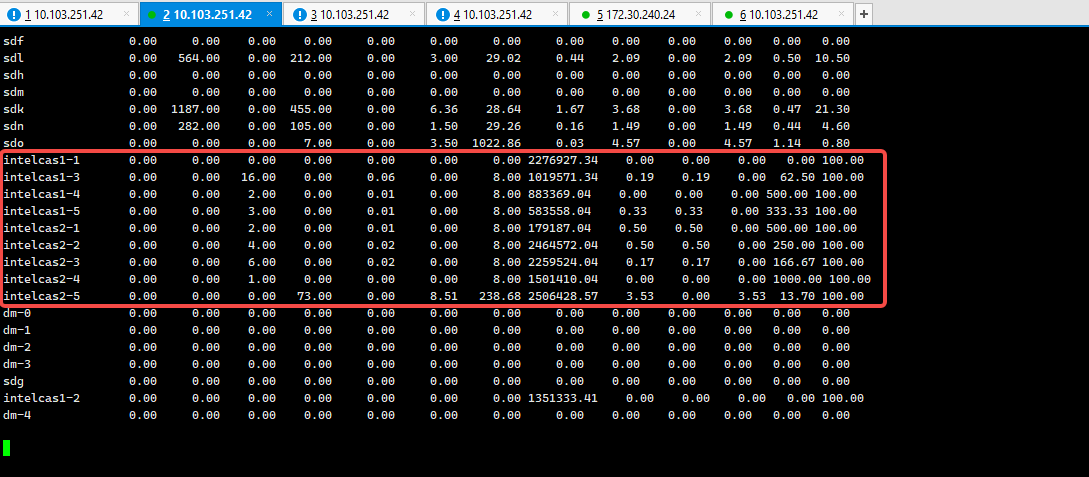


发现此时所有osd进程内存大小还是9G左右，未恢复，再查看free命令



2.9GB+4.5GB 这是可用内存大小。

查看对应节点的iostat

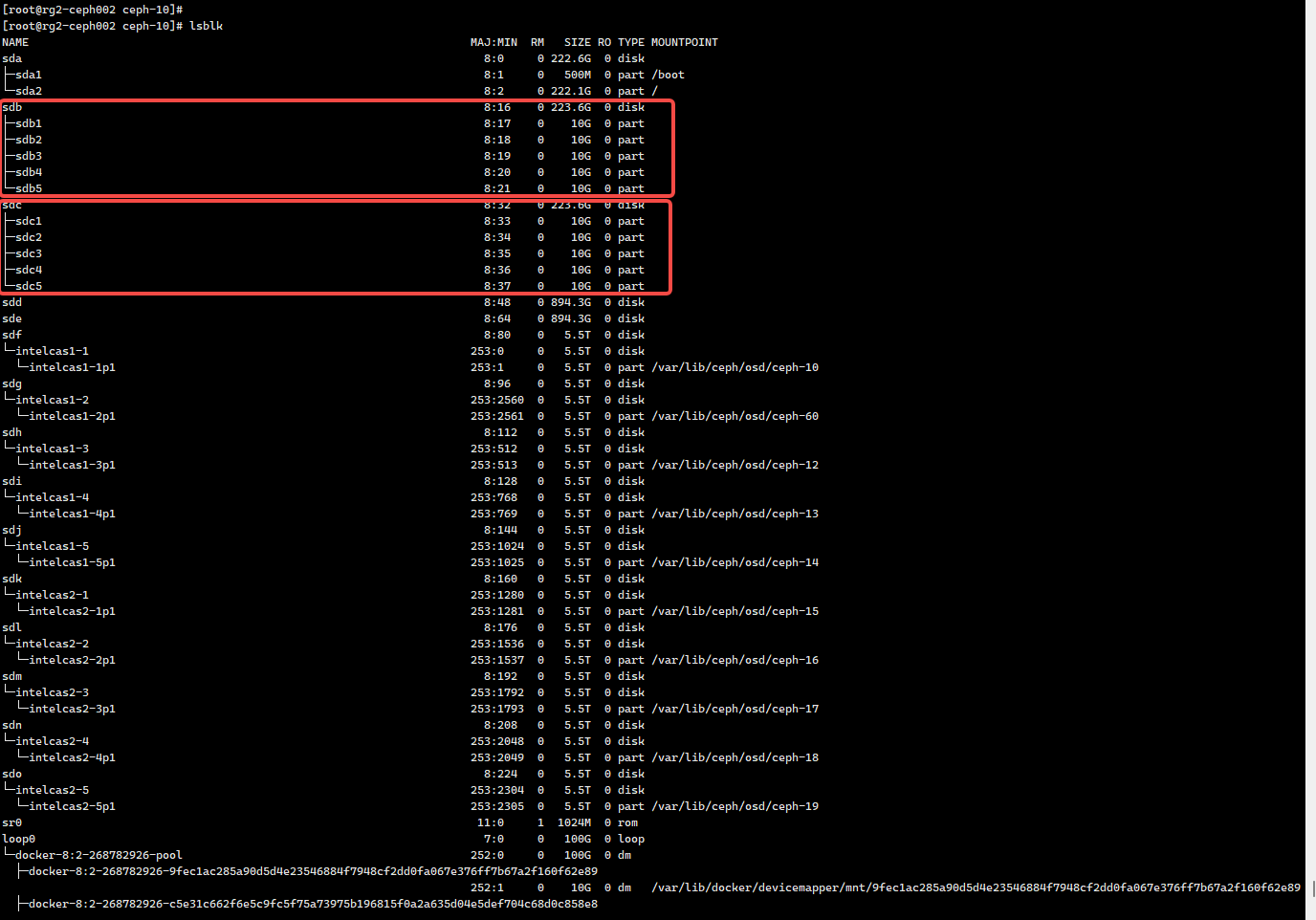


可以看到所有缓存盘跑满100%



从io业务来看，缓存盘队列深度排队严重，所有盘几乎跑满，达到盘瓶颈，肯定哪里不合理，继续分析缓存盘配置，如下所示，红框起来的是journal日志盘10GB，sdd、sde是缓存数据盘，未做分区，即单节点一块缓存盘拖5块hdd，有几点建议：

1. 建议缓存盘做分区;
2. 建议换nvme或大容量ssd缓存盘;
3. 建议扩容减少单节点缓存盘压力;





但是从free -h 确认不是缓存占用内存过多，排除了缓存因素。

### **执行[root@rg2-ceph001 sa]# cat /proc/buddyinfo**

Node 0, zone DMA 1 1 0 1 0 0 1 0 0 1 3

Node 0, zone DMA32 13 25 828 1647 529 136 22 5 2 2 29

Node 0, zone Normal 13631 1 0 1 0 0 0 0 0 0 0

Node 1, zone Normal 229420 87774 15152 1699 284 0 0 0 0 0 0

分析如下:



Numa1节点内存碎片更加严重，但是在其他节点验证尝试通过内核强制进行内存聚合，但是效果不明显，暂时不做修改。

### **5.通过执行sar -f sa21 -r命令获取历史时间点的内存情况**

暂无更进一步的线索。

[root@rg2-ceph001 sa]# sar -f sa21 -r

Linux 3.10.0-693.el7.x86\_64 (rg2-ceph001.cloud.hfa2.iflytek.net) 04/21/2025 \_x86\_64\_ (32 CPU)

12:00:01 AM kbmemfree kbmemused %memused kbbuffers kbcached kbcommit %commit kbactive kbinact kbdirty

