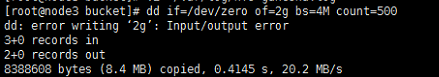
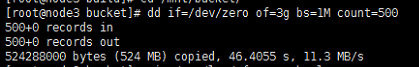
1. 现象

cp数据，rsync拷贝均正常

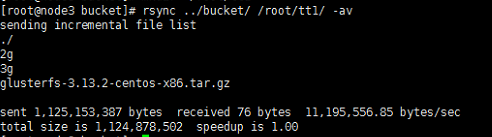
dd测试写入bs超过1M则偶现input io error







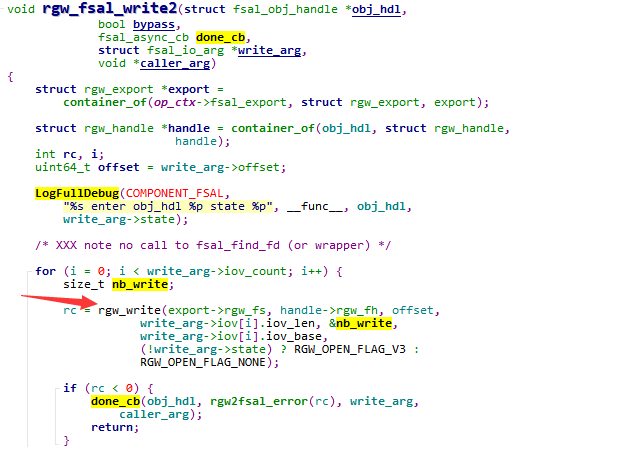




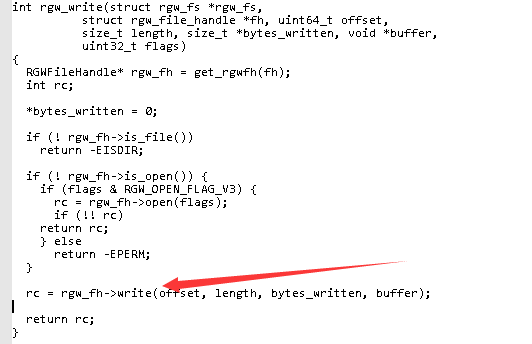
1. 分析

2.1 错误由dec\_nfs4\_state\_ref打印，调用点为写回调nfs4\_write\_cb

2.2 ganesha写函数rgw\_fsal\_write2最终调用ceph-rgw模块rgw\_write()



2.3 从反馈的错误码来看 -5 Input/output error，表示write步骤写入错误





2.4 打印offset偏移消息， 固定每次写入1048576

如下是写入bs为1M时选取的消息：

offset 519045120 size 1048576

offset 520093696 size 1048576

offset 521142272 size 1048576

offset 522190848 size 1048576

offset 523239424 size 1048576

如下时写入bs为4M时选取的消息：

1 offset 6291456 size 1048576

2 offset 7340032 size 1048576

3 offset 11534336 size 1048576

4 offset 8388608 size 1048576

5 offset 9437184 size 1048576

6 offset 10485760 size 1048576

可以看到写入4M时，由于每次写入大小为1M，则需要切分io 由ganesha统一下刷到rgw\_file, 而buffer长度被设置为了8，按照rgw的逻辑当3写入到来时会放入队列固定位置,当4写入时在找不到5，6没有下刷，则判断队列为空但是3仍然存在，则执行4失败

1. 解决方案

3.1 从ceph-rgw入手，修改rgw处理未对齐和无序的写入

--参考：

<https://github.com/ceph/ceph/pull/15623/commits/6025f6dd6c8af6981711397d666a231b2547f819>

3.2 优先符合使用场景，目前ganesha目录的功能设计为备份对象集群数据，即只有读需求。