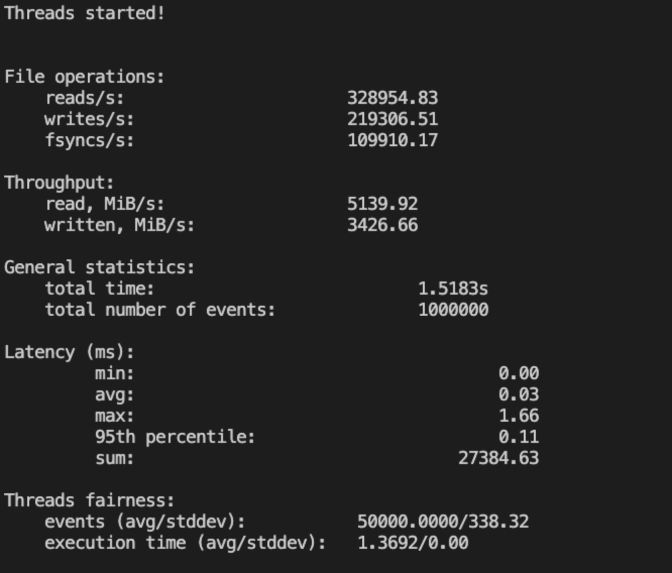
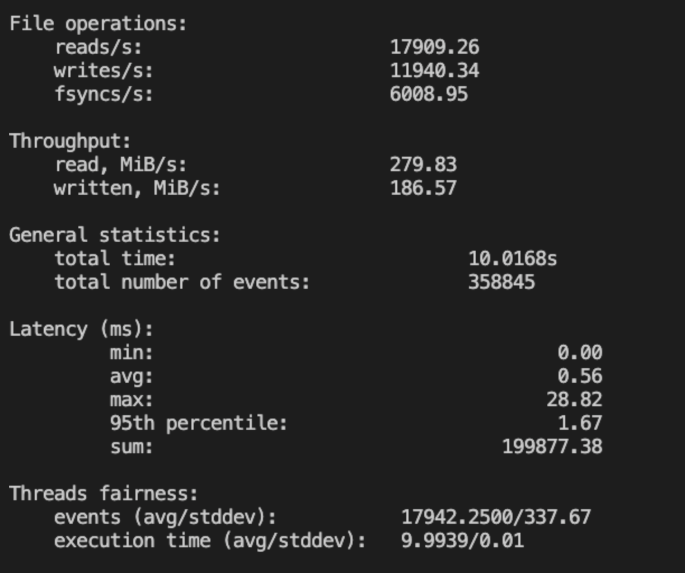
## 问题描述

1、现网支撑环境某台物理机上面运行的 PostgreSQL 实例写入速度比正常慢 4~5 倍。

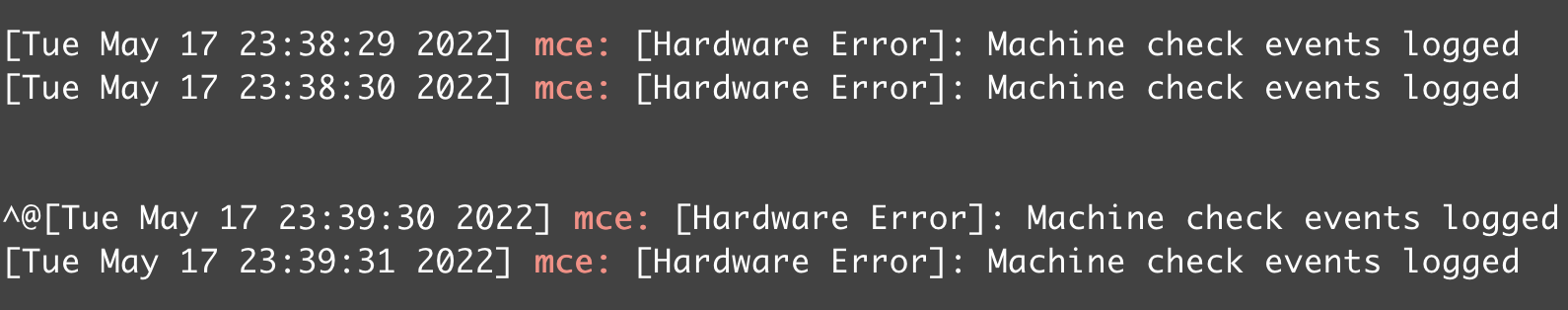
2、该机器包括 4 个 raid 盘（每个 raid 盘关联 4 个 NVMe 固态盘）md1 ~ md4，分别挂载目录 /data1 ~ /data4。使用 sysbench fileio 分别压测 /data1、/data2，/data1 性能结果符合预期，/data2 比较慢，不到 /data1 的 1/5。

