基于PCM的GPU存储系统设计与优化

穆 帅[1] 单书畅[2] 邓仰东[1] 王志华[1]
(清华大学微电子所 北京 100084)[1] (中国科学院计算技术研究所 北京 100084)[2]

摘 要 以相变存储器(PCM)为代表的新型非易失存储器,具有存储密度高和静态功耗低等传统动态随机存取存储器(DRAM)不具备的优势,但是过长的写操作延时会严重影响访存的性能。设计了基于PCM的图形处理器(GPU)中的存储系统。仿真结果显示,GPU程序中的内存写请求分布极不均匀,对少量的内存地址有非常高的访问频率。面向访存分布不均匀特点的专用缓冲单元设计,能够有效地存储频繁访问的内存数据,从而减少对PCM的访问次数,消除过长的写操作延时对系统性能的负面影响。GPU仿真器上的结果显示,基于缓冲单元的PCM存储系统能够有效地提高GPU的运算性能。

关键词 相变存储器,图形处理器,缓冲单元,存储系统
中图法分类号 TP391.9 **文献标识码** A

Evaluating and Optimizing of PCM Based GPU Memory Architecture

MU Shuai[1] SHAN Shu-chang[2] DENG Yang-dong[1] WANG Zhi-hua[1]
(Institute of Microelectronics, Tsinghua University, Beijing 100084, China)[1]
(Institute of Computing Technology Chinese Academy of Science, Beijing 100084, China)[2]

Abstract Recently, the emerging non-volatile memory(NVM) exemplified by Phase Change Memory(PCM) has been considered to take the place of conventional DRAM in the processors due to their characteristics of large capacity and little static power. However, the overhead of long write latency can cause severe performance degradation. We evaluated the feasibility of PCM based GPU memory architecture. Based on the analysis of unique memory access behaviors captured from GPU benchmarks, a dedicated buffer was designed to alleviate the pressure of frequently accessing PCM. Simulation results prove the efficiency of our proposed dedicated buffer and show its great potential for PCM based GPU architecture.

Keywords Phase change memory(PCM), Graphics processing unit(GPU), Buffer, Memory hierarchy

1 引言

在现代处理器中,存储系统的性能和功耗逐渐成为至关重要的问题。高性能处理器对于存储系统的性能、容量和功耗都提出了更高的要求[1]。尤其是对于图形处理器(GPU)这种需要大量数据访问的计算系统,存储器的性能将直接决定着整个系统的运算效率。传统的高性能计算系统常采用动态随机存储器(DRAM)作为主存储器。然而,随着晶体管特征尺寸的不断缩小,DRAM面临着严峻的存储密度和静态功耗问题。一方面,由于制造工艺的限制,DRAM的基本存储单元无法进一步缩小物理尺寸(ITRS数据显示,DRAM将无法缩放到22nm工艺以下),存储密度的进一步增加受到了很大的限制;另一方面,电容式存取特性使得DRAM的漏电流功耗和数据刷新功耗问题愈发严重[2]。例如,NVIDIA厂商开发了一款代号为Fermi的GPU处理器,其访问主存数据产生的功耗占据了系统总功耗的40%以上。为了解决传统DRAM所面临的问题,以相变存储器(PCM)为代表的新型非易失存储器(non-volatile memory)得到越来越广泛的关注和应用[3]。相对于传统的DRAM存储器,PCM利用材料自身物理特性存储数据,具有高存储密度、低漏电流功耗等优点。然而,其也具有读写性能差的缺点。本文将讨论在GPU计算系统中,使用PCM代替DRAM的可行性。仿真结果显示,直接使用PCM作为主存储器会带来性能的损失,但是基于GPU程序访存特点设计的专用缓冲单元可以极大地减少访问PCM的次数,消除过长的写操作延时对系统性能的负面影响,从而有效地提高GPU程序的总体性能。同时,本文的仿真结果对于未来进一步设计和优化基于PCM的GPU存储系统提供了重要的实验基础。

