一种自适应数据布局的混合硬盘结构

廖学良[1,2]，白石[1]，朱龙云[2]
(1. 清华大学计算机科学与技术系，北京 100084;
2. 可视媒体智能处理与内容安全北京市高等学校工程研究中心，北京 100084)

摘要：提出一种自适应数据布局的混合硬盘结构，其内部包含闪存介质和磁介质，而上层提供统一的逻辑地址空间。在混合硬盘内部，地址转换层统一管理闪存介质和磁介质的物理存储空间，并将逻辑地址转换为物理地址。混合硬盘结构采用自适应的方法，将读次数多和写次数多的数据分别分布在闪存和磁介质上，从而综合利用了闪存和磁介质的不同优点。

关键词：混合硬盘；自适应数据布局；地址转换
中图分类号：TP391 **文献标识码：**A

A hybrid disk structure with adaptive data distribution

Liao Xueliang[1,2], Bai Shi[1], Zhu Longyun[2]
(1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. Beijing Higher Institution Engineering Research Center of Intelligent Processing of Visual Media and Content Security, Tsinghua University, Beijing 100084, China)

Abstract: We propose a hybrid disk structure, which contains flash medium and magnetic medium, and provides unified logical address space to upper layer such as file system. An address translation layer inner the hybrid disk manages the physical addresses of the flash and magnetic medium, and translates logical address to physical address. This structure can make use of the advantages of flash and magnetic medium, and adaptively distribute the frequently read and written data to flash medium and magnetic medium respectively.

Key words: hybrid disk; adaptive data distribution; address translation

0 引言

磁盘是目前使用最普遍的外存介质，其容量不断提升，接口速度不断提高，内部控制单元的性能也逐渐进步。然而，由于其内部寻址时依靠固有的、速度较慢的磁头以及盘片的机械运动，磁盘和CPU、内存之间的速度差异也在不断增大，因此，外存访问速度慢的问题成为计算机系统发展的主要瓶颈之一。近年来，随着闪存技术的进步，闪存容量增大速度很快，以NAND闪存为介质的固态硬盘（solid state disk，SSD）得到了较多的应用。与磁盘相比，固态硬盘具有随机访问速度快、抗震动、能耗低等优点。但是由于闪存写前必须擦除且擦除次数有限，固态硬盘也因此存在擦除时速度慢和写操作次数有限等缺点。对固态硬盘的研究已经成为外存领域研究的热点。另一方面，磁盘具有本地更新、写次数大、顺序访问速度快的特点。为了综合利用磁盘和固态硬盘的优点，提出了一种混合硬盘结构。在混合硬盘结构中存在2种介质：闪存介质和磁介质。混合硬盘对外依然支持SATA等接口访问，并为文件系统提供统一的地址空间（本文称之为逻辑地址空间）。混合硬盘的内部通过一个控制单元统一管理闪存介质和磁介质（闪存介质和磁介质的地址空间称为物理地址空间）。在工业上，已有包含闪存介质和磁介质混合硬盘结构（例如三星公司的混合硬盘[1]中，将闪存介质作为磁介质的缓存），因此将2种介质放在一个硬盘中的设计在工业上是可行的。然而，已有的混合硬盘结构中闪存介质和磁介质不处于同等的地位（例如磁盘作为闪存的缓存[2]或闪存作为

基金项目：国家自然科学基金(61103020)
作者简介：廖学良(1984－)，男，博士研究生，主要研究方向：操作系统和存储系统，E-mail：liaoxl@oslab.cs.tsinghua.edu.cn