   vs  

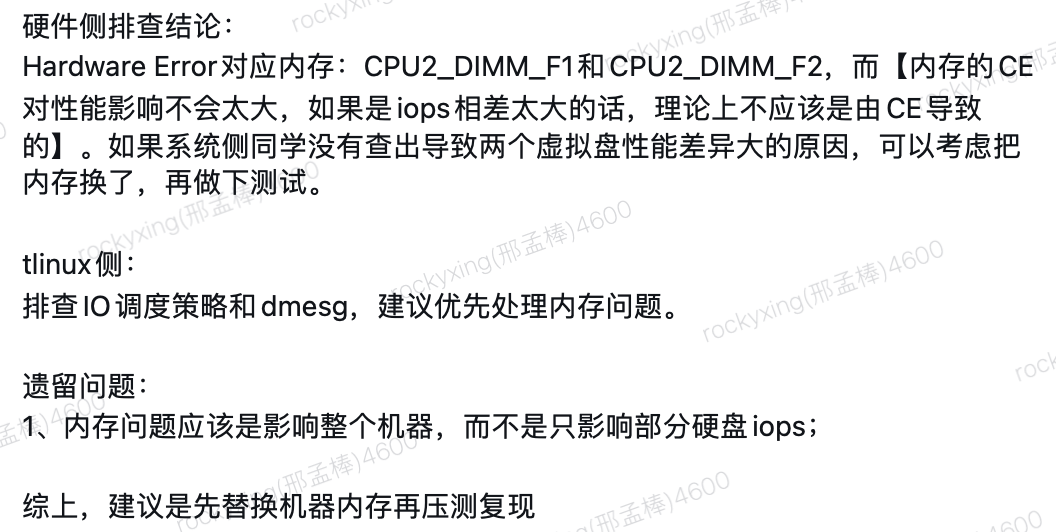
                sysbench fileio /data1                                                     sysbench fileio /data2

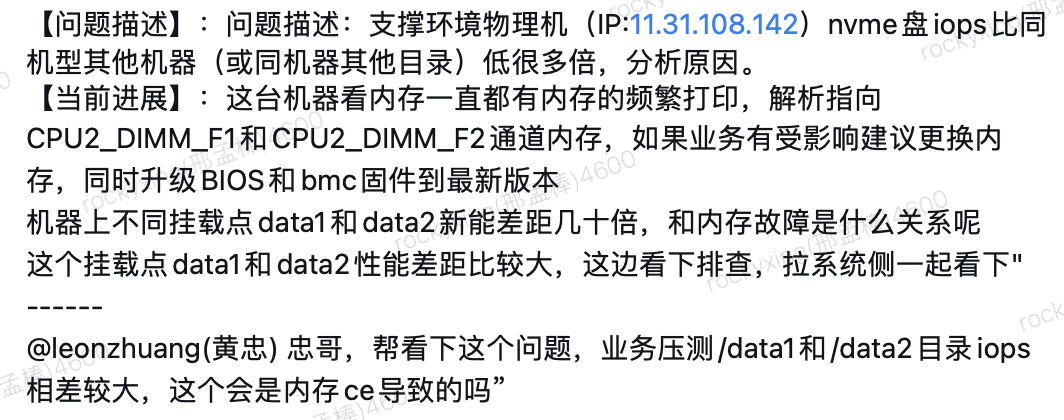
压测命令：sysbench --file-num=20 --threads=20 fileio --file-total-size=200G --events=1000000 --file-test-mode=rndrw run