2 相关工作

采用新型非易失存储器作为主存储器是当前处理器设计和仿真研究的热点,文献[3]调研了各种非易失存储器的特点

到稿日期:2013-01-07 返修日期:2013-04-03 本文受国家自然科学基金(61272085)资助。
穆 帅(1986—),男,博士生,主要研究方向为GPU架构,E-mail:mus04ster@gmail.com;单书畅(1985—),男,助理研究员,主要研究方向为处理器可靠性设计与缓存优化;邓仰东(1973—),男,副教授,主要研究方向为并行计算和仿真、众核体系架构;王志华(1960—),男,教授,主要研究方向为集成电路和系统的设计方法学、用于医疗和通信的低功耗模拟与射频集成电路设计、高速实时信号处理等。

和典型的实际应用。虽然非易失存储器具有传统 DRAM 不具备的众多优点，但是其过长的写操作延时也会严重影响访存的效率。对于写操作密集的程序而言，性能会受到极大的影响，这就限制了非易失存储器直接应用于当前的处理器中。针对不同的处理器结构，很多工作提出了有效的方法来解决写操作对程序性能的影响。文献[4]评估了使用 PCM 代替 CPU 系统中的 DRAM 的可行性并且定量分析了 PCM 对性能和功耗的影响。该文同时提出了通过重新组织内存访问请求的顺序来有效地降低写操作对性能的影响。文献[5]设计了一种面向 CPU 平台的 DRAM/PCM 混合型存储系统，该工作通过对内存访问的请求进行分类，把频繁写的数据存储在 DRAM 中，而把其余的数据存储在 PCM 中，从而极大地提升 PCM 存储系统的性能。文献[6]提出了在 CPU-GPU 异构系统中实现 DRAM/PRAM 混合存储系统，该设计使用部分 DRAM 存储空间存放 CPU 需要的数据，而其余的 DRAM 和 PCM 存储空间存放 GPU 需要的数据，从而很好地解决 CPU-GPU 异构系统对内存访问延时和带宽的需要。文献[7]考察了使用 PCM 作为流处理器 Imagine 的主存储器，并且对于各种内存调度算法进行了评估。本文的主要贡献在于首次系统地评估使用 PCM 作为 GPU 主存储器的可行性以及性能优化的方法。

3 基于 PCM 的存储系统性能仿真与分析

本节将介绍 PCM 存储器的基本特性，分析使用 PCM 代替 DRAM 对 GPU 程序性能的影响，并通过仿真给出 GPU 程序中访存的分布特点。

3.1 PCM 存储器

相变存储器(PCM)[8]是一种利用相变材料的晶态和非晶态的区别来存储二进制数据的新型存储技术。表 1[9]列出了 PCM 与传统的 SRAM 和 DRAM 在主要性能参数上的差异。与 DRAM 相比，PCM 的主要优势在于更高的存储密度，当作为处理器的主存时，可以有效地扩充存储空间。除此之外，PCM 中的数据是非易失的，不需要额外的能量来维持数据的有效性，因此不会产生漏电流功耗和数据刷新功耗。在数据访问速度上，3 种存储器的读取数据的延时基本保持一致，但是在 PCM 上的数据写回的延时要比 DRAM 长 10 倍以上。因此，直接使用 PCM 代替 DRAM，数据写回操作将会损害程序性能。

表 1 3 种存储技术的参数比较

存储技术	SRAM	DRAM	PCM
单元大小(F^2)	>100	6～8	4～20
读延时(ns)	～10	～10	10～100
写延时(ns)	～10	～10	100～1000
静态功耗	漏电流	漏电流 & 刷新	无
非易失	否	否	是

3.2 基于 PCM 的主存储器

本节中我们将定量考察使用 PCM 作为主存储器对于程序性能的影响。

3.2.1 测试环境

我们使用由 UBC(University of British Columbia)大学开发的具有时钟精度的模拟器 GPGPU-sim[10]。该模拟器能够精确地模拟当前主流 GPU 架构的特性，并且其中的各个模块以及各种参数能够根据仿真要求灵活配置。本次测试模拟了 NVIDIA 厂商开发的代号为 Fermi 的 GPU 架构，表 2 列出了该 GPU 的主要参数配置，包括处理器个数、缓存容量和内存的时序参数。其中内存时序参数一栏显示的是 DRAM 的参数，当考察使用 PCM 作为主存储器时，只需把 DRAM 的时序参数设置成 PCM 的时序参数(即写操作的延时设置成 DRAM 配置下的 10 倍)。