磁盘的缓存[3-4])。本文提出的混合硬盘结构中，闪存介质和磁介质处于同等的地位，它们都是外存的存储介质。对闪存介质的访问采用与固态硬盘相同的方法，即访问固态电子存储芯片阵列；对磁介质的访问采用与磁盘相同的方法，即磁头移动和盘片转动。与固态硬盘类似，在混合硬盘中也存在一个地址转换层。然而，混合磁盘中地址转换层所管理的地址包括闪存介质和磁介质二者的物理地址空间和对外提供的逻辑地址空间。采用地址转换层的好处是可以根据数据访问的特点，将数据在不同介质之间，或在同一介质的不同物理地址之间进行迁移，同时不影响对外提供的接口和更上层的文件系统。混合硬盘采用自适应的数据迁移机制，将读次数多的数据迁移到闪存介质上，将写次数多的数据迁移到磁介质上。

1 相关工作

下面介绍综合利用固态硬盘和磁盘的相关研究工作。

PB-PDC方法[5]针对一个固态硬盘和一个磁盘的组织方式，通过记录文件的读写次数，将频繁读的文件迁移到固态硬盘，将频繁写的文件迁移到磁盘。SAIL方法[6]中分析了在多个固态硬盘和磁盘的结构下如何对数据进行布局，并对性能和能量消耗进行了建模分析。PDC-NH[7]将大的、顺序访问的文件存于磁盘，而将访问频率高的数据存于带有固态硬盘缓存的磁盘中。文献[8]提出了将固态硬盘和磁盘统一管理的结构，其中固态硬盘存储逻辑地址小的数据，而磁盘存储逻辑地址大的数据，并根据文件类型和文件访问特点开发了文件布局的工具。文献[9]提出了将固态硬盘和磁盘统一管理并向上层提供虚拟硬盘的结构。与上述工作相比，笔者提出的混合硬盘结构并不是用固态硬盘和磁盘，而是使用它们的存储介质：闪存和磁介质，而且磁介质和闪存介质处于同等的地位。混合硬盘结构将这2种介质包装在一起，在一个硬盘内部实现闪存介质和磁介质的并存，对外提供和现有硬盘同样的接口，因此不会影响上层结构，也不需要使用专门的工具进行文件的重新布局。

在上述方法中，尽管固态硬盘和磁盘的特性不同，但二者处于同等的地位，即二者都是作为普通的外存设备使用，只是基于二者不同的特性，试图在其上分布不同的文件或者数据。还有一些研究工作也是基于二者的不同特性，但二者处于不同等的地位，主要分为两类：固态硬盘作为磁盘的缓存；磁盘作为固态硬盘的缓存。由于固态硬盘读取速度快，所以将固态硬盘（或者闪存）作为磁盘的缓存是一种比较自然的结构 [3-4]。同时，这种结构也得到了较多的应用，比如Intel的Turbo Memory技术[10]，Windows的ReadyBoost技术[11]，Windows的ReadyDrive技术[12]，三星的混合硬盘[1]等。而将磁盘作为固态硬盘的缓存是一种非常独特的结构，最著名的研究是Griffin [2]。在Griffin中，将磁盘的使用定位为类似于存储日志的存储介质，因此，对磁盘的写操作都是顺序性的，可以较好地利用磁盘顺序访问性能高的特点。

2 混合硬盘的地址转换

在笔者提出的混合硬盘中，地址转换层统一管理闪存介质和磁介质的物理地址空间，并向上层提供统一的逻辑地址空间。由于闪存的读写单位是 1 个页，整个逻辑地址空间和物理地址空间的读写单位也是 1 个页。在这里，称逻辑地址空间的 1 个页为逻辑页，物理地址空间的 1 个页为物理页。与固态硬盘中的闪存转换层（flash translation layer）一样，本文提出的混合硬盘需要记录每 1 个逻辑地址（即每 1 个逻辑页）的数据在物理介质上的存放地址（即存放在哪 1 个物理页上）。

在本文提出的混合硬盘内部有一个称为地址转换层（address translation layer）的控制单元。地址转换层记录每个逻辑地址所对应介质的物理地址。具体来说，对每 1 个逻辑页号，

有 1 个对应的物理页号，因此，这里的地址转换是页粒度的。物理页号由介质标志和介质上的物理页号两部分组成。其中，介质标志占用 1 bit 数据，例如，该比特为 0 和 1 分别表示闪存介质和磁介质。介质上的物理页号采用与当前使用的磁盘和固态硬盘同样的顺序编址方法。

