高密度磁盘性能优化技术研究综述

伍卫国,张 驰,于芳星,聂世强,李孟涵,牛 洁
(西安交通大学计算机科学与技术学院,陕西西安 710000)

摘 要: 在当今数字化时代,海量数据的生成和积累呈现出爆炸式增长的趋势,因此对存储容量的需求急速上升.传统磁记录磁盘CMR因其高容量和低成本而被视为解决海量数据存储的首选.然而,由于超顺磁效应的制约,CMR(Conventional Magnetic Recording)磁盘面密度的提升已触及极限.为了突破这一限制,叠瓦式磁记录技术SMR(Shingled Magnetic Recording)应运而生.基于传统硬盘架构,该技术以重叠磁道的方式,显著提升了磁盘面密度.但SMR磁盘在处理随机写时,会产生不可预测的写放大效应,从而严重影响I/O性能.为解决这一问题,业界随即提出了交错式磁记录技术IMR(Interlaced Magnetic Recording),利用优化的磁道布局和热辅助磁记录技术,有效实现了存储容量与性能的平衡.本文首先详细介绍了SMR和IMR的技术原理和磁盘类型,并量化分析了影响设备I/O性能的关键问题.然后,重点介绍了设备级优化方案,分析并总结了不同策略的优缺点与优化目标.接着,概述了面向设备的系统级和应用级设计方案,如文件系统、独立磁盘阵列技术和数据库等.最后讨论了在未来优化SMR磁盘和IMR磁盘性能可能的研究方向.

关键词: 超顺磁效应;叠瓦式磁记录;交错式磁记录;磁道布局;写干扰;写放大
基金项目: 国家自然科学基金(No.61972311,No.62202368)
中图分类号: TP333.3 **文献标识码:** A **文章编号:** 0372-2112(2024)05-1759-24
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20231129

A Survey on the Research of Novel High-Density Disk Performance Optimization Techniques

WU Wei-guo, ZHANG Chi, YU Fang-xing, NIE Shi-qiang, LI Meng-han, NIU Jie
(*Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi* 710000, *China*)

Abstract: In the digital age, the generation and accumulation of massive amounts of data are exploding, driving the demand for storage capacity sharply. Conventional magnetic recording disks (CMR) are considered the preferred solution for massive data storage due to their high capacity and low cost. However, the presence of the superparamagnetic effect (SPE) limits further improvements in CMR disk density. To overcome this limitation, shingled magnetic recording (SMR) technology was developed. Based on the conventional hard disc architecture, this technology dramatically increases disk areal density by overlapping tracks. However, SMR disks produce unpredictable write amplification effects when performing random writes, which can severely impact I/O performance. To solve this problem, the industry then proposed interlaced magnetic recording (IMR) technology, which uses an optimized track layout and heat-assisted magnetic recording technology to effectively balance storage capacity and performance. In this paper, we first introduce the technical principles, disk types, and applications of SMR and IMR in detail, and quantitatively analyzes the write interference problem that affects their I/O performance. It then focuses on solutions to optimize their I/O performance at the device level, analyzing and summarizing the advantages, disadvantages, and optimization objectives of different strategies in each type of solution. An overview of device-oriented system-level and application-level optimization solutions, such as file systems, redundant array of independent disk (RAID) technologies, and databases, is then provided. Finally, possible research directions for optimizing the performance of SMR and IMR disks in the future are discussed.

Key words: superparamagnetic effect; shingled magnetic recording; interlaced magnetic recording; track layout; write interference; write amplification

Foundation Item(s): National Natural Science Foundation of China (No.61972311, No.62202368)

收稿日期:2023-12-05;修回日期:2024-03-25;责任编辑:宋甲英

1 前言

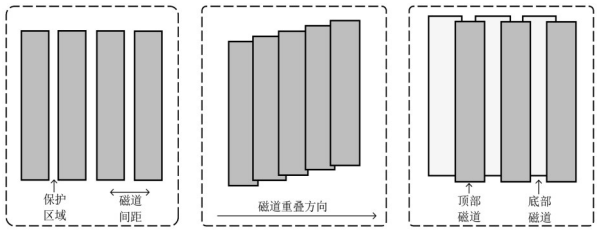
在当今时代，伴随着人工智能、航天航空、生物信息和能源勘探等领域的快速发展，边缘、核心及云端的数据呈现出爆炸式增长．根据国际数据公司IDC的2021年报告预测，未来五年，数据将以21.2%的复合年增长率上升，到2026年将达到超过221 000艾字节[1]．面对数据量飞速扩增，存储设备作为其关键载体，正在朝着高密度和高性能方向前进．目前，存储市场主要被闪存固态盘(Solid State Drive，SSD)和机械硬盘(Hard Disk Drive，HDD)所主导．其中，SSD的出货量超过了15%，而HDD出货量更是高达60%[2]．海量数据量对存储设备的容量可扩展性和成本效益提出了更高的要求，但凭借容量与价格的优势，机械硬盘目前仍然是大数据存储的主要设备，其霸主地位还将维持一段相当长时间[3]．

过去的数十年中，得益于记录技术、制造工艺和介质材料的进步，机械硬盘的存储容量增长了近六个数量级．然而，超顺磁效应的存在使得垂直磁记录技术(Perpendicular Magnetic Recording，PMR)的面密度极限最高约为1.1 Tb/in^2[4]．为突破这一限制，业界探索了各种新型记录技术来进一步增加面密度，如改变磁性层结构的比特图案化磁记录[5](Bit-Patterned Magnetic Recording，BPMR)、减少存储位尺寸的微波辅助磁记录[6](Microwave-Assisted Magnetic Recording，MAMR)以及利用激光改变磁性状态的热辅助磁记录[7](Heat-Assisted Magnetic Recording，HAMR)等．虽然这些新技术提供了额外的容量增益，但也给结构设计和制造成本带来了巨大挑战．考虑到成本效益和市场需求，叠瓦式磁记录[8](Shingled Magnetic Recording，SMR)技术被视为传统磁记录(Conventional Magnetic Recording，CMR)的替代者．与其他技术不同，SMR无需大幅改变介质或磁盘的基础结构，而是依赖于非对称的读写磁场和更紧凑的磁道间距．为实现此目标，SMR采用了扩大的写磁头和重叠的磁道布局，从而显著提升了面密度．例如，在使用SMR技术的Ultrastar DC HC680磁盘中，其面密度达到了1.43 Tb/in^2[9]．如果将SMR技术与其他辅助技术相结合，理论上能够将面密度提升至约6~8 Tb/in^2[10,11]．自2014年希捷和日立公司发布首款商业化SMR磁盘，SMR磁盘已经被广泛部署在国内外的大型数据中心[12,13]．例如，美国的云存储服务商Dropbox率先将SMR磁盘部署到他们的EB级存储系统，并且SMR磁盘的规模已占据其总硬盘数量的90%[14]；而国内的阿里云，在其云对象存储服务改用SMR磁盘后，成功将总体成本降低了15%，且保持了与之前系统相当的性能[15]．近期，业界基于SMR技术提出了交错式磁记录[16](Interlaced Magnetic Recording，IMR)技术．与SMR技术相比，IMR技术对磁道布局进行了优化，有效缓解了磁道重叠布局所导致的随机写性能下降．结合HAMR技术，IMR的存储密度甚至超越了SMR，为进一步实现存储容量与性能的平衡开创了新机遇[17,18]．虽然目前IMR尚处于原型设计阶段，但一系列研究已开始讨论如何从软件层面优化其读写性能[19]．

2 叠瓦式与交错式磁盘背景介绍

2.1 SMR技术和IMR技术基本原理

提高面密度最直接的方式就是减小磁颗粒尺寸，从而增加单位区域内记录的数据量．然而，这种方式给介质材料、写入/读取技术以及存储稳定性带来了挑战．与需要对底层介质结构进行大幅改变的技术相比，SMR对制造工艺和介质材料的改动最小[20]．基于传统的垂直磁记录技术，SMR通过重叠磁盘盘面上的磁道来提高每英寸磁道数(Tracks Per Inch，TPI)，并结合减小的磁颗粒尺寸，有效增加了面密度．实际上，SMR技术可以将磁道部分重叠的根本原因在于磁盘读写数据时磁场强度的差异化，即读操作所需要的磁场强度弱于写操作[21]．为此，磁盘厂商设计了更窄的读磁头和更宽的写磁头来满足读写数据时所需的最低磁场强度要求．如图1(*b*)所示，在SMR磁盘中，磁道像瓦片一样依次部分重叠，通过减少磁道间距来增加同一盘面上的磁道数量．然而，这种磁道布局方式导致磁盘只能沿着叠瓦方向顺序写入数据，因为对指定磁道的写入会覆盖相邻K(K通常为4~8)个磁道的有效数据，从而造成数据丢失问题[22]．因此，为了限制这种影响，磁盘表面通常被划分为若干独立区域，并由预留的空磁道将各独立区域相互隔离．而对于任意位置的读请求，则不会产生上述问题．



(*a*) CMR磁道布局 (*b*) SMR磁道布局 (*c*) IMR磁道布局

图1 CMR、SMR以及IMR磁盘的磁道布局示意图

严格的顺序写入约束限制了SMR磁盘的应用场景．为此，业界在SMR技术的基础上提出了下一代高密度磁记录技术——IMR．IMR技术优化了磁道布局，并结合HAMR技术来克服因磁性颗粒尺寸减小所导致的矫顽力限制问题，从而提升了面密度．IMR技术不再仅从单侧重叠磁道，而是从磁道两侧以上下“交错”的顺序重叠，以减少磁道间距并放宽顺序写约束[23~25]．如

图1(*c*)所示,IMR技术从逻辑上将所有磁道划分为数量相等的顶部磁道和底部磁道,每个底部磁道只与两个相邻的顶部磁道重叠.为了保证底部磁道的数据可以被正常读取,顶部磁道略窄于底部磁道,这使得顶部磁道的容量相对较小.对于数据的写入,顶部磁道所需要的激光强度要低于底部磁道,保证了顶部磁道的写入不会破坏底部磁道数据.因此,相较于SMR技术,IMR技术放松了顺序写约束.对任意位置的数据读取,IMR技术同样不施加任何限制.

2.2　SMR磁盘和IMR磁盘的基本类型

受随机写操作的影响,SMR磁盘无法像传统磁盘那样直接集成到存储系统中.为了适应多样化的存储场景,业界相继提出了三种主流的SMR磁盘模型,赋予用户不同级别的设备控制能力,从而提高了设备的使用效率.

为了实现向后兼容,驱动管理型叠瓦式磁盘(Drive-Managed SMR,DM-SMR)利用嵌入固件的叠瓦式翻译层[26~28](Shingle Translation Layer,STL)来隐藏顺序写约束,从而为上层应用提供标准的块接口.如图2所示,在DM-SMR磁盘中,盘面上的各个独立区域被称为带(band),它们是由物理上连续的磁道组织,大小在15~40 MB之间[29].对于任何提交到SMR磁盘的读写请求,STL首先将"逻辑地址"转换为实际的"物理地址",再根据物理地址对目标band读写数据.此外,DM-SMR内部配置了一个由少量叠瓦式磁道组成的持久缓冲区.为了保证磁盘写性能的稳定性,写入的数据被首先定向至持久缓冲区,随后由STL在设备空闲期间其批量迁移回本机存储区域.尽管DM-SMR可以像传统磁盘一样在存储系统中即插即用,但是在随机写密集型工作负载下,磁盘内部不可控的数据迁移会造成性能骤降.因此,DM-SMR主要用于写负载较低的消费级市场.

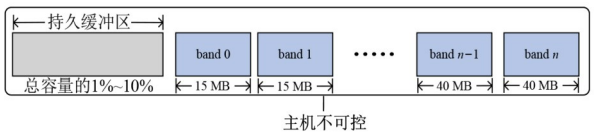


图2　DM-SMR磁盘内部结构示意图

尽管DM-SMR提供了较好的灵活性和适应性,但是作为黑盒设备,它无法为上层应用提供可预测的性能.鉴于此,业界提出了主机管理型叠瓦式磁盘(Host-Managed SMR,HM-SMR),并采用了一种全新的分区存储模式来管理设备内部存储空间[30].为了实现系统主机与分区存储设备的交互,国际信息技术标准委员会(InterNational Committee for Information Technology Standards,INCITS)分别制定了分区设备块指令(Zoned Block Commands,ZBC)[31]和分区设备指令(Zoned-device ATA Commands,ZAC)[32],明确描述了设备管理命令集,包括打开分区、关闭分区、结束分区、报告分区以及重置分区写指针等.如图3所示,该模式将SMR磁盘的连续地址空间划分为若干大小为256 MB的分区(zone),并将zone作为基本管理单元.每个zone均与一个写指针相关联,并由写指针指示当前可写入的逻辑地址.用户可以通过特定的库函数(例如libzbc[33],libzbd[34])获取zone的起始地址、写指针的偏移和状态等实时信息.此外,为了减少磁盘内部数据迁移造成的性能抖动,HM-SMR磁盘内部不再配置持久缓冲区,这意味着主机应用必须严格将数据顺序写入zone.HM-SMR简化了设备端管理数据复杂度,但其代价是需要对整个存储栈进行大幅适配性优化,以遵循顺序写约束.尽管如此,对于具备完全数据流控制能力的用户而言,主机管理型SMR是一种理想的存储设备[35~38].

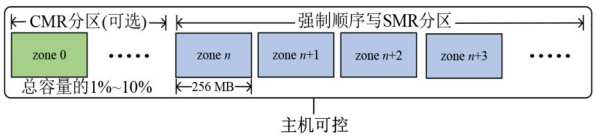


图3　HM-SMR磁盘内部结构示意图

作为HM-SMR磁盘和DM-SMR磁盘的超集,主机感知型叠瓦式磁盘(Host-Aware SMR,HA-SMR)允许主机系统和磁盘固件的STL共同参与存储管理.因此,HA-SMR被视为是在兼容传统存储堆栈和保持性能稳定性之间的一种权衡.如图4所示,这种权衡的努力来自两方面:一方面,通过保留持久缓冲区,不仅减轻了主机系统处理非顺序写的负担,而且还提供了与DM-SMR一样标准的块访问接口;另一方面,利用分区存储模式的zone接口,为主机端的数据流优化提供磁盘内部的实时数据布局信息,实现了性能的可预测性.上述特征使得HA-SMR磁盘在提供与传统存储堆栈的兼容性的同时,还可以利用主机资源实现不同程度的性能优化.然而,为了获取更高的性能和稳定性,用户仍然需要谨慎处理非顺序写入的数据流,因为持久缓冲区的数据迁移会对磁盘性能产生干扰[39~41].

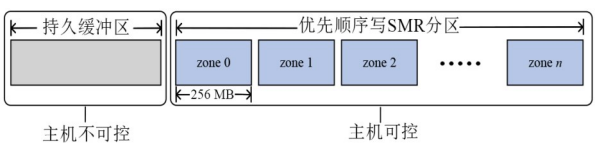


图4　HA-SMR磁盘内部结构示意图

最近,谷歌和西部数据提出了一种新的SMR模型,名为混合叠瓦式磁盘[42,43](Hybrid SMR,H-SMR).在H-SMR模型中,可以同时存在常规分区和强制顺序写分区,而且允许主机通过特定的应用程序编程接口灵活

地调节两种分区的比例. 就操作方式和物理组成而言，H-SMR 磁盘非常接近 HM-SMR 磁盘. 因此，H-SMR 磁盘既可以发挥SMR技术的高密度优势又可以利用CMR技术无写限制特点，满足用户对存储容量和性能的需求，因此被认为是一种有前景的解决方案[44,45]. 表1总结了不同类型SMR磁盘特征.

