希捷瓦记录磁盘评测

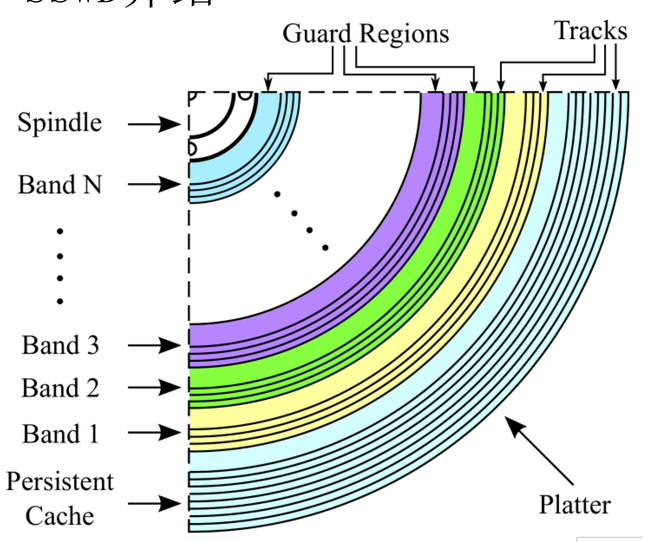
张强，董欢庆，刘振军，马留英，肖文健，李猛坤

中国科学院计算技术研究所 北京，中国科学院大学 北京 ，首都师范大学管理学院 北京

背景

随着大数据时代的到来，数据量的爆炸式增长使得数据中心的存储能力面临空前的考验，而作为数据中心主要存储介质的磁盘，其单碟片面密度已经接近1TB/in^{2}的理论极限，即单盘的存储容量很难再有大幅提升。因此数据中心要想大规模扩容就需要使用更多的磁盘，从而占用更多的物理空间和付出高额的能耗代价。在目前众多可提升磁盘存储密度的新技术当中，瓦记录(Shingled Magnetic Recording，SMR)技术因其对现有磁盘的物理构造改动极小便可大幅提升磁盘的存储空间，从而首先有产品面世，其中有代表性的是希捷的瓦记录磁盘(Seagate Shingled Write Disk，SSWD)。SMR采用部分叠加相邻磁道的方式增大了磁盘的面密度，但是这也带来了写覆盖（Write Overlay）问题，即更新某一磁道上的数据时会覆盖相邻磁道上的有效数据，导致SWD无法支持原地更新。 SMR的这一特性，使得在非顺序写场景下，数据不能直接被写入到磁盘中，需要配合其他技术手段来防止写入数据时破坏磁盘上已经存在的有效数据，这必然导致其非顺序写性能受到影响。本文通过Fio和Filebench测试工具，对SSWD做了比较全面的测试，并对各种应用场景下SSWD的性能做了详细的测试和分析，为SSWD在数据中心的应用提供了重要的参考。

SSWD介绍



SSWD将磁盘物理空间分成了三部分：持久缓存区、数据区和Map区。

持久缓存区位于SSWD的外圈，如左图所示，占用磁盘存储空间几十GB左右。持久缓存区以Log的方式管理，发往磁盘的所有非顺序写请求会先被顺序地写入磁盘的持久缓存区，待缓存区满了或者磁盘空闲一定时间后，SSWD会将持久缓存区内的数据合并然后移动到磁盘的数据区。

数据区为SSWD内数据最终存放的区域，区域按带顺序划分，带大小在17MiB~36MiB左右，相邻的带之间预留了一部分空间（Guard Regions），使得带间数据的写入不会互相覆盖。持久缓存区被写满后会以带为单位将数据移动到数据区,即下刷。下刷前会先将目标带内的旧数据读出，与持久缓存区内的新数据合并，最终将新的带内数据整体顺序写回带中，即RMW，这种以带为单位的RMW过程避免了更新数据区时可能产生的写覆盖问题，但是依然引入了额外的数据迁移开销。

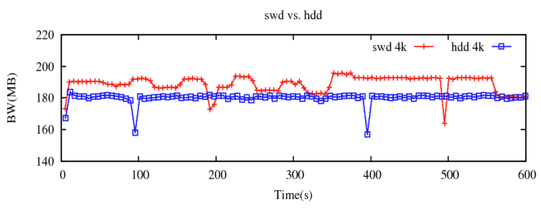
Map区位于盘片的中间，主要存放持久缓存的映射信息。由于持久缓存区采用Log方式写入数据，因此需要维护数据块在持久缓存区和带内的位置信息。并且为了尽可能避免掉电后持久缓存区写入的数据丢失，需要周期性的持久化映射信息到Map区域，而这一过程必然会导致磁头的抖动。

