计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20200690
58(7):1518-1532,2021

闪存固态硬盘系统结构与技术

高聪明[1] 石 亮[2] 刘 凯[3] 薛 春[4] 舒继武[1]
[1](清华大学计算机科学与技术系 北京 100083)
[2](华东师范大学计算机科学与技术学院 上海 200062)
[3](重庆大学计算机学院 重庆 400044)
[4](香港城市大学计算机系 香港 999077)
(gaocm92@gmail.com)

Architecture and Technologies of Flash Memory Based Solid State Drives

Gao Congming[1], Shi Liang[2], Liu Kai[3], Xue Chun[4], and Shu Jiwu[1]
[1](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100083)
[2](*School of Computer Science and Technology, East China Normal University, Shanghai* 200062)
[3](*College of Computer Science, Chongqing University, Chongqing* 400044)
[4](*Department of Computer Science, City University of Hong Kong, Hongkong* 999077)

Abstract Flash memory based solid state drives (SSDs) are widely deployed in personal computers, data centers, and cloud storages given their well identified advantages, such as high performance, low power consumption and non-volatile property. In recent years, with the development of process technology and micro electronic technology, the features of SSDs are greatly changed. First, detailed storage mechanism of flash cell is introduced, including architecture of flash cell and flash block, programming method, and basic operation of SSDs. Then, several SSD controller key technologies are presented, including buffer device, flash translation layer, garbage collection, data allocation, wear leveling and error correction code. These technologies are used to well support normal operation of SSDs. Additionally, the parallel architecture of SSDs which is used to boost the performance of SSDs, is discussed and related constraints are also presented while several previous works on parallelism exploration are analyzed. Next, since the scaling of SSD has evolved from planar (2D) to 3D stacking, 3D SSD is introduced as a new type of SSDs that can provide larger capacity compared with traditional planar SSD. In this paper, the characteristics of 3D SSDs' vertical architecture, performance and lifetime are analyzed. Also, the disadvantages of previous works on 3D SSD performance and lifetime optimizations are discussed. Finally, current state of SSDs is summarized and possible future research works are given.

Key words flash memory; solid state drives; buffer; flash translation layer; parallelism; 3D solid state drives

摘 要 闪存固态硬盘凭借其高性能、低功耗、非易失等特点已经被广泛应用于个人电脑、数据中心和云存储服务等.近10年来,随着制程工艺和微电子技术的发展,闪存固态硬盘的特性发生了显著的变化.

收稿日期:2020-09-07;**修回日期**:2020-11-04
基金项目:中国博士后科学基金资助项目(2020M680568,2021T140376);国家自然科学基金项目(62072177)
This work was supported by China Postdoctoral Science Foundation (2020M680568, 2021T140376) and the National Natural Science Foundation of China (62072177).
通信作者:舒继武(shujw@tsinghua.edu.cn)

首先介绍了闪存存储单元的基本结构和存储原理.然后讨论了闪存固态硬盘的多项控制器关键技术,包括缓存设备、地址转换层、垃圾回收、数据分配、磨损均衡以及纠错码等.这些关键技术将支撑闪存固态硬盘的正常运作.此外,探讨了闪存固态硬盘的并行结构,并分析了闪存固态硬盘并行性利用的限制条件以及最新的并行性优化工作.接着,分析了3D闪存固态硬盘的发展和堆叠式结构,并针对3D固态硬盘的性能和寿命优化工作进行了归纳和分析,提出了现有3D固态硬盘性能和寿命优化工作的不足.最后,总结了当前闪存固态硬盘的现状,并提出可能的未来研究方向.

关键词 闪存;固态硬盘;缓冲区;闪存转换层;并行性;3D固态硬盘

中图法分类号 TP391

20世纪80年代,东芝公司发明了闪存(flash memory).随后,英特尔公司推出了世界上第一款商用闪存芯片,旨在应用于计算机数据存储.在20世纪80年代末,多家闪存厂商推出了基于闪存介质的固态硬盘.进入21世纪后,闪存固态硬盘(flash memory based solid state drives, SSDs)进入了发展的快车道,其在性能、存储容量等方面全面超越机械硬盘.闪存固态硬盘正逐步取代机械硬盘,作为外部存储设备已被广泛应用于各类嵌入式设备、个人电脑和数据中心等.

相比于机械硬盘,闪存固态硬盘内部采用电子器件实现数据存储,在结构设计方面具有完全不同的设计原理.在固态硬盘内部,闪存介质被封装为闪存芯片,并以多级并行结构进行组织,满足多个读写操作在固态硬盘内部的并行处理.为充分利用闪存芯片的低读写时延以及高并行结构等优势,固态硬盘通过专用控制器实现对闪存芯片的管理.因此,固态硬盘控制器具有多项关键技术用于实现数据布局和管理、空间管理、寿命维护等[1].

近10年来,随着闪存制作工艺和微电子技术的发展,闪存固态硬盘正呈现着新的特性和形态.当前,为进一步提高闪存固态硬盘的存储容量,3D闪存固态硬盘已经成为下一代固态硬盘的主要表现形态[2].相比于传统固态硬盘,3D固态硬盘采用堆叠结构实现单位面积内存储容量的增长.基于该堆叠结构,3D固态硬盘在可靠性、数据写入方式等方面具有不同的特性.因此,针对3D闪存固态硬盘所展现的特性进行固态硬盘性能和寿命优化将成新的研究热点.

1 闪存固态硬盘基本特性

1.1 闪存单元基本结构特征

当前,使用最为广泛的闪存单元(flash memory cell)是一种基于浮栅(floating gate)技术的场效应管(MOS晶体管)[3].根据浮栅内存储的电子数量所表现的电压值,闪存单元能够表现不同的数据值.当前,闪存单元主要可分为4种类型:单级闪存单元(single-level cell, SLC)、二级闪存单元(multi-level cell, MLC)、三级闪存单元(triple-level cell, TLC)和四级闪存单元(quad-level cell, QLC).这4种类型闪存单元分别可存储1,2,3和4 b数据.

1.2 闪存单元基本操作

闪存单元存在3种基本操作,分别为“编程”“读”和“擦除”.首先,编程操作是指电子被注入闪存单元的过程.传统闪存单元编程操作普遍采用递进增压编程(increment step pulse programming, ISPP)方式[4].ISPP编程方式可将闪存单元电压值控制在预定电压范围内,从而提高闪存单元编程操作的可靠性.其次,读操作是指获取当前闪存单元电压值所属的电压范围,从而获得当前闪存单元内存储的数据信息[3].最后,擦除操作是指将闪存单元数据信息恢复至初始值.在物理实现过程中,擦除操作通过在闪存单元上构建高电场将电子从浮栅中释放,从而降低闪存单元的电压值至初始值.编程后的闪存单元需要先被擦除至初始状态,才能进行下一次编程,该现象被称之为“非就地更新”.

1.3 闪存阵列结构

为了实现批量闪存单元存储数据的读、写和擦除操作,闪存固态硬盘将闪存单元以阵列结构进行组织.如图1所示,一定数量的闪存单元构成了闪存固态硬盘中的最小闪存阵列——闪存块(flash block).该闪存阵列存在2个方向,分别是字线(word line, WL)方向和位线(bit line, BL)方向[3].在字线方向,根据闪存单元的类型(SLC,MLC,TLC或QLC),一条字线中的闪存单元构成1,2,3或4个闪存页(flash page).

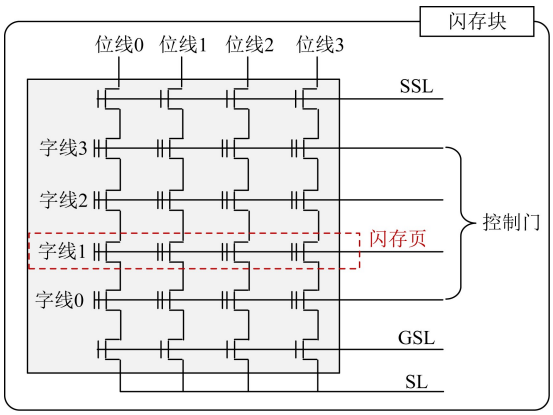


Fig. 1 Architecture of flash memory array[3]

图 1 闪存阵列结构[3]

1.4 闪存固态硬盘的基本操作

根据闪存阵列结构，闪存固态硬盘存在 3 种基本操作，分别为“写”“读”和“擦除”.写、读操作的最小执行单位是闪存页，即同一条字线中的所有闪存单元.擦除操作的最小执行单位是闪存块[5].

针对闪存页的写操作.以图 1 为例，假设当前闪存单元为 SLC 类型，图 1 闪存块中的每条字线构成一个闪存页.首先，写操作通过地址解码器确定当前写入字线位置.随后，根据写操作数据内容，对应的位线打开闪存单元充入电子，该过程表示为对当前字线对应的闪存单元进行写“0”操作.

针对闪存页读操作.同样以图 1，闪存阵列地址解码器定位读取闪存页位置.随后，读操作对所有位线施加探测电压.根据探测结果获得当前闪存页的数据内容.

针对闪存块擦除操作.所有闪存块内字线对应的控制门电路均施加 0V 电压，随后对闪存块内的所有闪存单元基质端施加高电压，从而在闪存单元上构建高电场进行电子释放.

