

基于新型存储的大数据存储管理

金培权^{1,2}

1. 中国科学技术大学计算机科学与技术学院, 安徽 合肥 230027
2. 中国科学院电磁空间信息重点实验室, 安徽 合肥 230027

摘要

如何高效地存储大数据并支持实时大数据处理与分析是大数据技术发展面临的首要问题。近年来, 以相变存储器、闪存等为代表的新型存储为实现高效的大数据存储和管理提供了新思路。以相变存储器为代表的存储级主存技术为切入点, 针对大数据存储与管理中的高效存储、实时处理等存在的挑战, 讨论了面向新型存储的大数据存储管理研究现状, 并对未来基于新型存储的大数据研究进行了展望。

关键词

相变存储器; 大数据管理; 新型存储; 存储管理

中图分类号: TP311

文献标识码: A

doi: 10.11959/j.issn.2096-0271.2017053

Big data storage management based on new storage

JIN Peiquan^{1,2}

1. School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China
2. Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Hefei 230027, China

Abstract

How to efficiently store big data and support real-time big data processing and analysis has been the most critical issue in the development of big data technologies. Recently, new storage media such as phase change memory and flash memory provides new opportunities for developing an efficient framework for big data storage and management. Based on the challenges of efficient storage and real-time processing in big data storage and management, storage class memories were focused, which were represented by phase change memory, and the state of the art of new-storage-based big data storage management was discussed. Finally, some future research directions for new-storage-based big data storage management were proposed.

Key words

phase change memory, big data management, new storage, storage management

1 引言

大数据已经成为目前的一个研究热点^[1]。如何改进现有的数据存储与管理技术或者设计全新的体系结构,以满足大数据应用中的大数据量和高速数据流实时处理需求,是大数据技术中的核心问题之一。如果采用传统数据库管理系统(database management system, DBMS)的集中式数据存储方式,大数据存取性能就会受到极大的影响。Hadoop技术虽然提供了对大规模数据的快速、低成本存储和管理,但它是一个离线、批量的数据处理系统,对于实时数据处理与分析的支持较弱,难以满足许多应用的要求。例如,在城市公共安全中,通常要求能够对高达每秒几千帧的高清监控视频流进行实时处理与分析。但目前在传统计算体系结构下,单台计算机只能支持每秒150~300帧的低分辨率图像实时异常事件检测^[2,3]。如果要做进一步的目标识别,根据目前的处理技术,性能将下降到每秒16帧左右^[4,5],远远不能满足每秒几千帧高清图像的实时处理要求。因此,迫切需要研究能够满足大数据高效存储与实时处理的新型体系结构与新方法。

针对大数据高效存储与管理问题,目前除了Hadoop技术之外,学术界和工业界也提出了一些其他的设计,包括以NoSQL数据库为代表的大规模分布式数据库系统设计^[6]、基于动态随机存取存储器(dynamic random access memory, DRAM)的内存数据库技术^[7]等。但现有的NoSQL分布式数据库技术仍以磁盘存储或者“磁盘+闪存(flash memory)”混合存储的方式存储数据,本质上还是传统的“CPU-DRAM-二级存储”的存

储架构,依然存在着内存和磁盘之间的“存储墙”问题,难以从本质上解决大数据实时存取的问题。此外,由于DRAM能耗和成本较高,也限制了其在大规模数据处理中的应用。

过去5年来,闪存作为新型存储的代表性技术取得了快速发展,对现有的数据管理技术提出了极大的挑战,同时也带来了许多新的机遇^[8]。但是,闪存由于其存取方式(按页)、存取性能(1次存取通常需要约 2^{17} 个CPU时钟周期)的限制,仍适合作为二级存储器。基于闪存的数据管理只是优化了I/O延迟,并没有从本质上改变计算架构。

除了闪存之外,近年来另一种新型存储介质——相变存储器(phase change memory, PCM)引起了学术界和工业界的广泛关注^[9]。与闪存相比,PCM可以被CPU直接按位存取,而且存取性能更高。因此PCM可以与DRAM一样与CPU交互。但与DRAM相比,PCM存储具有非易失性,能够进行持久的数据存储。传统硬盘基于磁性存储机理存储数据,闪存基于微型电容储存电荷的机理存储数据,存储密度都有理论上限,而PCM基于微型相变单元存储数据的机理使其能够迅速超越固态盘的存储密度,并且在未来还有更大的提升空间。IBM公司把PCM这一类具有DRAM的存取性能,同时又具有持久存储能力的介质称为存储级主存(storage class memory, SCM)^[10,11]。PCM等存储级主存以其非挥发、存储速度快、易实现高密度等技术特点,在高速与海量存储方面具有巨大的潜能,被认为是下一代非易失存储技术的发展方向。另外,因该技术兼有DRAM的高速随机访问和闪存的非易失特性,模糊了主存和外存的界限,有望突破原有的存储架构,实现更高性能的存储。

因此,如果能够利用PCM等新型存储

器件设计出适合大数据存储与管理的新型存储架构(如图1所示),同时设计新的分布式多节点存储技术,则可以将大数据存取集中在DRAM和PCM上,充分发挥DRAM和PCM的高性能特性以及PCM的随机存取和非易失优点,而且可以利用分布式多节点存储的优势建立高扩展的大数据存储系统,从而有望彻底解决大数据存取中的性能与容量问题,为大规模的大数据分析与应用提供有力的支撑。