表 2 GPU 模拟器的主要参数配置

名称	参数
SM 的个数	16
每个 SM 支持的线程数	2048
每个 SM 中 SP 的个数	32
L1 缓存的大小	48kB
内存大小	2GB
内存带宽	8bytes/cycle
内存调度算法	FR-RCFS
内存时序参数	tCL=9,tRP=13,tRC=34 tRAS=21,tRCD=12,tRRD=8

3.2.2 测试程序

本文选择 8 个典型的 GPU 程序进行测试。这些程序来自于各种实际应用，例如信号处理、数据挖掘、高性能计算等。所有的测试程序都使用 CUDA 编程语言实现。表 3 给出了测试程序的功能描述。

表 3 测试程序描述

测试程序	功能描述
AES	高级数据加密算法
BFS	广度优先遍历搜索算法
Kmeans	基于划分的聚类算法
Hotspot	估算芯片温度分布的算法
Laplace	信号从时域到复频域的变换
Libor	金融市场中的利率模型算法
Neuralnet	神经网络算法
Raytracing	光线追踪算法

3.2.3 测试结果

图 1 给出了分别使用 PCM 和 DRAM 作为 GPU 主存储器的性能比较，性能的测试指标是 IPC(Instructions Per Cycle)。为了更直观地观察性能变化，我们把 PCM 配置下的程序性能归一化到 DRAM 配置下的程序性能，图中给出的是 PCM 与 DRAM 的性能比值。仿真结果显示，一方面，在 PCM 作为主存储器时，几乎所有的程序都有性能的损失，平均损失高达 35%。这表明过长的写操作延时确实严重影响了整个系统的性能；另一方面，不同的程序性能损失的比例也各不相同。计算密集型程序，例如 AES 和 Neuralnet，由于在执行过程中只产生少量内存访问请求，因此使用 PCM 只会带来很小的性能损失(性能分别减少了 2%和 8%)。但是，对于内存访问密集的程序(BFS 和 Raytracing 等)，性能的恶化非常严重。尤其是 BFS 程序，使用 PCM 的性能只有使用 DRAM 时的 25%左右。

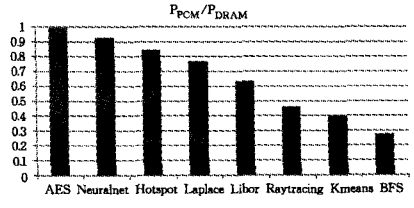
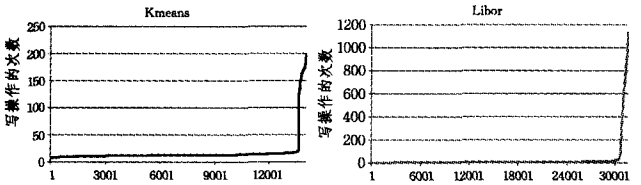


图 1 使用 PCM 和 DRAM 作为主存的性能比较

为了优化 PCM 存储系统的性能，需要进一步研究 GPU

程序中内存写操作的特点。我们通过模拟器统计了内存中发生在不同行的写操作的次数(这里内存行的容量为 1kB)。我们发现,在大部分程序中,访存分布非常不均匀,即写操作的访问地址集中分布在极少数的内存行中。图 2 给出了两个典型测试程序中内存写操作的访问次数分布。其中横坐标表示行的编号,纵坐标表示对该行进行写操作的次数。为了更直观地显示访存特点,访存次数按照从少到多进行了排序。从中可以看出,两个测试程序的访存分布都极不均匀。在 Kmeans 程序中,大约对 3%的数据行的访问次数占据了总访问次数的 30%以上。而在 Libor 程序中,这种不均匀分布的特点更明显:对 2.6%的数据行访问的次数占总访问次数的 66.8%,约 93%的内存行被访问的次数不到 20,然而被访问最频繁的内存行有超过 1000 次的访问次数。在其余的测试程序中,也发现了相似的分布不均匀的特点。