2 闪存固态硬盘内部组件及结构特征

在本节中，我们主要介绍闪存固态硬盘的内部关键组件以及高并发的结构特征.如图 2 所示，闪存固态硬盘主要由 2 部分构成，分别是固态硬盘控制器(SSD controller)和闪存芯片(flash chip).控制器内部存在多个关键组件，用于维护闪存固态硬盘的正常运作.闪存芯片用于存储数据内容.

2.1 闪存固态硬盘控制器

根据闪存介质的独有特性，固态硬盘内部配置有专用控制器用于维护固态硬盘的正常工作，主要

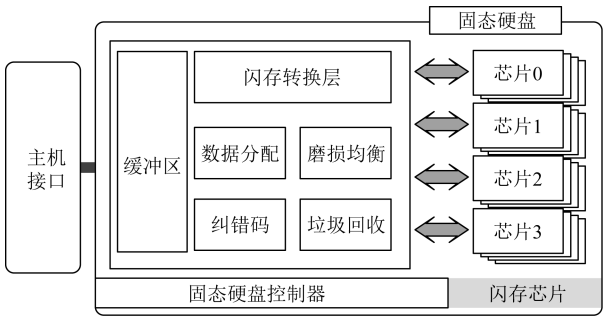


Fig. 2 Architecture of solid state drives[6]

图 2 固态硬盘结构[6]

包括 6 个关键组件：缓存(buffer)、闪存转换层(flash translation layer，FTL)、垃圾回收(garbage collection，GC)、数据分配(data allocation，DA)、磨损均衡(wear leveling，WL)和纠错码(error correction code，ECC).

2.1.1 缓 存

当前，固态硬盘的一个主要发展趋势是提升读写性能.为此，固态硬盘通过在控制器内部配置缓存设备的方式提高固态硬盘整体性能.根据缓存存储介质的易失特性，缓存设备可分为 2 类：易失性缓存和非易失性缓存.易失性缓存具有更低的访问时延，但存在设备断电数据丢失的问题.非易失性缓存访问时延介于闪存访问时延与易失性缓存访问时延之间，但其优点在于设备断电数据不丢失.

针对易失性缓存，缓存设备普遍采用 DRAM 或 SRAM 作为缓存介质[7].相比于闪存 us 级别数据访问时延，DRAM 存储介质具有更低的访问时延(ns 级别).因此，主机的读写请求可通过访问缓存设备的方式完成数据在固态硬盘内部的读写操作，从而提高固态硬盘的整体性能.由于易失性缓存存储空间有限，缓存设备根据数据特征进行数据剔除，实现有限缓存空间的高效利用.目前，已有大量工作提出基于固态硬盘特征感知的缓存数据剔除策略.从数据剔除粒度考虑，缓存数据剔除策略可分为 3 类：基于页剔除策略、基于块剔除策略以及混合剔除策略.1)页剔除策略.CFLRU(clean first least recently used)策略考虑了闪存页非对称读写时延的特征，数据剔除过程以页作为剔除粒度.其主要思想是，优先剔除缓存中的干净页(clean page)，减少脏页(dirty page)数据写回带来的闪存设备写操作时间开销[8].测试结果表示，相比于传统 LRU 策略，CFLRU 能够减小缓存数据页剔除时间达 26%的.进一步考虑到频繁写操作对闪存固态硬盘擦除操作的影响，WIPPA(write intensive page preserving algorithm)[9]

提出延长频繁更新的数据在缓存中的保存时间,从而提高数据在缓存中被更新的概率,进而减少固态硬盘内部的写操作次数.2)块剔除策略.FAB(flash aware buffer policy)策略将缓存数据以块的方式进行组织[10],剔除过程选择当前 MRU(most recently used)端具有最多数据页个数的块.测试结果显示,相比于传统 LRU 策略,FAB 减小总处理时间达 17%.基于 FAB 策略,BPLRU(block padding least recently used)进一步提出页填充机制,在缓存设备中将随机写填充至碎片化的数据块中,从而减少数据剔除开销.测试结果显示,相比于 FAB,BPLRU 提升系统性能最高达 43%.此外,VBBMS(virtual-block-based buffer management scheme)[11]在缓存中提出虚拟块结构,并将数据分为随机请求和顺序请求,针对不同类型的访问请求分别利用其空间局域性和时间局域性,从而实现性能优化的目的.3)混合剔除策略.BPAC(block-page adaptive cache)提出在缓存中按照数据冷热程度分别以页和块的方式进行组织[12].由于热数据具有更强的时间局域性,冷数据具有更强的空间局域性,因此,热数据采用页剔除策略,冷数据采用块剔除策略.测试结果显示,相比于 BPLRU,BAPC 减少缓存剔除次数最高达 79.6%.

然而,由于闪存固态硬盘缓存空间有限,且传统的缓存数据管理策略已经能取得较为出色的性能表现,因此,研究人员对缓存数据管理策略的关注逐步下降.取而代之的是,非易失性缓存设计方案逐步进入到研究人员视野中.

针对非易失性缓存,固态硬盘内部存在 2 种设计方案.1)使用非易失性存储器(non-volatile memory, NVM)作为缓存存储介质,例如相变位存储器(phase change memory, PCM),自旋转移矩存储器(spin-transfer torque RAM, STT-RAM),以及可变电阻式存储器(resistive RAM, RRAM)等[13-14].然而,由于 NVM 访问时延低于 DRAM 等易失性存储介质,因此,固态硬盘内部采用混合式存储方法作为非易失性缓存设计方法.以 PCM/DRAM 混合缓存设计方法为例,文献[13]提出将热数据存储在 DRAM 缓存区域,将冷数据存储在 PCM 缓存区域.该方法降低请求响应时间最多达 26%.2)使用电容器件对易失性缓存进行断电数据写回保护[15].在固态硬盘正常工作期间,DRAM 缓存设备提供快速数据访问性能.在固态硬盘断电之后,固态硬盘供电转换器激活电容器件为固态硬盘进行供电,使得缓存数据可被写回闪存[16].目前,基于电容器件的非易失性缓存方案已经被广泛应用于高端固态硬盘.

缓存设备能有效协调闪存与内存之间的性能差异,提高固态硬盘访问性能.面对目前各类新型非易失存储器,如何选择合适的新型非易失存储器作为固态硬盘的缓存设备将是一个研究的热点.并且,如何设计高效的基于新型非易失存储器的缓存管理策略也将是一个研究的难点.

2.1.2 闪存转换层

基于闪存“非就地更新”的特点,固态硬盘在控制器内部通过构建闪存转换层的方式实现“逻辑地址(logical address)”到“物理地址(physical address)”的地址映射转换.闪存转换层的设计主要考虑 2 方面的影响:固态硬盘性能和 RAM 空间需求.因此,根据地址映射单位的不同,闪存转换层中的地址映射转换可分为 3 种类型:闪存页级别地址映射(page-level mapping)、闪存块级别地址映射(block-level mapping)以及混合地址映射(hybrid mapping).

1) 闪存页级别地址映射[17].该映射方式中逻辑闪存页地址与物理闪存页地址存在唯一对应关系,其通过构建页粒度的映射条目实现闪存存储空间的灵活、充分使用.然而,该映射方式需要在固态硬盘缓存中保存大量映射条目,从而造成较大的缓存空间开销.

为减少对缓存空间的消耗,大量相关工作提出了针对闪存页级别地址映射缓存空间优化的设计方案,核心思想是:将当前访问概率低的映射条目写回闪存,减少缓存空间开销.DFTL(demand-based page-mapped flash translation layer)是一种典型的闪存页级别地址映射空间优化方案[18].为进一步优化 DFTL,S-FTL(spatial-locality-aware FTL)设计了基于数据空间局域性感知的 DFTL 优化方案[19].S-FTL 提出将顺序访问数据的映射条目写入同一转换页,提高转换页内部映射条目的命中率,并进一步在缓存中保存较少的随机访问数据映射条目,从而减少缓存数据剔除开销.测试结果显示,相比于 DFTL,S-FTL 减少系统想响应时间最高达 25%.在此基础上,CDFTL(two-level caching DFTL)优化方案提出构建 2 层 CMT 缓存,分别进行基于数据时间局域性和空间局域性的 DFTL 优化方案[20].测试结果显示,相比于 DFTL,CDFTL 提高 CMT 命中率达 31.51%,平均减少 31.11%系统响应时间.

2) 闪存块级别地址映射[21-23].该映射方式中逻辑闪存块地址与物理闪存块地址的对应关系符合

组相联关系.对于闪存页地址映射,闪存块级别地址映射采用固定映射机制,即逻辑块内页偏移地址与物理块内页偏移地址相同.因此,闪存块级别地址映射表仅保存闪存块地址映射条目,然而,受限于闪存块内部仅支持闪存页顺序写[24],在该映射方式下,同一逻辑闪存块内部的闪存页更新操作将导致闪存块数据频繁重组并迁移,从而增加垃圾回收开销,降低固态硬盘性能.基于这些问题,闪存块级别地址映射难以被广泛应用.