目前,公共安全、智能交通、物联网等许多应用都要求实现大数据的实时存取。但是,现有的Hadoop等技术还很难达到这一目标,主要的困难在于无法提供低延迟、高吞吐的大数据实时存取能力。新型存储的出现为解决这一难题提供了可能。首先,PCM等非易失内存的出现为实现大规模的内存计算奠定了基础,使得人们有可能在内存中支持高并发的事务处理,而不需要传统DRAM导致的大量I/O操作,从而实现低延迟的大数据存取。其次,借助基于新型存储的分布式内存文件系统等技术^[12],可以大规模提升外存和内存的写吞吐速率。

本文综述了基于新型存储的大数据存储管理技术,分析了现有大数据存储技术

的局限性,介绍了新型存储的特点和发展概况,总结了基于新型存储的大数据存储架构、基于新型存储的大数据存储管理等方向的研究现状,在此基础上给出了基于新型存储的大数据存储与管理的若干未来研究方向。

2 大数据存储技术

2.1 常见的大数据存储技术

目前,大数据存储一般采用分布式存储技术,主要应用在NoSQL数据库系统中。现有的主流的NoSQL数据库系统,例如文档数据库系统MongoDB^①、列存储数据库系统HBase^[13,14]、内存数据库系统Redis^②等,均采用了分布式集群架构实现大数据的存储。也有一些分布式数据库系统在存储架构设计上考虑了异构存储的特性,例如RAMCloud^[15]和RethinkDB^③,从而有效提升了系统的存取性能。

但是,现有的大数据存储技术还存在着以下的局限性^[16]。

- 以NoSQL数据库为代表的大规模分布式数据库系统设计了基于磁盘存储的读写方式、索引结构、查询执行、查询优化、恢复策略,但是磁盘固有的读写性能差等弊端限制了大数据存取尤其是大数据分析性能的提升。

- 在以Hadoop分布式文件系统(Hadoop distributed file system, HDFS)为代表的大规模分布式文件系统中,虽然它们提供了大数据的存储支持能力,但由于这些文件系统在设计时并没有考虑对实时、高性能的数据处理的支持,因此无法满足日益增长的大数据在线分析的需求。此外,随着数据量的急剧增加,元数据的大小也急剧增加,传统的元数据架

①
<http://www.mongodb.org/>

②
<http://redis.io/>

③
<http://rethinkdb.com/>

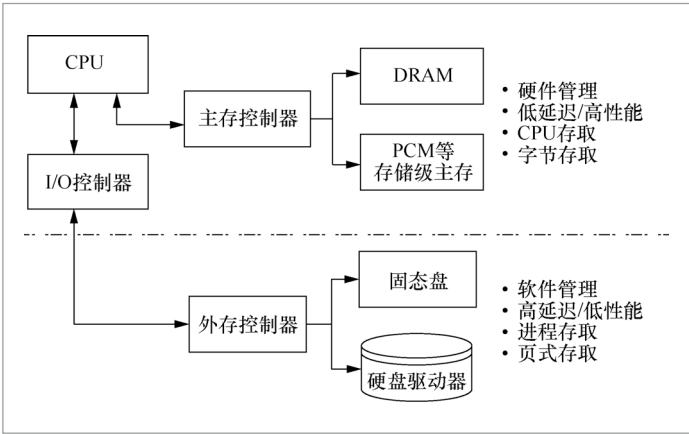


图1 引入 PCM 等存储级主存后的存储体系结构

构、元数据备份管理、元数据动态负载均衡等越来越难适应大数据应用需求。

- 基于DRAM的内存数据管理技术旨在通过海量的内存提高大数据的处理性能。但是,由于DRAM本身能耗高、价格相对昂贵,使得构建基于大内存的大数据存储集群在环境支持、成本上存在较大的困难。此外,DRAM的掉电易失特性导致的大数据环境下的数据一致性也是一个棘手的问题。

2.2 新型存储技术

鉴于磁盘存储、内存存储在面临大数据管理与分析时的困难,学术界和工业界开始将目光转向新型存储技术。目前,从技术成熟度和应用前景上看,闪存和相变存储器最有可能形成大规模应用,因此也吸引了国内外学者的关注^[8,9,16,17]。

闪存是一种可以被电子化擦除和重写的非易失性存储设备^[8]。基于闪存的固态硬盘(solid state drive, SSD)是目前市场上常见的闪存存储设备。与传统的磁存储介质相比,闪存具有传输速率高、低延迟、低能耗、低噪音、抗震等优良特性。同时也有一些特殊性质:写前擦除,对闪存的写操作不是简单地改变某个二进制位,而是需要将整个擦除块的所有二进制位置改为1,这带来了闪存的读写不对称性,一般采用异地更新的方式缓解写前擦除带来的延迟,减少读写不对称带来的影响;寿命限制,目前企业级闪存能耐受3万次写循环,消费级闪存仅为3 000次;读写与擦除的单位不一致,一个擦除块中包含若干个闪存页,擦除的单位是一个闪存擦除块,读写的单位是闪存页。

