Skylight—A Window on Shingled Disk Operation

郑小川

摘要

我们介绍了Skylight，一种结合软件和硬件技术对驱动器管理型叠瓦式磁记录驱动器的关键特性进行逆向工程的方法。Skylight的软件部分通过测量I/O操作的延迟时间推断驱动器管理型叠瓦式磁记录盘的重要属性，其中包括驱动器类型、持久缓存结构及其大小、块映射、带的大小、清理算法、区域结构。Skylight的硬件部分在这些测试中使用高速摄像机从一个开在驱动器盖子上的观测窗口跟踪驱动器头部的运动，通过观测驱动器在进行I/O操作时驱动器头部的位置对叠瓦式磁记录驱动器的硬件结构进行分析。Skylight的观察结果证实了其从测量得出的推论，并且归纳总结出了叠瓦式磁记录驱动器的特性及工作原理。我们根据Skylight特征化所归纳总结出的叠瓦式磁记录驱动器特性及工作原理，模拟复现一台仿真叠瓦式磁记录驱动器。

关键词：叠瓦式磁记录、 逆向工程

1 引言

背景：硬盘驱动器是目前计算机存储系统的支柱。尽管现今NAND基于闪存的固态硬盘在性能方面给磁盘的地位带来了巨大的冲击，但磁盘在总比特容量和比特单位成本方面都比闪存更具有优势，如果密度继续以目前的速度提高，这一优势将持续下去。最近磁盘容量的增长是垂直磁记录改进的结果，它通过在70纳米宽的磁道中实现短至20纳米的比特，产生了tb级的驱动器，但进一步增加将需要新的技术[2]。叠瓦式磁记录是第一种应运而生并成功进入到市场的一种技术。叠瓦式磁记录空间像屋顶上的一排排瓦片重叠，从而提高了每个盘片上的比特容量及其密度。但这种密度的增加带来了管理上的复杂性。因为每当需要修改一个磁盘扇区便会破坏重叠轨道上的其他数据，必须进行复制以避免数据丢失。叠瓦式磁记录驱动器到目前为止通过实现隐藏了这种复杂性的叠瓦翻译层来保持与现有驱动器的兼容性[3][4]。

动机：与固态硬盘一样，叠瓦式磁记录驱动器将异地写入与动态映射相结合，以便有效地更新数据。但由于叠瓦式磁记录驱动器修改数据带来的数据保护开销，导致驱动器的性能与常规磁记录驱动器有很大不同。然而，与固态硬盘不同的是，叠瓦式磁记录驱动器及其转换层的行为和性能，以及如何优化文件系统、存储阵列和应用程序，以最佳地使用它们，人们知之甚少[5][6]。另外，固态硬盘加叠瓦式磁记录混合架构是当前存储方面比较具有研究可行性的技术，但这一技术还未进入市场，当缺少实际设备时对这一项技术的研究将缺少实验评估，这不利于我们对它的研究。

选题意义：Skylight介绍了一种测量和表征这种驱动器的方法，为这一表征过程开发了一系列特定的微观基准。Skylight通过一种新技术增强了这些计时测量，该技术通过高速摄像机和图像处理

跟踪实际头部运动，在计时结果不明确的情况下提供了可靠的信息来源[1]。然后，Skylight将此方法应用于希捷提供的5tb和8tb 叠瓦式磁记录驱动器，推断叠瓦翻译层算法及其属性，从而对此类驱动器进行归纳总结。通过根据Skylight的工程思想以及其归纳总结的驱动器特性、工作原理复现叠瓦式磁记录驱动器，可以帮助我们更加清楚地了解叠瓦式磁记录驱动器的工作原理、细节、性能属性，并且使得我们在缺少实际物理设备时可以通过仿真实验进行研究的实验评估。

2 相关工作

2.1 明确特征化目标

为了更好地确定特征化方法，skylight首先对特征化目标进行一个描述。Skylight测量的目标是确定叠瓦式磁记录驱动器的特性以及其关键属性，包括：驱动器类型、持久缓存类型及其大小、清理规则、带的大小、块的映射规则以及区域结构

2.2 特征化方法

Skylight的软件部分使用Linux fio命令对磁盘进行微基准测试，通过该测试结果分析驱动器的特性，归纳总结特征化目标。微基准测试主要是获得延迟时间这一数据，通过磁盘本身特性导致的不同延迟时间分析磁盘的特征。

Skylight的硬件部分在驱动器外壳上安装了一个透明的窗口，使用高速摄像机透过该窗口跟踪驱动器头部运动，并使用开源软件ffmpeg对记录进行处理，为每一视频帧生成头部位置值。通过分析头部位置值，可以解析出使用延迟时间对磁盘分析中的模糊点，最终对磁盘有价值的关键特征尽可能全面地进行了特征化