3) 混合地址映射.为解决上述2种映射方式的缺点,大量相关研究工作提出闪存页级别地址映射和闪存块级别地址映射相结合的地址映射方式[25-27].基于混合地址映射的固态硬盘闪存块分为2种类型,分别为数据块和日志块,其中数据块采用闪存块级别地址映射,日志块采用闪存页级别地址映射.针对数据块的数据更新写将通过闪存页级别地址映射方式写入日志块.当日志块空间不足时,固态硬盘触发垃圾回收(在2.1.3节详述)实现数据块与日志块的合并操作.由于合并操作存在大量数据迁移以及擦除操作,因此,研究人员提出相关工作,旨在优化混合地址映射中的合并操作开销.BAST(block associative sector translation)策略[25]提出了将日志块与数据块一一对应的映射关系.然而,当面对大量随机更新写,BAST将引起大量垃圾回收操作,导致性能下降.基于此,FAST(fully associative sector translation)策略[26]提出仅维护一个顺序写日志块,剩余日志块负责随机更新写,减少垃圾回收频率.然而,FAST却无法满足多个顺序更新写的数据流,且仍然无法利用随机更新写的时间局域性.为克服FAST所存在的问题,研究人员进一步提出LAST(locality-aware sector translation)策略[27].该策略维护了多个顺序日志块,用于满足多个顺序更新写数据流.此外,LAST进一步将随机写日志块分为冷日志块和热日志块,减少混合地址映射中的合并操作.

因此,3种类型的地址映射方案在性能和空间开销上各有优势,具体特征如表1所示.因此,针对具有较大缓存空间的闪存固态硬盘,闪存页映射方案可取得最优的性能表现.但从缓存空间考虑,闪存块映射方案可有效减少缓存空间开销.由于闪存块级别地址映射表仅保存闪存块地址映射条目,相比闪存页级别地址映射表,其所需缓存空间开销可减少"闪存块尺寸/闪存页尺寸".为实现性能和空间开销之间的权衡,混合映射根据数据的访问特征,实现闪存页映射和闪存块映射的混合管理.

Table 1 Comparison Among Page-Level, Block-Level and Hybrid Mappings

表1 闪存页级别、闪存块级别和混合映射方案对比

地址映射方案	性能	空间开销	典型方案
闪存页映射	高	高	DFTL[18]
闪存块映射	低	低	WAB-FTL[23]
混合映射	中	中	LAST[27]

闪存转换层的效率与固态硬盘性能成正相关.受限于有限的缓存空间,研制具有高命中率的地址映射表以及低空间开销的地址映射方式将有效提高固态硬盘整体性能.然而,面对新型固态硬盘(例如,3D固态硬盘)的独有特征,设计适配新型固态硬盘的闪存转换层将成为后续的一个重要研究点.

2.1.3 垃圾回收

闪存的"非就地更新"特征要求闪存单元仅在被擦除之后才可进行下一次写操作.基于此,更新写操作将在固态硬盘内部产生大量无效闪存页,从而快速消耗固态硬盘剩余可用空间.为保障固态硬盘存储空间的充分利用,固态硬盘采用垃圾回收机制实现无效空间的回收再利用[28].由于垃圾回收过程将导致大量数据读写以及高时间开销的擦除操作,因此,固态硬盘性能与垃圾回收效率成正相关[29].目前,垃圾回收优化类型可分为3类,如表2所示:

Table 2 Summary of Garbage Collection Optimizations

表2 垃圾回收优化方案总结

优化类型	主要措施	典型方案
时间开销	减小数据迁移开销	GreedyGC[30]
执行时机	避免主机请求阻塞	PGC[31]
访问预测	数据未来热度预测	LSTM/KM[32]

针对时间开销优化类型,目前,应用最为广泛的策略是贪婪式垃圾回收(greedy garbage collection)[30],该策略选择当前具有最少有效页个数的闪存块作为待回收闪存块,减少有效页迁移次数.基于此,相关研究工作提出通过冷热数据分离的方式减少待回收闪存块的有效页个数[33].根据访问数据的时间局域性,基于闪存块粒度的冷热分离策略通过将相同热度的数据在同一闪存块内进行聚集,从而实现同一闪存块内部数据在相近时间内同时无效,从而减少垃圾回收所需的有效页迁移次数.贪婪式垃圾回收方案旨在减小垃圾回收迁移的有效页个数,实现垃圾回收效率优化.为进一步提高垃圾回收效率,研究人员提出一系列垃圾回收有效页迁移过程优化方案,通过减少有效页迁移的时间开销提高

固态硬盘垃圾回收效率.针对有效页迁移过程,相关工作提出了固态硬盘的高并发特性减少有效页迁移开销.Gao 等人[34]提出了利用闪存芯片高并发的特点同时迁移多个有效页,减少有效页迁移的写操作时间开销.Shahidi 等人[35]提出利用闪存芯片内部分组并行性实现并行垃圾回收、降低垃圾回收有效页迁移时间开销和垃圾回收触发次数.

针对时间开销优化的垃圾回收方案从垃圾回收时间开销角度出发,并未考虑垃圾回收对主机读写请求的影响.因此,相关工作考虑了主机请求与垃圾回收之间的冲突关系,通过优化垃圾回收执行时机减少垃圾回收对主机读写请求的阻塞影响,实现固态硬盘性能优化[31,36].例如,Lee 等人[31]提出了可中断式垃圾回收策略.根据主机请求到达时间点,该策略将有效页迁移过程拆分为几个子阶段,并优先处理主机请求,从而减少主机请求的等待时间,提高固态硬盘性能.

针对时间开销优化的垃圾回收方案和可中断式垃圾回收策略 2 种垃圾回收优化思路呈正交关系,二者分别通过优化垃圾回收有效页迁移过程,以及选择合适的垃圾回收时机,实现固态硬盘性能优化的目的.目前,研究人员进一步提出了基于数据访问特征预测的垃圾回收优化方案[32,37].例如 Yang 等人[32]提出利用神经网络进行数据热度预测,从而进行将具有相同预测热度的数据写入同一闪存块,提高垃圾回收效率.然而,基于神经网络的预测算法具有较高的时间和空间开销,因此,在保证预测算法高准确度的前提下,如何优化预测过程的开销是一项关键技术.

2.1.4 数据分配

针对写请求数据,为确定其数据的物理写入位置,固态硬盘通过数据分配策略确定逻辑地址到物理地址的映射关系.如图 3 所示,固态硬盘具有 4 级并行层次,分别是通道(channel)级、芯片(chip)级、晶圆(die)级和分组(plane)级.数据分配策略需依次定位写请求在各个并行层次的物理单元位置.分配策略的目标是最大化固态硬盘的并行性,提高固态硬盘性能.

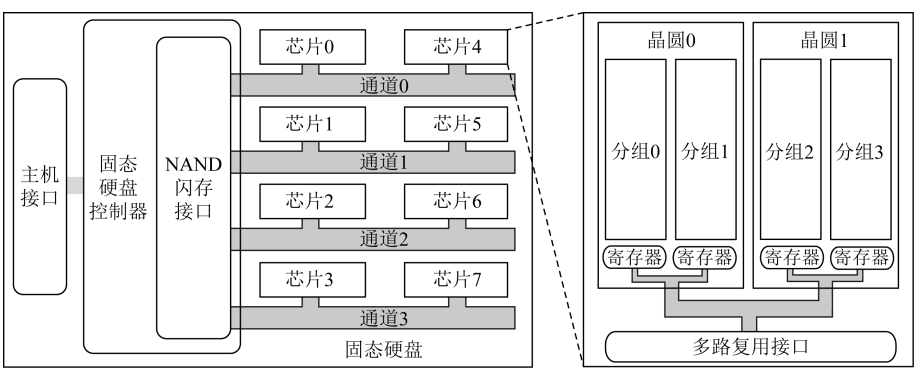


Fig. 3 Parallel architecture of solid state drives[38]

图 3 固态硬盘并行结构[38]

目前,固态硬盘内部的数据分配策略可分为 3 种类型:静态分配(static allocation)、动态分配(dynamic allocation)[39]以及混合式分配策略.静态分配策略具有固定的地址转换算法,实现过程开销低,且根据 4 级并行层次的分配优先级可确定多种分配方式.Hu 等人[39]验证了多种分配优先级下的静态数据分配策略,发现最优分配顺序为:通道、晶圆、分组、芯片.然而,由于静态分配策略不感知闪存并行单元的当前状态,若请求访问正处于忙碌状态的闪存并行单元,则导致请求受到阻塞,增加等待时间.为此,Gao 等人[40]提出基于并行发送队列(parallel issue queue, PIQ)的请求调度器.PIQ 利用静态分配策略的地址转换算法提前感知请求发送位置,并将可并行处理的多个请求合并为一个并行批次,从而减少请求之间的相互阻塞影响.

相比静态分配策略,动态分配策略更为灵活,且具有更好的性能表现,其核心思想是:将请求数据写入空闲闪存页[41].在该过程中,动态分配策略要求数据写入空闲闪存页路径上的通道、芯片、晶圆以及分组均处于空闲状态,因此,动态分配策略需要对各级闪存并行单元进行空闲/忙碌状态监测,从而增加实现开销.