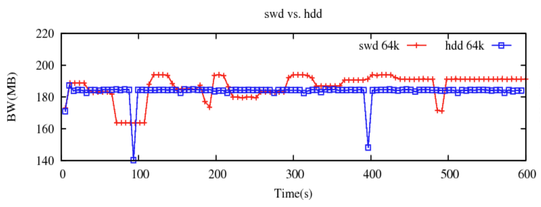
对于非顺序写，数据先被写入持久缓存区然后再下刷，这种设计原理必然导致在持久缓存未满之前SWD的非顺序写性能会很好，一旦启动下刷，性能会急剧下降。而对于顺序写，则无需写入持久缓存区，可直接写入数据区，并不会导致写覆盖，因此可保持较高的写性能。

从磁盘架构上来看，SSWD采用了将in-place update和out-place update结合的方式，但是依然没有很好的解决两种技术的缺点，从希捷将其定位在桌面应用和动态归档领域也能看出，SSWD在写压力较大的非顺序场景下的持续IO性能是较差的。

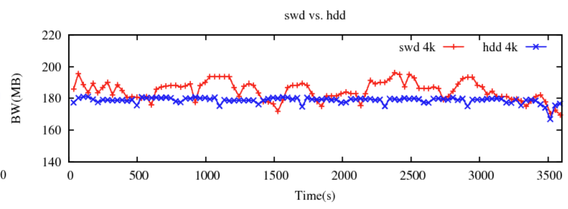
Fio压力测试

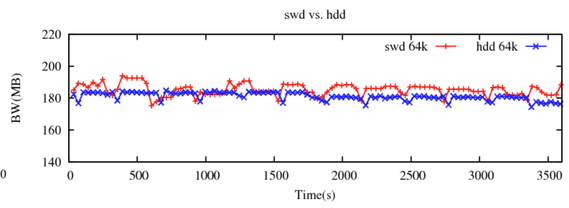
单盘顺序读测试



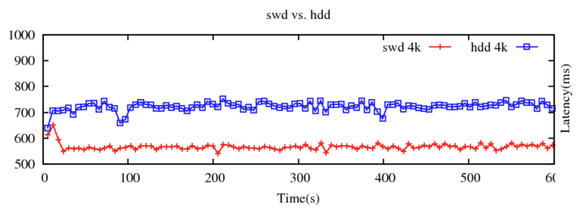


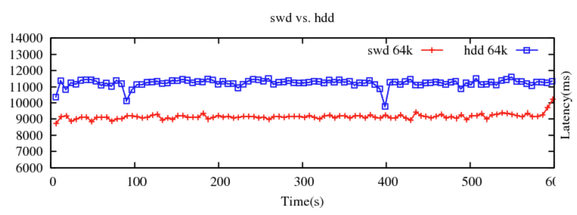
单盘顺序写测试



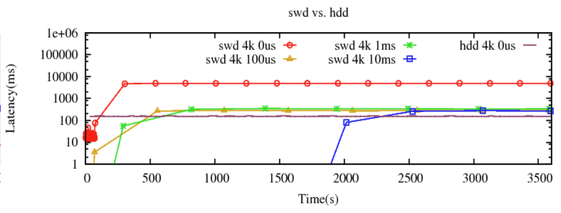


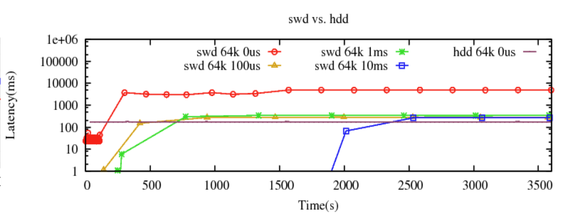
单盘随机读测试



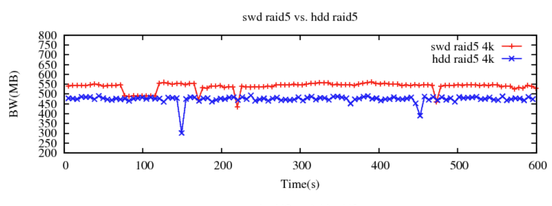


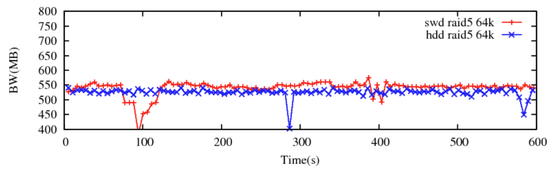
单盘随机写测试



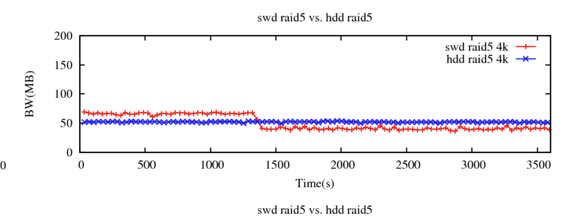


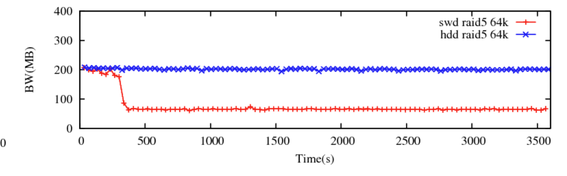