12:10:01 AM 3116916 128368668 97.63 64 5861968 481811796 366.44 96648408 2559764 28648

12:20:02 AM 2311880 129173704 98.24 64 6529168 481834708 366.45 96703100 3173704 14392

12:30:01 AM 2249700 129235884 98.29 64 6947932 482011320 366.59 96684240 3636852 19036

12:40:01 AM 2354156 129131428 98.21 64 6846604 482088604 366.65 96750132 3487368 19432

12:50:01 AM 1580472 129905112 98.80 64 7366304 481821424 366.44 96719008 3995300 39612

01:00:01 AM 1265488 130220096 99.04 64 7646492 482123488 366.67 96773552 4258176 22232

01:10:01 AM 2088492 129397092 98.41 64 7156604 481808472 366.43 96611852 3902224 66660

01:20:01 AM 2870760 128614824 97.82 64 6277264 481832700 366.45 96713044 2937504 11296

01:30:01 AM 1042592 130442992 99.21 64 7891012 481824408 366.45 96760844 4499680 30936

01:40:01 AM 1932628 129552956 98.53 64 7155488 481836904 366.46 96731988 3808976 32820

01:50:01 AM 1636732 129848852 98.76 64 7444996 481823060 366.45 96749376 4086748 16684

02:00:01 AM 1686036 129799548 98.72 64 7446980 482065800 366.63 96736360 4137300 31220

02:10:01 AM 2799620 128685964 97.87 64 6336092 481842160 366.46 96685100 3063376 10768

02:20:01 AM 1499208 129986376 98.86 64 7662248 481829668 366.45 96800332 4268124 39632

02:30:01 AM 1150288 130335296 99.13 64 7807180 481847844 366.46 96843260 4390776 20776

02:40:01 AM 1321316 130164268 99.00 64 7676488 482097720 366.65 96856580 4244772 20620

02:50:01 AM 1444880 130040704 98.90 64 7500288 481847352 366.46 96819600 4123156 11564

03:00:01 AM 1497200 129988384 98.86 64 7333044 482141372 366.69 96888928 3922276 11728

03:10:01 AM 1502476 129983108 98.86 64 7574988 481845640 366.46 96779452 4248004 7396

03:20:01 AM 1511996 129973588 98.85 64 7514604 481835200 366.45 96824768 4112856 30272

03:30:01 AM 1025192 130460392 99.22 64 8021832 482108996 366.66 96856600 4625696 124944

03:40:01 AM 3028536 128457048 97.70 64 5868484 482105824 366.66 96816672 2526568 10844

03:50:01 AM 2763484 128722100 97.90 64 6166272 481855668 366.47 96799300 2824620 29288

04:00:01 AM 1624828 129860756 98.76 64 7301532 482064208 366.63 96912316 3886968 18156

04:10:01 AM 1342068 130143516 98.98 64 7486180 481997612 366.58 96914712 4042408 18968

04:20:01 AM 2353484 129132100 98.21 64 6720788 481844060 366.46 96739468 3445360 35972

04:30:01 AM 2332244 129153340 98.23 64 6781324 482009628 366.59 96746480 3512588 4828

04:30:01 AM kbmemfree kbmemused %memused kbbuffers kbcached kbcommit %commit kbactive kbinact kbdirty