表1 不同类型的SMR磁盘模型规格说明

磁盘类型	设备形态	持久缓冲区	分区类型	CMR分区	管理复杂度	性能稳定性
DM-SMR	常规块设备	存在	不存在	不存在	低	不可控
HM-SMR	分区块设备	不存在	强制顺序写分区	可选	高	完全可控高
HA-SMR	常规/分区块设备	存在	优先顺序写分区	可选	中	部分可控
H-SMR	分区块设备	未知	强制/优先顺序写分区	可选	高	完全可控

不同于已经被广泛应用的SMR磁盘，IMR技术目前仍处于理论研究阶段. 作为SMR的衍生技术，业界可能会采用类似SMR磁盘的数据管理方式来解决IMR的写入限制. 这种方式可以在I/O堆栈的不同层次中实施，例如，主机系统、存储控制器和驱动器内部. 根据磁道布局的方式和磁盘结构的设计来看，IMR磁盘的使用方式可能不会较SMR磁盘有大幅度的变化. 因此，受益于现有SMR磁盘设计，驱动管理型的IMR磁盘和主机管理型IMR磁盘有希望出现在未来存储市场中[46]. 目前，为了验证所提出的数据管理策略，研究者只能通过模拟和仿真的方式来构建IMR磁盘模型，例如，基于知名的硬盘模拟软件DiskSim[47~49]进行修改，或者是使用类似于dm-zoned的方式在真实的CMR磁盘上构建仿真设备[50,51].

2.3 SMR磁盘和IMR磁盘的I/O性能挑战

SMR磁盘在采用叠瓦式磁道布局方式增加磁盘面密度的同时，引入了写干扰问题. 实际上，产生写干扰问题的原因在于SMR磁盘使用的写磁头宽度大于单个磁道的宽度，使得在对目标磁道执行就地更新操作时会覆盖相邻的“下游”磁道，从而破坏有效数据. 为了实现数据的就地更新，磁盘需要重写目标磁道所属的band/zone，以避免数据丢失. 在数据重写过程中，磁盘将band/zone内所有数据首先加载到内部的内存缓冲区，然后将被更新的数据合并至缓冲区，最后再将全部数据一并写回原始磁道，因此这个过程也被称为读-改-写操作(Read-Modify-Write，RMW). 然而，叠瓦式磁道布局导致即使只有一个物理块(4 KB)被更新，RMW操作也需要进行band/zone(15~40 MB，256 MB)级别的数据迁移. 由于实际写入的数据量超过了原本待写入数据量，因此RMW操作产生了写放大(最差情况下写放大率为16 384倍). 在式(1)中，$\text{rewrite_tracks}_{\text{smr}}$表示在SMR磁盘中每次更新时因写干扰而产生的磁道重写数目.

$$\text{rewrite\_tracks}_{\text{smr}}=\sum_{i=1}^{N}P_i(N-i) \tag{1}$$

其中，N表示每个zone内的磁道数目；$N-i$表示需要重写的下游磁道数；P_i表示更新特定磁道的概率. 如果每个磁道的更新请求是等概率的，那么$P_i=1/N$，因此平均每次更新需要重写$(N-1)/2$个磁道[17]. 假设每个磁道为2 MB，一个典型的256 MB zone则包含128个磁道，根据式(2)计算可得出平均每更新一个磁道需要额外重写的磁道数目为64. 在写密集型应用中，频繁的磁道重写会导致不可预测的写放大效应，极大影响了用户写请求的完成时间，是造成磁盘性能抖动的直接原因[52~54].

$$\text{rewrite\_tracks}_{\text{imr}}=2\times P_{\text{bottom}} \tag{2}$$

为了缓解写干扰问题，IMR磁盘采用上下交错的磁道布局来代替从单侧依次重叠的叠瓦式磁道布局. 在这种新布局中，写入顶部磁道所需的激光强度要低于写入底部磁道，从而保护了底部磁道数据不受影响；而更高强度的激光写入底部磁道时，却会破坏顶部磁道的数据. 这种对激光强度需求的差异性，是导致IMR磁盘产生写干扰问题的根本原因. 需要说明的是，在对底部磁道进行就地更新时，IMR磁盘也需要执行RMW操作以保证数据完整性. 相较于SMR磁盘，IMR磁盘将写干扰所影响的范围从band/zone级别降至磁道级别，大幅减少了就地更新的开销. 在式(2)中，$\text{rewrite_tracks}_{\text{imr}}$表示在IMR磁盘中每次更新时由于写干扰而产生的磁道重写数目. 其中，P_{bottom}表示更新底部磁道的概率. 如果更新请求是等概率的，即P_{bottom}和P_{top}均1/2，那么平均每次更新仅需要重写1个磁道[17]. 显然，IMR磁盘的重写代价要远低于SMR磁盘. 尽管如此，随着磁盘空间利用率的增加，越来越多的顶部磁道会受到写干扰影响[19,47,55,56]，倘若频繁进行RMW操作将会导致磁盘的性能抖动问题.

3 国内外研究现状与分析

SMR和IMR磁盘的重叠磁道布局在处理写请求时可能会引发写干扰效应，进而影响数据完整性. 虽然RMW操作可以确保数据完整性，但这会引入额外的读写操作，进而导致写放大效应. 写放大增加了用户请求的响应延迟，尤其是在处理大量细粒度随机写请求时，会造成磁盘吞吐量的急剧下降. 如图5所示，为应对此问题，工业界与学术界从多层面开展了针对SMR和

IMR磁盘I/O性能的优化研究.应用级优化方案更容易满足特定应用场景的需求,但是由于实现复杂性和长期维护成本而受到限制;系统级优化方案能够提供良好的数据管理和访问控制(例如设计特定文件系统),但可能与其他系统软件存在兼容性问题.相比之下,设备级优化方案对上层应用和操作系统透明而具备广泛的兼容性(例如优化磁盘固件程序),并且能够直接提升存储设备的读写性能,故备受国内外研究者的关注.因此本文将重点讨论设备级优化,同时也会简要回顾系统级和应用级优化的典型案例,以便读者全面了解SMR和IMR磁盘领域的研究动态.

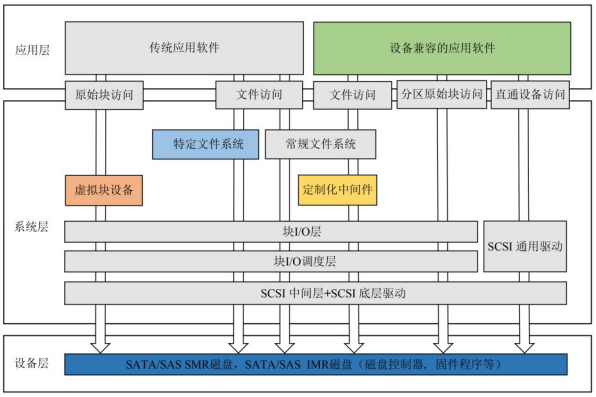


图5　SMR和IMR磁盘在Linux系统中的优化层级示意图[57]

3.1　设备级I/O性能优化

3.1.1　SMR磁盘性能优化方案

针对SMR磁盘的写干扰问题,业界聚焦于三个优化方向:磁盘数据布局、持久缓冲区以及面向SMR磁盘的混合存储系统.在SMR磁盘推出的早期阶段,由于容量较小,设计新型磁盘数据布局能够在不显著增加额外成本的情况下,减少重写操作产生的数据迁移.然而,随着SMR磁盘容量的快速提升,元数据管理的复杂度和数据迁移开销变得愈加突出.而在磁盘内部建立持久缓冲区,则可以进一步降低数据迁移频率,提供更好的持续写入性能.但是在面对I/O密集型工作负载时,机械寻道开销是限制磁盘性能的关键因素.对于此,结合高性能存储设备构建混合存储系统,能够有效弥补SMR磁盘的随机读写能力不足的缺陷,同时提升数据迁移效率,实现更高的整体性能.实际上,缓解非顺序写入产生的写干扰问题是上述优化手段的共同目标,不同之处在于它们面对不同强度、大小、访问类型的工作负载时优化效果和实现开销存在一定的差异.因此,它们既可以单独部署,也可以协同工作,来满足不同的成本需求和性能要求.接下来本小节将对这三个优化方向进行具体阐述和分析.

(1)磁盘数据布局策略

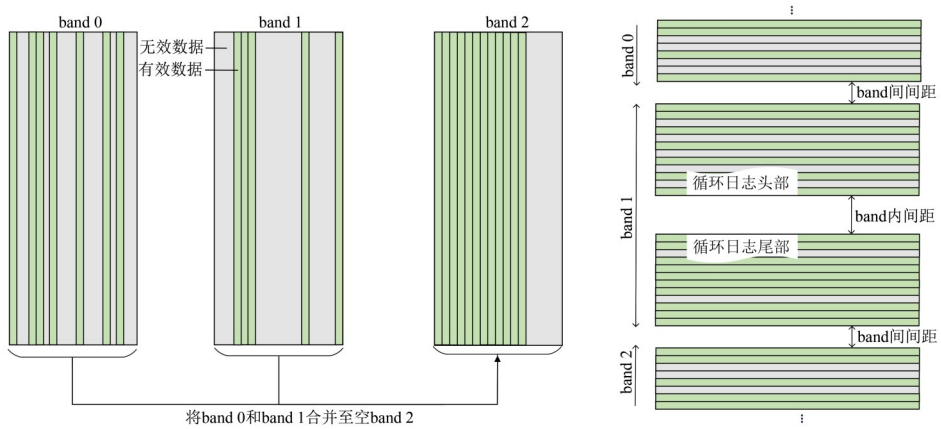
SMR磁盘无法在硬件级别支持就地随机写入/更新,这促使磁盘系统采用软件(固件)级别的数据块布局优化变得十分必要.数据布局方案的核心功能是建立映射机制来记录逻辑块地址(Logical Block Address,LBA)与物理块地址(Physical Block Address,PBA)的关系,因此映射机制包括静态映射机制和动态映射机制,而与这些映射机制相对应的数据更新模式分别为就地更新和异地更新.在就地更新模式下,磁盘通过RMW操作对目标LBA块进行就地更新,因此每个LBA被静态映射到与其相一致的PBA;在异地更新模式下,磁盘以追加写(Append Write)的方式将更新数据写入新的物理块,并动态地记录LBA和新PBA之间的关系.静态映射机制的实现开销和元数据开销极低,但是频繁的RMW操作将会严重降低磁盘性能稳定性.为了减少就地更新模式的开销,研究者通过适当增加存储成本,为SMR磁盘配置内部持久缓冲区或外部缓存来聚合非顺序写数据,从而减少数据迁移频率.与静态映射机制相比,动态映射机制能够将随机写转换为顺序写,减少RMW操作频率,因此具有较高的写入效率.但是,动态映射机制也带来了垃圾回收、内存开销以及碎片化空间等问题.本小节主要介绍动态映射机制的相关研究.如表2所示,其优化目标主要包括提升空间效率、减少垃圾回收(Garbage Collection,GC)代价以及降低内存开销.

表2　SMR磁盘数据布局优化策略分类与总结

优化类型	优化目标	核心思想	典型方案	映射级别	优化效果
动态映射	空间效率	使用循环日志方式管理数据区,提高空间利用率	文献[21]	块级	强
	垃圾回收代价	基于特定冷热识别机制实现数据分类,降低数据迁移量和GC频率	文献[58~60]	块级	较强
	内存开销	利用缓冲区对连续数据进行聚合操作,减少碎片化数据布局	文献[61,62]	块级	不稳定
		采用磁道级动态映射机制,限制元数据记录规模	文献[63~65]	磁道级	强

在块级动态映射方案中,不断到达的更新请求使得先前数据逐渐失效,因此磁盘需要定期进行数据迁移,以回收被无效数据占据的存储空间.Amer等人[21]提出了两种采用动态映射机制的数据布局方案.如图6(a)所示,第一种方案采用仅追加写模式处理写入band的数据,并通过带间清理操作将多个band内的有效数据进行合并来回收完整的空band.如图6(b)所示,第二种方案则是对每个band均采用循环日志模式进行管理,并使用头指针和尾指针指示数据写入位置和垃圾回收位置.当需要回收连续的可用空间时,磁盘执行带

内清理操作从尾指针处开始,将有效数据依次移动到头指针处. 相较于第一种方案,后者可以跟随头指针移动来利用无效数据占据的空间,提升空间利用率,进而减少数据迁移频率.



(*a*) 基于仅追加写的band管理方式 (*b*) 基于循环日志的band管理方式

图6 动态映射机制的两种典型数据布局示意图[21]

Lin等人[58]提出了一种名为H-SWD的解决方案,旨在通过冷热数据分离降低数据块的移动频率. 为了将热度相似的数据聚集在同一区域,H-SWD将整个磁盘空间的1%的区域用作热数据区,其余则是冷数据区,并基于时间窗口来识别写入数据的热度. 在此基础上,H-SWD建立了一种多区域动态垃圾回收机制,通过综合考虑空间利用率、数据热度和无效数据比例,来确定最佳的数据迁移位置. 然而,为了实现垃圾回收时的冷热数据分离,H-SWD需要在不同区域间频繁移动数据,引入了昂贵的寻道开销,从而造成不可控的数据迁移耗时.

为了从长远角度减少垃圾回收的数据迁移量,文献[59]提出了一种启发式的band压缩策略. 该策略引入了一种基于更新频率的热数据识别机制,将更新过的数据块标记为热数据,而未被更新的则归为冷数据,从而在写入阶段实现冷热数据分离. 在垃圾回收阶段,策略采用贪婪算法,优先清理充满无效数据的空band,以尽量避免进行显式的数据迁移. 在无法找到符合条件的候选者时,该策略则依据band内冷数据、热数据以及无效数据的比例和权重来计算综合权值. 由于热数据通常具有较短的生命周期,因此冷数据被赋予了更低的权重,以尽量减少热数据移动. 然而,基于写频率的热数据识别机制缺乏时效性,只适用于顺序写或周期性更新为主的工作负载. 当出现数据漂移的情况时,这种机制会增加热数据的误判概率,进而影响垃圾回收效率.

Chuang等人[60]指出,大多数现有块级动态映射机制在管理映射条目和磁盘空间时,没有考虑数据的访问模式和更新频率,从而导致了不必要的性能下降. 为解决这一问题,他们对映射管理和数据布局进行联合设计,并提出了一种名为DSTL的转换层策略. 为了减少因更新映射条目而产生的垃圾回收开销,DSTL引入了一种两级映射方案. 该方案将一部分映射条目放置在基于区间树的RAM缓存中,另一部分映射条目放置在所对应band末尾的磁道上,通过实现元数据的就地更新,减少了垃圾回收的频率. 此外,DSTL使用K均值大小检查方法预测数据热度,主动聚集更新频率相似的数据来增加band中无效数据比例,以减少数据迁移量. 然而,在垃圾回收期间,DSTL可能会多次执行数据合并操作以实现冷热数据隔离,这会导致用户请求被长时间阻塞.

异地更新模式可以将随机写转换为顺序写,这与HM-SMR磁盘的强制顺序写约束高度契合. 鉴于此,Ma等[61]提出了一项名为HMSS的解决方案. 为了减少碎片化的空间布局和映射条目的数量,HMSS基于盘内的CMR zone构建了写缓冲区,用于缓存频繁更新的数据块以及对LBA连续的数据块进行聚合操作. 此外,HMSS还使用B+树管理地址映射表,加速对指定LBA映射条目的访问. 当叠瓦分区的空间利用率达到预设阈值时,HMSS按照贪婪策略优先清理有效数据最少的分区,以最小化数据迁移次数. 然而,HMSS仅依据写入数据是否与写指针对齐来决定缓冲区的准入,忽略了数据的热度. 不仅如此,分区级LRU缓存替换算法无法精准捕捉块级别的热数据. 这些因素导致热数据被过早写回,从而触发更多的数据迁移,降低了磁盘性能.

Hao等[62]发现,碎片化数据布局会频繁触发垃圾回收,是导致磁盘性能下降的重要原因. 为解决这一问题,他们提出了一种在主机端实现映射管理的方案. 该方案使用一部分主机内存作为一级写缓冲区,通过数

据排序和请求合并来缓解LBA-PBA映射的碎片化问题.为了进一步降低地址映射表的内存开销,该方案引入了基于段的自适应LBA-PBA地址转换机制.通过将连续的多个块级映射条目合并为单个段级映射条目,有效降低了映射条目的规模.在本地数据区域进行垃圾回收时,该方案采用了一种碎片感知的回收策略,综合考虑有效数据比例和映射碎片数量来选择需要清理的分区.然而,该方案需要使用较多的内存作为写缓冲区,故增加了存储系统成本.另外,该方案在清理写缓冲区时未考虑数据的热度,这也会潜在地增加垃圾回收频率.

