计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2014.20120928
51(2):445-453, 2014

新型非易失存储研究

沈志荣[1] 薛 巍[1,2] 舒继武[1,2]
[1](清华大学计算机科学与技术系 北京 100084)
[2](清华大学信息科学与技术国家实验室(筹) 北京 100084)
(czr10@mails.tsinghua.edu.cn)

Research on New Non-Volatile Storage

Shen Zhirong[1], Xue Wei[1,2], and Shu Jiwu[1,2]
[1](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)
[2](*Tsinghua National Laboratory for Information Science and Technology (Tsinghua University), Beijing* 100084)

Abstract Recently, the performance gap between CPU and storage system has been continually increasing, resulting in the consequence that the storage system becomes the bottleneck of performance improvement of the overall computer systems. With the rapid development of microelectronics technology, new non-volatile storage devices that have the metrics of non-volatility, low power consumption, good scalability and shock resistance, are attracting a great attention from academia and industry. This paper introduces several new non-volatile storage devices (i.e., STT-RAM, RRAM, PCRAM and FeRAM) and compares their performance characteristics with those of traditional storage devices. We further discuss the current exploratory works that seek for lower power consumption, higher reliability and better scalability by applying the new non-volatile storage devices to the current three levels of storage architecture (i.e., cache-level, main-memory-level and external-storage-level). A detailed analysis is then presented which focuses on some strategies to mitigate the inherent drawbacks of the new non-volatile storage devices in the application, such as the limited write endurance and the performance imbalance between the read and write operations. Finally, a panoramic summary is given and the possible future development tendencies are discussed.

Key words non-volatile storage; PCRAM; STT-RAM; cache; main memory; external storage

摘 要 近年来,由于处理器性能和存储性能之间的差距不断扩大,存储系统成为计算机整体系统性能提升的瓶颈.随着微电子技术的迅速发展,新型非易失存储器件由于具有非易失、低能耗、良好的可扩展性和抗震等优良特性,得到了学术界和工业界的广泛关注.介绍了4种新型非易失存储器件,分别是STT-RAM,RRAM,PCRAM和FeRAM,对比了其与传统存储器件的性能参数.讨论了目前在存储架构中的不同层面(即缓存层、主存层和外存层)针对这些非易失存储器件的利用所开展的一些探索性工作,并分析了其中针对非易失存储器件的写次数有限、读写性能不均衡等不足所作出的一些策略设计.最后,对新型非易失存储器件的研究现状进行了总结,并提出了未来可能的发展方向.

关键词 非易失存储;PCRAM;STT-RAM;缓存;主存;外存

中图法分类号 TP302

收稿日期:2012-10-17;**修回日期**:2013-04-24

基金项目:国家杰出青年科学基金项目(60925006);国家"八六三"高技术研究发展计划基金项目(2013AA013201)

近年来,随着硬件技术的不断发展,基于传统硬件而搭设的现有计算机体系结构也开始遇到了一些挑战.从访问速度方面考虑,首先在主存和 CPU 之间,由于近年来 CPU 时钟频率的增强、多核技术的快速发展以及多线程技术的广泛使用,CPU 的计算处理能力得到了大幅度的提升,但是主存(DRAM)访问速度的提升则较为缓慢,造成长期存在的计算性能和主存访问性能之间的差距变得越来越大,严重限制了 CPU 计算能力的充分使用.虽然缓存和预取技术的引进在一定程度上缓解了这个问题,但是由于缓存容量小而价格偏贵,因此并未从根本上对这个问题进行解决.其次,在主存与磁盘之间,现今所广泛使用的磁盘由于其机械寻道特性,造成其随机访问速度很难进一步提高,进一步限制了磁盘 I/O 性能的提升.从能耗开销方面考虑,随着数据中心逐渐成为现今计算平台的重要组成部分,数据中心的能耗问题也得到了广泛的关注,而在一个数据中心中,主存耗电和磁盘耗电已经达到了整个数据中心的 40%,而这个比例也将会随着因为寻求更大的吞吐量所导致的更多主存和磁盘的引入而变得更高[1],高能耗的开销也将严重增加系统运营的负担.以上问题共同导致了存储子系统对计算机整体性能的制约越来越突出.

在这种背景条件下,随着近几年微电子技术的飞速发展,一些具有优良特性的新型存储器件陆续推出并得到了学术界和产业界的广泛关注,并为计算机的发展和存储能效的提高带来了新的契机.相比于传统存储器件,新型存储器件具有高集成度、低功耗(包括数据访问和待机情况)、高读写访问速度、非易失、体积小和抗震等优良特性,但是由于其材质和设计原理的不同,这些新型存储器件分别在功耗、读写性能、访问速度和读写寿命上具有不同的特点.现今主流的新型存储器件主要包括 FeRAM(铁电存储器)、PCRAM(相变存储器)、STT-RAM(自旋矩传输磁存储器)以及 RRAM(电阻式存储器).近年来,一些著名高校和企业也开始尝试将这些新型存储器件替代原来所使用的传统存储器件,并移植到现今所使用的计算机体系中,虽然通过这些尝试取得了一些可喜的成果,但是这样依然存在着一些问题.首先,由于新型存储器件介质本身存在的一些特性,例如读写性能不对称、写次数有限等,这些特性为新型非易失存储器件在计算机存储系统中的应用提出了挑战;其次,因为现今所设计的操作系统、文件系统和数据库管理系统等软件都是基于原来磁盘存储系统而构建,而一些相应的数据结构和算法设计的本质目的也是为搭建在磁盘系统之上的相应软件进行功能和性能上的优化,但是由于新型非易失存储器件和磁盘系统之间不同的特性,例如新型存储器件和传统磁盘所不同的寻址机制,因此如果直接将新型非易失性存储器件直接应用于传统的存储系统中时,这些软件和算法将有可能无法充分发挥新器件的特性;再次,正是由于新型非易失存储器件同传统磁盘所不同的特性,一些在原来计算机系统中不存在的问题也将有可能出现.因此,针对新型非易失存储器件的特性进行数据结构和算法的重新设计是一个需要重点研究的问题.

