计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20180264
55(9):2016-2028,2018

高密度磁记录技术研究综述

王国华[1,2] 杜宏章[2] 吴凤刚[2] 刘石勇[2,3]
[1](华南理工大学软件学院 广州 510006)
[2](美国明尼苏达大学计算机科学与工程系 明尼苏达州明尼阿波利斯 55455)
[3](中国海洋大学信息科学与工程学院 山东青岛 266100)
(ghwang@scut.edu.cn)

Survey on High Density Magnetic Recording Technology

Wang Guohua[1,2], David Hung-Chang Du[2], Wu Fenggang[2], and Liu Shiyong[2,3]
[1](*School of Software Engineering*, *South China University of Technology*, *Guangzhou* 510006)
[2](*Department of Computer Science and Engineering*, *University of Minnesota*, *Minneapolis*, *Minnesota* 55455)
[3](*College of Information Science and Engineering*, *Ocean University of China*, *Qingdao*, *Shandong* 266100)

Abstract In the era of big data, the demand for large-capacity disks has been growing. With minimal technology changes to the existing disk head and storage media of hard disks, the shingled magnetic recording (SMR) technology is the best choice to increase the disk storage capacity. The interlaced magnetic recording (IMR) technology is a newly developed technology in recent years, which can achieve higher storage density and random write performance than SMR. In this paper, we first introduce the shingled track layout of SMR drive and the resulting write amplification problem. We also review the data management methods that mitigate write amplification problem, the evaluation of performance characterizations, and the research on SMR-based upper applications. Then we introduce the interlaced track layout of IMR drive and its data write amplification problem. We also analyze the future research topics of IMR drive. Finally, we compare SMR drive and IMR drive from the storage density, random write performance, and other aspects. A variety of SMR-based upper applications, like file system, database, and RAID, prove that SMR drive can be effectively used to replace conventional disks to build large-scale storage systems. The advantages of IMR drive over SMR drive will make it have a bright future.

Key words high density magnetic recording; shingled magnetic recording (SMR); interlaced magnetic recording (IMR); storage density; write amplification

摘 要 大数据时代对大容量磁盘的需求日益增长,而在对现有的磁盘不进行较大改动的前提下,叠瓦式磁记录技术SMR是提高磁盘存储容量的最佳选择.近年来,兴起了一种新的磁记录技术——交错式磁记录技术IMR,它可以获得比SMR更高的存储密度和随机写性能.首先介绍了SMR磁盘的内部叠瓦式结构以及由此带来的数据写放大问题,并对缓解数据写放大问题的数据管理方式、性能特性评测以及基于SMR的上层应用系统方面的研究进展进行了概述;然后对新兴的IMR磁盘内部结构及其数据写放大问题进行了介绍,并对其将来的研究方向做了一定的分析和展望;最后对SMR磁盘和IMR磁盘在存储密度、数据写性能等方面进行了比较分析.当前有很多基于SMR磁盘的上层应用系统,这表

收稿日期:2018-04-06;**修回日期**:2018-06-15
基金项目:国家留学基金委青年骨干教师出国研修项目(201706155079)
This work was supported by the Young Scholars Study Abroad Program of China Scholarship Council (201706155079).

明 SMR 磁盘可以高效地替代传统磁盘来构建大型的存储系统,而 IMR 磁盘的优势也将使其未来的发展前景可期.

关键词 高密度磁记录技术;叠瓦式磁记录;交错式磁记录;存储密度;写放大

中图法分类号 TP391

云存储、移动计算、视频监控、社交媒体、大数据等的发展导致了数据量的爆炸式增长.根据国际数据公司(International Data Corporation, IDC)预测,2020 年全球数据的存储需求将会达到 44 ZB 的规模[1],并且还将继续呈指数级增长趋势.而低成本大容量的磁盘是当今以及未来信息存储的最基本需求.

几十年以来,磁盘的发展遵循其自身的摩尔定律——Kryder 定律[2],其存储容量按照每年 30%~50%的速度增长.但受超顺磁效应[3]的影响,现有磁盘的存储密度已达到 1Tb/in^2 的极限[4-5],仅仅依靠缩小每个记录位尺寸的方法已经无法适用.为了进一步提高磁盘的存储容量,存储厂商提出了各种新型的磁记录方法,如热辅助磁记录(heat-assisted magnetic recording, HAMR)[6-7]、微波辅助磁记录(microwave-assisted magnetic recording, WAMR)[8-9]以及比特模式磁记录(bit-patterned magnetic recording, BPMR)[10-11]等.HAMR 和 WAMR 需要在读写头上增加能量产生及消除装置,在写数据时需要产生额外的能量来降低磁记录介质的矫顽力,在介质磁化后需要迅速消除这些能量以防止磁化信息挥发,所以 HAMR 和 WAMR 会增加额外的技术和工艺成本.BPMR 虽然不需要在磁头上增加新的装置,但是其记录介质制作成本较高,寻道难度大.这些技术都需要对现有的磁头或者磁记录介质等进行改变,需要更高水平的工艺,厂家也需要投入巨大的研发成本,其在商业化之前还有一些重大的挑战需要克服.

而叠瓦式磁记录(shingled magnetic recording, SMR)[12-13]是一种对现有磁盘的伺服系统、读写头以及存储介质进行较小改动就能够大幅提高存储密度的技术.相应地,传统的磁记录技术我们通常称之为普通磁记录技术(conventional magnetic recording, CMR).叠瓦式磁记录技术早在多年前就被提出,但是获得关注却是始于 2008 年美国日立环球存储科技公司的 Roger Wood 在磁记录技术大会的演讲[14-15],直到 2014 年 8 月希捷公司推出全球首款 8 TB 的 SMR 磁盘 Archive HDD[16]之后,SMR 磁盘才开始进入商业化阶段.

最近,一种新兴的被称为交错式磁记录(interlaced magnetic recording, IMR)[17-18]的磁盘存储技术被提出,它可以获得比 SMR 更高的存储密度和随机写性能.目前,IMR 磁盘还没有上市,工业界和学术界还没有对其展开广泛的研究.本文将在总结 SMR 技术研究现状的基础上,提出未来对 IMR 技术的研究方向.

1 叠瓦式磁记录技术

叠瓦式磁记录技术是将相邻磁道像屋顶的瓦片一样进行部分重叠,消除了磁道和磁道之间的间隙,因而增加了每个盘面所能容纳的磁道数,使得磁记录密度大幅度增加,可以达到 2~3 Tb/in^2[19],这样,在不改变磁盘物理尺寸的情况下可以大幅度提升磁盘的存储容量.

1.1 SMR 磁盘的基本结构

传统的磁盘有 2 个磁头:一个用于写数据;一个用于读数据.通常写数据比读数据需要更强的磁场,从而写磁头比读磁头要大,这样磁道的宽度就需要与写磁头的宽度一致,如图 1 所示.此外,磁道和磁道之间需要存在一定的安全间隙(guard space),使得数据的写入不会影响或干扰到相邻磁道中的数据.

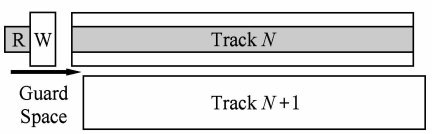


Fig. 1 Magnetic recording track layout for CMR

图 1 传统磁盘的磁道布局

而叠瓦式磁记录技术利用了读磁头的宽度可以小于写磁头的宽度这一特性来将磁道重叠起来,其中每个磁道未被其他磁道覆盖的部分设计成读磁头的宽度,而整个磁道的宽度则设计成写磁头的宽度[20],如图 2 所示.这样,对 SMR 的读操作将与传统磁盘一样,且其顺序写性能较好,适合应用于数据的归档和备份.

