# 现网ceph10.2.10版本出现slow\_req问题分析

## 1.问题背景

IM客服现网环境，版本如下：



## **2.问题现象**

IM客户现网环境概率偶现slow request，影响客服业务使用，当前已将有怀疑的host节点的osd全部down掉，本次抓取是2025年0321凌晨，重新拉起异常osd进行复现slow req现象，并抓取响应数据进行分析。

## **3.定位方案**

### 3.1 安装atop工具

现网环境需要安装atop工具，建议推广到块存储云主机所有集群，帮助分析当时系统资源使用情况

### **3.2 后台检测抓取异常osd数据**

1.讯飞云存储监控第一时间捕获到异常osd

2.后台触发脚本一键搜集相关信息，执行脚本内容如下：

d=$(date "+%Y-%m-%d\_%H-%M-%S")

cd /root/nvme9\_trace

blktrace -d /dev/nvme9n1 -w 15

blkparse -i nvme9n1 -d nvme9n1.blktrace\_${d}.bin > /dev/null

btt -i nvme9n1.blktrace\_${d}.bin > nvme9n1\_${d}.txt

ceph daemon osd.109 perf dump > 109\_perf\_dump\_${d}.txt

ceph daemon osd.109 dump\_ops\_in\_flight > 109\_dump\_ops\_in\_flight\_${d}.txt

ceph daemon osd.109 dump\_historic\_ops > 109\_dump\_historic\_ops\_${d}.txt

cd /root/nvme10\_trace

blktrace -d /dev/nvme10n1 -w 15

blkparse -i nvme10n1 -d nvme10n1.blktrace\_${d}.bin > /dev/null

btt -i nvme10n1.blktrace\_${d}.bin > nvme10n1\_${d}.txt

ceph daemon osd.110 perf dump > 110\_perf\_dump\_${d}.txt

ceph daemon osd.110 dump\_ops\_in\_flight > 110\_dump\_ops\_in\_flight\_${d}.txt

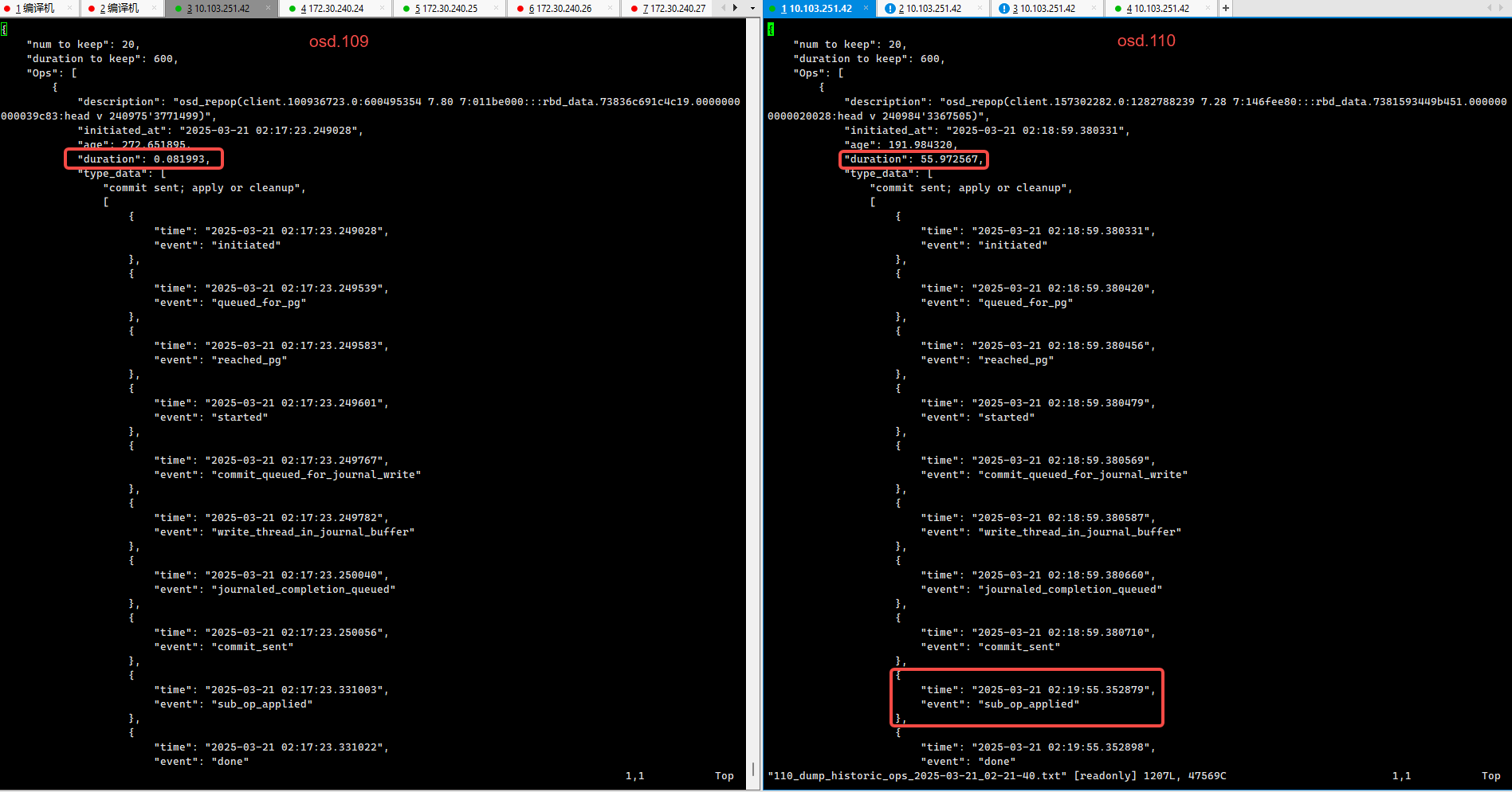
ceph daemon osd.110 dump\_historic\_ops > 110\_dump\_historic\_ops\_${d}.txt

