**纠删码存储系统中的投机性部分写—背景综述**

**纠删码与部分写简介**

多副本和纠删码（EC，Erasure Code）是存储系统中常见的两种数据可靠性方法。与多副本冗余不同，EC将m个原始数据块编码生成k个检验块，形成一个EC组，之后系统可最多容忍任意k个原始数据块或校验块损坏，都不会产生数据丢失。纠删码可将数据存储的冗余度降低50%以上，大大降低了存储成本，在许多大规模分布式存储系统中已得到实际应用。

然而纠删码技术（EC）给写操作带来了很大的额外开销，包括编解码计算开销和流程性开销两部分。在向量指令集SSE、AVX等的帮助下，一个现代CPU核心的EC编解码能力就可以达到几GB到十几GB每秒，远远大于存储设备的I/O吞吐率。这使得流程性开销成为EC写入性能的最重要制约因素。若一次写操作的偏移和长度没有对齐EC组，就需要部分更新涉及的EC组，因而将此类操作称为部分写。部分写带来了大部分的流程性开销。

**早期处理部分写的方法**

处理部分写的最直接办法是将不被写的数据块读出来，跟新数据组合在一起，然后再整体编码并写入。目前Ceph、Sheepdog等系统都采用了这种办法。

一种简单有效的改进是将被覆盖数据的原始值读出来，然后根据新旧数据的差值来进行增量编码，得到各个校验块的差值，并“加”到各个校验块上。这种方法可以显著减少系统总体I/O次数，然而它需要对涉及的数据块和所有校验块进行原地读写（即在同一位置进行先读后写），在没有缓存的情况下（常态），HDD需要花费8.3毫秒（7200RPM）的时间旋转一周才能完成写入请求，跟一次随机I/O的平均寻道时间相当。这样的流程极大地影响了写入效率，即便应用层面发出的是顺序写操作，最终得到的性能也跟随机写差不多。

**最新进展**

目前业内已有许多工作对部分写这一问题进行改进，其中最具代表性的工作是2014年发表在[FAST](https://www.usenix.org/conference/fast14)技术会议上的，由香港中文大学的Jeremy C. W. Chan, Qian Ding, Patrick P. C. Lee, and Helen H. W. Chan提出的[《Parity Logging with Reserved Space: Towards Efficient Updates and Recovery in Erasure-coded Clustered Storage](https://www.usenix.org/node/179852)》，它的核心思路是通过在校验块上记变更日志的方式来减少其上不必要的读操作，同时将随机写变成顺序追加写，以改善写入性能。然而它并不能改善原始数据块上的写流程，这构成了大多数的写操作，所以系统总体写性能改善有限。

**面临的问题与挑战**

在实际应用当中，只有写操作的偏移和长度都恰好跟EC组对齐才可以避免部分写，然而应用往往无法照顾到底层存储的实现细节和参数，所以部分写构成了写操作的主体，决定了EC存储系统的实际写性能。EC模式的部分写性能大大低于三副本写，这使得EC尚不适用于写操作较多的场合，如云硬盘。

因此我们希望设计出一种改善的方法，可以提高系统的总体写性能，让EC模式的部分写性能可以与三副本写相媲美，让EC存储系统具有更高的适用性。