(a) Kmeans 程序　　(b) Libor 程序

图 2　GPU 内存写操作的分布特点

4　基于 PCM 的存储系统优化

在本节中,我们讨论优化 PCM 存储系统的方法,并且评估优化之后的性能改善。

4.1　基于 PCM 的专用缓冲单元设计

由第 3 节的仿真结果得知,GPU 程序的写操作集中分布在少量的内存行中,如果能够把这些行的数据尽可能地存储在专用缓冲单元中,那么就可以极大地减少对 PCM 主存的访问,从而提高程序的执行性能。为此,该缓冲单元的设计需要满足以下两个方面的要求:1)由于程序在执行之前,缓冲单元无法得知哪些内存地址行会被频繁地访问,因此,需要缓冲单元能够在程序运行的过程中动态地识别出频繁被访问的内存行;2)由于对 PCM 的写操作的延时远远大于读操作的延时,因此该缓冲单元应该尽量多地存储内存写操作的数据。图 3 给出了一种满足上述要求的专用缓冲单元(buffer)设计。

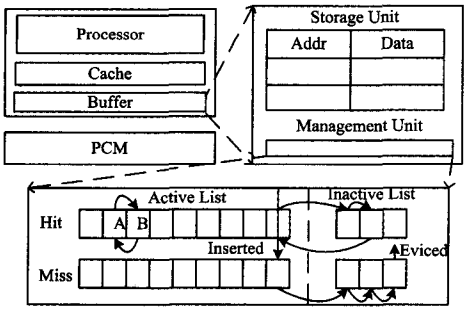


图 3　基于 PCM 的专用缓冲单元结构

该缓冲单元位于缓存(cache)和 PCM 主存之间,由数据存储单元和数据管理单元两部分组成。数据存储单元记录了内存访问请求的行地址和行数据,其中,行数据的空间大小等同于内存行的大小,即 1kB。数据管理单元负责管理存储单元中的数据访问频率以及数据替换的策略。数据管理单元依靠两个列表统计内存地址的访问频繁程度。如图 3 所示,一个列表记录活跃的写访存行地址,而另一个列表记录不太活跃的行地址。列表中的每一项对应于数据存储单元中的行地址,而两个列表的大小等同于数据存储单元中存储的总行数。在每个列表中,列表头部(最左边)表示访问最频繁的内存行,而尾部(最右边)表示访问最不频繁的内存行。图 3 详细给出了列表处理新的访存请求时的策略。

与传统 cache 处理方式不同,该 buffer 在处理访存请求时,读操作和写操作采用不同的方式更新地址列表。具体来说,当访存请求在 cache 访问缺失后,该请求将由专用的 buffer进行处理。访存请求的地址和 buffer 中存放的地址进行比较,将会出现以下两种结果:1)访存请求在访存地址列表中命中,即该访存地址已经存在 buffer 中,则访存请求的数据可以直接从 buffer 中获得,而不必去 PCM 访问数据。如果该请求是读操作,则不对列表进行任何处理。如果该请求是写操作,则按照如图 3 所示的规则操作,即当访存地址属于活跃列表时,该地址在列表中的顺序左移一位;当访存地址属于不活跃列表时,则把该地址替换到活跃列表的尾部,而原尾部的地址移到不活跃列表的头部,其余的地址依次右移一位。2)若该访存地址不在 buffer 中,则需要到 PCM 中读取数据,并且把得到的数据存放在 buffer 中。当 buffer 空间被全部占用时,则需要替换出原有的一行数据来存放新的数据。被替换的数据地址为不活跃列表的尾部,而新的数据地址存放在活跃列表的尾部,不活跃列表中其余的数据地址依次右移一位。