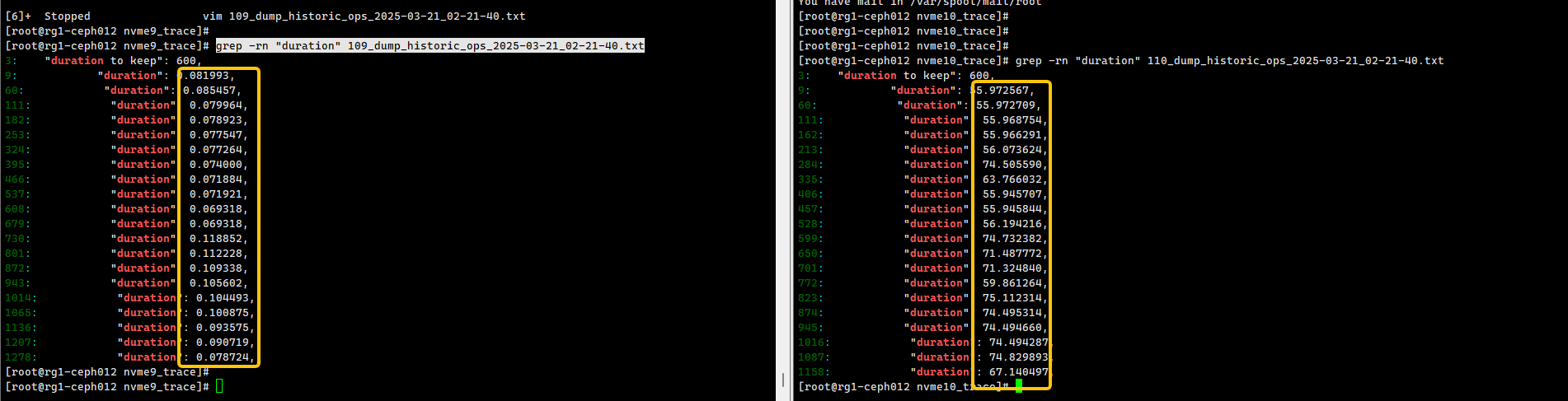
注意：该脚本是已知nvme9、nvme10两块盘有异常，所以需要修改该脚本针对监控捕获有异常的osd进行创建目录，并收集对应osd的信息