3 本文方法

3.1 本文方法概述

本文根据skylight逆向工程推断出的特征以及其归纳总结的叠瓦式磁记录驱动器工作原理，通过C++代码模拟复现skylight所用叠瓦式磁记录驱动器。主要包括定义叠瓦式磁记录驱动器物理属性如盘片上带的数目、带上的轨道数目、轨道上的扇区数目、磁头的旋转速度。由于本方法对skylight进行创新，模拟实现固态硬盘与叠瓦式磁记录驱动器混合结构，故还包括定义固态硬盘的擦除时间、写入时间、读取时间。另外还需要定义模拟叠瓦式磁记录驱动器中的地址映射结构。模拟定义物理属性后对叠瓦式磁记录驱动器的基本工作原理进行定义实现，包括模拟磁头移动、读写扇区、持久缓存查询、读取持久缓存内页面数据、清除地址映射、修改数据、叠瓦式磁记录驱动器特有的读合并写操作、持久缓存清理。利用代码模拟实现叠瓦式磁记录驱动器，并在模拟实现的叠瓦式磁记录驱动器上进行大量的I/O操作，记录进行I/O操作后的运行时间、阻塞时间、清理时间、平均响应时间、平均清理时间、读写次数、擦除次数、读合并写操作次数。根据记录的这些结果数据，比较skylight特征化结果，评估复现出的叠瓦式磁记录驱动器。

3.2 驱动器模拟

首先，由于仿真驱动器的物理属性是固定不变的，是驱动器操作的基础，故需先对其进行设定。由于本文基于skylight复现，设定物理属性时以skylight所使用的实际设备物理属性为参考，根据skylight实验结果得出的平均结果作为仿真驱动器的属性。

设定完成基本的物理属性后，模拟实现仿真驱动器的基本操作。首先我们需要明确本文复现的采用固态硬盘加叠瓦式磁记录混合架构的仿真驱动器的工作原理。当驱动器接受到I/O操作时，区分对应写操作还是读操作。对于读操作，首先判断数据是否在持久缓存中，若不存在则通过磁头移动到数据所在磁盘区域读取数据。对于写操作，首先判断持久缓存内是否还有可供写入的空间，当持久缓存内存在写入空间时，直接写入即可，由于采用固态硬盘作为持久缓存，这一行为带来的开销相对而言微乎其微。当持久缓存内不存在写入空间时，需要进行持久缓存的清理回收，即将持久缓存的数据通过磁头移动写入到磁盘区域。当写入磁盘区域时，若因为写入会影响到原本数据则需要进行读合并写操作，即先将磁盘带里的数据读出合并后再进行写入，这一行为将带来巨大的时间开销，即叠瓦式磁记录驱动器的性能缺陷。

明确了仿真驱动器的基本操作后，我们根据研究所需考察的属性，对基本操作进行模拟。例如模拟实现仿真驱动器的磁头移动，由于模拟磁头移动主要在于探究其移动所带来的时间开销以及位置变化，故而我们只需根据仿真驱动器的物理属性，在每次使用到磁头移动时，记录下磁头移动所带来的时间开销以及其位置变化。基本流程为根据访问的物理扇区地址，找到其所在带、轨道、扇区位置，并且对比当前的带、轨道、扇区位置，计算时间开销以及记录位置变化。

4 复现细节

4.1 与已有开源代码对比

本文未使用已有开源代码，主要根据skylight逆向工程所得特征化模拟实现叠瓦式磁记录驱动器，并借助skylight特征化方法对复现出的叠瓦式磁记录驱动器进行仿真评估。

4.2 实验环境搭建

本文基于Windows10环境下编码实现，主要实验环境如下

系统：windows10 64位操作系统

处理器：12th Gen Intel(R) Core(TM) i5-12600K   3.70 GHz

运行内存：32GB

存储器：1T固态硬盘

编程语言：C++

编译平台：CLION 2022.2.4

4.3 界面分析与使用说明

如图1所示，run_test函数的第一个参数为数据输入路径，根据实验数据所在实际路径填写，第二个参数为实验结果输出路径。一次选择一个I/O数据集进行I/O操作，并将结果记录到结果文件，对多个数据集实验后分析所得结果。

图1：运行界面示意

4.4 创新点

本文模拟实现的叠瓦式磁记录驱动器区别于一般叠瓦式磁记录驱动器使用磁盘的一部分作为持久缓存，本文模拟实现了一个采用了固态硬盘作为持久缓存的一个固态硬盘加叠瓦式磁记录驱动器的混合架构磁盘。同时，模拟实现的磁盘也可以作为一般叠瓦式磁记录驱动器使用，即仍然使用磁盘的一部分作为持久缓存。

使用固态硬盘作为持久缓存是目前对叠瓦式磁记录驱动器性能进行优化的最新科研成果，主要在于减少叠瓦式磁记录盘驱动器清理持久缓存时的开销，兼顾了成本与性能。本文复现固态硬盘+叠瓦式磁记录混合架构磁盘不仅可以帮助了解一般叠瓦式磁记录驱动器的工作原理以及工作性能，还可以基于目前最新研究成果，发掘混合结构中仍存在的优化点。