2.1 地址转换信息的管理

当某一逻辑地址的数据发生以下 4 种情况时，所对应的物理地址发生相应的变化，所以也需要更新地址转换信息：1) 数据从磁介质迁移到闪存介质；2) 从闪存介质迁移到磁介质；3) 在闪存介质上进行写操作；4) 在闪存介质上进行垃圾收集（garbage collection）。其中，前 2 种情况是由混合硬盘的数据迁移机制引起的，后 2 种情况与当前固态硬盘上的数据写操作以及垃圾收集过程一样。值得注意的是，在磁盘介质上写操作不需要更新地址信息，因为此时可以将数据写在旧的磁盘介质上。

地址转换信息的管理采用与操作系统中内存页管理类似的方法，即使用多级页表。假设 1 个页的大小为 2^a byte，逻辑地址空间大小为 2^n byte，每 1 个页表项占用 2^b byte，则 1 个页表可以存储 2^{a-b} 个页表项，所需的多级页表层数为 $n/(a-b)$（向上取整）。例如，当页大小为 4 K byte，每个页表项占用 8byte，逻辑地址空间大小为 4 T byte 时，a=12，b=3，n=42，需要的多级页表层数为 5。

为了页表的快速读取，页表都存储在闪存介质上，并且在 RAM 中保存最近使用的页表。为了适应数据访问的空间局部性，对地址转换信息的读取采用预取机制，将 1 个物理页上的页表都预取到 RAM 中。RAM 中页表的替换使用 LRU 算法，替换的粒度为 1 个页所能存储的页表。

2.2 数据分配和垃圾收集

对于闪存介质，地址转换层维护 1 个当前更新块，用于将需要更新的页写到该块的空闲页上。在闪存介质上进行写操作时，地址转换层试图在当前更新块中分配 1 个空闲物理页。如果没有可以分配的空闲物理页，则试图从系统的空闲块分配 1 个空闲块作为新的当前更新块，并将数据更新到该块的空闲页上。如果在试图分配空闲块时，当前系统中的空闲块少于系统至少需要持有的空闲块数 N_{th}，则启动垃圾收集过程回收一些物理块，并获得一些空闲块，进而满足空闲块分配的需要。在磁介质进行写操作时，将数据写到原来的物理页上。

在闪存上进行垃圾收集时，首先选择有效页最少的数据块，并将其上的有效页依次拷贝到当前更新块上。如果当前更新块中不存在空闲页，则分配 1 个空闲块作为新的当前更新块，最后擦除该数据块。持续上述过程，直到系统中的空闲块数等于 N_{th}。

3 自适应数据迁移

采用混合硬盘最大的优点是可以根据数据的读写访问特点，对数据进行重新布局。数据的读写特点是指数据在一定的时期内，要么没有被访问，要么主要被读，要么主要被写，要么是读、写相间。在笔者设计的混合硬盘结构中，用 1 bit 表示数据最近访问的读写标志，例如，该比特为 0 和 1 分别表示上次操作为读操作和写操作。如果对某页有连续多次读或者写，则将读写次数加 1（如果已经达到读写次数的最大值，则不加 1，而是保持最大值）。如果在多次读之后写，或者在多次写之后读，则将读写标志取反，并将读写次数置 0。因此，结合混合磁盘的地址转换机制，地址转换信息中的页表项具有如图 1 所示的格式。

读写标志(1 bit)	读写次数	介质标志(1 bit)	物理页号

图 1 页表项格式
Fig.1 The format of page table entry

例如，使用 2 bit 即可表示每个数据所在的介质和最近 1 次的访问情况。具体来说，这 2 bit 数据的含义如表 1 所示。

表1 2 bit 表示的数据状态
Table 1 Data State Represented with 2 bits

2 个比特的值	数据的状态
00	数据位于闪存介质，上次访问为读操作
01	数据位于闪存介质，上次访问为写操作
10	数据位于磁介质，上次访问为读操作
11	数据位于磁介质，上次访问为写操作