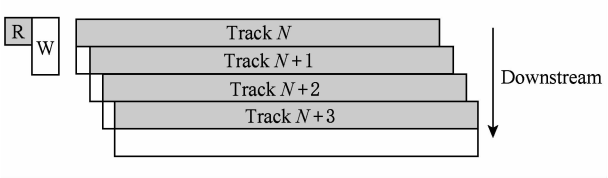


Fig. 2 Magnetic recording track layout for SMR

图 2 叠瓦式磁盘的磁道布局

1.2 SMR 磁盘的数据写放大问题

磁道重叠虽然可以解决磁盘容量增长缓慢的问题,但它却是以牺牲随机写数据的能力为代价的.

如图 2 所示,只有最后一个磁道上的数据可以被任意修改,如果要修改其他磁道上的数据,则需要对其所有下游磁道(downstream)上的数据进行重写,这是因为写磁头需要跨越多个磁道,因而会对下游叠加磁道上的数据造成破坏. 为了避免下游磁道上的数据被覆盖,则需要将下游所有磁道上的数据先读到内存,待修改操作完成之后再将下游磁道上的数据写回,因此一次修改操作引入了多次额外的读、写操作,这就是数据写放大问题,而数据写的效率与下游需要读出和重写的磁道数有关.

为了尽可能地减少数据写放大问题,通常将 SMR 磁盘的磁道划分为几个磁道带(band)[21],每个磁道带包含多个连续的、相互重叠的磁道,而磁道带与磁道带之间保持一定的安全间隙,安全间隙的宽度只需要刚好保证一个磁道带中最后一个磁道的写操作不会干扰或破坏下一个磁道带中第 1 个磁道上的数据即可. 这样,就可以将每个磁道带的数据写放大问题限制在本磁道带内,而不会影响到其他磁道带的数据.

1.3 SMR 磁盘的管理方式

由于 SMR 磁盘不同于传统磁盘的特殊结构,为对其进行合理地控制与管理并更好地解决其数据写放大问题,目前业界提出了 3 种类型的 SMR:磁盘管理式 SMR(drive managed SMR, DM-SMR)、主机管理式 SMR(host managed SMR, HM-SMR)以及主机感知式 SMR(host aware SMR, HA-SMR)[22-23].

DM-SMR 是为了满足 SMR 磁盘与现有的系统相互兼容而提出的一种管理方式,它将 SMR 的管理放到磁盘内部,通过引入一个类似于 SSD 闪存转换层(flash translation layer, FTL)[24]的软件层[25-29],来向主机系统隐藏 SMR 磁盘的内部结构以及数据访问限制等,该软件层被称为叠瓦转换层(shingled translation layer, STL)[30-31].

DM-SMR 磁盘中通常有一个持久缓存(persistent cache,也称为 media cache),STL 将来自主机的随机写操作缓存在持久缓存中,随后在持久缓存满了或者系统空闲时再通过清除(cleaning)操作将数据顺序写回到相应位置. 这就需要维护一个将逻辑块地址(logical block address, LBA)转换为物理块地址(physical block address, PBA)的映射表,向主机文件系统、数据库等提供一个线性的 LBA 地址空间及标准的块接口,从而使得 DM-SMR 磁盘可以像传统的非 SMR 磁盘一样与主机系统交互,而无需修改上层应用. 但是,随着数据量的变大,STL 上的映射方式以及数据清除策略使得 STL 变得越来越复杂,严重影响了磁盘的整体性能.

国际标准化组织(InterNational Committee on Information Technology Standards, INCITS) T10 和 T13 通过制定 ZBC[32] 和 ZAC 标准[33],将 SMR 磁盘的物理磁道带抽象化为逻辑的磁道区(zone),每个磁道区表示一段连续的非重叠的逻辑块地址空间. 本文中会交叉使用磁道带和磁道区,非特殊情况下它们代表相同的含义.

通过 ZBC 或 ZAC 所提供的 API,主机可以以磁道区为单位来对 SMR 磁盘进行管理,以便更好地保持 SMR 磁盘的性能可预测性.

其中,HM-SMR 便将 SMR 的管理全部放到主机端,通过 ZBC 或 ZAC 提供的 API 向主机提供 SMR 磁盘内部的数据布局信息,比如磁道区的数量及特性、LBA 的起始位置、写指针等信息,然后由主机完成 LBA 到 PBA 的转换. 但它在磁盘上没有保留持久缓存,因而 HM-SMR 不接受非顺序写操作,其磁道区通常被称为顺序写磁道区(sequential-write-required zones)[32].

而 HA-SMR 则结合了 DM-SMR 和 HM-SMR 两者的功能,它将对 SMR 磁盘的管理一部分放在磁盘端,另一部分放在主机端,这样它就可以像 DM-SMR 那样通过内置的 STL 来处理非顺序写操作,同时它还可以像 HM-SMR 那样向主机提供其内部信息,以便于主机更好地组织 I/O 请求. 为完成非顺序写操作的处理,HA-SMR 磁盘中需要像 DM-SMR 那样保留一个持久缓存,而其剩余部分同样被划分为多个磁道区,这些磁道区被称为优先顺序写磁道区(sequential-write-preferred zones)[32].

顺序写磁道区和优先顺序写磁道区都有各自的写指针(write pointer, WP),用来指明下一个顺序写操作所要写入的位置,但 HA-SMR 磁盘可以接收

顺序写操作,也可以处理非顺序写操作.对于顺序写操作,HA-SMR 直接将其追加到相应磁道区中写指针所指示的位置,如图 3 中的 d_2 所示,并将写指针移动到新追加的数据之后.但是对于非顺序写操作,HA-SMR 则将其暂存在持久缓存中,如图 3 中的 d_0 和 d_1 所示,同时将对应的磁道区 Z_0 和 Z_1 置为"非顺序"状态,并将其写指针置为无效.此后对处于"非顺序"状态的磁道区的所有写操作都将被缓存到持久缓存中.当持久缓存满了或者系统空闲时,再通过清除操作将所有缓存的数据写入到其目标位置.

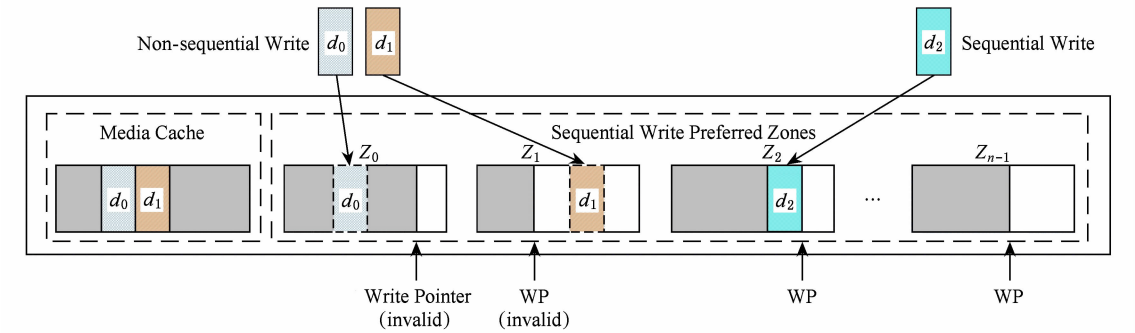


Fig. 3 Write requests of HA-SMR

图 3 HA-SMR 的写操作

这 3 种类型的 SMR 各有优势.DM-SMR 无需主机软件的更改即可与现有系统兼容;HM-SMR 类似磁带存储器,可以适应顺序的 I/O 负载,同时没有任何效能缺失;HA-SMR 可以与传统的 I/O 栈兼容,并且可以使用新的、SMR 所特有的 API 来实现主机 I/O 栈的优化.

表 1 对这 3 种类型 SMR 的异同和优缺点进行了总结.