相变存储器是一种非易失类型的存储器,由硫系玻璃材质构成^[17]。由于这种材质的特质,通过施以电脉冲热,它可以在非

晶态和多晶态这两种状态之间进行切换。PCM兼具速度快、耐用、非挥发性和高密度性等多种优势,其读写数据和恢复数据的速度是闪存的100倍^[18]。

随着云计算和物联网等新一代信息技术的涌现,对海量存储系统的低能耗、高速及高可靠性的需求日益凸显,以新型存储取代传统存储介质的呼声越来越高,而PCM有望成为未来新型存储的主要技术。与DRAM、闪存等存储介质相比,PCM具有非易失性、存取速度快、节能、可字节寻址、写寿命长等优点^[19]。韩国三星(Samsung)公司与美国美光(Micron)公司是目前在PCM技术方面较为领先的两家公司,其中三星公司开发出的65 nm制程、512 MB容量的PCM芯片已投入量产,并应用在三星公司的手机存储卡中;同时三星公司已经推出了20 nm制程、8 GB容量的相变内存颗粒。美光公司已经成功研制了45 nm制程、1 GB容量的LPDDR2接口的PCM芯片产品,并已经量产^[20]。我国中国科学院上海微系统与信息技术研究所近年来也研制了中国自主知识产权的PCM芯片(8 MB),为研制我国自主知识产权的新型存储系统奠定了基础^[21]。此外,华中科技大学自2007年开始研究高密度低功耗的电阻式相变存储器、相变存储器功能芯片、相变存储器芯片的关键材料以及相关专用测试设备等,已经自主研制出具有简单读、擦、写功能的相变存储器功能芯片^[22]。

总体而言,传统的磁盘存储技术在大数据存储与管理方面面临着严重的性能瓶颈。内存数据管理技术由于价格、容量以及易失等特点难以作为PB级大数据存储的最终解决方案,但在大数据存储与管理中可以借鉴内存数据处理的一些思路。闪存、PCM等新型存储器件提供了高性能、非易失的数据存储支持。从目前的技术发展现状看,PCM是现有最为成熟,且性能、

容量与DRAM最为接近的存储技术。PCM以其非挥发、存储速度快、易实现高密度等技术特点以及与CMOS工艺兼容性好、易于与CPU集成形成片上系统(system on chip, SoC)芯片等优点,具有广泛的应用前景。

3 大数据存储架构

新型存储的出现为构建新的大数据存储架构提供了可能。目前,学术界针对基于新型存储的大数据存储架构提出了多种设计,包括基于PCM的主存架构、基于闪存的主存扩展架构、分布式存储与缓存架构等。

3.1 基于PCM的主存架构

PCM与闪存相比,其存取延迟更短,而且可以直接按位存取,因此能够被CPU直接存取,更适合作为DRAM的扩展。与DRAM相比,PCM具有非易失性特点,因此适合存储文件等静态数据。总而言之,PCM可以看作兼有DRAM和闪存的优点。从存储架构设计的角度来看,PCM既可以作为主存使用,也可以作为外存使用。但由于PCM的可字节寻址特性(与闪存不同),目前学术界对基于PCM的主存架构研究相对较多。

在利用PCM替代DRAM方面,理论上可以有两种架构,即纯PCM主存架构和DRAM/PCM混合主存架构。在纯PCM主存架构中,PCM完全替代DRAM作为唯一的主存,而在DRAM/PCM混合主存架构中,DRAM和PCM共同作为主存。在后一种架构中,又存在着两种可能的设计:一是将DRAM作为PCM缓存的层次架构,另一种是DRAM和PCM并列的平等架构。目前,大多数的研究都假设DRAM/PCM的混合主

存架构^[10,11,23-25]。研究者针对DRAM/PCM的混合主存架构,提出了多种PCM写操作优化^[23]以及负载均衡算法^[10,11,24]。由于PCM的写次数有限制,因此如何在混合主存中减少PCM上的写操作是目前的研究重点。

基于PCM的主存架构为实现大数据的实时处理提供了可能。首先,PCM的低能耗特性使得在集群系统中使用大量的PCM存储代替DRAM成为可能,从而降低系统成本。其次,PCM的持久存储特性可以通过设计有效的算法提高分布式存储环境中的数据一致性。第三,PCM的高密度特性可以为内存计算提供有力的支持。

3.2 基于闪存的主存扩展架构

闪存是目前相对较成熟的新型存储技术。基于闪存的SSD已经大量装备在服务器上,成为企业级存储解决方案中的重要组成。由于闪存的整体存取性能优于磁盘,因此理论上可以借助闪存提升大数据存储和管理的性能。在早期的一些研究工作中,研究人员往往假设未来存储系统中闪存可以完全替代磁盘作为外存,但是,由于闪存的读写不均衡特性以及寿命问题,目前实际的系统中往往是DRAM、闪存和磁盘共存。

在DRAM、闪存、磁盘共存的存储架构下,闪存通常作为主存的扩展,即作为DRAM和磁盘之间的中间层,提升大数据存取的性能^[26-28]。SSDAlloc^[26]是基于闪存的主存扩展系统,它将闪存作为磁盘的缓存,实现了系统整体性能的提升。也有一些学者提出了将闪存作为虚拟内存,在DRAM容量不够的情况下,将闪存作为虚拟内存设备进行页面交换^[27]。由于闪存性能总体优于磁盘,因此这种以闪存作为虚拟内存的架构理论上在大数据应用场景下