块级动态地址映射可以灵活、充分地利用存储空间,但这种方式需要在磁盘缓存中保存大量映射条目,尤其是对于TB级别的SMR磁盘,引入了高昂的内存开销.相较于块级地址映射,磁道级动态地址映射显著降低了元数据对内存的需求.在磁道级动态地址映射中,逻辑磁道和物理磁道采用动态映射关系,而逻辑块在磁道内的偏移则保持静态映射关系,即逻辑磁道内的块偏移与物理磁道内的块偏移相同.为了利用这种优势,Hall等[63]设计了一种磁道级数据布局优化方案.该方案将磁盘划分为两个容量不等的区域,其中,容量大的I区用于数据存储,而容量小的E区用于数据缓冲.其中E区采用循环日志的方式来管理所有的写入数据,以减少写入延迟.当E区达到存储容量上限时,该方案将启动磁道级的数据迁移,从缓冲区的日志尾部开始,依次将属于同一逻辑磁道的数据和E区剩余的有效数据迁移到I区的新磁道内.然而,该方案并未对I区中的热冷数据进行分离,这会导致碎片化的空间布局,进而增加碎片整理开销.

为了降低磁道级动态映射的垃圾回收频率,He等[64]提出了一种名为SMaRT的解决方案.SMaRT允许就地和异地两种更新操作,如果目标磁道下游没有有效数据,则采用就地更新;否则,数据会被异地更新到空闲磁道.由于频繁执行异地更新操作会导致空间碎片化,为缓解此问题,SMaRT通过主动迁移高更新频率磁道下游数据,以增加就地更新的机会.此外,SMaRT也会根据空间碎片率定期执行空间合并.然而,在进行异地更新时,与之相关联的原始磁道的有效数据也需迁移,增加了数据迁移的成本.随着空间利用率增加,SMaRT需要频繁进行碎片整理和持久化映射信息,这对磁盘性能造成了不可忽略的影响.

为了进一步减少磁道级动态映射的内存开销,Shafaei等[65]提出了一种名为Virtual Guard的解决方案.Virtual Guard的核心思想是在将数据写入目标磁道之前,主动将相邻的下游磁道迁移至持久缓冲区,来为目标磁道提供就地更新的条件.相较于SMaRT追踪全部磁道映射信息的方式,Virtual Guard只需要记录被迁移的磁道号,极大地降低了元数据规模.然而,Virtual Guard规定每个缓存磁道只能被数据磁道独占使用,这导致其无法充分利用有限的缓冲区空间.在清理缓冲区时,Virtual Guard也未考虑磁道热度,而是简单地迁移相邻磁道.即使目标磁道只有少量数据更新,也必须备份相邻磁道的所有数据.因此,在非顺序写主导的应用场景中,这些因素将频繁触发内部数据迁移,进而影响磁盘的吞吐量.

在对磁盘数据布局策略的研究中,空间效率优化方案致力于提高磁盘空间利用率,以减少数据迁移频率;而垃圾回收代价优化方案则通过区分冷热数据来减少热数据的移动,提升数据迁移效率.然而,空间效率优化方案并未深入优化冷热数据分布,效果不及垃圾回收代价优化方案.此外,针对内存开销的优化方案更多关注于如何减少动态映射机制下的元数据规模,以减轻由频繁更新元数据引起的性能抖动.但面对SMR硬盘日益增长的容量,元数据管理的复杂性也面临更大的挑战.因此,如何在不牺牲空间效率和垃圾回收代价的前提下有效降低内存开销将成为关键研究内容.

(2)持久缓冲区管理机制

为了减少RMW操作的频率,DM-SMR磁盘和HA-SMR磁盘将大约1%~10%的叠瓦式磁道用于构建持久缓冲区[29,30](也被称作介质缓存),以缓存非顺序写数据.通常,持久缓冲区以循环日志或基于段的数据组织方式管理缓冲区空间.基于循环日志的空间管理机制通过将数据追加到日志的尾部来提高写入速率.但这种方式会破坏数据块的空间局部性,致使逻辑上连续的数据块在物理空间上过于分散,进而增加之后顺序请求的完成时间.此外,为了减少清理时的写放大,该机制按照先进先出的原则批量写回属于相同band/zone的数据块,但这可能导致频繁更新的数据块过早离开缓冲区.与此同时,在基于段的空间管理机制中,数据块被限制在缓冲区的特定范围内,减少了后续顺序请求的寻道开销.其中,段的大小通常小于或等于一个band/zone大小.段的分配方式包括独占式和共享式两种.在独占式段分配方式中,每个段仅存储指定SMR zone内的数据,因此在写回段内数据时最多触发一次RMW操作.但在实际工作负载中,写请求并不总是均匀地访问LBA空间,这导致段内部分空间未被写入数据,进而降低了缓冲区空间利用率.相比之下,共享式段分配方式能够容纳来自多个SMR zone的数据块,显著提高了缓冲区资源利用率.然而,这种方式产生了新的问题,即在写回段的过程中可能会多次触发RMW操作,从而产生不可预测的清理开销.为了减少持久缓冲

区的清理开销,现有研究主要集中在缓冲区清理策略和清理调度策略的优化.如表3所示,缓冲区清理策略的优化目标包括减少写放大率、减少RMW次数以及缩短数据迁移耗时,而清理调度策略的优化目标则是通过合理调度清理操作来减少数据迁移过程引入的长尾延迟.

虽然持久缓冲区可以提高RMW操作的效率,但是不合理的数据清理顺序会造成严重的写放大,从而增加写回数据的耗时.为了尽可能减少写放大效应,Yang等[66]设计了一种名为SACE的叠瓦式磁道感知的持久缓冲区清理策略.在SACE的工作机制下,如果某缓存数据所属的逻辑磁道下方不存在有效数据,即认为其符合叠瓦式写入约束.一旦检测到持久缓冲区中存在符合叠瓦式写入约束的数据,SACE便会主动进行写回操作,从而避免了写放大效应.然而,该策略并不能从根本上改善磁盘性能,原因包括两方面:首先,SACE的主动写回机制没有考虑数据的访问热度,导致热数据频繁地参与清理操作;其次,随着有效数据比例的增加,机会主义的数据回写策略将逐渐失效,最终将重新产生写放大效应.

表3 SMR磁盘持久缓冲区优化策略分类与总结

优化类型	优化目标	核心思想	典型方案	管理形式	优化效果
动态映射	写放大	机会性的写回不会产生写放大的缓存数据	文献[66]	段式	不稳定
		预测数据热度,为热数据动态分配缓冲区	文献[67]	段式	较强
		将存储区内的热数据集中放置在防护磁道附近	文献[68]	循环日志	不稳定
	RMW次数	建立自适应工作负载的双层缓冲区,分离冷热数据	文献[69]	循环日志	强
	迁移耗时	优化缓冲区空间分配单元,降低清理规模	文献[70]	循环日志	强
		采用贪婪策略选择有效数据最少的受害者	文献[71]	循环日志	较强
		基于CMR分区构建可就地更新的缓冲区	文献[72]	段式	不稳定
	长尾延迟	区分工作集数据,提高缓冲区的写入频率	文献[39]	循环日志	较强
清理调度		人为创造空闲时间,减少阻塞清理的频率	文献[41]	—	强
		利用强化学习模型调度缓冲区的清理操作	文献[73~75]	—	强

Liang等[67]提出了一种顺序写约束的缓冲区管理机制(SWC^2),以减少清理过程中的写放大效应.该机制将持久缓冲区划分为静态缓冲区和动态缓冲区两部分.在初始阶段,SWC^2以哈希映射的方式将数据定向到静态缓冲区.在数据写入前,SWC^2会根据请求大小预测数据热度并进行分类存储,提高清理效率.当静态缓冲区已满后,SWC^2将数据分配到动态缓冲区,以延迟触发清理操作.直到缓冲区全部耗尽,SWC^2将根据成本模型选择被写回的数据.通过上述方法,SWC^2有效降低了清理频率和写放大效应.但在数据访问高度集中的场景下,基于哈希的空间分配方式可能无法充分利用静态缓冲区的存储空间.同时,被动清理方式会导致数据迁移时间的不确定性,从而引发磁盘性能波动.

针对现实工作负载中磁道更新概率的不均匀性,Lee等[68]提出了一种自适应写干扰管理的STL设计方案.为了减少写放大效应,该方案采用磁道级热度识别机制,主动将高频更新磁道的数据迁移到预留的静态保护磁道附近,从而减少参与清理的磁道数目.对于更新频率相对较低的磁道,该方案引入了延迟重写机制,通过设置触发清理操作的条件,来进一步降低缓冲区清理频率.另外,该方案还专门设计了一种支持就地更新的映射项缓存管理策略,通过连续的磁盘读写操作来更新和检索映射元数据,减少了随机访问产生的寻道成本.然而,磁道级热数据识别机制的准确率较低,可能会导致大量冷数据参与数据迁移,进而降低回收效率.

传统的持久缓冲区管理模式通常使用单层缓冲区结构,并以循环日志的形式组织写入数据.然而,在这种模式下,清理策略只能从日志头部依次写回缓存数据,这不可避免地产生了热数据回写问题,从而触发冗余的RMW操作.为了缓解这一问题,Ma等[69]创新性提出了一种双缓冲区管理方案.该方案将持久缓存划分为过滤区和持久区,分别用于缓存热数据和处理写请求.当触发持久区的清理操作时,双缓冲区管理方案对数据块按照更新频率排序,将高更新频率的数据块迁移至过滤区,而其余的数据块写回所属的band区域,以此缓解热数据回写问题.为了进一步优化空间利用率,该方案还加入了一个基于采样的动态分区机制,能够根据工作负载调整各缓冲区的容量.尽管这一机制减少了热数据回写频率,但动态分区操作和大规模的持久区域清理操作会导致不可预测的时间和性能开销.

在清理持久缓冲区时,磁盘会将传入的用户请求暂时挂起,从而引入了磁盘性能恢复时间(Performance Recovery Time,PRT)的问题.为了缩短PRT,Ma等[70]提出了一种名为K-帧回收(KFR)的高效持久缓冲区管理策略.如图7所示,KFR将沿着叠瓦方向的任意k个连续扇区组成单元定义为一个K帧.在分配缓存空间时,除了采用传统的追加写模式,还会根据写磁头的覆盖

范围,将无效K帧中的扇区分配给即将到来的数据,以延迟触发回收操作. 与此同时,KFR通过将待写回的数据从band级别分解至更小的K帧级别,有效减少了持久缓冲区的清理负荷. 此外,如果没有其他可调度的扇区时,KFR以按需的方式将包含有效数据最少的K帧写回,以减少参与RMW操作的数据量.

Liu等[71]基于应用程序视角,为HA-SMR磁盘设计了一种名为CLDM的持久缓冲区清理策略. 为了加速清理过程,CLDM基于缓存分区的热度和数据量设立了双重条件来筛选最佳候选分区. 在写回缓存分区之前,CLDM首先根据缓存分区的平均访问频率来设立热度阈值. 随后,CLDM根据数据量对缓存分区的升序进行排序,并从有序列表起点开始,逐一写回低于热度阈值的缓存分区. 然而,一个256 MB的分区包含了65 536个4 KB的数据块. 这意味着,基于分区访问频率识别机制可能会因为个别热数据块而将整个分区都视为热数据,从而延迟冷数据的写回,降低了缓冲区的空间效率.

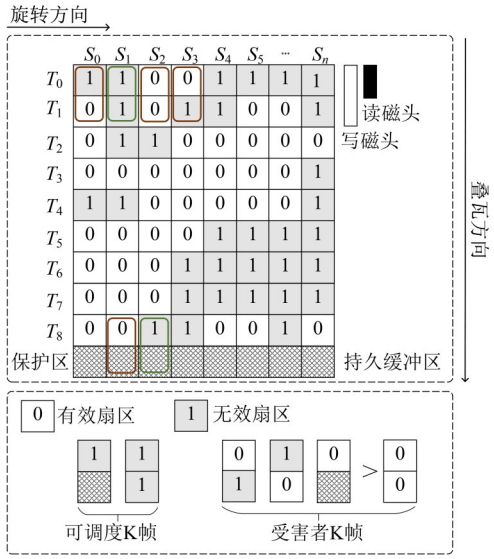


图7　KFR策略中的持久缓冲区结构示意图[70]

基于叠瓦式磁道构建的持久缓冲区通常采用异地更新的方式来处理更新请求. 因此,密集的数据更新会加速缓冲区空间消耗,从而频繁触发缓冲区清理. 为了降低清理频率,Hajkazemi等[72]利用HM-SMR磁盘内部配置的CMR zone设计了一种名为μCache的解决方案. μCache将LBA空间划分为若干固定大小的桶,并使用CMR zone构建了一个支持就地更新的可重写持久缓冲区. 当写请求到达后,如果其所属的桶仍驻留在SMR zone,μCache会先将待写入数据和桶内剩余有效数据通过晋升操作迁移至可重写持久缓冲区,以利用工作负载的空间局部性. 虽然该方案为SMR磁盘的持久缓冲区的设计提供了新思路,但是在时间局部性主导的工作负载下,可能会频繁对桶进行晋升和降级操作,从而产生大量数据迁移开销,造成磁盘性能下降.

持久缓冲区的清理时机也是影响磁盘性能的重要因素. 通常,持久缓冲区的清理模式包括空闲清理模式和阻塞清理模式. 空闲清理模式利用工作负载中较长的请求间隔来调度清理操作,但是在写密集型工作负载下,该模式难以寻找到合适的时机,进而会触发阻塞清理模式. 然而,阻塞清理模式的启动时间和持续时间无法预知,这导致了长尾延迟现象. 为了减少阻塞清理的频率,Zhang等[41]提出了一种名为Idler的方案. Idler综合考虑了缓冲区资源利用率和当前工作负载特性,如请求到达速率、非顺序请求的比例和请求大小,以自适应地触发空闲清理模式. 为防止I/O响应时间的显著增加,Idler根据缓存空间和映射条目的利用率设置清理阈值,并采用启发式方法动态调节空闲周期长度. 此外,Idler维护了一个幽灵缓存以记录每个分区的缓存资源使用信息以及写指针位置,从而实现了低成本、实时的缓存资源利用率监控. 总体而言,该方案以可控方式人为创造空闲时间,引导持久缓冲区进入空闲清理模式,从而有效地减少了长尾延迟的发生.

启发式方法主要依赖于设备端或主机端的计算和管理资源,以及对工作负载的离线分析来调度清理任务. 在面对黑盒型DM-SMR磁盘时,这种方法缺乏灵活性和适应性. 强化学习方法能够通过迭代的方式对未知环境观察和探索来做出最合适的决策,这对时间敏感和资源受限的DM-SMR磁盘非常适合. 为了解决DM-SMR磁盘清理过程中的长尾延迟问题,Pan等[73]基于Q-learning模型,提出了一个名为RL-Cleaning的清理调度方案. 首先,RL-Cleaning通过监测I/O请求的时间间隔和持久缓冲区的空间利用率来构建状态空间. 随后,RL-Cleaning根据待清理数据band的数量定义可执行的动作,并通过分解清理规模优化空闲窗口的使用效率. 最后,RL-Cleaning构建了一个基于I/O延迟的分段奖励函数来不断提升决策精度. 为进一步降低长尾延迟的发生,该方案设置了一个保留缓冲区,以处理清理期间的I/O请求.