与传统的 cache 结构相比,本文设计的面向 GPU 架构的专用 buffer 有两个明显的特点:1)每行存储数据的粒度更大,buffer 中每一行存储的数据对应于内存中的一行数据(1kB)。因此,GPU 程序中对于不同行访问不均匀的特点能够直观地反映在对 buffer 中行的访问上。而在传统 cache 中,内存行的数据将分布在多条 cache 行中,内存行访问分布不均匀的特点不能直观反映在对 cache 行的访问上;2)专用的 buffer 对读操作和写操作进行差异化的处理,即频繁的写操作会增加该地址在地址列表中的优先级,而读操作不对地址列表优先级产生影响。这种差异化的处理能够实现频繁被写的数据存储在 buffer 中,适应 PCM 读写延时不同的特点,从而减少对 PCM 访问的次数。而在传统的 cache 中,读写操作一般是按照相同的方式处理,并不涉及对于读写操作差异化的处理,因此,不能达到上述的目的。在传统的 cache 结构之外增添了一级专用的存储空间,这样的设计会带来额外的硬件开销,下一节将分析缓冲单元大小对性能的影响以及如何取得性能和硬件开销之间的折中。

4.2　性能测试

本文在 GPGPU-sim 仿真器上实现了 4.1 节介绍的缓冲单元模块,并且考察了不同缓冲单元尺寸下(64kB,256kB 和 1024kB)性能的变化。我们观察了两个重要的指标,即程序总体性能的变化和对 PCM 访问次数的变化。为了更直观地反映变化趋势,所有配置下的测试结果都归一化到没有缓冲单元配置时的测量结果。

图 4 给出了在 3 种缓冲单元尺寸下两个指标的变化。数据表明,增加专用缓冲单元之后,大部分程序的性能都有了明显的提升,这得益于缓冲单元能够极大地减少对于 PCM 读

(下转第 71 页)

点作为中间节点，在中间节点的引导下绕过空洞，由于数据包沿凸包绕过空洞，路径最短。实验结果表明，RTGR 算法相对于 GPSR 算法在平均跳数和分组投递率方面有更好的性能，适用于对网络实时性要求较高、节点能量相对充足的 WMSN。RTGR 算法使用局部广播方式通告空洞信息，增加了节点的能量消耗，在下一步研究中，将就如何有效降低能量消耗进行研究，以进一步提高算法性能。

参 考 文 献

[1] 孙利民，李建中，陈渝. 无线传感器网络[M]. 北京：清华大学出版社，2005：20-24

[2] 张耀，贾振红. 求解路由空洞问题的 GEAR 改进算法[J]. 计算机工程，2008，34(12)：94-96

[3] 田乐，谢东亮，任彪，等. 无线传感器网络贪婪转发策略中的路由空洞问题[J]. 电子与信息学报，2007，29(12)：2996-3000

[4] Karp B，Kung H T. GPSR：Greedy Perimeter Stateless Routing for Wireless Networks[C]//Proc of 6th Annual International Conference on Mobile Computing and Networking. Boston：ACM press，2000：243-254

[5] Ma Xiao-li，Sun Min-te，Zhao Gang，et al. Improving geographical routing for wireless networks with an efficient path pruning algorithm[J]. IEEE Transactions on Vehicle Technology，2008，57(4)：2474-2488

[6] Yan Yu，Govindan R，Estrin D. Geographical and energy-aware routing：A recursive data dissemination protocol for Wireless Sensor Networks[R]. UCLA-CSD TR-01-0023. UCLA Computer Science Department，2001

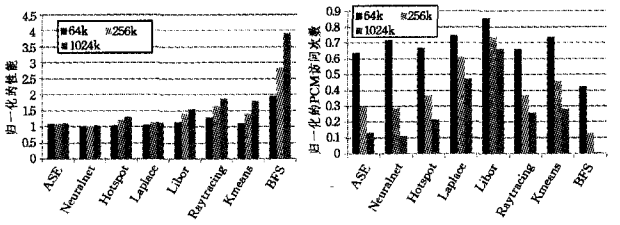
[7] Fang Qing，Gao Jie，Guibas L. Locating and bypassing routing holes in sensor networks[C]//Proc of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies. Hong Kong：IEEE press，2004：2458-2468

[8] Yu Fu-cai，Soo-chang P，Ye Tian，et al. Efficient hole detour scheme for geographic routing in Wireless Sensor Networks[C]//Proc of the 68th Annual IEEE Vehicular Technology Conference. Orlando：IEEE press，2008：153-157