增加osd线程cpu占比高的火焰图，排查内存紧张问题

## **4.定位分析**

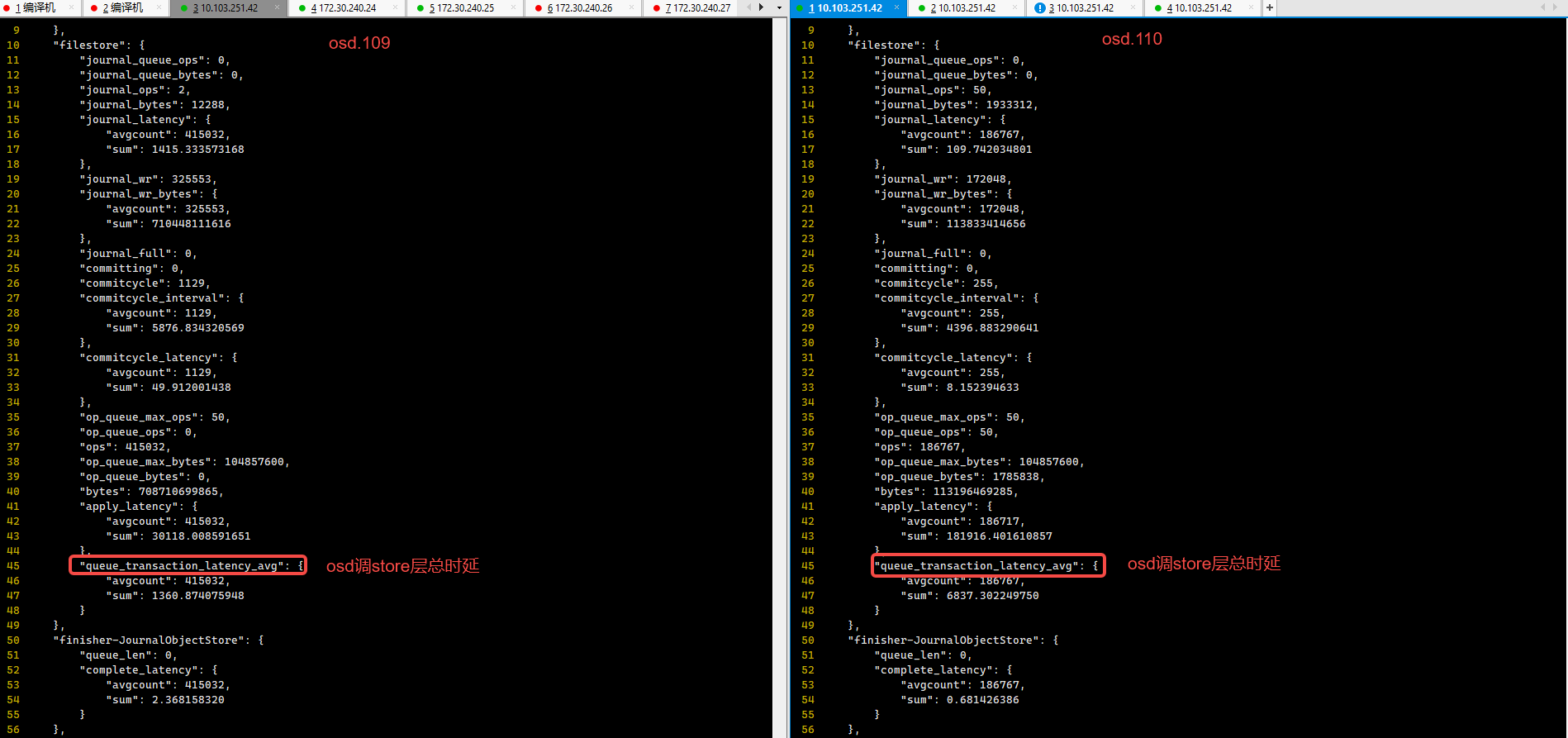
1.现网执行的动作是重新拉起osd109、osd110两个，进行对比分析，先看ceph daemon osd.x dump\_historic\_ops命令输入结果如下：