一些研究尝试将闪存嵌入SMR磁盘中充当持久缓冲区,以减少垃圾回收的耗时. 然而,闪存的“先擦后写”特性会周期性地触发垃圾回收操作,从而产生长尾延迟. 为了解决上述问题,Ma等[74]基于Q-learning模型,设计了一个名为MAID-Q的智能清理调度机制. 该机制通过三个主要方面协同工作来减少长尾延迟的发生. 首先,MAID-Q把回收规模从粗粒度的band级别降低至细粒度的闪存块级别,以缩短数据迁移时间. 然后,MAID-Q利用闪存的高并行优势,将块级回收分为

可挂起和不可挂起两个阶段,进一步减少用户请求的排队时间. 最后,MAID-Q集成了一个基于Q-learning的辅助调度器,该调度器能通过与运行环境的互动学习来确定清理过程中可以接受的RMW操作次数,从而充分利用系统的空闲时间执行缓冲区清理操作.

在实际应用场景中,SMR磁盘可能会被多个应用程序共享使用. 由于应用程序访问特征的多样性,磁盘上不同区域的访问模式会随着空间和时间维度变化而变化. 因此,仅依靠单代理的强化学习清理调度方案往往难以及时适应不断变化的访问模式. 为了解决这一局限性,Shen等[75]在RL-Cleaning的基础上研发了一种名为MARL-Cleaning的多代理强化学习清理方案. 该方案使用分布式Q-learning方式来减少Q表的存储开销,并且指定每个代理负责调度特定区域的清理操作. 不仅如此,MARL-Cleaning还为代理间增加了协调机制,以确保在多代理环境下能够最大化系统的整体效益. 此外,MARL-Cleaning对每个代理的开发-探索策略设置了差异化的ε值,避免对某些区域探索和开发时出现不平衡的现象. 最后,为了更好地适应多变的应用访问模式,MARL-Cleaning利用DBSCAN聚类算法动态地调整奖励函数阈值,以进一步降低长尾延迟的发生.

不同于优化清理调度时机的方案,Yang等提出,可以适当利用主机端的计算和管理资源,来提高HA-SMR磁盘性能的可预测性. 为此,他们设计了一种名为VPC的解决方案[39],其核心思想是将少量叠瓦式分区用于构建容量可配置的虚拟缓存,以隔离不同更新频率的数据. 为了提高数据迁移效率,VPC基于时间窗口将更新数据划分为工作集数据和非工作集数据,并将它们分别重定向至虚拟缓存的特定区域. 其中,高更新频率的工作集数据,由于其失效速度较快,被直接将定向至持久缓冲区,以降低持久缓冲区的清理频率. 为了减少长尾延迟,VPC还引入了一种多阶段的虚拟缓存清理机制,利用空闲时间逐批次写回缓存数据. 总之,该方案从数据分配和清理策略两方面协同增强持久缓冲区的清理效率,尽管牺牲了一定的存储容量,但却有效减少了由持久缓冲区清理导致的长尾延迟.

在针对持久缓冲区管理机制的研究中,清理策略优化方案通过参考数据局部性、数据迁移量以及数据布局等因素来选择最合适的受害者,从而降低缓冲区清理的时间开销和频率. 相比之下,清理调度优化方案则关注的是何时对选中的受害者进行数据迁移,这需要考虑缓冲区使用模式、工作负载强度和任务优先级等因素,以尽可能缓解数据迁移对用户请求的干扰. 实际上,清理策略优化方案和清理调度优化方案具有正交关系,如何有效地结合上述方案来提高SMR磁盘性能稳定性是一个重要研究议题.

(3)面向SMR磁盘的混合存储系统

随着闪存价格下降和容量增加,利用闪存固态硬盘(SSD)来改善由SMR磁盘构成的存储系统随机写性能成为了一种经济高效的解决方案. 混合存储系统主要包括分层式[76~80]存储系统和缓存式[81~85]存储系统. 在缓存式存储系统中,SSD用于充当主存和磁盘之间的非易失性缓存,根据局部性原理将访问频繁的数据块放置在缓存. 在分层式存储系统中,SSD被用来与SMR磁盘构建统一的线性地址空间,其中,访问频繁的LBA区间被整体迁移到SSD性能层,以加速I/O请求的访问. 然而,分层式混合存储系统需要定期在SSD和SMR磁盘之间进行数据迁移,对于读写性能不对称的SMR磁盘而言,不当的分层策略会引发剧烈的性能抖动. 此外,在一些时间局部性特征主导的工作负载中,频繁访问的数据块可能分散在不同的LBA区间,而冷数据占据着SSD的大部分空间,降低了SSD的空间效率. 因此,面向SMR磁盘的混合存储技术的研究重点是如何将SSD以缓存的方式与SMR磁盘集成,从而平衡混合存储系统的成本和性能. 在面向SMR磁盘的缓存管理策略中,根据缓存功能可以分为读写缓存优化和只写缓存优化两类. 将SSD作为读写缓存可以充分利用其并行能力,扩大SMR磁盘的应用范围,但这可能会因读脱靶而触发缓存替换,进而引发SMR磁盘内部的数据迁移. 鉴于此,部分研究主张将SSD作为只写缓存,以减少SMR磁盘内部的数据迁移频率. 需要注意的是,在面对写密集型工作负载时,如果SSD缓存无差别吸收写入流量,将会加速闪存磨损,进而缩短SSD的使用寿命. 如表4[86~99]所示,面向SMR的缓存式存储系统的优化目标主要包括平衡命中率与写放大率、减少RMW次数、延长闪存寿命以及降低持久缓冲区的清理开销.

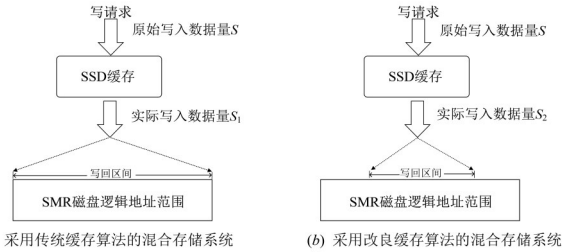
文献[86]提出了一种混合波状瓦片记录(HWSR)磁盘系统,旨在提升DM-SMR磁盘的性能与容量. 为了减少写放大率,HWSR采用了一种基于段的数据布局,以取代传统的分区式数据布局. 在所提出的布局方式中,HWSR沿磁盘径向将数据写入特定的段,有效地将随机写入引发的写放大限制在有限范围内. 同时,HWSR分别使用RAM和SSD作为写缓冲区和读缓存. 作为读缓存,SSD采用LRU链表管理缓存块,以尽可能维持高命中率. 与以命中率为优化导向的读缓存不同,写缓冲区的LRU链表分为工作区和替换区,通过驱逐替换区内缓存块最多的段来减少分段数据布局对逻辑数据连续性的影响. 尽管HWSR利用新的磁盘布局有效抑制了写放大,但实际上却牺牲了数据的空间局部性,增加了顺序I/O的访问延迟.

Wang等[87]经过实验发现,随着从SSD缓存层淘汰的数据块地址范围增加,磁盘性能逐渐恶化. 为了提高

表4　基于SMR磁盘混合存储系统优化策略分类与总结

优化类型	优化目标	核心思想	典型方案	缓存类型	优化效果
混合存储管理方案	平衡写放大与命中率	对外部缓存采用分级管理机制,以隔离不同局部性特征的数据块	文献[86]	写缓存	强
		重新设计磁盘布局以减少写放大,并将SSD作为读缓存维持高命中率	文献[87]	读缓存	较强
		限制从SSD缓存层淘汰的数据块地址范围,减少参与数据迁移的分区	文献[88]	写缓存	较强
	RMW次数	基于局部性、读写开销等因素建立成本模型,根据代价选择受害者	文献[89~91]	读写缓存	强
		优先回收无效的闪存块,以减少写回次数	文献[96]	写缓存	不稳定
	持久缓冲区清理耗时	对外部缓存数据进行重排序,以减少清理持久缓冲区时数据收集开销	文献[92~94]	写缓存	强
		将闪存嵌入到SMR磁盘内部,以代替磁介质构建持久缓冲区	文献[95,96,98,99]	写缓存	强
	闪存寿命	根据历史访问信息和工作负载特征,将数据块加入缓存中指定区域	文献[97~99]	文献[97]为读写缓存,文献[98,99]为写缓存	强

SMR磁盘性能的稳定性,他们设计了一种名为PORE的混合存储解决方案. PORE的核心理念是通过限制被驱逐脏块的LBA地址范围,来平衡命中率和写放大率. 如图8所示,根据PORE的设计,只有开放区域的脏数据块会被淘汰,而禁止区域的脏数据块则被暂时保留在缓存,通过减少缓存替换时参与RMW操作的分区数量来降低写放大率. 为适应工作负载访问特征的变化,PORE会定期更新开放区域集合. 在进行缓存替换时,根据记录的数据块访问时间来驱逐最老的数据块,以维护缓存命中率. 然而,由于淘汰数据块会被随机地写入磁盘,这种驱逐方式可能会增加持久缓冲区清理时的收集开销,从而导致磁盘性能进一步恶化.



(a) 采用传统缓存算法的混合存储系统　(b) 采用改良缓存算法的混合存储系统

图8　传统缓存算法和面向SMR磁盘的缓存算法对比示意图[87]

传统的外部缓存管理策略对持久缓冲区管理机制缺乏感知能力,无法按需进行数据迁移,进而可能导致不可预测的写放大. 此外,为了降低写放大率,外部缓存通常采用分区级的替换策略,这不可避免地会导致一些热数据被驱逐,从而降低了缓存命中率. 为了解决上述问题,Zhang等[88]提出了一种名为MCB的方法,旨在调和混合存储系统中写放大和缓存命中率之间的冲突. 为了保持外部缓存的命中率,MCB根据I/O请求的特征将数据定向到指定的存储区域,通过减少外部缓存的清理频率,以尽可能将热数据长时间保留在外部缓存. 不仅如此,MCB还引入了一个量化数据块流行度的指标,利用该指标将外部缓存内的冷热数据相互隔离,以进一步减少分区级缓存替换造成的命中率下降. 此外,MCB基于磁盘内部预留的CMR分区构建了一个外部缓存感知的持久缓冲区,并通过与外部缓存协同调度缓冲区的数据迁移,将写放大率控制在可接受的范围内.

通常,面向SMR的缓存式混合存储系统旨在降低写放大率或者提高缓存命中率,但通常无法实现明显的性能改进. 为了解决这一挑战,Sun等[89]提出了一种新型SMR磁盘感知的缓存管理策略SAC. 该策略通过深入分析DM-SMR磁盘内部持久缓冲区和band的物理布局,以更精准地指导外部缓存替换决策,从而有效地缓解了写放大效应. 此外,为了在写放大和数据流行度之间取得良好的平衡,SAC利用皮尔逊相关系数来分析缓存命中率、写放大率以及RMW次数与I/O时间之间的相关性,并得到降低RMW操作的频率是解决上述问题最有效方式的结论. 此外,考虑到SMR磁盘读写性能之间存在显著的差异,SAC引入了一个基于工作负载特征和实时读写性能的驱逐成本模型,用于动态平衡干净数据与脏数据在外部缓存的占比. 总体来说,这种软硬系统设计的方法能更全面地解析影响面向SMR的缓存式混合存储系统性能关键因素,对提高SSD的部署效率具有重要价值.

最近,Sun等[90]在SAC的基础上又提出了一种名为SAC+的通用缓存式混合存储框架,以提高混合存储系统中DM-SMR和HM-SMR磁盘之间的兼容性. SAC+遵

循SAC框架的核心理念,即以减少RMW操作数量为目标来优化写性能.不同的是,该框架引入了一个设备类型感知的SMR设备代理,专门用于提高不同类型SMR磁盘的写入效率.SMR设备代理是基于POSIX系统调用和libzbc用户库实现的,能根据SMR磁盘类型调用相应的读写函数,并将I/O请求传递到相应的设备.更进一步,SMR设备代理还通过封装针对不同设备类型的高级操作,实现了缓存替换策略与设备类型的解耦.例如,在HM-SMR的情境下,该框架提供了"partRMW"操作,以减小合并到目标区域时的读写数据量.对于DM-SMR,则提出了"写时排序"操作,以减少清理持久缓冲区时的数据收集开销.通过以上措施,SAC+有效地提高了外部缓存和内部持久缓冲区的协作效率,从而全面提升了混合存储系统性能.

Zheng等发现,基于时间局部性的LRU算法及其变体在减少RMW次数的问题上表现不佳.针对这一问题,他们深入研究了缓存块的空间局部性和干净缓存块的写回次序对RMW次数的影响,并提出了一种名为SLA的解决方案[91].SLA采用SSD作为DM-SMR磁盘的读写缓存,并将缓存块以band为单位组织,来减少缓存替换时收集数据块的开销.该策略综合考虑了缓存块的空间和时间局部性,为每个缓存band建立了基于访问时间和空间局部性的权重模型.当进行缓存清理时,为了维持高缓存命中率,SLA首先依据各缓存band的权重构建一个指定大小的候选者集合,然后选择包含最多干净块的缓存band进行写回,以降低RMW操作频率.然而,需要指出的是,在读写平衡型工作负载下,由于SLA总是优先淘汰干净块最多的集合,因此会导致持久缓冲区中存在大量不同band的数据块,从而增加RMW操作频次.

在持久缓冲区清理过程中,数据块收集效率不佳是导致设备在处理持续非顺序写时性能下降的原因.为了应对上述问题,Xiao等[92]提出了一种band感知混合存储管理策略HS-BAS.HS-BAS采用SSD作为写缓存以加速随机写请求的响应.对于顺序写请求,HS-BAS则将其旁路至SMR磁盘,以减少SSD缓存的写压力.当触发外部缓存替换时,HS-BAS将同一band中缓存块最多的块组写回到SMR,以尽可能释放更多的缓存空间并减少在清理持久缓冲区时收集数据块的开销.然而,HS-BAS在缓存替换时完全没有考虑数据热度,这不仅会频繁引发外部缓存的清理,还会加速SSD的磨损,从而降低混合存储系统整体的性能和可靠性.

在混合存储系统中,密集的非顺序写操作会频繁触发持久缓冲区清理,从而导致性能下降.为解决此问题,Liu等研究者对非顺序写操作的特性进行了详细分析,并识别出包括更新写、随机写和乱序写在内的多种非顺序写请求类型.在这些非顺序写请求类型中,乱序写请求具有良好的空间局部性,因此有机会被重新排序为顺序写请求.基于此,Liu等[93]设计了一种名为ROCO的缓存框架.ROCO采用SSD作为写缓存,将所有非顺序写入的数据按照逻辑块地址重排序后插入到相应分区的缓存块组中.当缓存数据需要被写回HA-SMR磁盘时,ROCO利用磁盘提供给主机的分区地址信息,优先将符合顺序写条件的缓存块组写回其所属的分区,而未能满足这些条件的数据则被写入磁盘的持久缓冲区.这种设计有效减轻了SMR磁盘的持久缓冲区清理压力,从而提升了磁盘整体性能.然而,ROCO只考虑了被驱逐的受害者是否会触发RMW操作,导致热数据被过早驱逐.热数据反复进出SSD缓存,不仅增加了缓存替换的频率,而且也会缩短SSD的使用寿命.

虽然SSD缓存有助于延迟触发SMR磁盘内部的持久缓冲区清理操作,但是并不能显著减少由非顺序写请求导致的RMW操作.为了解决这个问题,Xie等[94]设计了一种名为ZoneTier的策略.不同于传统的单一存储架构,ZoneTier充分利用SSD随机读写能力强的优势,将其划分为性能层和缓存层,构建了一个性能层与缓存层共存的混合存储架构.为了防止大量的非顺序写数据进入持久缓冲区,ZoneTier定期将非顺序写密集的区域迁移到SSD性能层,有效地减少持久缓冲区的清理开销.除此之外,ZoneTier也对未迁移到SSD的SMR磁盘分区实施了特殊处理,通过将这些区域的非顺序写请求重定向至SSD的缓存层,并进行数据重组来进一步减缓持久缓冲区的清理压力.