[9] 彭玉旭，郭月，胡立辉. WMSN 中的一种有效地理路由空洞迂回算法[J]. 计算机工程与应用，2012，48(12)：58-62

(上接第 31 页)

写的次数。同时实验结果显示缓冲单元的容积越大，性能提升越明显。具体来说，性能提升幅度在不同测试程序上差异很大。对于读写操作分布极不均匀的程序，例如 Raytracing 和 Kmeans，能够达到接近 2 倍的性能提升。对于读写相对不均匀的程序来说，例如 Hotspot 和 Laplace，性能的提升只有 20%到 30%。AES 和 Neuralnet 这两个程序只产生很少的内存访问并且程序性能受访存延时影响很小，因此，虽然在访存次数上降低的幅度很大，但是总体性能变化不明显。BFS 程序性能提升幅度最大，由于该程序只访问到数百个内存行，当 buffer 尺寸为 1024kB 时，几乎所有的数据都能存放在专用的 buffer 中，因此有非常明显的性能改善。



(a) 程序性能的影响　　(b) PCM 主存访问次数的影响

图 4　不同尺寸的缓冲单元对于性能的影响

实验显示，当 buffer 的容积为 256kB 时，程序的平均性能提升为 46%，基本上接近于 DRAM 作为主存时的性能。当 buffer 的尺寸大于 256kB 时，程序性能可以获得进一步的提升，但是会带来更大的硬件开销；当 buffer 的尺寸小于 256kB 时，性能的改善不是特别理想。当前主流的 GPU 架构中，二级 cache 的设计尺寸至少在 1MB 以上，该专用的 buffer 属于辅助的存储空间，因此尺寸为 256kB 的缓冲单元能够取得性能和硬件开销之间的良好折中。

结束语　本文定量分析了使用相变存储器(PCM)作为 GPU 主存储器带来的优势和挑战。仿真结果显示，直接使用 PCM 作为主存会带来性能上的损失。通过分析 GPU 程序的访存特性可知，提出的专用的缓冲单元的设计能够极大地改善 PCM 作为主存的性能，同时保持 PCM 在存储密度和静态功耗方面的优势。本文的仿真结果对于未来进一步设计和优化基于 PCM 的 GPU 存储系统提供了重要的研究基础和方向。

参 考 文 献

[1] Hennessy J L，Patterson D A. Computer Architecture A Quantitative Approach [M]. San Mateo：Morgan Kaufmann，2011

[2] Freitas R，Wilcke W. Storage-class memory：The next storage system technology [J]. IBM Journal of Research and Development，2008，52(4/5)：439-447

[3] 刘金磊，李琼. 新型非易失相变存储器 PCM 应用研究[J]. 计算机研究与发展，2012，49(S1)：90-93

[4] Lee B C，Ipek E，Mutlu O，et al. Architecting Phase Change Memory as a Scalable DRAM Architecture[J]. Proc. of International Symposium on Computer Architecture，2009，37(3)：2-13

[5] Qureshi M K，Srinivasan V，Rivers J A. Scalable High Performance Main Memory System Using Phase Change Memory Technology [J]. Proc. of International Symposium on Computer Architecture，2009，37(3)：24-33

[6] Kim D，Lee S，Chung J，et al. Hybrid DRAM/PRAM-based Main Memory for Single-Chip CPU/GPU [C]//Proc. of Design Automation Conference. 2012：888-896

[7] 郝秀蕊，安虹，李小强，等. 流处理器的相变存储器主存性能优化[J]. 计算机工程，2011，37(24)：251-256

[8] Raoux R，Burr G W，Breitwisch M J，et al. Phase Change Random Access Memory A Scalable Technology [J]. IBM Journal of Research and Development，2008，52(4/5)：465-479

[9] Dong X Y，Jouppi N P，Xie Y. PCRAMsim System Level Performance Energy and Area Modeling for Phase Change RAM [C]//Proc. of the International Conference on Computer Aided Design. 2009：269-275

[10] Bakhoda A，Yuan G L，Fung W L，et al. Analyzing CUDA Workloads Using a Dedicated GPU Simulator [C]//Proc. of International Symposium on Performance Analysis of Systems and Software. 2009：163-171