04:40:01 AM 1215260 130270324 99.08 64 7904832 481850468 366.47 96693672 4674712 99160

04:50:01 AM 1815712 129669872 98.62 64 7452100 481828176 366.45 96701140 4184328 10276

05:00:01 AM 2814336 128671248 97.86 64 6247420 481893208 366.50 96797832 2939196 41428

05:10:01 AM 1269056 130216528 99.03 64 7678856 482435492 366.91 96928364 4182596 22876

05:20:02 AM 2403384 129082200 98.17 64 6585244 481867972 366.48 96761488 3255156 17244

05:30:01 AM 1387160 130098424 98.95 64 7494464 481863460 366.48 96799268 4126268 24780

05:40:01 AM 1217552 130268032 99.07 64 7590620 482119936 366.67 96914044 4144700 19248

05:50:01 AM 2032532 129453052 98.45 64 6945292 481861748 366.47 96879508 3514652 16636

06:00:02 AM 2025320 129460264 98.46 64 6981116 481898596 366.50 96861544 3605920 17768

06:10:01 AM 936092 130549492 99.29 64 8008904 482122828 366.67 96986448 4509024 57784

06:20:01 AM 3299916 128185668 97.49 64 5804920 481869980 366.48 97041088 2234824 24280

06:30:01 AM 1341868 130143716 98.98 64 7831368 481877876 366.49 97076188 4210868 23344

06:40:01 AM 2660092 128825492 97.98 64 6399456 481859024 366.47 96805460 3029716 31708

06:50:01 AM 2328432 129157152 98.23 64 6758132 481865364 366.48 97120660 3115136 13748

07:00:01 AM 876020 130609564 99.33 64 8139680 482151136 366.70 97172816 4481564 29868

07:10:02 AM 1431720 130053864 98.91 64 7533392 481871512 366.48 96848112 4144240 23108

07:20:01 AM 2863436 128622148 97.82 64 6203060 481877148 366.49 97181672 2477972 17764

07:30:01 AM 950924 130534660 99.28 64 7963404 481878908 366.49 97296128 4123788 25340

07:40:02 AM 1508204 129977380 98.85 64 7529660 481900492 366.50 96853756 4146968 42600

07:50:01 AM 876572 130609012 99.33 64 8246720 482117292 366.67 96997700 4710112 15252

08:00:01 AM 1332968 130152616 98.99 64 7724208 481888576 366.50 97152744 4046636 15236

08:10:01 AM 1457724 130027860 98.89 64 7633940 481852404 366.47 97173488 3904848 26988

08:20:01 AM 1204472 130281112 99.08 64 7856908 482029960 366.60 97165480 4159920 34420

08:30:01 AM 2706096 128779488 97.94 64 6360748 481863080 366.48 96774852 3065120 7660

08:40:02 AM 1735604 129749980 98.68 64 7349620 481883644 366.49 96857912 3998024 39092

08:50:01 AM 1512816 129972768 98.85 64 7524840 481869184 366.48 96854312 4170792 68848

09:00:01 AM 1778660 129706924 98.65 64 7142908 482439624 366.91 97043220 3644168 32552

09:00:01 AM kbmemfree kbmemused %memused kbbuffers kbcached kbcommit %commit kbactive kbinact kbdirty

09:10:01 AM 1136808 130348776 99.14 64 7839168 482098660 366.66 97273376 4067564 32648

09:20:01 AM 3344752 128140832 97.46 64 5696432 481857708 366.47 96946312 2256924 14632

09:30:01 AM 1982208 129503376 98.49 64 6672096 482036324 366.61 96892344 3192456 20784

09:40:01 AM 2388172 129097412 98.18 64 6313408 482337644 366.84 96787480 2952168 23148

09:50:01 AM 1699156 129786428 98.71 64 6883372 481964336 366.55 96924908 3375060 15892

10:00:02 AM 1944836 129540748 98.52 64 6670256 481426952 366.14 96911120 3199888 28796

10:10:01 AM 2139392 129346192 98.37 64 6456736 482056368 366.62 96899320 3008292 20140

10:20:01 AM 2505648 128979936 98.09 64 6093632 482007800 366.59 96839588 2699016 15532

10:30:01 AM 2491036 128994548 98.11 64 6072644 482141220 366.69 96821480 2726808 14752

10:40:01 AM 2577668 128907916 98.04 64 6085464 481956180 366.55 96817548 2724776 16452

10:50:01 AM 1819884 129665700 98.62 64 6762460 482109324 366.66 96918556 3306364 24916

11:00:01 AM 2520088 128965496 98.08 64 6087424 482332640 366.83 96836316 2738932 21608

11:10:01 AM 2337816 129147768 98.22 64 6251528 482319308 366.82 96920368 2797304 5064

11:20:01 AM 1768080 129717504 98.66 64 6775128 482151168 366.70 96918080 3331984 18936

11:30:01 AM 2083280 129402304 98.42 64 6463744 482167260 366.71 96875652 3067172 21308

11:40:01 AM 1700880 129784704 98.71 64 6922440 481975296 366.56 96911932 3471320 27924

11:50:01 AM 2447724 129037860 98.14 64 6162492 482364616 366.86 96853048 2797272 32628

12:00:01 PM 1628692 129856892 98.76 64 6823124 482029140 366.60 96958068 3360336 23096

12:10:01 PM 1293840 130191744 99.02 64 7169628 482146600 366.69 96960160 3688736 15544

12:20:02 PM 1494908 129990676 98.86 64 7013912 482014352 366.59 96932216 3551720 31428

12:30:01 PM 1688152 129797432 98.72 64 6821868 482178568 366.72 96925864 3369648 39976

12:40:01 PM 1423316 130062268 98.92 64 7149600 482025092 366.60 96946192 3687052 25396

12:50:01 PM 2326908 129158676 98.23 64 6272540 482251704 366.77 96857064 2897796 13488

01:00:01 PM 1967244 129518340 98.50 64 6978312 481830968 366.45 97007276 3509912 23092

01:10:01 PM 2040396 129445188 98.45 64 7016916 481798916 366.43 96989860 3541664 24704

01:20:01 PM 1387212 130098372 98.94 64 7611688 481797412 366.43 96985528 4142488 14600

01:30:01 PM 1045084 130440500 99.21 64 7932524 481985000 366.57 97266592 4193556 26000

01:30:01 PM kbmemfree kbmemused %memused kbbuffers kbcached kbcommit %commit kbactive kbinact kbdirty