性能优于传统的DRAM+磁盘的架构。

对于大数据处理而言,基于PCM的主存扩展总体上比基于闪存的主存扩展更具可行性。这是因为大容量的闪存本身仍然采用按页存取的方式,与CPU按位存取模式之间存在不一致性,而且在存取性能上PCM也高于闪存,因此更有望减小与CPU之间的性能差距,构建能够充分发挥CPU、DRAM和PCM各自优势的高性能数据处理系统。

3.3 分布式存储与缓存架构

目前,基于分布式观点的数据管理是大数据存储与管理研究中的一个热点。一种观点是将闪存应用于分布式文件系统中进行元数据存储。元数据对于整个大数据管理系统的性能起着决定性作用,对于大数据解析、大数据统计、大数据操作优化等起着重要作用。基于闪存的分布式文件系统元数据管理的基本思路是在元数据服务器(metadata server, MDS)上使用SSD作为存储设备加速文件系统,如参考文献[29]在Lustre分布式文件系统架构中的元数据服务器上使用闪存作为存储介质,加速元数据的读写速度。此外,基于Memcached的内存分布式缓存技术也被广泛用来加速大规模数据的访问,而在更为复杂的大数据环境下,其局限性主要体现在:一方面内存分布式缓存受限于集群内存容量,只能服务容量较小的热点数据,会造成性能下降;另一方面,如果采取扩大集群内存容量满足更多数据缓存需求,会带来高额的成本和巨大的能耗。现阶段解决方法是将小容量、高I/O负载的缓存处理与大容量、中低等I/O负载的缓存处理分离,形成“热缓存”与“冷缓存”的缓存策略,其中在“冷缓存”方面主要采用了闪存技术。例如,Facebook设计了基于

闪存的键-值存储系统McDipper,代替Memcached为大量访问频率较低的图片提供缓存服务,降低成本和能耗,为了减少闪存I/O延迟,将闪存层分成两个区域,一个区域存放数据,另一个区域配置散列桶存放键值数据的指针,并将散列桶元数据放入内存。

分布式存储技术将是解决大数据存储与管理问题的主要途径之一。一方面是由于Hadoop分布式技术已经为现有的大数据管理提供了一种行之有效的存储方案,而且已经在Google、Facebook等公司的实际应用中得到了验证,为大数据未来研究提供了有用的借鉴;另一方面也是因为在大数据应用中数据来源、用户等本身存在天然分布特性,适合采用分布式存储技术。

4 大数据存储管理

闪存、PCM等新型存储的物理特性、读写特性等均与磁盘有着非常显著的不同,而目前已有的大数据数据库,其设计理念均是基于磁盘存储,在面对闪存、PCM等新型存储时,并不能最大限度地发挥新型存储的性能。目前,在基于新型存储的大数据存储管理方面也有一些研究工作。

在基于PCM的存储管理方面,Ramos L E等人^[30]提出了一种针对DRAM/PCM混合主存的硬件驱动的页面置换策略。该策略依赖一个内存控制器(memory controller, MC)监控内存页面的使用频率和写密集程度。MC在DRAM和PCM之间进行页面迁移,保证性能攸关的页面和频繁写的页面保存在DRAM中,而性能不太敏感以及很少写的页面存储在PCM中。Qureshi M K等人^[23]提出了一个层次型混合主存系统。他们将DRAM设计为

CPU和PCM之间的缓冲区。所有的数据页都存储在PCM中,只有当DRAM发生页面置换或者需要访问新的页面时系统才存取PCM。Wu Z L等人^[31,32]在PCM存储管理方面提出了动态桶列表(dynamic bucket list)以及写敏感的混合时钟存储管理方法。

索引作为优化数据存取性能的重要技术,是数据存储管理中的关键问题之一。传统的B+树索引在数据库系统和文件系统中被广泛应用,近年来在云计算^[33-36]、位置服务^[37,38]等应用中也有一些针对B+树的优化工作。虽然B+树具有很好的搜索性能,但它常常导致较高的更新代价。在面向闪存的数据库领域,研究人员提出了多种针对B+树的改进设计,例如 μ^* -Tree^[39]、BF-Tree^[40]、LA-Tree^[41]、HashTree^[42]、BloomTree^[43]等。这些方法以减少对闪存的写操作为主要目标,采用了利用溢出节点延迟更新、利用额外的缓存节点的更新等方法,最终减少B+树叶节点的更新次数以及索引的合并和分裂操作。

虽然目前在基于闪存的索引设计方面已经有了不少的工作,但由于在大数据存储中引入了PCM等其他类型的新型存储介质,而且在计算架构上产生了根本性的变化(闪存定位在二级存储,而PCM则可以用于直接的内存扩展),因此,近年来研究人员也探讨了针对PCM的B+树索引优化问题。Chen S M等人^[44]最早在2011年的国际创新数据库研究会议(International Conference on Innovative Database Research, CIDR)上测试了B+树在采用了PCM主存技术的服务器上的性能。其研究结果表明,当PCM技术应用到数据库服务器上后,因其具备高速随机访问特性,传统的索引技术应进行新的设计。他们在后续的工作中继续研究了针对PCM等非易失内存的B+树索引结构^[45],类似的工作还有