1 新型非易失存储器件

1.1 一些新型非易失存储器件原理

现有的主流新型存储器件主要是根据存储介质在某些外加条件下所产生的不同特性来存储数据,现今引起广泛关注的新型存储器件主要有 FeRAM, PCRAM,STT-RAM 以及 RRAM.其中 FeRAM 的存储原理是基于铁电材料的高介电常数和铁电极化特性,并根据工作模式的不同,可以分为破坏性读出(DRO)和非破坏性读出(NDRO).而 FeRAM 则是基于 DRO 模式,通过利用铁电薄膜的极化反转实现数据的写入和读取.PCRAM 则是利用硫系玻璃(chalcogenide glass)在不同温度下所展现出不同状态的特性,通过电脉冲控制 PCRAM 介质单元的温度来改变单元中介质的状态,利用不同状态所体现出的不同的反光特性和电阻值来表征逻辑值,由于其单元尺寸较小,而且在每个单元之内能够保存多位比特数值,因此存储密度较高[2].STT-RAM 则是使用磁隧道结(magnetic tunnel junctions, MTJ)来实现二元存储[3],即通过改变 MTJ 中的自由层(free layer)和参考层(reference layer)的相对磁力方向来改变 MTJ 的电阻值,以此来记录不同的逻辑状态.相对传统的 MRAM,作为第 2 代的 MRAM 的 STT-RAM 具有更低的能耗和更好的可扩展性[4].RRAM 则是利用强相关电子类材料(例如 NiO 氧化物等)由于施加电压大小和方向的不同而造成电阻变化的原理来记录所需要存储的逻辑值.

1.2 新型存储器件与传统存储器件的比较

新型非易失存储器件由于介质组成和设计原理的不同,因此在单元大小、单元存储位数、读写访问时间、读写寿命和能耗开销等性能参数上都有各自

的特点.表1列举了多种新型非易失存储器件(包括FeRAM,PCRAM,STT-RAM和MRAM)分别同SRAM,DRAM,NOR Flash和NAND Flash在不同性能参数上的对比(在Flash方面,文献[5]详细介绍了Flash存储介质的特性,并讨论了管理Flash的两种软件体系结构,同时着重分析了工业界和学术界针对Flash的特性所发展的一些关键技术).

从表1的性能参数对比中可以分析出一些介质的特性和可能所适用的环境,并得到如下的结论.

1)针对单个存储器件的性能指标进行纵向比较,可以发现:

① 一些新型非易失存储器件写操作所需要的时间比读操作更长,存在着读写不均衡的问题,例如PCRAM,RRAM和STT-RAM;

② 一些非易失存储器件的写操作所需能耗较高,例如MRAM和PCRAM.

2)针对不同的存储器件的相同性能指标进行横向比较,可以得到以下结论:

① 从耐久性方面(endurance)考虑,相比于传统存储器件(即SRAM和DRAM),一些新型非易失存储器件存在着写耐久性较差的缺点,例如PCRAM和RRAM;另外,STT-RAM的写耐久性和SRAM,DRAM相当;

② 从功耗和集成度方面考虑,相比于传统存储器件,一些新型存储器件具有低功耗和高集成度的特点,例如PCRAM的片上单元大小为4~4.8F^2(其中F是单元密度度量单位),同时每个单元含有1~2个位等;

③ 从写操作所需时间上考虑,新型非易失存储器件的写操作所需要的时间普遍高于传统存储器件;

④ 从读操作所需时间方面上考虑,新型非易失存储器件和传统存储器件之间存在着一些相似的情况,例如STT-RAM的访问时间和SRAM相似,PCRAM的访问时间接近于DRAM;

⑤ 从非易失方面考虑,相比于传统存储器件,新型非易失存储器件具有非易失特性,而不需要定时充电保存数据.

Table1 Characteristics Comparison of Many Storage Devices[6]

表1 多种存储介质参数对比[6]

Metrics	Existing Products						Prototype		
	SRAM	DRAM	Flash(NOR)	Flash(NAND)	FeRAM	MRAM	PRAM	RRAM	STT-RAM
Non-Volatile	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cell Size/F^2	50~120	6~10	10	5	15~34	16~40	6~12	6~10	6~20
Read Time/ns	1~100	30	10	50	20~80	3~20	20~50	10~50	2~20
Write/Erase Time/ns	1~100	15	1μs/10ms	1ms/0.1ms	50	3~20	60~120	10~50	2~20
Endurance	1016	1016	105	105	1012	>1015	108	108	>1015
Write Power	Low	Low	Very High	Very High	Low	High	High	Low	Low
Other Power Consumption	Current Leakage	Refresh Current	None	None	None	None	None	None	None
High Voltage Required/V	No	3	6~8	16~20	2~3	3	1.5~3	1.5~3	<1.5

Note: The device labeling the value "a/b" in "write/erase time" comparison means its cell needs to be erased before being re-written and the value "a" (resp. "b") is the corresponding write (resp. erase) time. The device labeling the value "c" indicates its cell can be overwritten without being erased first and "c" denotes the write time.