01:40:01 PM 2365368 129120216 98.20 64 6609288 481820976 366.44 96934092 3196596 30440

01:50:01 PM 2342712 129142872 98.22 64 6680180 481812568 366.44 97322044 2873736 20648

02:00:01 PM 1483052 130002532 98.87 64 7647088 481861772 366.47 97454332 3733648 61924

02:10:01 PM 4253828 127231756 96.76 0 4635276 481987624 366.57 93745092 4297252 22840

02:20:01 PM 3989068 127496516 96.97 0 4816888 482266100 366.78 93764752 4471104 26452

Average: 1929440 129556144 98.53 63 6983735 481977934 366.56 96825592 3613383 26541



### **6.分析ceph osd进程堆内存分配信息**

通过执行

[root@rg2-ceph001 sa]# ceph tell osd.0 heap start\_profiler

osd.0 started profiler

[root@rg2-ceph001 sa]#

[root@rg2-ceph001 sa]# ceph tell osd.0 heap dump

osd.0 dumping heap profile now.

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

MALLOC: 4741122136 ( 4521.5 MiB) Bytes in use by application

MALLOC: + 1011515392 ( 964.7 MiB) Bytes in page heap freelist

MALLOC: + 1621235320 ( 1546.1 MiB) Bytes in central cache freelist

MALLOC: + 145152 ( 0.1 MiB) Bytes in transfer cache freelist

MALLOC: + 1123289648 ( 1071.3 MiB) Bytes in thread cache freelists

MALLOC: + 72744960 ( 69.4 MiB) Bytes in malloc metadata

MALLOC: ------------

MALLOC: = 8570052608 ( 8173.0 MiB) Actual memory used (physical + swap)

MALLOC: + 150052864 ( 143.1 MiB) Bytes released to OS (aka unmapped)

MALLOC: ------------

MALLOC: = 8720105472 ( 8316.1 MiB) Virtual address space used

MALLOC:

MALLOC: 870970 Spans in use

MALLOC: 5049 Thread heaps in use

MALLOC: 8192 Tcmalloc page size

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

Call ReleaseFreeMemory() to release freelist memory to the OS (via madvise()).

Bytes released to the OS take up virtual address space but no physical memory.

[root@rg2-ceph001 ~]# ceph tell osd.0 heap stats

osd.0 tcmalloc heap stats:------------------------------------------------

MALLOC: 4762546280 ( 4541.9 MiB) Bytes in use by application --------------------------> **应用程序实际使用的内存 4.5GB, OSD 进程自身（数据缓存、元数据、对象映射等），这是osd filestore自身对于数据和元数据缓存的内存大小，无论是否重启osd进程，这部分内存是不会变化的，除非从进程资源层面做限制，目前讯飞云存储均没有做这方面的限制，默认即可**

MALLOC: + 1104404480 ( 1053.2 MiB) Bytes in page heap freelist ----------------> 内存页堆管理的空闲空间链表,和tcmalloc内存管理有关

MALLOC: + 1495831608 ( 1426.5 MiB) Bytes in central cache freelist ----------------> 中心缓冲区占用，和tcmalloc内存管理有关

MALLOC: + 368128 ( 0.4 MiB) Bytes in transfer cache freelist -----------------------> 交换缓冲区占用

MALLOC: + 1109982560 ( 1058.6 MiB) Bytes in thread cache freelists ---------------------------------------------> 线程缓冲区占用 有3.6GB内存和ceph OSD进程无关，主要是和内存库有关系，这些未归还给OS内核

MALLOC: + 72744960 ( 69.4 MiB) Bytes in malloc metadata -----------------------------> malloc为元数据分配的空间

MALLOC: ------------

MALLOC: = 8545878016 ( 8150.0 MiB) Actual memory used (physical + swap)

MALLOC: + 174227456 ( 166.2 MiB) Bytes released to OS (aka unmapped)

MALLOC: ------------

MALLOC: = 8720105472 ( 8316.1 MiB) Virtual address space used

MALLOC:

MALLOC: 865556 Spans in use

MALLOC: 5008 Thread heaps in use

MALLOC: 8192 Tcmalloc page size

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

Call ReleaseFreeMemory() to release freelist memory to the OS (via madvise()).