Hu W W等人^[46]提出的B^P-tree、Chi P等人^[47]提出的写优化B+树以及Li L等人^[48]提出的面向PCM的读写趋势感知的CB+-tree索引。这些工作基本都采用了针对PCM特性优化传统的B+树的思路。

5 未来研究展望

5.1 基于新型存储的大数据存储架构

以PCM为代表的新型存储技术进一步提升了非易失存储的性能极限。PCM类似于DRAM的高速随机访问模式使其有机会直接与CPU连接,而其高密度潜力也使它能够适应大数据时代的容量需求。当存储静态数据的非易失存储允许CPU通过直接寻址的方式访问时,存储体系的进化不仅仅带来性能的大幅提升,同时还将改变应用程序访问数据的方式。由于PCM等存储级主存能够直接支持随机读写,因此可以将其与DRAM共同连接于主存控制器上,与DRAM实现统一编址,CPU可直接寻址到PCM的任何地址。

由于计算机系统的系统集成度较高、构成复杂,不易完成架构改动,因此基于新型存储的大数据存储架构可以采用嵌入式系统方式构建验证用的硬件平台,在平台上直接实现新存储架构及相应软件,从而能够准确地评估新存储架构带来的性能优势。通过搭建新型嵌入式存储架构软硬件验证平台,实现对存储系统的硬件级访问检测,为验证软件系统性能提供准确的数据。

5.2 基于新型存储的分布式内存文件系统

PCM等存储级主存的出现及应用打破

了传统的硬盘驱动器(hard disk drive, HDD)/SDD+DRAM的存储架构,为适应PCM等存储级主存PCM和DRAM共存的新存储架构,需研究新型的可支持以内存访问形式访问各种文件数据的新型文件系统。同时,由于大数据时代数据一般需要分布式存储与计算,因此在文件管理上还需要考虑对分布式环境的支持。因此,需要结合新型存储架构和分布式环境的需求,研究新型的大数据文件系统。该方向的一些研究要点包括以下几方面。

(1) 支持新型存储架构的单节点文件系统

单节点文件系统是研制分布式随机访问内存文件系统的基础,具体包括新型存储架构下的文件原位访问技术、文件系统管理与控制技术、基于新型存储架构的内存管理机制等。

(2) 支持新型存储架构的分布式文件系统

本地节点的数据访问仅能够提升应用程序访问本地数据时的效率。分布式存储技术可以基于新型存储架构搭建支持海量数据存储的分布式环境,从而满足大数据存储的容量需求。因此,将单节点文件系统向多节点扩充,完成支持新型存储架构的分布式内存文件系统,是实现基于新型存储的大数据存储管理的关键,研究要点包括分布式文件系统虚拟访问接口、基于统一寻址的分布式文件管理技术、存储空间的全局划分和寻址技术等。

5.3 基于新型存储的大数据管理

PCM等存储级主存给存储与计算架构带来了极大的挑战,包括异构存储上的数据分配与调整机制、异质缓存管理机制、基于新型存储的大数据索引技术等。

(1) 基于新型存储架构的数据存储分配与调整机制

由于DRAM、PCM、SSD/HDD等多种存储介质同时用于数据存储,因此需要研究一种自适应的多粒度数据存储分配机制。具体而言,该机制首先根据数据访问频度将数据划分为3种状态:热(hot)、温(warm)、冷(cold),然后根据数据的状态进行存储分配与调整。所谓多粒度是指在存储分配时,同时采用文件和页两种粒度。在PCM与SSD/HDD之间进行数据分配时,PCM作为持久存储介质,采用文件粒度进行数据分配;在DRAM与SSD/HDD之间进行数据分配时,DRAM作为缓存,采用页粒度进行数据分配;在DRAM与PCM之间进行数据分配时,以键值记录粒度进行数据迁移和交换。

在数据存储调整方面,一种可能的方法是基于应用对数据的访问模式变化,自适应、动态地调整数据存储策略。访问模式的度量基于数据的访问频度以及存取方式(读/写)两类因素,通过周期性考察的方法确定当前数据访问模式的变化程度,并基于访问模式的变化程度确定是否重新执行数据存储分配。一旦确定了新的数据存储分配策略,将对相应的数据进行介质之间的迁移操作。

(2) 基于新型存储架构的异质缓存管理

数据缓存是传统数据库领域中的核心技术之一,它对于提升系统存取性能有着非常重要的作用。在大数据环境下,由于数据量的急剧增加,数据缓存的重要性尤为突出,因为如果让每个应用直接在全部的大数据上运行将很难保证访问性能。目前一种普遍的观点认为,虽然大数据环境下数据量很大,但对一个具体应用而言,涉及的只是大数据集合中的一部分(小数据)。但是,在新型存储架构下,数据存储涉及了

DRAM、PCM、SSD等具有完全不同访问特性的存储介质,在缓存层也同样面临着多种介质共存的局面,例如数据既可以缓存在DRAM中,也可以缓存在PCM中,甚至也可以缓存在SSD中。这类异质缓存管理问题是传统数据缓存研究中不曾面临的新问题,也是构建基于新型存储的高效大数据管理系统的关键所在,需要首先分析异质缓存管理中的普遍性问题,阐明异质缓存管理的一些新的准则,在此基础上研究新的方法。