为增加数据分配策略的灵活性,研究人员提出基于静态分配和动态分配策略的混合式分配策略.该策略可在不同并行层次实现静态分配和动态分配策略,从而实现性能与实现开销之间的权衡.例如

TBM(twin block management)[42],其提出在通道、芯片和晶圆层次实现静态分配策略,在分组层次采用动态分配策略,从而提高固态硬盘的并行性利用率.

然而,固态硬盘的4级并行层次中,晶圆级和分组级并行存在限制条件,导致固态硬盘并行性利用不充分,具体原因将在2.2节详述.为实现固态硬盘性能优化,数据分配策略应进一步考虑请求访问特征与固态硬盘内部数据活动,保障数据分配过程中最大化利用固态硬盘并行性,且最小化请求数据之间以及请求数据与内部数据活动的相互阻塞.3种数据分配策略性能及实现开销如表3所示:

Table 3 Summary of Data Allocation Schemes

表3 数据分配策略总结

分配策略	性能	实现开销
静态分配	低	低
动态分配	高	高
混合式	中	中

2.1.5 磨损均衡

闪存单元具有有限的可擦除次数,一旦达到擦除次数上限,闪存单元内部存储数据将无法保障其可靠性.从固态硬盘角度来看,若部分闪存单元被快速磨损,则将导致固态硬盘存储寿命退化,性能波动,以及数据存储不可靠.因此,固态硬盘采用磨损均衡保障固态硬盘内部闪存块具有相近磨损程度,从而最大化固态硬盘使用寿命.磨损均衡按照实现算法可分为动态磨损均衡(dynamic wear leveling)和静态磨损均衡(static wear leveling)[43],按照作用域可分为全局磨损均衡(global wear leveling)和本地磨损均衡(local wear leveling).4种类型的磨损均衡策略具体侧重点和性能效果如表4所示:

Table 4 Summary of Wear Leveling Schemes

表4 磨损均衡策略总结

磨损均衡	技术侧重点	性能
动态策略	空闲空间内实现磨损均衡	高
静态策略	所有空间内实现磨损均衡	低
全局策略	实现于全盘/存在数据迁移	低
本地策略	实现于芯片/减小数据迁移	高

1) 动态磨损均衡[44].该磨损均衡算法的做法是将作用域内可用闪存块列入闪存块池,在其内部选择擦除次数最少的闪存块进行数据写入.基于此,磨损较少的闪存块能够被写入更多数据内容,从而在该闪存块中触发垃圾回收,实现磨损均衡.然而,动态磨损均衡的不足在于,当作用域内存在数据长期不更新,则这部分数据所在的闪存块将无法列入可用闪存块池,从而导致动态磨损均衡无法作用于这部分闪存块.因此,动态磨损均衡难以实现作用域内所有闪存块之间的磨损均衡.

2) 静态磨损均衡[45-46].为解决动态磨损均衡所存在的不足,静态磨损均衡算法提出对作用域内所有闪存块实现磨损均衡,其思想是:将作用域内所有闪存块列入磨损均衡闪存块池,并选择擦除次数最少的闪存块进行数据写入.若所选闪存块内部存在数据内容,则将数据内容迁移至其他可用闪存块.相比于动态磨损均衡,静态磨损均衡可更大程度延长闪存块使用寿命,但其引入的数据迁移将影响固态硬盘性能.

3) 全局磨损均衡和本地磨损均衡[47].在传统闪存固态硬盘中,由于制程精度更大,闪存单元最大可容忍擦除次数更高,因此,固态硬盘可将磨损均衡算法的作用域设置为闪存芯片,保障闪存芯片之间相对磨损均衡.然而,目前闪存制程精度越来越小,导致闪存单元最大可容忍擦除次数逐步下降.因此,固态硬盘对闪存芯片之间的磨损均衡要求更为严格,从而将磨损均衡算法作用域扩充至全盘,实现全盘内部的动态磨损均衡或静态磨损均衡算法.然而,全局磨损均衡将导致芯片之间的数据迁移,从而导致固态硬盘性能下降.

磨损均衡算法的高效性与固态硬盘寿命成直接关系.动态磨损均衡和静态磨损均衡在性能和寿命方面各有优势,因此,如何高效合并2种磨损均衡算法,使得在提升固态硬盘寿命的同时不影响其性能将成为关键研究问题.此外,面向固态硬盘日益增长的存储容量,全局磨损均衡在信息维护,数据处理等方面的开销也将逐步增加.因此,如何减小大容量环境下的全局磨损均衡开销也将成为一个关键研究内容.

2.1.6 纠错码

随着制程工艺的进步,闪存单元可靠性面临更为严峻的挑战.如图4所示,当闪存单元受到错误源干扰,其表现的电压值将发生改变,从而导致闪存单元所代表的数据内容出错[48].目前,闪存单元受到的出错影响主要包括:保存时间[49]、编程干扰[50]、读干扰[51]和物理干扰[52]等.具体导致的闪存单元出错情况如表5所示.因此,为避免错误源导致的数据错误,提高数据可靠性,固态硬盘采用纠错码实现数据出错之后的纠正恢复.

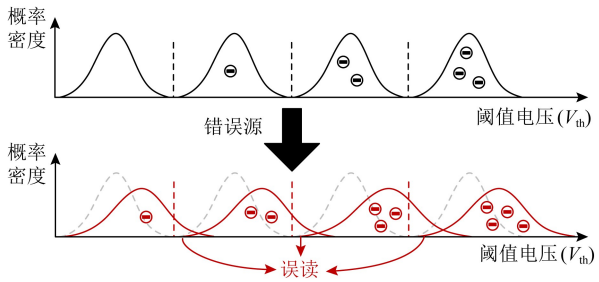


Fig. 4 Threshold voltage distribution shifts of flash cell[48]

图 4 闪存单元电压状态分布偏移[48]

Table 5 Summary of Flash Cell Error Sources[49-52]

表 5 闪存单元错误源总结[49-52]

错误源	闪存单元电压状态	可靠性影响
保持时间	电子流失、状态左偏	高
编程干扰	电子注入、状态右偏	高
读干扰	电子注入、状态右偏	低
物理干扰	电子注入、状态右偏	高

当前,固态硬盘纠错码主要 2 种类型:BCH(Bose-Chaudhuri-Hocquenghem)[53-54] 和 LDPC(low-density parity-check)[55-56].纠错码的效率决定于其是否能够有效恢复错误数据并降低数据错误率至 JEDEC 标准(10^{-5})[57].由于 BCH 纠错能力较低,其主要应用于制程精度较低,闪存单元密度较小的固态硬盘或者嵌入式闪存设备[58].针对 LDPC,其已被广泛应用于制程精度更高、密度更大、可靠性要求更高的固态硬盘,例如 3D 闪存固态硬盘.

LDPC 采用 2 种方式进行数据纠错,包括硬判决和软判决[59],其实现过程是:首先,固态硬盘控制器利用 LDPC 进行解码纠错,若解码成功则返回数据内容.该过程称之为硬判决.若硬判决失败,则进入 LDPC 软判决过程.软判决通过不同探测电压的读操作获得闪存单元数据的软信息.软信息表示闪存单元存储数据是“0”或“1”的概率.根据软信息数据,LDPC 再次进行数据解码纠错,若解码成功则返回数据内容,否则调整读操作探测电压,从而获得更多软信息.因此,LDPC 的时间开销与纠错过程的读操作次数成正比.目前,已有大量相关工作提出对 LDPC 纠错开销进行优化[60-62].

纠错码的纠错能力与数据内容和出错特征相关,面对闪存单元制程精度和存储密度的提高,纠错码的纠错能力正面临着更大的挑战.因此,一方面,如何设计具有更高纠错能力的纠错码将成为研究方向之一;另一方面,如何充分利用数据内容和出错特征,提高现有纠错码的纠错能力和减少其时间开销将成为重要关注点.

2.2 闪存固态硬盘结构特征

固态硬盘具有高并行结构特性.如图 3 所示,固态硬盘内部存在 4 级并行结构,数据通过访问不同并行单元实现高并发访问[63].在 4 级并行结构中,“通道”和“芯片”级别并行结构称之为外部并行.“晶圆”和“分组”级别并行结构称之为内部并行.外部并行是固态硬盘的固有并行结构,读写请求通过数据分配策略可完全利用该并行结构.然而,针对内部并行,为完全利用该级并行结构,固态硬盘需支持高级命令,且读写请求需满足相关限制条件[64].

固态硬盘高级命令包括 3 种类型:Copyback,Multiplane 及 Interleave.首先,针对 Copyback 命令,其可将闪存页数据内容迁移至当前分组中的可用闪存页.该命令的优势在于数据迁移过程可不经过 I/O 总线,从而减小数据迁移时间开销[65];其次,Multiplane 命令通过在同一晶圆内的不同分组中同时执行读、写及擦除操作实现分组级别并行性.然而,Multiplane 命令存在 2 个限制条件:1)不同分组中执行的操作类型相同,即同时在不同分组中执行读操作或写操作或擦除操作;2)访问的分组内偏移位置相同[66];最后,针对 Interleave 命令,该命令允许读、写及擦除操作在同一闪存芯片内的不同分组中同时执行,且不存在限制条件[64].为完全利用固态硬盘的 4 级并行结构,研究人员分别在外部并行方面和内部并行方面进行了相关研究.