对于基于磁介质的持久缓冲区而言,寻道开销是制约清理效率的瓶颈.尤其是在进行大规模清理操作时,磁头需要在缓冲区与磁盘的本地存储区之间频繁移动,严重影响了磁盘性能.为了解决这个问题,Ma等[95]提出了一种名为FC的新型存储架构,通过在磁盘内部嵌入了NAND闪存,以取代传统的基于磁介质的持久缓冲区.FC将所有到来的数据直写入到闪存缓冲区,以增加闪存擦除次数为代价,显著降低了数据写入响应延迟.当磁盘进行持久缓冲区的垃圾回收时,同一回写路径的有效闪存页被批量写回所属band,而剩余有效闪存页则被复制到预留闪存区域.然而,在写入密集型工作负载下,闪存作为一级缓存会遭受巨大的写压力,加速闪存磨损.

在基于闪存的持久缓冲区中,一个闪存块可能会包含多个不同的band/zone的数据页.这种存储方式在执行闪存块的垃圾回收时会触发多次RMW操作,降低持久缓冲区的清理效率.为此,文献[96]提出了一种名为MU-RMW的策略.MU-RMW明确规定,每个闪存块仅存储来自同一band的数据页,并通过链表集中管理

隶属于同一band的闪存块. 同时,MU-RMW采用混合级集中地址映射机制,将扇区级映射表分散存储在各闪存块的OOB备用区,以减少对内存空间的消耗. 在进行垃圾回收时,为避免触发额外的RMW操作,MU-RMW优先回收那些不包含有效数据的闪存块. 如果不存在这样的闪存块,MU-RMW将会迁移占用闪存块最多的band的数据页,因为这些闪存块内可能包含大量无效的闪存页,能够释放更多的存储空间. 与FC一样,MU-RMW同样将闪存缓冲区视为一级缓存,缺乏对闪存磨损均衡的考虑.

使用SSD作为外部缓存可以有效地改善SMR磁盘的性能,但大量仅访问一次的数据块进入SSD缓存后,会产生虚假的写入压力问题. 这不仅加速了闪存磨损,还会延长SMR磁盘性能恢复时间. Xie等[97]设计了一种名为Duchy的混合存储系统框架,以实现SSD耐久性和SMR磁盘清理开销之间的平衡. Duchy将SMR磁盘分为"开放区域"和"非开放区域",只允许非开放区域的脏数据块、重复访问的脏、干净块进入缓存,以减少SSD缓存的写入量. 所有缓存块被按照LRU算法排序,防止热数据被过早驱逐. 当触发SSD的缓存替换时,Duchy优先从指定的窗口范围内选择靠近LRU端且属于开放区域的干净块作为受害者. 若无符合条件的缓存块,Duchy将直接驱逐LRU端的非开放区域缓存块,以尽快完成缓存清理. 此外,每当非开放区域缓存块被写回,Duchy将最久未被访问的开放区域变更为非开放区域,以限制RMW操作的范围. 然而,持久缓冲区的清理操作完全由设备内部的磁盘控制器负责,外部缓存采用的细粒度驱逐策略可能会引发不可预测的写放大,从而降低磁盘性能.

减少写入闪存的数据量可以在一定程度上延长SSD的使用寿命,但是设备管理型固态盘是以黑盒的形式呈现给上层应用,导致外部缓存管理机制无法从根本上实现闪存块的磨损均衡优化. 为了解决这一问题,Ma等[98]提议将闪存嵌入到SMR磁盘内部,并提出了一种名为RMW-F的策略. 为了最小化闪存的写入量,RMW-F根据写入请求是否会触发RMW操作来决定数据应写入的位置. 如果当前写请求会触发RMW操作,数据将被重定向到一个在闪存缓冲区中特定的闪存块. 相反,如果写请求不会触发RMW操作,则数据会直接被写入到目标存储区. 当触发垃圾回收时,RMW-F策略综合考虑了每个闪存页的访问频率和写回成本,并优先写回成本最低的闪存块. 然而,为了提高闪存块内数据的局部性,RMW-F策略规定每个闪存块只能存储同一叠瓦方向上的扇区,这极大降低了闪存块的空间利用率.

为了减少RMW操作并延长闪存的使用寿命,文献[99]提出了一种名为MCM的策略,用于协同管理SMR磁盘和NAND闪存. 为了提升垃圾回收的效率,MCM参考band的历史访问序列,并结合基于时间窗口的双重布隆过滤器,来评估数据的顺序性和访问热度,以实现不同特征数据的分类存储. 为了确保闪存块的磨损均衡,MCM综合考虑闪存块的擦除次数和无效页的比例,将持久缓冲区内权值最小且已满的闪存块迁移至空闲的未满区域. 当持久缓冲区需要进行垃圾回收时,MCM会重新评估数据热度,以减少不必要的RMW操作. 然而,MCM可能会产生较大的垃圾回收开销,原因在于,MCM为了分离冷热数据,需要在多个存储区域进行数据迁移,延长了数据迁移的周期,进而导致用户请求被长时间阻塞. 此外,由于闪存块内可能包含多个来自不同band的数据页,因此使得擦除一个闪存块需要多次调用RMW操作.

在面向SSD-SMR磁盘缓存式存储系统研究中,一种方法是使用SSD作为外部缓存,另一种则是将闪存直接嵌入到SMR磁盘中. 使用SSD作为外部缓存的方式允许策略根据工作负载的变化动态调整缓存容量,从而满足上层应用的性能需求. 相比之下,将闪存嵌入SMR磁盘的方法在感知数据布局方面表现更出色,有助于实施有效的磨损均衡策略,从而延长闪存寿命. 然而,受设备内部计算和缓存资源的限制,这种方法在缓存管理策略设计时面临更多挑战. 两者在性能和寿命方面各有优势,因此未来研究在设计混合存储系统方案时需要充分考虑工作负载的特征,以达成灵活性与性能之间的平衡.

3.1.2　IMR磁盘性能优化方案

SMR磁盘在进行细粒度随机写时会遭受严重的写入惩罚,导致其应用场景有限,无法从根本上取代CMR磁盘. 作为CMR磁盘的替代品,IMR磁盘因其更高的容量和更低的写入代价而开始受到关注. 但是,SMR和IMR技术在磁道布局方面存在差异性,直接将为SMR磁盘设计的方法应用在IMR磁盘上,会产生昂贵的磁道重写开销. 如表5所示,当前在系统软件和算法方面的研究主要集中在数据放置和数据洗牌策略的设计.

(1)数据放置策略

在IMR磁盘中,顶部磁道和底部磁道数量一致,如果按照逻辑磁道顺序依次分配磁道,那么会过早地产生写干扰效应. 因此,研究者提出了一系列数据放置策略来优化磁道的分配顺序,以尽可能避免数据放置时产生额外的I/O开销. 当前数据放置策略主要分为两类:静态数据放置策略和动态数据放置策略. 然而,无论是静态数据放置策略还是动态数据放置策略,数据所在的逻辑磁道号不一定与其所分配的物理磁道号相一致,因此它们均需要维护磁道级的映射信息.

表5 IMR磁盘数据放置与数据洗牌优化策略分类与总结

优化类型	实现方式	实现开销	灵活性	性能	核心思想	典型方案
数据放置	静态放置	低	低	中等	根据空间利用率为新数据顺序分配磁道,减少算法内存开销	文献[47,55,100,101]
	动态放置	高	高	高	基于磁道热度为新写入数据动态分配磁道,降低磁道重写概率	文献[56,102]
	块级迁移	高	高	高	将热数据迁移至无写干扰区域,提高无写干扰区域空间利用率	文献[47]
数据洗牌	磁道级迁移	低	低	中等	将热磁道内数据块迁移至无写干扰区域,缩短数据迁移周期	文献[19,55,56,102]

静态数据放置策略主要是根据空间占用率来决定磁道的分配顺序,以尽可能将数据均匀放置在顶部磁道,从而降低磁道重写的概率. 例如,Gao等[100,101]提出了两阶段和三阶段数据放置策略. 在磁盘空间占用率不超过50%时,两阶段数据放置策略沿着磁盘的径向,由外径至内径来分配底部磁道. 数据被按照它们在逻辑磁道内的偏移,依次放置在所分配的物理磁道上. 在此阶段,只有底部磁道存在有效数据,因此向底部磁道写入数据不会产生写干扰问题. 在底部磁道耗尽后,再按照相同的方式分配所有剩余的顶部磁道. 而对于三阶段数据放置策略,在磁盘空间占用率不超过50%时,它与两阶段数据放置策略的分配方式相同. 如图9(*a*)所示,当开始分配顶部磁道时,三阶段数据放置策略将以跳跃的形式来分配顶部磁道. 例如,在分配顶部磁道1后,下一个分配是顶部磁道5,而不是顶部磁道3. 由于在50%~75%的空间占用率范围内,每次重写最多只涉及一条顶部磁道,因此三阶段数据放置策略进一步降低了磁道重写开销. 而当空间占用率超过75%后,三阶段数据放置策略将依次分配剩余的顶部磁道. 在三阶段数据放置策略的基础上,Wu等[47]提出了一种"之"字形数据放置策略. 如图9(*b*)所示,在磁道分配的第二个阶段,该策略将磁道分配顺序改为从内到外,以维护逻辑数据块的连续性,进而缩短顺序请求的寻道距离.

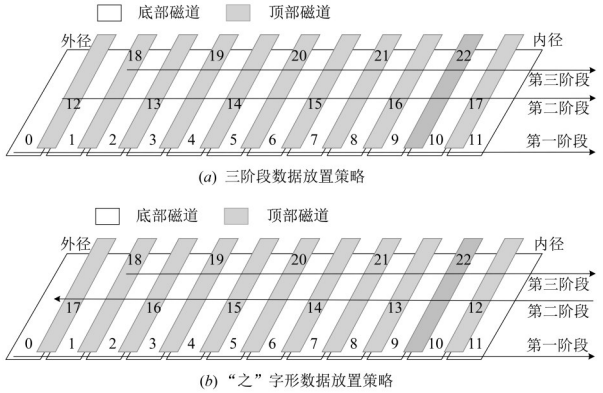


图9 三阶段数据放置过程和"之"字形数据放置过程对比示意图[47]

尽管静态数据放置策略显著降低了RMW频率,但它忽略了与所分配磁道的相邻底部磁道更新频率,从而会产生额外的磁道重写开销. 为此,Liang等[56]提出了一种自下而上的数据放置策略. 该策略将数据放置分为了底部阶段和顶部阶段. 在底部阶段,为了延迟磁道重写操作,策略优先将数据放置在底部磁道,并使用不同的地址映射粒度来保持逻辑数据块的连续性. 进入顶部阶段后,则根据记录的磁道更新频率信息,尽可能将数据放置在靠近冷底部磁道的顶部磁道,以减少不必要的磁道重写. 为了追踪磁道热度,该策略需要记录底部磁道的更新频率. 而每TB存储空间需消耗2 MB内存来记录磁道的更新频率,对于数十TB容量的IMR磁盘而言,这会产生显著的内存开销. 此外,基于频率的热数据识别机制容易受长期积累的历史数据影响,无法及时适应工作负载的访问模式改变.

对于资源受限的存储设备,传统基于频率的磁道热度评估机制会产生显著的计算和存储开销. 为缓解这一问题,Zhang等[102]设计了一种自适应写干扰的数据放置策略. 该策略在两阶段机制基础上,进一步考虑了寻道时间和磁道热度等因素. 在底部磁道分配阶段,策略从上次I/O请求访问的磁道附近开始寻找候选者,从而降低寻道成本. 在顶部磁道分配阶段,策略根据磁道热度进行选择合适的候选者. 为此,该策略建立了一个基于多布隆过滤器的磁道热度评估模型,使用磁道级别的决策来指导磁道分配,以平衡热度识别准确性和内存开销. 为了进一步提高评估的准确率,策略依据被评估磁道在布隆过滤器出现"阳"性结果的频率,将其热度细化为冷、温和热三个级别. 此外,该策略还会定期重置指定的布隆过滤器,以捕捉磁道热度的新近性.

在针对IMR磁盘的数据放置策略中,现有研究均采用磁道级别的动态映射机制. 其中,静态放置方案只需按照预定的顺序依次分配,因此实现开销极低,但缺乏感知数据特征的能力,无法动态适应工作负载的访问模式变化. 动态放置方案在分配磁道时综合考虑了数据热度、寻道距离以及空间局部性等因素,能够最大化资源利用率,提升整体性能,但在数十TB级别的磁盘容量环境下,所需要的实现开销更大. 因此,如何将静态放置与动态放置方案结合进行优势互补,成为了值得探索的问题.

(2)数据洗牌策略

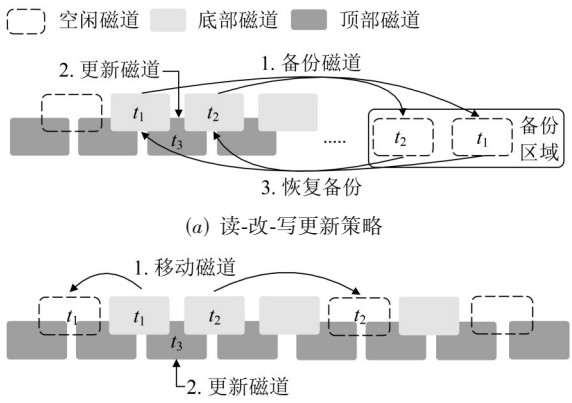
工作负载的访问模式可能会随着时间推移而变化,例如,一些处在底部磁道的数据逐渐被频繁更新,而一些处在顶部磁道的数据几乎不会再被更新. 因此,数据洗牌的目标是主动或者周期性在顶部磁道和底部

磁道之间交换冷热数据,以尽量将热数据存储在顶部磁道、冷数据存储在底部磁道,从而减少磁道重写的频率. 数据洗牌策略主要分为块级数据迁移和磁道级数据迁移.

为了解决磁道重写过程中性能下降的问题,Hajkazemi等[19]对RMW操作对磁盘性能的影响进行了深入分析. 研究结果指出两条重要发现:首先,对同一底部磁道的多次更新可以通过命令队列或缓存合并成单次写入,来提高磁道重写的效率;其次,为防止系统重启时,数据备份区域内的旧数据被再次拷贝回原始磁道造成数据丢失的问题,磁盘需要采取懒惰或主动的方式清除旧数据. 基于这些关键发现,他们提出了三种数据洗牌机制:磁道翻转机制、动态磁道映射机制以及选择性磁道缓存机制. 这些机制将磁道作为基本管理单位,以此降低内存开销. 具体来说,磁道翻转机制通过交换底部磁道的高频数据与相邻顶部磁道的数据,来优化冷热数据布局;动态磁道映射机制则在预设区域内,允许频繁和不频繁更新的底部与顶部磁道数据进行交换,进一步扩大了冷热数据的交换范围;选择性磁道缓存机制则通过将高频更新磁道的数据迁移到专门的持久缓冲区,以减少数据迁移的成本. 尽管作者尚未详细研究这些新机制之间的协同效应,但它们为未来数据洗牌策略的设计提供了有益的启发.

虽然RMW操作确保了磁道重写时的数据一致性,但该过程引入的额外I/O影响了写入性能. 为了降低磁道重写的开销,Liang等[55]提出了一种名为MoM的更新策略,用于取代RMW操作. 如图10所示,MoM策略的核心思想是,在更新底部磁道之前,将相邻顶部磁道的有效数据迁移至空闲顶部磁道,而不是进行额外的备份操作. 通过这种方式,磁盘无需再将备份数据迁移写回原始磁道,从而减少了1/2的额外I/O次数. 当需要移动顶部磁道的数据时,MoM根据底部磁道的更新频率来选择目标顶部磁道,以尽可能减少数据迁移的频率. 实际上,MoM更适合作为其他数据管理策略的补充,因为其单独部署时对IMR磁盘的性能提升有限. 造成这种问题的根本原因在于,MoM策略的数据洗牌范围仅局限于顶部磁道之间,并未从根本上减少数据迁移的频率,而底部磁道的热数据仍保留在原处. 因此,当没有空闲磁道可用于容纳迁移的数据时,RMW操作将被重新激活.