3、dmesg 包含 mce 错误信息，具体如下。



4、硬件侧、tlinux 侧给出的结论（只是怀疑跟 MCE 有关系，无法明确解释 /data1、/data2 性能差别很大的原因）





## 问题分析

1、首先确认 /data1、/data2 挂载的磁盘（md1、md2）自身是否存在性能差别。

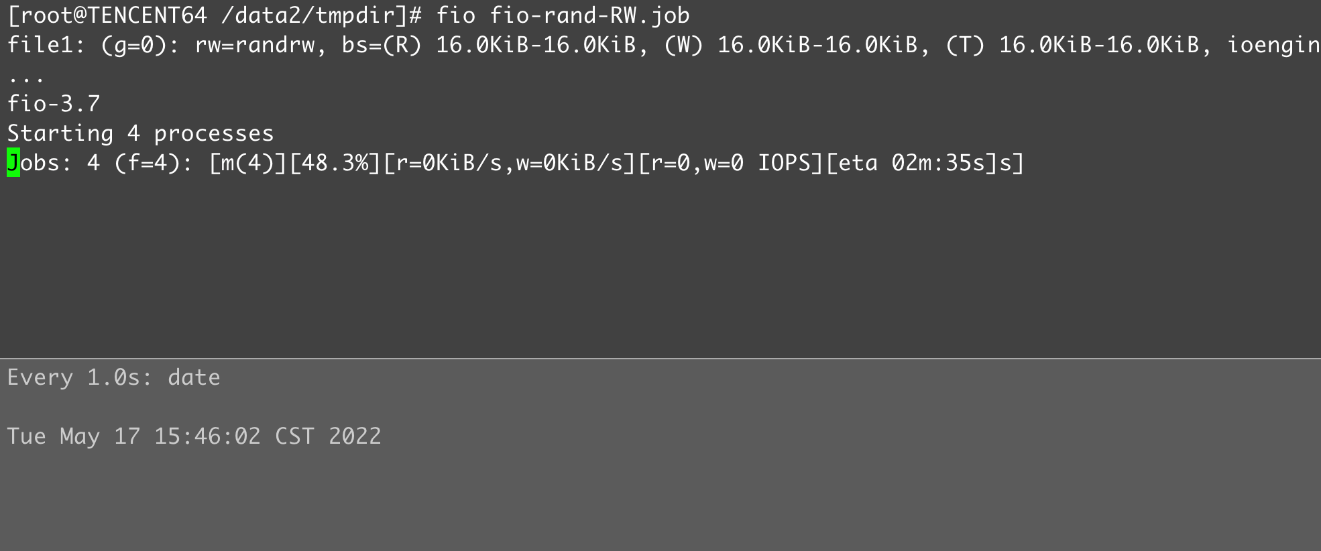
fio 直写 md1、md2，结果基本相同，且与其他物理机上同规格磁盘基本相同，说明磁盘自身不存在 I/O 性能问题。

2、步骤 1 已排除磁盘自身因素，由于 sysbench fileio、PostgreSQL redo log 写入均采用 bufferd I/O（先写 page cache，再异步刷新至磁盘），接下来验证内存（ page cache）对结果的影响。