外部并行挖掘:基于传统调度策略的固态硬盘无法充分利用外部并行优势,如图 5 所示,在配置有 64 个闪存芯片的固态硬盘中,大部分负载下的闪存芯片利用率均小于 30%[40].为解决该问题,大量相关工作提出了一系列优化方案.例如:动态数据分配

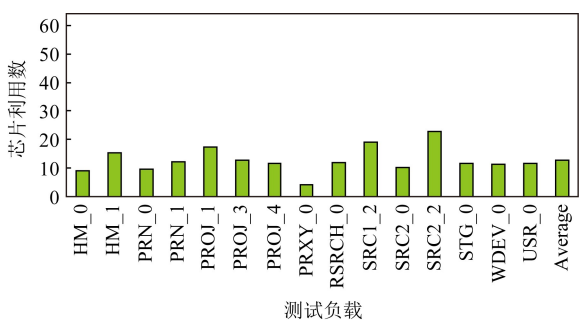


Fig. 5 Results of utilized flash chips in SSD[40]

图 5 闪存芯片利用率测试结果[40]

策略通过监控闪存芯片状态,利用贪婪算法将写请求数据写入当前空闲闪存芯片,从而有效提升固态硬盘外部并行性利用率[39].PIQ(parallel issue queuing)提出在主机端根据请求逻辑地址确定访问的闪存芯片地址,从而通过调度器同时发送多个无访问冲突的请求数据,充分利用固态硬盘外部并行[40].文献[67]提出将冲突概率较大的数据页复制迁移至其他闪存芯片,利用闪存芯片级别并行性减小读操作受到的冲突阻塞影响.

内部并行挖掘:内部并行挖掘的目标是如何高效利用分组级别并行性.目前,如图 6 所示,在当前固态硬盘中,由于高级命令的限制,其内部并行性利用率仍然较低.在支持 Multiplane 高级命令情况下的并行写操作平均比例仅为 0.088 7%.

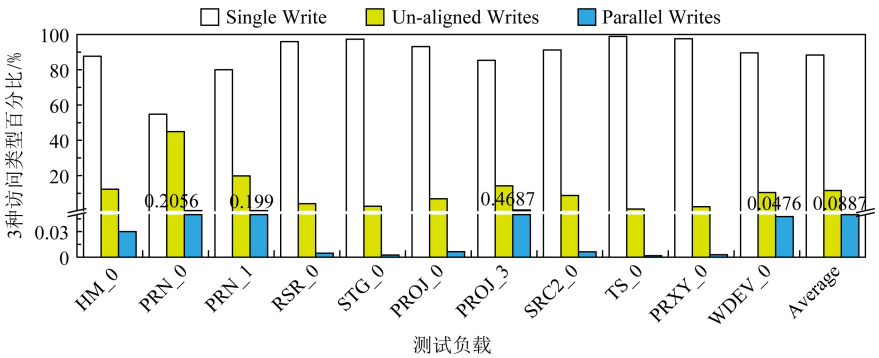


Fig. 6 The percentages of parallel writes in SSD[38]

图 6 固态硬盘并行写测试结果[38]

为此,相关工作提出一系列优化方法,实现基于 Multiplane 高级命令的并行性优化方案.例如:PAQ(physically addressed queuing)提出在固态硬盘控制器层次实现基于物理访问地址的调度策略,该策略通过地址映射表和数据分配策略提前确定当前请求访问的分组位置,从而将访问不同分组的请求并行处理,提高分组级别并行性[68].TBM(twin block management)提出在不同分组的写入点维护机制中采用循环策略(round-robin policy),将写入数据循环写至不同分组内偏移位置,从而主动适配分组并行的 2 个限制条件[69].然而,由于写请求写入数量的不可预测性,不同分组的写入点无法保障时刻对齐状态,从而限制了分组并行的挖掘.除此之外,研究人员进一步提出通过主动构造满足分组并行限制条件的数据访问模式.PaGC(parallel GC)提出在不同分组中维护额外的写入点,这些写入点具有相同的偏移位置.因此,垃圾回收的数据迁移可利用分组并行性同时进行数据迁移,从而提高垃圾回收效率[38].类似地,SPD(an SSD from plane to die)提出利用固态硬盘缓存设备进行访问模式重构[38].SPD 通过从缓存设备中剔除与一个晶圆内分组数量相同的数据页个数,并将剔除数据对齐写回不同分组的相同偏移位置实现分组并行挖掘.基于此,SPD 对垃圾回收过程的数据迁移采取相同做法,从而保证固态硬盘的一个晶圆内部的所有分组写入点均处于对齐状态,从而完全利用分组并行性.相比于 TBM 和 PaGC,SPD 写请求时延分别下降 42.05%和 28.58%.

固态硬盘的并行性是性能优化的一个重要途径.当前大量相关工作针对数据的访问特征以及并行性的限制条件进行请求调度、数据分配等优化设计,然而,仅从主机读写请求访问特征角度进行并行性挖掘仍存在一定的局限性.由于固态硬盘存在大量内部活动,例如垃圾回收、磨损均衡、映射条目读写等,主机读写请求的访问特征极易受到固态硬盘内部活动的干扰,因此,在固态硬盘并行性挖掘过程中,研究人员应当充分考虑固态硬盘内部活动带来的数据迁移影响.

3 新型闪存固态硬盘

随着制程工艺的进步,在保障数据可靠存储的前提下,传统闪存固态硬盘的存储容量难以进一步增加.因此,各大闪存厂商提出利用 3D 堆叠式结构增加单位物理面积下的存储容量[70].当前主流的堆叠技术包括 BiCS[71],P-BiCS[72],VNAND[73] 和 TCAT[74].以三星公司推出的 TCAT 结构为例,其结构如图 7 所示.闪存单元在 3D 闪存块中呈立体堆叠式构造,同一水平面上的闪存单元构成“层”级结构[75-76].目前,

3D 固态硬盘内部的闪存单元采用 TLC 或 QLC 类型,使得存储容量进一步提升.基于这 2 种方式,3D 固态硬盘存储容量已达到 TB 级别.

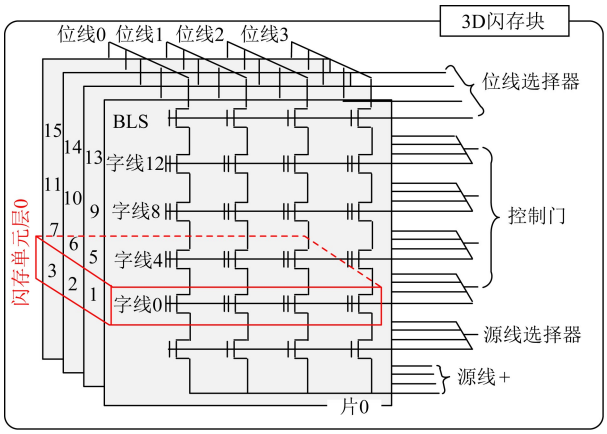


Fig. 7 The architecture of TCAT based 3D flash block[74]

图 7 基于 TCAT 的 3D 闪存块结构示意图[74]

3.1 3D 固态硬盘特征

相比于传统平面固态硬盘,3D 固态硬盘在结构方面的差异主要体现在闪存块结构.传统平面固态硬盘的闪存块呈现平面结构,但 3D 固态硬盘闪存块采用 3D 垂直堆叠式结构,该结构可有效提高单位面积存储容量.现有 3D 闪存芯片已实现 128 层的堆叠技术.2019 年,三星、美光等国外公司相继推出 128 层 3D 闪存芯片[77-78].2020 年我国长江存储研制出 128 层 3D 闪存芯片,其推出的 XTacking 结构通过控制电路与存储电路堆叠的方式实现了电路面积的进一步缩小,并提高了存储颗粒生产效率[79].三星公司表示,在未来 5 年内将有望实现 500 层的 3D 闪存芯片堆叠技术,在未来 10 年,3D 堆叠层数有望突破 1 000 层.

3D 固态硬盘的另一结构特征是更大尺寸的闪存页.当前 3D 固态硬盘最大闪存页尺寸可达 32KB.这一特征使得上层应用的访问数据粒度与 3D 固态硬盘闪存页粒度难以匹配,从而在小尺寸数据访问环境下,引起 3D 固态硬盘性能下降以及寿命退化.类似地,3D 固态硬盘闪存块内部的闪存页数量也呈现增长趋势,从而导致 3D 固态硬盘闪存块的数据迁移开销增大,引起性能下降.

因此,基于 3D 固态硬盘结构特征的改变,3D 固态硬盘研究领域将出现新的挑战:1)3D 固态硬盘作为最新的仍受限于有限的电路空间,3D 闪存芯片堆叠层数无法无限制增加,因此,如何在现有 3D 结构中进一步提高存储容量将成为 3D 固态硬盘的一个研究重点;2)3D 固态硬盘闪存页具有更大尺寸,闪存块具有更多闪存页,因此,如何减少大闪存页结构下的写放大问题,以及如何减少闪存块垃圾回收导致的数据迁移时间开销将成为 3D 固态硬盘的寿命和性能优化的研究热点.