在磁道级数据洗牌过程中,即使磁道仅包含少量热数据块,数据迁移也会涉及大量冷数据块,这降低了数据迁移的效率. 块级数据混洗策略能够更准确地识别热点数据块,从而提高每次交换的效率. 与磁道级数据洗牌策略相比,这种方法需要消耗更多内存来存储元数据,并且可能会产生更多的寻道开销. 为此,Wu



(*a*) 读-改-写更新策略

(*b*) Move-on-Modify更新策略

图10　典型的RMW操作和MoM策略的更新操作对比示意图[55]

等[47]提出了两种块级数据洗牌策略:Top-Buffer和Block-Swap,以平衡数据迁移效率和内存开销. Top-Buffer能够利用未分配的顶部磁道来缓存底部磁道的更新数据,从而减少磁道重写频率. 当磁盘空间占用率提升,这些缓冲区空间被逐渐归还给用户存储新数据. 为了解决由缓冲区容量不足引发的性能下降,TrackLace设计了Block-Swap机制,通过将底部磁道的热数据块主动交换至顶部磁道来适应不同的工作负载. 此外,TrackLace还提出了虚拟帧和自适应缓冲区这两种优化措施,通过限制数据块交换范围和自适应的"打开"和"关闭"缓冲区,有效减少了数据迁移的寻道开销和缓冲区的维护开销. Top-Buffer和Block-Swap提升了数据洗牌效率,但也破坏了逻辑数据块的连续性,导致顺序I/O请求被拆分为多个子请求,从而增加了响应延迟.

为了提高数据洗牌的效率,Liang等[56]在其所设计的数据管理框架Magic中引入了一个最小化重写代价的持久缓冲区. Magic将磁盘外围的少量底部磁道用于构建块级持久缓冲区,并以循环队列管理数据块. 引发写干扰效应的数据被首先放置在该缓冲区,随后再按照FIFO方式迁移到所属的底部磁道. 在驱逐数据块的过程中,同一目标磁道的所有数据会被一起写回,以减少写放大率. 为了进一步提升数据洗牌的效率,Magic还提出了读-换-写技术. 具体来说,如果被驱逐的数据块属于热磁道,那么Magic会将该磁道与相邻冷顶部磁道进行交换,以防止未来磁道重写操作. 但该技术的有效性很大程度上依赖于Magic对热数据的识别效率,若无法准确识别热磁道,可能导致低效的磁道交换操作,从而影响性能稳定性.

尽管持久缓冲区能够减少磁道重写频率,但由于其容量有限,频繁的清理操作会降低磁盘性能. 为了应对这一问题,Zhang等[102]提出了名为Balloon的数据管理框架,旨在优化数据放置与数据洗牌的协作效率. Balloon采取了类似于Magic的方式,将部分磁道用于构

建持久缓冲区. 不同的是,Balloon能够主动将未被使用的顶部磁道作为磁道级缓冲区,从而缓解持久缓冲区的写入压力,并为实现无成本的数据迁移创造机会. 因此,Balloon同时包含了磁道级和块级的数据洗牌策略. 为了最大程度减少数据迁移对写请求的影响,Balloon实现了按需的数据洗牌策略. 当需要回收缓冲磁道时,它会根据磁道的访问热度和更新块覆盖率来确定数据迁移的目标区域. 同时,Balloon能够根据空间占用率动态调整缓冲区的准入限制,并结合热数据重定向技术,避免因频繁缓冲区清理而产生的磁道重写.

目前,关于IMR磁盘的研究正处于兴起阶段,主要集中在数据放置和数据洗牌策略的设计. 前者旨在将数据放置在重写概率低的磁道上,这可能会破坏数据的空间局部性,增加大型请求的响应延迟;后者则是在数据写入一段时间后,利用记录的历史访问信息,重新调整冷热数据的分布,但在写密集型场景下,数据迁移过程可能会对系统性能产生干扰. 实际上,数据放置和数据洗牌是从不同角度减少磁道重写的频率,如何优化这两种策略的协同效果来实现更高的性能收益,将成为研究者的重要课题.

3.2 系统级和应用级I/O性能优化

3.2.1 面向SMR和IMR磁盘系统级优化方案

(1)面向SMR和IMR磁盘文件系统

与设备级优化方案相比,文件系统在感知磁盘块使用情况这方面更具优势,能够获取更多工作负载的语义信息,因此可以进行深层次的性能优化. 如表6所示,基于SMR磁盘的文件系统优化重点集中在降低元数据更新开销、优化数据布局、减少长尾延迟和元数据回写频率这四个方面. 鉴于文件系统元数据的更新开销是影响SMR磁盘性能的主要因素,因此Le等[103]提出了SFS方案. 该方案将文件元数据存于随机zone,而文件数据则存储于顺序zone. 但对于小文件工作负载,可能加剧碎片化问题. 为此,Jin等[104]提出了HiSMRfs,通过分离元数据和文件数据,并且利用缓存算法,根据文件数据的大小和访问频率进行冷热数据分离来提高系统性能. Zhou等[15]设计并实现了一种名为SMRSTORE的用户空间存储引擎,用于直接管理HM-SMR的地址空间. SMRSTORE采用了自定义的日志格式组织元数据和数据,并结合三级内存数据结构来维护块和记录之间的映射关系,从而实现了高效的读写操作. 不仅如此,其特有的zone分配策略能够根据对象存储服务流的特征分离和优化数据放置,减少GC对用户请求的干扰. 与此同时,Xu等[40]设计的AMFS使用高性能SSD存储元数据,并在空闲时间主动回收缓存空间,减少了长尾延迟的发生. Aghayev等[52]基于Ext4提出了改进的Ext4-lazy,通过在日志中使用jmap映射管理元数据,以减少回写过程中的随机写次数.

针对IMR磁盘写放大问题,现有设备级解决方案无法充分理解文件系统层面的数据语义,因此对性能的提升有限. 为了克服这一限制,研究者开始尝试面向IMR磁盘的文件系统设计. 例如,Lien等[25]深入研究了文件系统数据的特点,并提出了一种文件系统感知的IMR磁盘数据管理方案FSIMR. FSIMR通过分析文件的目录结构和数据热度,来实现目录文件、文件元数据和内容数据的合理布局. 同时,该方案利用异地更新策略来减少RMW操作频率,从而提高了性能稳定性. 此外,FSIMR还设计了一种基于目录结构的垃圾回收机制,能够高效地回收无效数据和释放存储空间,进一步减少RMW操作成本.

表6 基于SMR磁盘的文件系统优化策略分类与总结

优化方向	系统/算法	设备类型	核心思想
降低元数据更新开销	SFS[103]	DM-SMR	随机zone存储元数据,顺序zone存储文件数据
	HiSMR[104]	HM-SMR	元数据与热文件数据存储在非瓦状zone中,降低更新开销
优化数据布局	FSIMR[25]	IMR	根据文件目录和数据热度,合理分配目录文件,文件元数据及内容数据
	SMRSTORE[15]	HM-SMR	自定义日志格式组织数据,建立三级内存数据结构维护映射
减少长尾延迟	AMFS[40]	HA-SMR	利用空闲窗口主动释放缓存空间,并限制数据转存大小,降低清理开销
减少冗数据回写频率	Ext4-lazy[52]	DM-SMR	建立映射管理元数据,避免多次回写产生随机写入

(2)面向SMR磁盘的RAID系统

虽然基于SMR磁盘的RAID系统能够提供可观的存储容量,但这也可能进一步加剧写放大问题,进而导致存储系统在处理细粒度写请求时性能下降. 因此,目前研究工作主要集中在减少RAID奇偶校验更新开销上. 表7对现有基于SMR磁盘RAID系统优化策略如何减少奇偶校验开销进行了分类和总结.

Luo等[105]在提出的DVS方案中,通过采用基于段的数据布局管理方法和追加写方式来处理待写入数据. 为了降低奇偶校验更新开销,该方案按照写入模式动态调整条带宽度. 此外,还引入了一种考虑局部性和访问频率的写缓存管理算法,通过优化淘汰策略来提高缓存利用率和减少垃圾回收开销. 为了提高存储性能并降低随机写入开销,Lu等[106]利用SSD作为SMR磁盘的数据缓存,设计了一种混合RAID系统. 为了避免在SMR磁盘中进行更新操作,Le等[107]在提出的

RAID4 SMR方案中,使用额外的CMR磁盘存储奇偶校验和数据. 同时,该方案采用基于映射表和无效块计数的垃圾回收策略,降低了数据迁移和重写开销. 吴坤尧等[38]在所设计的FT-RAID策略中,利用可覆盖写磁道的就地更新特性来更新热校验块. 同时,还提出了一种双层持久缓冲区架构来优化数据块管理,从而提高了系统吞吐量和响应时间. 而Lin等[36]基于HM-SMR实现了RAID5系统,采用文件为基础的混合映射结构,以降低映射成本. 此外,他们还分别提出了文件导向的数据放置策略和频率导向的垃圾回收策略,通过计算延迟奇偶校验,有效降低了数据移动和计算开销.

表7　基于SMR磁盘的RAID系统优化策略分类与总结

优化方向	系统/算法	设备类型	核心思想
降低奇偶校验更新开销	DVS[105]	DM-SMR	根据写入模式自适应调整条带宽度,减少奇偶校验更新次数和延迟
	FT-RAID[38]	DM-SMR	利用可覆盖写磁道更新热校验块,建立双层持久化缓冲区管理数据块
	HSMR-RAID[36]	HM-SMR	建立基于文件的混合映射结构,降低文件与物理条带地址之间映射成本
	混合RAID[106]	DM-SMR	将SSD作为外部援存,降低随机写入开销
	RAID 4SMR[107]	DM-SMR	利用HDD存储奇偶校验和更新数据,避免SMR产生更新操作

3.2.2　面向SMR和IMR磁盘应用级优化方案

(1)面向SMR和IMR磁盘键值存储系统

基于日志结构合并树(Log-Structured Merge tree, LSM-tree)的键值存储方案因其独特的写缓冲机制和分层结构,能够有效将随机写操作转化为顺序写操作,这与SMR磁盘的顺序写约束高度匹配. 然而,为了维护数据的有序性和清除过期数据,LSM-tree必须执行"合并"操作. 该操作不仅会产生额外的写开销和不可预测的写放大,还会导致SMR磁盘内部出现碎片化的数据布局,从而加重GC的负担. 如表8所示,当前研究主要集中在基于SMR磁盘的LSM-tree型键值存储系统优化合并策略和降低GC开销.

表8　基于SMR和IMR磁盘的键值存储优化策略分类与总结

优化方向	系统/算法	设备类型	核心思想
优化合并效率	SMRD[108]	HM-SMR	减少LSM-tree层数,降低合并频率
	LWC-tree[109]	HM-SMR	只合并元数据,减少参与合并的数据量
	SEALDB[110]	HM-SMR	扩大合并粒度,减少SMR磁盘的随机读写
	KVIMR[23]	IMR	优化数据布局,减少合并触发的数据迁移
降低GC开销	GearDB[111]	HM-SMR	采用轮询合并算法,去除垃圾回收
	KVSTL[28]	DM-SMR	选择空间占用率最低的分区进行GC,减少数据迁移

为了尽可能避免随机写,文献[108]在提出的SMRDB策略中规定,每个分区中都只存放单个文件,并对LSM-tree的层数从多层简化至两层,同时在层内部允许数据不完全按序分布. 这种方式有效地减少了数据的合并次数. 而在Yao等[109]的研究中,他们设计的LWC-tree只针对元数据执行合并操作,实际的键值对则被追加到文件的尾部,从而显著降低了合并的数据量和额外开销. 进一步地,为了解决传统文件系统中逻辑相邻的文件在SMR磁盘上物理位置上分散的问题,他们提出了SEALDB策略[110]. 该策略通过将待合并的文件聚合为Set,从而使得后续合并时可以把多次细粒度的随机写操作转变为单次粗粒度的顺序写,极大减少了元数据管理开销和随机读写操作的频次. 与SMR磁盘相比,IMR磁盘的顺序写约束更为宽松,但是合并操作仍会触发数据迁移. 为此,Liang等[23]提出了KVIMR方案,通过将不易失效的数据分配给底部磁道,易失效的数据分配给顶部磁道,优化了IMR磁盘上的数据布局,从而有效地降低了触发数据迁移操作的概率.

LSM-tree的顺序写模式导致不同生命周期的数据混杂在同一分区内,从而造成了碎片化的数据布局. 为了减少垃圾回收开销,Yao等[111]提出了GearDB. 考虑到同层级数据具有相似的合并频率,GearDB将文件的来源限制在相同层级. GearDB还进一步推出了齿轮合并算法,确保合并操作的范围和所产生的空间碎片都限制在特定分区内. 基于此,可以自动回收被无效数据占据的分区,避免了显式数据迁移,从而减少GC的负担. 但GearDB的一大挑战是需要重新设计LSM-tree,这增加了实现的复杂性. 相比之下,Chen等[28]基于DM-SMR设计了KVSTL. 该策略利用设备内部的STL来隐藏顺序写约束,而无需对软件系统做出任何修改. 在KVSTL中,每个层次都有专门的分区列表,且预留了一定空间. 因此,策略通过优先回收空间利用率低且无效文件比例高的分区,有效地减少了垃圾回收过程中的数据迁移量.

(2)面向SMR和IMR磁盘的文件安全删除策略

对磁盘存储空间进行多次重写是安全、有效地删

除文件数据的常用方法. 如表9所示,现有基于SMR和IMR磁盘的文件安全删除策略主要从数据布局和冷热数据分离这两个方面进行优化研究.

SMR磁盘的顺序写约束在安全删除时会引起写放大效应,导致性能急剧下降. 为解决这个问题,Chen等[54]提出了FFSD策略. 该策略首先将元数据存放在非叠瓦区域,并引入了自适应保护屏障机制. 随后,将zone分割成较小的保护空间,并以追加更新的方式来代替持久缓存中的随机写操作. 最后,通过文件标识自适应分配存储空间,实现了快速、安全的删除操作. 为了进一步优化元数据在安全删除中带来的性能损耗,Chen等[112]提出了SSSA策略. 该策略采用元数据重定向机制,将文件元数据与待删除文件关联,从而减少了RMW操作次数. 同时,通过动态分配段存储空间,有效地隔离了不同文件的数据,进而降低了写放大. 尽管IMR磁盘已降低了写入惩罚,但磁道重写在进行安全删除时仍会对磁盘性能产生影响. 现有针对IMR写放大问题的数据放置策略,对于安全删除场景并非完全适用. Chen等[24]提出的P-SD策略,通过将磁盘空间分为多个虚拟组并根据文件大小分离冷热数据,来提高随机写入效率. 为了实现安全删除,该策略还引入了一种自适应日志机制,通过将日志流引导至无效或空闲磁道来覆盖待删除的数据. 此外,所设计的双交换机制也能够有效区分冷热数据,从而减少了数据写入和安全删除时的磁道重写频率.

表9 基于SMR和IMR磁盘的文件安全删除优化策略分类与总结

优化方向	系统/算法	设备类型	核心思想
优化数据布局	FFSD[54]	DM-SMR	根据文件标识,自适应分配存储空间
	SSSA[112]	DM-SMR	聚合文件数据与元数据,动态分配段空间
分离冷热数据	P-SD[24]	IMR	采用特定算法分离冷热文件数据,利用写干扰效应实现快速安全删除

4 总结与展望

大数据应用对存储容量的需求与日俱增,但传统磁记录硬盘在提高面密度方面遇到了瓶颈. 随着SMR技术的普及和IMR技术的兴起,这两种高密度磁记录技术预计将在海量数据的存储和管理中发挥更大作用. 本文从SMR和IMR技术的基本原理、基本类型以及国内外研究现状三个维度,详细介绍了关键技术内容和发展趋势. 其中,在国内外研究现状章节中,分别介绍了设备级、系统级以及应用级I/O性能优化方案的典型案例,特别是针对设备级的优化研究进行了深入剖析.