1.3 新型非易失存储器件的地位和影响

新型非易失存储器件的快速发展为未来计算机体系结构的发展提供了新的选择,针对这些新型非易失存储器件的特性,近年来的一些新型非易失存储器件的工作主要可以划分为以下3个方面:

1)在不同的存储层次中尝试使用与传统存储器件性能相近的新型非易失存储器件,从而达到减少能耗、提高集成度等目的,并探索如何缓解计算性能和存储性能之间的差距.另外,在使用这些新型存储器件替代传统存储器件的过程中,针对一些新型非易失存储器件本身所存在的另一些缺陷(例如写次数有限、读写开销不均衡等)和该存储层次的访问特性,设计相应的优化策略,从而达到最大化利用新型存储器件优势,同时又尽量降低其他特性所带来的性能影响.

2)针对新型非易失存储器件的引入所带来的

一些在以前的体系结构所不存在的问题进行发掘和研究,例如由于新型存储器件的非易失性所带来的数据安全[7-8]问题.

3) 根据新型非易失存储器件的特性,对基于传统存储器件上构建的软件进行重新设计和优化等,例如根据新型非易失存储器件的以字节寻址特性,设计相应的文件系统等.

2 新型非易失存储器件在体系结构中的应用

由于一些新型非易失存储器件在一些性能指标上和传统存储器件相似,同时在能耗开销和可扩展性上较传统存储器件更具有一定的优势,因此研究者开始尝试分别在体系结构的多个存储层次中(即缓存、主存和外存)引入新型非易失存储器件,并根据每个层次的数据访问特征,针对新型非易失存储器件存在的写能耗高和写次数有限等缺陷制定相应的访问策略.本节分别对近年来针对新型非易失存储器件在体系结构中的应用进行介绍和分析,即分别在缓存、主存和外存3个层面上介绍相应的工作.

2.1 基于 STT-RAM 的片上缓存系统

随着多核技术的发展,处理速度越来越快的CPU对缓存容量的需求逐渐增大,造成了当前所使用的缓存器件SRAM面临着以下几个挑战:首先,由于SRAM是通过其上电容电荷的数目来保存数据,需要依靠充放电过程进行数据写入,而这种电荷的特性也导致其集成度有限,当需要增加缓存容量时,SRAM的体积大小则成为需要考虑的问题;其次,SRAM的集成度有限,目前认定的极限是5 nm,并且随着集成度的增高,其功耗也将随之增大,从而带来片上的散热难题.由于STT-RAM集成了SRAM的访问速度、DRAM的集成度,具有非易失性的同时又具有很好的可扩展性,因此被认为是用来替代SRAM较为合适的选择[3].但是由于STT-RAM具有更高的写访问延迟和写能耗,这些特性将极大影响缓存系统的性能,因此在设计基于STT-RAM的缓存系统时并不直接将数据写入,而是采取一些降低写入次数和写入开销的策略.

2.1.1 基于 STT-RAM 的片上缓存系统的写延迟和写能耗

虽然STT-RAM相比于SRAM具有更高集成度,且没有静态能耗,但是由于STT-RAM具有更高的写延迟,极大地影响了基于STT-RAM的片上缓存系统的性能.针对STT-RAM用作缓存时导致的写性能不理想的情况,文献[9]采用了读强占的写缓冲策略,即当缓存收到读请求时,如果写请求才刚刚开始则中止写操作,先执行读操作,通过这种方法来降低因为写延迟所带来的系统阻塞问题.文献[10]尝试在片上网络对STT-RAM的缓存系统的写性能进行改善,提出了根据cache bank的负载情况来决定写请求的调度,即尽可能将耗时的写操作通过调度而在更加空闲的cache bank上执行,这样不仅可以避免传统的Round-Robin算法那样需要依序询问每个虚拟通道的请求,而且可以通过写操作的动态调度而提高缓存的读写性能.

针对STT-RAM的写能耗过高的问题,文献[11]利用STT-RAM的读写性能不均衡特性提出了一种写提前终止策略,即当收到对STT-RAM的写操作时,则先从准备写入单元中读出其中的状态,并将其与准备写入的状态进行比较,如果发现读出的状态和写入的状态相同,则终止写操作,这种方法能够明显避免不必要的写操作,因此达到了降低写能耗的目的.文献[12]设计了一种基于STT-RAM缓存层次化(cache hierarchy),能够降低动态功耗.需要注意的是,以上针对缓存系统的写延迟研究在一定程度上将会给缓存系统的写能耗带来影响.