在 3D 固态硬盘制程粒度方面,由于当前 3D 堆叠技术采用的制程粒度较大(30 nm 至 50 nm),因此,3D 闪存块内部的层与层之间具有较大的容错空间,从而使得闪存单元之间的错误干扰(例如编程干扰等)影响较小[50].基于此,3D 固态硬盘可采用“一次编程到位(one-shot programming)”方式提升写操作吞吐率和性能[80].相比于传统 ISPP 编程方式,“一次编程到位”方式可将闪存单元从初始擦除位直接编程至目标位.该编程过程要求,所有待编程数据需提前确定,从而通过一次编程操作在闪存单元中同时写入多个比特数据.

然而,由于 3D 固态硬盘采用的闪存单元类型从 TLC,发展至 QLC,甚至 PLC,因此,闪存单元的电压状态分布更为密集,从而使得 3D 固态硬盘对可靠性的要求更高.此外,在更高的制程精度下,“一次编程到位”方式将导致 3D 固态硬盘比特错误率增加,进一步影响其可靠性,因此,如何进一步提高 3D 固态硬盘的可靠性将成为未来 3D 固态硬盘发展的重要支撑点.

3.2 3D 固态硬盘优化方案

针对 3D 固态硬盘性能优化,ApproxFTL 提出利用 3D 固态硬盘的高容错能力,通过近似数据写操作的方式,缩短闪存单元编程电压范围,从而在保障数据一定可靠性的前提下,实现更快的编程速度[81].OSPADA 提出“一次编程到位”编程方式感知的数据布局方法[82].由于“一次编程到位”编程方式在单条字线的闪存单元中同时写入多个数据,从而导致顺序访问的热读数据的顺序读性能差.因此,OSPADA 提出将热读数据写入不同的字线,实现并行读操作,提升 3D 固态硬盘读性能.

针对 3D 固态硬盘寿命优化,相比于同种闪存单元类型(例如 TLC)的传统固态硬盘,3D 固态硬盘制程粒度较大,因此,其可容忍的编程/擦除次数更大,使用寿命更长.为进一步优化 3D 固态硬盘使用寿命,ReSSD 提出利用 3D 闪存单元重编程技术增加闪存单元在一个擦除周期内的存储比特量,从而减少垃圾回收次数,进而提高 3D 固态硬盘使用

寿命[83].相比于传统重编程技术，由于3D闪存单元具有更大的制程粒度，且闪存单元之间存储间距更大，因此，3D闪存单元具有更高的容错能力，从而针对重编程技术具有更好的适配度，可有效解决重编程问题带来的可靠性影响.ApproxFTL提出的缩短闪存单元编程电压范围的方法同样可提高3D固态硬盘的使用寿命，其方法是：对闪存单元写操作施加小步幅编程电压，从而减少闪存单元磨损，从而提高3D固态硬盘寿命[81].

目前，3D固态硬盘正处于兴起阶段，其具有不同于传统固态硬盘的独有特征，例如大闪存页、大闪存块以及高容错能力等.在未考虑3D固态硬盘独有特征的情况下，现有的性能和寿命优化工作难以取得显著效果，从而制约了3D固态硬盘的发展.因此，如何充分考虑3D固态硬盘的独有特征，设计高效的寿命和性能优化技术将成为3D固态硬盘发展过程中的关键挑战.

4 总结与展望

本文从闪存固态硬盘的基本结构特征、固态硬盘控制器组件、3D固态硬盘的结构与优化3个角度出发，详细分析了固态硬盘的结构特征和发展趋势，并针对关键技术内容和发展趋势进行了剖析.现有的固态硬盘关键技术能有效提高其性能和寿命，但面对上层高性能应用的兴起，固态硬盘现有的访问性能和使用寿命仍然无法满足数据长期有效存储以及快速响应的严苛需求.因此，如何利用传统固态硬盘的结构特征以及高性能应用环境下的数据访问特征，提升固态硬盘性能和寿命将是一个重要挑战.此外，面对3D固态硬盘，如何实现存储容量和性能以及寿命之间的权衡，在满足大容量数据存储的同时，实现高性能访问和长使用寿命也将是3D固态硬盘的重要研究方向.

首先，针对固态硬盘控制器内部组件，基于各类新型非易失存储器，如何在控制器缓存设备中利用新型非易失存储器的性能优势将对固态硬盘的性能起到至关重要的作用.并且，基于固态硬盘的缓存优势，如何进一步实现基于固体硬盘的智能存储，将处理数据在缓存中进行提前处理，从而避免数据在传统I/O栈上的高软件传输开销；其次，针对闪存转换层，如何在3D固态硬盘中设计高效的闪存转换层，从而减小空间、时间开销，提升命中率将是目前的一个研究难点.并且，如何针对新型应用，例如神经网络应用，设计匹配神经网络应用的闪存转换层也将成为一个研究热点；此外，针对垃圾回收和磨损均衡，考虑到固态硬盘内部活动的复杂性，如何将请求数据特征与固态硬盘内部结构和内部数据活动想结合，减小垃圾回收频率以及提高磨损均衡效率将有效提高固态硬盘的性能和寿命.最后，针对3D固态硬盘，传统固态硬盘优化技术将难以取得有效的性能和寿命优化，原因在于，相比于传统固态硬盘，3D固态硬盘在内部结构、容错能力、以及编程方式等方面具有不同的特征.因此，为有效提升3D固态硬盘的性能和寿命，研究人员需充分考虑3D固态硬盘的独有特征，发挥3D堆叠式结构、高容错能力等优势，实现高性能、高寿命的3D固态硬盘.此外，针对国产3D固态硬盘，研究其特有的堆叠结构，发挥其独有优势，从而实现国产3D固态硬盘的优化设计将有助于推动国产3D固态硬盘的普及，助力存储行业的发展.

综上，固态硬盘将成为存储系统的主流存储设备.在高性能应用环境下，进一步优化传统固态硬盘和新型3D固态硬盘的性能和寿命将有助于推动固态硬盘的广泛应用和存储产业的发展.

作者贡献声明：高聪明进行了该论文相关实验设计、编码及测试、论文撰写等工作；石亮进行了前期实验方案的讨论设计与论文修改；刘凯进行了论文结构讨论和修改；薛春进行了实验补充设计讨论及论文修改；舒继武进行了论文修改以及新型固态硬盘研究讨论.

参 考 文 献

[1] Zheng Wenjing, Li Mingqiang, Shu Jiwu. Flash storage technology [J]. Journal of Computer Research and Development, 2010, 47(4): 716-726 (in Chinese)
(郑文静，李明强，舒继武. Flash存储技术[J]. 计算机研究与发展，2010，47(4)：716-726)

[2] Micheloni R. 3D Flash Memories [M]. Berlin: Springer, 2016

[3] Micheloni R, Crippa L, Marelli A. Inside NAND Flash Memories [M]. Berlin: Springer Science & Business Media, 2010

[4] Suh K D, Suh B H, Lim Y H, et al. A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme [J]. IEEE Journal of Solid-State Circuits, 1995, 30(11): 1149-1156

[5] Agrawal N, Prabhakaran V, Wobber T, et al. Design tradeoffs for SSD performance [C] //Proc of the 2008 USENIX Annual Technical Conf. Berkeley, CA: USENIX Association, 2008: 57-70

[6] Gao Congming, Shi Liang, Xue Chun, et al. Parallel all the time: Plane level parallelism exploration for high performance SSDs [C] //Proc of the 2019 Symp on Mass Storage Systems and Technologies. Piscataway, NJ: IEEE, 2019: 172-184

[7] Cheong W, Yoon C, Woo S, et al. A flash memory controller for 15μs ultra-low-latency SSD using high-speed 3D NAND flash with 3μs read time [C] //Proc of the 2018 Int Solid-State Circuits Conf. Piscataway, NJ: IEEE, 2018: 338-340

[8] Park S, Jung D, Kang J, et al. CFLRU: A replacement algorithm for flash memory [C] //Proc of the 2006 Int Conf on Compilers, Architecture and Synthesis for Embedded Systems. New York: ACM, 2006: 234-241

[9] Jin Xin, Jung S, Song Y H. Write-aware buffer management policy for performance and durability enhancement in NAND flash memory [J]. IEEE Transactions on Consumer Electronics, 2010, 56(4): 2393-2399

[10] Jo H, Kang J U, Park S Y, et al. FAB: Flash-aware buffer management policy for portable media players [J]. IEEE Transactions on Consumer Electronics, 2006, 52(2): 485-493

[11] Du Chenjie, Yao Yingbiao, Zhou Jie, et al. VBBMS: A novel buffer management strategy for NAND flash storage devices [J]. IEEE Transactions on Consumer Electronics, 2019, 65(2): 134-141

[12] Wu Guanying, He Xubin, Eckart B. An adaptive write buffer management scheme for flash-based SSDs [J]. ACM Transactions on Storage, 2012, 8(1): DOI: 10.1145/2093139.2093140

[13] Tarihi M, Asadi H, Haghdoost A, et al. A hybrid non-volatile cache design for solid-state drives using comprehensive I/O characterization [J]. IEEE Transactions on Computers, 2015, 65(6): 1678-1691

[14] Lee E, Kim J, Bahn H, et al. Reducing write amplification of flash storage through cooperative data management with NVM [J]. ACM Transactions on Storage, 2017, 13(2): 1-13