具体而言,早期关于SMR磁盘的研究主要集中在动态映射机制下的数据布局和垃圾回收策略优化,以降低写放大效应. 然而,现代SMR磁盘容量已经达到了数十TB,这种方法极大地增加了内存需求和数据迁移开销,使得仅依靠设计数据布局策略来提升磁盘性能变得愈加困难. 为此,研究者利用静态映射机制的低内存需求优势,通过牺牲磁盘内少量存储空间作为持久缓冲区来减少RMW操作的频率,有效降低了数据迁移规模. 然而,持久缓冲区清理时会阻塞用户的I/O请求,并产生不可预测的寻道开销,导致磁盘在写密集型场景下性能大幅下降. 与此同时,闪存技术的广泛应用使得研究者考虑将具有更优随机读写性能的闪存作为SMR磁盘的高性能缓存,通过内嵌或外置的方式构建混合存储系统,以适应I/O密集型应用场景. 然而,缓存管理算法效率和闪存寿命成为了平衡混合存储系统性能和成本的关键挑战. 此外,对于IMR磁盘的研究尚处于初期阶段,主要关注于磁盘数据布局策略设计. 基于对SMR磁盘研究的启发,研究者在磁道级动态映射机制的基础上,为IMR磁盘设计了一系列数据放置和洗牌策略,优化了数据的冷热分布. 考虑到IMR技术的创新性,研究者还需要进一步探索其他解决方案,以满足不同应用场景对磁盘性能的需求. 与此同时,分区存储模式在SMR磁盘上的成功应用,使得升数据管理在灵活性、安全性和稳定性方面取得显著进步. 随着分区存储模式标准逐渐完善,未来的研究可以围绕以下方向展开.

(1)具备性能隔离的数据布局

随着热辅助磁记录技术的逐渐成熟,将其与SMR技术相结合时,能够进一步提升磁盘容量. 目前,大部分研究聚焦在如何从设备的视角进行数据布局优化,来提升垃圾回收效率. 为了提高存储空间利用率和降低存储成本,一个大容量SMR磁盘可能会被多个租户共享使用. 而同一分区内不同租户对数据的访问模式可能存在差异性,这会在垃圾回收时引起不必要的数据迁移和性能干扰. 分区存储模式为实现数据隔离提供了助力,能够提升数据安全性,但这也给存储空间分配和数据迁移带来了挑战. 因此,如何利用分区存储模式优势设计可动态调整的数据布局方案,以确保资源和性能平衡,有待进一步深入研究.

(2)基于传统分区构建持久缓冲区

先前的研究主要集中于叠瓦式持久缓冲区优化,但受限于硬件的顺序写约束,此类型缓冲区无法支持就地更新操作. 而现有设计多采用追加写模式,影响了清理策略的灵活性和效率. 同时,由于磁盘固件的内存和计算资源有限,多数优化策略难以显著提升磁盘性能,而且也不利于实现工作负载感知的清理调度策略. 最近,部分HM-SMR磁盘开始配备少量的CMR分区,这

为研究者实现主机可控的持久缓冲区清理调度和缓冲区空间分配策略提供了机会. 因此,如何自适应地利用主机提供的计算和管理资源以及设备内配置的CMR分区,设计高效协同的清理策略和清理调度机制是一个有潜力的研究方向.

(3)使用ZNS SSD构建混合存储系统

面向SMR磁盘的混合存储系统研究,通常采用驱动管理型固态盘作为性能层或缓存层. 考虑到闪存的擦写次数限制,这种设计可能难以满足写密集型负载的长期可靠性要求. 此外,传统的缓存管理策略因缺乏对内部数据活动的感知,可能会将I/O请求发送至正在执行垃圾回收的闪存芯片,进而影响性能稳定性. 将闪存集成到SMR磁盘内可以部分解决这一问题,但存在利用率和扩展性的限制. 而基于分区存储模型的ZNS SSD能够减少对垃圾回收的依赖,降低写入和擦除操作频率,提升了硬件的耐用性. 同时,它允许主机参与管理数据布局,深入优化混合存储系统的数据管理策略,确保了I/O性能的稳定性和可预测性. 因此,利用ZNS SSD和HM/HA-SMR构建主机系统完全可控的新型混合存储系统,将成为未来研究的一个重要方向.

(4)利用SMR技术的研究成果优化IMR技术

尽管IMR和SMR技术在磁道重叠方式上存在差异,但它们均受到写干扰导致性能降低的挑战. 考虑到技术相似性,现有针对SMR磁盘的研究可以为IMR磁盘性能优化提供设计参考. 例如,一些研究开始将分区存储的思想融入IMR磁盘数据管理,通过引入“虚拟帧”和“区域”这类概念,以达到数据空间局部性与迁移开销的平衡. 同时,为IMR磁盘配置类似于SMR磁盘的持久缓冲区也被证实可以降低磁道重写代价. 未来,如何充分利用现有SMR磁盘研究成果,针对IMR磁盘特性进行改进与重新设计需要进一步探索.

5　结束语

SMR和IMR技术在尽可能不改变现有磁盘结构的前提下,通过减少磁道间距和结合热辅助磁记录技术,显著提升了磁盘面密度并保持了成本优势. 但这两项技术都遭遇了写干扰引发的磁盘I/O性能下降问题. 本文总结了SMR和IMR技术的技术背景、所面临的挑战、针对这两种磁盘在设备级、系统级和应用级的策略设计,重点分类讨论了主流的设备级性能优化方案的优缺点,为研究者基于现有技术设计新的解决方案提供思路.

目前,SMR磁盘已广泛应用在个人电脑和数据中心,而IMR磁盘还被广泛商业化. 尽管如此,由于两者在磁盘结构和生产工艺上的相似性,SMR磁盘的研究成果可以为IMR磁盘的未来优化提供宝贵参考. 随着分区存储模式的逐渐成熟与完善,研究者在I/O栈的不同层面相继提出了各种解决写干扰问题的解决方案,这些措施极大地促进了SMR磁盘和IMR磁盘的优化和应用. 综上所述,机械硬盘作为存储系统主流的存储设备,进一步优化它们的性能以及设计定制化的应用将为高密度磁盘的推广和存储产业的多元发展注入新动力.

参考文献

[1] BURGENER E, RYDNING J. High data growth and modern applications drive new storage requirements in digitally transformed enterprises[R/OL]. [2024-01-27]. https://www.delltechnologies.com/asset/en-in/products/storage/industry-market/h19267-wp-idc-storage-reqs-digital-enterprise.pdf.

[2] REINSEL D, GANTZ J, RYDNING J. Data age 2025: The evolution of data to life-critical don't focus on big data; focus on the data that's big[R/OL]. [2024-01-27]. https://www. seagate. com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf.

[3] LI Y, CHEN X B, ZHENG N, et al. An exploratory study on software-defined data center hard disk drives[J]. ACM Transactions on Storage, 2019, 15(3): 18.

[4] BHUSHAN B. Current status and outlook of magnetic data storage devices[J]. Microsystem Technologies, 2023, 29 (11): 1529-1546.

[5] WHITE R L, NEWT R M H, PEASE R F W. Patterned media: A viable route to 50 Gbit/in^2 and up for magnetic recording?[J]. IEEE Transactions on Magnetics, 1997, 33(1): 990-995.

[6] ZHU J G, ZHU X C, TANG Y H. Microwave assisted magnetic recording[J]. IEEE Transactions on Magnetics, 2008, 44(1): 125-131.

[7] KRYDER M H, GAGE E C, MCDANIEL T W, et al. Heat assisted magnetic recording[J]. Proceedings of the IEEE, 2008, 96(11): 1810-1835.

[8] CASSUTO Y, SANVIDO M A A, GUYOT C, et al. Indirection systems for shingled-recording disk drives[C]// 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST). Piscataway: IEEE, 2010: 1-14.

[9] Digital Western. Data sheet of ultrastar DC HC680[EB/OL]. (2023-12-01)[2024-01-27]. https://www.westerndigital. com/tools/documentRequestHandler? docPath=/content/dam/doc-library/en_us/assets/public/western-digital/product/data-center-drives/ultrastar-dc-hc600-series/data-sheet-ultrastar-dc-hc680.pdf.

[10] HSU W H, VICTORA R H. Rotated read head design for high-density heat-assisted shingled magnetic recording [J]. Applied Physics Letters, 2021, 118(7): 072406.

[11] GREAVES S J, KANAI Y, MURAOKA H. Shingled

thermally assisted magnetic recording for 8 Tbit/in^2[J]. IEEE Transactions on Magnetics, 2014, 50(11): 3001204.

[12] Seagate Technology. Seagate ships worlds first 8TB hard drives[EB/OL]. (2014-08-26) [2024-01-27]. http://www.seagate.com/about/newsroom/.

[13] BALAKRISHNAN S, BLACK R, DONNELLY A, et al. Pelican: A building block for exascale cold data storage [C]//Proceedings of the 11th USENIX conference on Operating Systems Design and Implementation. New York: ACM, 2014: 351-365.

[14] Dropbox. Extending magic pocket innovation with the first petabyte scale SMR drive deployment[EB/OL]. (2018-06-12) [2024-01-27]. https://blogs.dropbox.com/tech/2018/06/extending-magic-pocket-innovation-withthe-firstpetabyte-scale-smr-drive-deployment/.

[15] ZHOU S, XU E, WU H, et al. SMRSTORE: A storage engine for cloud object storage on HM-SMR drives[C]//Proceedings of the 21st USENIX Conference on File and Storage Technologies. Berkeley: USENIX Association, 2023: 395-408.

[16] HWANG E, PARK J, RAUSCHMAYER R, et al. Interlaced magnetic recording[J]. IEEE Transactions on Magnetics, 2017, 53(4): 3101407.

[17] GRANZ S, ZHU W Z, SENG E C S, et al. Heat-assisted interlaced magnetic recording[J]. IEEE Transactions on Magnetics, 2018, 54(2): 3100504.

[18] GRANZ S, JURY J, REA C, et al. Areal density comparison between conventional, shingled, and interlaced heat-assisted magnetic recording with multiple sensor magnetic recording[J]. IEEE Transactions on Magnetics, 2019, 55(3): 3100203.

[19] HAJKAZEMI M H, KULKARNI A N, DESNOYERS P, et al. Track-based translation layers for interlaced magnetic recording[C]//Proceedings of the 2019 USENIX Conference on Usenix Annual Technical Conference. Berkeley: USENIX Association, 2019: 821-832.

[20] SHAFAEI M, HAJKAZEMI M H, DESNOYERS P, et al. Modeling drive-managed SMR performance[J]. ACM Transactions on Storage, 2017,13(4): 38.

[21] AMER A, HOLLIDAY J, LONG D D E, et al. Data management and layout for shingled magnetic recording[J]. IEEE Transactions on Magnetics, 2011, 47(10): 3691-3697.

[22] AMER A, LONG D D E, MILLER E L, et al. Design issues for a shingled write disk system[C]//2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST). Piscataway: IEEE, 2010: 1-12.

[23] LIANG Y, YANG T Y, YANG M C. KVIMR: Key-value store aware data management middleware for interlaced magnetic recording based hard disk drive[C]//Proceedings of the 2021 USENIX Annual Technical Conference. Berkeley: USENIX Association, 2021: 657-671.

[24] CHEN S H, HUANG K H. Leveraging journaling file system for prompt secure deletion on interlaced recording drives[J]. IEEE Transactions on Emerging Topics in Computing, 2023, 11(3): 619-634.

[25] LIEN Y H, CHEN Y T, CHANG Y H, et al. FSIMR: File-system-aware data management for interlaced magnetic recording[J]. ACM Transactions on Embedded Computing Systems, 2023, 22(5s): 128.

[26] HAJKAZEMI M H, ABDI M, SHAFAEI M, et al. FSTL: A framework to design and explore shingled magnetic recording translation layers[C]//2018 IEEE 26th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS). Piscataway: IEEE, 2018: 40-52.

[27] LIN Y S, LIANG Y P, YEN Y S, et al. Adaptive mode-switching for write-amplification reduction of SMR disks [C]//2023 IEEE 20th International SoC Design Conference (ISOCC). Piscataway: IEEE, 2023: 247-248.

[28] CHEN S H, LIANG Y H, YANG M C. KVSTL: An application support to LSM-tree based key-value store via shingled translation layer data management[J]. IEEE Transactions on Computers, 2022, 71(7): 1598-1611.

[29] AGHAYEV A, SHAFAEI M, DESNOYERS P. Skylight—A window on shingled disk operation[J]. ACM Transactions on Storage, 2015, 11(4): 16.

[30] WU F G, FAN Z Q, YANG M C, et al. Performance evaluation of host aware shingled magnetic recording (HA-SMR) drives[J]. IEEE Transactions on Computers, 2017, 66(11): 1932-1945.

[31] NCITS T10 Technical Committee. Information technology-zoned block commands-2 (ZBC-2)[EB/OL]. (2017-12-15)[2024-01-27]. http://www.t10.org/drafts.htm.

[32] INCITS T13 Technical Committee. Information technology-zoned device ATA command set-2 (ZAC-3) [EB/OL]. (2024)[2024]. http://www.t13.org/Documents/.

[33] Digital Western. Libzbc version 5.13.0[EB/OL]. (2016-02-26)[2024-01-27]. https://github.com/westerndigitalcorporation/libzbc.

[34] Western Digital Corporation. Libzbd[EB/OL]. (2020-05-26) [2024-01-27]. https://github. com/westerndigitalcorporation/libzbd.

[35] WU C F, YANG M C, CHANG Y H. Improving runtime performance of deduplication system with host-managed SMR storage drives[C]//2018 55th ACM/ESDA/IEEE De-

sign Automation Conference (DAC). Piscataway: IEEE, 2018: 1-6.

[36] LIN T Y, CHEN T Y. HSMR-RAID: Enabling a low overhead RAID-5 system over a host-managed shingled magnetic recording disk array[C]//Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing. New York: ACM, 2023: 294-296.

[37] 姚婷. 基于新型存储器件的键值存储系统性能优化研究[D]. 武汉: 华中科技大学, 2020.
YAO T. Research on High Performance Key-value Stores for Systems with Emerging Storage Devices[D]. Wuhan: Huazhong University of Science and Technology, 2020. (in Chinese)

[38] 吴坤尧, 柴云鹏, 张大方, 等. 一种新型瓦记录磁盘的高可靠数据存储方法[J]. 软件学报, 2022, 33(12): 4851-4868.
WU K Y, CHAI Y P, ZHANG D F, et al. Highly reliable data storage method based on novel shingled magnetic disks[J]. Journal of Software, 2022, 33(12): 4851-4868. (in Chinese)

[39] YANG M C, CHANG Y H, WU F G, et al. On improving the write responsiveness for host-aware SMR drives[J]. IEEE Transactions on Computers, 2019, 68(1): 111-124.

[40] XU P, WAN J G, HUANG P, et al. An active method to mitigate the long latencies for host-aware shingle magnetic recording drives[C]//2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS). Piscataway: IEEE, 2019: 17-26.

[41] ZHANG B Q, YANG M H, XIE X C, et al. Idler: I/O workload controlling for better responsiveness on host-aware shingled magnetic recording drives[J]. IEEE Transactions on Computers, 2020, 69(6): 777-788.

[42] Digital Western. Dynamic hybrid SMR[EB/OL]. (2017-12-13) [2024-01-27]. https://blog. westerndigital. com/dynamic-hybrid-smr/.

[43] Google. Dynamic Hybrid-SMR: An OCP proposal to improve data center disk drives[EB/OL]. (2017-12-13) [2024-01-27]. https://blog. google/products/google-cloud/dynamic-hybrid-smr-ocp-proposal-improve-data-center-disk-drives/.

[44] WU F G, LI B Z, CAO Z C, et al. ZoneAlloy: Elastic data and space management for hybrid SMR drives[C]//Proceedings of the 11th USENIX Conference on Hot Topics in Storage and File Systems. Berkeley: USENIX Association, 2019: 2.

[45] WU F G, LI B Z, DU D H C. FluidSMR: Adaptive management for hybrid SMR drives[J]. ACM Transactions on Storage, 2021, 17(4): 32.

[46] 王国华, 杜宏章, 吴凤刚, 等. 高密度磁记录技术研究综述[J]. 计算机研究与发展, 2018, 55(9): 2016-2028.
WANG G H, DU H Z, WU F G, et al. Survey on high density magnetic recording technology[J]. Journal of Computer Research and Development, 2018, 55(9): 2016-2028. (in Chinese)

[47] WU F G, LI B Z, ZHANG B Q, et al. TrackLace: Data management for interlaced magnetic recording[J]. IEEE Transactions on Computers, 2021, 70(3): 347-358.