在混合硬盘中，采用的迁移策略为：当数据被连续读取 *N* 次时，认为数据主要被读，因此此时如果数据在磁介质上，则将其迁移到闪存介质上；当数据被连续写 *N* 次时，认为数据主要被写，因此此时如果数据在闪存上，则将其迁移到磁介质上；当数据为读、写间隔的访问时，不将其进行迁移。当 *N* 为 2 时，数据的迁移及状态转换图如图 2 所示。在图 2 中，每个状态包括 2 bit。其中第 1 bit 为介质标志，第 2 bit 为读写标志。以 01 状态转换到 11 状态为例，当数据位于 01 状态时，表示数据位于闪存介质，而且上次操作为写操作。当 01 状态的数据被再次写时，该数据已经被连续写 2 次，因此需要将其迁移到磁介质上，状态变为 11，表示数据位于磁介质，上次访问为写操作。

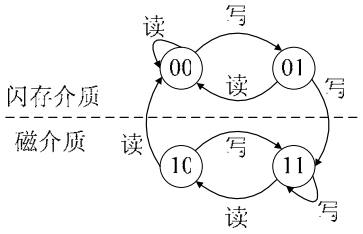


图 2 *N* 为 2 时数据的迁移和状态转换示意图
Fig.2 Data Transition and State Translation when *N* is 2

当数据需要从磁介质迁移到闪存介质时，如果将引起闪存介质的垃圾收集过程，则不进行迁移，而是仅仅从磁盘读取数据，满足上层的请求。当数据需要从闪存介质迁移到磁介质时，如果磁介质已达到容量上限，则也不进行迁移，而是仍然将数据写在闪存介质上。

4 实验结果

实现了一个 trace 驱动的模拟器，用于测试混合硬盘的读写性能和其中闪存介质的擦写情况。测试中使用的地址转换信息存储方式、闪存数据页的分配方式、垃圾收集机制与 DFTL 一样。笔者使用的数据包括在清华大学计算机系一台服务器上收集的 tucs09 和 SNIA[14]及微软提供的 MSN Storage File Sever trace（简称为 msn），Exchange Server trace（简称为 ex），RadiusBackEndSQLServer（简称为 rad），其中 tucs09 包含 1 块硬盘的数据，msn trace 和 ex trace 包含多块硬盘的数据，笔者只测试写操作超过 100 万次的硬盘。在实验中，闪存介质和磁介质的大小设为一样，且 2 种介质上各有 3%的冗余空间。

4.1 读写性能

表2 混合硬盘中在 2 种介质上的读、写页数。R_F，R_D，W_F，W_D 分别表示读取闪存介质、读取磁介质、写闪存介质、写磁介质的页数。

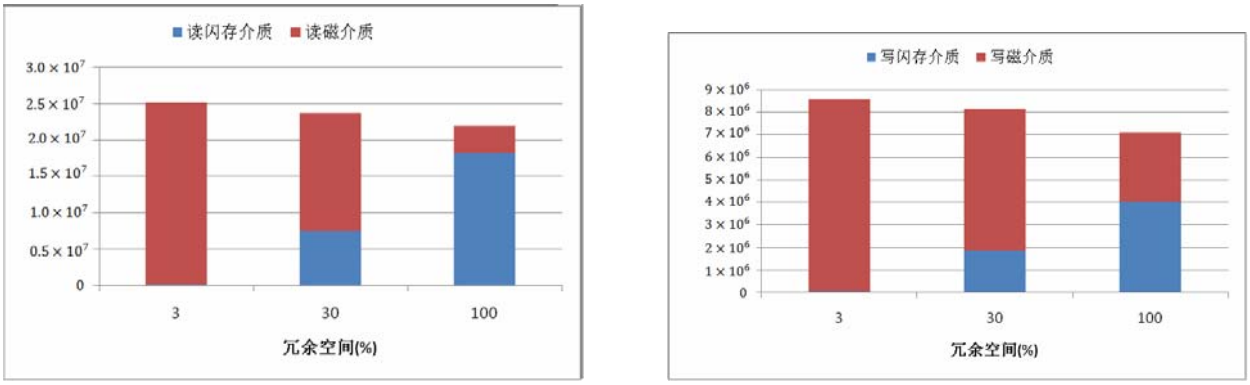
Table 2 Read/Write Number to the two media. R_F: Read Number to Flash，R_D: Read Number to Disk，W_F: Write Number to Flash，W_D: Write Number to Disk.