2.1.2 基于 STT-RAM 的片上缓存系统的性能

由于新型非易失存储器件STT-RAM比SRAM具有更高的集成度,可以扩展SRAM以增大缓存容量.文献[13]尝试通过采用基于新型非易失存储器件的混合Cache技术缩短处理器与存储器之间的性能差距,并提出了两种混合Cache结构:层间混合缓存(level hybrid cache architecture, LHCA)和层内混合缓存(region hybrid cache architecture, RHCA),同时讨论了最新的3D堆叠技术,最后通过实验分别测试了这3种技术组合后的情况,得到的结论显示:利用混合缓存结构后处理器性能指标IPC提高了7%~18%.文献[4]提出STT-RAM的芯片设计,从寄存器、缓存、主存控制器、浮点部件、逻辑控制甚至是流水部件都给出了基于STT-RAM的设计方案,实验结果显示当采用提出的方法后,静态功耗为原功耗的47.6%,而性能可以达到原来的97%.

2.2 基于 PCRAM 的主存系统

过去的几十年,主存的集成度越来越高,其容量越来越大,虽然其访问延迟并没有显著的改善,但是容量的提升在性能、成本上也给计算机系统的发展带来了很大的益处.由于DRAM为充电型器件,其电容需要达到一定的大小才能保存足够的电荷而被

感知.同时其上的半导体器件只有达到某些特定的大小才能产生有效的通道控制信号,且需要动态刷新来保存数据,由于目前预测理论上的最小值为40 nm,而当前工艺已经接近极限,导致传统主存的可扩展性遇到瓶颈.另外,主存容量增大也使得主存耗电和散热问题变得越来越突出,而主存系统的断电也将造成其中所存储数据的全部丢失,导致了一些大型实验对主存可靠性的要求较高.随着新型非易失存储器件研究的逐步发展,研究者们开始希望通过设计基于新型非易失存储器件的主存系统来解决上述问题.由于PCRAM具有同DRAM相接近的读访问性能同时具有较高的集成度,所以被认为是替代DRAM候选方案[2].例如文献[14]设计了一种PCRAM与DRAM混合的主存系统,其中DRAM用来作PCRAM的缓冲区(buffer),然后针对存在的问题提出了一系列解决方案,包括延迟写操作(lazy-write operation)、线级别写操作(line-level writes)、细粒度的磨损均衡(fine-grained wear leveling)等.相对于传统的主存系统,基于PCRAM的主存系统虽然具有非易失、存储集成度较高等优点,但是同时也存在着高延迟、读写不均衡和写次数有限等不足,同时其非易失性也带来了主存系统中的数据安全性问题.针对这些问题,目前主要从以下几个方面来进行研究.

2.2.1 基于PCRAM主存系统的读写访问延迟

由于对PCRAM的写分为Program和Verification两个阶段,如此迭代反复进行,不同单元将数据写入所需要的迭代次数是不同的,但为了保证能将数据正确写入,目前PCRAM所取的是最大迭代次数,这大大延长了写延迟.文献[15]提出了一种通过减少写的迭代次数来减少写延迟的办法,但是迭代次数变少后有些数据并没有正确写入,导致其数据的出错率变高.作者后来通过采用能容忍更多位错误的较强ECC来进行补救,为了避免强ECC带来的开销,提供了ECC的动态转换开关,当存储数据放在具有较高可靠性的SLC中时,可以采用具有低开销的弱ECC.文献[16]针对PCRAM主存系统所采用的ECC计算开销过大的情况,设计了一种按需分配的ECC策略,即根据不同单元的可靠性差异,为其分配不同强弱程度的ECC策略,该方法能够明显减少ECC操作所带来的开销.针对基于MLC的PCRAM在增大容量的同时会导致性能出现降低的情况,文献[17]在容量和性能之间进行权衡,即针对容量要求不高的情况,将设备作为SLC使用,从而达到提高读写速率的目的;针对容量要求增大的情况,将设备作为MLC使用,但是这种方法需要对主存硬件控制器和相应的管理软件进行重新设计,复杂度较高.PCRAM较高的写延迟不仅给写性能带来了影响,同时也因为阻塞读操作而影响读性能.针对这个问题,文献[2]提出了一种基于写取消和写暂停的策略,即当收到读请求时,若此时响应的写请求还没开始,则取消写请求,转而先执行读请求,若写请求开始执行而并未结束,则在某个迭代结束后暂停写请求操作,转而先执行读请求操作.虽然这种方法可以提高系统处理操作请求的效率,但是亦容易导致读写操作不一致的情况.

2.2.2 基于PCRAM的主存系统的动态能耗

根据每次写操作的写入数据有可能和之前存储数据相同的发现,文献[18]提出了一种写冗余去除的方法来减少写的数据量.在执行每次写操作之前,将写入之前的数据读出并进行比较,最后只将修改的部分写入,从而达到降低写能耗并延长PCRAM使用寿命的目的.这种方法虽然能够有效减少写入的数据量,但是PCRAM局部写的情况也将变得更加严重.针对这个额外问题,该文增加一个偏移器,通过将每次写入的部分依次偏移,从而达到避免对某个块进行频繁写的效果,这种损耗均衡实现起来十分简单并且可以达到较好的效果.文献[19]在文献[18]所提方法上作了进一步的改进,即在每次写入时依然比较写入的数据和与准备写入单元上存储的原始数据,如果差异数目小于1/2,则将差异的数据直接写入;如果差异数量大于或者等于1/2,将要写入的数据全部取反,则取反后的数据和原始数据的差异小于或者等于1/2,并在写入数据的同时需要在缓存线或缓存段(cache line)上增加一个标记位并将其置位.在读取数据时需要将标记位读出,如果显示该标记位被置位,则读出的数据需要经过取反才能恢复原来写入的真实数据.这种方法能够明显地降低写入的数据量,不仅能够降低写延迟带来的性能损耗,同时也有益于PCRAM的耐久性的增加.另外,文献[20]中提出使用主存控制器(memory controller)检测访问模式,在DRAM和PCRAM之间实现了基于页迁移的混合主存系统,提高了系统的性能.