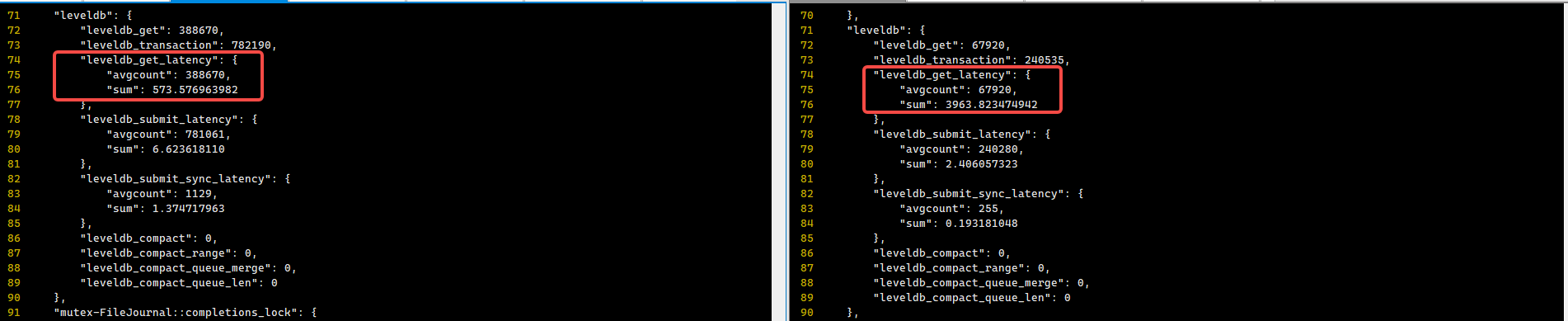


可以看到osd.109时延是正常的，没有超过slow op的阈值，所以osd.109是正常拉起的，但是osd.110明显出现时延较大的op，通过第一张截图，可以明确大时延主要在写从osd的store层，data部分较慢。

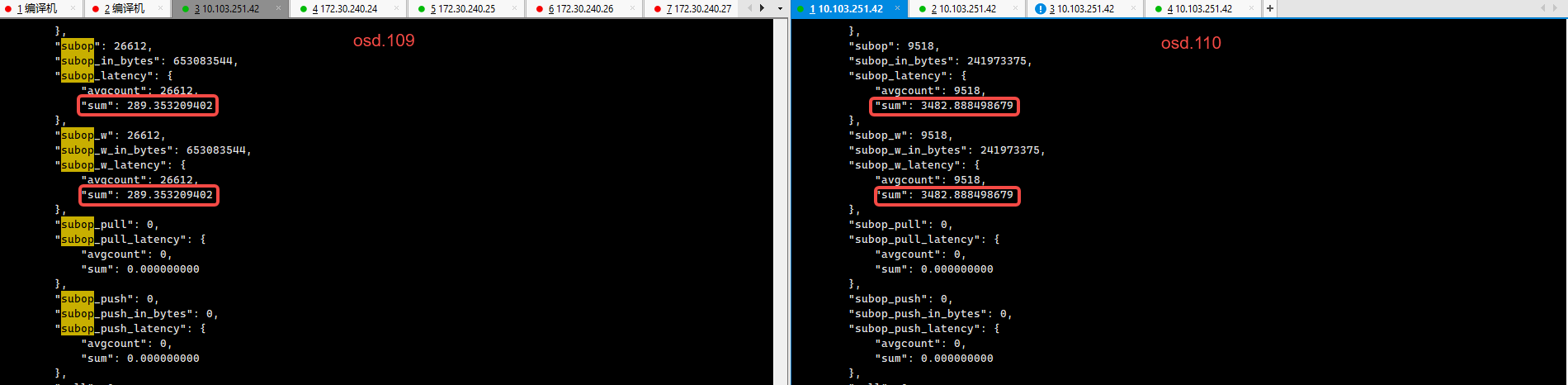
2.继续查看对应osd的perf dump文件



osd.110耗时为6837.302249750 / 186767 = 0.03660s = 36.6ms (单个io)