Trace	R_F (10^4)	R_D (10^4)	W_F (10^4)	W_D (10^4)	$R_F/(R_F+R_D)$	$W_D/(W_F+R_D)$	$(R_F+W_D)/(R_F+R_D+W_F+W_D)$
tucs09	3361.1	15.2	89.6	9949.6	99.5%	99.1%	99.2%
msn_disk0	1.8	47.0	4.0	779.8	3.7%	99.5%	93.9%
msn_disk1	203.1	172.2	148.3	396.7	54.1%	72.8%	65.2%
msn_disk5	1843.3	553.2	181.2	562.9	76.9%	75.6%	76.6%
ex_disk0	83.9	9.9	33.4	2404.5	89.4%	98.6%	98.3%
ex_disk2	698.2	251.1	145.0	893.6	73.5%	86.0%	80.1%
ex_disk3	666.7	268.6	110.3	745.6	71.3%	87.1%	78.8%
ex_disk4	644.8	329.4	130.3	975.1	66.2%	88.2%	77.9%
ex_disk5	10.1	681.	29.4	1529.1	1.5%	98.1%	68.4%
ex_disk6	696.4	374.6	166.8	963.6	65.0%	85.2%	75.4%
ex_disk7	598.3	446.9	149.4	923.8	57.2%	86.1%	71.9%
ex_disk8	511.5	452.0	136.3	1027.2	53.1%	88.3%	72.3%
ex_disk9	16.4	877.6	36.0	1972.5	1.8%	98.2%	68.5%
rad_disk4	15.1	2496.9	3.9	852.4	0.6%	99.5%	25.8%

表 2 列出了对 2 种介质的读写页面情况。从图中可以看到，在大部分 trace 中，R_F 和 W_D 之和占 65%以上，W_D 占写页面数的 70%以上。因此，使用数据迁移方法较好地实现了研究目的。

从表 2 也可以看到，对于 msn_disk0、ex_disk5、ex_disk9、rad_disk4，R_F 占所有读页面数的比例很小，这是因为，这 4 个 trace 都是以写操作为主，很少发生数据从磁介质迁移到闪存介质的情况。

测试了冗余空间大小对读写性能的影响。图 3 表示对于 rad_disk6，在不同冗余空间下的读写次数。从图中可以看到，随着冗余空间的增加，R_F 所占读操作的比例增加。由于读闪存介质的次数增加很多，所以 W_F 也随之增加，但是此时 W_F 占所有写操作的比例并不大，只比 W_D 稍多。例如，当冗余空间为 100%时，W_F 所占的比例为 56.3%。



(a) 读操作　　(b) 写操作

图 3 不同冗余空间时对 rad_disk6 的读写闪存介质、磁介质次数

Fig.3 Read and write number to flash and magnetic media with varying spare area under rad_disk6

4.2 闪存擦除情况

在混合硬盘中，也存在闪存擦除的情况。表 3 列出了在混合硬盘和固态硬盘下，各个 trace 模拟完成时的擦除次数。可以看出，采用混合硬盘结构和采用固态硬盘相比，擦除次数降低率都在 90%以上。因此，混合硬盘中闪存介质的寿命得到了很大提升。

表3 混合硬盘和固态硬盘的擦除次数对比
Table 3 Erase Number of Our Hybrid Disk and SSD.