Table 1 Three Management Techniques for SMR

表 1 SMR 磁盘的 3 种管理方式比较

SMR Types	ZBC/ZAC Commands	Media Cache	Non-Sequential Write	Advantages	Disadvantages
DM-SMR (Autonomous)	No	Yes	Allowed	Backwards compatible with all existing host systems	Drive handles all data requests internally, which creates unpredictable performance degradation under some conditions.
HM-SMR (Restrictive)	Yes	No	Disallowed	High and predictable performance	Not backwards compatible with all existing host systems and changes are required to hardware, filesystems, applications and operating systems.
HA-SMR (Cooperatively-Managed)	Yes	Yes	Allowed	Backwards compatible with host systems, best performance	Low and unpredictable performance when host issues sub-optimal commands.

1.4 SMR 磁盘的研究现状

1.4.1 数据写放大问题的研究

虽然将 SMR 磁盘划分为磁道带的方式可以将数据写放大问题限制在本磁道带内,但即使如此,数据写放大问题的开销依然很大.例如假设一个磁道带的大小为 64 MB,如果要修改本磁道带内最开始的 4 KB 数据,则最终需要写的数据为 64 MB,为应写数据 4 KB 的 16 384 倍,而如果采用较大的磁道带,其写放大问题将会更加严重[34].

为了在现有的存储系统中使用 SMR 磁盘,并且不会带来明显的性能损失,早期对 SMR 的研究主要是针对数据写放大问题的.

SMR 磁盘的数据写放大问题主要是由随机写操作或数据更新操作引起的,而对 SMR 磁盘的随机写操作或更新操作主要有 2 种处理方式:1)就地更新(in-place update);2)异地更新(out-of-place update)[20].在一些 SMR 数据管理的设计中,SMR 磁盘会保留一小部分的磁盘空间(1%~3%)来保存元数据,为了元数据访问的高效性,这部分空间采用 CMR 的方式,磁道之间没有重叠,可以随机访问,因而称为随机访问区域[28,35-37],而剩余部分则称为叠瓦式访问区域.

采用异地更新方式的 SMR 磁盘是通过将更新数据追加到新的位置并且将原数据置为无效的方式来缓解写放大问题的.它将叠瓦式访问区域划分为一个 E 区域和多个 I 区域[27-28].E 区域用来缓存或重组数据,I 区域则用来永久地保存数据.所有的写数据都先被缓存到 E 区域中,随后在需要时再将其写回到相应的 I 区域.对 E 区域和 I 区域的写操作都必须是顺序操作.E 区域和 I 区域都需要执行垃圾回收(garbage collection, GC)操作来回收无效的数据块或者整理磁盘碎片.此外,还需要维护一个 LBA-PBA 的映射表来记录数据的移动和访问,当前有很多关于减少 GC 操作和维护映射表开销的研究[26,29,38-39].

而就地更新方式则是将待修改数据所在的整个磁道带读入内存,在内存中将该磁道带中的数据修改后再将整个磁道带写回到它原本的位置,将其原有的数据覆盖,因此就地更新方式无需 GC 操作,也无需维护映射表.但这样一次数据更新操作会引起该磁道带内后续多个磁道的读写操作,如果磁道带较大,写放大问题就不会得到很好地缓解,因而可以采用较小的磁道带来进一步减小一次写操作所波及的范围,可由此却造成磁道带间隙所占空间的浪费过大.

文献[28]提出了一种新的 LBA 到 PBA 的静态映射机制.通常情况下,一个磁道带内的磁道是按磁道重叠的顺序依次进行分配的,而该静态映射机制则是对一个磁道带内的磁道分配顺序进行重新编排,以此来缓解数据写放大问题.该文作者假设磁道的重叠度为 2,且一个磁道带内有 4 个磁道 1,2,3,4,如果按照 4,1,2,3 的顺序分配磁道,总体性能将会有很大的提高.也就是说,将前 25%的 LBA 映射到第 4 磁道上,然后将第 25%~50%部分的 LBA 映射到第 1 磁道上、50%~75%部分映射到第 2 磁道上、75%~100%部分映射到第 3 磁道上.这样当该磁道带的空间使用率小于 50%时,只有第 1 磁道和第 4 磁道上有数据,由于这 2 个磁道没有重叠,所以不会有写放大问题;当该磁道带的空间使用率为 50%~75%时,数据存放在第 1、2、4 磁道上,由于第 1 磁道和第 2 磁道有重叠,所以第 1 磁道上的数据更新操作会导致一次额外的读和一次额外的写操作,而第 2 和第 4 磁道上的数据更新操作不会引起写放大问题;当该磁道带的空间使用率为 75%~100%时,所有磁道上都存有数据,此时每一磁道的数据更新操作都会引起写放大问题,其性能将会明显下降,与按 1,2,3,4 的顺序分配磁道时的开销差不多.

SMaRT[29]则提出了一种就地更新和异地更新相结合的混合性更新策略,主要采用异地更新的方式,当有数据修改操作时,将该数据写到缓冲区中并且将原数据置为无效.被置为无效的磁道不能马上变为空闲的可分配磁道,SMaRT 将无效磁道暂时作为安全间隙,使得其前面一个磁道上的数据可以进行就地更新,直到该无效磁道后面的磁道变为空闲磁道后,该磁道才从无效状态变为空闲状态,此时无需激活 GC 操作来进行无效空间的回收.除了安全间隙之前的那一个磁道,其他磁道上的数据更新都采用异地更新的方式.此外,SMaRT 还根据每个磁道的最近更新情况将磁道划分为热磁道(更新频繁)和冷磁道(较少更新),并为热磁道分配一个额外的安全间隙磁道,从而使其可以进行就地更新,减少写放大问题所带来的开销.为了减少垃圾回收的开销,SMaRT 中的垃圾回收操作仅在磁盘空间碎片较严重的情况下进行,通常在磁盘空闲时进行或者采用请求式垃圾回收策略,即在每一个写请求到达后先检查磁盘的碎片化情况,必要时才进行垃圾回收,将有效磁道上的数据进行移动,将空闲的磁道合并到一起,形成较大的空闲空间.

文献[28]对磁道的重叠程度做了假设,其假设磁道的重叠度为 2,但是随着 SMR 技术的发展,磁道的重叠度远超过 2,这就使得文献中所提出的静态映射机制变得不可行;同样地,在 SMaRT[29]中,如果磁道的重叠度超过 2,则在安全间隙只有一个磁道的情况下,其前面一个磁道上的数据将不能进行就地更新.但是文献[28-29]的思想可以借鉴,在确定磁道重叠度的情况下,合理设计 LBA 和 PBA 的映射机制,以及确定 SMaRT 中安全间隙包含多少个磁道时其前面一个磁道上的数据才可以进行就地更新.

1.4.2 SMR 磁盘特性的评测研究

早期的 SMR 研究主要从数据管理方面对随机写放大问题进行研究,而鲜少有对 SMR 磁盘内部的结构信息、性能参数等进行研究的,但是,如果上层应用能了解这些信息,则可以更好地使用 SMR 磁盘,尽可能地发挥其优势.

Skylight[40-41]通过软硬件相结合的方式提供了一套解密 DM-SMR 磁盘内部结构信息的方法.软件部分通过测量软件 I/O 操作的延迟来推断 DM-SMR 磁盘的一些特性,如持久缓存的类型、结构、大

小和位置,清除算法的类型,映射方法以及磁道带的大小等.而硬件部分则通过一个高速的照相机来跟踪读写头的运动轨迹,以此来确认软件部分所分析出的磁盘性能,同时也解决了仅仅通过软件延时测量来推测磁盘性能的不确定性.Skylight 所分析出的磁盘特性结果,可以为后续构建基于 SMR 磁盘的上层应用提供参考.