Bytes released to the OS take up virtual address space but no physical memory.

101节点所有osd节点内存情况，目前是一直没有释放内存大小



通过执行命令: **ceph tell osd.0 heap release**

用于释放osd tcmalloc内存，输出如下：

[root@rg2-ceph001 ~]# ceph tell osd.0 heap stats

osd.0 tcmalloc heap stats:------------------------------------------------

MALLOC: 4758613376 ( 4538.2 MiB) Bytes in use by application -----------------------> osd 缓存数据和元数据未发生变化，这部分是和业务强相关

MALLOC: + 69763072 ( 66.5 MiB) Bytes in page heap freelist -----------------------------> 内存页堆管理释放内存归还给OS内核

MALLOC: + 1495624048 ( 1426.3 MiB) Bytes in central cache freelist -----------------------------------> 未发生变化

MALLOC: + 593920 ( 0.6 MiB) Bytes in transfer cache freelist

MALLOC: + 1112307984 ( 1060.8 MiB) Bytes in thread cache freelists -----------------------------------> 未发生变化

MALLOC: + 72744960 ( 69.4 MiB) Bytes in malloc metadata

MALLOC: ------------

MALLOC: = 7509647360 ( 7161.8 MiB) Actual memory used (physical + swap)

MALLOC: + 1210458112 ( 1154.4 MiB) Bytes released to OS (aka unmapped)

MALLOC: ------------

MALLOC: = 8720105472 ( 8316.1 MiB) Virtual address space used

MALLOC:

MALLOC: 865244 Spans in use

MALLOC: 5009 Thread heaps in use

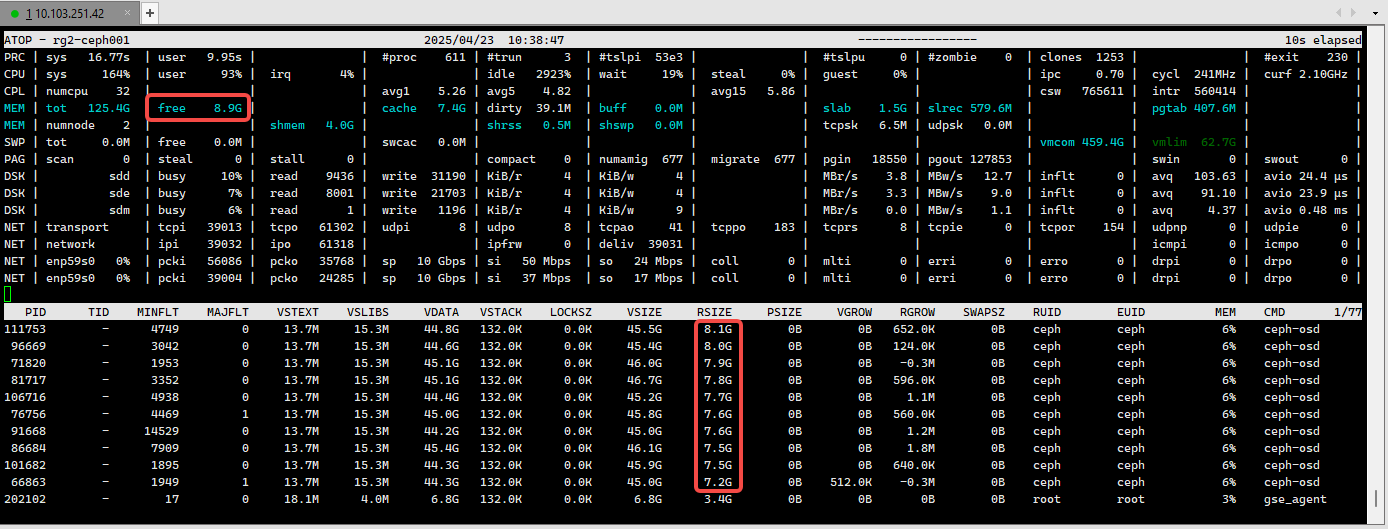
MALLOC: 8192 Tcmalloc page size

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

Call ReleaseFreeMemory() to release freelist memory to the OS (via madvise()).