[15] Kang W H, Lee S W, Moon B, et al. Durable write cache in flash memory SSD for relational and NoSQL databases [C] //Proc of the 2014 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2014: 529-540

[16] Gao Congming, Shi Liang, Di Yejia, et al. An efficient cache management scheme for capacitor equipped solid state drives [C] //Proc of the 2018 Great Lakes Symp on VLSI. New York: ACM, 2018: 463-466

[17] Ban A. Flash file system: United States, US5404485A[P]. 1995-04-04

[18] Gupta A, Kim Y, Urgaonkar B. DFTL: A flash translation layer employing demand-based selective caching of page-level address mappings [J]. ACM Sigplan Notices, 2009, 44(3): 229-240

[19] Jiang Song, Zhang Lei, Yuan Xinhao, et al. S-FTL: An efficient address translation for flash memory by exploiting spatial locality [C] //Proc of the 2011 Symp on Mass Storage Systems and Technologies. Piscataway, NJ: IEEE, 2011: 1-12

[20] Qin Zhiwei, Wang Yi, Liu Duo, et al. A two-level caching mechanism for demand-based page-level address mapping in NAND flash memory storage systems [C] //Proc of the 2011 Real-Time and Embedded Technology and Applications Symp. Piscataway, NJ: IEEE, 2011: 157-166

[21] Estakhri P, Iman B. Moving sequential sectors within a block of information in a flash memory mass storage architecture: United States, US5930815[P]. 1999-07-27

[22] Shinohara T. Flash memory card with block memory address arrangement: United States, US5905993[P]. 1999-05-18

[23] Liu Duo, Wang Tianzheng, Wang Yi, et al. A block-level flash memory management scheme for reducing write activities in PCM-based embedded systems[C]//Proc of the 2012 Design, Automation & Test in Europe Conf & Exhibition. Piscataway, NJ: IEEE, 2012: 1447-1450

[24] Chen Feng, Koufaty D A, Zhang Xiaodong. Understanding intrinsic characteristics and system implications of flash memory based solid state drives [J]. ACM SIGMETRICS Performance Evaluation Review, 2009, 37(1): 181-192

[25] Chung T S, Park D J, Park S, et al. System softwarefor flash memory: A survey [C] //Proc of the 2006 Int Conf on Embedded and Ubiquitous Computing. Berlin: Springer, 2006: 394-404

[26] Lee S W, Park D J, Chung T S, et al. A log buffer-based flash translation layer using fully-associative sector translation [J]. ACM Transactions on Embedded Computing Systems, 2007, 6(3): DOI: 10.1145/1275986.1275990

[27] Lee S, Shin D, Kim Y J, et al. LAST: Locality-aware sector translation for NAND flash memory-based storage systems [J]. ACM SIGOPS Operating Systems Review, 2008, 42(6): 36-42

[28] Zhang Qi, Li Xuandong, Wang Linzhang, et al. Lazy real time garbage collection method: United States, US9971681 [P]. 2018-05-15

[29] Yan Shiqin, Li Huaicheng, Hao Mingzhe, et al. Tiny-tail flash: Near-perfect elimination of garbage collection tail latencies in NAND SSDs [J]. ACM Transactions on Storage, 2017, 13(3): DOI: 10.1145/3121133

[30] Muchherla K K, Feeley P S, Ratnam S K, et al. Periodic flush in memory component that is using greedy garbage collection: United States, US 16175005[P]. 2020-04-30

[31] Lee J, Kim Y, Shipman G M, et al. A semi-preemptive garbage collector for solid state drives [C] //Proc of the 2011 Int Symp on Performance Analysis of Systems and Software. Piscataway, NJ: IEEE, 2011: 12-21

[32] Yang Pan，Xue Ni，Zhang Yuqi，et al. Reducing garbage collection overhead in SSD based on workload prediction [C] //Proc of the 2019 USENIX Workshop on Hot Topics in Storage and File Systems. Berkeley，CA：USENIX Association，2019：DOI：10.5555/3357062.3357089

[33] Wu Suzhen，Lin Yanping，Mao Bo，et al. GCaR：Garbage collection aware cache management with improved performance for flash-based SSDs [C] //Proc of the 2016 Int Conf on Supercomputing. New York：ACM，2016：DOI：10.1145/2925426.2926263

[34] Gao Congming，Shi Liang，Di Yejia，et al. Exploiting chip idleness for minimizing garbage collection-induced chip access conflict on SSDs [J]. ACM Transactions on Design Automation of Electronic Systems，2017，23(2)：DOI：10.1145/3131850

[35] Shahidi N，Kandemir M T，Arjomand M，et al. Exploring the potentials of parallel garbage collection in SSDs for enterprise storage systems [C] //Proc of the 2016 Int Conf for High Performance Computing，Networking，Storage and Analysis. Piscataway，NJ：IEEE，2016：561-572

[36] Gao Congming，Shi Liang，Liu Kai，et al. Boosting the performance of SSDs via fully exploiting the plane level parallelism [J]. IEEE Transactions on Parallel and Distributed Systems，2020，31(9)：2185-2200

[37] Kang W，Yoo S. Q-value prediction for reinforcement learning assisted garbage collection to reduce long tail latency in SSD [J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems，2019，39(10)：2240-2253

[38] Gao Congming，Shi Liang，Ji Cheng，et al. Exploiting parallelism for access conflict minimization in flash-based solid state drives [J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems，2017，37(1)：168-181

[39] Hu Yang，Jiang Hong，Feng Dan，et al. Performance impact and interplay of SSD parallelism through advanced commands，allocation strategy and data granularity [C] //Proc of the 2011 Int Conf on Supercomputing. New York：ACM，2011：96-107

[40] Gao Congming，Shi Liang，Zhao Mengying，et al. Exploiting parallelism in I/O scheduling for access conflict minimization in flash-based solid state drives [C] //Proc of the 2014 Symp on Mass Storage Systems and Technologies. Piscataway，NJ：IEEE，2014：DOI：10.1109/MSST.2014.6855544

[41] Tavakkol A，Mehrvarzy P，Arjomand M，et al. Performance evaluation of dynamic page allocation strategies in SSDs [J]. ACM Transactions on Modeling and Performance Evaluation of Computing Systems，2016，1(2)：DOI：10.1145/2829974

[42] Tavakkol A，Mehrvarzy P，Sarbazi-Azad H. TBM：Twin block management policy to enhance the utilization of plane-level parallelism in SSDs [J]. IEEE Computer Architecture Letters，2015，15(2)：121-124

[43] Kwon J H，Dong-Gun K I M，Sang-Gu J O. Memory system and wear-leveling method thereof：United States，US9990153 [P]. 2018-06-05

[44] Chang L P，Kuo T W. An adaptive striping architecture for flash memory storage systems of embedded systems [C] //Proc of the 2002 Real-Time and Embedded Technology and Applications Symp. Piscataway，NJ：IEEE，2002：187-196

[45] Murugan M，Du D H C. Rejuvenator：A static wear leveling algorithm for NAND flash memory with minimized overhead [C] //Proc of the 2011 Symp on Mass Storage Systems and Technologies. Piscataway，NJ：IEEE，2011：DOI：10.1109/MSST.2011.5937225

[46] Wang Chundong，Wong W F. Observational wear leveling：An efficient algorithm for flash memory management [C] //Proc of the 2012 Annual Design Automation Conf. New York：ACM，2012：235-242

[47] Yang W M，Kim K R，Kwak J S. Solid state storage system using global wear leveling and method of controlling the solid state storage system：United States，US12344702[P]. 2010-04-08

[48] Nakamura T，Deguchi Y，Takeuchi K. AEP-LDPC ECC with error dispersion coding for burst error reduction of 2D and 3D NAND flash memories [C] //Proc of the 2017 Int Memory Workshop. Piscataway，NJ：IEEE，2017：DOI：10.1109/IMW.2017.7939070

[49] Cai Yu，Yalcin G，Mutlu O，et al. Flash correct-and-refresh：Retention-aware error management for increased flash memory lifetime [C] //Proc of the 2012 Int Conf on Computer Design. Piscataway，NJ：IEEE，2012：94-101

[50] Cai Yu，Ghose S，Luo Yixin，et al. Vulnerabilities in MLC NAND flash memory programming：Experimental analysis，exploits，and mitigation techniques [C] //Proc of the 2017 Int Symp on High Performance Computer Architecture. Piscataway，NJ：IEEE，2017：49-60

[51] Cai Yu，Luo Yixin，Ghose S，et al. Read disturb errors in MLC NAND flash memory：characterization，mitigation，and recovery [C] //Proc of the 2015 Int Conf on Dependable Systems and Networks. Piscataway，NJ：IEEE，2015：438-449

[52] Luo Yixin，Ghose S，Cai Yu，et al. Heat Watch：Improving 3D NAND flash memory device reliability by exploiting self-recovery and temperature awareness [C] //Proc of the 2018 Int Symp on High Performance Computer Architecture. Piscataway，NJ：IEEE，2018：504-517

[53] Bose R C，Ray-Chaudhuri D K. On a class of error correcting binary group codes [J]. Information and Control，1960，3(1)：68-79