DM-SMR 磁盘相对于主机来说是封闭的,所以需要 Skylight 这样的评测系统来解密磁盘的内部信息,而对于 HA-SMR 磁盘来说,通过 ZBC 或 ZAC 提供的 API 即可获取磁盘的相应参数,但是如果能对影响其性能的其他因素做出评测,则对构建基于 HA-SMR 磁盘的存储系统有很大的参考价值.文献[42-43]就对影响 HA-SMR 磁盘性能的一些其他独有特性进行了评测,比如可同时打开的磁道区的最大数量、处于"非顺序"状态的磁道区的最大数量以及影响持久缓存清除效率的相关因素等,评测结果表明这些参数最大值的设置对系统性能影响的合理性,若违背这些规定,系统性能就会严重下降.针对可同时打开的磁道区最大数量这一参数的限制,作者还提出了一个主机控制的间接缓冲区(host-controlled indirection buffer,H-Buffer),H-Buffer 类似磁盘中的持久缓存,当某些上层应用要求同时进行顺序写操作的磁道区超过这个最大值时,可通过 H-Buffer 来保存一些数据,之后再被迁移到其最终位置,并且 H-Buffer 是由主机控制的,这就可以利用主机中较大的内存空间和较强的处理能力来支持更高效的数据迁移算法,从而可以提升 I/O 性能,使得 HA-SMR 可以应用到大型的存储系统中.但 H-Buffer 如何与持久缓存结合起来以发挥更大的作用,还有待进一步的研究.

1.4.3 数据清除策略的研究

DM-SMR 和 HA-SMR 磁盘通过持久缓存来暂存非顺序写操作,随后在适当的时候需要通过清除操作来将所缓存的数据写回到其最终位置.持久缓存的清除算法一般是按照 FIFO 的顺序从持久缓存中读出最老的数据块,以及持久缓存中与这个最老的数据块同属一个磁道带的其他数据块,同时也将相对应的目的磁道带中已存储的数据读出来,把这 3 部分数据合并成连续的内容后再写回相应的磁道带中.然后再从持久缓存中选择下一个最老的数据块按照上述方法进行清除,直到所有的数据都被清除或者清除过程被中断[40-43].

这个读-合并-写(read-merge-write)操作的过程比较耗时,DM-SMR 的数据清除工作是由 STL 利用磁盘端有限的计算能力和资源来完成的,因而会严重影响系统的性能;HA-SMR 则可以借助主机强大的计算能力来完成数据清除工作.当前主要有 2 种类型的清除算法:1)积极清除(aggressive cleaning)算法;2)懒惰清除(lazy cleaning)算法[40-41].

积极清除方式是在系统空闲时进行清除操作,因而也被称为空闲清除(idle cleaning)[42-43],当有其他的 I/O 请求时,清除过程将被终止.Skylight 和文献[42-43]所评测的 DM-SMR 和 HA-SMR 采用的都是空闲清除方式,按照 FIFO 的顺序依次清除.

懒惰清除方式是在持久缓存或映射表的空闲空间较低时才进行清除操作,一旦开始,将一直持续到持久缓存被清空或者映射表有足够的空间时才结束,这段时间较长,系统无法响应外界的 I/O 请求,因而也被称为阻塞清除(blocking cleaning)[42-43].

当前业界对持久缓存清除策略的研究较少,但根据文献[43],数据清除策略跟具体的工作负载(workload)有关.例如,有一个随机写密集且其随机写操作涉及多个磁道区的工作负载,磁盘持久缓存的空间会随着该工作负载的执行很快被耗尽,因而将触发阻塞清除操作,但该清除操作的持续时间较长,且这段时间内的系统吞吐率较低(从 100 MB/s 降到 0.1 MB/s),后续 I/O 请求的响应延迟较大,严重影响了系统的读写性能.

而空闲清除方式将会比阻塞清除方式更有优势,它可以充分利用系统的空闲时间来进行数据清除操作,对后续的 I/O 请求没有影响.但是如果一个工作负载中 I/O 请求之间的空闲时间较短且 I/O 请求较密集,空闲清除方式的优势将无法体现.比如,如果工作负载中 I/O 请求之间的时间间隔小于执行空闲清除方式所需的最小空闲时间,空闲清除算法就根本不会被触发,即使被触发,后续的 I/O 请求也会中断数据清除过程,使得持久缓存中不会有太多的数据被清除.但是我们可以通过对工作负载进行处理,比如将一些 I/O 请求延迟执行来人为地制造出较长的空闲时间段,以便有充分的时间完成足够多的数据清除,而那些被延迟的 I/O 请求将在数据清除操作结束之后再执行,其响应时间被拉长.

而文献[42]还指出,非顺序写数据发送给磁盘的顺序将会影响数据清除的效率.因为传统的 I/O 栈无法感知磁道区的边界,而工作负载中的非顺序写请求可能会在不同的磁道区中跳来跳去,如果主

机能根据这些写请求的目标磁道区将其分组,并在累积一段时间后再将这些请求发送给磁盘,那么持久缓存的清除效率将会大大提高,工作负载的完成时间也将被缩短.但是同样地,这些 I/O 请求会被延迟写入,因而响应时间较长.

文献[44]通过实验发现了由于 HA-SMR 数据清除过程中的读-合并-写操作所引起的长延迟问题,提出了一种虚拟持久缓存(virtual persistent cache)方法.该方法是在优先顺序写磁道区中划分出一小部分空间来作为虚拟缓存,将非顺序写操作中经常更新的数据和不经常更新的数据进行分离,不经常更新的数据保存在虚拟缓存中,而只有经常更新的数据才暂存在持久缓存中,从而可以减轻持久缓存的负担,进而降低持久缓存的清除开销.

1.4.4 SMR 磁盘的上层应用研究

随着 SMR 磁盘的商品化,很多研究者开始使用 SMR 磁盘构建应用系统,以证明 SMR 磁盘可以有效地替代传统的 CMR 磁盘.

SMRfs[45]与 HiSMRfs[46]都是运行在 HM-SMR 磁盘上的文件系统,它们无需在 SMR 磁盘内部实现重映射层来重定向读写请求即可管理 SMR 磁盘和支持随机写操作,可以完成数据存储空间的分配、垃圾回收、读写请求的调度、磁道带布局信息的获取等.由于元数据的修改是随机且比较频繁的,所以它们都将元数据和文件数据分开存储和管理,将元数据存放在非叠瓦式磁盘上.但它们最大的区别就是 HiSMRfs 采用树型结构来管理其元数据,并通过 Hash 表来加快元数据的查找,从而可以获得比 SMRfs 更高的性能.此外,HiSMRfs 还在文件系统层面实现了一个 SMR 磁盘阵列的 RAID 模块,提供了良好的容错性能.

文献[47]中也实现了 2 种 HM-SMR 的数据管理模式:1)严格追加模式,它将磁盘分成固定大小的磁道带,每个磁道带都有一个写指针,用于将新数据按顺序追加在尾部,类似日志结构文件系统(log-structured file system, LFS)[48];2)为基于 HM-SMR 的 Caveat-Scriptor[22]所实现的数据管理模式,它定义了 SMR 磁盘的 2 个专有参数:磁盘隔离间距(drive isolation distance, DID)和磁盘前缀隔离间距(drive prefix isolation distance, DPID),主机通过这 2 个参数可以动态地确定磁道带的边界,而不是采用事先划分磁道带的方法.这样,在为要写入的数据(多个连续的逻辑块)分配磁盘空间时,通过在该数据前后分别插入 DID 和 DPID 个逻辑块来对该数据加以保护,当该数据区域随着数据的追加或删除而扩张或收缩时,其前后保护空间的位置和状态也动态地调整,这是一种磁道带大小动态可变的方法,可以充分利用磁盘空间,但参数 DID 和 DPID 跟具体的磁盘结构有关.