Bytes released to the OS take up virtual address space but no physical memory.

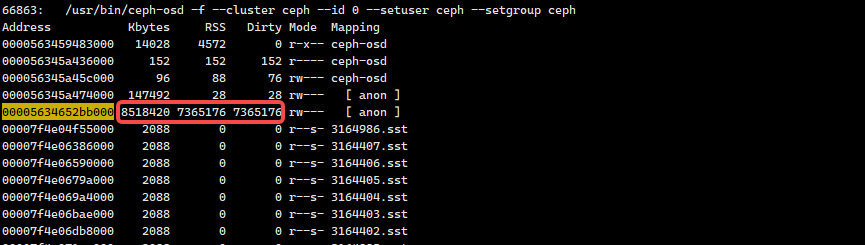
通过atop观察osd进程内存有明显下降，如下图所示：



Ds分析



继续分析pmap ceph-osd进程堆内存排布，简单截图如下图所示：



可以看到虚拟地址00005634652bb000的匿名内存占用8518420， 实际物理内存占用7365176， 通过pmap无法进一步分析，部分文件如下：



### **7.重启ceph OSD进程服务**

暂时无更进一步的分析，重启101节点 osd.0 ，查看一下进程内存前后变化的关系如下所示：

重启前：

[root@rg2-ceph001 ~]# ceph tell osd.0 heap stats

osd.0 tcmalloc heap stats:------------------------------------------------

MALLOC: 4760644592 ( 4540.1 MiB) Bytes in use by application ------------------------------------> 未发生明显变化

MALLOC: + 138149888 ( 131.8 MiB) Bytes in page heap freelist

MALLOC: + 1492099536 ( 1423.0 MiB) Bytes in central cache freelist

MALLOC: + 290816 ( 0.3 MiB) Bytes in transfer cache freelist

MALLOC: + 1113121344 ( 1061.6 MiB) Bytes in thread cache freelists

MALLOC: + 72744960 ( 69.4 MiB) Bytes in malloc metadata

MALLOC: ------------

MALLOC: = 7577051136 ( 7226.0 MiB) Actual memory used (physical + swap)

MALLOC: + 1143054336 ( 1090.1 MiB) Bytes released to OS (aka unmapped)

MALLOC: ------------

MALLOC: = 8720105472 ( 8316.1 MiB) Virtual address space used

MALLOC:

MALLOC: 865144 Spans in use

MALLOC: 5004 Thread heaps in use

MALLOC: 8192 Tcmalloc page size

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

Call ReleaseFreeMemory() to release freelist memory to the OS (via madvise()).

Bytes released to the OS take up virtual address space but no physical memory.

重启后：

[root@rg2-ceph001 ~]# ceph tell osd.0 heap stats

osd.0 tcmalloc heap stats:------------------------------------------------

MALLOC: 4619533576 ( 4405.5 MiB) Bytes in use by application ------------------------------------> 未发生明显变化

MALLOC: + 51912704 ( 49.5 MiB) Bytes in page heap freelist

MALLOC: + 80496112 ( 76.8 MiB) Bytes in central cache freelist -----------------------> 中央缓存链表占用内存减少

MALLOC: + 6662912 ( 6.4 MiB) Bytes in transfer cache freelist

MALLOC: + 46690824 ( 44.5 MiB) Bytes in thread cache freelists ---------------------------------------> 线程缓存链表占用内存减少

MALLOC: + 50331648 ( 48.0 MiB) Bytes in malloc metadata

MALLOC: ------------

MALLOC: = 4855627776 ( 4630.7 MiB) Actual memory used (physical + swap)

MALLOC: + 0 ( 0.0 MiB) Bytes released to OS (aka unmapped)

MALLOC: ------------

MALLOC: = 4855627776 ( 4630.7 MiB) Virtual address space used

MALLOC:

MALLOC: 554089 Spans in use

MALLOC: 4905 Thread heaps in use

MALLOC: 8192 Tcmalloc page size

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

Call ReleaseFreeMemory() to release freelist memory to the OS (via madvise()).