2.2.3 基于PCRAM的主存系统的耐久性

由于PCRAM具有擦写次数的限制,因此设计策略延长PCRAM的耐久性成为一个需要考虑的问题.前面所提到的关于针对减少PCRAM与DRAM

的延迟差距、减少 PCRAM 的写操作能耗等方面中所使用的一些技术,同样能够对 PCRAM 的耐久性延长起到一定的作用.除此之外,文献[21]引入了随机化的思想来分散热点数据的位置,以此来降低热点数据密集区域因为热点数据频繁写操作而带来的介质磨损,从而将介质的磨损程度均匀化.文献[22]通过改进 buffer 组织和部分写(partial write)策略来减少对 PCRAM 的写次数,从而增加了 PCRAM 的循环可擦写次数.为了避免一些无用的写操作,文献[23]引入了无用写回(useless write-back)的概念,即在对主存的两次写操作期间,某个块在这个时间内被其他块挤出缓存而导致的写回操作,并定义这个时间段为 dead region,但是该方法需要系统进行一些额外的操作,例如维护一些额外的信息用于记录 dead region,并需要增加一条专门的额外的指令来把这些信息传递给底层的存储设备等.

2.2.4 基于 PCRAM 的主存系统中的数据安全

传统的基于 DRAM 的主存系统需要定期充电来保存数据状态,而系统断电将会导致数据从主存中丢失,因此在系统重新启动的阶段,主存需要重新装载原来的数据以恢复到断电之前的状态.虽然新型非易失存储器件能够在系统断电之后保留数据,从而加快系统启动和休眠恢复的时间,但是其中的数据也可能被非法用户通过物理途径获取,例如通过窃取主存等手段,从而得知系统当前的状态和其中所使用的数据,这对数据的安全性和用户的隐私都将带来一定的威胁.

针对这类问题,文献[7]提出基于 PCRAM 主存系统加密的方案,并对加密产生的损耗均衡方法的改变进行了说明,给出了基于加密系统的损耗均衡改进方案.这种加密虽然保护了数据安全,但是其开销太大.针对这个问题,文献[8]提出了增量式加密方法,所根据的原则是系统中真正所在使用的主存数据在整个主存数据中所占的比例较低,因此可以对暂时不用的数据进行加密,而使用的数据采用明文进行存储,在这种系统中的数据大部分都是以密文进行存储,同时掉电后也可以快速完成加密过程,达到与 DRAM 一样的数据保护时间.但是该方法需要精确的动态预测以及使用主存中的数据集,这种预测不仅需要一定的开销,同时其预测效果在不同的应用下也有很大的区别.

2.3 基于新型非易失存储器件的外存系统

由于传统的机械磁盘在读写数据时需要进行寻道等慢速机械操作,导致随机访问性能不佳,相反,新型非易失存储器件并没有磁头寻址所带来的时间开销,同时具有更快的访问速度,因此使用新型非易失存储器件来替代传统的磁盘将带来极大的访问性能提升.但是由于新型非易失存储器件的字节访问寻址机制不同于传统磁盘的块级别访问寻址,同时考虑到其非易失特性对传统I/O所带来的影响,因此当设计基于新型存储器件的外存系统时,主要开展了以下几个方面的研究.

2.3.1 基于新型非易失存储器件的外存系统接口

由于现今计算机访问外存的接口是使用 seek/read/write 三种基本原语,通过基于块的接口来访问外存.当以非易失存储器件作为外存时,虽然可以通过使用这些接口来达到对现有系统的兼容,但是这种方法不能充分利用非易失存储器件字节寻址以及持久化的特性,从而提高性能及访问的灵活性.针对现有的文件系统并没有充分考虑和利用存储级内存(storage class memory, SCM)的特点,即 SCM 可以直接连接到内存总线上并具有字节寻址及非易失性等优良特性,文献[24]提出了在虚拟地址空间实现的文件系统 SCMFS. SCMFS 利用已有的操作系统中的主存管理模块进行块管理,使得每个文件所占用的空间保持连续性,这种简单性使得 SCMFS 不仅简单易实现而且提高了性能.文献[25]利用持久化、字节寻址特性提出了一种文件系统 BPFS,使用了短路影子分页(short-circuit shadow paging)的技术以对持久化存储进行原子的、细粒度的更新. BPFS 比传统的文件系统提供了更强的可靠性保证及更好的性能.除此之外,该文还提出了一种硬件架构,不仅保证了 BPFS 所需要的原子性,而且还能继续利用 L1 和 L2 Cache 所带来的性能优势.

文献[26]也从操作系统的角度,通过改进页调度来配合 PCRAM,降低其写入次数.文献[27]认为在目前的存储系统中,若用存储级内存替换硬盘,由于基于块的接口不能向设备提供足够的信息,因此将无法针对特定设备的特性进行数据管理的优化.针对此问题,他们设计了基于对象的 SCMs,并实现了一个基于对象的原型系统.