(3) 基于新型存储架构的大数据索引

在传统的基于“DRAM+SSD/HDD”的存储架构下,DRAM与外存之间的I/O是影响系统查询处理性能的瓶颈。但在基于新型存储架构的大数据应用系统中,索引的设计不仅要考虑内外存之间的I/O代价,还要考虑异质内存之间的数据迁移代价(从DRAM到PCM以及从PCM到DRAM),此外还要考虑PCM等新型存储的器件特性(例如芯片写次数有限制)。另一方面,大数据应用系统往往构建在分布式环境之上,由于数据的分布以及涉及的数据量过大,传统的单一索引机制不能从根本上解决问题。因此,需要针对新型存储和分布式查询处理要求,设计相应的大数据索引结构以及操作算法。

6 结束语

高效的大数据存储与管理如果仅从软件体系结构考虑很难取得本质性突破,因为在大数据环境下内存与外存之间的I/O瓶颈很难克服。以PCM为代表的新型存储为大数据高效存储与实时处理提供了可能。研究适合高效大数据存储和管理的新型存储架构,借助创新的系统软件设计,改变大数据处理过程中对外存I/O的依赖,

有望克服目前大数据存储与管理中的性能瓶颈,并进一步带动大数据技术的未来发展。本文讨论了新型存储的特点以及现有大数据存储技术的局限性,在此基础上综述了基于新型存储的大数据存储管理领域的研究现状,最后给出了未来研究展望,以期能对新型存储与大数据管理的未来研究提供有价值的参考。

目前,由于非易失内存技术仍处于研发阶段,工业界还没有推出真正可用的新型存储系统,因此目前的研究还只能在新型存储模拟器^[49]上展开。随着非易失内存芯片工艺上的突破,预计几年内会出现可用的新型存储系统。届时,可以基于实际的平台开展理论和实验,对基于新型存储的理论研究成果进行验证。

参考文献:

- [1] 孟小峰, 慈祥. 大数据管理: 概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1): 146-169.
MENG X F, CI X. Big data management: concepts, techniques and challenges[J]. Journal of Computer Research and Development, 2013, 50(1): 146-169.
- [2] LU C, SHI J P, JIA J Y. Abnormal event detection at 150 FPS in MATLAB[C]// 2013 IEEE International Conference on Computer Vision(ICCV), December 1-8, 2013, Sydney, Australia. New Jersey: IEEE Press, 2013: 2720-2727.
- [3] CHENG M M, ZHANG Z M, LIN W Y, et al. BING: binarized normed gradients for objectness estimation at 300fps[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), June 23-28, 2014, Columbus, USA. New Jersey: IEEE Press, 2014: 3286-3293.
- [4] YAN J J, LEI Z, WEN L Y, et al. The fastest deformable part model for object

- detection[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), June 23-28, 2014, Columbus, USA. New Jersey: IEEE Press, 2014: 2497-2504
- [5] ROSS B. Girshick: fast R-CNN[C]// International Conference on Computer Vision(ICCV), December 13-16, 2015, Santiago, Chile. [S.l.:s.n.], 2015: 1440-1448
- [6] 申德荣, 于戈, 王习特, 等. 支持大数据管理的NoSQL系统研究综述[J]. 软件学报, 2013, 24(8): 1786-1803.
- SHEN D R, YU G, WANG X T, et al. Survey on NoSQL for management of big data[J]. Journal of Software, 2013, 24(8): 1786-1803.
- [7] 哈索·普拉特纳, 亚历山大·蔡尔. 内存数据管理(第2版)[M]. SAP, 译. 北京: 清华大学出版社, 2012.
- PLATTNER H, ZEIER A. In-memory data management(2nd edition)[M]. Translated by SAP. Beijing: Tsinghua University Press, 2012.
- [8] 王江涛, 赖文豫, 孟小峰. 闪存数据库: 现状、技术与展望[J]. 计算机学报, 2013, 36(8): 1549-1567.
- WANG W T, LAI W Y, MENG X F. Flash-based database: studies, techniques and forecasts[J]. Chinese Journal of Computers, 2013, 36(8): 1549-1567.
- [9] 吴章玲, 金培权, 岳丽华, 等. 基于PCM的大数据存储与管理研究综述[J]. 计算机研究与发展, 2015, 52(2): 343-361.
- WU Z L, JIN P Q, YUE L H, et al. A survey on PCM-based big data storage and management[J]. Journal of Computer Research and Development, 2015, 52(2): 343-361.
- [10] FREITAS R F, WILCKE W W. Storage-class memory: the next storage system technology[J]. IBM Journal of Research and Development, 2008, 52(4): 439-447
- [11] BURR G W, KURDI B N, SCOTT J C, et al. Overview of candidate device technologies for storage-class memory[J]. IBM Journal of Research and Development, 2008, 52(4): 449-464.
- [12] HAO X J, JIN P Q, YUE L H. Efficient storage of multi-sensor object-tracking data[J]. IEEE Transactions on Parallel and Distributed Systems, 2016, 27(10): 2881-2894.
- [13] HARTE T, BORTHAKUR D, DONG S Y, et al. Analysis of HDFS under HBase: a Facebook messages case study[C]//The 12th USENIX Conference on File and Storage Technologies(FAST), February 17-20, 2014, Santa Clara, USA. Berkeley: USENIX Association Berkeley, 2014: 199-212.
- [14] O'NEIL P E, CHENG E, GAWLICK D, et al. The log-structured merge-tree (LSM-Tree) [J]. Acta Informatica, 1996, 33(4): 351-385.
- [15] OUSTERHOUT J K, GOPALAN A, GUPTA A, et al. The RAMCloud storage system[J]. ACM Transactions on Computer Systems, 2015, 33(3): 7.
- [16] 金培权, 郝行军, 岳丽华. 面向新型存储的大数据存储架构与核心算法综述[J]. 计算机工程与科学, 2013, 35(10): 12-24.
- JIN P Q, HAO X J, YUE L H. A survey on storage architectures and core algorithms for big data management on new storages[J]. Computer Engineering and Science, 2013, 35(10): 12-24.
- [17] 张鸿斌, 范捷, 舒继武, 等. 基于相变存储器的存储系统与技术综述[J]. 计算机研究与发展, 2014, 51(8): 1647-1662.
- ZHANG H B, FAN J, SHU J W, et al. Summary of storage system and technology based on phase change memory[J]. Journal of Computer Research and Development, 2014, 51(8): 1647-1662.
- [18] VIGLAS S D. Write-limited sorts and joins for persistent memory[J]. Proceedings of the VLDB Endowment, 2014, 7(5): 413-424.
- [19] XIA F, JIANG D J, XIONG J, et al. A survey of phase change memory systems[J]. Journal of Computer Science and Technology, 2015, 30(1): 121-144.
- [20] 冒伟, 刘景宁, 童薇, 等. 基于相变存储器的存储技术研究综述[J]. 计算机学报, 2015, 38(5): 944-960.