RAID5顺序读测试



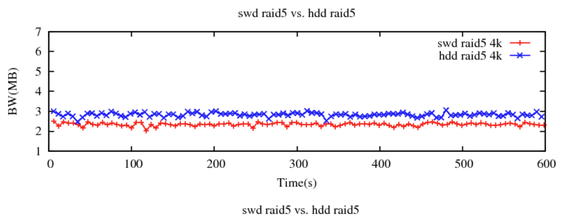


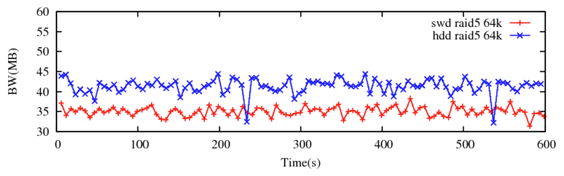
RAID5顺序写测试



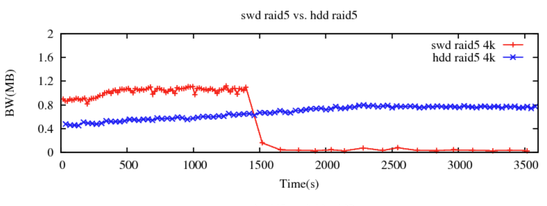


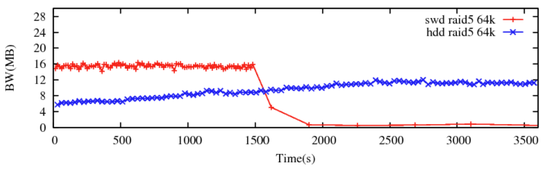
RAID5随机读测试





RAID5随机写测试





结论

通过压力测试可以看到，由于磁盘面密度增大了，希捷SWD在顺序读写方面性能较好，而随机读写方面由于需要检查请求是否命中持久缓存，引入了从Map区域读取地址映射表的开销，因此导致磁盘整体性能有所下降。对于SSWD RAID5，顺序读性能依然较好，但是在顺序写场景下，性能不佳，特别是由于SSWD RAID5会产生读改写或重构写，导致单盘上的IO顺序性降低，因此性能会大幅降低。从对盘阵的随机写测试可以看出，单盘回收带来了巨大的性能影响，因此如果想提升SSWD RAID5的非顺序写性能，需要对现有RAID5系统重新做设计，以尽量利用SSWD的磁盘特征，尽可能避免回收带来的性能影响。

希捷将其SSWD定位为适用于桌面应用和动态归档场景。从本文的一系列测试也可以验证这一点。对于顺序性较好的场景，例如动态归档、视频存储等，IO会绕过SSWD的持久缓存区，顺序的写入磁盘的带中，且不会产生写覆盖，因此可以最大化SSWD的写性能；对于桌面应用等IO密度较小的场景，即使IO可能随机性较强，SSWD依然可以将随机写转换为顺序写，并且磁盘有足够的时间做回收，保证了持久缓存区始终有较多的空间接收非顺序写，从而大幅降低应用的IO延迟。