5 实验结果分析

如图2所示，本实验使用多个数据集对模拟实现的叠瓦式磁记录驱动器进行I/O测试。限于篇幅，本文仅就实验结果的运行时间与读写操作次数进行分析讨论。

由于叠瓦式磁记录驱动器本身特性，其时间开销主要是其写操作带来的读合并写操作开销，而传统磁盘的时间开销主要是磁头旋转寻道带来的开销，故本结果分析选用4个在写操作次数相近的数据集进行比较分析。由实验结果可以看出，在四个数据集的读操作次数各不相同且最多相差6倍的情况下，其在模拟实现的叠瓦式磁记录驱动器上的运行时间仍然相近。实验结果表示，对于写操作次数相近但读操作次数相差很多的I/O操作，在叠瓦式磁记录驱动器上运行的时间是相近的。这一实验结果与skylight特征化后归纳出的叠瓦式磁记录驱动器特性相符合，即所复现的驱动器能够较好地模拟出叠瓦式磁记录驱动器的工作原理，可以使用本文复现的叠瓦式磁记录驱动器进行其他的基于叠瓦式磁记录驱动器的实验研究。

	A	B	C	D
1	Run_time: 1099117954.561697	Run_time: 1076009743.651034	Run_time: 1199226368.524643	Run_time: 1105186666.361892
2	BLOCK_TIME: 1088272028.646484	BLOCK_TIME: 1064365614.775391	BLOCK_TIME: 1188011658.544922	BLOCK_TIME: 1091474414.593750
3	Cleaning time: 8924653.551725	Cleaning time: 9498942.472494	Cleaning time: 8761602.678125	Cleaning time: 11259636.134928
4	AVG_respond_time: 705.551468	AVG_respond_time: 597.207881	AVG_respond_time: 535.873928	AVG_respond_time: 544.181645
5	AVG_cleaning time: 16650.473044	AVG_cleaning time: 16491.219570	AVG_cleaning time: 16979.850151	AVG_cleaning time: 16830.547287
6	IO_count: 1557814	IO_count: 1801734	IO_count: 2237889	IO_count: 2030915
7	read_count: 176729	read_count: 316692	read_count: 904483	read_count: 308437
8	write_count: 1381085	write_count: 1485042	write_count: 1333406	write_count: 1722478
9	GC_count: 536	GC_count: 576	GC_count: 516	GC_count: 669
10	RMW_band_count: 52552	RMW_band_count: 55918	RMW_band_count: 51583	RMW_band_count: 66259
11	erase_count: 10720	erase_count: 11520	erase_count: 10320	erase_count: 13380
12	total_valid_pages_copy: 109657	total_valid_pages_copy: 121291	total_valid_pages_copy: 118209	total_valid_pages_copy: 144224
13	tl_incur_pos:	tl_incur_pos:	tl_incur_pos:	tl_incur_pos:
14	1281700605472816 78 53119.721875	1281715273538925 24 47970.553906	1281690366337603 25 34303.992969	1281706308494396 20 67631.376172
15	RW_move: 228858.590902	RW_move: 309884.774089	RW_move: 715955.620931	RW_move: 334163.490153
16				
17				

图2：实验结果示意

6 总结与展望

本文根据skylight对叠瓦式磁记录驱动器特征化后得到特性及工作原理，使用代码模拟复现了一个叠瓦式磁记录驱动器，且结合最新的研究成果，模拟复现运用了固态硬盘加叠瓦式磁记录混合架构的叠瓦式磁记录驱动器。这一复现有利于在缺少实际物理设备时，对叠瓦式磁记录驱动器进行研究。特别的，由于固态硬盘加叠瓦式磁记录混合结构是目前最有前景的对叠瓦式磁记录驱动器优化方法且其还未进入市场，模拟复现采用这一架构的叠瓦式磁记录驱动器有利于未来对叠瓦式磁记录驱动器的研究，避免了研究停留在设想但缺乏实验的问题。未来可利用本文复现的叠瓦式磁记录驱动器对目前混合架构在兼顾成本和性能上的优化进行研究，并通过这一仿真驱动器进行研究成果的实验评估。

参考文献

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

[2] S. N. Piramanayagam. Perpendicular recording media for hard disk drives. Journal of Applied Physics, 102(1):011301, July 2007.

[3] Y uval Cassuto, Marco A. A. Sanvido, Cyril Guyot, David R. Hall, and Zvonimir Z. Bandic. Indirection Systems for Shingled-recording Disk Drives. In Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), MSST ’10, pages 1–14, Washington, DC, USA, 2010. IEEE Computer Society.

[4] David Hall, John H Marcos, and Jonathan D Coker. Data handling algorithms for autonomous shingled magnetic recording hdds. IEEE Transactions on Magnetics, 48(5):1777–1781, 2012.

[5] Luc Bouganim, Bjorn Jnsson, and Philippe Bonnet. uFLIP: understanding flash IO patterns. In Proceedings of the Int’l Conf. on Innovative Data Systems Research (CIDR), Asilomar, California, 2009.

[6] Feng Chen, David A. Koufaty, and Xiaodong Zhang. Understanding Intrinsic Characteristics and System Implications of Flash Memory Based Solid State Drives. In Proceedings of the Eleventh

International Joint Conference on Measurement and Modeling of Computer Systems, SIGMETRICS ’ 09, pages 181–192, New Y ork, NY , USA, 2009. ACM.