[54] Lee Y，Yoo H，Yoo I，et al. 6.4 Gb/s multi-threaded BCH encoder and decoder for multi-channel SSD controllers [C] //Proc of the 2012 Int Solid-State Circuits Conf. Piscataway，NJ：IEEE，2012：426-428

[55] MacKay D J C，Neal R M. Near Shannon limit performance of low density parity check codes [J]. Electronics letters，1996，32(18)：1645-1646

[56] Zhao Kai, Zhao Wenzhe, Sun Hongbin, et al. LDPC-in-SSD: Making advanced error correction codes work effectively in solid state drives [C] //Proc of the 2013 USENIX Conf on File and Storage Technologies. Berkeley, CA: USENIX Association, 2013: 243-256

[57] Council J. Failure mechanisms and models for semiconductor devices [R]. Arlington: JEDEC Solid State Technology Association, 2002

[58] Sun Fei, Rose K, Zhang Tong. On the use of strong BCH codes for improving multilevel NAND flash memory storage capacity [C] //Proc of the 2006 Workshop on Signal Processing Systems: Design and Implementation. Piscataway, NJ: IEEE, 2006: 5-10

[59] Tanakamaru S, Yanagihara Y, Takeuchi K. Over-10x-extended-lifetime 76%-reduced-error solid-state drives (SSDs) with error-prediction LDPC architecture and error-recovery scheme [C] //Proc of the 2012 Int Solid-State Circuits Conf. Piscataway, NJ: IEEE, 2012: 424-426

[60] Li Qiao, Shi Liang, Di Yejia, et al. Process variation aware read performance improvement for LDPC-based nand flash memory [J]. IEEE Transactions on Reliability, 2019, 69 (1): 310-321

[61] Li Qiao, Shi Liang, Cui Yufei, et al. Exploiting asymmetric errors for LDPC decoding optimization on 3D NAND flash memory [J]. IEEE Transactions on Computers, 2019, 69 (4): 475-488

[62] Lv Yina, Shi Liang, Li Qiao, et al. Optimizing tail latency of LDPC based flash memory storage systems via smart refresh [C] //Proc of the 2019 Int Conf on Networking, Architecture and Storage. Piscataway, NJ: IEEE, 2019: DOI: 10.1109/NAS.2019.8834728

[63] Chen Feng, Lee R, Zhang Xiaodong. Essential roles of exploiting internal parallelism of flash memory based solid state drives in high-speed data Processing [C] //Proc of the 2011 Int Symp on High Performance Computer Architecture. Piscataway, NJ: IEEE, 2011: 266-277

[64] Hu Yang, Jiang Hong, Feng Dan, et al. Exploring and exploiting the multilevel parallelism inside SSDs for improved performance and endurance [J]. IEEE Transactions on Computers, 2012, 62(6): 1141-1155

[65] Wu Fei, Zhou Jiaona, Wang Shunzhuo, et al. FastGC: Accelerate garbage collection via an efficient copyback-based data migration in SSDs [C] //Proc of the 2018 Design Automation Conf. Piscataway, NJ: IEEE, 2018: DOI: 10.1109/DAC.2018.8465931

[66] Park S, Seo E, Shin J Y, et al. Exploiting internal parallelism of flash-based SSDs [J]. IEEE Computer Architecture Letters, 2010, 9(1): 9-12

[67] He Bingsheng, Yu Xu, Zhou Chi. Improving update-intensive workloads on flash disks through exploiting multi-chip parallelism [J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 26(1): 152-162

[68] Jung M, Wilson III E H, Kandemir M. Physically addressed queueing (PAQ) improving parallelism in solid state disks [J]. ACM SIGARCH Computer Architecture News, 2012, 40(3): 404-415

[69] Tavakkol A, Mehrvarzy P, Sarbazi-Azad H. TBM: Twin block management policy to enhance the utilization of plane-level parallelism in SSDs [J]. IEEE Computer Architecture Letters, 2015, 15(2): 121-124

[70] Micheloni R, Crippa L, Zambelli C, et al. Architectural and integration options for 3D NAND flash memories [J]. Computers, 2017, 6(3): 27-35

[71] Nitayama A, Aochi H. Vertical 3D NAND flash memory technology [J]. ECS Transactions, 2011, 41(7): 15-25

[72] Katsumata R, Kito M, Fukuzumi Y, et al. Pipe-shaped BiCS flash memory with 16 stacked layers and multi-level-cell operation for ultra high density storage devices [C] //Proc of the 2009 Symp on VLSI Technology. Piscataway, NJ: IEEE, 2009: 136-137

[73] Kang D, Jeong W, Kim C, et al. 256 Gb 3 b/cell V-NAND flash memory with 48 stacked WL layers [J]. IEEE Journal of Solid-State Circuits, 2016, 52(1): 210-217

[74] Jang J, Kim H S, Cho W, et al. Vertical cell array using TCAT (Terabit Cell Array Transistor) technology for ultrahigh density NAND flash memory [C] //Proc of the 2009 Symp on VLSI Technology. Piscataway, NJ: IEEE, 2009: 192-193

[75] Liu Chunyi, Kotra J, Jung M, et al. PEN: Design and evaluation of partial-erase for 3D nand-based high density SSDs [C] //Proc of the 2018 USENIX Conf on File and Storage Technologies. Berkeley, CA: USENIX Association, 2018: 67-82

[76] Liu Chunyi, Kotra J B, Jung M, et al. SOML read: Rethinking the read operation granularity of 3D NAND SSDs [C] //Proc of the 2019 Int Conf on Architectural Support for Programming Languages and Operating Systems. New York: ACM, 2019: 955-969

[77] CDRinfo. Samsung says 6th generation, 128-layer V-NAND coming in 2020 [EB/OL]. [2019-11-25]. https://www.cdrinfo.com/d7/content/samsung-says-6th-generation-128-layer-v-nand-coming-2020

[78] Btarunr. Micron tapes out 128-layer 3D NAND flash memory [EB/OL]. [2019-10-09]. https://www.techpowerup.com/259941/micron-tapes-out-128-layer-3d-nand-flash-memory

[79] YMTC launched 128-layer QLC flash memory with a large capacity of 1.33Tb [EB/OL]. [2020-04-10]. http://www.ymtc.com/cn/index.php?s=/cms/index/detail/id/215.html (in Chinese)
(长江存储. 长江存储推出 128 层 QLC 闪存,单颗容量达 1.33Tb [EB/OL]. [2020-04-10]. http://www.ymtc.com/cn/index.php?s=/cms/index/detail/id/215.html)

[80] Ho C C, Li Y C, Chang Y H, et al. Achieving defect-free multilevel 3D flash memories with one-shot program design [C] //Proc of the 2018 Design Automation Conf. Piscataway, NJ: IEEE, 2018: DOI: 10.1109/DAC.2018.8465858

[81] Cui Jinhua，Zhang Youtao，Shi Liang，et al. Approxftl：On the performance and lifetime improvement of 3-D nand flash-based SSDs [J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems，2017，37(10)：1957-1970

[82] Wu Fei，Lu Zuo，Zhou You，et al. OSPADA：One-shot programming aware data allocation policy to improve 3D NAND flash read performance [C] //Proc of the 2018 Int Conf on Computer Design. Piscataway，NJ：IEEE，2018：51-58

[83] Gao Congming，Ye Min，Li Qiao，et al. Constructing large，durable and fast ssd system via reprogramming 3D TLC flash memory [C] //Proc of the 2019 Int Symp on Microarchitecture. New York：ACM，2019：493-505

Gao Congming，born in 1992. PhD. His main research interests include flash memory，storage system，embedded systems.

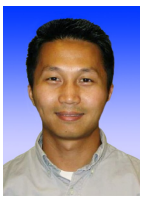
高聪明，1992 年生.博士.主要研究方向为闪存、存储系统、嵌入式系统.

Shi Liang，born in 1987. PhD，professor，PhD supervisor. Member of CCF. His main research interests include flash memory，embedded systems，and emerging non-volatile memory technology.（shi.liang.hk@gmail.com）

石　亮，1987 年生.博士，教授，博士生导师.CCF 会员.主要研究方向为闪存、嵌入式系统、新型非易失性存储技术.

Liu Kai，born in 1984. PhD，professor，PhD supervisor. Member of CCF. His main research interests include Internet of vehicles，mobile computing，pervasive computing and big data.（liukai0807@gmail.com）

刘　凯，1984 年生.博士，教授，博士生导师.CCF 会员.主要研究方向为车联网、移动计算、普适计算、大数据.



Xue Chun，born in 1974. PhD，professor，PhD supervisor. His main research interests include memory and parallelism optimization for embedded systems，software/hardware co-design，real time systems and computer security.（jasonxue@cityu.edu.hk）

薛　春，1974 年生.博士，教授，博士生导师.主要研究方向为嵌入式系统存储和并行优化、软硬件协同设计、实时系统、计算机安全.

Shu Jiwu，born in 1968. PhD，professor，PhD supervisor. Member of CCF. He is an IEEE Fellow. His main research interests include storage security and reliability，non-volatile memory based storage systems，and parallel and distributed computing.

舒继武，1968 年生.博士，教授，博士生导师.CCF 会员，电子工程师协会会士.主要研究方向为存储安全与可靠性、基于非易失内存的存储系统、并行与分布式计算.