SMR 磁盘相对于传统 CMR 磁盘来说存在随机写放大问题,但是其读性能及顺序写性能跟 CMR 磁盘一样,SMRDB[49]就是这样一个充分利用 SMR 磁盘顺序写的高效性的键值对(key-value)数据库引擎,它不依赖于文件系统,可以直接运行在 HM-SMR 磁盘上,对下层的磁盘进行管理.SMRDB 不需要磁盘固件提供任何形式的数据管理,只需要将磁道划分成一个小的随机访问区域和一些固定大小的叠瓦式磁道带,随机访问区域用来存储叠瓦式磁道带的信息,而键值对及相关的元数据则存储在叠瓦式磁道带中.SMRDB 可以被用作单独的数据库引擎,现有的文件系统也可以使用它在 SMR 磁盘上存储固定大小的键值对.SMRDB 的实现表明 SMR 磁盘能够在大部分应用上高效地代替传统磁盘.

SMORE[50]是一个可靠、高效的冷数据对象存储系统,它将较大且较少改动的冷数据对象,如娱乐、医疗影像、监控等多媒体数据以及备份数据、虚拟机映像等,划分为多个条带,分布存储到由 HM-SMR 或 HA-SMR 磁盘所构成的阵列中.阵列中的每个 SMR 磁盘都必须采用日志结构的方式顺序写入数据.如果有数据被删除,其所释放的空间不能马上被分配,而是通过垃圾回收操作将其写入顺序磁道区之后才可以被继续分配.

2 交错式磁记录技术

交错式磁记录 IMR 技术是最近被提出的磁盘存储技术,并在 HAMR 系统中进行测试[51-52],它可以提供比 SMR 更高的存储密度,同时又能缓解 SMR 中随机写所带来的数据放大问题.

2.1 IMR 磁盘的结构及特性

与 SMR 磁盘中相邻磁道依次重叠的方式不同,IMR 磁盘中的磁道分为上下 2 层,如图 4 所示,上层磁道和下层磁道相互交错,下层磁道被 2 个相邻的上层磁道部分覆盖,通常下层磁道较宽,上层磁道较窄,这样,下层磁道未被覆盖的部分即为下层磁道的读磁头宽度.如果上层磁道与下层磁道宽度相同,则下层磁道可能会被 2 个相邻的上层磁道完全覆盖,或者下层磁道未被覆盖的部分小于读磁头的

宽度以至于无法完成下层磁道的读操作.而传统的垂直磁记录(perpendicular magnetic recording, PMR)技术无法控制磁道的宽度,因此只能用激光或微波的方式来控制上下层磁道宽度的不同,即在 HAMR 或 MAMR 上实现 IMR,下层较宽的磁道需要较高强度的激光或微波才能将数据写入,而上层较窄的磁道,写入数据的激光或微波的强度相对较低.

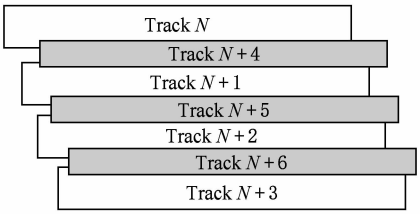


Fig. 4 Magnetic recording track layout for IMR

图 4 交错式磁盘的磁道布局

对 IMR 下层磁道的写操作会破坏其临近的上层磁道中的已有数据,因而,为了避免数据的丢失或者损坏,需要先将受影响的上层磁道中的数据读出,在将数据写入下层磁道后再将上层磁道中的数据重新写回,这就是类似于 SMR 中的数据写放大问题.而对上层磁道的写操作则不会引入额外的数据读和重写操作.这样,IMR 的随机写放大问题就比 SMR 要小的多.

由于下层磁道的写操作需要较高的激光功率,因而下层磁道的存储密度要比上层磁道高,从而下层磁道比其相邻的上层磁道具有更高的存储容量.此外,下层磁道也将比上层磁道具有较高的数据传输率,这是因为磁盘旋转一周,下层磁道可以读写的数据量比上层磁道多.

2.2 IMR 磁盘的数据分配方式

根据文献[18,53],对 IMR 磁盘的数据写操作可以按 2 阶段或者 3 阶段的方式进行磁道的分配.

1) 2 阶段分配方式

阶段 1 将所有的数据都写入下层磁道,如图 5

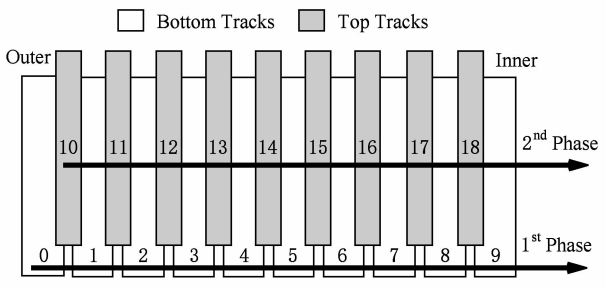


Fig. 5 Two-phase allocation method

图 5 2 阶段分配方式

中的磁道 0~9,直到所写入的数据量达到一定的容量,比如整个磁盘或者某个磁道带中所有的下层磁道都已分配完毕,则阶段 1 结束.在阶段 1 中,每一个下层磁道可以被随机写入或修改数据,而不会导致相邻磁道的数据读出和重新写入操作.

当整个磁盘或一个磁道带中的所有下层磁道都被写满时,阶段 2 的数据分配就开始了.在阶段 2 中,所有需要写入的数据都将被分配到上层磁道中,对上层磁道的写操作也不会引入额外的读写操作.按照文献[18,53],上层磁道的分配顺序跟下层磁道一样,都是从磁盘外圈到内圈进行分配.

但是在阶段 2 中,如果要对下层磁道的数据进行修改,则有可能需要首先将其上层磁道的数据进行转移,待下层磁道的写操作完成后再将上层磁道的数据重新写回,这就依赖于其上层的 2 个磁道是否已经存储了有效数据.例如图 5 中,对磁道 1 上的数据进行修改时,如果其上层磁道 10 和 11 已经被写入有效数据了,则需要先将磁道 10 和 11 中的数据读出,然后对磁道 1 进行修改操作,最后再将磁道 10 和 11 中的数据写回.这样,下层磁道的一次写操作将会引入额外的 2 次读和 2 次写操作.

2) 3 阶段分配方式

采用这种分配方式时,阶段 1 跟 2 阶段分配方式类似,都是先分配整个磁盘或某个磁道带中的下层磁道.

阶段 2 中上层磁道的分配则采用跳跃的方式,如图 6 所示,在分配了磁道 10 之后,将跳过磁道 15 继续分配磁道 11,依此类推.

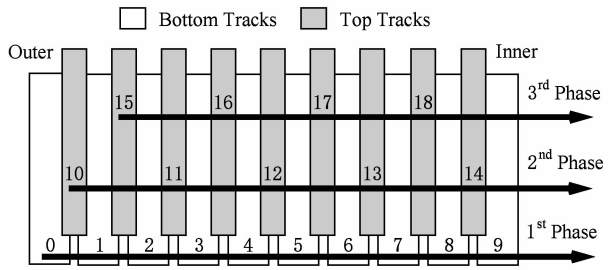


Fig. 6 Three-phase allocation method

图 6 3 阶段分配方式

阶段 3 的分配则将剩下的上层磁道按照从外圈到内圈的顺序依次进行分配.

3 阶段分配方式相对于 2 阶段分配方式来说,其主要的优势是当阶段 2 结束,也就是上层磁道空间使用了一半时,下层磁道的数据写操作只会引入额外的一次读和一次写操作.例如图 6 中,在阶段 2 分配结束时,如果要对磁道 4 或磁道 5 上的数据进

行修改,则只需要将上层的磁道 12 中的数据进行读出(此时磁道 16 和磁道 17 还没有被分配),待磁道 4 或磁道 5 修改结束后再将磁道 12 的数据写回.

2.3 IMR 磁盘的未来研究方向

由于 IMR 技术是最近兴起的,当前业界还没有针对 IMR 磁盘的相关问题展开深入的研究,但是可以借鉴 SMR 磁盘的相关研究方案对其进行研究.