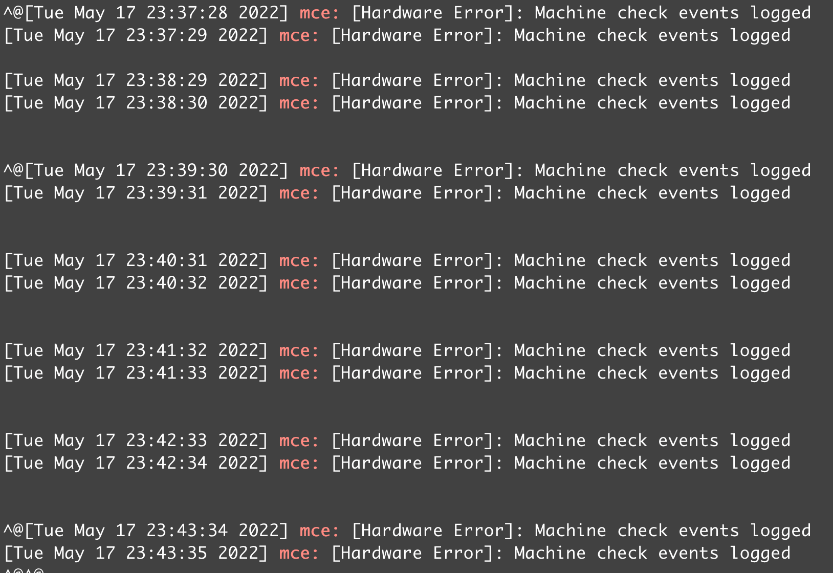
使用 fio 压测，direct=0，无论 data1 还是 data2：  
> IOPS 周期性降低至 0，然后逐渐增加，达到峰值后又会掉落下来。测试时长 5 分钟之内，均出现 6 ~ 7 次下降至 0 的情况。期间 mce 错误日志刷了 5 ~ 6 次。  
> 平均 IOPS 低于 direct=1 测试结果（不及 1/2），先写文件缓存比直写硬盘性能要差，内存是影响写入速度的主因。