Leveldb get操作时延较高， 单io时延为0.0583s = 58.3ms



左侧是osd.109, 右侧是osd.110，可以看到subop latency

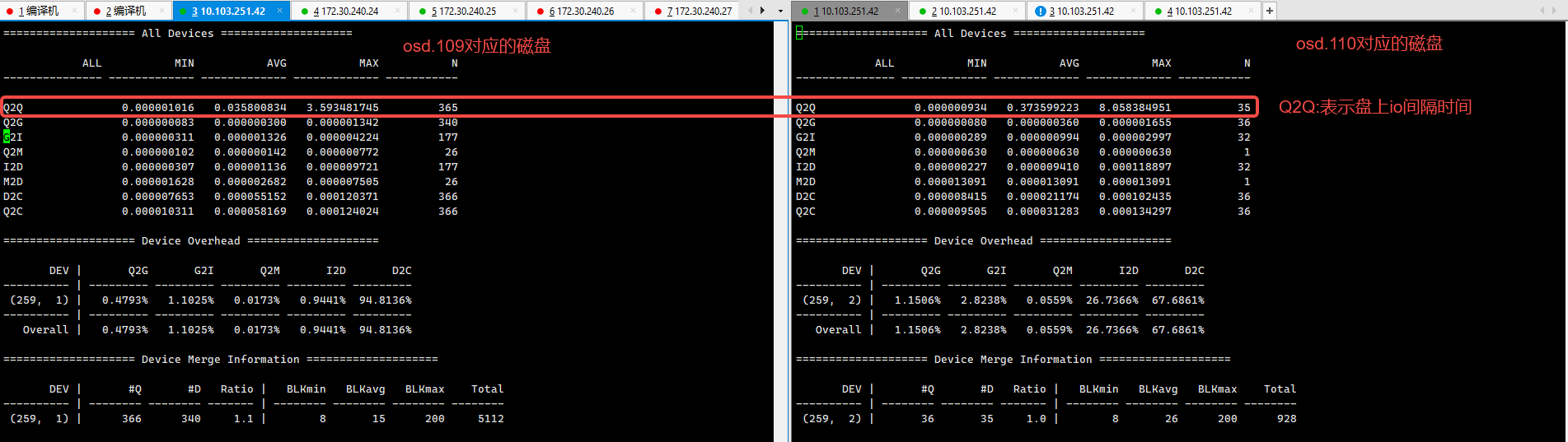
Osd.109从osd一共是写26612个io，耗时是289.353209402s， 单个io耗时0.01087s=10.87ms

osd.110一共是写9518, 耗时是3482.888498679s, 单个io耗时0.36592s=365ms

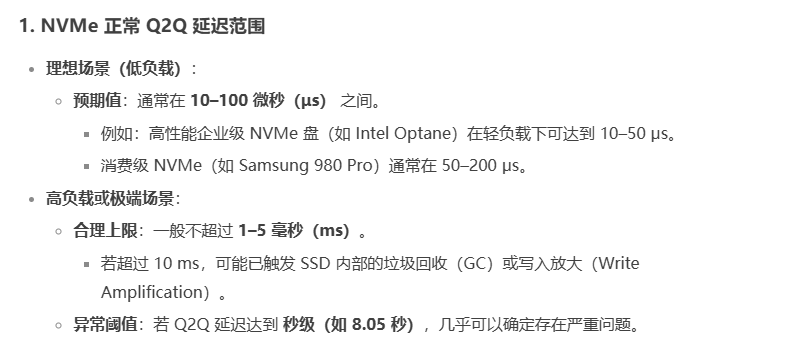
可以看到osd.110比osd.109 从osd写时延高了一个数量级，非常明显；注意的是，ceph daemon osd.x dump\_historic\_ops 和perf dump命令不能对应起来看，dump\_historic\_ops是确定的一个op时延，perf dump是该osd上处理的所有op的均值，只能大致对应来看。

从以上perf dump数据来看，osd.110写db和写data的时延相对osd.109来说，都是比较高的，从osd统计来看，osd.110承载的io业务并不比osd109的多，所以不太可能是osd.110 业务负载高导致，而且所有的软件配置都是一样的，也排除因软件配置不同导致问题。