1) 数据更新方式

最基本的数据更新方式是就地更新. 与 SMR 磁盘不同的是,IMR 磁盘中一个下层磁道的更新操作只会影响覆盖它的 2 个上层磁道,因而就地更新操作的开销要比 SMR 小的多.

在采用就地更新方式时,可以对文献[18,53]中的 2 阶段和 3 阶段分配方式进行改进,如图 7 和图 8 所示.

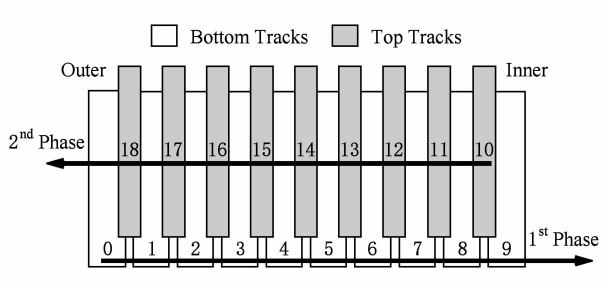


Fig. 7 Improved two-phase allocation method

图 7 改进的 2 阶段分配方式

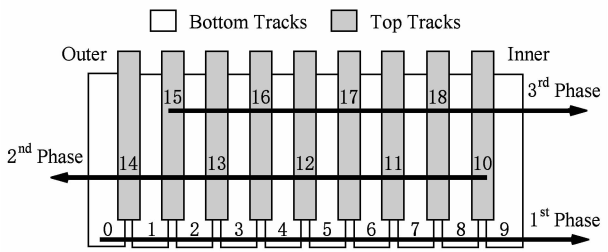


Fig. 8 Improved three-phase allocation method

图 8 改进的 3 阶段分配方式

图 7 中的 2 阶段分配方式将阶段 2 中上层磁道的分配按与下层磁道相反的顺序进行,即上层磁道的分配按照从内圈到外圈的顺序进行分配,这样,对于一个需要跨越下层磁道和上层磁道的数据流,其前半部分保存在磁盘内圈的下层磁道中,而后半部分则可以保存在磁盘内圈的上层磁道中,从而可以减少寻道时间,保持数据流的空间局部性.

同样地,在图 8 所示的 3 阶段分配方式中,其阶段 2 中上层磁道的分配可以采用与阶段 1 相反的方向,以保留数据的空间局部性,提高数据的访问效率.

IMR 磁盘中的磁道也可以像 SMR 磁盘一样,采用类似磁道带的分组方式,这样,在进行空间分配时,先将一个磁道带分配完毕,再分配下一个磁道带,从而可以将数据流的空间局部性保持在磁道带内.

另外,IMR 磁盘也可采取类似 SMR 磁盘的异地更新方式,对下层磁道的数据更新操作不在原来的位置上进行,而是将其写到上层磁道中的一个新位置上,然后将其原始位置上的数据置为无效. 但同样地,异地更新方式虽然可以避免数据写放大问题,但是却引入了回收下层磁道无效数据块以及频繁更新映射表的开销. 与 SMR 磁盘不同的是,采用就地更新方式对 IMR 磁盘上层磁道的更新操作不会影响任何磁道,而对下层磁道的更新操作也最多只会影响其上层的 2 个磁道,因而,需要权衡异地更新方式所引入的开销与它所能缓解的数据写放大问题.

2) 数据写缓冲区

不管是 2 阶段分配方式还是 3 阶段分配方式,都是先分配下层磁道,所以当需要对下层磁道上的已有数据进行更新时,便可以选择空闲的上层磁道作为临时缓冲区(类似 SMR 磁盘中的持久缓存),当缓冲区满了或者达到其他条件时再将缓冲区的内容写回到其位于下层磁道的最终位置. 但是选择哪部分上层磁道作为缓冲区将会对后续的数据清除性能有很大的影响. 如果缓冲区的位置离所缓存数据的原始位置较近,则数据清除时的开销就较小;如果缓冲区的位置离所缓存数据的原始位置较远,则数据清除时的开销就会很大. 此外,缓冲区的大小也直接影响到映射表的管理开销,数据清除策略的设计也对系统性能有着很大的影响.

DM-IMR[54]将磁道划分成多个磁道组(track group, TG),每个磁道组中的磁道分配方式采用图 8 所示的 3 阶段分配方式,其上层磁道缓存(Top-Buffer)方法将每个磁道组中最后几个未被分配的上层磁道作为缓冲区,并且该缓冲区可以缓存对同一下层磁道数据块的多次修改,在数据清除时一次性完成多个修改操作,减少了多次写操作所带来的开销. Top-Buffer 将该缓冲区的大小设置为磁道组大小的 2%,从而可以限制映射表的大小,使其能够保存到内存中,以加速对映射表的操作. Top-Buffer 方法在缓冲区满时采用顺序清除策略,按一次清除一个磁道的方式将该磁道中的数据块写入其目的磁道中,这样就可以回收连续的存储空间,从而避免缓冲区的碎片化.

作为数据更新操作的缓冲区,其大小可以保持不变,则该区域作为永久的缓冲区,不能用来存储有

效的用户数据.而 DM-IMR 中的 Top-Buffer 则随着磁道组空间使用率的增长而动态缩减,越来越小的 Top-Buffer 将会失去它的优势,会导致频繁的数据清除操作,因而系统性能会越来越低.

3) 冷热数据交换

IMR 上层磁道的写操作跟传统磁盘一样,不会影响任何磁道,根据文献[18],可以采用将下层磁道中的热数据与上层磁道中的冷数据进行交换的方法,将下层磁道中更新频繁的数据置换到上层磁道中,这样之后的数据更新操作将在上层磁道进行,不再需要额外的数据读和重写操作.

但是由于下层磁道跟上层磁道的容量不同,不能简单地将一个包含热数据的下层磁道与一个包含冷数据的上层磁道进行置换,所以数据交换单位的选择以及由此带来的管理开销有待进一步的研究.此外,相对于数据写缓存方法来说,数据交换操作的开销更大,因而冷热数据的跟踪确定,以及被置换后的数据其冷热变化对后续冷热数据的选择都将会影响系统的性能.

DM-IMR[54]在 Top-Buffer 缩减到一定程度时,开始启用块置换(Block-Swap)的方法以数据块为单位进行上下层磁道冷热数据的交换.2 种方法共存时,Block-Swap 可以在对 Top-Buffer 进行数据清除操作时通过映射表中的访问计数来确定它所缓存的下层磁道中的热数据块;当 Top-Buffer 退出时,其映射表为 Block-Swap 所用,可以通过在映射表中为每个数据块维护一个更新计数器来跟踪数据块的冷热情况.由于为一个磁道组中的所有数据块维护一个完整的映射表的开销太大,所以上层磁道中冷数据块的跟踪选择则采用了一种启发式的随机选择算法,将同一个磁道组中的上层磁道分布到多个计数(统计该磁道的更新情况)范围不相交的桶中,从计数值较低的桶中随机选择一个磁道上的数据块作为冷数据进行交换.

数据写缓存和冷热数据交换本身都可以作为缓解数据写放大问题的单独方法,而 DM-IMR 则将这 2 种方法结合起来,从数据写缓存方法过渡到冷热数据交换方法,但 2 种方法进行切换的最佳时机、映射表中的表项替换等问题还有待进一步的研究.

4) IMR 的应用研究

IMR 技术推出的主要目的也是能够替代传统磁盘甚至 SMR 磁盘,从而为存储系统提供大容量的存储设备.为此,IMR 在上层系统的应用研究同样对 IMR 磁盘的普及有着至关重要的作用,主要包括基于 IMR 的文件系统、数据库系统、RAID 系统等的研究.