- MAO W, LIU J N, TONG W, et al. A review of storage technology research based on phase change memory[J]. Chinese Journal of Computers, 2015, 38(5): 944-960.
- [21] 陈小刚, 许林海, 陈一峰, 等. 一种非易失性随机存储器及其操作方法: CN201110202834.7[P]. 2013-01-23.
- CHEN X G, XU L H, CHEN Y F, et al. A non-volatile random memory and its operations: CN201110202834.7[P]. 2013-01-23.
- [22] 缪向水, 余念念, 童浩, 等. 相变存储器: CN105247677A[P]. 2015-11-05.
- MIAO X S, YU N N, DONG H, et al. Phase change memory: CN105247677A[P]. 2015-11-05.
- [23] QURESHI M K, SRINIVASAN V, RIVERS J A. Scalable high performance main memory system using phase-change memory technology[C]//The 36th Annual International Symposium on Computer Architecture(ISCA), June 20-24, 2009, Austin, USA. New York: ACM Press, 2009: 24-33.
- [24] QURESHI M K, KARIDIS J P, FRANCESCHINI M, et al. Enhancing lifetime and security of PCM-based main memory with start-gap wear leveling[C]//The 42nd Annual IEEE/ACM International Symposium on Microarchitecture(MICRO), December 12-16, 2009, New York, USA. New Jersey: IEEE Press, 2009: 14-23.
- [25] RAMOS L E, BIANCHINI R. Exploiting phase-change memory in cooperative caches[C]// 2012 IEEE 24th International Symposium on Computer Architecture and High Performance Computing(SBAC-PAD), October 24-26, 2012, New York, USA. New Jersey: IEEE Press, 2012: 227-234.
- [26] BADAM A, PAI V S. SSDAlloc: hybrid SSD/RAM memory management made easy[C]// The 8th USENIX Conference on Networked Systems Design and Implementation(NSDI), March 30-April 1, 2011, Boston, USA. New York: ACM Press, 2011: 16-30.
- [27] 韩旭, 曹巍, 孟小峰. 使用固态硬盘管理主存KV数据库的虚拟内存[J]. 计算机科学与探索, 2011, 5(8): 686-694.
- HAN X, CAO W, MENG X F. Virtual memory management for main-memory KV database using solid state disk[J]. Journal of Frontiers of Computer Science & Technology, 2011, 5(8): 686-694.
- [28] O U Y A N G X Y, I S L A M N S, RAJACHANDRASEKAR R, et al. SSD-assisted hybrid memory to accelerate memcached over high performance networks[C]//The 41st International Conference on Parallel Processing(ICPP), September 10-13, 2012, Pittsburgh, USA. New Jersey: IEEE Press, 2012: 470-479.
- [29] 陈卓, 熊劲, 马灿, 等. 基于SSD的机群文件系统元数据存储系统[J]. 计算机研究与发展, 2012, 49(z1): 269-275.
- CHEN Z, XIONG J, MA C, et al. SSD-based metadata storage system for cluster file system[J]. Journal of Computer Research and Development, 2012, 49(z1): 269-275.
- [30] RAMOS L E, GORBATOV E, BIANCHINI R. Page placement in hybrid memory systems[C]// The International Conference on Supercomputing(ICS), May 31 - June 4, 2011, Tucson, USA. New York: ACM Press, 2011: 85-95.
- [31] WU Z L, JIN P Q, YUE L H. Efficient space management and wear leveling for pcm-based storage systems[C]// The 15th International Conference on Algorithms and Architectures for Parallel Processing(ICA3PP), November 18-20, 2015, Zhangjiajie, China. Berlin: Springer, 2015: 784-798.
- [32] WU Z L, JIN P Q, YANG C C, et al. Efficient memory management for NVM-based hybrid memory systems[J]. International Journal of Control and Automation, 2016, 9(1): 445-458.
- [33] WU S, JIANG D W, OOI B C, et al. Efficient B-tree based indexing for cloud data processing[J]. Proceedings of the