结果附图：

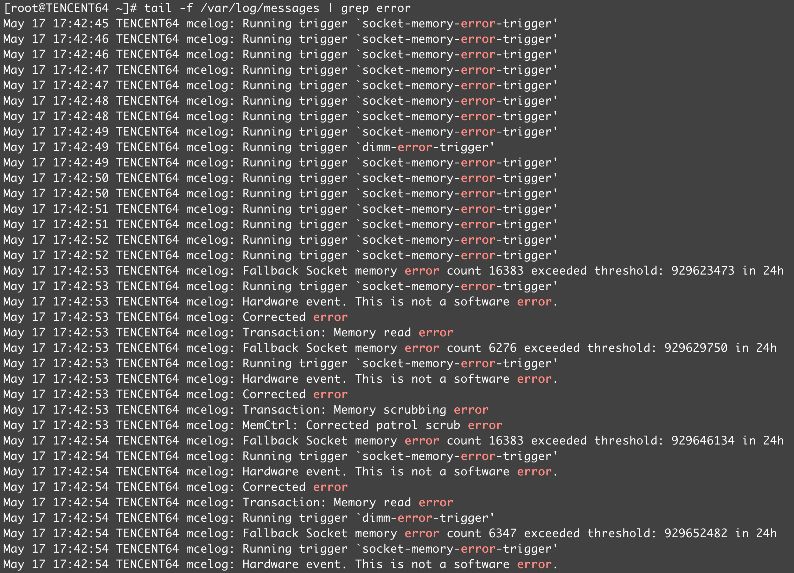
> fio 速度降为 0 的其中一次，持续几秒钟。



> mce 错误信息（看起来是每分钟输出一次）



> mcelog 具体错误

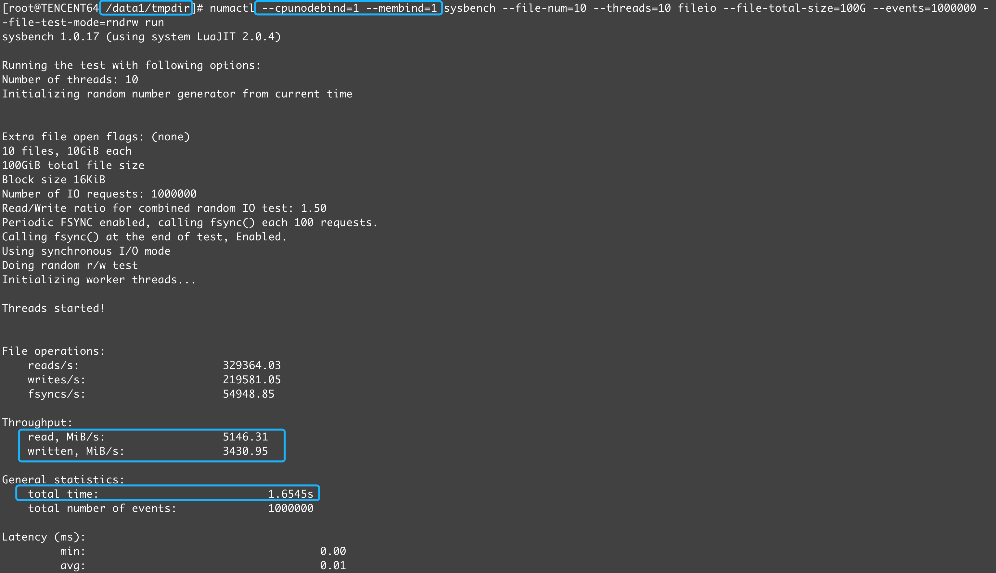
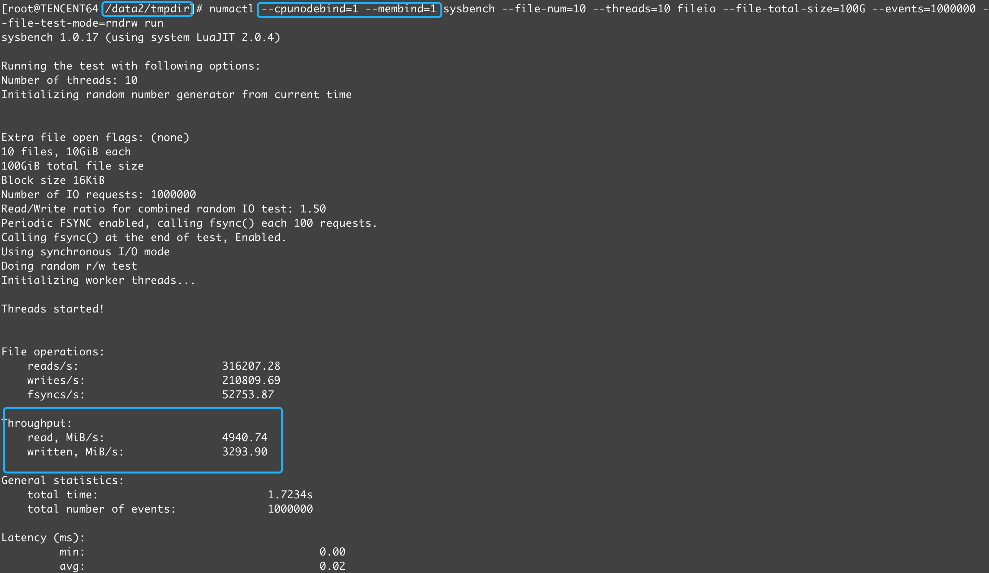


3、以上分析并不能解释为什么 /data1、/data2 分别 sysbench fileio 压测结果却存在较大差异。接下来进一步做以下分析。

怀疑点：/data2 上进行 sysbench fileio 压测时，page cache 不巧命中有问题的内存条（机器共 2 个 NUMA 节点，一共 24 根内存条）。

验证：绑定 NUMA 1，sysbench fio 无论在 data1 还是 data2，表现基本一样，差距很大的现象消失。

结论：NUMA 0 上面的内存条有问题，NUMA 1 看起来问题不大，解释了为什么 /data1、/data2 分别 sysbench fileio 压测结果却存在较大差异的现象。

 vs 

                                   sysbench fileio /data1                                                                                     sysbench fileio /data2

## 结果验证

根据上述分析结果，PostgreSQL DBA 换了 2 根内存条，再次使用 sysbench fileio 压测，无论 /data1 还是 /data2，结果均表现正常，证明了 MCE 是导致写入性能变差的主因。