3.继续查看两个盘本身的处理时延

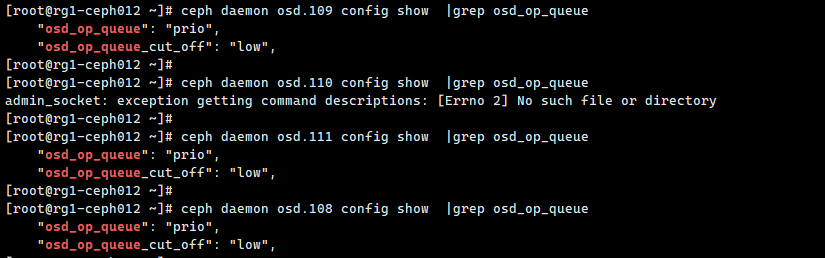


从Q2Q的数据来看，似乎是异常的，这是deepseek给出的结论：



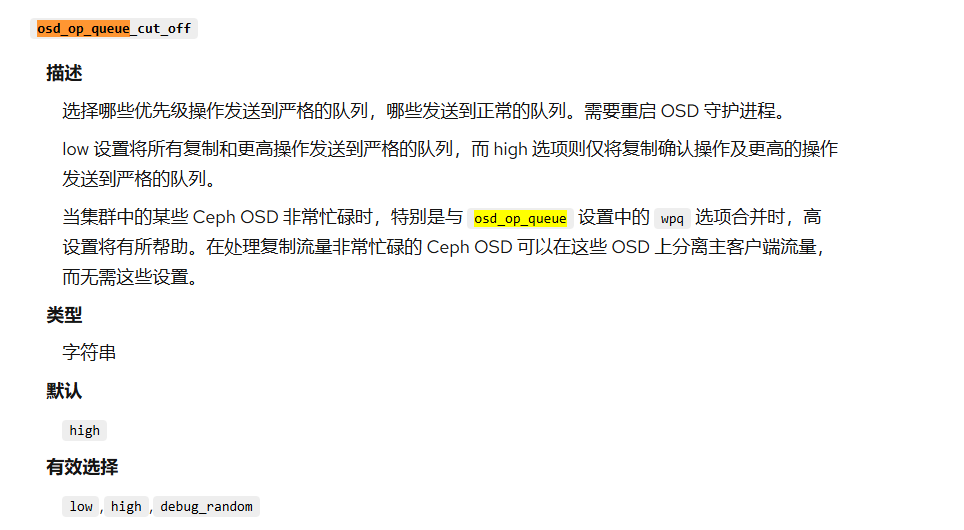
但是从Q2C、D2C数据来看，又似乎磁盘本身没问题，Q2C表示请求到完成的整个时延，D2C表示磁盘驱动从接收到op到处理完成的时延，所以这两组数据来看，又是矛盾的。

4.继续分析ceph osd配置，发现有如下差异



其中ceph osd的两个配置项如下:





5.从目前的几个怀疑点出发，store层的处理时延较异常，20250321凌晨IM客服业务基本停了，理论上不应该有大量的io下来，当时osd.110出现的slow op推测是我们自己数据迁移出现的，可能是盘本身的问题(只是怀疑，从抓取数据看，有一定的怀疑度)，也可能是store层本身软件的问题，但是从htop来看，cpu远远没有达到瓶颈，目前结合dump\_historic\_ops、perf dump命令无更进一步的数据统计。

6.介于当前集群概率偶现slow op问题，一定是当数据量一定时，物理资源存在瓶颈，物理资源无外乎磁盘、网络、cpu、内存，从blktrace看，磁盘可能存在问题，网络推测应该也不可能存在瓶颈，cpu通过htop分析远达到瓶颈，还有一个就是物理内存，这个需要再抓一下火焰图，或者直接增加内存条到256G

7.下一步还需要再抓一下出现slow op时，osd cpu占用高线程的火焰图。

## **5.解决方案**

1.现网集群节点统一安装atop命令

2.继续复现和抓取数据，出现slow op时，osd cpu占用高线程的火焰图(同时观测网络是否达到瓶颈上限、同时抓取上述数据)

按照硬件资源配置难易程度

提升cpu利用率

2.修改ceph osd\_op\_queue、osd\_op\_queue\_cut\_off配置改为如下参数



3.修改osd线程数，提高cpu使用率

增加内存

4.从128G增加到256G

5.置换nvme10n1盘

策略: 从研发环境替换一块同等容量nvme盘，继续观测分析