- VLDB Endowment, 2010, 3(1): 1207–1218.
- [34] WANG J B, WU S, GAO H, et al. Indexing multi-dimensional data in a cloud system[C]// The 2010 ACM SIGMOD International Conference on Management of Data(SIGMOD), June 6–10, 2010, Indianapolis, USA. New York: ACM Press, 2010: 591–602.
- [35] DING L L, QIAO B Y, WANG G R, et al. An efficient quad-tree based index structure for cloud data management[C]// International Conference on Web-Age Information Management (WAIM), September 14–16, 2011, Wuhan, China. Berlin: Springer Berlin Heidelberg, 2011: 238–250.
- [36] ZOU Y Q, LIU J, WANG S C, et al. CCIndex: a complementary clustering index on distributed ordered tables for multi-dimensional range queries[C]// IFIP International Conference on Network and Parallel Computing(NPC), April 28, 2011, Changsha, China. Berlin: Springer, 2010: 247–261.
- [37] NISHIMURA S, DAS S, AGRAWAL D, et al. MD-HBase: a scalable multi-dimensional data infrastructure for location aware services[C]//The 12th IEEE International Conference on Mobile Data Management(MDM), June 6–9, 2011, Lulea, Sweden. New Jersey: IEEE Press, 2011: 7–16
- [38] MA Y Z, RAO J, HU W S, et al. An efficient index for massive IOT data in cloud environment[C]// The 21st ACM International Conference on Information and Knowledge Management(CIKM), October 29–November 2, 2012, Maui, USA. New York: ACM Press, 2012: 2129–2133.
- [39] AHN J S, KANG D W, JUNG D, et al. μ^* -Tree: an ordered index structure for NAND flash memory with adaptive page layout scheme[J]. IEEE Transactions on Computers, 2013, 62(4): 784–797.
- [40] ATHANASSOULIS M, AILAMAKI A. BF-Tree: approximate tree indexing[J]. Proceedings of the VLDB Endowment, 2014, 7(14): 1881–1892.
- [41] ON S T, HU H B, LI Y, et al. Flash-optimized B+-tree[J]. Journal of Computer Science and Technology (JCST), 2010, 25(3): 509–522.
- [42] CUI K, JIN P Q, YUE L H. HashTree: a new hybrid index for flash disks[C]// The 12th International Asia-Pacific Web Conference(APWeb), April 6–8, 2010, Busan, Korea. New Jersey: IEEE Press, 2010: 45–51.
- [43] JIN P Q, YANG C C, JENSEN C S, et al. Read/Write-optimized tree indexing for solid state drives[J]. VLDB Journal, 2016, 25(5): 695–717.
- [44] CHEN S M, GIBBONS P B. SumanNath: rethinking database algorithms for phase change memory[C]//The 5th Biennial Conference on Innovative Data Systems Research(CIDR), January 9–12, 2011, Asilomar, USA. [S.l.:s.n.], 2011: 21–31.
- [45] CHEN S M, JIN Q. Persistent B+-Trees in non-volatile main memory[J]. Proceedings of the VLDB Endowment, 2015, 8(7): 786–797.
- [46] HU W W, LI G L, NI J C, et al. Bp-tree: a predictive B+-tree for reducing writes on phase change memory[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(10): 2368–2381.
- [47] CHI P, LEE W C, XIE Y. Making B+-tree efficient in PCM-based main memory[C]// International Symposium on Low Power Electronics and Design(ISLPED), August 11–13, 2014, La Jolla, USA. [S.l.:s.n.], 2014: 69–74.
- [48] LI L, JIN P Q, YANG C C, et al. Optimizing B+-Tree for PCM-based hybrid memory[C]//The 19th International Conference on Extending Database Technology(EDBT), March 15–18, 2016, Bordeaux, France. [S.l.:s.n.], 2016: 662–663.
- [49] ZHANG D Z, JIN P Q, WANG X L, et al. DPHSim: a flexible simulator for DRAM/

PCM-based hybrid memory[C]// Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data(APWeb/WAIM), July 7-9, 2017, Beijing, China. Berlin: Springer, 2017: 319-323.

作者简介



金培权 (1975-), 男, 博士, 中国科学技术大学计算机科学与技术学院和中国科学院电磁空间信息重点实验室副教授、硕士生导师, 目前主要从事大数据与数据库领域的研究工作, 近年来主持了20余项科研项目, 包括5项国家自然科学基金项目和2项“863”计划项目, 在VLDB Journal、TKDE、TPDS、ICDE、WWW等本领域著名期刊和会议上发表论文20余篇, 曾获中国科学院院长奖、NPC 2014最佳论文奖、DASFAA 2015最佳海报奖、NDBC 2012最佳论文提名奖以及NDBC 2011最佳系统演示奖。

收稿日期: 2017-05-10
基金项目: 国家自然科学基金资助项目 (No.61672479)
Foundation Item: The National Natural Science Foundation of China (No.61672479)