Bytes released to the OS take up virtual address space but no physical memory.

# **三、ceph osd进程内存分析**

1.ceph osd进程堆内存占用分为如下几部分：

(1) osd进程业务占用内存，包括了数据部分缓存和元数据缓存，对应Bytes in use by application

(2) Tcmalloc 内存页堆管理缓存， 对应Bytes in page heap freelist

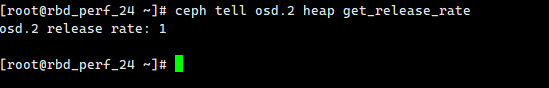
(3) Tcmalloc 中央缓存，对应Bytes in central cache freelist

(4) Tcmalloc 线程缓存，对应Bytes in thread cache freelists

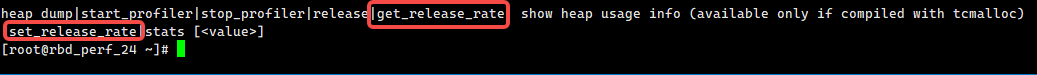
问题结论：

1.OSD进程业务内存占用差不多是4GB左右，这部分是正常的，当集群缩容后，业务迁移到这4个节点上，平均每个OSD承载PG数从80增到120，这4节点集群业务压力突增，tcmalloc分配更多的缓存用于业务IO，这部分缓存本身归还给OS的速率较慢，所以看起来是内存一直占着，未释放；

2.tcmalloc缓存本身会释放，但是释放单位粒度过小，从ceph15.2.17版本来看，释放粒度是1MB/s，如下图所示，推测tcmalloc本身释放内存给OS粒度过小；



Ceph15.2.17版本本身支持动态修改内存释放大小，命令如下所示：



但是ceph10.2.10版本不支持动态修改，命令仅支持释放页缓存，如下所示：



3.可以通过修改ceph-osd配置文件方式修改内存释放粒度，但是要重启osd进程服务，步骤如下：

3.1编辑 OSD 服务配置（以 osd.0 为例）

systemctl edit ceph-osd@0

输入: [Service]

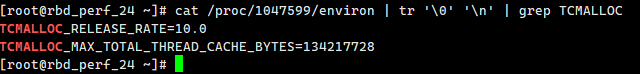
Environment="TCMALLOC\_RELEASE\_RATE=10.0"

3.2按 Ctrl + O 保存文件

3.3按 Enter 确认文件名（默认会保存到 /etc/systemd/system/ceph-osd@0.service.d/override.conf）

3.4按 Ctrl + X 退出编辑器

重启OSD进程服务，查看如下所示：



注意:该值单位是MB, 要根据具体业务情况进行调整。

# **四、问题方案**

1.单节点缓存盘%util达100%，已成瓶颈，磁盘排队严重，需要消耗更多内存且不释放，建议:

a.扩容集群，减少目前4节点业务压力;

b.更换nvme缓存盘或增大缓存容量，提升缓存盘处理效率，减少因IO排队带来的内存占用：

2.执行ceph tell osd.0 heap release 释放缓存页管理内存，可以缓解部分内存占用压力,对现网业务无影响，在禁止直接重启OSD服务情况下可用;

3.如果是ceph15.2.17版本可通过执行ceph tell osd.1 heap set\_release\_rate 100 命令动态调整内存释放速率降低内存占用高情况;

4.如果是ceph10.2.10版本，只能通过修改ceph-osd配置文件调整内存释放速率，且需要重启OSD进程服务;

操作步骤: a.systemctl edit ceph-osd@0 (以OSD.0 为例)

b.[Service]

Environment="TCMALLOC\_RELEASE\_RATE=100.0" (暂时设置内存释放单位为100MB/s)

c.按 Ctrl + O 保存文件

d.按 Enter 确认文件名

e.按 Ctrl + X 退出编辑器

5.直接重启OSD进程服务释放tcmalloc内存，降低内存占用;

6.使用cgroup限制OSD系统资源占用; 但是这种做法可能会导致OSD进程内存耗尽直接触发OOM kill，可能会带来风险，暂时先不用吧