Trace	混合硬盘	固态硬盘
tucs09	24,681	162,256
msn_disk0	369	415,755
msn_disk1	40,078	125,309
msn_disk5	100,127	1,527,673
ex_disk0	6,463	380,027
ex_disk2	2,201	150,897
ex_disk3	1,467	113,155
ex_disk4	1,955	165,369
ex_disk5	6,204	220,582
ex_disk6	4,037	197,418
ex_disk7	2,105	169,311
ex_disk8	2,567	182,261
ex_disk9	7,245	298,207
rad_disk4	1,355	393,495

5 结束语

提出了一种混合硬盘结构，其内部包含闪存介质和磁介质，并采用自适应的方法，将读次数多的数据和写次数多的数据分别布局到闪存介质和磁介质。实验结果显示该混合硬盘较好地达到了数据布局的目标，混合硬盘内部的闪存寿命也得到了很大的提升。

下一步工作将研究数据如何在磁介质内部布局，以提高访问磁介质的性能。

[参考文献] (References)

[1] SamSung. Samsung Hybrid Hard Drive [EB/OL]. http://www.samsung.com/global/business/semiconductor/support/brochures/downloads/hdd/hdd_datasheet_200708.pdf

[2] Soundararajan G, Prabhakaran V, Balakrishnan M, et al. Extending SSD lifetimes with disk-Based write caches [C]// 8th USENIX Conference on File and Storage Technologies. San Jose, CA, USA, 2010: 101-114.

[3] Michael W, Willy Z. eNVy: a non-volatile, main memory storage system [C]// Proceedings of the sixth international conference on Architectural support or programming languages and operating systems. San Jose, CA, USA, 1994: 86-97.

[4] Taeho K, David R, Trevor M. Improving NAND Flash Based Disk caches [C]// 35th International Symposium on Computer Architecture. Beijing, China, 2008: 327-338.

[5] Kim Y, Kwon K, Kim J. Energy-efficient file placement techniques for heterogeneous mobile storage systems [C]// 6th ACM & IEEE International Conference on Embedded Software. Seoul, Korea, 2006: 171-177.

[6] Xie T, Madathil D. SAIL: Self-adaptive file reallocation on hybrid disk arrays [J]. LNCS 5374, 2008: 529-540.

[7] Lee D, Koh K. PDC-NH: Popular data concentration on NAND flash and hard disk drive [C]// 10th IEEE/ACM International Conference on Grid Computing. Banff, Alberta, Canada, 2009: 196-200.

[8] Payer H, Sanvido M, Bandic Z, et al. Combo drive: Optimizing cost and performance in a heterogeneous storage device [C]// First Workshop on Integrating Solid-state Memory into the Storage Hierarchy. Washington DC, USA, 2009: 1-8.

[9] Jo H, Kwon Y, Kim H, et al. SSD-HDD-Hybrid virtual disk in consolidated environments [J]. LNCS, 2010, 6043: 375-384.

[10] Matthews J, Trika S, Hensgen D, et al. Intel turbo memory: Nonvolatile disk caches in the storage hierarchy of mainstream computer systems [J]. ACM Transactions on Storage, 2008, 4(2): 1-24.

[11] Microsoft Corporation. Microsoft Windows Ready-Boost [EB/OL]. http://www.microsoft.com/windows/windows-vista/features/readyboost.aspx

[12] Panabaker, Ruston. Hybrid Hard Disk and ReadyDrive Technology: Improving Performance and Power for Windows Vista Mobile PCs [EB/OL]. http://www.microsoft.com/whdc/system/sysperf/accelerator.mspx

[13] Gupta A, Kim Y, Urgaonkar B. DFTL: A flash translation layer employing demand-based selective caching of page-level address mappings [C]// Proceeding of the 14th international conference on Architectural support for programming languages and operating systems. Washington, DC, USA, 2009: 229-240.

[14] SNIA. Storage networking industry association [EB/OL]. http://www.snia.org/home/