3 SMR 与 IMR 的比较

SMR 和 IMR 都是为了增加磁盘密度的磁记录技术,IMR 技术与 SMR 有很多相似的特征,所以对 IMR 技术的研究内容和相关方案可以参考 SMR,如管理方式、数据更新方式、数据写放大问题解决方案、映射方式等.但 IMR 目前还没有产品上市,其与 SMR 的区别主要有 2 个方面:

1) 磁盘结构及数据写放大问题

SMR 磁盘中磁道的重叠方式是级联的,一个磁道的写操作将会破坏后续所有磁道上的数据;而 IMR 磁盘中上下 2 层磁道的重叠方式是交错式的,一个下层磁道的边缘仅被 2 个上层磁道所覆盖,下层磁道的写操作只会影响其上层的 2 个磁道,而上层磁道的写操作将不会影响任何磁道,因而 IMR 磁盘的写放大问题要比 SMR 磁盘小的多.

2) 存储密度

由于 SMR 和 IMR 两种磁盘的工艺与 CMR 磁盘不同,它们的磁道宽度、磁道存储密度也与 CMR 磁盘不同,再加上 SMR 磁盘中持久缓存和磁道带间隙的开销等,使得 SMR 和 IMR 两种磁盘的存储密度无法简单地进行计算.

而文献[51]对基于 HAMR 实现的 CMR,SMR 和 IMR 磁盘在存储密度方面进行了比较.由实验结果可知,CMR 磁盘的面密度(areal density capability)为 1.15Tb/in^2,SMR 和 IMR 磁盘的面密度分别为 1.46 和 1.50 Tb/in^2;SMR 磁盘的线密度(linear density)比 CMR 磁盘高 4.2%,而 IMR 磁盘的线密度比 CMR 磁盘高 11.9%,因而 IMR 磁盘可以获得比 SMR 磁盘稍高的存储密度.

4 总 结

SMR 和 IMR 都是通过磁道重叠来提高磁盘存储容量的技术,SMR 中的相邻磁道是以单边挤压的方式相互重叠,而 IMR 的上层磁道则是以双边挤压的方式与下层磁道重叠,虽然它们的磁道重叠方式不同,但它们都导致了数据写放大问题.本文通过对 SMR 技术在数据写放大问题的解决方案及其在上层文件系统或数据库中的应用方面进行总结分析,为 IMR 技术未来的研究方向提供参考.

参 考 文 献

[1] IDC. The digital universe of opportunities: Rich data and the increasing value of the Internet of things [EB/OL]. [2016-01-05]. http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

[2] Walter C. Kryder's law [J]. Scientific American, 2005, 293(2): 20-21

[3] Thompson D, Best J. The future of magnetic data storage technology [J]. IBM Journal of Research and Development, 2000, 44(3): 311-322

[4] Shen Xiao, Hernandez S, Victora R. Feasibility of recording 1 Tb/in^2 areal density [J]. IEEE Trans on Magnetics, 2008, 44(1):163-168

[5] Wood R. The feasibility of magnetic recording at 1 Terabit per square inch [J]. IEEE Trans on Magnetics, 2000, 36(1): 36-42

[6] Rottmayer R, Batra S, Buechel D, et al. Heat-assisted magnetic recording [J]. IEEE Trans on Magnetics, 2006, 42(10): 2417-2421

[7] Kryder M, Gage E, McDaniel T, et al. Heat assisted magnetic recording [J]. Proceeding of the IEEE: Advances in Magnetic Data Storage Technologies, 2008, 96(11): 1810-1835

[8] Zhu Jian-Gang, Zhu Xiaochun, Tang Yuhui. Microwave assisted magnetic recording [J]. IEEE Trans on Magnetics, 2008, 44(1): 125-131

[9] Zhu Jian-Gang, Wang Yiming. Microwave assisted magnetic recording utilizing perpendicular spin torque oscillator with switchable perpendicular electrodes [J]. IEEE Trans on Magnetics, 2010, 46(3): 751-757

[10] White R, Newt R, Pease R. Patterned media: A viable route to 50 Gbit/in^2 and up for magnetic recording? [J]. IEEE Trans on Magnetics, 1997, 33(1): 990-995

[11] Dobisz E, Bandic Z, Wu Tsai-Wei, et al. Patterned media: Nanofabrication challenges of future disk drives [J]. Proceedings of the IEEE: Advances in Magnetic Data Storage Technologies, 2008, 96(11): 1836-1846

[12] Wood R, Williams M, Kavcic A, et al. The feasibility of magnetic recording at 10 Terabits per square inch on conventional media [J]. IEEE Trans on Magnetics, 2009, 45(2): 917-923

[13] Tagawa I, Williams M. High density data-storage using shingle-write [C] //Proc of IEEE Int Magnetics Conf. Piscataway, NJ: IEEE, 2009

[14] Wood R, Williams M, Kavcic A, et al. The feasibility of magnetic recording at 10 Terabits per square inch on conventional media [C] //Proc of the 19th Magnetics Recording Conf. Piscataway, NJ: IEEE, 2008

[15] Takenoiri S, Matsuo S, Fujihira T. Magnetic recording media: Technical trends and future outlook [J]. Fuji Electric Review, 2011, 57(2): 32-36

[16] Seagate. Seagate ships world's first 8TB hard drives [EB/OL]. [2016-03-16]. https://www.seagate.com/about-seagate/news/Seagate-ships-worlds-first-8TB-hard-drives-pr-master/?paramChannelName=newsroom

[17] Hwang E, Park J, Rauschmayer R, et al. Interlaced magnetic recording [J]. IEEE Trans on Magnetics, 2017, 53(4): Article Sequence Number 3101407

[18] Gao Kaizhong, Zhu Wenzhong, Gage E. Interlaced magnetic recording: United States, US9728206B2 [P]. 2017-08-08

[19] Greaves S, Kanai Y, Muraoka H. Shingled recording for 2-3 Tbit/in^2 [J]. IEEE Trans on Magnetics, 2009, 45(10): 3823-3829

[20] Tan Yujuan, Liu Tao, Zhao Yajun. Review on shingled magnetic recording disk [J]. China Sciencepaper, 2016, 11(14): 1661-1667 (in Chinese)
(谭玉娟, 刘涛, 赵亚军. 叠瓦式磁记录磁盘的研究进展[J]. 中国科技论文, 2016, 11(14): 1661-1667)

[21] Kasiraj P, New R, De Souza J, et al. System and method for writing data to dedicated bands of a hard disk drive: United States, US7490212B2 [P]. 2009-02-10

[22] Feldman T, Gibson G. Shingled magnetic recording: Areal density increase requires new data management [J]. ;login:, 2013, 38(3): 22-30

[23] Alcorn P. SMR (shingled magnetic recording) 101[EB/OL]. [2017-11-22]. http://www.tomsitpro.com/articles/shingled-magnetic-recoding-smr-101-basics,2-933.html

[24] Gal E, Toledo S. Mapping structures for flash memories: Techniques and open problems [C] //Proc of IEEE Int Conf on Software—Science, Technology & Engineering (SwSTE'05). Piscataway, NJ: IEEE, 2005: 83-92

[25] Gibson G, Polte M. Directions for shingled-write and two dimensional magnetic recording system architectures: Synergies with solid-state disks, CMU-PDL-09-104 [R/OL]. Pittsburgh, PA: Parallel Data Laboratory, Carnegie Mellon University, 2009. [2017-12-05]. http://repository.cmu.edu/cgi/viewcontent.cgi?article=1004&context=pdl

[26] Cassuto Y, Sanvido M, Guyot C, et al. Indirection systems for shingled-recording disk drives [C] //Proc of the 26th IEEE Symp on Massive Storage Systems and Technologies. Piscataway, NJ: IEEE, 2010

[27] Hall D, Marcos J, Coker J. Data handling algorithms for autonomous shingled magnetic recording HDDs [J]. IEEE Trans on Magnetics, 2012, 48(5): 1777-1781

[28] He Weiping, Du H-C D. Novel address mappings for shingled write disks [C] //Proc of the 6th USENIX Workshop on Hot Topics in Storage and File Systems. Berkeley, CA: USENIX Association, 2014