2.3.2 基于新型非易失存储器件的外存系统的访问机制

由于非易失存储器件的高速、字节寻址特性与主存类似,而其具有的持久性与外存类似,使得以非易失存储作为数据存储介质时,可以打破传统存储系统中主存与外存的界限.考虑到非易失存储器件的高速低延迟,如果采用传统的数据访问方式,则操

作系统的 I/O 调度在总开销中占用的比例过大,根据这个问题,文献[28]提出了面向非易失存储的事务性接口 NVTM,它允许编程人员实现快速的、可伸缩的、可持久的数据结构,并能面对各种系统失效保持鲁棒性. NVTM 把非易失存储直接映射到应用程序的地址空间,允许易失及非易失数据结构在程序中无缝交互,从而在基本读写操作等关键性路径中移除操作系统的介入,大幅提高了性能. 为了充分发挥非易失存储的性能,文献[29]认为在简单阻塞 I/O 原语之上需要提供新的存储原语,该文提出的"原子写"原语可以把多个 I/O 操作组成一个独立的逻辑组,并进行整体持久化或者失败回滚. 通过将写原子操作下放到存储设备中,从而达到显著地减小上层应用、文件系统、操作系统为保证数据的一致性和完整性所需要做的工作. 文献[30]则指出,由于 PCRAM 等非易失存储的访问延迟比基于 Flash 的 SSD 低一个数量级,因而在基于 Flash 的系统中影响很少的软件开销在使用非易失存储的系统中则会成为严重的性能瓶颈. 他们由此提出一种新型非易失的存储硬件和软件架构,能够消除两种引起开销的源头,即进入内核以及进行文件系统权限检查. 该架构为每个进程提供了一个私有的虚拟接口,并把文件系统保护检查移到硬件中,因此应用程序可以不经过操作系统的干预访问文件数据,对于多数访问来说可以彻底去除操作系统和文件系统的开销. 文献[31]发现在分布式缓存系统 memcached 及 NoSQL 系统等框架下,易失数据及数据的持久化拷贝之间并无不同. 为通过使用非易失存储以实现性能最大化,提出了一致持久的数据结构(consistent and durable data structures, CDDSs). CDDSs 使用版本化以允许无需日志的原子更新,同样的版本化方式也用来进行失效恢复的回滚,使得系统的吞吐率能够提高数倍. 文献[32]则从操作系统层次上改变了写入策略,当向某个地址写入内容,先读出这个地址中的内容与当前需要写入的内容进行比较,仅仅写入那些被改变的位,从而降低了总体写入能耗.

2.3.3 基于新型非易失存储器件的混合硬盘设计

由于新型非易失存储器件同传统的机械磁盘二者之间存在着各自的优缺点,例如新型非易失存储器件具有更高的可靠性、更好的并发读性能和更低的能耗开销,但是其又具有读写性能不均衡和写次数有限等限制,而机械磁盘不具有写次数的限制,价格低廉并且技术成熟,因此将新型非易失存储器件与机械磁盘一起构建混合磁盘以同时发挥二者的优势,成为近年来的一个研究热点. 这方面的研究工作根据不同的设计宗旨可以主要分为以下 2 类.

1) 构建混合磁盘以降低外存的能耗开销和可靠性. 文献[30]设计了基于 PCM 的存储阵列原型 Onyx,并对比了其与基于 Flash 的固态硬盘的读写性能,其中 Onyx 可以通过 PCIe 接口同主机相连. 但是该原型容量较小,仅为 10 GB. 文献[33]为了更好地延长磁盘的备用(standby)时间并提高可靠性,提出了一种基于新型非易失存储器件的磁盘缓存——NVCache. 在 NVCache 中,有写缓存(write-cache)和预取读缓存(prefetch read cache),其中在写缓存中会存放一份数据列表,在进行写操作时,如果磁盘处于启动(active)状态且写入的数据包括在列表中,则将数据直接写入并删除列表中的数据项. 如果磁盘是处于备用(standby)状态时且有读请求到达,则先尝试所需要读的数据是否包含在列表中,不包含才启动磁盘. 最终所进行的实验表明当结合 NVCache 和自适应磁盘的 spin-down 算法,磁盘的能耗能够降低 90%. 文献[34]提出将 PCRAM 作为磁盘的写缓存并构建出基于 PCRAM 和磁盘的混合存储结构. 为了提高 PCRAM 的写耐久性,该文将全局地址空间视为多维几何空间(multidimensional geometric space),并利用基于空间填充曲线算法(space filling curve-based algorithm)让数据访问均匀地分布在不同的维度,从而达到磨损均衡的效果;另外,该文还提出一种基于 Hash 的写缓存方案以提高磁盘的随机写性能.

2) 针对新型非易失存储器件的引入而分析外存的可靠性. 文献[35-36]针对 PCM 的可靠性较磁盘更高的特点,考虑了在几种不同的 RAID 方案中引入 PCM 作为冗余盘来增加数据的可靠性,并分别根据磁盘可靠性和 PCM 可靠性分析了在不同 RAID 方案中数据丢失的概率.