[48] BUCY J S, JIRI S, SCHLOSSER S W, et al. The Disksim Simulation Environment Version 4.0 Reference Manual [M]. Pittsburgh: Carnegie Mellon University, 2008.

[49] WU F G, ZHANG B Q, CAO Z C, et al. Data management design for interlaced magnetic recording[C]//Proceedings of the 10th USENIX Conference on Hot Topics in Storage and File Systems. Berkeley: USENIX Association, 2018: 14.

[50] Digital Western. Dm-zoned[EB/OL]. (2018-05-14)[2024-01-27]. http://github.com/westerndigitalcorporation/dm-zoned-tools.

[51] ZENG Z M, CHEN X Y, YANG L T, et al. IMRSim: A disk simulator for interlaced magnetic recording technology[C]//IFIP International Conference on Network and Parallel Computing. Cham: Springer, 2022: 267-273.

[52] AGHAYEV A, TS'O T, GIBSON G, et al. Evolving Ext4 for shingled disks[C]//Proceedings of the 15th USENIX Conference on File and Storage Technologies. Berkeley: USENIX Association, 2017: 105-119.

[53] MA C L, SHEN Z Y, WANG J H, et al. Tiler: An autonomous region-based scheme for SMR storage[J]. IEEE Transactions on Computers, 2021, 70(2): 291-304.

[54] CHEN S H, WU C F, YANG M C, et al. A file-oriented fast secure deletion strategy for shingled magnetic recording drives[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2022, 41(8): 2463-2476.

[55] LIANG Y H, YANG M C. Move-on-modify: An efficient yet crash-consistent update strategy for interlaced magnetic recording[C]//2021 58th ACM/IEEE Design Automation Conference (DAC). Piscataway: IEEE, 2021: 97-102.

[56] LIANG Y H, YANG M C, CHEN S H. MAGIC: Making IMR-based HDD perform like CMR-based HDD[J]. IEEE Transactions on Computers, 2022, 71(3): 643-657.

[57] Storage Zoned. Linux zoned storage support overview [EB/OL]. (2019.06.14) [2024-01-27]. https://zonedstorage.io/docs/linux/overview.

[58] LIN C I, PARK D, HE W P, et al. H-SWD: Incorporating hot data identification into shingled write disks[C]//2012

IEEE 20th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems. Piscataway: IEEE, 2012: 321-330.

[59] JONES S N, AMER A, MILLER E L, et al. Classifying data to reduce long term data movement in shingled write disks[C]//2015 31st Symposium on Mass Storage Systems and Technologies (MSST). Piscataway: IEEE, 2015: 1-9.

[60] CHUANG Y J, CHEN S H, CHANG Y H, et al. DSTL: A demand-based shingled translation layer for enabling adaptive address mapping on SMR drives[J]. ACM Transactions on Embedded Computing Systems, 19(4): 25.

[61] MA L Y, XU L. HMSS: A high performance host-managed shingled storage system based on awareness of SMR on block layer[C]//2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS). Piscataway: IEEE, 2016: 570-577.

[62] HAO J P, CHEN X B, QIAO Y F, et al. On the design of SMR HDD block device driver[C]//2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). Piscataway: IEEE, 2020: 1291-1299.

[63] HALL D, MARCOS J, COKER J. Data handling algorithms for autonomous shingled magnetic recording HDDs[J]. IEEE Transactions on Magnetics, 2012, 48(5): 1777-1781.

[64] HE W P, DU D H C. SMaRT: An approach to shingled magnetic recording translation[C]//Proceedings of the 15th Usenix Conference on File and Storage Technologies. Berkeley: USENIX Association, 2017: 121-133.

[65] SHAFAEI M, DESNOYERS P. Virtual guard: A track-based translation layer for shingled disks[C]//Proceedings of the 9th USENIX Conference on Hot Topics in Storage and File Systems. Berkeley: USENIX Association, 2017: 18.

[66] YANG T M, WU H T, HUANG P, et al. A shingle-aware persistent cache management scheme for DM-SMR disks [C]//2017 IEEE International Conference on Computer Design (ICCD). Piscataway: IEEE, 2017: 81-88.

[67] LIANG Y P, CHEN S H, CHANG Y H, et al. Mitigating write amplification issue of SMR drives via the design of sequential-write-constrained cache[J]. Journal of Systems Architecture, 2019, 99: 101634.

[68] LEE M C, CHANG L P, WU S M, et al. Adaptive write interference management with efficient mapping for shingled recording disks[C]//2019 IEEE 37th International Conference on Computer Design (ICCD). Piscataway: IEEE, 2019: 181-189.

[69] MA C L, SHEN Z Y, WANG Y, et al. Alleviating hot data write back effect for shingled magnetic recording storage systems[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2019, 38(12): 2243-2254.

[70] MA C L, WANG Y, SHEN Z Y, et al. KFR: Optimal cache management with K-framed reclamation for drive-managed SMR disks[C]//2020 57th ACM/IEEE Design Automation Conference (DAC). Piscataway: IEEE, 2020: 1-6.

[71] LIU W G, ZENG L F, FENG D. CLDM: A cache cleaning algorithm for host aware SMR drives[C]//International Conference on Algorithms and Architectures for Parallel Processing. Cham: Springer, 2018: 608-620.

[72] HAJKAZEMI M H, ABDI M, DESNOYERS P. μCache: A mutable cache for SMR translation layer[C]//2020 28th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS). Piscataway: IEEE, 2020: 1-8.

[73] PAN Y G, JIA Z P, SHEN Z Y, et al. Reinforcement learning-assisted cache cleaning to mitigate long-tail latency in DM-SMR[C]//2021 58th ACM/IEEE Design Automation Conference (DAC). Piscataway: IEEE, 2021: 103-108.

[74] MA C L, ZHOU Z K, WANG Y P, et al. MAID-Q: Minimizing tail latency in embedded flash with SMR disk via learning model[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2022, 41 (11): 3709-3720.

[75] SHEN Z Y, YANG Y H, PAN Y G, et al. A multiagent reinforcement learning-assisted cache cleaning scheme for DM-SMR[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023, 42(8): 2500-2513.

[76] CHERUBINI G, KIM Y, LANTZ M, et al. Data prefetching for large tiered storage systems[C]//2017 IEEE International Conference on Data Mining (ICDM). Piscataway: IEEE, 2017: 823-828.

[77] KAKOULLI E, KARMIRIS N, HERODOTOU H. OctopusFS in action: Tiered storage management for data intensive computing[J]. Proceedings of the VLDB Endowment, 2018, 11(12): 1914-1917.

[78] KOUGKAS A, DEVARAJAN H, SUN X H. Hermes: A

heterogeneous-aware multi-tiered distributed I/O buffering system[C]//Proceedings of the 27th International Symposium on High-Performance Parallel and Distributed Computing. New York: ACM, 2018: 219-230.

[79] CAO Z C, WEN H, GE X Z, et al. TDDFS: A tier-aware data deduplication-based file system[J]. ACM Transactions on Storage, 2019, 15(1): 4.

[80] GE X Z, CAO Z C, DU D H C, et al. HintStor: A framework to study I/O hints in heterogeneous storage[J]. ACM Transactions on Storage, 2022, 18(2): 18.

[81] ZHOU K, ZHANG Y, HUANG P, et al. Efficient SSD cache for cloud block storage via leveraging block reuse distances[J]. IEEE Transactions on Parallel and Distributed Systems, 2020, 31(11): 2496-2509.

[82] LIN H D, LI J, SHA Z B, et al. Adaptive management with request granularity for DRAM cache inside nand-based SSDs[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023, 42(8): 2475-2487.

[83] YANG T W, POLLEN S, UYSAL M, et al. CacheSack: Admission optimization for google data center flash caches[C]//Proceedings of the 2022 USENIX Annual Technical Conference. Berkeley: USENIX Association, 2022: 1021-1036.

[84] SUN H, DAI S S, HUANG J Z, et al. DAC: A dynamic active and collaborative cache management scheme for solid state disks[J]. Journal of Systems Architecture, 2023, 140: 102896.

[85] YANG J C, ZHANG Y Z, QIU Z Y, et al. FIFO queues are all you need for cache eviction[C]//Proceedings of the 29th Symposium on Operating Systems Principles. New York: ACM, 2023: 130-149.

[86] LUO D, WAN J G, ZHU Y F, et al. Design and implementation of a hybrid shingled write disk system[J]. IEEE Transactions on Parallel and Distributed Systems, 2016, 27(4): 1017-1029.

[87] WANG C, WANG D, CHAI Y, et al. Larger cheaper but faster: SSD-SMR hybrid storage boosted by a new SMR-oriented cache framework[C]//Proceedings of the 33rd IEEE International Conference on Massive Storage Systems and Technology. Piscataway: IEEE, 2017: 1-16.

[88] ZHANG C, NIE S Q, WANG J Y, et al. MCB: A multidevice cooperative buffer management strategy for boosting the write performance of the SSD-SMR hybrid storage[J]. The Journal of Supercomputing, 2023, 79(12): 13462-13489.

[89] SUN D S, CHAI Y P. SAC: A co-design cache algorithm for emerging SMR-based high-density disks[C]//Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems. New York: ACM, 2020: 1047-1061.

[90] SUN D S, TAN R X, CHAI Y P. A universal SMR-aware cache framework with deep optimization for DM-SMR and HM-SMR disks[J]. ACM Transactions on Storage, 2023, 19(3): 26.

[91] ZHENG X D, ZHANG C, DUAN K Q, et al. SLA: A cache algorithm for SSD-SMR storage system with minimum RMWs[C]//International Conference on Algorithms and Architectures for Parallel Processing. Cham: Springer, 2022: 587-601.

[92] XIAO W J, DONG H Q, MA L Y, et al. HS-BAS: A hybrid storage system based on band awareness of Shingled Write Disk[C]//2016 IEEE 34th International Conference on Computer Design (ICCD). Piscataway: IEEE, 2016: 64-71.

[93] LIU W G, ZENG L F, FENG D, et al. ROCO: Using a solid state drive cache to improve the performance of a host-aware shingled magnetic recording drive[J]. Journal of Computer Science and Technology, 2019, 34(1): 61-76.

[94] XIE X C, XIAO L Q, DU D H C. ZoneTier: A zone-based storage tiering and caching co-design to integrate SSDs with SMR drives[J]. ACM Transactions on Storage, 2019, 15(3): 19.

[95] MA C L, SHEN Z Y, HAN L, et al. FC: Built-in flash cache with fast cleaning for SMR storage systems[J]. Journal of Systems Architecture, 2019, 98: 214-220.

[96] MA C L, ZHOU Z K, WANG Y P, et al. MU-RMW: Minimizing unnecessary RMW operations in the embedded flash with SMR disk[C]//2022 Design, Automation & Test in Europe Conference & Exhibition (DATE). Piscataway: IEEE, 2022: 490-495.

[97] XIE X C, YANG T Y, LI Q, et al. Duchy: Achieving both SSD durability and controllable SMR cleaning overhead in hybrid storage systems[C]//Proceedings of the 47th International Conference on Parallel Processing. New York: ACM, 2018: 1-9.

[98] MA C L, SHEN Z Y, HAN L, et al. RMW-F: A design of RMW-free cache using built-in NAND-flash for SMR storage[J]. ACM Transactions on Embedded Computing Systems, 18(5s): 65.

[99] CHEN Z G, WANG G H, SHI Z P, et al. Region-based flash caching with joint latency and lifetime optimization in hybrid SMR storage systems[C]//2023 Design, Automation & Test in Europe Conference & Exhibition (DATE). Piscataway: IEEE, 2023: 1-6.

[100] GAO K Z, ZHU W Z, GAGE E. Write management for

interlaced magnetic recording devices: US9508362[P]. 2016-11-29.
[101] GAO K Z, ZHU W Z, GAGE E. Interlaced Magnetic Recording: US9728206[P]. 2017-08-08.
[102] ZHANG C, LIU S, YU F X, et al. Balloon: An elastic data management strategy for interlaced magnetic recording[J]. Applied Sciences, 2023, 13(17): 9767.
[103] MOAL D L, BANDIC Z, GUYOT C. Shingled file system host-side management of Shingled Magnetic Recording disks[C]//2012 IEEE International Conference on Consumer Electronics (ICCE). Piscataway: IEEE, 2012: 425-426.
[104] JIN C, XI W Y, CHING Z Y, et al. HiSMRfs: A high performance file system for shingled storage array[C]// 2014 30th Symposium on Mass Storage Systems and Technologies (MSST). Piscataway: IEEE, 2014: 1-6.
[105] LUO D, YAO T, QU X Y, et al. DVS: Dynamic variable-width striping RAID for shingled write disks[C]// 2016 IEEE International Conference on Networking, Architecture and Storage (NAS). Piscataway: IEEE, 2016: 1-10.
[106] LU Z W, ZHOU G. Design and implementation of hybrid shingled recording RAID system[C]//2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th International Conference on Pervasive Intelligence and Computing, 2nd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech). Piscataway: IEEE, 2016: 937-942.
[107] LE Q, AMER A, HOLLIDAY J. RAID 4SMR: RAID array with shingled magnetic recording disk for mass storage systems[J]. Journal of Computer Science and Technology, 2019, 34(4): 854-868.
[108] PITCHUMANI R, HUGHES J, MILLER E L. SMRDB: Key-value data store for shingled magnetic recording disks[C]//Proceedings of the 8th ACM International Systems and Storage Conference. New York: ACM, 2015: 1-11.
[109] YAO T, WAN J G, HUANG P, et al. Building efficient key-value stores via a lightweight compaction tree[J]. ACM Transactions on Storage, 2017, 13(4): 29.
[110] YAO T, TAN Z H, WAN J G, et al. SEALDB: An efficient LSM-tree based KV store on SMR drives with sets and dynamic bands[J]. IEEE Transactions on Parallel and Distributed Systems, 2019, 30(11): 2595-2607.
[111] YAO T, WAN J G, HUANG P, et al. GearDB: A GC-free key-value store on HM-SMR drives with gear compaction[C]//Proceedings of the ACM Turing Award Celebration Conference - China 2023. New York: ACM, 2023: 51-52.
[112] CHEN P X, CHEN S H, CHANG Y H, et al. Facilitating the efficiency of secure file data and metadata deletion on SMR-based Ext4 file system[C]//2021 26th Asia and South Pacific Design Automation Conference (ASP-DAC). Piscataway: IEEE, 2021: 728-733.

作者简介

伍卫国 男,1963年出生于江西安福. 现为西安交通大学计算机科学与技术学院教授、博士生导师. 主要研究方向为高性能计算机体系结构、海量存储系统、云计算与嵌入式系统.
E-mail: wgwu@mail.xjtu.edu.cn

张 驰 男,1990年7月出生于河南安阳. 获得西安交通大学工学博士学位. 主要研究方向为计算机体系结构、海量存储系统以及新兴大容量存储器件性能优化等.
E-mail: chi.zhang@stu.xjtu.edu.cn

于芳星 男,1995年出生于河南鹤壁. 现为西安交通大学电信学部计算机科学与技术学院博士研究生. 主要研究方向为IMR磁盘性能优化、文件系统以及安全删除技术等.
E-mail: fangxingyu@stu.xjtu.edu.cn

聂世强 男,1993年出生于河南信阳. 现为西安交通大学计算机科学与技术学院助理教授. 主要研究方向为计算机体系结构、非易失性存储介质性能和可靠性优化等.
E-mail: shiqiang.nie@xjtu.edu.cn

李孟涵 男,2000年1月出生于辽宁葫芦岛. 现为西安交通大学计算机科学与技术学院硕士研究生. 主要研究方向为基于新型存储设备的键值存储优化.
E-mail: lmh13052690592@stu.xjtu.edu.cn

牛 洁 女,2000年9月出生于山西太原. 现为西安交通大学软件学院硕士研究生. 主要研究方向为非易失性存储介质性能和可靠性优化.