[29] He Weiping, Du H-C D. SMaRT: An approach to shingled magnetic recording translation [C] //Proc of the 15th USENIX Conf on File and Storage Technologies. Berkeley, CA: USENIX Association, 2017: 121-134

[30] Gibson G, Ganger G. Principles of operation for shingled disk devices, CMU-PDL-11-107 [R/OL]. Pittsburgh, PA: Parallel Data Laboratory, Carnegie Mellon University, 2011. [2017-11-28]. http://www.pdl.cmu.edu/PDL-FTP/Storage/CMU-PDL-11-107.pdf

[31] Chen Xiang. Technologies research of the shingled-recording hard drivers data organization and its design and implementation [D]. Wuhan: Huazhong University of Science and Technology, 2011 (in Chinese)
(陈祥. 瓦记录磁盘驱动器的数据组织技术研究及其实现[D]. 武汉: 华中科技大学, 2011)

[32] INCITS T10 Technical Committee. Information technology-zoned block commands-2 (ZBC-2)[EB/OL]. [2017-12-15]. http://www.t10.org/drafts.htm

[33] INCITS T13 Technical Committee. Information technology-zoned device ATA command set-2 (ZAC-2)[EB/OL]. [2017-12-15]. http://www.t13.org/Documents/MinutesDefault.aspx?DocumentType=4&DocumentStage=2

[34] Suresh A, Gibson G, Ganger G. Shingled magnetic recording for big data applications, CMU-PDL-12-105 [R/OL]. Pittsburgh, PA: Parallel Data Laboratory, Carnegie Mellon University, 2012. [2018-01-22]. http://www.pdl.cmu.edu/PDL-FTP/FS/CMU-PDL-12-105.pdf

[35] Amer A, Long D, Miller E, et al. Design issues for a shingled write disk system [C] //Proc of the 26th IEEE Symp on Mass Storage Systems and Technologies. Piscataway, NJ: IEEE, 2010

[36] Amer A, Holliday J, Long D, et al. Data management and layout for shingled magnetic recording [J]. IEEE Trans on Magnetics, 2011, 47(10): 3691-3697

[37] Le Moal D, Bandic Z, Guyot C. Shingled file system host-side management of shingled magnetic recording disks [C] // Proc of 2012 IEEE Int Conf on Consumer Electronics. Piscataway, NJ: IEEE, 2012: 425-426

[38] Lin Chung-I, Park D, He Weiping, et al. H-SWD: Incorporating hot data identification into shingled write disks [C] //Proc of the 20th IEEE Int Symp on Modeling, Analysis and Simulation of Computer and Telecommunication Systems. Piscataway, NJ: IEEE, 2012: 321-330

[39] Jones S, Amer A, Miller E, et al. Classifying data to reduce long-term data movement in shingled write disks [J]. ACM Trans on Storage, 2016, 12(1): 2:1-2:17

[40] Aghayev A, Desnoyers P. Skylight—a window on shingled disk operation [C] //Proc of the 13th USENIX Conf on File and Storage Technologies. Berkeley, CA: USENIX Association, 2015: 135-149

[41] Aghayev A, Shafaei M, Desnoyers P. Skylight—A window on shingled disk operation [J]. ACM Trans on Storage, 2015, 11(4): 16:1-16:28

[42] Wu Fenggang, Fan Ziqi, Yang Ming-Chang, et al. Performance evaluation of host Aware shingled magnetic recording (HA-SMR) drives [J]. IEEE Trans on Computers, 2017, 66(11): 1932-1945

[43] Wu Fenggang, Yang Ming-Chang, Fan Ziqi, et al. Evaluating host aware SMR drives [C] //Proc of the 8th USENIX Workshop on Hot Topics in Storage and File Systems. Berkeley, CA: USENIX Association, 2016

[44] Yang Mingchang, Chang Yuanhao, Wu Fenggang, et al. Virtual persistent cache: Remedy the long latency behavior of host-aware shingled magnetic recording drives [C] //Proc of IEEE/ACM Int Conf on Computer-Aided Design. Piscataway, NJ: IEEE, 2017: 17-24

[45] CMU. SMR Wiki [EB/OL]. [2017-11-20]. https://wiki.pdl.cmu.edu/SMR/WebHome

[46] Jin Chao, Xi Weiya, Ching Zhiyong, et al. HiSMRfs: A high performance file system for shingled storage array [C] // Proc of the 30th Symp on Mass Storage Systems and Technologies. Piscataway, NJ: IEEE, 2014

[47] Kadekodi S, Pimpale S, Gibson G. Caveat-scriptor: Write anywhere shingled disks [C] //Proc of the 7th USENIX Workshop on Hot Topics in Storage and File Systems. Berkeley, CA: USENIX Association, 2015

[48] Rosenblum M, Ousterhout J. The design and implementation of a log-structured file system [J]. ACM Trans on Computer Systems, 1992, 10(1): 26-52

[49] Pitchumani R, Hughes J, Miller E. SMRDB: Key-value data store for shingled magnetic recording disks [C] //Proc of the 8th ACM Int Systems and Storage Conf. New York: ACM, 2015

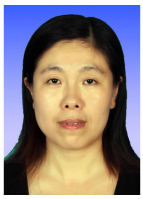
[50] Macko P, Ge X, Haskins J, et al. SMORE: A cold data object store for SMR drives [C] //Proc of the 33rd Int Conf on Massive Storage Systems and Technology. Piscataway, NJ: IEEE, 2017

[51] Granz S, Zhu Wenzhong, Seng E, et al. Heat-assisted interlaced magnetic recording [J]. IEEE Trans on Magnetics, 2018, 54(2): Article Sequence Number: 3100504

[52] Krichevsky A. Heat assisted magnetic recording with interlaced high-power heated and low-power heated tracks: United States, US9099103B1 [P]. 2015-08-04

[53] Gao Kaizhong, Zhu Wenzhong, Gage E. Write management for interlaced magnetic recording devices: United States, US9508362B2 [P]. 2016-11-29

[54] Wu Fenggang, Zhang Baoquan, Cao Zhichao, et al. Data management design for interlaced magnetic recording [C] // Proc of the 10th USENIX Workshop on Hot Topics in Storage and File Systems. Berkeley, CA: USENIX Association, 2018



Wang Guohua, born in 1977. PhD, lecturer. Her main research interests include storage systems, magnetic recording, data deduplication, and big data.

David Hung-Chang Du, born in 1951. PhD, professor, PhD supervisor. IEEE Fellow. His main research interests include storage systems, data center power management, high speed networks, distributed systems, intelligent storage, etc (du@umn.edu).

Wu Fenggang, born in 1987. PhD candidate. His main research interests include SMR drives performance characterization and application design, non-volatile memory based storage system, data deduplication, and key-value store (wuxx0835@umn.edu).

Liu Shiyong, born in 1986. PhD candidate. His main research interests include big data, intelligent storage, and distributed storage (lshyouc@163.com).

勘误启事

《计算机研究与发展》2018 年第 2 期发表的“CPU 和 DRAM 加速任务划分方法:大数据处理中 Hash Joins 的加速实例”(第 289～304 页)一文中,因作者疏漏出现 2 处错误,对由此给广大读者和《计算机研究与发展》编辑部带来的麻烦深表歉意.现更正如下:

1) 更正作者单位:前 3 位作者吴林阳、罗蓉、郭雪婷新增所属单位“中国科学院大学(University of Chinese Academy of Sciences, Beijing 100049)”.新增单位列为第一单位,中国科学院计算技术研究所应为第二单位.

2) 更正作者简介:前 3 位作者吴林阳、罗蓉、郭雪婷新增所属单位“University of Chinese Academy of Sciences”.

再次致歉!

作者:吴林阳 罗蓉 郭雪婷 郭崎