3 总结和展望

本文分别从缓存、主存和外存 3 个存储结构层面介绍了目前的一些研究进展,包括各个存储层面针对新型非易失存储器件的特性所作的一些探索性工作. 通过以上的分析可以看出,新型存储器件相对于传统的存储介质而言具有较为明显的优势,但是同时其本身也存在着一些缺陷. 我们认为未来的新型非易失存储器件研究主要在以下 2 个方面:1)在使用新型非易失存储器件时,如何能够根据应用特性

来设计有效策略,以此提高其使用寿命、降低其写能耗和写延迟依然是未来一段时间内的主流方向;2)新型非易失存储器件的特性也将对基于传统存储器件所搭建的相关应用体系带来冲击,如何对这些应用的体系进行调整,从而能够达到无缝地发挥新型非易失存储器件的优良特性,也是未来的一个值得考虑的问题.

由于新型非易失存储器件在体系结构中的应用尚处于起步阶段,很多的成果都局限于实验,并没有进行大规模的推广.这需要新型非易失存储器件的设计、开发和生产的工艺进一步成熟化,同时要求其性能更加稳定,而耐久性能够得到进一步提高方可以实现.从短期来看,新型非易失存储器件还不可能完全替代传统的 SRAM,DRAM 和磁盘,但是由于新型非易失存储器件的特性,它们依然有可能对现今所广泛采用的存储体系结构进行一些冲击.从长期来看,随着新型非易失存储器件研究的不断深入、一些应用需求的驱使和市场化的逐步推广,新型非易失存储器件将为用户的不同应用需求提供更加多样化的存储服务,从而发挥更加重要的作用.

参 考 文 献

[1] Roberts D, Kgil T, Mudge T. Using non-volatile memory to save energy in servers [C] //Proc of the Conf on Design, Automation and Test in Europe. Belgium: European Design and Automation Association, 2009: 743-748

[2] Qureshi M, Franceschini M, Lastras-Montaño L. Improving read performance of phase change memories via write cancellation and write pausing [C] //Proc of the 16th IEEE Int Symp on High Performance Computer Architecture. Piscataway, NJ: IEEE, 2010: 1-11

[3] Mishra A, Dong X, Sun G, et al. Architecting on-chip interconnects for stacked 3D STT-RAM caches in CMPs [C] //Proc of the 38th Int Symp on Computer Architecture. New York: ACM, 2011: 69-80

[4] Guo X, Ipek E, Soyata T. Resistive computation: Avoiding the power wall with low-leakage, stt-mram based computing [C] //Proc of the 37th Int Symp on Computer Architecture. New York: ACM, 2010: 371-382

[5] 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)

[6] Nigam A, Munira K, Ghosh A, et al. Model based study on energy and performance optimization for STT-RAM [C/OL] //2011 Non-Volatile Memories Workshop. 2011. [2013-04-01]. http://nvmw.ucsd.edu/2011/

[7] Kong J, Zhou H. Improving privacy and lifetime of PCM based main memory [C] //Proc of the 40th Annual IEEE/IFIP Int Conf on Dependable Systems and Networks. Piscataway, NJ: IEEE, 2010: 333-342

[8] Chhabra S, Solihin Y. i-NVMM: A secure non-volatile main memory system with incremental encryption [C] //Proc of the 38th Int Symp on Computer Architecture. New York: ACM, 2011: 177-188

[9] Sun G, Dong X, Xie Y, et al. A novel architecture of the 3D stacked MRAM L2 cache for CMPs [C] //Proc of the 15th IEEE Int Symp on High Performance Computer Architecture. Piscataway, NJ: IEEE, 2009: 239-249

[10] Kang H, Ryu K, Lee D, et al. Process variation tolerant all-digital multiphase DLL for DDR3 interface [C] //Proc of IEEE Custom Integrated Circuits Conference. Piscataway, NJ: IEEE, 2010: 1-4

[11] Zhou P, Zhao B, Yang J, et al. Energy reduction for STTRAM using early write termination [C] //Proc of IEEE/ACM 2009 Int Conf on Computer-Aided Design. Piscataway, NJ: IEEE, 2009: 264-268

[12] Smullen C, Mohan V, Nigam A, et al. Relaxing non-volatility for fast and energy-efficient STTRAM caches [C] //Proc of the 17th IEEE Int Symp on High Performance Computer Architecture. Piscataway, NJ: IEEE, 2011: 50-61

[13] Wu X, Li J, Zhang L, et al. Hybrid cache architecture with disparate memory technologies [C] //Proc of the 36th Int Symp on Computer Architecture. New York: ACM, 2009: 34-45

[14] Qureshi M, Srinivasan V, Rivers J. Scalable high performance main memory system using phase-change memory technology [C] //Proc of the 36th Int Symp on Computer Architecture. New York: ACM, 2009: 24-33

[15] Jiang L, Zhao B, Zhang Y, et al. Improving write operations in MLC phase change memory [C] //Proc of the 18th IEEE Int Symp on High Performance Computer Architecture. Piscataway, NJ: IEEE, 2012: 1-10

[16] Qureshi M. Pay-As-You-Go: Low-overhead hard-error correction for phase change memories [C] //Proc of the 44th IEEE/ACM Int Symp on Microarchitecture. New York: ACM, 2011: 318-328

[17] Qureshi M, Franceschini M, Lastras L, et al. Memory system: A robust architecture for exploiting multi-Level phase change memories [C] //Proc of the 37th Int Symp on Computer Architecture. New York: ACM, 2010: 153-162

[18] Zhou P, Zhao B, Yang J, et al. A durable and energy efficient main memory using phase change memory technology [C] //Proc of the 36th Int Symp on Computer Architecture. New York: ACM, 2009: 14-23

[19] Cho S, Lee H. Flip-N-Writes: A simple deterministic technique to improve pram write performance, energy and endurance [C] //Proc of the 42nd IEEE/ACM Int Symp on Microarchitecture. New York: ACM, 2009: 347-357

[20] Ramos L, Gorbatov E, Bianchini R. Page placement in hybrid memory systems [C] //Proc of 25th Int Conf on Supercomputing. New York: ACM, 2011: 85-95

[21] Franceschin M, Qureshi M, Karidis J. Architectural solutions for storage-class memory in main memory [C/OL] //2010 Non-volatile Memories Workshop. 2010. [2013-04-01]. http://nvmw.ucsd.edu/2010/

[22] Lee B, Ipek E, Mutlu O, et al. Architecting phase change memory as a scalable dram alternative [C] //Proc of the 36th Int Symp on Computer Architecture. New York: ACM, 2009: 2-13

[23] Bock S, Childers B, Melhem R, et al. Analyzing the impact of useless write-backs on endurance and energy consumption of PCM main memory [C] //Proc of IEEE Int Symp on Performance Analysis of Systems and Software. Piscataway, NJ: IEEE, 2011: 56-65

[24] Wu X, Reddy N. SCMFS: A file system for storage class memory [C] //Proc of 2011 Int Conf for High Performance Computing, Networking, Storage and Analysis. New York: ACM, 2011: 69-80

[25] Condit J, Nightingale E, Frost C. Better I/O through byte-addressable, persistent memory [C] //Proc of the 22nd ACM Symp on Operating Systems Principles. New York: ACM, 2009: 133-146

[26] Zhang W, Li T. Exploring phase change memory and 3D die-stacking for power/thermal friendly, fast and durable memory architectures [C] //Proc of the 18th Int Conf on Parallel Architectures and Compilation Techniques. Piscataway, NJ: IEEE, 2009: 101-112

[27] Kang Y, Yang J, Miller E. Object-based SCM: An efficient interface for storage class memories [C] //Proc of the 27th IEEE Symp on Mass Storage Systems and Technologies. Piscataway, NJ: IEEE, 2011: 1-12

[28] Coburn J, Caulfield A, Grupp L, et al. NVTM: A transactional interface for next-generation non-volatile memories [R/OL]. [2013-04-01]. http://csetechrep.ucsd.edu/Dienst/UI/2.0/Describe/ncstrl.ucsd_cse/CS2009-0948

[29] Ouyang X, Nellans D, Wipfel R, et al. Beyond block I/O: Rethinking traditional storage primitives [C] //Proc of the 17th IEEE Int Symp on High Performance Computer Architecture. Piscataway, NJ: IEEE, 2011: 301-311

[30] Akel A, Caulfield A, Mollov T, et al. Onyx: A protoype phase change memory storage array [C] //Proc of the 3rd USENIX Workshop on Hot Topics in Storage and File Systems. Berkeley: USENIX Association, 2011: 2-2

[31] Venkataraman S, Tolia N, Ranganathan P, et al. Consistent and durable data structures for non-volatile byte-addressable memory [C] //Proc of the 9th USENIX Conf on File and Storage Technologies. Berkeley: USENIX Association, 2011: 5-5

[32] Xu W, Liu J, Zhang T. Data manipulation techniques to reduce phase change memory write energy [C] //Proc of the 14th ACM/IEEE Int Symp on Low Power Electronics and Design. New York: ACM, 2009: 237-242

[33] Bisson T, Brandt S, Long D. NVCache: Increasing the effectiveness of disk spin-down algorithms with caching [C] //Proc of the 14th IEEE Int Symp on Modeling, Analysis, and Simulation. Piscataway, NJ: IEEE, 2006: 422-432

[34] Liu Z, Wang B, Carpenter P, et al. PCM-Based durable write cache for fast disk I/O [C] //Proc of the 20th IEEE Int Symp on Modeling, Analysis and Simulation of Computer and Telecommunication Systems. Piscataway, NJ: IEEE, 2012: 451-458

[35] Paris J, Amer A, Long D. Using storage class memories to increase the reliability of two-dimensional RAID arrays [C] //Proc of the 17th IEEE Int Symp on Modeling, Analysis, and Simulation. Piscataway, NJ: IEEE, 2009: 1-8

[36] Chaarawi S, Paris J, Amer A, et al. Using a shared storage class memory device to improve the reliability of RAID arrays [C] //Proc of the 5th Petascale Data Storage Workshop. Piscataway, NJ: IEEE, 2010: 1-5

Shen Zhirong, born in 1987. PhD candidate. His research interests include storage reliability and storage security (czr10@mails.tsinghua.edu.cn).

Xue Wei, born in 1974. PhD, associate professor. His research interests include parallel algorithm design, network storage and cluster computing (xuewei@tsinghua.edu.cn).

Shu Jiwu, born in 1968. PhD, professor and PhD supervisor. Senior member of China Computer Federation. His main research interests include network storage and cloud storage, storage security, parallel process technologies, and so on (shujw@tsinghua.edu.cn).